CORO-IMARO • AUVE
Autonomous Vehicles. Theoretical basis
CORO-SIP CORO-EPICO • STATES
Statistical Signal Processing and Estimation Theory

October 19, 2020

E. LE CARPENTIER

CENTRALE NANTES

Preface

Chapter contents

Context and objectives

In th everyday life, we often see the hazard intervention:

- it does not take always the same time to go from home to work;
- a smoker can have a cancer or not;
- the fishing is not always good.

Such phenomenons are said to be **random**, or **stochastic**. To quantify them leads us to use the **probability theory**.

- Let's consider again the nicotine addiction example. Let's imagine that the doctor does not trust his patient about the daily number of cigarettes. He asks to the medical analysis laboratory to measure the nicotine blood level. The probability theory provides some tools to quantify the stochastic link between the daily number of cigarettes and the nicotine blood level.

- Let's consider again the nicotine addiction example. Let's imagine that the doctor does not trust his patient about the daily number of cigarettes. He asks to the medical analysis laboratory to measure the nicotine blood level. The probability theory provides some tools to quantify the stochastic link between the daily number of cigarettes and the nicotine blood level.

- From this nicotine level, the doctor will be able to estimate the number of cigarettes. The **estimation theory** propose several solutions:
  - the most **likely** value;
  - the most **probable** value;
  - the **expected** value.

  These notions seem similar, but have different meanings in estimation theory, in which we will distinguish the classical estimation from the **Bayesian** estimation.

- Let's go back to the fishing example. The shoal of fish path depends of numerous factors, thus depends of the hazard; it is a **random signal**, also called a **stochastic process**. The problem is to estimate this path along time, by means of the sonar onboard. The Bayesian estimation remains tractable, if we are able to write a **Markovian** representation of the path, and it becomes the **Bayesian filtering**.

- Let's go back to the fishing example. The shoal of fish path depends of numerous factors, thus depends of the hazard; it is a **random signal**, also called a **stochastic process**. The problem is to estimate this path along time, by means of the sonar onboard. The Bayesian estimation remains tractable, if we are able to write a **Markovian** representation of the path, and it becomes the **Bayesian filtering**.
- We have numerous applications:
  - spacecrafts or mobile robots localization;
  - from handwriting to text;
  - DNA sequencing...

- At the end of this course:
  1. you will understand that there is no magic algorithm to solve such problems;
  2. you will be able to question the domain specialist to elaborate a Markovian model which link the hidden quantity to the observed data;
  3. you will know how to write a Bayesian filter (or a reasonable approximation) to estimate the hidden quantity from the observed data.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\top} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\top}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\top} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\top}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^\top A\,x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\,R^\top$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\mathsf{T}} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\mathsf{T}}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\mathsf{T}} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\mathsf{T}}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

chol

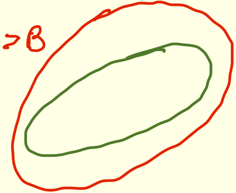
Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\intercal} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\intercal}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

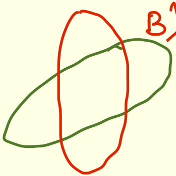
$\{x \mid x^T A x = k\}$ $\{x \mid x^T B x = k\}$

$A > B$



$A \not\geq B$

$B \not\geq A$



Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\intercal} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\intercal}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

To avoid the distributions theory which would be necessary for a rigorous approach to probability theory, we will use in this book the intuitive concept or Dirac delta function.

Requirements

It is advised to review some mathematical reminders on:

- differentiation;
- signal theory;
- (semi-)definite square matrix;
  - a symmetric matrix A is positive semi-definite if, for all vector x, $x^{\mathsf{T}} A\, x \geqslant 0$;
  - every positive semi-definite matrix A has square roots R such that $A = R\, R^{\mathsf{T}}$;
  - the order between positive semi-definite matrices $A \geqslant B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

To avoid the distributions theory which would be necessary for a rigorous approach to probability theory, we will use in this book the intuitive concept or Dirac delta function.

We will link it to the Kronecker delta.

- **The Kronecker delta** tests the equality of two discrete variables.
  For all $(x, \bar{x}) \in \mathbb{X}^2$ where $\mathbb{X}$ is a countable set:

$$\delta(x - \bar{x}) = \begin{cases} 1 & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{and then} \quad \sum_{x \in \mathbb{X}} \delta(x - \bar{x}) = 1 \tag{1}$$

- **The Kronecker delta** tests the equality of two discrete variables.
  For all $(x, \bar{x}) \in \mathbb{X}^2$ where $\mathbb{X}$ is a countable set:

$$\delta(x-\bar{x}) = \begin{cases} 1 & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{and then} \quad \sum_{x \in \mathbb{X}} \delta(x-\bar{x}) = 1 \tag{1}$$

- **The Dirac delta** tests the equality of two real-valued variables.
  For all $(x, \bar{x}) \in \mathbb{R}^2$ (so x and $\bar{x}$ are continuous variables):

$$\delta(x-\bar{x}) = \begin{cases} +\infty & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{under the condition} \quad \int_{\mathbb{R}} \delta(x-\bar{x}) \, \mathrm{d}\, x = 1 \tag{2}$$

  The function $x \mapsto \delta(x-\bar{x})$ is the Dirac delta located in $\bar{x}$.
  The function $x \mapsto \alpha\delta(x-\bar{x})$ is the Dirac delta with weight α located in $\bar{x}$ (its integral is α).

- **The Kronecker delta** tests the equality of two discrete variables.
For all $(x, \bar{x}) \in \mathbb{X}^2$ where $\mathbb{X}$ is a countable set:

$$\delta(x-\bar{x}) = \begin{cases} 1 & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{and then} \quad \sum_{x \in \mathbb{X}} \delta(x-\bar{x}) = 1 \tag{1}$$

- **The Dirac delta** tests the equality of two real-valued variables.
For all $(x, \bar{x}) \in \mathbb{R}^2$ (so x and $\bar{x}$ are continuous variables):

$$\delta(x-\bar{x}) = \begin{cases} +\infty & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{under the condition} \quad \int_{\mathbb{R}} \delta(x-\bar{x}) \, \mathrm{d}\, x = 1 \tag{2}$$

The function $x \mapsto \delta(x-\bar{x})$ is the Dirac delta located in $\bar{x}$.
The function $x \mapsto \alpha\delta(x-\bar{x})$ is the Dirac delta with weight α located in $\bar{x}$ (its integral is α).

- The Dirac delta permits to extend the derivative in case of discontinuities.
In a discontinuity point, the derivative exhibits a Dirac pulse whose weight is the jump magnitude.

- The Dirac delta and the Kronecker delta fulfill the **sifting property**:
  - for all function f from $\mathbb{X}$ and for all $\bar{x} \in \mathbb{X}$:

$$\sum_{x \in \mathbb{X}} f(x)\ \delta(x - \bar{x}) = f(\bar{x}) \tag{3}$$

  - for all function f from $\mathbb{R}$ and for all $\bar{x} \in \mathbb{R}$:

$$\int_{\mathbb{R}} f(x)\ \delta(x - \bar{x})\ \mathrm{d}\,x = f(\bar{x}) \tag{4}$$

- The Dirac delta and the Kronecker delta fulfill the **sifting property**:
  - for all function f from $\mathbb{X}$ and for all $\bar{x} \in \mathbb{X}$:

$$\sum_{x \in \mathbb{X}} f(x)\ \delta(x - \bar{x}) = f(\bar{x}) \tag{3}$$

  - for all function f from $\mathbb{R}$ and for all $\bar{x} \in \mathbb{R}$:

$$\int_{\mathbb{R}} f(x)\ \delta(x - \bar{x})\ \mathrm{d}\,x = f(\bar{x}) \tag{4}$$

- The Dirac delta and the Kronecker delta fulfill the **sifting property**:
  - for all function f from $\mathbb{X}$ and for all $\bar{x} \in \mathbb{X}$:
$$\sum_{x \in \mathbb{X}} f(x)\ \delta(x - \bar{x}) = f(\bar{x}) \tag{3}$$
  - for all function f from $\mathbb{R}$ and for all $\bar{x} \in \mathbb{R}$:
$$\int_{\mathbb{R}} f(x)\ \delta(x - \bar{x})\, \mathrm{d}\, x = f(\bar{x}) \tag{4}$$

- We generalize the delta to functions of a variable which contains continuous and discrete components [ref]:

$$\delta\left(\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{bmatrix}\right) = \delta(x_1 - \bar{x}_1)\ldots\delta(x_d - \bar{x}_d) \tag{5}$$

- We generalize the delta to functions of a variable which contains continuous and discrete components [ref]:

$$\delta\left(\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{bmatrix}\right) = \delta(x_1 - \bar{x}_1)\ldots\delta(x_d - \bar{x}_d) \tag{5}$$

- In this book, δ can designate an hybrid Dirac-Kronecker delta, function of a vector variable with discrete and continuous components.

  The unit integral (or sum), and the sifting property hold with this hybrid pulse:

  - we integrate with respect to continuous variables;
  - we sum with respect to discrete variables.

T1 Chapter 1

Probability theory

Chapter contents

T2 Probability space

Let Ω be the set of the students of the university.
A student ω is drawn at random in this population.
BSc, MSc, PhD are the students in bachelor degree, master degree, doctor degree.
FR, IN, CN... are the French, Indian, Chinese... students.

T2 Probability space

Let Ω be the set of the students of the university.
A student ω is drawn at random in this population.
BSc, MSc, PhD are the students in bachelor degree, master degree, doctor degree.
FR, IN, CN... are the French, Indian, Chinese... students.

➔ With the set theory language:

- Ω is the set;
- ω is an element ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some subsets ($\subset \Omega$).

With the probability theory language:

- Ω is the universe;
- ω is an elementary event ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some events ($\subset \Omega$).

T2 Probability space

Let Ω be the set of the students of the university.
A student ω is drawn at random in this population.
BSc, MSc, PhD are the students in bachelor degree, master degree, doctor degree.
FR, IN, CN... are the French, Indian, Chinese... students.

With the set theory language:

- Ω is the set;
- ω is an element ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some subsets ($\subset \Omega$).

With the probability theory language:

- Ω is the universe;
- ω is an elementary event ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some events ($\subset \Omega$).

T3 If all the students have the same chance to be drawn at random (equiprobability assumption), then the probability of the event $\Phi \subset \Omega$ is the proportion of elementary events of Ω belonging to Φ:

$$\mathrm{Prob}\,(\Phi) = \frac{\mathrm{Card}\,\Phi}{\mathrm{Card}\,\Omega}$$

T3 If all the students have the same chance to be drawn at random (equiprobability assumption), then the probability of the event $\Phi \subset \Omega$ is the proportion of elementary events of Ω belonging to Φ:

$$\mathrm{Prob}\,(\Phi) = \frac{\mathrm{Card}\,\Phi}{\mathrm{Card}\,\Omega}$$

- We just defined, in a natural way, in the case of a finite universe, by means of a counting interpretation, a **probability measure**:

$$\Phi \subset \Omega \longmapsto \mathrm{Prob}\,(\Phi)$$

The universe, with this probability measure, is a **probability space**.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\,(\emptyset) \leqslant \mathrm{Prob}\,(\Phi) \leqslant \mathrm{Prob}\,(\Omega) = 1$
- $\mathrm{Prob}\,(\Phi) + \mathrm{Prob}\left(\overline{\Phi}\right) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\,(\Phi_1 \cup \Phi_2) = \mathrm{Prob}\,(\Phi_1) + \mathrm{Prob}\,(\Phi_2) - \mathrm{Prob}\,(\Phi_1 \cap \Phi_2)$
- $\mathrm{Prob}\,(\bigcup_i \Phi_i) = \sum_i \mathrm{Prob}\,(\Phi_i)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).

1. The Kolmogorov axioms.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\,(\emptyset) \leqslant \mathrm{Prob}\,(\Phi) \leqslant \mathrm{Prob}\,(\Omega) = 1$
- $\mathrm{Prob}\,(\Phi) + \mathrm{Prob}\,(\overline{\Phi}) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\,(\Phi_1 \cup \Phi_2) = \mathrm{Prob}\,(\Phi_1) + \mathrm{Prob}\,(\Phi_2) - \mathrm{Prob}\,(\Phi_1 \cap \Phi_2)$
- $\mathrm{Prob}\,(\bigcup_i \Phi_i) = \sum_i \mathrm{Prob}\,(\Phi_i)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).

1. The Kolmogorov axioms.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\left(\emptyset\right) \leqslant \mathrm{Prob}\left(\Phi\right) \leqslant \mathrm{Prob}\left(\Omega\right) = 1$
- $\mathrm{Prob}\left(\Phi\right) + \mathrm{Prob}\left(\overline{\Phi}\right) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\left(\Phi_1 \cup \Phi_2\right) = \mathrm{Prob}\left(\Phi_1\right) + \mathrm{Prob}\left(\Phi_2\right) - \mathrm{Prob}\left(\Phi_1 \cap \Phi_2\right)$
- $\mathrm{Prob}\left(\bigcup_i \Phi_i\right) = \sum_i \mathrm{Prob}\left(\Phi_i\right)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\left(\Psi\right) = \sum_i \mathrm{Prob}\left(\Psi \cap \Phi_i\right)$ if the sets Φ_i form a partition of Ω (total probability law).

1. The Kolmogorov axioms.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}(\emptyset) \leqslant \mathrm{Prob}(\Phi) \leqslant \mathrm{Prob}(\Omega) = 1$
- $\mathrm{Prob}(\Phi) + \mathrm{Prob}(\overline{\Phi}) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}(\Phi_1 \cup \Phi_2) = \mathrm{Prob}(\Phi_1) + \mathrm{Prob}(\Phi_2) - \mathrm{Prob}(\Phi_1 \cap \Phi_2)$
- $\mathrm{Prob}(\bigcup_i \Phi_i) = \sum_i \mathrm{Prob}(\Phi_i)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}(\Psi) = \sum_i \mathrm{Prob}(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).

1. The Kolmogorov axioms.

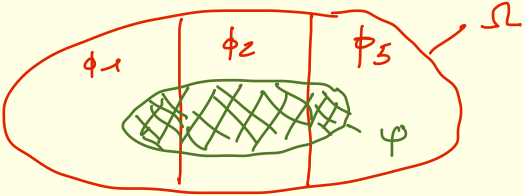
T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\left(\emptyset\right) \leqslant \mathrm{Prob}\left(\Phi\right) \leqslant \mathrm{Prob}\left(\Omega\right) = 1$
- $\mathrm{Prob}\left(\Phi\right) + \mathrm{Prob}\left(\overline{\Phi}\right) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\left(\Phi_1 \cup \Phi_2\right) = \mathrm{Prob}\left(\Phi_1\right) + \mathrm{Prob}\left(\Phi_2\right) - \mathrm{Prob}\left(\Phi_1 \cap \Phi_2\right)$
- $\mathrm{Prob}\left(\bigcup_i \Phi_i\right) = \sum_i \mathrm{Prob}\left(\Phi_i\right)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\left(\Psi\right) = \sum_i \mathrm{Prob}\left(\Psi \cap \Phi_i\right)$ if the sets Φ_i form a partition of Ω (total probability law).

1. The Kolmogorov axioms.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\,(\emptyset) \leqslant \mathrm{Prob}\,(\Phi) \leqslant \mathrm{Prob}\,(\Omega) = 1$
- $\mathrm{Prob}\,(\Phi) + \mathrm{Prob}\,(\overline{\Phi}) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\,(\Phi_1 \cup \Phi_2) = \mathrm{Prob}\,(\Phi_1) + \mathrm{Prob}\,(\Phi_2) - \mathrm{Prob}\,(\Phi_1 \cap \Phi_2)$
- $\mathrm{Prob}\,(\bigcup_i \Phi_i) = \sum_i \mathrm{Prob}\,(\Phi_i)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).



1. The Kolmogorov axioms.

T4 This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:[1]

- $0 = \mathrm{Prob}\,(\emptyset) \leqslant \mathrm{Prob}\,(\Phi) \leqslant \mathrm{Prob}\,(\Omega) = 1$
- $\mathrm{Prob}\,(\Phi) + \mathrm{Prob}\,(\overline{\Phi}) = 1$ (where $\overline{\Phi}$ is the complementary set of Φ in Ω);
- $\mathrm{Prob}\,(\Phi_1 \cup \Phi_2) = \mathrm{Prob}\,(\Phi_1) + \mathrm{Prob}\,(\Phi_2) - \mathrm{Prob}\,(\Phi_1 \cap \Phi_2)$
- $\mathrm{Prob}\,(\bigcup_i \Phi_i) = \sum_i \mathrm{Prob}\,(\Phi_i)$ if the Φ_i sets are disjoint;
- $\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).

➔ We will prefer to write the total probability law by means of the **conditional probabilities.**

1. The Kolmogorov axioms.

given

T5 $\mathrm{Prob}\,(\Phi \mid \Psi)$ is the **conditional** probability to belong to Φ under the assumption to belong to Ψ.
By means of the frequentist interpretation, it is obviously:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Phi \cap \Psi)}{\mathrm{Prob}\,(\Psi)} \tag{1.1}$$

T5 $\mathrm{Prob}\,(\Phi \mid \Psi)$ is the **conditional** probability to belong to Φ under the assumption to belong to Ψ.
By means of the frequentist interpretation, it is obviously:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Phi \cap \Psi)}{\mathrm{Prob}\,(\Psi)} \tag{1.1}$$

- $\mathrm{Prob}\,(\mathsf{FR} \mid \mathsf{MSc})$ is the probability that a master student is French;
  $\mathrm{Prob}\,(\mathsf{MSc} \mid \mathsf{FR})$ is the probability that a French student is registered for a master degree.

Both probabilities are in general different.

T5 $\mathrm{Prob}\,(\Phi \mid \Psi)$ is the **conditional** probability to belong to Φ under the assumption to belong to Ψ.
By means of the frequentist interpretation, it is obviously:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Phi \cap \Psi)}{\mathrm{Prob}\,(\Psi)} \tag{1.1}$$

- $\mathrm{Prob}\,(\text{FR} \mid \text{MSc})$ is the probability that a master student is French;
  $\mathrm{Prob}\,(\text{MSc} \mid \text{FR})$ is the probability that a French student is registered for a master degree.

⚠ Both probabilities are in general different.

- For all event Ψ, the conditional probability measure $\Phi \mapsto \mathrm{Prob}\,(\Phi \mid \Psi)$ fulfills the same properties than a probability measure; for example, $\mathrm{Prob}\left(\overline{\Phi} \mid \Psi\right) = 1 - \mathrm{Prob}\,(\Phi \mid \Psi)$.

⚠ There is no general relation between $\mathrm{Prob}\left(\Phi \mid \overline{\Psi}\right)$ and $\mathrm{Prob}\,(\Phi \mid \Psi)$.

T5 $\mathrm{Prob}\,(\Phi \mid \Psi)$ is the **conditional** probability to belong to Φ under the assumption to belong to Ψ.
By means of the frequentist interpretation, it is obviously:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Phi \cap \Psi)}{\mathrm{Prob}\,(\Psi)} \tag{1.1}$$

- $\mathrm{Prob}\,(\text{FR} \mid \text{MSc})$ is the probability that a master student is French;
  $\mathrm{Prob}\,(\text{MSc} \mid \text{FR})$ is the probability that a French student is registered for a master degree.

⚠ Both probabilities are in general different.

- For all event Ψ, the conditional probability measure $\Phi \mapsto \mathrm{Prob}\,(\Phi \mid \Psi)$ fulfills the same properties than a probability measure; for example, $\mathrm{Prob}\,(\overline{\Phi} \mid \Psi) = 1 - \mathrm{Prob}\,(\Phi \mid \Psi)$.

⚠ There is no general relation between $\mathrm{Prob}\,(\Phi \mid \overline{\Psi})$ and $\mathrm{Prob}\,(\Phi \mid \Psi)$.

In the everyday life, there is often some ambiguity in the meaning of probabilities, since the universe, or the conditioning event, is not properly defined.

For example, a disease can be rare in the whole population, but frequent for the persons who are exposed to the agent which provokes this disease.

T6 The events Φ and Ψ are **independent** if $\mathrm{Prob}\,(\Phi \cap \Psi) = \mathrm{Prob}\,(\Phi)\ \mathrm{Prob}\,(\Psi)$,
that is if $\mathrm{Prob}\,(\Phi \mid \Psi) = \mathrm{Prob}\,(\Phi)$,
that is if $\mathrm{Prob}\,(\Psi \mid \Phi) = \mathrm{Prob}\,(\Psi)$.

T6 The events Φ and Ψ are **independent** if $\mathrm{Prob}\,(\Phi \cap \Psi) = \mathrm{Prob}\,(\Phi)\ \mathrm{Prob}\,(\Psi)$,
that is if $\mathrm{Prob}\,(\Phi \mid \Psi) = \mathrm{Prob}\,(\Phi)$,
that is if $\mathrm{Prob}\,(\Psi \mid \Phi) = \mathrm{Prob}\,(\Psi)$.

- If FR and MSc are independent:
  the proportion of French students if the same, considering either the MSc students only or the whole university;
  the proportion of MSc students is the same, considering either the French students only or the whole university.

T7 The total probability law and the **Bayes law** are:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Psi \mid \Phi)\ \mathrm{Prob}\,(\Phi)}{\mathrm{Prob}\,(\Psi)} \qquad \text{(Bayes law)} \tag{1.2}$$

$$\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \mid \Phi_i)\ \mathrm{Prob}\,(\Phi_i) \qquad \text{(if the sets } \Phi_i \text{ form a partition of } \Omega\text{, total probability law)} \tag{1.3}$$

T7 The total probability law and the **Bayes law** are:

$$\mathrm{Prob}\,(\Phi \mid \Psi) = \frac{\mathrm{Prob}\,(\Psi \mid \Phi)\ \mathrm{Prob}\,(\Phi)}{\mathrm{Prob}\,(\Psi)} \qquad \text{(Bayes law)} \tag{1.2}$$

$$\mathrm{Prob}\,(\Psi) = \sum_i \mathrm{Prob}\,(\Psi \mid \Phi_i)\ \mathrm{Prob}\,(\Phi_i) \qquad \text{(if the sets } \Phi_i \text{ form a partition of } \Omega\text{, total probability law)} \tag{1.3}$$

T8

▷ **Exercise 1.** In the university, 60% of the students are registered in a BSc degree, 30% of them in a MSc degree, 10% in a PhD degree.
30% of the BSc students, 40% of the MSc ones and 20% of the PhD ones are Chinese.

a) What is the percentage of Chinese students in the university?

b) What is the percentage of MSc students among the Chinese students?

Prob(BSc) = 0.6 — —

Prob(CN|BSc) = 0.3 — —

a) Prob(CN) = 32%

b) Prob(MSc | CN) = 37.5 %

T9 ◁ ▷ **Evaluation, module 1** (screening test). The **prevalence** of a disease is the probability to be sick. The quality of a screening test is measured through the **sensibility** and the **specificity**.

Prevalence	Probability to be sick	
Sensibility	Probability that a sick person has a positive test	
Specificity	Probability that a healthy person has a negative test	

When the test is done, we can calculate the **predictive values**.

Positive predictive value	Probability to be sick if the test is positive	
Negative predictive value	Probability to be healthy is the test is negative	

a) M is the set of sick persons ($\overline{M}$ the healthy persons).
T is the set of persons whose test is positive ($\overline{T}$ the persons with negative test).
Complete the tables above with the suitable probabilities (for example, $\mathrm{Prob}\,(M \mid T)$, etc.)

b) For a disease with prevalence 0.1%, and a test with sensibility 96% et specificity 98%, calculate the predictive values.

T10 As a complement to this evaluation:

$M \cap T$ are the "true positives",
$\overline{M} \cap \overline{T}$ are the "true negatives",
$\overline{M} \cap T$ are the "false positives",
$M \cap \overline{T}$ are the "false negatives".

The test accuracy is the probability that the test provides a correct result, that is $\mathrm{Prob}\left((M \cap T) \cup (\overline{M} \cap \overline{T})\right)$.

◁ ▷ **Exercise 2.** Give the accuracy in function of the prevalence, the sensitivity and the specificity.

T11 Random variable

Definition

A random variable (r.v.) X is a function from a probability space to a given set $\mathbb{X}$:

$$\begin{aligned} \mathrm{X} : \Omega &\longrightarrow \mathbb{X} \\ \omega &\longmapsto \mathrm{X}(\omega) \end{aligned}$$

For example:

- Ω is the university, that is the set of students;
- the student ω is drawn at random in Ω;
- let x be the triplet made with his mean mark, his rank, the prepared diploma (we suppose that a student prepares exactly one diploma, $\{\text{BSc}, \text{MSc}, \text{PhD}\}$ is a partition of Ω).

We have defined a r.v. X from Ω to $\mathbb{X} = \mathbb{R}^{+} \times \mathbb{N} \times \{\text{“BSc”}, \text{“MSc”}, \text{“PhD”}\}$, such that $\mathrm{X}(\omega) = x$.
We say that X took the value x, or that x is a realization of X.

We will use upper case letters for the r.v., and lower case letters for the realizations.

T11 # Random variable

Definition

A random variable (r.v.) X is a function from a probability space to a given set $\mathbb{X}$:

$$\begin{aligned} \mathrm{X} : \Omega &\longrightarrow \mathbb{X} \\ \omega &\longmapsto \mathrm{X}(\omega) \end{aligned}$$

For example:

- Ω is the university, that is the set of students;
- the student ω is drawn at random in Ω;
- let x be the triplet made with his mean mark, his rank, the prepared diploma (we suppose that a student prepares exactly one diploma, $\{\text{BSc}, \text{MSc}, \text{PhD}\}$ is a partition of Ω).

We have defined a r.v. X from Ω to $\mathbb{X} = \mathbb{R}^+ \times \mathbb{N} \times \{\text{“BSc”}, \text{“MSc”}, \text{“PhD”}\}$, such that $\mathrm{X}(\omega) = x$.
We say that X took the value x, or that x is a realization of X.

We will use upper case letters for the r.v., and lower case letters for the realizations.

T11 # Random variable

Definition

A random variable (r.v.) X is a function from a probability space to a given set $\mathbb{X}$:

$$\begin{aligned} \mathrm{X} : \Omega &\longrightarrow \mathbb{X} \\ \omega &\longmapsto \mathrm{X}(\omega) \end{aligned}$$

For example:

- Ω is the university, that is the set of students;
- the student ω is drawn at random in Ω;
- let x be the triplet made with his mean mark, his rank, the prepared diploma (we suppose that a student prepares exactly one diploma, $\{\text{BSc}, \text{MSc}, \text{PhD}\}$ is a partition of Ω).

We have defined a r.v. X from Ω to $\mathbb{X} = \mathbb{R}^+ \times \mathbb{N} \times \{\text{“BSc”}, \text{“MSc”}, \text{“PhD”}\}$, such that $\mathrm{X}(\omega) = x$.
We say that X took the value x, or that x is a realization of X.

We will use upper case letters for the r.v., and lower case letters for the realizations.

T12 **Probability distribution**

Let $\mathbb{A}$ be a subset of $\mathbb{X}$.
The event $\boxed{\{\omega \in \Omega \mid \mathrm{X}(\omega) \in \mathbb{A}\} \text{ is noted } \mathrm{X} \in \mathbb{A}.}$
In particular, $\{\omega \in \Omega \mid \mathrm{X}(\omega) = x\}$ is noted $\mathrm{X} = x$.

$x \in [70kg, 80kg]$

T12 Probability distribution

Let $\mathbb{A}$ be a subset of $\mathbb{X}$.
The event $\{\omega \in \Omega \mid \mathrm{X}(\omega) \in \mathbb{A}\}$ is noted $\mathrm{X} \in \mathbb{A}$.
In particular, $\{\omega \in \Omega \mid \mathrm{X}(\omega) = x\}$ is noted $\mathrm{X} = x$.

- From the probability measure of Ω, we derive the **probability distribution** of X:

$$\mathbb{A} \subset \mathbb{X} \longmapsto \mathrm{Prob}\,(\mathrm{X} \in \mathbb{A})$$

T13 **Exercise 3.** Pierre, Paul and Jacques left the nursery school, Pierre and Paul with 3 marbles in the pocket, Jacques with 4 ones. One of them was victim of an extorsion, but his name was not revealed by the police. There are many assumptions at school about the victim identity, but, finally, people thinks that the 3 boys have the same chance of being attacked. What is the probability distribution of the loot marbles amount?

$\Omega = \{Pierre, Paul, Jacques\}$

$Prob(\{Pierre\}) = Prob(\{Paul\}) = Prob(\{Jacques\}) = \frac{1}{3}$

X number of marbles

$Prob(X=3) = Prob(\{\omega \mid X(\omega)=3\}) = Prob(\{Pierre, Paul\}) = \frac{2}{3}$

$Prob(X=4) = \frac{1}{3}$

$Prob(X=n) = 0$ if $n \notin \{3,4\}$

T14 ◁ ▷ **Exercise 4.** The indicator function $\mathbf{1}_\Phi$ of a subset Φ of Ω takes the value 1 on this subset, 0 outside. Write the probability distribution of this r.v. which takes its value in $\{0, 1\}$ by means of the probability of Φ.

$$\text{Prob}(1_\phi = 1) = \text{Prob}(\underbrace{\{\omega \mid 1_\phi(\omega) = 1\}}_{\phi}) = \text{Prob}(\phi)$$

$$\text{Prob}(1_\phi = 0) = 1 - \text{Prob}(\phi)$$

T15 **Probability mass function**

If $\mathbb{X}$ is countable, we say that x is a **discrete r.v.**.
Its distribution is described by its **probability mass function** (PMF):

$$p_{\mathrm{x}} : x \in \mathbb{X} \longmapsto \mathrm{Prob}\,(\mathrm{x} = x)$$

$p_x(x) = \mathrm{Prob}(X = x)$

Thus, for all $\mathbb{A} \in \mathbb{X}$ such that the sum below is meaningful:

$$\mathrm{Prob}\,(\mathrm{x} \in \mathbb{A}) = \sum_{x \in \mathbb{A}} p_{\mathrm{x}}(x) \tag{1.4}$$

T15 Probability mass function

If $\mathbb{X}$ is countable, we say that x is a **discrete r.v.**.
Its distribution is described by its **probability mass function** (PMF):

$$p_{\mathrm{x}} : x \in \mathbb{X} \longmapsto \operatorname{Prob}(\mathrm{x} = x)$$

Thus, for all $\mathbb{A} \in \mathbb{X}$ such that the sum below is meaningful:

$$\operatorname{Prob}(\mathrm{x} \in \mathbb{A}) = \sum_{x \in \mathbb{A}} p_{\mathrm{x}}(x) \tag{1.4}$$

- For example, let's throw a die; the r.v. corresponds to the number of the upper side, which takes its value in $\mathbb{X} = \{1, 2, 3, 4, 5, 6\}$; for all $x \in \mathbb{X}$, $p_{\mathrm{x}}(x) = \frac{1}{6}$; then $\operatorname{Prob}(\mathrm{x} \in \{1, 2\}) = \frac{2}{6}$.

T16 Let's define the r.v. DIPL which is the name of the prepared diploma (the "label" of the corresponding event):

$$\text{DIPL}(\omega) = \begin{cases} \text{"PhD"} & \text{if } \omega \in \text{PhD} \\ \text{"MSc"} & \text{if } \omega \in \text{MSc} \\ \text{"BSc"} & \text{if } \omega \in \text{BSc} \end{cases}$$

Then, $p_{\text{DIPL}}(\text{"PhD"})$ is the proportion of students who prepare a PhD.

$$p_{\text{DIPL}}(\text{"PhD"}) = \underbrace{\text{Prob}\left(\text{DIPL} = \text{"PhD"}\right) = \text{Prob}\left(\text{PhD}\right)}$$

T17 # Probability density function

If $\mathbb{X} = \mathbb{R}^d$, we say that X is a **continuous r.v.**
This book will reduce to the case where there exists a **probability density function** (PDF) $p_{\mathrm{X}} : \mathbb{X} \to \mathbb{R}^+$ such that for all $\mathbb{A} \in \mathbb{X}$ such that the integral below is meaningful:[2]

$$\mathrm{Prob}\left(\underbrace{\mathrm{X} \in \mathbb{A}}\right) = \int_{\mathbb{A}} \underbrace{p_{\mathrm{X}}(x)}\,\mathrm{d}\,x \tag{1.5}$$

Intuitively, $p_{\mathrm{X}}(x)\,\mathrm{d}\,x$ is the probability that X belongs to an hypervolume $\mathrm{d}\,x$ around x.

2. In the formula (1.5), if $d > 1$, the integral is a multiple integral and $\mathrm{d}\,x$ means the hypervolume $\mathrm{d}\,x_1 \ldots \mathrm{d}\,x_d$.

T17 Probability density function

If $\mathbb{X} = \mathbb{R}^d$, we say that X is a **continuous r.v.**
This book will reduce to the case where there exists a **probability density function** (PDF) $p_{\mathrm{X}} : \mathbb{X} \to \mathbb{R}^+$ such that for all $\mathbb{A} \in \mathbb{X}$ such that the integral below is meaningful:[2]

$$\mathrm{Prob}\,(\mathrm{X} \in \mathbb{A}) = \int_{\mathbb{A}} p_{\mathrm{X}}(x)\,\mathrm{d}\,x \tag{1.5}$$

Intuitively, $p_{\mathrm{X}}(x)\,\mathrm{d}\,x$ is the probability that X belongs to an hypervolume $\mathrm{d}\,x$ around x.

- The **cumulative distribution function** F_{X} (CDF) is defined, for all $x \in \mathbb{R}^d$, by:[3]

$$F_{\mathrm{X}}(x) = \mathrm{Prob}\,(\mathrm{X}_1 \leqslant x_1 \text{ and } \ldots \text{ and } \mathrm{X}_d \leqslant x_d) \qquad F_X(x) = \mathrm{Prob}(X \leq x) \tag{1.6}$$

Necessarily, the PDF is the derivative of the CDF:

$$p_{\mathrm{X}} = \frac{\partial^d F_{\mathrm{X}}}{\partial x_1 \ldots \partial x_d} \qquad p_X(x) = \frac{d F_X}{dx}(x) \tag{1.7}$$

2. In the formula (1.5), if $d > 1$, the integral is a multiple integral and $\mathrm{d}\,x$ means the hypervolume $\mathrm{d}\,x_1 \ldots \mathrm{d}\,x_d$.
3. In the formula (1.6), X_i (resp. x_i) means the ith scalar component of X (resp. x).

T18 Generalization

PMF and PDF can be extended to the case where the r.v. x is hybrid, that is to say with discrete and continuous components: $\mathrm{x} = \begin{bmatrix} \mathrm{x}_{\textbf{cont}} \\ \mathrm{x}_{\textbf{disc}} \end{bmatrix}$.

T18 Generalization

PMF and PDF can be extended to the case where the r.v. x is hybrid, that is to say with discrete and continuous components: $\mathrm{x} = \left[\begin{smallmatrix} \mathrm{x}_{\text{cont}} \\ \mathrm{x}_{\text{disc}} \end{smallmatrix} \right]$.

- Then, for all $\mathbb{A} \subset \mathbb{X}$ such that the formula below is meaningful:

$$\operatorname{Prob}\left(\mathrm{x} \in \mathbb{A}\right) = \underbrace{\sum_{x_{\text{disc}}} \int}_{\mathbb{A}} p_{\mathrm{x}}\left(\left[\begin{smallmatrix} x_{\text{cont}} \\ x_{\text{disc}} \end{smallmatrix} \right]\right) \mathrm{d}\, x_{\text{cont}} \tag{1.8}$$

T18 ## Generalization

PMF and PDF can be extended to the case where the r.v. x is hybrid, that is to say with discrete and continuous components: $\mathrm{x} = \left[\begin{smallmatrix} \mathrm{x}_{\mathsf{cont}} \\ \mathrm{x}_{\mathsf{disc}} \end{smallmatrix}\right]$.

- Then, for all $\mathbb{A} \subset \mathbb{X}$ such that the formula below is meaningful:

$$\mathrm{Prob}\left(\mathrm{x} \in \mathbb{A}\right) = \underbrace{\sum_{x_{\mathsf{disc}}} \int}_{\mathbb{A}} p_{\mathrm{x}}\left(\left[\begin{smallmatrix} x_{\mathsf{cont}} \\ x_{\mathsf{disc}} \end{smallmatrix}\right]\right) \mathrm{d}\, x_{\mathsf{cont}} \tag{1.8}$$

In this document, most of the formulae are written for continuous r.v., but are easily transposed to the discrete or hybrid cases; just remind that we integrate with respect to continuous variables, we sum with respect to discrete variables.

T19 Remind that:

$$p_{\mathrm{x}}(x) \geqslant 0 \qquad \underbrace{\sum_{x_{\mathsf{disc}}} \int}_{\mathbb{X}} p_{\mathrm{x}}\left(\left[\begin{smallmatrix} x_{\mathsf{cont}} \\ x_{\mathsf{disc}} \end{smallmatrix}\right]\right) \mathrm{d}\, x_{\mathsf{cont}} = 1$$

T20 # In practice

We rarely know the exact probability distribution.
How to obtain an approximate PMF or PDF?

T20 **In practice**

We rarely know the exact probability distribution.
How to obtain an approximate PMF or PDF?

- **Using situation analysis (physical law)** For a correct die, the probability of the visible number is $\frac{1}{6}$.

T20 In practice

We rarely know the exact probability distribution.
How to obtain an approximate PMF or PDF?

- **Using situation analysis (physical law)** For a correct die, the probability of the visible number is $\frac{1}{6}$.

- **With a model** A parameterized distribution is assumed, then we estimate the parameters.

T20 # In practice

We rarely know the exact probability distribution.
How to obtain an approximate PMF or PDF?

- **Using situation analysis (physical law)** For a correct die, the probability of the visible number is $\frac{1}{6}$.

- **With a model** A parameterized distribution is assumed, then we estimate the parameters.

- **With a descriptive approach (statistics, discrete r.v. case)** We observe n_r realizations of the r.v., and we measure, for each possible value, the proportion of trials which gave this value.

  The PMF is the limit of these proportions when n_r tends to ∞.[4]

4. Matlab. To plot of a PMF estimated over a scalar valued population in the vector x.
`ux = unique(x); px = hist(x, ux)/length(x); stem(ux, px)`

T21 **With a descriptive approch (statistics, continuous r.v. case)** We assume that there is an infinite number of students in the university.

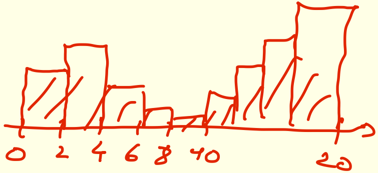
A finite number n_r of students is drawn at random, and we plot the **normalized histogram** of their average mark:

- the axis between the minimal mark and the maximal one is separated into n_b intervals with the same width;
- above each interval, a rectangle (bin) whose support is the interval and the area is equal to the proportion of students whose average mark belongs to the interval is drawn.

Thus, the histogram area is 1.

The PDF is the limit histogram when n_r, n_b and $\frac{n_\mathrm{r}}{n_\mathrm{b}}$ tend to ∞.[5]

We can choose $n_\mathrm{b} = \lfloor \sqrt{n_\mathrm{r}} \rfloor$: the number of intervals is the square root of the population size. See [ref] for other choices.



5. Matlab. To plot a normalized histogram with n_b bins of a scalar valued population in the vector x.

```
[n,b] = hist(x,nb); bar(b, n/(b(2)-b(1))/sum(n), 1)
```

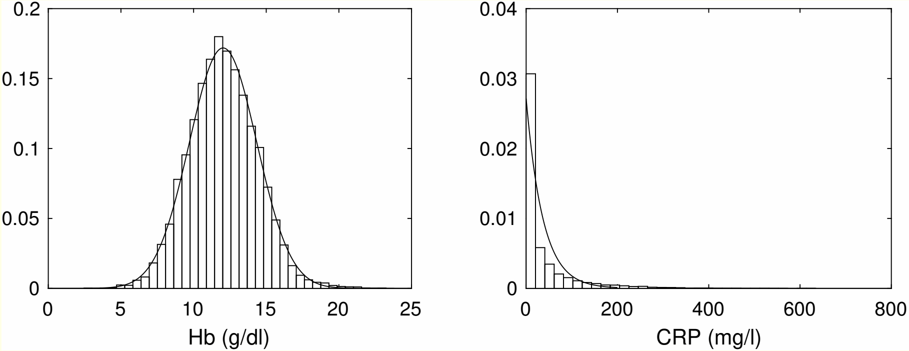


Figure 1.1: Histogram hemoglobin (Hb) and C-reactive protein (CRP). Gaussian model for Hb. Tempted exponential model for CRP

T22 The figure 1.1 displays the histogram of the hemoglobin (Hb) and the C-reactive protein (CRP) for patients of an hospital in Paris.
The Gauss distribution model (page 50) seems to fit in the Hb case.
The exponential distribution model (page 52), tried on the CRP, does not fit.

T23 Support

The **support** $\mathrm{S}(\mathrm{x})$ of the r.v. x is the set of value on which the PDF (or PMF) is strictly positive; this is the set of values that a r.v. can take:

$$\mathrm{S}(\mathrm{x}) = \{x \in \mathbb{X} \mid p_{\mathrm{x}}(x) > 0\} \tag{1.9}$$

T23 Support

The **support** $\mathrm{S}(\mathrm{x})$ of the r.v. x is the set of value on which the PDF (or PMF) is strictly positive; this is the set of values that a r.v. can take:

$$\mathrm{S}(\mathrm{x}) = \{x \in \mathbb{X} \mid p_{\mathrm{x}}(x) > 0\} \tag{1.9}$$

Let g be a function of a variable in $\mathbb{X}$.

For all realization x of x, we associate $g(x)$, realization of the r.v. $g \circ \mathrm{x}$; this new r.v. is noted $g(\mathrm{x})$.

Let's note that we need to define g only on the support $\mathrm{S}(\mathrm{x})$, since all realizations of x belong to this support; so, we can shorten the definition of a function.

Rather than writing: "let g be the function defined, for all $x \in \mathrm{S}(\mathrm{x})$, by $g(x) = x^2$", we write:
"let g be the function defined by $g(\mathrm{x}) = \mathrm{x}^2$".

T23 Support

The **support** $\mathrm{S}(\mathrm{x})$ of the r.v. x is the set of value on which the PDF (or PMF) is strictly positive; this is the set of values that a r.v. can take:

$$\mathrm{S}(\mathrm{x}) = \{x \in \mathbb{X} \mid p_{\mathrm{x}}(x) > 0\} \tag{1.9}$$

Let g be a function of a variable in $\mathbb{X}$.

For all realization x of x, we associate $g(x)$, realization of the r.v. $g \circ \mathrm{x}$; this new r.v. is noted $g(\mathrm{x})$.

Let's note that we need to define g only on the support $\mathrm{S}(\mathrm{x})$, since all realizations of x belong to this support; so, we can shorten the definition of a function.

Rather than writing: "let g be the function defined, for all $x \in \mathrm{S}(\mathrm{x})$, by $g(x) = x^2$", we write:
"let g be the function defined by $g(\mathrm{x}) = \mathrm{x}^2$".

- Let's see the r.v. $\begin{bmatrix} \text{WEIGHT} \\ \text{SIZE} \end{bmatrix}$. We obtain a new r.v., the "body mass index", with $\text{BMI} = \frac{\text{WEIGHT}}{\text{SIZE}^2}$, used to quantify the stoutness of a person. We don't have to define the BMI for a 0 size, since nobody has 0 size.

T24 ◁ ▷ **Evaluation, module 2.** Among the functions below, defined over $\mathbb{Z}$ or $\mathbb{R}$, give those which can be considered as a PMF or a PDF.

a) Is this function (over $\mathbb{Z}$) a PMF? $p_{\mathrm{x}}(x) = \begin{cases} 2 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$

b) Is this function (over $\mathbb{Z}$) a PMF? $p_{\mathrm{x}}(x) = \begin{cases} 2 & \text{if } x = 0 \\ -1 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$

c) Is this function (over $\mathbb{Z}$) a PMF? $p_{\mathrm{x}}(x) = \begin{cases} \frac{1}{3} & \text{if } x = -1 \text{ ou } x = 0 \text{ ou } x = 1 \\ 0 & \text{otherwise} \end{cases}$

d) For which α is this function (over $\mathbb{R}$) a PDF? $p_{\mathrm{x}}(x) = \begin{cases} \alpha\, x & \text{if } 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$

T25 # Expectation, mean, variance

Mathematical expectation

The **expectation** operator, for all r.v. x and all function g of a variable in $\mathbb{X}$ with numerical value, such that the integral below converges, gives a value noted $\mathrm{E}\left(g(\mathrm{x})\right)$ defined by:[6]

$$\mathrm{E}\left(g(\mathrm{x})\right) = \int_{\mathbb{X}} g(x)\, p_{\mathrm{x}}(x)\ \mathrm{d}\, x \tag{1.10}$$

6. In the discrete valued case:

$$\mathrm{E}\left(g(\mathrm{x})\right) = \sum_{x\in\mathbb{X}} g(x)\, p_{\mathrm{x}}(x) = \sum_{x\in\mathbb{X}} g(x)\ \mathrm{Prob}\left(\mathrm{x} = x\right)$$

T25 Expectation, mean, variance

Mathematical expectation

The **expectation** operator, for all r.v. x and all function g of a variable in $\mathbb{X}$ with numerical value, such that the integral below converges, gives a value noted $\mathrm{E}\left(g(\mathrm{x})\right)$ defined by:[6]

$$\mathrm{E}\left(g(\mathrm{x})\right) = \int_{\mathbb{X}} g(x)\, p_{\mathrm{x}}(x)\, \mathrm{d}\, x \tag{1.10}$$

- Its purpose is to calculate a mean value.

6. In the discrete valued case:

$$\mathrm{E}\left(g(\mathrm{x})\right) = \sum_{x \in \mathbb{X}} g(x)\, p_{\mathrm{x}}(x) = \sum_{x \in \mathbb{X}} g(x)\ \mathrm{Prob}\left(\mathrm{x} = x\right)$$

T26 Given the distribution of $\begin{bmatrix} \text{WEIGHT} \\ \text{SIZE} \end{bmatrix}$, we can calculate the BMI mean value.[7]

$$BMI = g\left(\begin{bmatrix} w \\ s \end{bmatrix}\right) = \frac{w}{s^2}$$

$$E(BMI) = \iint g\left(\begin{bmatrix} w \\ s \end{bmatrix}\right) p_{\begin{bmatrix} w \\ s \end{bmatrix}}\left(\begin{bmatrix} w \\ s \end{bmatrix}\right) dw\, ds$$

$$E(BMI) = \int b\, p_{BMI}(b)\, db$$

7. This implies that we obtain the same result by means of the calculation below (law of the unconscious statistician):

$$\mathrm{E}\,(g(\mathrm{x})) = \int_{\mathbb{Y}} y\, p_{g(\mathrm{x})}(y)\ \mathrm{d}\,y$$

T27 In linear transforms, the expectation operator fulfills the property below:[8]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{x})\right) = \left(\mathrm{E}\left(g(\mathrm{x})\right)\right)^{\intercal} \tag{1.11}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{x}) & g_{12}(\mathrm{x}) \\ g_{21}(\mathrm{x}) & g_{22}(\mathrm{x}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{x})) & \mathrm{E}(g_{12}(\mathrm{x})) \\ \mathrm{E}(g_{21}(\mathrm{x})) & \mathrm{E}(g_{22}(\mathrm{x})) \end{bmatrix} \tag{1.12}$$

$$\mathrm{E}\left(g_1(\mathrm{x}) + g_2(\mathrm{x})\right) = \mathrm{E}\left(g_1(\mathrm{x})\right) + \mathrm{E}\left(g_2(\mathrm{x})\right) \tag{1.13}$$

$$\mathrm{E}\left(A\, g(\mathrm{x})\, C + B\right) = A\, \mathrm{E}\left(g(\mathrm{x})\right)\, C + B \tag{1.14}$$

$$\mathrm{E}\left(A\right) = A \tag{1.15}$$

8. A, B, C are constant matrices such that sums and products are meaningful.

T27 In linear transforms, the expectation operator fulfills the property below:[8]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{x})\right) = \left(\mathrm{E}\left(g(\mathrm{x})\right)\right)^{\intercal} \tag{1.11}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{x}) & g_{12}(\mathrm{x}) \\ g_{21}(\mathrm{x}) & g_{22}(\mathrm{x}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{x})) & \mathrm{E}(g_{12}(\mathrm{x})) \\ \mathrm{E}(g_{21}(\mathrm{x})) & \mathrm{E}(g_{22}(\mathrm{x})) \end{bmatrix} \tag{1.12}$$

$$\mathrm{E}\left(g_1(\mathrm{x}) + g_2(\mathrm{x})\right) = \mathrm{E}\left(g_1(\mathrm{x})\right) + \mathrm{E}\left(g_2(\mathrm{x})\right) \tag{1.13}$$

$$\mathrm{E}\left(A\, g(\mathrm{x})\, C + B\right) = A\, \mathrm{E}\left(g(\mathrm{x})\right)\, C + B \tag{1.14}$$

$$\mathrm{E}\left(A\right) = A \tag{1.15}$$

8. A, B, C are constant matrices such that sums and products are meaningful.

T27 In linear transforms, the expectation operator fulfills the property below:[8]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{x})\right) = \left(\mathrm{E}\left(g(\mathrm{x})\right)\right)^{\intercal} \tag{1.11}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{x}) & g_{12}(\mathrm{x}) \\ g_{21}(\mathrm{x}) & g_{22}(\mathrm{x}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{x})) & \mathrm{E}(g_{12}(\mathrm{x})) \\ \mathrm{E}(g_{21}(\mathrm{x})) & \mathrm{E}(g_{22}(\mathrm{x})) \end{bmatrix} \tag{1.12}$$

$$\mathrm{E}\left(g_1(\mathrm{x}) + g_2(\mathrm{x})\right) = \mathrm{E}\left(g_1(\mathrm{x})\right) + \mathrm{E}\left(g_2(\mathrm{x})\right) \tag{1.13}$$

$$\mathrm{E}\left(A\, g(\mathrm{x})\, C + B\right) = A\, \mathrm{E}\left(g(\mathrm{x})\right)\, C + B \tag{1.14}$$

$$\mathrm{E}\left(A\right) = A \tag{1.15}$$

8. A, B, C are constant matrices such that sums and products are meaningful.

T27 In linear transforms, the expectation operator fulfills the property below:[8]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{x})\right) = \left(\mathrm{E}\left(g(\mathrm{x})\right)\right)^{\intercal} \tag{1.11}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{x}) & g_{12}(\mathrm{x}) \\ g_{21}(\mathrm{x}) & g_{22}(\mathrm{x}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{x})) & \mathrm{E}(g_{12}(\mathrm{x})) \\ \mathrm{E}(g_{21}(\mathrm{x})) & \mathrm{E}(g_{22}(\mathrm{x})) \end{bmatrix} \tag{1.12}$$

$$\mathrm{E}\left(g_1(\mathrm{x}) + g_2(\mathrm{x})\right) = \mathrm{E}\left(g_1(\mathrm{x})\right) + \mathrm{E}\left(g_2(\mathrm{x})\right) \tag{1.13}$$

$$\mathrm{E}\left(A\, g(\mathrm{x})\, C + B\right) = A\ \mathrm{E}\left(g(\mathrm{x})\right)\ C + B \tag{1.14}$$

$$\mathrm{E}\left(A\right) = A \tag{1.15}$$

8. A, B, C are constant matrices such that sums and products are meaningful.

T27 In linear transforms, the expectation operator fulfills the property below:[8]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{x})\right) = \left(\mathrm{E}\left(g(\mathrm{x})\right)\right)^{\intercal} \tag{1.11}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{x}) & g_{12}(\mathrm{x}) \\ g_{21}(\mathrm{x}) & g_{22}(\mathrm{x}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{x})) & \mathrm{E}(g_{12}(\mathrm{x})) \\ \mathrm{E}(g_{21}(\mathrm{x})) & \mathrm{E}(g_{22}(\mathrm{x})) \end{bmatrix} \tag{1.12}$$

$$\mathrm{E}\left(g_1(\mathrm{x}) + g_2(\mathrm{x})\right) = \mathrm{E}\left(g_1(\mathrm{x})\right) + \mathrm{E}\left(g_2(\mathrm{x})\right) \tag{1.13}$$

$$\mathrm{E}\left(A\, g(\mathrm{x})\, C + B\right) = A\, \mathrm{E}\left(g(\mathrm{x})\right)\, C + B \tag{1.14}$$

$$\mathrm{E}\left(A\right) = A \tag{1.15}$$

8. A, B, C are constant matrices such that sums and products are meaningful.

T28 ◁ ▷ **Exercise 5.** Let Φ be an event. Express $\mathrm{Prob}\,(\Phi)$ by means of the expectation and the indicator function of Φ.
Let X be a r.v. Express $\mathrm{Prob}\,(\mathrm{X} \in \mathbb{A})$ by means of the expectation and the indicator function of $\mathbb{A}$.

$$E(1_\Phi) = \sum_n n \underbrace{\mathrm{Prob}(1_\Phi = n)}_{p_{1_\Phi}[n]}$$

$$= 0 \,\mathrm{Prob}(1_\Phi = 0) + 1 \,\mathrm{Prob}(1_\Phi = 1)$$

$$= \mathrm{Prob}(1_\Phi = 1)$$

$$= \mathrm{Prob}\Big(\underbrace{\{\omega \mid 1_\Phi(\omega) = 1\}}_{\Phi}\Big)$$

$$= \mathrm{Prob}(\Phi)$$

T29 The **variance** gives, for all r.v. x such that the expectation below exists, a positive semi-definite matrix $\mathrm{Var}(\mathrm{x})$ defined by:

$$\mathrm{Var}(\mathrm{x}) = \mathrm{E}\left(\underbrace{(\mathrm{x} - \mathrm{E}(\mathrm{x}))}\,(\mathrm{x} - \mathrm{E}(\mathrm{x}))^{\top}\right) \tag{1.16}$$

It is straightforward to show that:

$$\mathrm{Var}(\mathrm{x}) = \mathrm{E}\left(\mathrm{x}\,\mathrm{x}^{\top}\right) - \mathrm{E}(\mathrm{x})\,\mathrm{E}\left(\mathrm{x}^{\top}\right) \tag{1.17}$$

◁ ▷ **Exercise 6.** Prove the formula (1.17).

$$Var(x) = E\left(x x^T - E(x) x^T - x E(x^T) + E(x) E(x^T)\right)$$

$$= E(x x^T) - E(x) E(x^T) - \cancel{E(x) E(x^T)} + \cancel{E(x) E(x^T)}$$

T30 Under linear transforms, this operator fulfills the rule below:[9]

$$\mathrm{Var}\,(A\,\mathrm{x} + B) = A\ \mathrm{Var}\,(\mathrm{x})\ A^{\mathsf{T}} \tag{1.18}$$

Exercise 7. Prove the formula (1.18).

$$Y = AX + B$$

$$E(Y) = A\,E(X) + B$$

$$Y - E(Y) = A\,(X - E(X))$$

$$\mathrm{Var}(Y) = E\left((Y - E(Y))(Y - E(Y))^T\right)$$

$$= E\left(A\,(X - E(X))\,(X - E(X))^T A^T\right)$$

$$= A \quad \mathrm{Var}(X) \quad A^T$$

In dimension 1 : $\mathrm{Var}\,(aX) = a^2\,\mathrm{Var}(X)$

9. A and B are constant matrices are constant matrices such that sums and products are meaningful.

T31 Mean and variance

Let x be a numerical r.v. Its mean m_{x} and its variance $C_{\mathrm{x},\mathrm{x}}$, if they exist, are defined by:

$$m_{\mathrm{x}} = \mathrm{E}\left(\mathrm{x}\right) \qquad C_{\mathrm{x},\mathrm{x}} = \mathrm{Var}\left(\mathrm{x}\right) \tag{1.19}$$

T31 Mean and variance

Let x be a numerical r.v. Its mean m_{x} and its variance $C_{\mathrm{x},\mathrm{x}}$, if they exist, are defined by:

$$m_{\mathrm{x}} = \mathrm{E}\left(\mathrm{x}\right) \qquad C_{\mathrm{x},\mathrm{x}} = \mathrm{Var}\left(\mathrm{x}\right) \tag{1.19}$$

- The mean is also called expectation, or first **moment**, or first **cumulant**; it is a location parameter.

T31 Mean and variance

Let x be a numerical r.v. Its mean m_{x} and its variance $C_{\mathrm{x},\mathrm{x}}$, if they exist, are defined by:

$$m_{\mathrm{x}} = \mathrm{E}\left(\mathrm{x}\right) \qquad C_{\mathrm{x},\mathrm{x}} = \mathrm{Var}\left(\mathrm{x}\right) \tag{1.19}$$

- The mean is also called expectation, or first **moment**, or first **cumulant**; it is a location parameter.
- The variance is also called second cumulant; it is a scale parameter; it measures how the r.v. realizations can fall away from the mean value.

For a scalar valued r.v., the variance square root is noted σ_{x}, and called "**standard deviation**".[10]

10. If the mean is not zero, the dimensionless ratio $\frac{\sigma_{\mathrm{x}}}{m_{\mathrm{x}}}$ is called "coefficient of variation".

T31 Mean and variance

Let x be a numerical r.v. Its mean m_{x} and its variance $C_{\mathrm{x},\mathrm{x}}$, if they exist, are defined by:

$$m_{\mathrm{x}} = \mathrm{E}\left(\mathrm{x}\right) \qquad C_{\mathrm{x},\mathrm{x}} = \mathrm{Var}\left(\mathrm{x}\right) \tag{1.19}$$

- The mean is also called expectation, or first **moment**, or first **cumulant**; it is a location parameter.
- The variance is also called second cumulant; it is a scale parameter; it measures how the r.v. realizations can fall away from the mean value.

  For a scalar valued r.v., the variance square root is noted σ_{x}, and called "**standard deviation**".[10]
- If the variance is not invertible, the r.v. is **degenerate**.

  The second moment is $\mathrm{E}\left(\mathrm{x}\,\mathrm{x}^{\mathsf{T}}\right)$, that is $C_{\mathrm{x},\mathrm{x}} + m_{\mathrm{x}}\, m_{\mathrm{x}}^{\mathsf{T}}$.

  To give a comparison with solid mechanics, the mean is the center of mass, the variance is the inertia around the center of mass.

10. If the mean is not zero, the dimensionless ratio $\frac{\sigma_{\mathrm{x}}}{m_{\mathrm{x}}}$ is called "coefficient of variation".

T32 **Confidence domain**

$\text{Prob}(X \in A) = \int_A p_X(x)\, dx$

The formula (1.5) provides the probability that a r.v. belongs to a given set.

Conversely, for a given probability P_0, there exist multiple sets such that the probability that the r.v. belongs to them is P_0; these sets are the P_0 confidence domains.

T32 Confidence domain

The formula (1.5) provides the probability that a r.v. belongs to a given set.

Conversely, for a given probability P_0, there exist multiple sets such that the probability that the r.v. belongs to them is P_0; these sets are the P_0 confidence domains.

- In general, we look for the smallest confidence domain. For example, when searching a lost object in the sea, the probability to find the object being fixed, we look for the smallest domain in order to minimize the means to implement.

T32 Confidence domain

The formula (1.5) provides the probability that a r.v. belongs to a given set.

Conversely, for a given probability P_0, there exist multiple sets such that the probability that the r.v. belongs to them is P_0; these sets are the P_0 confidence domains.

- In general, we look for the smallest confidence domain. For example, when searching a lost object in the sea, the probability to find the object being fixed, we look for the smallest domain in order to minimize the means to implement.
- We can show that, if the PDF is known, with minor assumptions, the smallest confidence domain corresponds to a constant level PDF on its boundary [ref]. Such a domain do not necessarily contains the mean, and can be disconnected.

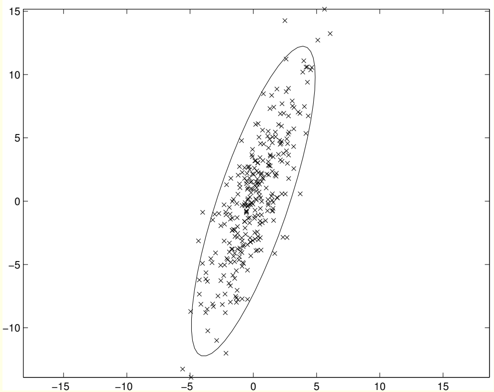


Figure 1.2: Confidence domain for a bivariate normal distribution

T33 For instance, the figure 1.2 represents a 300 realizations population of a bivariate normal r.v. (page 111) with zero mean and variance $\begin{bmatrix} 4 & 8 \\ 8 & 25 \end{bmatrix}$, and the 95% confidence domain which is a contour of the PDF. This contour est elliptic in the bivariate normal distribution case.

T34 The knowledge or the mean and the variance provides only confidence domains at at least P_0, by means of the Bienaymé-Tchebychev inequality, unless the variance is zero.

T35 Zero-variance case

◁ ▷ **Exercise 8.** Let Y be a positive r.v., such that $\mathrm{E}\left(\mathrm{Y}\right)$ exists.

Prove that $\displaystyle\int_0^{+\infty} \mathrm{Prob}\left(\mathrm{Y} \geqslant y\right) \mathrm{d}\, y = \mathrm{E}\left(\mathrm{Y}\right)$ (tip: invert integration order).

T35 Zero-variance case

◁ ▷ **Exercise 8.** Let Y be a positive r.v., such that $\mathrm{E}(\mathrm{Y})$ exists.
Prove that $\int_0^{+\infty} \mathrm{Prob}(\mathrm{Y} \geqslant y)\,\mathrm{d}\,y = \mathrm{E}(\mathrm{Y})$ (tip: invert integration order).

- Let Y be a positive r.v. whose mean exists.

  For all $y > 0$ (**Markov inequality**):[P1]

$$\mathrm{Prob}(\mathrm{Y} \geqslant y) \leqslant \frac{\mathrm{E}(\mathrm{Y})}{y} \qquad \text{for a positive r.v.} \tag{1.20}$$

T35 Zero-variance case

◁ ▷ **Exercise 8.** Let Y be a positive r.v., such that $\mathrm{E}(\mathrm{Y})$ exists.
Prove that $\int_0^{+\infty} \mathrm{Prob}(\mathrm{Y} \geqslant y)\,\mathrm{d}\,y = \mathrm{E}(\mathrm{Y})$ (tip: invert integration order).

- Let Y be a positive r.v. whose mean exists.

  For all $y > 0$ (**Markov inequality**):[P1]

$$\mathrm{Prob}(\mathrm{Y} \geqslant y) \leqslant \frac{\mathrm{E}(\mathrm{Y})}{y} \qquad \text{for a positive r.v.} \tag{1.20}$$

- By means of the Markov inequality and the exercise result:

$$\mathrm{E}(\mathrm{Y}) = 0 \quad \text{if and only if} \quad \forall y > 0,\ \mathrm{Prob}(\mathrm{Y} < y) = 1$$

  We say that the r.v. Y is null almost everywhere (or almost surely, or with probability 1); the almost sure equality is a weaker condition than the equality; but, since there is no practical consequence, this book will be simplified by saying that the r.v. Y is null, and we will note: $\mathrm{Y} = 0$.

T36 The straightforward corollary is that, for all r.v. x which takes its value in $\mathbb{R}^d$:

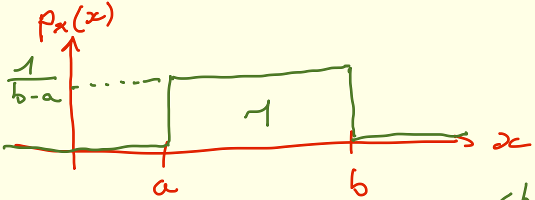
- x is (almost surely) null if and only if its square euclidean norm has zero mean:
$$\mathrm{E}\left(\mathrm{x}^\intercal \mathrm{x}\right) = 0 \quad \text{if and only if} \quad \mathrm{x} = 0_d$$
- x is (almost surely) null if and only if its 2th moment is null:
$$\mathrm{E}\left(\mathrm{x}\,\mathrm{x}^\intercal\right) = 0_{d\times d} \quad \text{if and only if} \quad \mathrm{x} = 0_d$$
- x is (almost surely) equal to its mean value, if and only if its variance is null:
$$\mathrm{Var}\left(\mathrm{x}\right) = 0_{d\times d} \quad \text{if and only if} \quad \mathrm{x} = \mathrm{E}\left(\mathrm{x}\right)$$

Many proofs in this book will be based on these equivalences.

T37 **Evaluation, module 3.** A uniform distribution is constant over the support. We consider the case of a uniformly distributed r.v. over $[a, b]$, with $a < b$.

a) Calculate the mean and the variance.

b) How to choose a and b such that the r.v. has zero mean and unit variance?



a) $E(X) = \int x \, p_x(x) \, dx = \frac{1}{b-a} \int_a^b x \, dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b$

$= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$

$E(X^2) = \frac{1}{b-a} \int_a^b x^2 \, dx$ $\quad Var(X) = \frac{(b-a)^2}{12}$ $\quad \sigma_X = \frac{b-a}{2\sqrt{3}}$

b) $\pm\sqrt{3}$

T38

Other features

◁ ▷ **Exercise 9.** Prove that, for a distribution with a PDF, the mean value minimizes the criterion $J(m) = \mathrm{E}\left(\|\mathrm{X} - m\|_W^2\right)$ with respect to m (independently from the symmetric positive definite matrix W used in the norm $\|x\|_W = \sqrt{x^\intercal W\, x}$).

T39 Mode, median, quartiles

There are several possibilities to define the "center" of a probability distribution of a r.v. X.

- the mean value minimizes $\mathrm{E}\left(\|\mathrm{X}-m\|_2^2\right)$ with respect to m; it does not always exist;
- for a scalar valued r.v., the **median** minimizes $\mathrm{E}\left(|\mathrm{X}-m|\right)$ with respect to m [ref];
  it is a value m such that $\mathrm{Prob}\left(\mathrm{X}<m\right) \leqslant \frac{1}{2}$ and $\mathrm{Prob}\left(\mathrm{X}>m\right) \leqslant \frac{1}{2}$;[11]
  an interval of solutions can exist, a common habit is to use the middle;
  the median can be extended to the vector-valued case [ref];
- the **mode**, which is the value which maximizes the PDF;
  it is not necessarily unique;
  if a PDF exhibits several local maxima, the distribution is a **multimodal** one.

11. These probabilities equal $\frac{1}{2}$ if $\mathrm{Prob}\left(\mathrm{x}=m\right)=0$

T39 Mode, median, quartiles

There are several possibilities to define the "center" of a probability distribution of a r.v. X.

- the mean value minimizes $\mathrm{E}\left(\|\mathrm{X}-m\|_2^2\right)$ with respect to m; it does not always exist;
- for a scalar valued r.v., the **median** minimizes $\mathrm{E}\left(|\mathrm{X}-m|\right)$ with respect to m [ref];
  it is a value m such that $\mathrm{Prob}\left(\mathrm{X}<m\right) \leqslant \frac{1}{2}$ and $\mathrm{Prob}\left(\mathrm{X}>m\right) \leqslant \frac{1}{2}$;[11]
  an interval of solutions can exist, a common habit is to use the middle;
  the median can be extended to the vector-valued case [ref];
- the **mode**, which is the value which maximizes the PDF;
  it is not necessarily unique;
  if a PDF exhibits several local maxima, the distribution is a **multimodal** one.

11. These probabilities equal $\frac{1}{2}$ if $\mathrm{Prob}\left(\mathrm{x}=m\right)=0$

T39 Mode, median, quartiles

There are several possibilities to define the "center" of a probability distribution of a r.v. X.

- the mean value minimizes $\mathrm{E}\left(\|\mathrm{X}-m\|_2^2\right)$ with respect to m; it does not always exist;
- for a scalar valued r.v., the **median** minimizes $\mathrm{E}\left(|\mathrm{X}-m|\right)$ with respect to m [ref];
  it is a value m such that $\mathrm{Prob}\left(\mathrm{X}<m\right) \leqslant \frac{1}{2}$ and $\mathrm{Prob}\left(\mathrm{X}>m\right) \leqslant \frac{1}{2}$;[11]
  an interval of solutions can exist, a common habit is to use the middle;
  the median can be extended to the vector-valued case [ref];
- the **mode**, which is the value which maximizes the PDF;
  it is not necessarily unique;
  if a PDF exhibits several local maxima, the distribution is a **multimodal** one.

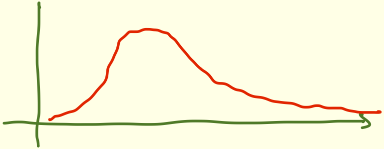
11. These probabilities equal $\frac{1}{2}$ if $\mathrm{Prob}\left(\mathrm{x}=m\right)=0$

T39 Mode, median, quartiles

There are several possibilities to define the "center" of a probability distribution of a r.v. X.

- the mean value minimizes $\mathrm{E}\left(\|\mathrm{X}-m\|_2^2\right)$ with respect to m; it does not always exist;
- for a scalar valued r.v., the **median** minimizes $\mathrm{E}\left(|\mathrm{X}-m|\right)$ with respect to m [ref];
  it is a value m such that $\mathrm{Prob}\left(\mathrm{X}<m\right)\leqslant\frac{1}{2}$ and $\mathrm{Prob}\left(\mathrm{X}>m\right)\leqslant\frac{1}{2}$;[11]
  an interval of solutions can exist, a common habit is to use the middle;
  the median can be extended to the vector-valued case [ref];
- the **mode**, which is the value which maximizes the PDF;
  it is not necessarily unique;
  if a PDF exhibits several local maxima, the distribution is a **multimodal** one.

➲ In general, in the scalar valued r.v. case, for a unimodal density which spreads over the right side: mode < median < mean.



11. These probabilities equal $\frac{1}{2}$ if $\mathrm{Prob}\left(\mathrm{x}=m\right)=0$

T40 In a scalar valued r.v. case, the quartiles try to summarize the PDF shape:

- the first quartile Q_1 is a value such that $\mathrm{Prob}\left(\mathrm{x} < Q_1\right) \leqslant \frac{1}{4}$ and $\mathrm{Prob}\left(\mathrm{x} > Q_1\right) \leqslant \frac{3}{4}$;
- the second quartile is the median;
- the third quartile Q_3 is a value such that $\mathrm{Prob}\left(\mathrm{x} < Q_3\right) \leqslant \frac{3}{4}$ and $\mathrm{Prob}\left(\mathrm{x} > Q_3\right) \leqslant \frac{1}{4}$;

For Q_1 and Q_3, an interval of solutions can exist; a common habit is to use the upper bound.
The interquartile range $Q_3 - Q_1$ is another way to measure the dispersion of a r.v.
The quartiles are a special case among **quantiles**.

T41 Higher order statistics

The moments and the cumulants of order 1 and 2 (if they exist) are only partial characteristics of a r.v.. They can be completed with higher order moments and cumulants, but the characterization remains partial.

T41 Higher order statistics

The moments and the cumulants of order 1 and 2 (if they exist) are only partial characteristics of a r.v.. They can be completed with higher order moments and cumulants, but the characterization remains partial.

In a nutshell:

- the moments are the coefficients of the Taylor series expansion around 0 of the first characteristic function, that is the Fourier transform of the PDF;
- the cumulants are the coefficients of the Taylor series expansion around 0 of the second characteristic function, that is the logarithm of the Fourier transform of the PDF.

For all $k \in \mathbb{N}$, there exists and invertible relation between the k first moments and the k first cumulants. With the cumulants, it is easier to obtain an interpretation which is independent of the location and the scale.

T42 In particular, for a scalar r.v. x, the kth moment is $\mathrm{E}\left(\mathrm{x}^k\right)$. Let's note $\mathrm{K}_k(\mathrm{x})$ the kth cumulant; the 4 first cumulants are written in function of the 4 first moments through the formulae:

$$
\begin{aligned}
m_{\mathrm{x}} &= \mathrm{K}_1(\mathrm{x}) = \mathrm{E}\,(\mathrm{x}) \\
\sigma_{\mathrm{x}}^2 &= \mathrm{K}_2(\mathrm{x}) = \mathrm{E}\left(\mathrm{x}^2\right) - m_{\mathrm{x}}^2 \\
&\mathrm{K}_3(\mathrm{x}) = \mathrm{E}\left(\mathrm{x}^3\right) - 3\,m_{\mathrm{x}}\;\mathrm{E}\left(\mathrm{x}^2\right) + 2\,m_{\mathrm{x}}^3 \\
&\mathrm{K}_4(\mathrm{x}) = \mathrm{E}\left(\mathrm{x}^4\right) - 4\,m_{\mathrm{x}}\;\mathrm{E}\left(\mathrm{x}^3\right) + 12\,m_{\mathrm{x}}^2\;\mathrm{E}\left(\mathrm{x}^2\right) - 6\,m_{\mathrm{x}}^4
\end{aligned}
$$

T43 For a non-zero variance, we can use the normalized (and dimensionless) cumulants $\widetilde{\mathrm{K}}_3(\mathrm{x})$ et $\widetilde{\mathrm{K}}_4(\mathrm{x})$:

$$\widetilde{\mathrm{K}}_3(\mathrm{x}) \hat{=} \frac{\mathrm{K}_3(\mathrm{x})}{\sigma_{\mathrm{x}}^3} = \mathrm{E}\left(\left(\frac{\mathrm{x}-m_{\mathrm{x}}}{\sigma_{\mathrm{x}}}\right)^3\right) \qquad \text{skewness}$$

$$\widetilde{\mathrm{K}}_4(\mathrm{x}) \hat{=} \frac{\mathrm{K}_4(\mathrm{x})}{\sigma_{\mathrm{x}}^4} = \mathrm{E}\left(\left(\frac{\mathrm{x}-m_{\mathrm{x}}}{\sigma_{\mathrm{x}}}\right)^4\right) - 3 \qquad \text{excess Kurtosis}$$

Both factors are independent of the mean and the variance, because they depend only of the standard score $\frac{\mathrm{x}-m_{\mathrm{x}}}{\sigma_{\mathrm{x}}}$; they only depend of the PDF shape.

T44 The **skewness** measures the asymmetry of the probability distribution: if the PDF is symmetric around the mean value, it is zero (but the reciprocal is false in general).
For a unimodal density which spreads over the right side, it is positive.

T44 The **skewness** measures the asymmetry of the probability distribution: if the PDF is symmetric around the mean value, it is zero (but the reciprocal is false in general).
For a unimodal density which spreads over the right side, it is positive.

T45 Distribution models

Two basic distributions

"Deterministic distribution" Although we often distinguish r.v. from deterministic variables, we can present the deterministic variables as a limit case of the r.v., in which a r.v. always takes the same value x_0. Therefore, a deterministic variable is a discrete r.v. whose PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

T45 Distribution models

Two basic distributions

"Deterministic distribution" Although we often distinguish r.v. from deterministic variables, we can present the deterministic variables as a limit case of the r.v., in which a r.v. always takes the same value x_0. Therefore, a deterministic variable is a discrete r.v. whose PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

- By means of Kronecker delta, a short way to write this PMF is:

$$p_{\mathrm{x}}(x) = \delta(x - x_0)$$

T45 Distribution models

Two basic distributions

"Deterministic distribution" Although we often distinguish r.v. from deterministic variables, we can present the deterministic variables as a limit case of the r.v., in which a r.v. always takes the same value x_0. Therefore, a deterministic variable is a discrete r.v. whose PMF is:

$$p_{\mathrm{X}}(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

- By means of Kronecker delta, a short way to write this PMF is:

$$p_{\mathrm{X}}(x) = \delta(x - x_0)$$

- But a numerical deterministic r.v. can also be considered as a continuous r.v.: in the formula above, p_{X} becomes the PDF, and δ becomes the Dirac delta.

T45 # Distribution models

Two basic distributions

"Deterministic distribution" Although we often distinguish r.v. from deterministic variables, we can present the deterministic variables as a limit case of the r.v., in which a r.v. always takes the same value x_0. Therefore, a deterministic variable is a discrete r.v. whose PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

- By means of Kronecker delta, a short way to write this PMF is:

$$p_{\mathrm{x}}(x) = \delta(x - x_0)$$

- But a numerical deterministic r.v. can also be considered as a continuous r.v.: in the formula above, p_{x} becomes the PDF, and δ becomes the Dirac delta.
- A deterministic variable has null variance.

T46 **Bernoulli distribution** A r.v. is Bernoulli-distributed if it only takes two distinct values x_1 and x_2, one with the probability $\lambda \in]0,1[$, the other with the probability $1-\lambda$; the PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} \lambda & \text{if } x = x_1 \\ 1-\lambda & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases}$$

T46 **Bernoulli distribution** A r.v. is Bernoulli-distributed if it only takes two distinct values x_1 and x_2, one with the probability $\lambda \in]0,1[$, the other with the probability $1-\lambda$; the PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} \lambda & \text{if } x = x_1 \\ 1-\lambda & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases}$$

- In the formula below, if p_{x} is a PMF, δ is the Kronecker delta; if p_{x} is a PDF, δ is the Dirac delta:

$$p_{\mathrm{x}}(x) = \lambda\,\delta(x-x_1) + (1-\lambda)\,\delta(x-x_2)$$

T46 **Bernoulli distribution** A r.v. is Bernoulli-distributed if it only takes two distinct values x_1 and x_2, one with the probability $\lambda \in]0, 1[$, the other with the probability $1 - \lambda$; the PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} \lambda & \text{if } x = x_1 \\ 1 - \lambda & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases}$$

- In the formula below, if p_{x} is a PMF, δ is the Kronecker delta; if p_{x} is a PDF, δ is the Dirac delta:

$$p_{\mathrm{x}}(x) = \lambda\,\delta(x - x_1) + (1 - \lambda)\,\delta(x - x_2)$$

- The mean is $\lambda\,x_1 + (1 - \lambda)\,x_2$, the variance is $\lambda\,(1 - \lambda)\,(x_2 - x_1)\,(x_2 - x_1)^{\mathsf{T}}$.

T46 **Bernoulli distribution** A r.v. is Bernoulli-distributed if it only takes two distinct values x_1 and x_2, one with the probability $\lambda \in]0,1[$, the other with the probability $1-\lambda$; the PMF is:

$$p_{\mathrm{x}}(x) = \begin{cases} \lambda & \text{if } x = x_1 \\ 1-\lambda & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases}$$

- In the formula below, if p_{x} is a PMF, δ is the Kronecker delta; if p_{x} is a PDF, δ is the Dirac delta:

$$p_{\mathrm{x}}(x) = \lambda\,\delta(x-x_1) + (1-\lambda)\,\delta(x-x_2)$$

- The mean is $\lambda\,x_1 + (1-\lambda)\,x_2$, the variance is $\lambda\,(1-\lambda)\,(x_2-x_1)\,(x_2-x_1)^{\intercal}$.
- In the scalar case, in general, skewness and kurtosis fulfill the inequality below:

$$\widetilde{\mathrm{K}}_4(\mathrm{x}) \geqslant [\widetilde{\mathrm{K}}_3(\mathrm{x})]^2 - 2$$

The equality holds only with the Bernoulli distribution.[P2]

T47 ◁ ▷ **Exercise 10** (Bernoulli distribution). Complete the figure 1.3 with the PDF of the Bernoulli distribution with equiprobability, zero mean and unit variance.

equiprobability : $\lambda = 1-\lambda \Rightarrow \lambda = \frac{1}{2}$

zero mean : $\frac{x_1+x_2}{2} = 0 \Rightarrow x_2 = -x_1$

unit variance : $\frac{1}{2}\,\frac{1}{2}\left(x_2-x_1\right)^2 = 1$

$\lambda = \frac{1}{2}$
$x_1 = -1$
$x_2 = +1$

T48 A few probability distributions

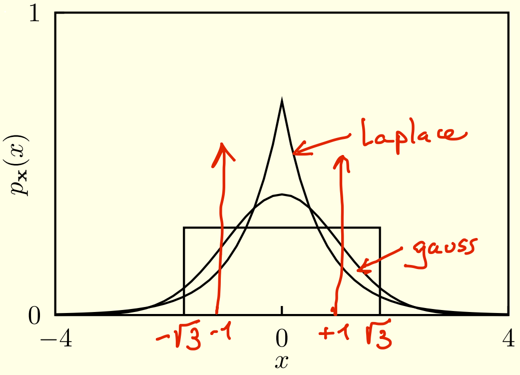


Figure 1.3: PDF of uniform, Gauss, Laplace distributions (zero mean, unit variance)

T49 **Uniform distribution** A real valued r.v. is uniformly distributed on the interval $[a,b]$ if its PDF is constant on this interval, and zero elsewhere:

$$p_{\mathrm{x}}(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

The excess kurtosis is $-\frac{6}{5}$.

T50 **Laplace distribution** A real valued r.v. is Laplace-distributed with mean m_{x} and standard deviation σ_{x} if its PDF is:

$$p_{\mathrm{x}}(x) = \frac{1}{\sigma_{\mathrm{x}}\sqrt{2}} \exp\left(-\sqrt{2}\,\frac{|x - m_{\mathrm{x}}|}{\sigma_{\mathrm{x}}}\right)$$

The excess kurtosis is 3.

T50 **Laplace distribution** A real valued r.v. is Laplace-distributed with mean m_x and standard deviation σ_x if its PDF is:

$$p_\mathrm{x}(x) = \frac{1}{\sigma_\mathrm{x}\sqrt{2}} \exp\left(-\sqrt{2}\,\frac{|x-m_\mathrm{x}|}{\sigma_\mathrm{x}}\right)$$

The excess kurtosis is 3.

Univariate normal distribution A real valued r.v. is normally distributed (or **Gauss**-distributed) with mean m_x and standard deviation σ_x if its PDF is:

$$p_\mathrm{x}(x) = \frac{1}{\sigma_\mathrm{x}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\,\frac{(x-m_\mathrm{x})^2}{\sigma_\mathrm{x}^2}\right)$$

All higher order cumulants are zero.[12]

12. A probability distribution is platikurtic (or sub-Gaussian) if the kurtosis is negative, mesokurtic if it is zero, leptokurtic (or super-Gaussian) if it is positive.

T51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.

T51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.

➲ From a pragmatic point of view, this distribution often leads to simple calculations which can be easily implemented with limited numerical means. For example, on the NASA Apollo program, in 1960, a trajectory estimation filter based on Gaussian assumption was implemented.

T51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.

- From a pragmatic point of view, this distribution often leads to simple calculations which can be easily implemented with limited numerical means. For example, on the NASA Apollo program, in 1960, a trajectory estimation filter based on Gaussian assumption was implemented.
- Least squares optimization are often encountered (for example in regression problems), they can be explained with a probabilistic interpretation based on Gaussian distributions.

T51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.

- From a pragmatic point of view, this distribution often leads to simple calculations which can be easily implemented with limited numerical means. For example, on the NASA Apollo program, in 1960, a trajectory estimation filter based on Gaussian assumption was implemented.
- Least squares optimization are often encountered (for example in regression problems), they can be explained with a probabilistic interpretation based on Gaussian distributions.
- These remarks also hold for the multivariate normal distribution, page 107, for which the PDF is, if the variance is invertible:

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{\det(2\pi\, C_{\mathrm{x},\mathrm{x}})}} \exp\left[-\frac{1}{2}(x - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x},\mathrm{x}}^{-1}\, (x - m_{\mathrm{x}})\right] \tag{1.21}$$

T51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.

- From a pragmatic point of view, this distribution often leads to simple calculations which can be easily implemented with limited numerical means. For example, on the NASA Apollo program, in 1960, a trajectory estimation filter based on Gaussian assumption was implemented.
- Least squares optimization are often encountered (for example in regression problems), they can be explained with a probabilistic interpretation based on Gaussian distributions.
- These remarks also hold for the multivariate normal distribution, page 107, for which the PDF is, if the variance is invertible:

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{\det(2\pi\, C_{\mathrm{x,x}})}} \exp\left[-\frac{1}{2}(x - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x,x}}^{-1}\,(x - m_{\mathrm{x}})\right] \tag{1.21}$$

- If x is a normal r.v. which takes its value in $\mathbb{R}^d$, $(\mathrm{x} - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x,x}}^{-1}\,(\mathrm{x} - m_{\mathrm{x}})$ is driven by a χ^2 distribution.

T52 χ^2 **distribution** A real valued r.v. is χ^2-distributed with d degres of freedom if its PDF is: [13]

$$p_{\mathrm{X}}(x) = \begin{cases} \frac{1}{2^{\frac{d}{2}}\,\Gamma\left(\frac{d}{2}\right)}\, x^{\frac{d}{2}-1}\, \exp\left(-\frac{x}{2}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is:

$$F_{\mathrm{X}}(x) = \begin{cases} \Gamma_{\mathsf{inc}}\left(\frac{x}{2}, \frac{d}{2}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The mean is d, the variance is $2\,d$.
For $d = 2$, we obtains the exponential distribution with mean 2.

13. Γ et Γ_{inc} are the Euler gamma function and the incomplete Euler gamma function:

$$\Gamma(a) = \int_0^{+\infty} t^{a-1}\, \mathrm{e}^{-t}\, \mathrm{d}\,t \qquad \Gamma_{\mathsf{inc}}(x, a) = \frac{1}{\Gamma(a)} \int_0^{x} t^{a-1}\, \mathrm{e}^{-t}\, \mathrm{d}\,t$$

if $a \in \mathbb{N}$, $\Gamma(a+1) = a!$ and $\Gamma\left(a + \frac{1}{2}\right) = \frac{\prod_{k=1}^{a}(2\,k-1)}{2^a}\sqrt{\pi}$

T53 **Exponential distribution** A real valued r.v. is exponentially distributed with mean m_x if its PDF is:

$$p_\mathrm{x}(x) = \begin{cases} \frac{1}{m_\mathrm{x}} \exp\left(-\frac{x}{m_\mathrm{x}}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is:

$$F_\mathrm{x}(x) = \begin{cases} 1 - \exp\left(-\frac{x}{m_\mathrm{x}}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The standard deviation is the mean value.

T54 ◁ ▷ **Exercise 11.** Let X be a real valued r.v., and Y is another r.v. such that $Y = \max(X, 0)$. Write the CDF of Y in function of the CDF of X. Deduce the PDF of Y in function of the PDF of X and a Dirac delta function with proper weight.

CDF of Y: $F_Y(y) = \text{Prob}(Y \leq y)$

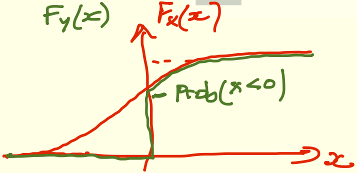
$$= \text{Prob}(\max(X,0) \leq y)$$

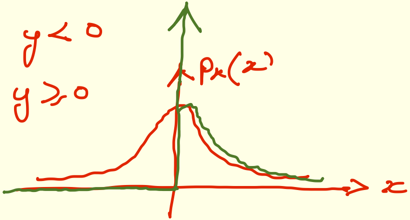
$$= \text{Prob}(X \leq y \text{ and } 0 \leq y)$$

$$= \text{Prob}((X \leq y) \cap (0 \leq y))$$

$$= \begin{cases} \text{Prob}((X \leq y) \cap \emptyset) & \text{if } y < 0 \\ \text{Prob}((X \leq y) \cap \Omega) & \text{if } y \geq 0 \end{cases}$$

$$= \begin{cases} 0 & \text{if } y < 0 \\ F_X(x) & \text{if } y \geq 0 \end{cases}$$





$$p_Y(x) = p_X(x)\, 1_{\mathbb{R}_+}(x) + \text{Prob}(X \leq 0)\, \delta(x)$$

T55 ◁ ▷ **Evaluation, module 5** (Binomial distribution). A coin is such that the probability to obtain "tail" is λ. If we throw this coin n times, the number of times for which we obtain "tail" is a binomial r.v. K with parameters n and λ, which takes its value in $\{0, \ldots, n\}$; the PMF is:

$$p_{\mathrm{K}}(k) = n! \, \frac{\lambda^k}{k!} \, \frac{(1-\lambda)^{n-k}}{(n-k)!}$$

We throw this coin twice.

a) What is the probability to never obtain "tail"?

b) What is the probability to obtain "tail" exactly once?

c) What is the probability to obtain "tail" twice?

d) What is the sum of the three preceding results?

e) Calculate the mean and the variance of the number of "tails".

T56 Pair of r.v.: joint and marginal distributions

Definitions

Let's consider a r.v. of dimension at least 2 whose components are split under the form $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$. X which takes its values in $\mathbb{X}$ and Y which takes its values in $\mathbb{Y}$ make up the pair (X, Y). The distribution of this pair is the distribution of $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$; but, this split being done, the distribution of this pair is called the **joint distribution**.

T56 Pair of r.v.: joint and marginal distributions

Definitions

Let's consider a r.v. of dimension at least 2 whose components are split under the form $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$. X which takes its values in $\mathbb{X}$ and Y which takes its values in $\mathbb{Y}$ make up the pair (X,Y). The distribution of this pair is the distribution of $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$; but, this split being done, the distribution of this pair is called the **joint distribution**.

- This book will focus on the case where the joint distribution has a PDF (or PMF, or hybrid) denoted $p_{\mathrm{X},\mathrm{Y}}$; for all (x,y) (note that $p_{\mathrm{X},\mathrm{Y}}(x,y)=p_{\mathrm{Y},\mathrm{X}}(y,x)$):

$$\boxed{p_{\mathrm{X},\mathrm{Y}}(x,y)=p_{\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]}\left(\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]\right)}$$

T56 Pair of r.v.: joint and marginal distributions

Definitions

Let's consider a r.v. of dimension at least 2 whose components are split under the form $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$. X which takes its values in $\mathbb{X}$ and Y which takes its values in $\mathbb{Y}$ make up the pair (X,Y). The distribution of this pair is the distribution of $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$; but, this split being done, the distribution of this pair is called the **joint distribution**.

- This book will focus on the case where the joint distribution has a PDF (or PMF, or hybrid) denoted $p_{\mathrm{X},\mathrm{Y}}$; for all (x,y) (note that $p_{\mathrm{X},\mathrm{Y}}(x,y)=p_{\mathrm{Y},\mathrm{X}}(y,x)$):

$$p_{\mathrm{X},\mathrm{Y}}(x,y)=p_{\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]}\left(\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]\right)$$

- The probability calculation is:

$$\mathrm{Prob}\left((\mathrm{X},\mathrm{Y})\in\mathbb{A}\right)=\int_{\mathbb{A}}p_{\mathrm{X},\mathrm{Y}}(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y \qquad (1.22)$$

T56 Pair of r.v.: joint and marginal distributions

Definitions

Let's consider a r.v. of dimension at least 2 whose components are split under the form $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$. X which takes its values in $\mathbb{X}$ and Y which takes its values in $\mathbb{Y}$ make up the pair (X,Y). The distribution of this pair is the distribution of $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$; but, this split being done, the distribution of this pair is called the **joint distribution**.

- This book will focus on the case where the joint distribution has a PDF (or PMF, or hybrid) denoted $p_{\mathrm{X},\mathrm{Y}}$; for all (x,y) (note that $p_{\mathrm{X},\mathrm{Y}}(x,y) = p_{\mathrm{Y},\mathrm{X}}(y,x)$):

$$p_{\mathrm{X},\mathrm{Y}}(x,y) = p_{\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]}\left(\left[\begin{smallmatrix}x\\ y\end{smallmatrix}\right]\right)$$

- The probability calculation is:

$$\mathrm{Prob}\left((\mathrm{X},\mathrm{Y}) \in \mathbb{A}\right) = \int_{\mathbb{A}} p_{\mathrm{X},\mathrm{Y}}(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y \qquad (1.22)$$

- Furthermore, the support is denoted $\mathrm{S}\left(\mathrm{X},\mathrm{Y}\right)$.

T57 The components X and Y of the pair are called **marginal** r.v.
When we observe a marginal r.v. X, we ignore the component Y in the pair (X, Y).

The PDF (or PMF) of the marginal r.v. are obtained by integration (or summation) of the joint PDF (or PMF) with respect to the ignored variable; for example, the PDF of X is, for all x:[P3]

$$\boxed{p_{\mathrm{X}}(x) = \int_{\mathbb{Y}} p_{\mathrm{X},\mathrm{Y}}(x, y)\, \mathrm{d}\, y} \tag{1.23}$$

T57 The components X and Y of the pair are called **marginal** r.v.
When we observe a marginal r.v. X, we ignore the component Y in the pair (X, Y).

The PDF (or PMF) of the marginal r.v. are obtained by integration (or summation) of the joint PDF (or PMF) with respect to the ignored variable; for example, the PDF of X is, for all x:[P3]

$$p_{\mathrm{X}}(x) = \int_{\mathbb{Y}} p_{\mathrm{X},\mathrm{Y}}(x, y) \, \mathrm{d}\, y \tag{1.23}$$

⚠ The knowledge of both marginal distributions is not sufficient to know the joint distribution.

T57 The components X and Y of the pair are called **marginal** r.v.
When we observe a marginal r.v. X, we ignore the component Y in the pair (X, Y).

The PDF (or PMF) of the marginal r.v. are obtained by integration (or summation) of the joint PDF (or PMF) with respect to the ignored variable; for example, the PDF of X is, for all x:[P3]

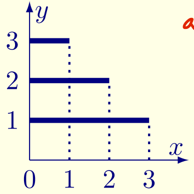
$$p_{\mathrm{X}}(x) = \int_{\mathbb{Y}} p_{\mathrm{X},\mathrm{Y}}(x, y)\, \mathrm{d}\, y \tag{1.23}$$

The knowledge of both marginal distributions is not sufficient to know the joint distribution.

The properties given in the section 1.2 hold:

- for the marginal r.v. X;
- for the marginal r.v Y;
- for the r.v. $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$, that is the pair (X, Y), through the change of notation $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right] \to (\mathrm{X}, \mathrm{Y})$.

T58 ◁ ▷ **Exercise 12** (Pair of r.v.: red thread). Let (X, Y) be a pair of r.v. where X is a continuous valued one and Y is a discrete valued one. We suppose that $p_{X,Y}$ is constant over the support $([0,3] \times \{1\}) \cup ([0,2] \times \{2\}) \cup ([0,1] \times \{3\})$.



a) Let note α the value of the constant PDF

$$\sum_y \int p_{x,y}(x,y)\,dx = 1$$

$$\int_0^3 \underbrace{p_{x,y}(x,1)}_{\alpha}\,dx + \int_0^2 \underbrace{p_{x,y}(x,2)}_{\alpha}\,dx + \int_0^1 \underbrace{p_{x,y}(x,3)}_{\alpha}\,dx = 1$$

$$3\alpha + 2\alpha + \alpha = 1$$

Solution : $\alpha = \frac{1}{6}$

a) What is the value of $p_{X,Y}$ on the support?
Plot $x \mapsto p_{X,Y}(x,y)$ for $y = 1$, for $y = 2$, and for $y = 3$
Plot $y \mapsto p_{X,Y}(x,y)$ for $0 < x < 1$, for $1 < x < 2$, and for $2 < x < 3$.

b) Give and plot the PDF of the marginal r.v. X; give its mean and its variance.

c) Give and plot the PMF of the marginal r.v. Y; give its mean and its variance.

T59 Mathematical expectation

With g a function from $\mathbb{X} \times \mathbb{Y}$ with numerical value, the expectation operator is:

$$\mathrm{E}\left(g(\mathrm{X}, \mathrm{Y})\right) = \int_{\mathbb{X}\times\mathbb{Y}} g(x, y)\, p_{\mathrm{X},\mathrm{Y}}(x, y)\; \mathrm{d}\, x\; \mathrm{d}\, y \qquad (1.24)$$

T59 **Mathematical expectation**

With g a function from $\mathbb{X} \times \mathbb{Y}$ with numerical value, the expectation operator is:

$$\mathrm{E}\left(g(\mathrm{X},\mathrm{Y})\right) = \int_{\mathbb{X}\times\mathbb{Y}} g(x,y)\, p_{\mathrm{X},\mathrm{Y}}(x,y)\, \mathrm{d}\,x\, \mathrm{d}\,y \tag{1.24}$$

- In linear transforms, this operator fulfills the following rules:[14]

$$\mathrm{E}\left(g^{\mathsf{T}}(\mathrm{X},\mathrm{Y})\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y})\right)\right)^{\mathsf{T}} \tag{1.25}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(g_{11}(\mathrm{X},\mathrm{Y})) & \mathrm{E}(g_{12}(\mathrm{X},\mathrm{Y})) \\ \mathrm{E}(g_{21}(\mathrm{X},\mathrm{Y})) & \mathrm{E}(g_{22}(\mathrm{X},\mathrm{Y})) \end{bmatrix} \tag{1.26}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y})\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y})\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y})\right) \tag{1.27}$$

$$\mathrm{E}\left(A\, g(\mathrm{X},\mathrm{Y})\, C + B\right) = A\ \mathrm{E}\left(g(\mathrm{X},\mathrm{Y})\right)\ C + B \tag{1.28}$$

14. A, B, C are constant matrices such that sums and products are meaningful.

T60 The variance operator can apply to both components X and Y. Furthermore, we define the **covariance** operator:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)=\mathrm{E}\left(\left(\mathrm{X}-\mathrm{E}\left(\mathrm{X}\right)\right)\left(\mathrm{Y}-\mathrm{E}\left(\mathrm{Y}\right)\right)^{\mathsf{T}}\right) \qquad d_x \times d_y \tag{1.29}$$

The alternate form below is straightforward to obtain:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)=\mathrm{E}\left(\mathrm{X}\,\mathrm{Y}^{\mathsf{T}}\right)-\mathrm{E}\left(\mathrm{X}\right)\,\mathrm{E}\left(\mathrm{Y}^{\mathsf{T}}\right) \tag{1.30}$$

T60 The variance operator can apply to both components X and Y. Furthermore, we define the **covariance** operator:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)=\mathrm{E}\left(\left(\mathrm{X}-\mathrm{E}\left(\mathrm{X}\right)\right)\left(\mathrm{Y}-\mathrm{E}\left(\mathrm{Y}\right)\right)^{\mathsf{T}}\right) \tag{1.29}$$

The alternate form below is straightforward to obtain:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)=\mathrm{E}\left(\mathrm{X}\,\mathrm{Y}^{\mathsf{T}}\right)-\mathrm{E}\left(\mathrm{X}\right)\,\mathrm{E}\left(\mathrm{Y}^{\mathsf{T}}\right) \tag{1.30}$$

- In linear transforms, the covariance operator fulfills the rules below:[15]

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{X}\right)=\mathrm{Var}\left(\mathrm{X}\right) \tag{1.31}$$

$$\mathrm{Cov}\left(\mathrm{Y},\mathrm{X}\right)=\left(\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)\right)^{\mathsf{T}} \tag{1.32}$$

$$\mathrm{Cov}\left(A\,\mathrm{X}+B,C\,\mathrm{Y}+D\right)=A\;\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)\;C^{\mathsf{T}} \tag{1.33}$$

15. A, B, C, D are constant matrices such that sums and products are meaningful.

T61 Furthermore, if the sums below are meaningful:

$$\mathrm{E}(\mathrm{X}+\mathrm{Y}) = \mathrm{E}(\mathrm{X}) + \mathrm{E}(\mathrm{Y}) \tag{1.34}$$

$$\mathrm{Var}(\mathrm{X}+\mathrm{Y}) = \mathrm{Var}(\mathrm{X}) + \mathrm{Var}(\mathrm{Y}) + \mathrm{Cov}(\mathrm{X},\mathrm{Y}) + \mathrm{Cov}(\mathrm{Y},\mathrm{X}) \tag{1.35}$$

$$\mathrm{Cov}\left(\mathrm{X}+\mathrm{Y},\mathrm{X}'+\mathrm{Y}'\right) = \mathrm{Cov}\left(\mathrm{X},\mathrm{X}'\right) + \mathrm{Cov}\left(\mathrm{X},\mathrm{Y}'\right) + \mathrm{Cov}\left(\mathrm{Y},\mathrm{X}'\right) + \mathrm{Cov}\left(\mathrm{Y},\mathrm{Y}'\right) \tag{1.36}$$

$$(X+Y)(X'+Y)^T = XX'^T + XY'^T + Y.X'^T + YY'^T$$

T62 Mean, variance, covariance

Means and variances of marginal r.v. X and Y are not sufficient to retrieve the mean and the variance of the r.v. $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$.

They have to be completed with the covariance $C_{\mathrm{X},\mathrm{Y}}$ defined by (note that $C_{\mathrm{Y},\mathrm{X}} = C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}$):

$$C_{\mathrm{X},\mathrm{Y}} = \mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right) \tag{1.37}$$

Thus, the mean and the variance of the r.v. $\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]$ are:

$$\mathrm{E}\left(\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]\right) = \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix} \qquad \mathrm{Var}\left(\left[\begin{smallmatrix}\mathrm{X}\\\mathrm{Y}\end{smallmatrix}\right]\right) = \begin{bmatrix} C_{\mathrm{X},\mathrm{X}} & C_{\mathrm{X},\mathrm{Y}} \\ C_{\mathrm{Y},\mathrm{X}} & C_{\mathrm{Y},\mathrm{Y}} \end{bmatrix} \tag{1.38}$$

T62 Mean, variance, covariance

Means and variances of marginal r.v. X and Y are not sufficient to retrieve the mean and the variance of the r.v. $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$.

They have to be completed with the covariance $C_{\mathrm{X},\mathrm{Y}}$ defined by (note that $C_{\mathrm{Y},\mathrm{X}} = C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}$):

$$C_{\mathrm{X},\mathrm{Y}} = \mathrm{Cov}\left(\mathrm{X}, \mathrm{Y}\right) \tag{1.37}$$

Thus, the mean and the variance of the r.v. $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$ are:

$$\mathrm{E}\left(\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]\right) = \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix} \qquad \mathrm{Var}\left(\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]\right) = \begin{bmatrix} C_{\mathrm{X},\mathrm{X}} & C_{\mathrm{X},\mathrm{Y}} \\ C_{\mathrm{Y},\mathrm{X}} & C_{\mathrm{Y},\mathrm{Y}} \end{bmatrix} \tag{1.38}$$

- The **Cauchy-Schwarz inequality** is (inequality for Loewner order):

$$C_{\mathrm{X},\mathrm{X}} \geqslant C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \tag{1.39}$$

The equality holds if and only if there is a linear relation between X and Y:[P4]

$$C_{\mathrm{X},\mathrm{X}} = C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \quad \text{if and only if} \quad \mathrm{X} = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\left(\mathrm{Y} - m_{\mathrm{Y}}\right) \tag{1.40}$$

T63 Independence, uncorrelatedness

We say that the r.v. X and Y are **independent**, and we note $\mathrm{X} \perp\!\!\!\perp \mathrm{Y}$, if $p_{\mathrm{X},\mathrm{Y}} = p_{\mathrm{X}}\, p_{\mathrm{Y}}$.

This probabilistic notion is similar to the intuitive notion; for example, the result of the blue die does not influence the result of the red die, the results are independent.

T63 Independence, uncorrelatedness

We say that the r.v. X and Y are **independent**, and we note $\mathrm{X} \perp\!\!\!\perp \mathrm{Y}$, if $p_{\mathrm{X},\mathrm{Y}} = p_{\mathrm{X}}\, p_{\mathrm{Y}}$.

This probabilistic notion is similar to the intuitive notion; for example, the result of the blue die does not influence the result of the red die, the results are independent.

- We say that the numerical r.v. X and Y are **uncorrelated** if their covariance $C_{\mathrm{X},\mathrm{Y}}$ is null.

T63 Independence, uncorrelatedness

We say that the r.v. X and Y are **independent**, and we note $\mathrm{X} \perp\!\!\!\perp \mathrm{Y}$, if $p_{\mathrm{X},\mathrm{Y}} = p_{\mathrm{X}}\, p_{\mathrm{Y}}$.

This probabilistic notion is similar to the intuitive notion; for example, the result of the blue die does not influence the result of the red die, the results are independent.

- We say that the numerical r.v. X and Y are **uncorrelated** if their covariance $C_{\mathrm{X},\mathrm{Y}}$ is null.

Independence implies uncorrelatedness, but the converse is false in general.

T64 ## Pearson's correlation coefficient

If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{\mathrm{X},\mathrm{Y}}$ by:

$$\rho_{\mathrm{X},\mathrm{Y}} = \frac{C_{\mathrm{X},\mathrm{Y}}}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} \tag{1.41}$$

T64 # Pearson's correlation coefficient

If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{\mathrm{X},\mathrm{Y}}$ by:

$$\rho_{\mathrm{X},\mathrm{Y}} = \frac{C_{\mathrm{X},\mathrm{Y}}}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} \tag{1.41}$$

➔ From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1. Obviously, $\rho_{\mathrm{X},\mathrm{X}} = 1$.

What can we deduce from its value concerning the dependence between two r.v.?

- $\rho_{\mathrm{X},\mathrm{Y}} = 0$: maybe independent;
- $\rho_{\mathrm{X},\mathrm{Y}} \neq 0$: dependent;
- $\rho_{\mathrm{X},\mathrm{Y}} = 1$: linearly dependent, the realizations of (X,Y) are on a rising line;
- $\rho_{\mathrm{X},\mathrm{Y}} = -1$: linearly dependent, the realizations of (X,Y) are on a falling line.

T64 Pearson's correlation coefficient

If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{\mathrm{X,Y}}$ by:

$$\rho_{\mathrm{X,Y}} = \frac{C_{\mathrm{X,Y}}}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} \tag{1.41}$$

- From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1. Obviously, $\rho_{\mathrm{X,X}} = 1$.

  What can we deduce from its value concerning the dependence between two r.v.?

  - $\rho_{\mathrm{X,Y}} = 0$: maybe independent;
  - $\rho_{\mathrm{X,Y}} \neq 0$: dependent;
  - $\rho_{\mathrm{X,Y}} = 1$: linearly dependent, the realizations of (X,Y) are on a rising line;
  - $\rho_{\mathrm{X,Y}} = -1$: linearly dependent, the realizations of (X,Y) are on a falling line.

T64 ## Pearson's correlation coefficient

If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{X,Y}$ by:

$$\rho_{X,Y} = \frac{C_{X,Y}}{\sigma_X \, \sigma_Y} \tag{1.41}$$

- From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1. Obviously, $\rho_{X,X} = 1$.

What can we deduce from its value concerning the dependence between two r.v.?

- $\rho_{X,Y} = 0$: maybe independent;
- $\rho_{X,Y} \neq 0$: dependent;
- $\rho_{X,Y} = 1$: linearly dependent, the realizations of (X, Y) are on a rising line;
- $\rho_{X,Y} = -1$: linearly dependent, the realizations of (X, Y) are on a falling line.

T64 # Pearson's correlation coefficient

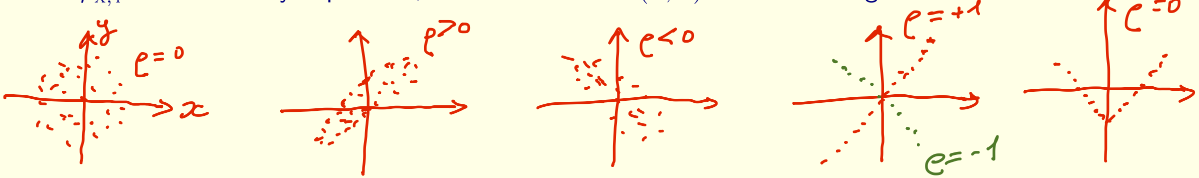
If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{\mathrm{X},\mathrm{Y}}$ by:

$$\rho_{\mathrm{X},\mathrm{Y}} = \frac{C_{\mathrm{X},\mathrm{Y}}}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} \tag{1.41}$$

➔ From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1. Obviously, $\rho_{\mathrm{X},\mathrm{X}} = 1$.

What can we deduce from its value concerning the dependence between two r.v.?

- $\rho_{\mathrm{X},\mathrm{Y}} = 0$: maybe independent;
- $\rho_{\mathrm{X},\mathrm{Y}} \neq 0$: dependent;
- $\rho_{\mathrm{X},\mathrm{Y}} = 1$: linearly dependent, the realizations of (X,Y) are on a rising line;
- $\rho_{\mathrm{X},\mathrm{Y}} = -1$: linearly dependent, the realizations of (X,Y) are on a falling line.



T64 Pearson's correlation coefficient

If X and Y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{\mathrm{X},\mathrm{Y}}$ by:

$$\rho_{\mathrm{X},\mathrm{Y}} = \frac{C_{\mathrm{X},\mathrm{Y}}}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} \tag{1.41}$$

- From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1. Obviously, $\rho_{\mathrm{X},\mathrm{X}} = 1$.

  What can we deduce from its value concerning the dependence between two r.v.?

  - $\rho_{\mathrm{X},\mathrm{Y}} = 0$: maybe independent;
  - $\rho_{\mathrm{X},\mathrm{Y}} \neq 0$: dependent;
  - $\rho_{\mathrm{X},\mathrm{Y}} = 1$: linearly dependent, the realizations of (X,Y) are on a rising line;
  - $\rho_{\mathrm{X},\mathrm{Y}} = -1$: linearly dependent, the realizations of (X,Y) are on a falling line.

- In both last cases, the line contains the point $(m_{\mathrm{X}}, m_{\mathrm{Y}})$, its slope is $\pm\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}$.

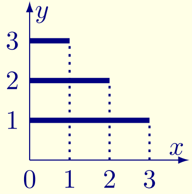
T65 Denoting U_k the kth scalar component of a r.v. U which takes its value in $\mathbb{R}^d$, the variance of U is:

$$C_{\mathrm{U},\mathrm{U}} = \left[\rho_{\mathrm{U}_\ell,\mathrm{U}_k}\, \sigma_{\mathrm{U}_\ell}\, \sigma_{\mathrm{U}_k}\right]_{\substack{1\leqslant \ell \leqslant d \\ 1\leqslant k\leqslant d}} \tag{1.42}$$

The main diagonal contains the variance of the scalar components.

$$U = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} \qquad C_{UU} = Var(U) = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}$$

T66 ◁ ▷ **Exercise 13** (Pair of r.v.: red thread). Let (X, Y) be a pair of r.v. where X is a continuous valued one and Y is a discrete valued one. We suppose that $p_{X,Y}$ is constant over the support $([0,3] \times \{1\}) \cup ([0,2] \times \{2\}) \cup ([0,1] \times \{3\})$.



a) What is the correlation coefficient between X and Y?

b) Are X and Y independent?

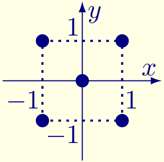
a) $E(XY) = \sum_y \int x y \, p_{XY}(x,y) \, dx = \frac{5}{3}$

$Cov(X,Y) = E(XY) - E(X)E(Y) = -\frac{5}{18}$

$\rho_{XY} = -\sqrt{\frac{5}{23}} \simeq -0,4663$

b) No! They are dependent since $\rho_{XY} \neq 0$!

T67 ◁ ▷ **Evaluation, module 6.** (X, Y) is a pair of discrete r.v. The PMF $p_{X,Y}$ is constant over the support $\{(0,0), (-1,-1), (1,1), (-1,1), (1,-1)\}$.



a) Fill the table below.

x	y	$p_{X,Y}(x,y)$	$p_X(x)$	$p_Y(y)$	$p_X(x)\,p_Y(y)$
-1	-1				
-1	1				
0	0				
1	-1				
1	1				

b) What is the correlation coefficient? Are X and Y independent?

T68 Pair of r.v.: conditional distributions

Définitions

Let (X, Y) be a pair of r.v.
We look at the distribution of X, when Y is fixed.
For example, what is the distribution of the mark of the math exam, in the universe restricted to the students whose mark of the physics exam is 5 over 10?

T68 # Pair of r.v.: conditional distributions

Définitions

Let (X, Y) be a pair of r.v.
We look at the distribution of X, when Y is fixed.
For example, what is the distribution of the mark of the math exam, in the universe restricted to the students whose mark of the physics exam is 5 over 10?

- From the conditional probability measure in the universe, we build the probability distribution of X given Y:

$$\left(y \in \mathrm{S}\left(\mathrm{Y}\right), \mathbb{A} \in \mathcal{P}(\mathbb{X})\right) \longmapsto \mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y} = y\right)$$

T68 # Pair of r.v.: conditional distributions

Définitions

Let (X, Y) be a pair of r.v.
We look at the distribution of X, when Y is fixed.
For example, what is the distribution of the mark of the math exam, in the universe restricted to the students whose mark of the physics exam is 5 over 10?

- From the conditional probability measure in the universe, we build the probability distribution of X given Y:

$$(y \in \mathrm{S}(\mathrm{Y}), \mathbb{A} \in \mathcal{P}(\mathbb{X})) \longmapsto \mathrm{Prob}(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y} = y)$$

In the theory of continuous r.v., the event $\mathrm{Y} = y$ has in general a null probability. We will admit that the conditional probability $\mathrm{Prob}(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y} = y) = \frac{\mathrm{Prob}((\mathrm{X} \in \mathbb{A}) \cap (\mathrm{Y} = y))}{\mathrm{Prob}(\mathrm{Y} = y)}$ which is the ratio between two null probabilities, is meaningful.

T69 It is often useful to calculate a conditional probability to belong to a set which depends of the condition. We will write the conditional probability distribution under the form:

$$\left(y \in \mathrm{S}\left(\mathrm{Y}\right), \mathbb{A} : \mathbb{Y} \to \mathcal{P}(\mathbb{X})\right) \longmapsto \mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}(\mathrm{Y}) \mid \mathrm{Y} = y\right)$$

T70 Let's suppose that it exists a PDF (or PMF) such that, for all $y \in \mathrm{S}\left(\mathrm{Y}\right)$:

$$\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}(\mathrm{Y}) \mid \mathrm{Y} = y\right) = \int_{\mathbb{A}(y)} p_{\mathrm{X}|\mathrm{Y}}(x, y)\, \mathrm{d}\, x \tag{1.43}$$

T70 Let's suppose that it exists a PDF (or PMF) such that, for all $y \in \mathrm{S}\,(\mathrm{Y})$:

$$\mathrm{Prob}\,(\mathrm{X} \in \mathbb{A}(\mathrm{Y}) \mid \mathrm{Y} = y) = \int_{\mathbb{A}(y)} p_{\mathrm{X}|\mathrm{Y}}(x, y)\,\mathrm{d}\,x \tag{1.43}$$

- This PDF is necessarily, for all $(x, y) \in \mathbb{X} \times \mathrm{S}\,(\mathrm{Y})$:[P5]

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X},\mathrm{Y}}(x, y)}{\int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}}(u, y)\,\mathrm{d}\,u} \tag{1.44}$$

$\leftarrow p_Y(y)$

T71 Remind that:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \geqslant 0 \qquad \int_{\mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x, y)\,\mathrm{d}\,x = 1$$

There is no general result on the integration with respect to the second variable.

T71 Remind that:

$$p_{\mathrm{X}|\mathrm{Y}}(x,y) \geqslant 0 \qquad \int_{\mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x,y)\,\mathrm{d}\,x = 1$$

There is no general result on the integration with respect to the second variable.

- For all fixed $y \in \mathrm{S}(\mathrm{Y})$, $\mathrm{S}(\mathrm{X} \mid \mathrm{Y}=y)$ is the support of the conditional distribution:

$$\mathrm{S}(\mathrm{X} \mid \mathrm{Y}=y) = \{x \in \mathbb{X} \mid p_{\mathrm{X}|\mathrm{Y}}(x,y) > 0\} \tag{1.45}$$

T72 The formulae (1.23) and (1.44) permit to obtain the marginal PDF p_{Y} and the conditional PDF $p_{\mathrm{X}|\mathrm{Y}}$ in function of the joint PDF $p_{\mathrm{X},\mathrm{Y}}$.

T72 The formulae (1.23) and (1.44) permit to obtain the marginal PDF p_{Y} and the conditional PDF $p_{\mathrm{X|Y}}$ in function of the joint PDF $p_{\mathrm{X,Y}}$.

- Conversely, by combining both formulae, we obtain the join PDF given the conditional and marginal PDFs; for all (x, y):

$$p_{\mathrm{X,Y}}(x,y) = \begin{cases} p_{\mathrm{X|Y}}(x,y)\, p_{\mathrm{Y}}(y) & \text{if } y \in \mathrm{S}(\mathrm{Y}) \\ 0 & \text{otherwise} \end{cases} \tag{1.46}$$

$$P_{X,Y} = P_{X|Y} \; P_Y$$
$$= P_X \; P_{Y|X}$$

T72 The formulae (1.23) and (1.44) permit to obtain the marginal PDF p_{Y} and the conditional PDF $p_{\mathrm{X}|\mathrm{Y}}$ in function of the joint PDF $p_{\mathrm{X},\mathrm{Y}}$.

- Conversely, by combining both formulae, we obtain the join PDF given the conditional and marginal PDFs; for all (x, y):

$$p_{\mathrm{X},\mathrm{Y}}(x,y) = \begin{cases} p_{\mathrm{X}|\mathrm{Y}}(x,y)\, p_{\mathrm{Y}}(y) & \text{if } y \in \mathrm{S}\,(\mathrm{Y}) \\ 0 & \text{otherwise} \end{cases} \tag{1.46}$$

- By means of this formula and the equivalent one obtained by the exchange of X and Y, we easily obtain the fundamental properties below, in which only the marginal and conditional PDFs appear; for all $(x, y) \in \mathrm{S}\,(\mathrm{X}) \times \mathrm{S}\,(\mathrm{Y})$:

$$p_{\mathrm{X}|\mathrm{Y}}(x,y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y,x)}{p_{\mathrm{Y}}(y)} \qquad \text{(Bayes law)} \tag{1.47}$$

$$p_{\mathrm{Y}}(y) = \int_{\mathrm{S}(\mathrm{X})} p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y,x)\, \mathrm{d}\,x \qquad \text{(total probability law)} \tag{1.48}$$

T73 The independence of two r.v. was defined in the previous section: X et Y are **independent** if $p_{\mathrm{X},\mathrm{Y}} = p_{\mathrm{X}}\, p_{\mathrm{Y}}$.

More naturally, the independence of X and Y means that the distribution of X given Y does not depend of the value taken by Y; thus, the conditional distribution reduces to the marginal one:

$$\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \quad \text{if and only if} \quad \forall (x,y),\ p_{\mathrm{X}|\mathrm{Y}}(x,y) = p_{\mathrm{X}}(x) \tag{1.49}$$

A deterministic variable is independent of every other r.v., since it takes always the same value.

Independence $\quad p_{X,Y} = p_X\, p_Y$

or $\quad p_{X|Y} = p_X$

or $\quad p_{Y|X} = p_Y$

T74 ◁ ▷ **Exercise 14.** Two balls are consecutively drawn from an urn containing 3 white balls and 5 black balls.
The 1st ball is not replaced in the urn before the 2nd sortition (drawing "without replacement").
For each drawing, we earn 1 € for a white ball, 0 € for a black one.
X is the amount earned at the 1st drawing, Y is the amount earned at the 2nd one.

a) Use the assumptions to fill the tables below:

x	$p_{\mathrm{X}}(x)$
0	5/8
1	3/8

x	y	$p_{\mathrm{Y}\vert\mathrm{X}}(y,x)$
0	0	/7
0	1	/7
1	0	/7
1	1	/7

b) Deduce the tables below (the last one correspond to the total earned amount):

y	$p_{\mathrm{Y}}(y)$
0	
1	

x	y	$p_{\mathrm{X},\mathrm{Y}}(x,y)$
0	0	
0	1	
1	0	
1	1	

z	$p_{\mathrm{X}+\mathrm{Y}}(z)$
0	10/28
1	15/28
2	3/28

T75 # Conditional expectation, mean and variance

The **conditional expectation** operator returns, for all numerical function g of a variable in $\mathbb{X} \times \mathrm{S}(\mathrm{Y})$, a function which, for all $y \in \mathrm{S}(\mathrm{Y})$, maps the value noted $\mathrm{E}\left(g(\mathrm{X}, \mathrm{Y}) \mid \mathrm{Y} = y\right)$ and defined by:

$$\mathrm{E}\left(g(\mathrm{X}, \mathrm{Y}) \mid \mathrm{Y} = y\right) = \int_{\mathbb{X}} g(x, y)\, p_{\mathrm{X}|\mathrm{Y}}(x, y)\, \mathrm{d}\, x \tag{1.50}$$

This is the mean of $g(\mathrm{X}, \mathrm{Y})$ given the event $\mathrm{Y} = y$.

T76 Let's temporarily note $\bar{g}$ the function $y \longmapsto \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}=y\right)$.
As every function from $\mathrm{S}(\mathrm{Y})$, it permits to transform the r.v. Y into another r.v. $\bar{g}(\mathrm{Y})$.
Thus, we defined:

- $\bar{g}$ a function defined over $\mathrm{S}(\mathrm{Y})$;
- $\bar{g}(y)$ the image of the realization $y \in \mathrm{S}(\mathrm{Y})$ by this function, noted $\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}=y\right)$;
- $\bar{g}(\mathrm{Y})$ the r.v. coming from the transformation of the r.v. Y by this function, noted $\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)$.

As mentioned above, the definition of the r.v. est equivalent to the definition of the function.

$\bar{g}(y) = E(g(x,y) \mid Y=y)$ image of y

$\bar{g}$ function $S(Y) \longrightarrow$

$\bar{g}(Y) = E(g(x,y) \mid Y)$ new r.v.

T76 Let's temporarily note $\bar{g}$ the function $y \longmapsto \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}=y\right)$.
As every function from $\mathrm{S}\left(\mathrm{Y}\right)$, it permits to transform the r.v. Y into another r.v. $\bar{g}(\mathrm{Y})$.
Thus, we defined:

- $\bar{g}$ a function defined over $\mathrm{S}\left(\mathrm{Y}\right)$;
- $\bar{g}(y)$ the image of the realization $y \in \mathrm{S}\left(\mathrm{Y}\right)$ by this function, noted $\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}=y\right)$;
- $\bar{g}(\mathrm{Y})$ the r.v. coming from the transformation of the r.v. Y by this function, noted $\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)$.

As mentioned above, the definition of the r.v. est equivalent to the definition of the function.

This convention can extend in many ways; for all realization y de Y, can be associated:

- the realization $\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}(\mathrm{Y}) \mid \mathrm{Y}=y\right)$ of the r.v. $\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}(\mathrm{Y}) \mid \mathrm{Y}\right)$;
- the realization $\mathrm{S}\left(\mathrm{X} \mid \mathrm{Y}=y\right)$ of the random support $\mathrm{S}\left(\mathrm{X} \mid \mathrm{Y}\right)$;
- the realization $x \mapsto p_{\mathrm{X}|\mathrm{Y}}(x,y)$ of the random PDF $x \mapsto p_{\mathrm{X}|\mathrm{Y}}(x)$.

T77 **Exercise 15.** Re-write the formulae (1.43) and (1.45) by means of this convention.

$\forall y \in S(Y)$

(1.43) $$\mathrm{Prob}(X \in A(Y) \mid Y = y) = \int_{A(y)} p_{X|Y}(x, y)\, dx$$

(1.45) $$S(X \mid Y = y) = \{x \in \mathcal{X} \mid p_{X|Y}(x, y) > 0\}$$

$$\mathrm{Prob}(X \in A(Y) \mid Y) = \int_{A(Y)} p_{X|Y}(x)\, dx$$

$$S(X \mid Y) = \{x \in \mathcal{X} \mid p_{X|Y}(x) > 0\}$$

T78 For linear transforms, the conditional expectation fulfills the rules below:[16]

$$\mathrm{E}\left(g^{\mathsf{T}}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\right)^{\mathsf{T}} \tag{1.51}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix} \mid \mathrm{Y}\right) = \begin{bmatrix} \mathrm{E}\left(g_{11}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{12}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \\ \mathrm{E}\left(g_{21}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{22}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \end{bmatrix} \tag{1.52}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \tag{1.53}$$

$$\mathrm{E}\left(A(\mathrm{Y})\, g(\mathrm{X},\mathrm{Y})\, C(\mathrm{Y}) + B(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y})\, \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\, C(\mathrm{Y}) + B(\mathrm{Y}) \tag{1.54}$$

$$\mathrm{E}\left(A(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y}) \tag{1.55}$$

16. A, B and C are matrix functions of over $\mathrm{S}(\mathrm{Y})$, such that the sums and the products below are meaningful.

T78 For linear transforms, the conditional expectation fulfills the rules below:[16]

$$\mathrm{E}\left(g^{\mathsf{T}}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\right)^{\mathsf{T}} \tag{1.51}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix} \mid \mathrm{Y}\right) = \begin{bmatrix} \mathrm{E}\left(g_{11}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{12}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \\ \mathrm{E}\left(g_{21}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{22}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \end{bmatrix} \tag{1.52}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \tag{1.53}$$

$$\mathrm{E}\left(A(\mathrm{Y})\, g(\mathrm{X},\mathrm{Y})\, C(\mathrm{Y}) + B(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y})\, \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\, C(\mathrm{Y}) + B(\mathrm{Y}) \tag{1.54}$$

$$\mathrm{E}\left(A(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y}) \tag{1.55}$$

16. A, B and C are matrix functions of over $\mathrm{S}(\mathrm{Y})$, such that the sums and the products below are meaningful.

T78 For linear transforms, the conditional expectation fulfills the rules below:[16]

$$\mathrm{E}\left(g^{\mathsf{T}}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\right)^{\mathsf{T}} \tag{1.51}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix} \mid \mathrm{Y}\right) = \begin{bmatrix} \mathrm{E}\left(g_{11}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{12}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \\ \mathrm{E}\left(g_{21}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{22}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \end{bmatrix} \tag{1.52}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \tag{1.53}$$

$$\mathrm{E}\left(A(\mathrm{Y})\, g(\mathrm{X},\mathrm{Y})\, C(\mathrm{Y}) + B(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y})\ \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\ C(\mathrm{Y}) + B(\mathrm{Y}) \tag{1.54}$$

$$\mathrm{E}\left(A(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y}) \tag{1.55}$$

16. A, B and C are matrix functions of over $\mathrm{S}\left(\mathrm{Y}\right)$, such that the sums and the products below are meaningful.

T78 For linear transforms, the conditional expectation fulfills the rules below:[16]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\right)^{\intercal} \tag{1.51}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix} \mid \mathrm{Y}\right) = \begin{bmatrix} \mathrm{E}\left(g_{11}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{12}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \\ \mathrm{E}\left(g_{21}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{22}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \end{bmatrix} \tag{1.52}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \tag{1.53}$$

$$\mathrm{E}\left(A(\mathrm{Y})\, g(\mathrm{X},\mathrm{Y})\, C(\mathrm{Y}) + B(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y})\ \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\ C(\mathrm{Y}) + B(\mathrm{Y}) \tag{1.54}$$

$$\mathrm{E}\left(A(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y}) \tag{1.55}$$

16. A, B and C are matrix functions of over $\mathrm{S}\left(\mathrm{Y}\right)$, such that the sums and the products below are meaningful.

T78 For linear transforms, the conditional expectation fulfills the rules below:[16]

$$\mathrm{E}\left(g^{\intercal}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \left(\mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\right)^{\intercal} \tag{1.51}$$

$$\mathrm{E}\left(\begin{bmatrix} g_{11}(\mathrm{X},\mathrm{Y}) & g_{12}(\mathrm{X},\mathrm{Y}) \\ g_{21}(\mathrm{X},\mathrm{Y}) & g_{22}(\mathrm{X},\mathrm{Y}) \end{bmatrix} \mid \mathrm{Y}\right) = \begin{bmatrix} \mathrm{E}\left(g_{11}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{12}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \\ \mathrm{E}\left(g_{21}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) & \mathrm{E}\left(g_{22}(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \end{bmatrix} \tag{1.52}$$

$$\mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) + g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) = \mathrm{E}\left(g_1(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) + \mathrm{E}\left(g_2(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right) \tag{1.53}$$

$$\mathrm{E}\left(A(\mathrm{Y})\, g(\mathrm{X},\mathrm{Y})\, C(\mathrm{Y}) + B(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y})\ \mathrm{E}\left(g(\mathrm{X},\mathrm{Y}) \mid \mathrm{Y}\right)\ C(\mathrm{Y}) + B(\mathrm{Y}) \tag{1.54}$$

$$\mathrm{E}\left(A(\mathrm{Y}) \mid \mathrm{Y}\right) = A(\mathrm{Y}) \tag{1.55}$$

16. A, B and C are matrix functions of over $\mathrm{S}\left(\mathrm{Y}\right)$, such that the sums and the products below are meaningful.

T79 $\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)$ is the **conditional mean**.[17]

The **conditional variance** is defined by:[18]

$$\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right) = \mathrm{E}\left(\left(\mathrm{X} - \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\left(\mathrm{X} - \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)^{\intercal} \mid \mathrm{Y}\right) \tag{1.56}$$

17. If $\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) = \mathrm{E}\left(\mathrm{X}\right)$, then X and Y are uncorrelated (ex. 19, page 89).

18. Thus, for all $y \in \mathrm{S}\left(\mathrm{Y}\right)$, we can write: $\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y} = y\right) = \mathrm{E}\left(\left(\mathrm{X} - \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y} = y\right)\right)\left(\mathrm{X} - \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y} = y\right)\right)^{\intercal} \mid \mathrm{Y} = y\right)$

T79 $\mathrm{E}(\mathrm{X} \mid \mathrm{Y})$ is the **conditional mean**.[17]

The **conditional variance** is defined by:[18]

$$\mathrm{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))^{\mathsf{T}} \mid \mathrm{Y}\right) \tag{1.56}$$

➔ The conditional variance can also be written as:

$$\mathrm{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left(\mathrm{X}\,\mathrm{X}^{\mathsf{T}} \mid \mathrm{Y}\right) - \mathrm{E}(\mathrm{X} \mid \mathrm{Y})\,\mathrm{E}\left(\mathrm{X}^{\mathsf{T}} \mid \mathrm{Y}\right) \tag{1.57}$$

17. If $\mathrm{E}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}(\mathrm{X})$, then X and Y are uncorrelated (ex. 19, page 89).

18. Thus, for all $y \in \mathrm{S}(\mathrm{Y})$, we can write: $\mathrm{Var}(\mathrm{X} \mid \mathrm{Y} = y) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))^{\mathsf{T}} \mid \mathrm{Y} = y\right)$

T79 $\mathrm{E}(\mathrm{X} \mid \mathrm{Y})$ is the **conditional mean**.[17]

The **conditional variance** is defined by:[18]

$$\operatorname{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))^{\intercal} \mid \mathrm{Y}\right) \tag{1.56}$$

- The conditional variance can also be written as:

$$\operatorname{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left(\mathrm{X}\,\mathrm{X}^{\intercal} \mid \mathrm{Y}\right) - \mathrm{E}(\mathrm{X} \mid \mathrm{Y})\;\mathrm{E}\left(\mathrm{X}^{\intercal} \mid \mathrm{Y}\right) \tag{1.57}$$

- For a linear transform, the conditional variance fulfills the rule below:[19]

$$\operatorname{Var}(A(\mathrm{Y})\,\mathrm{X} + B(\mathrm{Y}) \mid \mathrm{Y}) = A(\mathrm{Y})\;\operatorname{Var}(\mathrm{X} \mid \mathrm{Y})\;A^{\intercal}(\mathrm{Y}) \tag{1.58}$$

17. If $\mathrm{E}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}(\mathrm{X})$, then X and Y are uncorrelated (ex. 19, page 89).
18. Thus, for all $y \in \mathrm{S}(\mathrm{Y})$, we can write: $\operatorname{Var}(\mathrm{X} \mid \mathrm{Y} = y) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))^{\intercal} \mid \mathrm{Y} = y\right)$
19. A and B are some functions of a variable in $\mathbb{Y}$ such that the sums and products are meaningful.

T79 $\mathrm{E}(\mathrm{X} \mid \mathrm{Y})$ is the **conditional mean**.[17]

The **conditional variance** is defined by:[18]

$$\mathrm{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y}))^{\intercal} \mid \mathrm{Y}\right) \tag{1.56}$$

- The conditional variance can also be written as:

$$\mathrm{Var}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}\left(\mathrm{X}\,\mathrm{X}^{\intercal} \mid \mathrm{Y}\right) - \mathrm{E}(\mathrm{X} \mid \mathrm{Y})\;\mathrm{E}\left(\mathrm{X}^{\intercal} \mid \mathrm{Y}\right) \tag{1.57}$$

- For a linear transform, the conditional variance fulfills the rule below:[19]

$$\mathrm{Var}(A(\mathrm{Y})\,\mathrm{X} + B(\mathrm{Y}) \mid \mathrm{Y}) = A(\mathrm{Y})\;\mathrm{Var}(\mathrm{X} \mid \mathrm{Y})\;A^{\intercal}(\mathrm{Y}) \tag{1.58}$$

- As an immediate corollary:

$$\mathrm{Var}(B(\mathrm{Y}) \mid \mathrm{Y}) \text{ is null} \tag{1.59}$$

17. If $\mathrm{E}(\mathrm{X} \mid \mathrm{Y}) = \mathrm{E}(\mathrm{X})$, then X and Y are uncorrelated (ex. 19, page 89).
18. Thus, for all $y \in \mathrm{S}(\mathrm{Y})$, we can write: $\mathrm{Var}(\mathrm{X} \mid \mathrm{Y} = y) = \mathrm{E}\left((\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))\,(\mathrm{X} - \mathrm{E}(\mathrm{X} \mid \mathrm{Y} = y))^{\intercal} \mid \mathrm{Y} = y\right)$
19. A and B are some functions of a variable in $\mathbb{Y}$ such that the sums and products are meaningful.

T80 Laws based on the total probability law

The total probability law (1.48) can be re-written as (1.60) and permits to obtain the rules below [ref]:[20]

$$\forall x \in \mathbb{X},\ p_{\mathrm{X}}(x) = \mathrm{E}\left(p_{\mathrm{X}|\mathrm{Y}}(x)\right) \quad \text{(total probability)} \tag{1.60}$$

$$\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}\right) = \mathrm{E}\left(\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y}\right)\right) \quad \text{(total probability)} \tag{1.61}$$

$$\mathrm{E}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total expectation)} \tag{1.62}$$

$$\mathrm{Var}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total variance)} \tag{1.63}$$

20. In automatic clustering or in analysis of variance (ANOVA), the first term of the decomposition of the total variance (T, "total") is called within-groups variance (W, "within"), the second term is called between-groups variance (B, "between"): $T = W + B$.

T80 Laws based on the total probability law

The total probability law (1.48) can be re-written as (1.60) and permits to obtain the rules below [ref]:[20]

$$\forall x \in \mathbb{X},\ p_{\mathrm{X}}(x) = \mathrm{E}\left(p_{\mathrm{X}|\mathrm{Y}}(x)\right) \quad \text{(total probability)} \tag{1.60}$$

$$\mathrm{Prob}\,(\mathrm{X} \in \mathbb{A}) = \mathrm{E}\,(\mathrm{Prob}\,(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y})) \quad \text{(total probability)} \tag{1.61}$$

$$\mathrm{E}\,(\mathrm{X}) = \mathrm{E}\,(\mathrm{E}\,(\mathrm{X} \mid \mathrm{Y})) \quad \text{(total expectation)} \tag{1.62}$$

$$\mathrm{Var}\,(\mathrm{X}) = \mathrm{E}\,(\mathrm{Var}\,(\mathrm{X} \mid \mathrm{Y})) + \mathrm{Var}\,(\mathrm{E}\,(\mathrm{X} \mid \mathrm{Y})) \quad \text{(total variance)} \tag{1.63}$$

20. In automatic clustering or in analysis of variance (ANOVA), the first term of the decomposition of the total variance (T, "total") is called within-groups variance (W, "within"), the second term is called between-groups variance (B, "between"): $T = W + B$.

T80 Laws based on the total probability law

The total probability law (1.48) can be re-written as (1.60) and permits to obtain the rules below [ref]:[20]

$$\forall x \in \mathbb{X},\ p_{\mathrm{X}}(x) = \mathrm{E}\left(p_{\mathrm{X}|\mathrm{Y}}(x)\right) \quad \text{(total probability)} \tag{1.60}$$

$$\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}\right) = \mathrm{E}\left(\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y}\right)\right) \quad \text{(total probability)} \tag{1.61}$$

$$\mathrm{E}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total expectation)} \tag{1.62}$$

$$\mathrm{Var}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total variance)} \tag{1.63}$$

20. In automatic clustering or in analysis of variance (ANOVA), the first term of the decomposition of the total variance (T, "total") is called within-groups variance (W, "within"), the second term is called between-groups variance (B, "between"): $T = W + B$.

T80

Laws based on the total probability law

The total probability law (1.48) can be re-written as (1.60) and permits to obtain the rules below [ref]:[20]

$$\forall x \in \mathbb{X},\ p_{\mathrm{X}}(x) = \mathrm{E}\left(p_{\mathrm{X}|\mathrm{Y}}(x)\right) \quad \text{(total probability)} \tag{1.60}$$

$$\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A}\right) = \mathrm{E}\left(\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y}\right)\right) \quad \text{(total probability)} \tag{1.61}$$

$$\mathrm{E}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total expectation)} \tag{1.62}$$

$$\mathrm{Var}\left(\mathrm{X}\right) = \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \quad \text{(total variance)} \tag{1.63}$$

It is a good training to rewrite the total expectation formula by means of the functions $y \mapsto \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y} = y\right)$ and $y \mapsto \mathrm{Prob}\left(\mathrm{Y} = y\right)$ for a discrete valued r.v. Y, and to explain the result.

$$E(X) = E\big(\underbrace{E(X \mid Y)}_{g(Y)}\big)$$

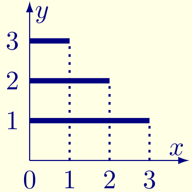
$$= \sum_y g(y)\, \mathrm{Prob}(Y = y)$$

$$= \sum_y E(X \mid Y = y)\ \mathrm{Prob}(Y = y)$$

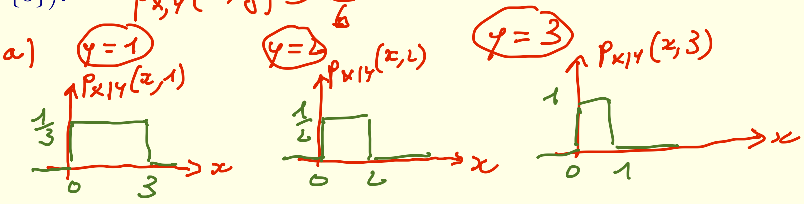
20. In automatic clustering or in analysis of variance (ANOVA), the first term of the decomposition of the total variance (T, "total") is called within-groups variance (W, "within"), the second term is called between-groups variance (B, "between"): $T = W + B$.

T81 ◁ ▷ **Exercise 16** (Pair of r.v.: red thread). Let (X, Y) be a pair of r.v. where X is a continuous valued one and Y is a discrete valued one. We suppose that $p_{X,Y}$ is constant over the support $([0,3] \times \{1\}) \cup ([0,2] \times \{2\}) \cup ([0,1] \times \{3\})$.

$p_{X,Y}(x,y) = \frac{1}{6}$



a)

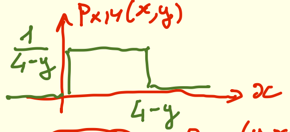


a) Give and plot the PDF of $X \mid Y$; give its mean and its variance.

b) Give the mean and the variance of X by means of the total expectation formula and the total variance one.

c) Give and plot the PMF of $Y \mid X$; give its mean and its variance.

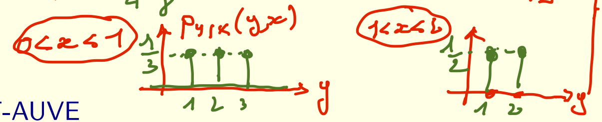
a)



$E(X|Y) = \frac{4-Y}{2}$

$Var(X|Y) = \frac{(4-Y)^2}{12}$

c)



b) $E(X) = E(E(X|Y))$

$= E\left(2 - \frac{Y}{2}\right)$

$= 2 - \frac{1}{2} E(Y)$

Check that $Var(X) = Var\left(\frac{4-Y}{2}\right) + E\left(\frac{(4-Y)^2}{12}\right)$

T82 ◁ ▷ **Evaluation, module 7.** Let's flip a correct coin.
At the first throw, we earn 1 € for "tail", 0 € otherwise.
If we did not earn anything at the first throw, we do not earn anything at the second one.
If we earned 1 € at the first throw, we flip the coin again, we earn again 1 € if we obtain "tail", 0 € otherwise.
X is the earned amount at the first throw, Y is the earned amount at the second throw.

a) Formalize the assumptions by filling the tables below.

x	$p_{\mathrm{X}}(x)$
0	
1	

x	y	$p_{\mathrm{Y}\vert\mathrm{X}}(y,x)$
0	0	
0	1	
1	0	
1	1	

b) Deduce the tables below (the last one corresponds to the total earned amount).

y	$p_{\mathrm{Y}}(y)$
0	
1	

x	y	$p_{\mathrm{X},\mathrm{Y}}(x,y)$
0	0	
0	1	
1	0	
1	1	

z	$p_{\mathrm{X}+\mathrm{Y}}(z)$
0	
1	
2	

T83 # Triplet of random variables

We know how to handle distributions (joint, marginal, conditional) of a pair of r.v..
For a triplet of r.v., we have also to consider, for example, the joint distribution of 2 r.v. given a third one.

T83 Triplet of random variables

We know how to handle distributions (joint, marginal, conditional) of a pair of r.v..
For a triplet of r.v., we have also to consider, for example, the joint distribution of 2 r.v. given a third one.

- Let (X, Y, Z) be a triplet of r.v.
  The PDF of the pair (X, Y) given Z is noted $p_{(X,Y)|Z}$
  The PDF of X given the pair (Y, Z) is noted $p_{X|(Y,Z)}$.
  In this section, we mainly examine distributions given the r.v. Z.

T84 **Joint, marginal, conditional distributions**

The PDF of the marginal v.a. Y given Z is, for all $y \in \mathbb{Y}$:[21]

$$p_{\mathrm{Y}|\mathrm{Z}}(y) = \int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y)\,\mathrm{d}\,x \tag{1.64}$$

21. That is, for all $(y, z) \in \mathbb{Y} \times \mathrm{S}(\mathrm{Z})$, $p_{\mathrm{Y}|\mathrm{Z}}(y, z) = \int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y, z)\,\mathrm{d}\,x$.

T84 Joint, marginal, conditional distributions

The PDF of the marginal v.a. Y given Z is, for all $y \in \mathbb{Y}$:[21]

$$p_{\mathrm{Y}|\mathrm{Z}}(y) = \int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y) \,\mathrm{d}\, x \tag{1.64}$$

➔ The PDF of X given the pair (Y, Z) is, for all $(x, y) \in \mathbb{X} \times \mathrm{S}\,(\mathrm{Y})$:

$$p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}}(x, y) = \frac{p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y)}{\displaystyle\int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(u, y) \,\mathrm{d}\, u} \tag{1.65}$$

21. That is, for all $(y, z) \in \mathbb{Y} \times \mathrm{S}\,(\mathrm{z})$, $p_{\mathrm{Y}|\mathrm{Z}}(y, z) = \displaystyle\int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y, z) \,\mathrm{d}\, x$.

T84 Joint, marginal, conditional distributions

The PDF of the marginal v.a. Y given Z is, for all $y \in \mathbb{Y}$:[21]

$$p_{\mathrm{Y}|\mathrm{Z}}(y) = \int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y)\,\mathrm{d}\,x \tag{1.64}$$

- The PDF of X given the pair (Y, Z) is, for all $(x, y) \in \mathbb{X} \times \mathrm{S}(\mathrm{Y})$:

$$p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}}(x, y) = \frac{p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y)}{\displaystyle\int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(u, y)\,\mathrm{d}\,u} \tag{1.65}$$

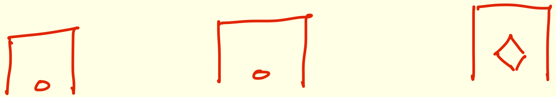
- From these PDFs, we can retrieve the joint distribution of the pair (X, Y) given Z, for all (x, y), using:

$$p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y) = \begin{cases} p_{\mathrm{X}|\mathrm{Y}\mathrm{Z}}(x, y)\, p_{\mathrm{Y}|\mathrm{Z}}(y) & \text{if } y \in \mathrm{S}(\mathrm{Y} \mid \mathrm{Z}) \\ 0 & \text{otherwise} \end{cases} \tag{1.66}$$

21. That is, for all $(y, z) \in \mathbb{Y} \times \mathrm{S}(\mathrm{Z})$, $p_{\mathrm{Y}|\mathrm{Z}}(y, z) = \int_{\mathbb{X}} p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}}(x, y, z)\,\mathrm{d}\,x$.

T85 ◁ ▷ **Exercise 17.** Write the Bayes law and the total probability law which link $p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}}$, $p_{\mathrm{X}|\mathrm{Z}}$, $p_{\mathrm{Y}|\mathrm{X},\mathrm{Z}}$ and $p_{\mathrm{Y}|\mathrm{Z}}$, for all $(x, y) \in \mathrm{S}(\mathrm{X} \mid \mathrm{Z}) \times \mathrm{S}(\mathrm{Y} \mid \mathrm{Z})$.

T86 We will apply the results of the previous exercise to solve the Monty Hall problem (name of the host of an American television game show).
Three opaque cups are upside down on the table. One of them hides a diamond, the other ones hide a bean. The quizmaster knows where is the diamond.
The gambler designates a cup (nobody lifts the cup).
The quizmaster lifts, among the two other cups, one which does not hide the diamond.
The gambler must lift a cup, he earns the object under it.



T87 ◁ ▷ **Exercise 18** (Monty Hall problem). The cups are numbered from 1 to 3; we note:

- D the number which hides the diamond,
- G the number pointed out by the gambler,
- Q the number lifted by the quizmaster

Fill the table. What is the probability to earn the diamond, if the gambler:

- lifts always the initially pointed out cup? 1/3
- lifts always the not initially pointed out cup? 2/3
- lifts one of them at random? 1/2

q	$p_{Q\|D,G}(q,d,g)$			$\sum_d p_{Q\|D,G}(q,d,g)$	$p_{D\|Q,G}(d,q,g)$			g
1								1
2								1
3								1
1								2
2								2
3								2
1								3
2								3
3								3
	1	2	3		1	2	3	
	d				d			

T88 Conditional covariance

The **conditional covariance** is defined by:

$$\mathrm{Cov}\left(\mathrm{X}, \mathrm{Y} \mid \mathrm{Z}\right) = \mathrm{E}\left(\left(\mathrm{X} - \mathrm{E}\left(\mathrm{X} \mid \mathrm{Z}\right)\right)\left(\mathrm{Y} - \mathrm{E}\left(\mathrm{Y} \mid \mathrm{Z}\right)\right)^{\intercal} \mid \mathrm{Z}\right) \tag{1.67}$$

T88 Conditional covariance

The **conditional covariance** is defined by:

$$\mathrm{Cov}\left(X, Y \mid Z\right) = \mathrm{E}\left(\left(X - \mathrm{E}\left(X \mid Z\right)\right)\left(Y - \mathrm{E}\left(Y \mid Z\right)\right)^{\mathsf{T}} \mid Z\right) \tag{1.67}$$

- It is straightforward to show that the conditional covariance can also be written as:

$$\mathrm{Cov}\left(X, Y \mid Z\right) = \mathrm{E}\left(X\,Y^{\mathsf{T}} \mid Z\right) - \mathrm{E}\left(X \mid Z\right)\,\mathrm{E}\left(Y^{\mathsf{T}} \mid Z\right) \tag{1.68}$$

T88 Conditional covariance

The **conditional covariance** is defined by:

$$\operatorname{Cov}(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}) = \mathrm{E}\left((\mathrm{X}-\mathrm{E}(\mathrm{X}\mid \mathrm{Z}))\,(\mathrm{Y}-\mathrm{E}(\mathrm{Y}\mid \mathrm{Z}))^{\mathsf{T}} \mid \mathrm{Z}\right) \tag{1.67}$$

- It is straightforward to show that the conditional covariance can also be written as:

$$\operatorname{Cov}(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}) = \mathrm{E}\left(\mathrm{X}\,\mathrm{Y}^{\mathsf{T}} \mid \mathrm{Z}\right) - \mathrm{E}(\mathrm{X}\mid \mathrm{Z})\;\mathrm{E}\left(\mathrm{Y}^{\mathsf{T}}\mid \mathrm{Z}\right) \tag{1.68}$$

- For linear transforms, this operator fulfills the rules below:[22]

$$\operatorname{Cov}(A(\mathrm{Z})\,\mathrm{X}+B(\mathrm{Z}), C(\mathrm{Z})\,\mathrm{Y}+D(\mathrm{Z})\mid \mathrm{Z}) = A(\mathrm{Z})\;\operatorname{Cov}(\mathrm{X},\mathrm{Y}\mid \mathrm{Z})\;C^{\mathsf{T}}(\mathrm{Z})$$

22. A, B, C and D are some functions, such that the sums and products below are meaningful.

T88 Conditional covariance

The **conditional covariance** is defined by:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}\right)=\mathrm{E}\left(\left(\mathrm{X}-\mathrm{E}\left(\mathrm{X}\mid \mathrm{Z}\right)\right)\,\left(\mathrm{Y}-\mathrm{E}\left(\mathrm{Y}\mid \mathrm{Z}\right)\right)^{\mathsf{T}}\;\mid \mathrm{Z}\right) \tag{1.67}$$

- It is straightforward to show that the conditional covariance can also be written as:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}\right)=\mathrm{E}\left(\mathrm{X}\,\mathrm{Y}^{\mathsf{T}}\mid \mathrm{Z}\right)-\mathrm{E}\left(\mathrm{X}\mid \mathrm{Z}\right)\,\mathrm{E}\left(\mathrm{Y}^{\mathsf{T}}\mid \mathrm{Z}\right) \tag{1.68}$$

- For linear transforms, this operator fulfills the rules below:[22]

$$\mathrm{Cov}\left(A(\mathrm{Z})\,\mathrm{X}+B(\mathrm{Z}),C(\mathrm{Z})\,\mathrm{Y}+D(\mathrm{Z})\mid \mathrm{Z}\right)=A(\mathrm{Z})\;\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}\right)\;C^{\mathsf{T}}(\mathrm{Z})$$

- The total expectation formula permits to obtain the **total covariance formula**:

$$\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\right)=\mathrm{E}\left(\mathrm{Cov}\left(\mathrm{X},\mathrm{Y}\mid \mathrm{Z}\right)\right)+\mathrm{Cov}\left(\mathrm{E}\left(\mathrm{X}\mid \mathrm{Z}\right),\mathrm{E}\left(\mathrm{Y}\mid \mathrm{Z}\right)\right) \tag{1.69}$$

22. A, B, C and D are some functions, such that the sums and products below are meaningful.

T89 ◁ ▷ **Exercise 19** (Laws of total expectation and total covariance). Give $\mathrm{Cov}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right), \mathrm{Y}\right)$ in function of $\mathrm{Cov}\left(\mathrm{X}, \mathrm{Y}\right)$:

a) by means of the law of total expectation (1.62);

b) ~~by means of the law of total covariance (1.69) (in which we take $\mathrm{Y} = \mathrm{Z}$).~~

c) If $\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) = \mathrm{E}\left(\mathrm{X}\right)$, what is the covariance $\mathrm{Cov}\left(\mathrm{X}, \mathrm{Y}\right)$?

a)
$$\begin{aligned}\mathrm{Cov}(E(X|Y), Y) &= E\left(E(X|Y)\,Y^T\right) - E\left(E(X|Y)\right) E(Y)\\ &= E\left(E(XY^T|Y)\right) - E\left(E(X|Y)\right) E(Y)\\ &= E(XY^T) - E(X)\,E(Y)\\ &= \mathrm{Cov}(X,Y)\end{aligned}$$

c)
$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(E(X|Y), Y) && \text{(see previous question)}\\ &= \mathrm{Cov}(E(X), Y) && \text{(assumption of (c))}\\ &= 0\end{aligned}$$

T90 Conditional independence

X et Y are independent given Z, which is noted $\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z}$, if $p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}}\, p_{\mathrm{Y}|\mathrm{Z}}$.

T90 Conditional independence

X et Y are independent given Z, which is noted $\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z}$, if $p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}}\, p_{\mathrm{Y}|\mathrm{Z}}$.

- The equivalence below holds [ref]:

$$\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z} \quad \text{if and only if} \quad p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}} \tag{1.70}$$

That means that, in the pair (Y, Z), Z catch all the information on X (we often say that X depends only of Z).

T90 Conditional independence

X et Y are independent given Z, which is noted $\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z}$, if $p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}}\, p_{\mathrm{Y}|\mathrm{Z}}$.

The equivalence below holds [ref]:

$$\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z} \quad \text{if and only if} \quad p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}} \tag{1.70}$$

That means that, in the pair (Y, Z), Z catch all the information on X (we often say that X depends only of Z).

The conditional independence do not imply the independence. For example, we suppose that in an exam, each sheet is marked by two professors. Z is the sheet value, X and Y are the marks. Intuitively, all sheets taken together, both marks are dependent with a positive correlation. However, for a given sheet, the professors have to mark independently (the best is that they do not communicate).

T90 # Conditional independence

X et Y are independent given Z, which is noted $\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z}$, if $p_{\mathrm{X},\mathrm{Y}|\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}}\, p_{\mathrm{Y}|\mathrm{Z}}$.

- The equivalence below holds [ref]:

$$\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z} \quad \text{if and only if} \quad p_{\mathrm{X}|\mathrm{Y},\mathrm{Z}} = p_{\mathrm{X}|\mathrm{Z}} \tag{1.70}$$

That means that, in the pair (Y, Z), Z catch all the information on X (we often say that X depends only of Z).

- The conditional independence do not imply the independence. For example, we suppose that in an exam, each sheet is marked by two professors. Z is the sheet value, X and Y are the marks. Intuitively, all sheets taken together, both marks are dependent with a positive correlation. However, for a given sheet, the professors have to mark independently (the best is that they do not communicate).

- The independence does not imply the conditional independence. Let's play to heads or tails with two coins. The two results are independent.
Let's introduce the boolean variable which is true is the results are equal. Given this r.v., the two results are not independent, since the value of one of them implies the value of the other one.

T91 The equivalence below holds:

$\Leftrightarrow X \perp\!\!\!\perp Z \mid Y$ and $X \perp\!\!\!\perp Y$

$$X \perp\!\!\!\perp (Y, Z) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z \text{ and } X \perp\!\!\!\perp Z \tag{1.71}$$

If X is independent of the pair (Y, Z), then X is independent of Y, and X is independent of Z (the independence of the whole implies the independence of each part), but the converse if false in general.

T92 # Recursive aspects of probability calculations

The formula (1.46) gives the joint distribution in function of the marginal distribution and the conditional distribution. By induction, we obtain (and that can be generalized to every number of r.v.):

$$p_{\mathrm{X,Y,Z}} = p_{\mathrm{X|Y,Z}}\, p_{\mathrm{Y|Z}}\, p_{\mathrm{Z}} \tag{1.72}$$

T92 **Recursive aspects of probability calculations**

The formula (1.46) gives the joint distribution in function of the marginal distribution and the conditional distribution. By induction, we obtain (and that can be generalized to every number of r.v.):

$$p_{\mathrm{X,Y,Z}} = p_{\mathrm{X|Y,Z}}\, p_{\mathrm{Y|Z}}\, p_{\mathrm{Z}} \tag{1.72}$$

- The rules (1.60) to (1.63) can be interpreted as the last stage of a recursive calculation whose preceding stage is:

$$\forall x \in \mathbb{X},\ p_{\mathrm{X|Z}}(x) = \mathrm{E}\left(p_{\mathrm{X|Y,Z}}(x) \mid \mathrm{Z}\right) \qquad \text{(total probability)} \tag{1.73}$$

$$\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Z}\right) = \mathrm{E}\left(\mathrm{Prob}\left(\mathrm{X} \in \mathbb{A} \mid \mathrm{Y,Z}\right) \mid \mathrm{Z}\right) \qquad \text{(total probability)} \tag{1.74}$$

$$\mathrm{E}\left(\mathrm{X} \mid \mathrm{Z}\right) = \mathrm{E}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y,Z}\right) \mid \mathrm{Z}\right) \qquad \text{(total expectation)} \tag{1.75}$$

$$\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Z}\right) = \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y,Z}\right) \mid \mathrm{Z}\right) + \mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y,Z}\right) \mid \mathrm{Z}\right) \qquad \text{(total variance)} \tag{1.76}$$

T93 Mutual independence

We introduced only the independence between 2 r.v..
The **mutual independence** between the 3 r.v. (X, Y, Z) can be defined as follows: X is independent of the pair (Y, Z) and Y is independent of Z.
By induction, we can define the mutual independence between any number of r.v..

T93 Mutual independence

We introduced only the independence between 2 r.v..
The **mutual independence** between the 3 r.v. $(\mathrm{X}, \mathrm{Y}, \mathrm{Z})$ can be defined as follows: X is independent of the pair (Y, Z) and Y is independent of Z.
By induction, we can define the mutual independence between any number of r.v..

- We easily check that some r.v. are mutually independent if and only if the PDF of the joint distribution is the product of the marginal PDFs, that is, in the case of 3 r.v., by means of the formula (1.72):

$$p_{\mathrm{X},\mathrm{Y},\mathrm{Z}} = p_{\mathrm{X}}\, p_{\mathrm{Y}}\, p_{\mathrm{Z}} \quad (\mathrm{X},\ \mathrm{Y},\ \mathrm{Z} \text{ mutually independent})$$

T93 Mutual independence

We introduced only the independence between 2 r.v..
The **mutual independence** between the 3 r.v. (X, Y, Z) can be defined as follows: X is independent of the pair (Y, Z) and Y is independent of Z.
By induction, we can define the mutual independence between any number of r.v..

- We easily check that some r.v. are mutually independent if and only if the PDF of the joint distribution is the product of the marginal PDFs, that is, in the case of 3 r.v., by means of the formula (1.72):

$$p_{X,Y,Z} = p_X \, p_Y \, p_Z \quad (X,\ Y,\ Z \text{ mutually independent})$$

- The mutual independence of the triplet (X, Y, Z) implies the independance of the pair (X, Y), of the pair (X, Z), and of the pair (Y, Z) (the mutual independence implies the pairwise independence), but the converse doest not hold in general.

T94 ◁ ▷ **Exercise 20** (Training for Markov chains)**.** Let's play to "heads or tails" with a correct coin. We earn 1 € if we obtain "tail", 0 € otherwise.
We throw the coin 3 times, consecutively.
Y_i is the amount of money earned at throw $i \in \{1, 2, 3\}$.
X_i is the cumulated amount of money till throw $\mathrm{X}_i = \sum_{j=1}^{i} \mathrm{Y}_j$.

a) Fill the left side table, and deduce the right side one.

x_1	x_2	$p_{\mathrm{X}_1,\mathrm{X}_2,\mathrm{X}_3}(x_1, x_2, x_3)$				$p_{\mathrm{X}_1,\mathrm{X}_2}(x_1, x_2)$
0	0					
0	1					
0	2					
1	0					
1	1					
1	2					
	x_3	0	1	2	3	

x_1	x_2	$p_{\mathrm{X}_3\|\mathrm{X}_1,\mathrm{X}_2}(x_3, x_1, x_2)$			
0	0				
0	1				
1	1				
1	2				
	x_3	0	1	2	3

b) Are X_3 and X_1 independent given X_2?

c) Are X_3 and X_2 independent given X_1?

T95 ◁ ▷ **Evaluation, module 8.** X, Y et Z are three r.v. which take their value in $\{0, 1\}$; the joint PMF is:

$$p_{\mathrm{X,Y,Z}}(x, y, z) = \frac{1}{2}\,\delta(y - x)\,\delta(z - 1 + x)$$

a) Fill the table below.

y	z	$p_{\mathrm{Y,Z} \mid \mathrm{X}}(y, z, x)$		$p_{\mathrm{Y} \mid \mathrm{X}}(y, x)$		$p_{\mathrm{Z} \mid \mathrm{X}}(z, x)$		$p_{\mathrm{Y,Z}}(y, z)$	$p_{\mathrm{Y}}(y)$	$p_{\mathrm{Z}}(z)$
0	0									
0	1									
1	0									
1	1									
		0	1	0	1	0	1			
		x		x		x				

b) Are Y and Z independent?

c) Are Y and Z independent given X?

T96 From probabilities to statistics

Random sampling

Let $\mathrm{x} : \Omega \to \mathbb{X}$ be a r.v.

It is equivalent to say:

- We obtained n_r realizations of x (from n_r draws with replacement in Ω);
- We obtained a realization of the n_r-tuple of r.v. $\vec{\mathrm{X}}_{n_\mathrm{r}} = (\mathrm{x}_1, \ldots, \mathrm{x}_{n_\mathrm{r}})$ where the r.v. x_q, $1 \leqslant q \leqslant n_\mathrm{r}$ are mutually independent with the same distribution as x.[23]

23. Strictly speaking, for all q, x_q is a r.v. from the universe Ω^{n_r} to $\mathbb{X}$ where the probability measure of Ω^{n_r} comes from the probability measure of Ω using $\mathrm{Prob}\left(\Phi_1 \times \ldots \times \Phi_{n_\mathrm{r}}\right) = \prod_{q=1}^{n_\mathrm{r}} \mathrm{Prob}\left(\Phi_q\right)$, and thus, $\mathrm{x}_q(\omega_1, \ldots, \omega_{n_\mathrm{r}}) = \mathrm{x}(\omega_q)$.

T96 From probabilities to statistics

Random sampling

Let $\mathrm{x} : \Omega \to \mathbb{X}$ be a r.v.

It is equivalent to say:

- We obtained n_r realizations of x (from n_r draws with replacement in Ω);
- We obtained a realization of the n_r-tuple of r.v. $\vec{\mathrm{X}}_{n_\mathrm{r}} = (\mathrm{x}_1, \ldots, \mathrm{x}_{n_\mathrm{r}})$ where the r.v. x_q, $1 \leqslant q \leqslant n_\mathrm{r}$ are mutually independent with the same distribution as x.[23]

➲ We also say that $\vec{\mathrm{X}}_{n_\mathrm{r}}$ is a **random sample** of x, with size n_r (or a n_r**-sample** of x).

23. Strictly speaking, for all q, x_q is a r.v. from the universe Ω^{n_r} to $\mathbb{X}$ where the probability measure of Ω^{n_r} comes from the probability measure of Ω using $\mathrm{Prob}\left(\Phi_1 \times \ldots \times \Phi_{n_\mathrm{r}}\right) = \prod_{q=1}^{n_\mathrm{r}} \mathrm{Prob}\left(\Phi_q\right)$, and thus, $\mathrm{x}_q(\omega_1, \ldots, \omega_{n_\mathrm{r}}) = \mathrm{x}(\omega_q)$.

T96 # From probabilities to statistics

Random sampling

Let $\mathrm{x} : \Omega \to \mathbb{X}$ be a r.v.

It is equivalent to say:

- We obtained n_r realizations of x (from n_r draws with replacement in Ω);
- We obtained a realization of the n_r-tuple of r.v. $\overrightarrow{\mathrm{X}}_{n_\mathrm{r}} = (\mathrm{x}_1, \ldots, \mathrm{x}_{n_\mathrm{r}})$ where the r.v. x_q, $1 \leqslant q \leqslant n_\mathrm{r}$ are mutually independent with the same distribution as x.[23]

➔ We also say that $\overrightarrow{\mathrm{X}}_{n_\mathrm{r}}$ is a **random sample** of x, with size n_r (or a n_r-**sample** of x).

➔ We say that $\overrightarrow{\mathrm{X}}_{n_\mathrm{r}}$ is a set of **independent and identically distributed (i.i.d.)** r.v.

23. Strictly speaking, for all q, x_q is a r.v. from the universe Ω^{n_r} to $\mathbb{X}$ where the probability measure of Ω^{n_r} comes from the probability measure of Ω using $\mathrm{Prob}\left(\Phi_1 \times \ldots \times \Phi_{n_\mathrm{r}}\right) = \prod_{q=1}^{n_\mathrm{r}} \mathrm{Prob}\left(\Phi_q\right)$, and thus, $\mathrm{x}_q(\omega_1, \ldots, \omega_{n_\mathrm{r}}) = \mathrm{x}(\omega_q)$.

T96 # From probabilities to statistics

Random sampling

Let $\mathrm{x} : \Omega \to \mathbb{X}$ be a r.v.

It is equivalent to say:

- We obtained n_r realizations of x (from n_r draws with replacement in Ω);
- We obtained a realization of the n_r-tuple of r.v. $\vec{\mathrm{X}}_{n_\mathrm{r}} = (\mathrm{x}_1, \ldots, \mathrm{x}_{n_\mathrm{r}})$ where the r.v. x_q, $1 \leqslant q \leqslant n_\mathrm{r}$ are mutually independent with the same distribution as x.[23]

- We also say that $\vec{\mathrm{X}}_{n_\mathrm{r}}$ is a **random sample** of x, with size n_r (or a n_r**-sample** of x).
- We say that $\vec{\mathrm{X}}_{n_\mathrm{r}}$ is a set of **independent and identically distributed (i.i.d.)** r.v.
- We also say that $\vec{\mathrm{X}}_{n_\mathrm{r}}$ is made of n_r independent copies of x.

23. Strictly speaking, for all q, x_q is a r.v. from the universe Ω^{n_r} to $\mathbb{X}$ where the probability measure of Ω^{n_r} comes from the probability measure of Ω using $\mathrm{Prob}\left(\Phi_1 \times \ldots \times \Phi_{n_\mathrm{r}}\right) = \prod_{q=1}^{n_\mathrm{r}} \mathrm{Prob}\left(\Phi_q\right)$, and thus, $\mathrm{x}_q(\omega_1, \ldots, \omega_{n_\mathrm{r}}) = \mathrm{x}(\omega_q)$.

T97 # Empirical distribution

From a random sample $\vec{\mathrm{X}}_{n_\mathrm{r}}$ of X, we obtain an **empirical** PDF (of PMF) of X with:

$$\hat{p}_\mathrm{X}(x) = \frac{1}{n_\mathrm{r}} \sum_{q=1}^{n_\mathrm{r}} \delta(x - \mathrm{X}_q) \tag{1.77}$$

T97 Empirical distribution

From a random sample $\vec{\mathrm{X}}_{n_\mathrm{r}}$ of X, we obtain an **empirical** PDF (of PMF) of X with:

$$\hat{p}_\mathrm{X}(x) = \frac{1}{n_\mathrm{r}} \sum_{q=1}^{n_\mathrm{r}} \delta(x - \mathrm{X}_q) \tag{1.77}$$

- For a discrete r.v. (δ is the Kronecker delta), the empirical PMF is and approximation of the exact PMF.

T97

Empirical distribution

From a random sample $\vec{\mathrm{X}}_{n_\mathrm{r}}$ of X, we obtain an **empirical** PDF (of PMF) of X with:

$$\hat{p}_{\mathrm{X}}(x) = \frac{1}{n_\mathrm{r}} \sum_{q=1}^{n_\mathrm{r}} \delta(x - \mathrm{x}_q) \tag{1.77}$$

- For a discrete r.v. (δ is the Kronecker delta), the empirical PMF is and approximation of the exact PMF.
- If a component of X is continuous-valued, the empirical PDF must be integrated with respect to the continuous valued component to obtain for example:
  - the empirical CDF;[24]
  - the empirical quantiles (median...);
  - the empirical moments (mean, variance...).

24. Matlab. To plot the empirical CDF.
```
stairs([min(x);sort(x(:))], (0:length(x))/length(x))
```

T98 ◁ ▷ **Exercise 21.** Let x be a r.v. which takes its value in $\mathbb{X}$, and let g be a function from $\mathbb{X}$ which takes a numerical value. Give the approximation of $\mathrm{E}\left(g(\mathrm{x})\right)$ which is obtained by replacing the exact PDF with the empirical PDF in the formula (1.10).

T99 Sample mean

Let X be a r.v. with a mean and a variance.
Let $\vec{\mathrm{X}}_{n_\mathrm{r}}$ be a sample of X.
The empirical mean is a r.v. which is nothing but the the arithmetic mean:

$$\overline{\mathrm{X}}_{n_\mathrm{r}} = \frac{1}{n_\mathrm{r}} \sum_{q=1}^{n_\mathrm{r}} \mathrm{X}_q$$

T99 Sample mean

Let X be a r.v. with a mean and a variance.
Let $\vec{\mathrm{X}}_{n_\mathrm{r}}$ be a sample of X.
The empirical mean is a r.v. which is nothing but the the arithmetic mean:

$$\overline{\mathrm{X}}_{n_\mathrm{r}} = \frac{1}{n_\mathrm{r}} \sum_{q=1}^{n_\mathrm{r}} \mathrm{X}_q$$

- $\overline{\mathrm{X}}_{n_\mathrm{r}}$ is a r.v. whose mean and variance are:[25]

$$\mathrm{E}\left(\overline{\mathrm{X}}_{n_\mathrm{r}}\right) = \mathrm{E}\left(\mathrm{X}\right) \qquad \mathrm{Var}\left(\overline{\mathrm{X}}_{n_\mathrm{r}}\right) = \frac{1}{n_\mathrm{r}} \mathrm{Var}\left(\mathrm{X}\right)$$

25. The mean formula does not necessitate any assumption. The variance formula just necessitates the uncorrelatedness between the r.v. in the sample.

T99 Sample mean

Let X be a r.v. with a mean and a variance.
Let $\vec{\mathrm{X}}_{n_{\mathrm{r}}}$ be a sample of X.
The empirical mean is a r.v. which is nothing but the the arithmetic mean:

$$\overline{\mathrm{X}}_{n_{\mathrm{r}}} = \frac{1}{n_{\mathrm{r}}} \sum_{q=1}^{n_{\mathrm{r}}} \mathrm{X}_q$$

- $\overline{\mathrm{X}}_{n_{\mathrm{r}}}$ is a r.v. whose mean and variance are:[25]

$$\mathrm{E}\left(\overline{\mathrm{X}}_{n_{\mathrm{r}}}\right) = \mathrm{E}\left(\mathrm{X}\right) \qquad \mathrm{Var}\left(\overline{\mathrm{X}}_{n_{\mathrm{r}}}\right) = \frac{1}{n_{\mathrm{r}}} \mathrm{Var}\left(\mathrm{X}\right)$$

- Thus, the distribution of $\sqrt{n_{\mathrm{r}}}\left(\overline{\mathrm{X}}_{n_{\mathrm{r}}} - \mathrm{E}\left(\mathrm{X}\right)\right)$ is zero-mean with variance $\mathrm{Var}\left(\mathrm{X}\right)$.
  The **central limit theorem** says that this distribution tends to a normal one when $n_{\mathrm{r}} \to +\infty$.

25. The mean formula does not necessitate any assumption. The variance formula just necessitates the uncorrelatedness between the r.v. in the sample.

T100 When n_r tends to $+\infty$, the variance of $\overline{\mathrm{x}}_{n_\mathrm{r}}$ tends to 0, which implies the **weak law of large numbers** (convergence in probability):[P6]

$$\forall y > 0 \quad \lim_{n_\mathrm{r} \to +\infty} \mathrm{Prob}\left(\left\| \overline{\mathrm{x}}_{n_\mathrm{r}} - \mathrm{E}\left(\mathrm{x}\right) \right\| < y \right) = 1$$

But the convergence in probability does not imply that a sample mean path tends to a limit.

T100 When n_{r} tends to $+\infty$, the variance of $\overline{\mathrm{x}}_{n_{\mathrm{r}}}$ tends to 0, which implies the **weak law of large numbers** (convergence in probability):[P6]

$$\forall y > 0 \quad \lim_{n_{\mathrm{r}} \to +\infty} \mathrm{Prob}\left(\left\|\overline{\mathrm{x}}_{n_{\mathrm{r}}} - \mathrm{E}\left(\mathrm{x}\right)\right\| < y\right) = 1$$

But the convergence in probability does not imply that a sample mean path tends to a limit.

- We admit the **strong law of large numbers** (almost sure convergence):

$$\mathrm{Prob}\left(\lim_{n_{\mathrm{r}} \to +\infty} \overline{\mathrm{x}}_{n_{\mathrm{r}}} = \mathrm{E}\left(\mathrm{x}\right)\right) = 1$$

T100 When n_r tends to $+\infty$, the variance of $\overline{\mathrm{X}}_{n_\mathrm{r}}$ tends to 0, which implies the **weak law of large numbers** (convergence in probability):[P6]

$$\forall y > 0 \quad \lim_{n_\mathrm{r}\to+\infty} \mathrm{Prob}\left(\left\|\overline{\mathrm{X}}_{n_\mathrm{r}} - \mathrm{E}\left(\mathrm{X}\right)\right\| < y\right) = 1$$

But the convergence in probability does not imply that a sample mean path tends to a limit.

- We admit the **strong law of large numbers** (almost sure convergence):

$$\mathrm{Prob}\left(\lim_{n_\mathrm{r}\to+\infty} \overline{\mathrm{X}}_{n_\mathrm{r}} = \mathrm{E}\left(\mathrm{X}\right)\right) = 1$$

- This means that for (almost) all sequence $(x_n)_{n\in\mathbb{N}^*}$ which is a realization of $(\mathrm{X}_n)_{n\in\mathbb{N}^*}$:

$$\lim_{n_\mathrm{r}\to+\infty} \overline{x}_{n_\mathrm{r}} = \mathrm{E}\left(\mathrm{X}\right)$$

T101 If the calculation of a given quantity can be interpreted as the calculation of the mean of a r.v. X, a way to calculate this mean is to draw at random a realization $\vec{x}_{n_\mathrm{r}}$ of $\vec{\mathrm{X}}_{n_\mathrm{r}}$ with a high n_r; then, the arithmetic mean $\overline{x}_{n_\mathrm{r}}$ is an approximation of the mean m.

Such a method is called a **Monte-Carlo method**.

T101 If the calculation of a given quantity can be interpreted as the calculation of the mean of a r.v. X, a way to calculate this mean is to draw at random a realization $\vec{x}_{n_\mathrm{r}}$ of $\vec{\mathrm{X}}_{n_\mathrm{r}}$ with a high n_r; then, the arithmetic mean $\overline{x}_{n_\mathrm{r}}$ is an approximation of the mean m.

Such a method is called a **Monte-Carlo method**.

- In practice, we simulate random drawings by means of pseudo-random numbers generators. This is called **Stochastic simulation.**[26]

26. Matlab. To obtain an approximation of π (with a high n_r).

```
4*mean(abs([1 j]*rand(2,nr))<1)
```

T102 From the linear model to the normal distribution

The stochastic link between two r.v. X and Y is the joint distribution (or the marginal X and the conditional $Y \mid X$).
In many cases, a simple model can provide a structure for the distribution of Y given X, in which a modeling error W appears.

T102 From the linear model to the normal distribution

The stochastic link between two r.v. X and Y is the joint distribution (or the marginal X and the conditional $\mathrm{Y} \mid \mathrm{X}$).
In many cases, a simple model can provide a structure for the distribution of Y given X, in which a modeling error W appears.

General model

A general enough formalization of this dependency is to suppose that there exist a function $\bar{h}$ such that:

$$\mathrm{Y} = \bar{h}(\mathrm{X}, \mathrm{W}) \text{ with } \mathrm{W} \perp\!\!\!\perp \mathrm{X} \tag{1.78}$$

T102
From the linear model to the normal distribution

The stochastic link between two r.v. X and Y is the joint distribution (or the marginal X and the conditional $\mathrm{Y} \mid \mathrm{X}$).
In many cases, a simple model can provide a structure for the distribution of Y given X, in which a modeling error W appears.

General model

A general enough formalization of this dependency is to suppose that there exist a function $\bar{h}$ such that:

$$\mathrm{Y} = \bar{h}(\mathrm{X}, \mathrm{W}) \text{ with } \mathrm{W} \perp\!\!\!\perp \mathrm{X} \tag{1.78}$$

- Thus, using the total probability law:

$$p_{\mathrm{Y}|\mathrm{X}}(y) = \int \delta(y - \bar{h}(\mathrm{X}, w))\, p_{\mathrm{W}}(w)\, \mathrm{d}\, w \tag{1.79}$$

T103 Additive error model

The error is often assumed an additive one ($\bar{h}(x, w) = h(x) + w$):

$$\mathrm{Y} = h(\mathrm{X}) + \mathrm{W} \text{ with } \mathrm{W} \perp\!\!\!\perp \mathrm{X} \tag{1.80}$$

T103 Additive error model

The error is often assumed an additive one ($\bar{h}(x, w) = h(x) + w$):

$$\mathrm{Y} = h(\mathrm{X}) + \mathrm{W} \text{ with } \mathrm{W} \perp\!\!\!\perp \mathrm{X} \tag{1.80}$$

- Thus, using the sifting property of the Dirac delta:

$$p_{\mathrm{Y}|\mathrm{X}}(y) = p_{\mathrm{W}}(y - h(\mathrm{X})) \tag{1.81}$$

T104 Linear model

h is assumed to be a linear function, and the independence of x et w is limited to the second order.
There exists H such that the mean and the variance given x of the r.v. $\mathrm{w} = \mathrm{y} - H\,\mathrm{x}$ are independent of x:

$$\mathrm{y} = H\,\mathrm{x} + \mathrm{w} \text{ with } \begin{cases} \mathrm{E}\left(\mathrm{w} \mid \mathrm{x}\right) & = m_{\mathrm{w}} \\ \mathrm{Var}\left(\mathrm{w} \mid \mathrm{x}\right) & = C_{\mathrm{w},\mathrm{w}} \end{cases} \tag{1.82}$$

T104 Linear model

h is assumed to be a linear function, and the independence of X et W is limited to the second order. There exists H such that the mean and the variance given X of the r.v. $\mathrm{W} = \mathrm{Y} - H\,\mathrm{X}$ are independent of X:

$$\mathrm{Y} = H\,\mathrm{X} + \mathrm{W} \text{ with } \begin{cases} \mathrm{E}\left(\mathrm{W} \mid \mathrm{X}\right) & = m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{W} \mid \mathrm{X}\right) & = C_{\mathrm{W},\mathrm{W}} \end{cases} \tag{1.82}$$

- It is equivalent to assume a linear conditional mean $\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right)$ and a uniform conditional variance $\mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right)$:

$$\text{there exist } H,\ m_{\mathrm{W}},\ C_{\mathrm{W},\mathrm{W}} \text{ such that } \begin{cases} \mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right) & = H\,\mathrm{X} + m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right) & = C_{\mathrm{W},\mathrm{W}} \end{cases} \tag{1.83}$$

T104 Linear model

h is assumed to be a linear function, and the independence of X et W is limited to the second order.
There exists H such that the mean and the variance given X of the r.v. $\mathrm{W} = \mathrm{Y} - H\,\mathrm{X}$ are independent of X:

$$\mathrm{Y} = H\,\mathrm{X} + \mathrm{W} \text{ with } \begin{cases} \mathrm{E}\left(\mathrm{W} \mid \mathrm{X}\right) & = m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{W} \mid \mathrm{X}\right) & = C_{\mathrm{W},\mathrm{W}} \end{cases} \tag{1.82}$$

- It is equivalent to assume a linear conditional mean $\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right)$ and a uniform conditional variance $\mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right)$:

$$\text{there exist } H,\ m_{\mathrm{W}},\ C_{\mathrm{W},\mathrm{W}} \text{ such that } \begin{cases} \mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right) & = H\,\mathrm{X} + m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right) & = C_{\mathrm{W},\mathrm{W}} \end{cases} \tag{1.83}$$

- In numerous problems, X is the unknown quantity of interest, for which the observation Y brings some information. The r.v. W usually represents a measurement noise. We refer to a **linear observation model**. H is the **observation matrix**.

T105 If X has a mean m_{X} and a variance $C_{\mathrm{X},\mathrm{X}}$:

- the mean and the variance of Y and the covariance are, in function of H, m_{W} and $C_{\mathrm{W},\mathrm{W}}$:[P7]

$$\begin{cases} m_{\mathrm{Y}} &= H\, m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y},\mathrm{Y}} &= H\, C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} + C_{\mathrm{W},\mathrm{W}} \\ C_{\mathrm{X},\mathrm{Y}} &= \quad C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} \end{cases} \tag{1.84}$$

- We can deduce the conditional mean and variance in function of the joint distribution parameters, if X is not degenerate:[P8]

$$\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right) = m_{\mathrm{Y}} + C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\left(\mathrm{X} - m_{\mathrm{X}}\right) \qquad \mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right) = C_{\mathrm{Y},\mathrm{Y}} - C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\, C_{\mathrm{X},\mathrm{Y}} \tag{1.85}$$

T105 If X has a mean m_{X} and a variance $C_{\mathrm{X},\mathrm{X}}$:

- the mean and the variance of Y and the covariance are, in function of H, m_{W} and $C_{\mathrm{W},\mathrm{W}}$:[P7]

$$\begin{cases} m_{\mathrm{Y}} = H\, m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y},\mathrm{Y}} = H\, C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} + C_{\mathrm{W},\mathrm{W}} \\ C_{\mathrm{X},\mathrm{Y}} = C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} \end{cases} \tag{1.84}$$

- We can deduce the conditional mean and variance in function of the joint distribution parameters, if X is not degenerate:[P8]

$$\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right) = m_{\mathrm{Y}} + C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1} \left(\mathrm{X} - m_{\mathrm{X}}\right) \qquad \mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right) = C_{\mathrm{Y},\mathrm{Y}} - C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\, C_{\mathrm{X},\mathrm{Y}} \tag{1.85}$$

T105 If X has a mean m_{X} and a variance $C_{\mathrm{X},\mathrm{X}}$:

- the mean and the variance of Y and the covariance are, in function of H, m_{W} and $C_{\mathrm{W},\mathrm{W}}$:[P7]

$$\begin{cases} m_{\mathrm{Y}} &= H\, m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y},\mathrm{Y}} &= H\, C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} + C_{\mathrm{W},\mathrm{W}} \\ C_{\mathrm{X},\mathrm{Y}} &= \quad C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} \end{cases} \tag{1.84}$$

- We can deduce the conditional mean and variance in function of the joint distribution parameters, if X is not degenerate:[P8]

$$\mathrm{E}\,(\mathrm{Y} \mid \mathrm{X}) = m_{\mathrm{Y}} + C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\, (\mathrm{X} - m_{\mathrm{X}}) \qquad \mathrm{Var}\,(\mathrm{Y} \mid \mathrm{X}) = C_{\mathrm{Y},\mathrm{Y}} - C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\, C_{\mathrm{X},\mathrm{Y}} \tag{1.85}$$

T105 If X has a mean m_{X} and a variance $C_{\mathrm{X},\mathrm{X}}$:

- the mean and the variance of Y and the covariance are, in function of H, m_{W} and $C_{\mathrm{W},\mathrm{W}}$:[P7]

$$\begin{cases} m_{\mathrm{Y}} &= H\, m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y},\mathrm{Y}} &= H\, C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} + C_{\mathrm{W},\mathrm{W}} \\ C_{\mathrm{X},\mathrm{Y}} &= \quad C_{\mathrm{X},\mathrm{X}}\, H^{\mathsf{T}} \end{cases} \tag{1.84}$$

- We can deduce the conditional mean and variance in function of the joint distribution parameters, if X is not degenerate:[P8]

$$\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}\right) = m_{\mathrm{Y}} + C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1} \left(\mathrm{X} - m_{\mathrm{X}}\right) \qquad \mathrm{Var}\left(\mathrm{Y} \mid \mathrm{X}\right) = C_{\mathrm{Y},\mathrm{Y}} - C_{\mathrm{Y},\mathrm{X}}\, C_{\mathrm{X},\mathrm{X}}^{-1}\, C_{\mathrm{X},\mathrm{Y}} \tag{1.85}$$

⚠ The linear assumption (1.83) (or (1.82)), which refer to the distribution of Y given X, do not imply any property about the distribution of X given Y.
We must distinguish the two following formulations:

- $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model (described above);
- $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model (obtained by exchanging X and Y in the formulas above).

T106 ◁ ▷ **Exercise 22.** Let's consider d_y thermometers. Each of them provides a measure y_i of the actual temperature x. The errors are supposed zero-mean, independent, with the same variance σ^2.

We define $\mathrm{Y} = \begin{bmatrix} \mathrm{Y}_1 \\ \vdots \\ \mathrm{Y}_{d_y} \end{bmatrix}$ $\quad H, m_w, C_{ww}$

Provide a linear model which fulfills these assumptions.

◁ ▷ **Exercise 23** (Training for estimation theory and Kalman filtering).
$\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model described by H, m_W and $C_{\mathrm{W},\mathrm{W}}$.
The mean and the variance of X are m_X and $C_{\mathrm{X},\mathrm{X}}$.

Furthermore, we assume that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
Give $\mathrm{E}(\mathrm{X} \mid \mathrm{Y})$ and $\mathrm{Var}(\mathrm{X} \mid \mathrm{Y})$ in function of H, m_W, $C_{\mathrm{W},\mathrm{W}}$, m_X and $C_{\mathrm{X},\mathrm{X}}$.

$$\begin{bmatrix} Y_1 \\ | \\ Y_{d_y} \end{bmatrix} = \begin{bmatrix} 1 \\ | \\ 1 \end{bmatrix} x + \begin{bmatrix} w_1 \\ | \\ w_{d_y} \end{bmatrix} \qquad \forall i \quad Y_i = x + w_i$$

$$E(w_i) = 0$$
$$Var(w_i) = \sigma^2$$

$$Y = H x + w$$

$$H = \begin{bmatrix} 1 \\ | \\ 1 \end{bmatrix} \qquad m_w = \begin{bmatrix} 0 \\ | \\ 0 \end{bmatrix} \qquad C_{ww} = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

T107 Multivariate normal distribution

A r.v. x which takes its value in $\mathbb{R}^d$ is normally distributed (or Gauss-distributed) if, for all vector $v \in \mathbb{R}^d$, $v^\intercal \mathrm{x}$ is normally distributed.[27]

Then, necessarily, x has a mean m_{x} and a variance $C_{\mathrm{x},\mathrm{x}}$.

A normal distribution is perfectly characterized by its mean and every root Σ_{x} of its variance $C_{\mathrm{x},\mathrm{x}} = \Sigma_{\mathrm{x}} \Sigma_{\mathrm{x}}^\intercal$.

Every linear function of a normal r.v. gives a normal r.v.

All higher order cumulants are null.

27. For all $v \in \mathbb{R}^d$, $v^\intercal \mathrm{x}$ is driven by an univariate normal distribution.

T108 The 3 formulations below are equivalent.

1. The pair (X, Y) is normally distributed. $\begin{bmatrix} x \\ y \end{bmatrix}$ is normally distributed
2. The conditions below hold:
   - X is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[28]
3. There exists a matrix H such that X and $\mathrm{Y} - H\,\mathrm{X}$ are normally distributed, and independent.

28. The means, variances and covariance can be calculated thanks to formulas (1.84) and (1.85). We can deduce that if $g(., C)$ if the PDF of the centered normal distribution with variance C, for all (x, y):

$$g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (H\,x + m_{\mathrm{W}}), C_{\mathrm{W,W}}) = g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ H\,m_{\mathrm{X}} + m_{\mathrm{W}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,X}}\,H^{\mathsf{T}} \\ H\,C_{\mathrm{X,X}} & H\,C_{\mathrm{X,X}}\,H^{\mathsf{T}} + C_{\mathrm{W,W}} \end{bmatrix}\right)$$

Conversely:

$$g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,Y}} \\ C_{\mathrm{Y,X}} & C_{\mathrm{Y,Y}} \end{bmatrix}\right) = g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (m_{\mathrm{Y}} + C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,(x - m_{\mathrm{X}})), C_{\mathrm{Y,Y}} - C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,C_{\mathrm{X,Y}})$$

T108 The 3 formulations below are equivalent.

1. The pair (X, Y) is normally distributed.
2. The conditions below hold:
   - X is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[28]
3. There exists a matrix H such that X and $\mathrm{Y} - H\,\mathrm{X}$ are normally distributed, and independent.

28. The means, variances and covariance can be calculated thanks to formulas (1.84) and (1.85). We can deduce that if $g(., C)$ if the PDF of the centered normal distribution with variance C, for all (x, y):

$$g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (H\,x + m_{\mathrm{W}}), C_{\mathrm{W,W}}) = g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ H\,m_{\mathrm{X}} + m_{\mathrm{W}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,X}}\,H^{\mathsf{T}} \\ H\,C_{\mathrm{X,X}} & H\,C_{\mathrm{X,X}}\,H^{\mathsf{T}} + C_{\mathrm{W,W}} \end{bmatrix}\right)$$

Conversely:

$$g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,Y}} \\ C_{\mathrm{Y,X}} & C_{\mathrm{Y,Y}} \end{bmatrix}\right) = g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (m_{\mathrm{Y}} + C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,(x - m_{\mathrm{X}})), C_{\mathrm{Y,Y}} - C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,C_{\mathrm{X,Y}})$$

T108 The 3 formulations below are equivalent.

1. The pair (X, Y) is normally distributed.
2. The conditions below hold:
    - X is normally distributed;
    - $\mathrm{Y} \mid \mathrm{X}$ is normally distributed;
    - $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[28]
3. There exists a matrix H such that X and $\mathrm{Y} - H\,\mathrm{X}$ are normally distributed, and independent.

28. The means, variances and covariance can be calculated thanks to formulas (1.84) and (1.85). We can deduce that if $g(., C)$ if the PDF of the centered normal distribution with variance C, for all (x, y):

$$g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (H\,x + m_{\mathrm{W}}), C_{\mathrm{W,W}}) = g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ H\,m_{\mathrm{X}} + m_{\mathrm{W}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,X}}\,H^{\mathsf{T}} \\ H\,C_{\mathrm{X,X}} & H\,C_{\mathrm{X,X}}\,H^{\mathsf{T}} + C_{\mathrm{W,W}} \end{bmatrix}\right)$$

Conversely:

$$g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X,X}} & C_{\mathrm{X,Y}} \\ C_{\mathrm{Y,X}} & C_{\mathrm{Y,Y}} \end{bmatrix}\right) = g(x - m_{\mathrm{X}}, C_{\mathrm{X,X}})\, g(y - (m_{\mathrm{Y}} + C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,(x - m_{\mathrm{X}})), C_{\mathrm{Y,Y}} - C_{\mathrm{Y,X}}\,C_{\mathrm{X,X}}^{-1}\,C_{\mathrm{X,Y}})$$

T108 The 3 formulations below are equivalent.

1. The pair (X, Y) is normally distributed.
2. The conditions below hold:
   - X is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is normally distributed;
   - $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[28]
3. There exists a matrix H such that X and $\mathrm{Y} - H\,\mathrm{X}$ are normally distributed, and independent.

➲ In the formulations 2 and 3, we can exchange X and Y.

28. The means, variances and covariance can be calculated thanks to formulas (1.84) and (1.85). We can deduce that if $g(.,C)$ if the PDF of the centered normal distribution with variance C, for all (x,y):

$$g(x - m_{\mathrm{X}}, C_{\mathrm{X},\mathrm{X}})\, g(y - (H\,x + m_{\mathrm{W}}), C_{\mathrm{W},\mathrm{W}}) = g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ H\,m_{\mathrm{X}} + m_{\mathrm{W}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X},\mathrm{X}} & C_{\mathrm{X},\mathrm{X}}\,H^{\mathsf{T}} \\ H\,C_{\mathrm{X},\mathrm{X}} & H\,C_{\mathrm{X},\mathrm{X}}\,H^{\mathsf{T}} + C_{\mathrm{W},\mathrm{W}} \end{bmatrix}\right)$$

Conversely:

$$g\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_{\mathrm{X}} \\ m_{\mathrm{Y}} \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X},\mathrm{X}} & C_{\mathrm{X},\mathrm{Y}} \\ C_{\mathrm{Y},\mathrm{X}} & C_{\mathrm{Y},\mathrm{Y}} \end{bmatrix}\right) = g(x - m_{\mathrm{X}}, C_{\mathrm{X},\mathrm{X}})\, g(y - (m_{\mathrm{Y}} + C_{\mathrm{Y},\mathrm{X}}\,C_{\mathrm{X},\mathrm{X}}^{-1}\,(x - m_{\mathrm{X}})), C_{\mathrm{Y},\mathrm{Y}} - C_{\mathrm{Y},\mathrm{X}}\,C_{\mathrm{X},\mathrm{X}}^{-1}\,C_{\mathrm{X},\mathrm{Y}})$$

T109 If the pair (X, Y) is normally distributed:

- the marginal and conditional distributions are normal;
- uncorrelatedness of X and Y implies their independence (thus, the uncorrelatedness and the independence are equivalent).

If the r.v. X and Y are normally distributed (each of both), and not independent:

- the joint and marginal distributions are not necessarily normal;
- the uncorrelatedness does not imply the independence.

T110 Non degenerate case

If the variance $C_{\mathrm{x,x}}$ is invertible, we can show that the PDF is, for all x: [29]

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{\det(2\pi\, C_{\mathrm{x,x}})}} \exp\left[-\frac{1}{2}(x - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x,x}}^{-1}\, (x - m_{\mathrm{x}})\right] \tag{1.86}$$

29. See the degenerate case in [ref].

T110 Non degenerate case

If the variance $C_{\mathrm{x},\mathrm{x}}$ is invertible, we can show that the PDF is, for all x: [29]

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{\det(2\pi\, C_{\mathrm{x},\mathrm{x}})}} \exp\left[-\frac{1}{2}(x - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x},\mathrm{x}}^{-1}\,(x - m_{\mathrm{x}})\right] \tag{1.86}$$

- If X takes its value in $\mathbb{R}^d$, the r.v. $\kappa^2 = (\mathrm{X} - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x},\mathrm{x}}^{-1}\,(\mathrm{X} - m_{\mathrm{x}})$ is χ^2-distributed with d degrees of freedom (page 51).

  The smallest confidence domain at level $P_0 \in [0, 1]$ is the ellipsoid:

$$\{x \in \mathbb{R}^d \mid (x - m_{\mathrm{x}})^{\intercal}\, C_{\mathrm{x},\mathrm{x}}^{-1}\,(x - m_{\mathrm{x}}) \leqslant F_{\kappa^2}^{-1}(P_0)\}$$

29. See the degenerate case in [ref].

T111 Non degenerate bivariate case

This is a the $d = 2$ case. $\mathrm{K}^2 = (\mathrm{X} - m_\mathrm{X})^\intercal C_{\mathrm{X},\mathrm{X}}^{-1} (\mathrm{X} - m_\mathrm{X})$ is χ^2-distributed with 2 degrees of freedom (2 mean exponential distribution).

T111 Non degenerate bivariate case

This is a the $d = 2$ case. $\mathrm{K}^2 = (\mathrm{X} - m_{\mathrm{X}})^{\intercal} C_{\mathrm{X},\mathrm{X}}^{-1} (\mathrm{X} - m_{\mathrm{X}})$ is χ^2-distributed with 2 degrees of freedom (2 mean exponential distribution).

- The confidence domain at level P_0 is the ellipse:[30] [31]

$$\{x \in \mathbb{R}^2 \mid (x - m_{\mathrm{X}})^{\intercal} C_{\mathrm{X},\mathrm{X}}^{-1} (x - m_{\mathrm{X}}) \leqslant -2 \log(1 - P_0)\}$$

30. The boundary of this domain is written in polar coordinates, by means of every square root Σ_{X} of $C_{\mathrm{X},\mathrm{X}}$:

$$\left\{\sqrt{-2 \log(1 - P_0)}\, \Sigma_{\mathrm{X}} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + m_{\mathrm{X}} \mid \theta \in [0, 2\pi] \right\}$$

31. Matlab. To plot the P_0 confidence ellipse (normal distribution with mean m and variance C).

```
teta = linspace(0,2*pi,100); X = sqrt(-2*log(1-P0))*chol(C,'lower')*[cos(teta); sin(teta)] +
m*ones(1,length(teta)); plot(X(1,:),X(2,:))
```

T112 ◁ ▷ **Exercise 24.** X is a zero mean unit variance r.v., and $\mathrm{Y} = \Sigma\,\mathrm{X} + m$, where m and Σ have correct dimensions.
Give the mean and the variance of Y.

$E(Y) = m$

$Var(Y) = \Sigma\Sigma^T$

T112 ◁ ▷ **Exercise 24.** X is a zero mean unit variance r.v., and $\mathrm{Y} = \Sigma\,\mathrm{X} + m$, where m and Σ have correct dimensions.
Give the mean and the variance of Y.

- From the result of this exercise, we have a way to simulate a normal r.v. with known mean and variance by means of the random generators available in scientific calculation softwares.[32]

32. Matlab. To simulate n_{r} realizations of a Gaussian r.v. in $\mathbb{R}^d$, with mean m and variance C.

```
x = chol(C,'lower')*randn(d,nr) + m*ones(1,nr);
```

T112 ◁ ▷ **Exercise 24.** X is a zero mean unit variance r.v., and $\mathrm{Y} = \Sigma\,\mathrm{X} + m$, where m and Σ have correct dimensions.
Give the mean and the variance of Y.

➲ From the result of this exercise, we have a way to simulate a normal r.v. with known mean and variance by means of the random generators available in scientific calculation softwares.[32]

◁ ▷ **Exercise 25.** With Matlab or Octave, simulate 100 realizations of a bivariate normal r.v. $\left[\begin{smallmatrix}\mathrm{X}\\ \mathrm{Y}\end{smallmatrix}\right]$ with zero mean, correlation coefficient 0.97, the first component with variance 4, the second one with variance 1.
Plot the confidence ellipse at 95%.
Plot the line $x \mapsto \mathrm{E}\left(\mathrm{Y} \mid \mathrm{X} = x\right)$ (formula (1.85)).
Plot the line $y \mapsto \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y} = y\right)$.

32. Matlab. To simulate n_{r} realizations of a Gaussian r.v. in $\mathbb{R}^d$, with mean m and variance C.

```
x = chol(C,'lower')*randn(d,nr) + m*ones(1,nr);
```

T113 Mixture distribution

A continuous r.v. is driven be a mixture distribution if its PDF writes:

$$p_{\mathrm{X}}(x) = \sum_{z \in \{\zeta_1, \dots, \zeta_{n_\mathrm{c}}\}} \lambda_z \, f_z(x) \tag{1.87}$$

with:

- $\{\zeta_1, \dots, \zeta_{n_\mathrm{c}}\}$ a labels set;
- for all z, f_z a positive valued function with unit integral;
- for all z, $\lambda_z \geqslant 0$, and $\sum_z \lambda_z = 1$.

T113 Mixture distribution

A continuous r.v. is driven be a mixture distribution if its PDF writes:

$$p_{\mathrm{X}}(x) = \sum_{z \in \{\zeta_1, \ldots, \zeta_{n_c}\}} \lambda_z \, f_z(x) \tag{1.87}$$

with:

- $\{\zeta_1, \ldots, \zeta_{n_c}\}$ a labels set;
- for all z, f_z a positive valued function with unit integral;
- for all z, $\lambda_z \geqslant 0$, and $\sum_z \lambda_z = 1$.

➔ A mixture distribution can be used when the r.v. X show phenomenons from different clusters:

- f_z is the PDF of the r.v. in the cluster labelled z;
- λ_z is the probability of the cluster labelled z.

For example, if X represent the mark for an exam in which there were good and bad students, but no average one: $\{\zeta_1, \zeta_2\} = \{$"good student", 'bad student"$\}$.

T113
Mixture distribution

A continuous r.v. is driven be a mixture distribution if its PDF writes:

$$p_{\mathrm{X}}(x) = \sum_{z \in \{\zeta_1, \ldots, \zeta_{n_c}\}} \lambda_z \, f_z(x) \tag{1.87}$$

with:

- $\{\zeta_1, \ldots, \zeta_{n_c}\}$ a labels set;
- for all z, f_z a positive valued function with unit integral;
- for all z, $\lambda_z \geqslant 0$, and $\sum_z \lambda_z = 1$.

➔ A mixture distribution can be used when the r.v. X show phenomenons from different clusters:

- f_z is the PDF of the r.v. in the cluster labelled z;
- λ_z is the probability of the cluster labelled z.

For example, if X represent the mark for an exam in which there were good and bad students, but no average one: $\{\zeta_1, \zeta_2\} = \{$"good student", 'bad student"$\}$.

➔ A mixture distribution can be used in problems not related to clustering ones; thus the cluster label is meaningless, we will use $\{\zeta_1, \ldots, \zeta_{n_c}\} = \{1, \ldots, n_c\}$.

T114 Let's introduce the discrete r.v. z which takes its value in $\{\zeta_1, \ldots, \zeta_{n_\mathrm{c}}\}$.
By means of the total probability law:

$$p_\mathrm{x}(x) = \sum_{z \in \{\zeta_1, \ldots, \zeta_{n_\mathrm{c}}\}} \underbrace{\mathrm{Prob}\,(\mathrm{z} = z)}_{\lambda_z} \underbrace{p_{\mathrm{x}|\mathrm{z}}(x, z)}_{f_z(x)} \tag{1.88}$$

A mixture distribution can be seen as the marginal distribution of the continuous-valued r.v. in a pair composed of a discrete-valued r.v. and a continuous-valued one; the joint distribution PDF is $p_{\mathrm{x},\mathrm{z}}(x, z) = \lambda_z\, f_z(x)$.

T115 On the figure 1.4 the PDF of a Gaussian mixture is represented with full line together with $p_{\mathrm{x,z}}(x,\zeta_1)$ and $p_{\mathrm{x,z}}(x,\zeta_2)$ in dotted line.

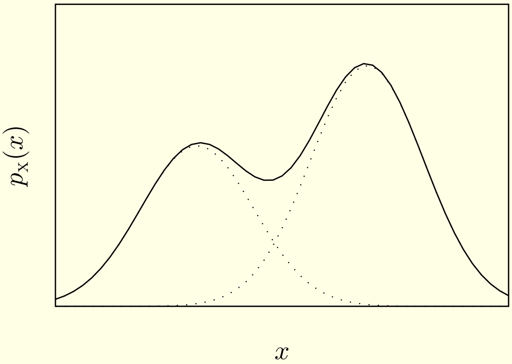


Figure 1.4: PDF of a Gaussian mixture

T116 Uncertainty propagation

Let x be a r.v. with known probability distribution.
Let $h : x \mapsto h(x)$ be a function defined on the support of x. A few ideas to determine the probability distribution of $h(\mathrm{x})$ are available in the appendix.

In this section, x has known mean and variance, and h is a non-linear function.
We want to approximate the mean and the variance of $h(\mathrm{x})$, and the covariance between $h(\mathrm{x})$ and x.

We want to define an **uncertainty propagation**, that is a transform UP such that:

T116 Uncertainty propagation

Let x be a r.v. with known probability distribution.
Let $h : x \mapsto h(x)$ be a function defined on the support of x. A few ideas to determine the probability distribution of $h(\mathrm{x})$ are available in the appendix.

In this section, x has known mean and variance, and h is a non-linear function.
We want to approximate the mean and the variance of $h(\mathrm{x})$, and the covariance between $h(\mathrm{x})$ and x.

We want to define an **uncertainty propagation**, that is a transform UP such that:

T116 Uncertainty propagation

Let X be a r.v. with known probability distribution.
Let $h : x \mapsto h(x)$ be a function defined on the support of X. A few ideas to determine the probability distribution of $h(\mathrm{X})$ are available in the appendix.

In this section, X has known mean and variance, and h is a non-linear function.
We want to approximate the mean and the variance of $h(\mathrm{X})$, and the covariance between $h(\mathrm{X})$ and X.

We want to define an **uncertainty propagation**, that is a transform UP such that:

$$\text{if } \mathrm{Y} = h(\mathrm{X}) \text{ then } \left(m_{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}\right) \simeq \mathrm{UP}\left(h, m_{\mathrm{X}}, C_{\mathrm{X},\mathrm{X}}\right) \tag{1.89}$$

T117 If h is differentiable, with Jacobian matrix $\frac{\partial h}{\partial x^{\mathsf{T}}}$, a natural solution consists in a linearization around the mean value of X:

$$h(\mathrm{X}) \simeq h(m_{\mathrm{X}}) + \frac{\partial h}{\partial x^{\mathsf{T}}}(m_{\mathrm{X}})\,(\mathrm{X} - m_{\mathrm{X}})$$

We obtain, if $\mathrm{Y} = h(\mathrm{X})$:

$$\begin{aligned} m_{\mathrm{Y}} &\simeq h(m_{\mathrm{X}}) \\ C_{\mathrm{Y},\mathrm{Y}} &\simeq \left[\frac{\partial h}{\partial x^{\mathsf{T}}}(m_{\mathrm{X}})\right] C_{\mathrm{X},\mathrm{X}} \left[\frac{\partial h}{\partial x^{\mathsf{T}}}(m_{\mathrm{X}})\right]^{\mathsf{T}} \\ C_{\mathrm{X},\mathrm{Y}} &\simeq C_{\mathrm{X},\mathrm{X}} \left[\frac{\partial h}{\partial x^{\mathsf{T}}}(m_{\mathrm{X}})\right]^{\mathsf{T}} \end{aligned} \tag{1.90}$$

T118 Another solution, the **unscented transform** (UT), consists in the determination of a set of n_σ vectors ξ_q (the so-called σ-points) and weights ω_q, $1 \leqslant q \leqslant n_\sigma$ such that:

$$\sum_{q=1}^{n_\sigma} \omega_q = 1 \qquad m_{\mathrm{x}} = \sum_{q=1}^{n_\sigma} \omega_q \, \xi_q \qquad C_{\mathrm{x,x}} = \sum_{q=1}^{n_\sigma} \omega_q \left(\xi_q - m_{\mathrm{x}}\right)\left(\xi_q - m_{\mathrm{x}}\right)^{\mathsf{T}} \tag{1.91}$$

T118 Another solution, the **unscented transform** (UT), consists in the determination of a set of n_σ vectors ξ_q (the so-called σ-points) and weights ω_q, $1 \leqslant q \leqslant n_\sigma$ such that:

$$\sum_{q=1}^{n_\sigma} \omega_q = 1 \qquad m_{\mathrm{X}} = \sum_{q=1}^{n_\sigma} \omega_q \, \xi_q \qquad C_{\mathrm{X},\mathrm{X}} = \sum_{q=1}^{n_\sigma} \omega_q \left(\xi_q - m_{\mathrm{X}}\right)\left(\xi_q - m_{\mathrm{X}}\right)^{\mathsf{T}} \tag{1.91}$$

Thus, if $\mathrm{Y} = h(\mathrm{X})$, $\left(m_{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}\right) \simeq \mathrm{UT}\left(h, m_{\mathrm{X}}, C_{\mathrm{X},\mathrm{X}}\right)$, that is:

$$\begin{aligned} m_{\mathrm{Y}} &\simeq \sum_{q=1}^{n_\sigma} \omega_q \, h(\xi_q) \text{ (noted } \bar{y} \text{ below)} \\ C_{\mathrm{Y},\mathrm{Y}} &\simeq \sum_{q=1}^{n_\sigma} \omega_q \left(h(\xi_q) - \bar{y}\right)\left(h(\xi_q) - \bar{y}\right)^{\mathsf{T}} \\ C_{\mathrm{X},\mathrm{Y}} &\simeq \sum_{q=1}^{n_\sigma} \omega_q \left(\xi_q - m_{\mathrm{X}}\right)\left(h(\xi_q) - \bar{y}\right)^{\mathsf{T}} \end{aligned} \tag{1.92}$$

We can easily check that, if h is linear, there is no approximation.

T119 How to obtain the σ-points? Let Σ_{x} be a square root of $C_{\mathrm{x,x}}$ ($C_{\mathrm{x,x}} = \Sigma_{\mathrm{x}}\,\Sigma_{\mathrm{x}}^{\mathsf{T}}$), and $\Sigma_{\mathrm{x}}^{(q)}$ be the qth column of Σ_{x}, d_x be the size of X, $\lambda > -d_x$ be a tuning scale parameter; we obtain $n_\sigma = 2\,d_x + 1$ σ-points, with:

$$\begin{aligned} \omega_q &= \tfrac{1}{2\,(d_x+\lambda)} & \xi_q &= m_{\mathrm{x}} + \sqrt{d_x+\lambda}\ \Sigma_{\mathrm{x}}^{(q)} & 1 \leqslant q \leqslant d_x \\ \omega_{q+d_x} &= \tfrac{1}{2\,(d_x+\lambda)} & \xi_{q+d_x} &= m_{\mathrm{x}} - \sqrt{d_x+\lambda}\ \Sigma_{\mathrm{x}}^{(q)} & 1 \leqslant q \leqslant d_x \\ \omega_{2\,d_x+1} &= \tfrac{\lambda}{d_x+\lambda} & \xi_{2\,d_x+1} &= m_{\mathrm{x}} \end{aligned} \tag{1.93}$$

The σ-points (but the central one) are distributed over the ellipsoid $\{x \mid (x-m_{\mathrm{x}})^{\mathsf{T}}\, C_{\mathrm{x,x}}^{-1}\,(x-m_{\mathrm{x}}) = d_x+\lambda\}$.

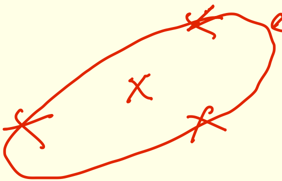
T119 How to obtain the σ-points? Let Σ_{X} be a square root of $C_{\mathrm{X,X}}$ ($C_{\mathrm{X,X}} = \Sigma_{\mathrm{X}}\,\Sigma_{\mathrm{X}}^{\mathsf{T}}$), and $\Sigma_{\mathrm{X}}^{(q)}$ be the qth column of Σ_{X}, d_x be the size of X, $\lambda > -d_x$ be a tuning scale parameter; we obtain $n_\sigma = 2\,d_x + 1$ σ-points, with:

$$
\begin{aligned}
\omega_q &= \frac{1}{2\,(d_x+\lambda)} & \xi_q &= m_{\mathrm{X}} + \sqrt{d_x+\lambda}\ \Sigma_{\mathrm{X}}^{(q)} & 1 \leqslant q \leqslant d_x \\
\omega_{q+d_x} &= \frac{1}{2\,(d_x+\lambda)} & \xi_{q+d_x} &= m_{\mathrm{X}} - \sqrt{d_x+\lambda}\ \Sigma_{\mathrm{X}}^{(q)} & 1 \leqslant q \leqslant d_x \\
\omega_{2\,d_x+1} &= \frac{\lambda}{d_x+\lambda} & \xi_{2\,d_x+1} &= m_{\mathrm{X}}
\end{aligned}
\tag{1.93}
$$

The σ-points (but the central one) are distributed over the ellipsoid $\{x \mid (x-m_{\mathrm{X}})^{\mathsf{T}}\, C_{\mathrm{X,X}}^{-1}\,(x-m_{\mathrm{X}}) = d_x+\lambda\}$.

- There is no convincing policy about the choice of the λ parameter [ref]. The simplest solution is $\lambda = 0$, so that the central σ-point is not used; this is a cubature [ref]. If $\lambda < 0$, the weight of the central σ-point is negative, so that we can obtain a non positive definite variance $C_{\mathrm{Y,Y}}$; a solution is to cancel the effect of the central σ-point in this variance calculation [ref].

$(x-m_x)^T C_{xx}^{-1}(x-m_x) = \text{cte}$



T120 Chapter 2

Parametric estimation

Chapter contents

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x}: \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} :\ & \mathbb{Y} \longrightarrow \mathbb{X} \\ & y \longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$.
Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

An **estimator** is a function $\widehat{x}$ which, for all observation, returns an estimate:

$$\begin{aligned} \widehat{x} : \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \widehat{x}(y) \end{aligned} \tag{2.1}$$

This definition is rather fuzzy, since an estimator is nothing but a statistic[1] which takes its value in $\mathbb{X}$. Intuitively, an estimator should give an estimate "close" to the actual value.

1. That is a function of the observation

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

The objective is now:

- to represent this stochastic link between the parameter X and the observation Y,
- to define an estimator $\widehat{x}(\mathrm{Y})$,
- to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

- The objective is now:
  - to represent this stochastic link between the parameter X and the observation Y,
  - to define an estimator $\widehat{x}(\mathrm{Y})$,
  - to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

The objective is now:

- to represent this stochastic link between the parameter X and the observation Y,
- to define an estimator $\widehat{x}(\mathrm{Y})$,
- to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

➔ The objective is now:

- to represent this stochastic link between the parameter X and the observation Y,
- to define an estimator $\widehat{x}(\mathrm{Y})$,
- to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

➔ The objective is now:

- to represent this stochastic link between the parameter X and the observation Y,
- to define an estimator $\widehat{x}(\mathrm{Y})$,
- to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y.

The process which produced this observation depends of a parameter X, random or not, which, in the experiment, took the unknown value x^*.

Thus, we have to make explicit the expression of the r.v. $\widehat{x}(\mathrm{Y})$.

➔ The objective is now:

- to represent this stochastic link between the parameter X and the observation Y,
- to define an estimator $\widehat{x}(\mathrm{Y})$,
- to measure its performances (we have to make explicit the notion of "closeness" mentioned above).

T123 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

➲ In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \qquad (2.2)$$

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123

Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

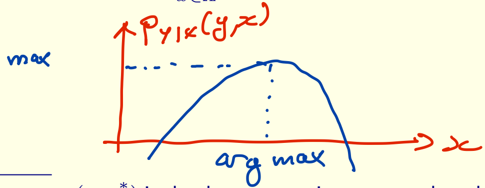
- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

➔ In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

y is Fixed



2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T123

Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $\mathrm{Y} \mid \mathrm{X}$, obtained by physical insight, or by intuition.[2]

- With fixed x, the function $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is the observation PDF, assuming that X is x.
- With fixed y, the function $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.
The likelihood measures the adequacy of the parameter to the data.

- In the classical estimation theory,[3] it is the unique formalization.
It permits to define the **Maximum Likelihood estimator** (**MLE**) estimator:

$$\widehat{x}_{\mathrm{MLE}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}, x) \tag{2.2}$$

- Other estimators which rely on a likelihood structure are possible, for exemple if $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.

2. The conditional distribution with PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.
3. R.A. Fisher, 1922.

T124 ◁ ▷ **Exercise 26.** The observation Y is uniformly distributed between 0 and X. We want to estimate X.

a) Plot the PDF $y \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ for a fixed parameter x,
the likelihood $x \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$, for an observation y,
and the function $(x, y) \mapsto p_{\mathrm{Y}|\mathrm{X}}(y, x)$ (perspective drawing).

b) Give $\widehat{x}_{\mathrm{MLE}}(\mathrm{Y})$. $=$ Y

T125

Prior distribution, predictors

In the Bayesian estimation theory (from Thomas Bayes's name, 1702-1761, but this theory developed in the fifties), the complementary assumption is that, if the experiment is repeated, the parameter itself is modified, and is driven by a probability distribution, the **prior distribution**.
This distribution characterizes the prior information about the parameter, before any observation.

- A **Bayesian estimator** makes a compromise between:
  - the information brought by the data (by means of the likelihood),
  - and this prior information.

The doctor diagnosis uses the biological examinations, but also the family medical history.

- In the extreme, the medical history could lead the doctor not to ask for further examination.

Thus, it is natural to propose an evaluation of the parameter with no data, that is a **predictor**, which returns a prediction only from the prior distribution.

T125 Prior distribution, predictors

In the Bayesian estimation theory (from Thomas Bayes's name, 1702-1761, but this theory developed in the fifties), the complementary assumption is that, if the experiment is repeated, the parameter itself is modified, and is driven by a probability distribution, the **prior distribution**.
This distribution characterizes the prior information about the parameter, before any observation.

- A **Bayesian estimator** makes a compromise between:
  - the information brought by the data (by means of the likelihood),
  - and this prior information.

The doctor diagnosis uses the biological examinations, but also the family medical history.

- In the extreme, the medical history could lead the doctor not to ask for further examination.

Thus, it is natural to propose an evaluation of the parameter with no data, that is a **predictor**, which returns a prediction only from the prior distribution.

T125 # Prior distribution, predictors

In the Bayesian estimation theory (from Thomas Bayes's name, 1702-1761, but this theory developed in the fifties), the complementary assumption is that, if the experiment is repeated, the parameter itself is modified, and is driven by a probability distribution, the **prior distribution**.
This distribution characterizes the prior information about the parameter, before any observation.

- A **Bayesian estimator** makes a compromise between:
  - the information brought by the data (by means of the likelihood),
  - and this prior information.

The doctor diagnosis uses the biological examinations, but also the family medical history.

- In the extreme, the medical history could lead the doctor not to ask for further examination.

Thus, it is natural to propose an evaluation of the parameter with no data, that is a **predictor**, which returns a prediction only from the prior distribution.

T126 The **MMSE** predictor returns[4] the mean value:

$$\check{x}_{\text{MMSE}} = \mathrm{E}\,(\mathrm{x}) \qquad \text{prior mean} \tag{2.3}$$

MMSE stands for **"Minimum Mean Square Error"**; the significance will appear further.

- The maximum *a priori* predictor (**MAP**) returns the most probable value:

$$\check{x}_{\text{MAP}} = \arg\max_{x \in \mathbb{X}} p_{\mathrm{x}}(x) \qquad \text{maximum } a\ priori \tag{2.4}$$

- But there are other policies. For example, if the parameter can be split into two components, $\mathrm{x} = (\mathrm{x}_{\text{C}}, \mathrm{x}_{\text{D}})$:[5]

$$\begin{aligned} \check{x}_{\text{D}} &= \arg\max_{x_{\text{d}}} p_{\mathrm{x}_{\text{D}}}(x_{\text{d}}) && \text{marginal maximum } a\ priori \\ \check{x}_{\text{C}} &= \mathrm{E}\,(\mathrm{x}_{\text{C}} \mid \mathrm{x}_{\text{D}} = \check{x}_{\text{D}}) && \text{conditional prior mean} \end{aligned} \tag{2.5}$$

4. For numerical parameters.
5. Let's consider an urn which contains balls, some of them in balsa and the others in lead. The lead balls have all the same weight. We get ready to draw a ball from the urn. What can we expect on the category, which is a discrete r.v. x_{d} (balsa or lead), and on the weight, which is a continuous variable x_{c}, with the knowledge of the joint distribution of the pair $(\mathrm{x}_{\text{c}}, \mathrm{x}_{\text{d}})$? A "rational" predictor consists in choosing the maximum *a priori* for the category (that is the most probable category), then the prior mean given this category for the weight (that is the mean weight in the most probable category).

T126 The **MMSE** predictor returns[4] the mean value:

$$\check{x}_{\mathrm{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \text{prior mean} \tag{2.3}$$

MMSE stands for **"Minimum Mean Square Error"**; the significance will appear further.

- The maximum *a priori* predictor (**MAP**) returns the most probable value:

$$\check{x}_{\mathrm{MAP}} = \arg\max_{x \in \mathbb{X}} p_{\mathrm{x}}(x) \qquad \text{maximum } \textit{a priori} \tag{2.4}$$

- But there are other policies. For example, if the parameter can be split into two components, $\mathrm{x} = (\mathrm{x}_{\mathrm{C}}, \mathrm{x}_{\mathrm{D}})$:[5]

$$\begin{aligned} \check{x}_{\mathrm{D}} &= \arg\max_{x_{\mathrm{d}}} p_{\mathrm{x}_{\mathrm{D}}}(x_{\mathrm{d}}) && \text{marginal maximum } \textit{a priori} \\ \check{x}_{\mathrm{C}} &= \mathrm{E}\left(\mathrm{x}_{\mathrm{C}} \mid \mathrm{x}_{\mathrm{D}} = \check{x}_{\mathrm{D}}\right) && \text{conditional prior mean} \end{aligned} \tag{2.5}$$

4. For numerical parameters.
5. Let's consider an urn which contains balls, some of them in balsa and the others in lead. The lead balls have all the same weight. We get ready to draw a ball from the urn. What can we expect on the category, which is a discrete r.v. x_{d} (balsa or lead), and on the weight, which is a continuous variable x_{c}, with the knowledge of the joint distribution of the pair $(\mathrm{x}_{\mathrm{c}}, \mathrm{x}_{\mathrm{d}})$? A "rational" predictor consists in choosing the maximum *a priori* for the category (that is the most probable category), then the prior mean given this category for the weight (that is the mean weight in the most probable category).

T126 The **MMSE** predictor returns[4] the mean value:

$$\check{x}_{\text{MMSE}} = \mathrm{E}\,(\mathrm{x}) \qquad \text{prior mean} \tag{2.3}$$

MMSE stands for **“Minimum Mean Square Error”**; the significance will appear further.

- The maximum *a priori* predictor (**MAP**) returns the most probable value:

$$\check{x}_{\text{MAP}} = \arg\max_{x\in\mathbb{X}} p_{\mathrm{x}}(x) \qquad \text{maximum } \textit{a priori} \tag{2.4}$$

- But there are other policies. For example, if the parameter can be split into two components, $\mathrm{x} = (\mathrm{x}_{\text{C}}, \mathrm{x}_{\text{D}})$:[5]

$$\begin{aligned} \check{x}_{\text{D}} &= \arg\max_{x_{\text{d}}} p_{\mathrm{x}_{\text{D}}}(x_{\text{d}}) && \text{marginal maximum } \textit{a priori} \\ \check{x}_{\text{C}} &= \mathrm{E}\,(\mathrm{x}_{\text{C}} \mid \mathrm{x}_{\text{D}} = \check{x}_{\text{D}}) && \text{conditional prior mean} \end{aligned} \tag{2.5}$$

4. For numerical parameters.
5. Let’s consider an urn which contains balls, some of them in balsa and the others in lead. The lead balls have all the same weight. We get ready to draw a ball from the urn. What can we expect on the category, which is a discrete r.v. x_{d} (balsa or lead), and on the weight, which is a continuous variable x_{c}, with the knowledge of the joint distribution of the pair $(\mathrm{x}_{\text{c}}, \mathrm{x}_{\text{d}})$? A “rational” predictor consists in choosing the maximum *a priori* for the category (that is the most probable category), then the prior mean given this category for the weight (that is the mean weight in the most probable category).

T127 The estimator (or predictor) choice can be pragmatic (in some situations, it can be easier to compute a maximum rather than an expectation).

It can depend of the interpretation we want to give to the result:

- the birth rate is 1.9 chidren per woman (that is the mean value);
- if we encounter an unknown woman, we can expect that she has 2 children (that is the most probable value).

T127 The estimator (or predictor) choice can be pragmatic (in some situations, it can be easier to compute a maximum rather than an expectation).

➜ It can depend of the interpretation we want to give to the result:

- the birth rate is 1.9 chidren per woman (that is the mean value);
- if we encounter an unknown woman, we can expect that she has 2 children (that is the most probable value).

T128 Posterior distribution, Bayesian estimators

We now observe the r.v. Y.
The **posterior distribution** is the distribution of the parameter X given the observation Y.

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

- The **MMSE** estimator is defined by:

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \text{posterior mean} \tag{2.6}$$

- The **Maximum *a Posteriori*** estimator (**MAP**) is defined by:

$$\widehat{x}_{\mathrm{MAP}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x) \qquad \text{maximum } \textit{a posteriori} \tag{2.7}$$

- If the r.v. X can be split as $\mathrm{X} = \left[\begin{smallmatrix} \mathrm{X}_{\mathrm{C}} \\ \mathrm{X}_{\mathrm{D}} \end{smallmatrix}\right]$, we can use the estimator below:

$$\begin{aligned} \widehat{x}_{\mathrm{D}}(\mathrm{Y}) &= \arg\max_{x_{\mathrm{d}}} p_{\mathrm{X}_{\mathrm{D}}|\mathrm{Y}}(x_{\mathrm{d}}) && \text{marginal maximum } \textit{a posteriori} \\ \widehat{x}_{\mathrm{C}}(\mathrm{Y}) &= \mathrm{E}\left(\mathrm{X}_{\mathrm{C}} \mid \mathrm{X}_{\mathrm{D}} = \widehat{x}_{\mathrm{D}}(\mathrm{Y}), \mathrm{Y}\right) && \text{conditional posterior mean} \end{aligned} \tag{2.8}$$

T128 Posterior distribution, Bayesian estimators

We now observe the r.v. Y.
The **posterior distribution** is the distribution of the parameter X given the observation Y.

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

- The **MMSE** estimator is defined by:

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \text{posterior mean} \tag{2.6}$$

- The **Maximum *a Posteriori*** estimator (**MAP**) is defined by:

$$\widehat{x}_{\mathrm{MAP}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x) \qquad \text{maximum } \textit{a posteriori} \tag{2.7}$$

- If the r.v. X can be split as $\mathrm{X} = \left[\begin{smallmatrix} \mathrm{X}_{\mathrm{C}} \\ \mathrm{X}_{\mathrm{D}} \end{smallmatrix}\right]$, we can use the estimator below:

$$\begin{aligned} \widehat{x}_{\mathrm{D}}(\mathrm{Y}) &= \arg\max_{x_{\mathrm{d}}} p_{\mathrm{X}_{\mathrm{D}}|\mathrm{Y}}(x_{\mathrm{d}}) && \text{marginal maximum } \textit{a posteriori} \\ \widehat{x}_{\mathrm{C}}(\mathrm{Y}) &= \mathrm{E}\left(\mathrm{X}_{\mathrm{C}} \mid \mathrm{X}_{\mathrm{D}} = \widehat{x}_{\mathrm{D}}(\mathrm{Y}), \mathrm{Y}\right) && \text{conditional posterior mean} \end{aligned} \tag{2.8}$$

T128 # Posterior distribution, Bayesian estimators

We now observe the r.v. Y.
The **posterior distribution** is the distribution of the parameter X given the observation Y.

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

- The **MMSE** estimator is defined by:

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \text{posterior mean} \tag{2.6}$$

- The **Maximum *a Posteriori*** estimator (**MAP**) is defined by:

$$\widehat{x}_{\mathrm{MAP}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x) \qquad \text{maximum } \textit{a posteriori} \tag{2.7}$$

- If the r.v. X can be split as $\mathrm{X} = \left[\begin{smallmatrix} \mathrm{X}_{\mathrm{C}} \\ \mathrm{X}_{\mathrm{D}} \end{smallmatrix}\right]$, we can use the estimator below:

$$\begin{aligned} \widehat{x}_{\mathrm{D}}(\mathrm{Y}) &= \arg\max_{x_{\mathrm{d}}} p_{\mathrm{X}_{\mathrm{D}}|\mathrm{Y}}(x_{\mathrm{d}}) && \text{marginal maximum } \textit{a posteriori} \\ \widehat{x}_{\mathrm{C}}(\mathrm{Y}) &= \mathrm{E}\left(\mathrm{X}_{\mathrm{C}} \mid \mathrm{X}_{\mathrm{D}} = \widehat{x}_{\mathrm{D}}(\mathrm{Y}), \mathrm{Y}\right) && \text{conditional posterior mean} \end{aligned} \tag{2.8}$$

T128 Posterior distribution, Bayesian estimators

We now observe the r.v. Y.
The **posterior distribution** is the distribution of the parameter X given the observation Y.

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

- The **MMSE** estimator is defined by:

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \text{posterior mean} \tag{2.6}$$

- The **Maximum *a Posteriori*** estimator (**MAP**) is defined by:

$$\widehat{x}_{\mathrm{MAP}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x) \qquad \text{maximum } \textit{a posteriori} \tag{2.7}$$

- If the r.v. X can be split as $\mathrm{X} = \begin{bmatrix} \mathrm{X_C} \\ \mathrm{X_D} \end{bmatrix}$, we can use the estimator below:

$$\begin{aligned} \widehat{x}_{\mathrm{D}}(\mathrm{Y}) &= \arg\max_{x_\mathrm{d}} p_{\mathrm{X_D}|\mathrm{Y}}(x_\mathrm{d}) && \text{marginal maximum } \textit{a posteriori} \\ \widehat{x}_{\mathrm{C}}(\mathrm{Y}) &= \mathrm{E}\left(\mathrm{X_C} \mid \mathrm{X_D} = \widehat{x}_{\mathrm{D}}(\mathrm{Y}), \mathrm{Y}\right) && \text{conditional posterior mean} \end{aligned} \tag{2.8}$$

T128
Posterior distribution, Bayesian estimators

We now observe the r.v. Y.
The **posterior distribution** is the distribution of the parameter X given the observation Y.

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

- The **MMSE** estimator is defined by:

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \text{posterior mean} \tag{2.6}$$

- The **Maximum *a Posteriori*** estimator (**MAP**) is defined by:

$$\widehat{x}_{\mathrm{MAP}}(\mathrm{Y}) = \arg\max_{x \in \mathbb{X}} p_{\mathrm{X}|\mathrm{Y}}(x) \qquad \text{maximum } \textit{a posteriori} \tag{2.7}$$

- If the r.v. X can be split as $\mathrm{X} = \left[\begin{smallmatrix} \mathrm{X}_{\mathrm{C}} \\ \mathrm{X}_{\mathrm{D}} \end{smallmatrix}\right]$, we can use the estimator below:

$$\begin{aligned} \widehat{x}_{\mathrm{D}}(\mathrm{Y}) &= \arg\max_{x_{\mathrm{d}}} p_{\mathrm{X}_{\mathrm{D}}|\mathrm{Y}}(x_{\mathrm{d}}) && \text{marginal maximum } \textit{a posteriori} \\ \widehat{x}_{\mathrm{C}}(\mathrm{Y}) &= \mathrm{E}\left(\mathrm{X}_{\mathrm{C}} \mid \mathrm{X}_{\mathrm{D}} = \widehat{x}_{\mathrm{D}}(\mathrm{Y}), \mathrm{Y}\right) && \text{conditional posterior mean} \end{aligned} \tag{2.8}$$

T129 The **LMMSE** ("Linear MMSE") estimator relies on the second order characteristics of the joint distribution of (X, Y):

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - m_{\mathrm{Y}}) \tag{2.9}$$

It will be rewritten page 145 when $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[6]

6. $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model if $\mathrm{Y} = H\,\mathrm{X} + \mathrm{W}$ with $\begin{cases} \mathrm{E}\,(\mathrm{w} \mid \mathrm{x}) & = m_{\mathrm{w}} \\ \mathrm{Var}\,(\mathrm{w} \mid \mathrm{x}) & = C_{\mathrm{w},\mathrm{w}} \end{cases}$

T129 The **LMMSE** ("Linear MMSE") estimator relies on the second order characteristics of the joint distribution of (X, Y):

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, (\mathrm{Y} - m_{\mathrm{Y}}) \tag{2.9}$$

It will be rewritten page 145 when $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[6]

6. $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model if $\mathrm{Y} = H\,\mathrm{X} + \mathrm{W}$ with $\begin{cases} \mathrm{E}\,(\mathrm{w} \mid \mathrm{x}) & = m_{\mathrm{w}} \\ \mathrm{Var}\,(\mathrm{w} \mid \mathrm{x}) & = C_{\mathrm{w},\mathrm{w}} \end{cases}$

T129 The **LMMSE** ("Linear MMSE") estimator relies on the second order characteristics of the joint distribution of (X, Y):

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - m_{\mathrm{Y}}) \tag{2.9}$$

It will be rewritten page 145 when $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model.[6]

◁ ▷ **Exercise 27.** In this exercise, the problem is inverted, we look for an estimator $\widehat{y}(\mathrm{X})$, for the simplified model $\mathrm{Y} = H\,\mathrm{X} + \mathrm{W}$ with $\mathrm{E}\,(\mathrm{W} \mid \mathrm{X}) = m_{\mathrm{W}}$ (no condition on $\mathrm{Var}\,(\mathrm{W} \mid \mathrm{X})$).

a) Write $\widehat{y}_{\mathrm{MMSE}}(\mathrm{X})$ and $\widehat{y}_{\mathrm{LMMSE}}(\mathrm{X})$.

b) Prove that $\widehat{y}_{\mathrm{MMSE}}(\mathrm{X}) = \widehat{y}_{\mathrm{LMMSE}}(\mathrm{X})$ if and only if this simplified model is fulfilled.

6. $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model if $\mathrm{Y} = H\,\mathrm{X} + \mathrm{W}$ with $\begin{cases} \mathrm{E}\,(\mathrm{W} \mid \mathrm{X}) & = m_{\mathrm{W}} \\ \mathrm{Var}\,(\mathrm{W} \mid \mathrm{X}) & = C_{\mathrm{W},\mathrm{W}} \end{cases}$

T130

How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{\mathrm{Y}|\mathrm{X}}$, then the distribution of $\mathrm{Y} \mid \mathrm{X}$,
- and the prior distribution, that is the distribution of X.

It is equivalent to say that we have the distribution of the pair (X, Y).

➲ By means of the Bayes law, we obtain the posterior distribution; for all (x, y):

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x)}{p_{\mathrm{Y}}(y)} \tag{2.10}$$

Since the denominator does not depend of x, we often write:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x) \tag{2.11}$$

This formula is the basis one in Bayesian inference.

T130 How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{\mathrm{Y}|\mathrm{X}}$, then the distribution of $\mathrm{Y} \mid \mathrm{X}$,
- and the prior distribution, that is the distribution of X.

It is equivalent to say that we have the distribution of the pair (X, Y).

By means of the Bayes law, we obtain the posterior distribution; for all (x, y):

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x)}{p_{\mathrm{Y}}(y)} \tag{2.10}$$

Since the denominator does not depend of x, we often write:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x) \tag{2.11}$$

This formula is the basis one in Bayesian inference.

T130 How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{\mathrm{Y}|\mathrm{X}}$, then the distribution of $\mathrm{Y} \mid \mathrm{X}$,
- and the prior distribution, that is the distribution of X.

It is equivalent to say that we have the distribution of the pair (X, Y).

By means of the Bayes law, we obtain the posterior distribution; for all (x, y):

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x)}{p_{\mathrm{Y}}(y)} \tag{2.10}$$

Since the denominator does not depend of x, we often write:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x) \tag{2.11}$$

This formula is the basis one in Bayesian inference.

T130 How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{\mathrm{Y}|\mathrm{X}}$, then the distribution of $\mathrm{Y} \mid \mathrm{X}$,
- and the prior distribution, that is the distribution of X.

It is equivalent to say that we have the distribution of the pair (X, Y).

By means of the Bayes law, we obtain the posterior distribution; for all (x, y):

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x)}{p_{\mathrm{Y}}(y)} \tag{2.10}$$

Since the denominator does not depend of x, we often write:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x) \tag{2.11}$$

This formula is the basis one in Bayesian inference.

T130

How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{\mathrm{Y}|\mathrm{X}}$, then the distribution of $\mathrm{Y} \mid \mathrm{X}$,
- and the prior distribution, that is the distribution of X.

It is equivalent to say that we have the distribution of the pair (X, Y).

By means of the Bayes law, we obtain the posterior distribution; for all (x, y):

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) = \frac{p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x)}{p_{\mathrm{Y}}(y)} \tag{2.10}$$

Since the denominator does not depend of x, we often write:

$$p_{\mathrm{X}|\mathrm{Y}}(x, y) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(y, x) \tag{2.11}$$

This formula is the basis one in Bayesian inference.

T131 How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

⚠ A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics.
Do not make a confusion between "Bayesian estimation" (in which the parameter is equipped with a prior distribution) and "Bayesian interpretation" (which refers to the subjective signification of the probabilities).

⚠ Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.
With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).
This is controversial [ref] and leads to technical problems when the domain $\mathbb{X}$ is not bounded (**improper** distributions). See [ref] for more information.

T131

How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

⚠ A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics.
Do not make a confusion between "Bayesian estimation" (in which the parameter is equipped with a prior distribution) and "Bayesian interpretation" (which refers to the subjective signification of the probabilities).

⚠ Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.
With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).
This is controversial [ref] and leads to technical problems when the domain $\mathbb{X}$ is not bounded (**improper** distributions). See [ref] for more information.

T131 # How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics.
Do not make a confusion between "Bayesian estimation" (in which the parameter is equipped with a prior distribution) and "Bayesian interpretation" (which refers to the subjective signification of the probabilities).

Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.
With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).
This is controversial [ref] and leads to technical problems when the domain $\mathbb{X}$ is not bounded (**improper** distributions). See [ref] for more information.

T131 How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics.
Do not make a confusion between "Bayesian estimation" (in which the parameter is equipped with a prior distribution) and "Bayesian interpretation" (which refers to the subjective signification of the probabilities).

Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.
With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).
This is controversial [ref] and leads to technical problems when the domain $\mathbb{X}$ is not bounded (**improper** distributions). See [ref] for more information.

T131 # How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

⚠ A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics.
Do not make a confusion between "Bayesian estimation" (in which the parameter is equipped with a prior distribution) and "Bayesian interpretation" (which refers to the subjective signification of the probabilities).

⚠ Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.
With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).
This is controversial [ref] and leads to technical problems when the domain $\mathbb{X}$ is not bounded (**improper** distributions). See [ref] for more information.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{lll} \text{prior} & p_{\mathrm{X}}(x) & \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{lll} \text{prior} & p_{\mathrm{X}}(x) & \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) & \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) & \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{lll} \text{prior} & p_{\mathrm{X}}(x) & \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) & \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) & \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{ll} \text{prior} & p_{\mathrm{X}}(x) \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

➔ Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$
\begin{array}{lll}
\text{prior} & p_{\mathrm{X}}(x) & \\
\text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) & \\
\text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) &
\end{array}
\tag{2.12}
$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

➔ Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{lll} \text{prior} & p_{\mathrm{X}}(x) & \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) & \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) & \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

➲ Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{lll} \text{prior} & p_{\mathrm{X}}(x) & \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

➲ Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$\begin{array}{ll} \text{prior} & p_{\mathrm{X}}(x) \\ \text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) \\ \text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x) \end{array} \tag{2.12}$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

- which permits to intuitively set a confidence level,[7]
- which is **conjugate** for the likelihood [ref], this means that the posterior distribution belongs to the same family.[8]

➲ Thus, if the observation is a sample $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ of a r.v. Y, i.i.d. given the parameter, sequentially processed, for all x:

$$
\begin{array}{ll}
\text{prior} & p_{\mathrm{X}}(x) \\
\text{taking } \mathrm{Y}_1 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1}(x) \propto p_{\mathrm{X}}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_1, x) \\
\text{taking } \mathrm{Y}_2 \text{ into account} & p_{\mathrm{X}|\mathrm{Y}_1,\mathrm{Y}_2}(x) \propto p_{\mathrm{X}|\mathrm{Y}_1}(x)\, p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_2, x)
\end{array}
\tag{2.12}
$$

The posterior distribution of one stage becomes the prior distribution of the next stage.
With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.

T133 Mean error analysis

We evaluate a parameter X where all the d_x components are numeric.
A predictor provides a prediction $\check{x}$.
From an observation Y, an estimator provides an estimate $\widehat{x}(\mathrm{Y})$.

- The prediction is corrupted by the **prediction error** $\check{x} - \mathrm{X}$.
The estimate is corrupted by the **estimation error** $\widehat{x}(\mathrm{Y}) - \mathrm{X}$.
If a component of the error is positive, it is an overestimation.

T133 Mean error analysis

We evaluate a parameter X where all the d_x components are numeric.
A predictor provides a prediction $\breve{x}$.
From an observation Y, an estimator provides an estimate $\widehat{x}(\mathrm{Y})$.

- The prediction is corrupted by the **prediction error** $\breve{x} - \mathrm{X}$.
  The estimate is corrupted by the **estimation error** $\widehat{x}(\mathrm{Y}) - \mathrm{X}$.
  If a component of the error is positive, it is an overestimation.

T134 The performances of a predictor or an estimator can be measured by means of:[9]

- the **bias** (Bias,vector valued), that is the mean error,
  and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^\intercal W\, x}$, where W is a symmetric positive definite matrix W.
The MSE is the mean of the square norm of the error.
Necessarily, $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\,(W\ \mathrm{Errvar})$

9. If the quantities exist.

T134 The performances of a predictor or an estimator can be measured by means of:[9]

- the **bias** (Bias,vector valued), that is the mean error, and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^\intercal W\, x}$, where W is a symmetric positive definite matrix W.
The MSE is the mean of the square norm of the error.
Necessarily, $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\,(W\ \mathrm{Errvar})$

9. If the quantities exist.

T134 The performances of a predictor or an estimator can be measured by means of:[9]

- the **bias** (Bias,vector valued), that is the mean error,
  and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^\intercal W\, x}$, where W is a symmetric positive definite matrix W.
The MSE is the mean of the square norm of the error.
Necessarily, $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\,(W\ \mathrm{Errvar})$

9. If the quantities exist.

T134 The performances of a predictor or an estimator can be measured by means of:[9]

- the **bias** (Bias,vector valued), that is the mean error,
  and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^\intercal W\, x}$, where W is a symmetric positive definite matrix W.
The MSE is the mean of the square norm of the error.
Necessarily, $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\,(W\ \mathrm{Errvar})$

9. If the quantities exist.

T134 The performances of a predictor or an estimator can be measured by means of:[9]

- the **bias** (Bias,vector valued), that is the mean error,
  and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^\intercal W\, x}$, where W is a symmetric positive definite matrix W.
The MSE is the mean of the square norm of the error.
Necessarily, $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\,(W\ \mathrm{Errvar})$

The biais, the error variance, the MSE are expressed:

- given the parameter ($^{|\mathsf{p}}$) in classical estimation;
- by letting (X, Y) jointly varying, in Bayesian estimation.

9. If the quantities exist.

T135 Mean error, given the parameter

The bias of the estimator is the function from $\mathbb{X}$ to $\mathbb{R}^{d_x}$ which returns the error mean value:

$$\begin{aligned}\mathrm{Bias}_{\hat{x}}^{|\mathrm{P}}(\mathrm{x}) &= \mathrm{E}\left(\hat{x}(\mathrm{Y}) - \mathrm{x} \mid \mathrm{x}\right) \\ &= \mathrm{E}\left(\hat{x}(\mathrm{Y}) \mid \mathrm{x}\right) - \mathrm{x}\end{aligned} \tag{2.13}$$

If the bias is uniformly null, the estimator is **unbiased**.
A positive bias component means a tendency to overestimate.

The variance of the estimator is the function from $\mathbb{X}$ to $\mathbb{R}^{d_x \times d_x}$ which returns the estimation error variance, that is a symmetric positive semi-definite matrix:

$$\begin{aligned}\mathrm{Errvar}_{\hat{x}}^{|\mathrm{P}}(\mathrm{x}) &= \mathrm{Var}\left(\hat{x}(\mathrm{Y}) - \mathrm{x} \mid \mathrm{x}\right) \\ &= \mathrm{Var}\left(\hat{x}(\mathrm{Y}) \mid \mathrm{x}\right)\end{aligned} \tag{2.14}$$

Given the parameter, the estimation variance and the estimation error variance coincide.

T135

Mean error, given the parameter

The bias of the estimator is the function from $\mathbb{X}$ to $\mathbb{R}^{d_x}$ which returns the error mean value:

$$\begin{aligned} \mathrm{Bias}_{\widehat{x}}^{|\mathrm{P}}(\mathrm{X}) &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X} \mid \mathrm{X}\right) \\ &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) \mid \mathrm{X}\right) - \mathrm{X} \end{aligned} \tag{2.13}$$

If the bias is uniformly null, the estimator is **unbiased**.
A positive bias component means a tendency to overestimate.

➔ **The variance** of the estimator is the function from $\mathbb{X}$ to $\mathbb{R}^{d_x \times d_x}$ which returns the estimation error variance, that is a symmetric positive semi-definite matrix:

$$\begin{aligned} \mathrm{Errvar}_{\widehat{x}}^{|\mathrm{P}}(\mathrm{X}) &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X} \mid \mathrm{X}\right) \\ &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) \mid \mathrm{X}\right) \end{aligned} \tag{2.14}$$

⚠ Given the parameter, the estimation variance and the estimation error variance coincide.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average.
The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

But, in an estimation problem such that:

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $\mathrm{S}\,(\mathrm{Y} \mid \mathrm{X})$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:[10]

$$\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) \geqslant [\mathrm{FI}(\mathrm{X})]^{-1} \qquad \text{(for an unbiased estimator)} \tag{2.15}$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:[11]

$$\mathrm{FI}(\mathrm{X}) = \mathrm{Var}\left(\frac{\partial \ln p_{\mathrm{Y}|\mathrm{X}}}{\partial x}(\mathrm{Y}, \mathrm{X}) \mid \mathrm{X}\right) \tag{2.16}$$

A high Fisher information means that a parameter variation implies a strong observation variation.

10. The Cramer-Rao inequality can be extended to the biased case.
11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average. The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

But, in an estimation problem such that:

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X}\right)$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:[10]

$$\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) \geqslant [\mathrm{FI}(\mathrm{X})]^{-1} \qquad \text{(for an unbiased estimator)} \tag{2.15}$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:[11]

$$\mathrm{FI}(\mathrm{X}) = \mathrm{Var}\left(\frac{\partial \ln p_{\mathrm{Y}|\mathrm{X}}}{\partial x}(\mathrm{Y}, \mathrm{X}) \mid \mathrm{X}\right) \tag{2.16}$$

A high Fisher information means that a parameter variation implies a strong observation variation.

10. The Cramer-Rao inequality can be extended to the biased case.
11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average.
The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

➔ But, in an estimation problem such that:

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X}\right)$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:[10]

$$\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) \geqslant \left[\mathrm{FI}(\mathrm{X})\right]^{-1} \qquad \text{(for an unbiased estimator)} \tag{2.15}$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:[11]

$$\mathrm{FI}(\mathrm{X}) = \mathrm{Var}\left(\frac{\partial \ln p_{\mathrm{Y}|\mathrm{X}}}{\partial x}(\mathrm{Y}, \mathrm{X}) \mid \mathrm{X}\right) \tag{2.16}$$

A high Fisher information means that a parameter variation implies a strong observation variation.

10. The Cramer-Rao inequality can be extended to the biased case.
11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average.
The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

➲ But, in an estimation problem such that:

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X}\right)$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:[10]

$$\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) \geqslant [\mathrm{FI}(\mathrm{X})]^{-1} \qquad \text{(for an unbiased estimator)} \tag{2.15}$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:[11]

$$\mathrm{FI}(\mathrm{X}) = \mathrm{Var}\left(\frac{\partial \ln p_{\mathrm{Y}|\mathrm{X}}}{\partial x}(\mathrm{Y},\mathrm{X}) \mid \mathrm{X}\right) \tag{2.16}$$

A high Fisher information means that a parameter variation implies a strong observation variation.

10. The Cramer-Rao inequality can be extended to the biased case.
11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average. The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

But, in an estimation problem such that:

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X}\right)$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:[10]

$$\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) \geqslant [\mathrm{FI}(\mathrm{X})]^{-1} \qquad \text{(for an unbiased estimator)} \tag{2.15}$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:[11]

$$\mathrm{FI}(\mathrm{X}) = \mathrm{Var}\left(\frac{\partial \ln p_{\mathrm{Y}|\mathrm{X}}}{\partial x}(\mathrm{Y}, \mathrm{X}) \mid \mathrm{X}\right) \tag{2.16}$$

A high Fisher information means that a parameter variation implies a strong observation variation.

10. The Cramer-Rao inequality can be extended to the biased case.
11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

⚠ Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

⚠ Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

➜ **The MVUE** ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

⚠ Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

⚠ Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

➲ **The MVUE** ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T137 **An efficient estimator** is an unbiased estimator which reaches this bound.
If an efficient estimator exists, it is unique, and it is the MLE.

Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [ref]);
- the MLE is not necessarily efficient.

Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;[12]
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [ref].[13]

The MVUE ("Minimum Variance Unbiased estimator") may exist without being efficient (the variance does not reach the CRLB).

The BLUE ("Best Linear Unbiased estimator") is a minimum variance unbiased estimator among the linear estimators.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.
13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

T138 The **Mean square error (MSE)** is a positive scalar valued function which quantifies the estimator quality:[14]

$$\begin{aligned} \mathrm{MSE}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x}) &= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \mathrm{x}\|_W^2 \mid \mathrm{x}\right) \\ &= \left\|\mathrm{Bias}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x})\right\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x})\right) \end{aligned} \tag{2.17}$$

If the goal is to minimize the MSE, there are usually neither unbiasedness nor minimal variance, but a compromise.

14. Reminder: W is a positive definite matrix and $\|x\|_W = \sqrt{x^\top W x}$.

T138 The **Mean square error (MSE)** is a positive scalar valued function which quantifies the estimator quality:[14]

$$\begin{aligned}\mathrm{MSE}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x}) &= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \mathrm{x}\|_W^2 \;\mid \mathrm{x}\right) \\ &= \left\|\mathrm{Bias}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x})\right\|_W^2 + \mathrm{trace}\left(W \;\mathrm{Errvar}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x})\right)\end{aligned} \tag{2.17}$$

If the goal is to minimize the MSE, there are usually neither unbiasedness nor minimal variance, but a compromise.

14. Reminder: W is a positive definite matrix and $\|x\|_W = \sqrt{x^\intercal W x}$.

T1 We have postulated the existence and the unicity of the MLE.
But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

We have postulated the existence and the unicity of the MLE.
But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

T1 We have postulated the existence and the unicity of the MLE.
But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

We have postulated the existence and the unicity of the MLE.
But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

T1[illegible] We have postulated the existence and the unicity of the MLE.
But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

T140 ◁ ▷ **Exercise 28.** The observation Y is uniformly distributed between 0 and X. We estimate X (see exercise 26).

a) Give the bias, the variance, the MSE of the estimator $\widehat{x}(\mathrm{Y}) = \alpha\,\mathrm{Y} + \beta$.

b) Write the bias, the variance, the MSE of:

- the MLE,
- the BLUE,
- the estimator $\widehat{x}(\mathrm{Y}) = \alpha\,\mathrm{Y}$ which minimizes the MSE.

◁ ▷ **Exercise 29.** The observation Y is uniformly distributed between 0 and $1/\mathrm{X}$. We estimate X.

a) Give the MLE.

b) Does it exist an unbiased estimator?

T141 Mean error, Bayesian point of view

A priori

The bias, the error variance, the mean square error of a predictor $\breve{x}$ are defined below; they are rewritten by means of the MMSE predictor:[15]

$$\begin{aligned} \mathrm{Bias}_{\breve{x}} &= \mathrm{E}\left(\breve{x} - \mathrm{x}\right) &&= \breve{x} - \breve{x}_{\text{MMSE}} \\ \mathrm{Errvar}_{\breve{x}} &= \mathrm{Var}\left(\breve{x} - \mathrm{x}\right) &&= \mathrm{Var}\left(\mathrm{x}\right) \\ \mathrm{MSE}_{\breve{x}} &= \mathrm{E}\left(\left\|\breve{x} - \mathrm{x}\right\|_W^2\right) &&= \left\|\breve{x} - \breve{x}_{\text{MMSE}}\right\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Var}\left(\mathrm{x}\right)\right) \end{aligned} \tag{2.18}$$

- The error variance does not depend of the predictor.
  Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\breve{x}_{\text{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \begin{cases} \mathrm{Bias}_{\breve{x}_{\text{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\breve{x}_{\text{MMSE}}} &= \mathrm{Var}\left(\mathrm{x}\right) \end{cases} \tag{2.19}$$

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^\intercal W x}$, and $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Errvar}\right)$.

T141 Mean error, Bayesian point of view

A priori

The bias, the error variance, the mean square error of a predictor $\breve{x}$ are defined below; they are rewritten by means of the MMSE predictor:[15]

$$\begin{aligned}
\mathrm{Bias}_{\breve{x}} &= \mathrm{E}\left(\breve{x} - \mathrm{x}\right) && = \breve{x} - \breve{x}_{\mathrm{MMSE}} \\
\mathrm{Errvar}_{\breve{x}} &= \mathrm{Var}\left(\breve{x} - \mathrm{x}\right) && = \mathrm{Var}\left(\mathrm{x}\right) \\
\mathrm{MSE}_{\breve{x}} &= \mathrm{E}\left(\|\breve{x} - \mathrm{x}\|_W^2\right) && = \|\breve{x} - \breve{x}_{\mathrm{MMSE}}\|_W^2 + \mathrm{trace}\left(W \ \mathrm{Var}\left(\mathrm{x}\right)\right)
\end{aligned} \tag{2.18}$$

- The error variance does not depend of the predictor.
  Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\breve{x}_{\mathrm{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \begin{cases} \mathrm{Bias}_{\breve{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\breve{x}_{\mathrm{MMSE}}} &= \mathrm{Var}\left(\mathrm{x}\right) \end{cases} \tag{2.19}$$

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^\top W x}$, and $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\left(W \ \mathrm{Errvar}\right)$.

T141 # Mean error, Bayesian point of view

A priori

The bias, the error variance, the mean square error of a predictor $\breve{x}$ are defined below; they are rewritten by means of the MMSE predictor:[15]

$$\begin{aligned} \mathrm{Bias}_{\breve{x}} &= \mathrm{E}\left(\breve{x} - \mathrm{x}\right) &&= \breve{x} - \breve{x}_{\mathrm{MMSE}} \\ \mathrm{Errvar}_{\breve{x}} &= \mathrm{Var}\left(\breve{x} - \mathrm{x}\right) &&= \mathrm{Var}\left(\mathrm{x}\right) \\ \mathrm{MSE}_{\breve{x}} &= \mathrm{E}\left(\|\breve{x} - \mathrm{x}\|_W^2\right) &&= \|\breve{x} - \breve{x}_{\mathrm{MMSE}}\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Var}\left(\mathrm{x}\right)\right) \end{aligned} \tag{2.18}$$

- The error variance does not depend of the predictor.
  Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\breve{x}_{\mathrm{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \begin{cases} \mathrm{Bias}_{\breve{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\breve{x}_{\mathrm{MMSE}}} &= \mathrm{Var}\left(\mathrm{x}\right) \end{cases} \tag{2.19}$$

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^\intercal W x}$, and $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Errvar}\right)$.

T141 # Mean error, Bayesian point of view

A priori

The bias, the error variance, the mean square error of a predictor $\breve{x}$ are defined below; they are rewritten by means of the MMSE predictor:[15]

$$\begin{aligned} \mathrm{Bias}_{\breve{x}} &= \mathrm{E}\left(\breve{x} - \mathrm{x}\right) && = \breve{x} - \breve{x}_{\mathrm{MMSE}} \\ \mathrm{Errvar}_{\breve{x}} &= \mathrm{Var}\left(\breve{x} - \mathrm{x}\right) && = \mathrm{Var}\left(\mathrm{x}\right) \\ \mathrm{MSE}_{\breve{x}} &= \mathrm{E}\left(\|\breve{x} - \mathrm{x}\|_W^2\right) && = \|\breve{x} - \breve{x}_{\mathrm{MMSE}}\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Var}\left(\mathrm{x}\right)\right) \end{aligned} \tag{2.18}$$

- The error variance does not depend of the predictor.
  Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\breve{x}_{\mathrm{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \begin{cases} \mathrm{Bias}_{\breve{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\breve{x}_{\mathrm{MMSE}}} &= \mathrm{Var}\left(\mathrm{x}\right) \end{cases} \tag{2.19}$$

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^\intercal W x}$, and $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\left(W\ \mathrm{Errvar}\right)$.

T141 # Mean error, Bayesian point of view

A priori
The bias, the error variance, the mean square error of a predictor $\breve{x}$ are defined below; they are rewritten by means of the MMSE predictor:[15]

$$\begin{aligned}
\mathrm{Bias}_{\breve{x}} &= \mathrm{E}\left(\breve{x} - \mathrm{x}\right) && = \breve{x} - \breve{x}_{\text{MMSE}} \\
\mathrm{Errvar}_{\breve{x}} &= \mathrm{Var}\left(\breve{x} - \mathrm{x}\right) && = \mathrm{Var}\left(\mathrm{x}\right) \\
\mathrm{MSE}_{\breve{x}} &= \mathrm{E}\left(\|\breve{x} - \mathrm{x}\|_W^2\right) && = \|\breve{x} - \breve{x}_{\text{MMSE}}\|_W^2 + \mathrm{trace}\left(W \ \mathrm{Var}\left(\mathrm{x}\right)\right)
\end{aligned} \tag{2.18}$$

➲ The error variance does not depend of the predictor.
Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\breve{x}_{\text{MMSE}} = \mathrm{E}\left(\mathrm{x}\right) \qquad \begin{cases} \mathrm{Bias}_{\breve{x}_{\text{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\breve{x}_{\text{MMSE}}} &= \mathrm{Var}\left(\mathrm{x}\right) \end{cases} \tag{2.19}$$

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^\intercal W x}$, and $\mathrm{MSE} = \|\mathrm{Bias}\|_W^2 + \mathrm{trace}\left(W \ \mathrm{Errvar}\right)$.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned}
\mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right)\\
\mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\\
\mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\|_W^2\right) &&= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right)
\end{aligned} \tag{2.20}$$

- Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} &= \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

⚠ Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned} \mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) && = \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) \\ \mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) && = \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \\ \mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\right\|_W^2\right) && = \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right) \end{aligned} \tag{2.20}$$

➔ Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} & = 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} & = \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

⚠ Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned}
\mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) \\
\mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \\
\mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\right\|_W^2\right) &&= \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right)
\end{aligned} \tag{2.20}$$

- Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} &= \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

⚠ Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned}
\mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) \\
\mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \\
\mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\right\|_W^2\right) &&= \mathrm{E}\left(\left\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right)
\end{aligned} \tag{2.20}$$

Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} &= \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned} \mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) \\ \mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) &&= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \\ \mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\|_W^2\right) &&= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right) \end{aligned} \tag{2.20}$$

➲ Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} &= \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

⚠ Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T142 **A posteriori**

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:[P1]

$$\begin{aligned}
\mathrm{Bias}_{\widehat{x}} &= \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) && = \mathrm{E}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) \\
\mathrm{Errvar}_{\widehat{x}} &= \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \mathrm{X}\right) && = \mathrm{Var}\left(\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\right) + \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \\
\mathrm{MSE}_{\widehat{x}} &= \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \mathrm{X}\|_W^2\right) && = \mathrm{E}\left(\|\widehat{x}(\mathrm{Y}) - \widehat{x}_{\mathrm{MMSE}}(\mathrm{Y})\|_W^2\right) + \mathrm{trace}\left(W\ \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)\right)
\end{aligned} \tag{2.20}$$

➜ Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.[16] [P2]

$$\widehat{x}_{\mathrm{MMSE}}(\mathrm{Y}) = \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{MMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{MMSE}}} &= \mathrm{E}\left(\mathrm{Var}\left(\mathrm{X} \mid \mathrm{Y}\right)\right) \end{cases} \tag{2.21}$$

⚠ Do not muddle the estimation variance and the estimation error variance!

16. Whatever the weighting matrix W is.

T143 If the observation is numeric with d_y components, we can decide to look for:

- an unbiased estimator with minimal error variance,
- or an estimator with minimal MSE,

among the linear estimators.

Both objectives lead to the LMMSE estimator: [P3]

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - m_{\mathrm{Y}}) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X},\mathrm{X}} - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \end{cases} \tag{2.22}$$

If $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model, the LMMSE estimator is the MMSE one.

T143 If the observation is numeric with d_y components, we can decide to look for:

- an unbiased estimator with minimal error variance,
- or an estimator with minimal MSE,

among the linear estimators.

➔ Both objectives lead to the LMMSE estimator: [P3]

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - m_{\mathrm{Y}}) \qquad \begin{cases} \text{Bias}_{\widehat{x}_{\text{LMMSE}}} &= 0_{d_x} \\ \text{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X},\mathrm{X}} - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \end{cases} \tag{2.22}$$

If $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model, the LMMSE estimator is the MMSE one.

T143 If the observation is numeric with d_y components, we can decide to look for:

- an unbiased estimator with minimal error variance,
- or an estimator with minimal MSE,

among the linear estimators.

Both objectives lead to the LMMSE estimator: [P3]

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, (\mathrm{Y} - m_{\mathrm{Y}}) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X},\mathrm{X}} - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \end{cases} \tag{2.22}$$

If $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model, the LMMSE estimator is the MMSE one.

T144 **From prior to posterior**

The MMSE predictor, the LMMSE estimator and the MMSE estimator are all unbiased.
The error variances are sorted in decreasing order below:[17]

$$\begin{aligned} &\text{MMSE predictor} && \mathrm{Errvar}_{\check{x}_{\text{MMSE}}} = C_{\text{X,X}} \\ &\text{LMMSE estimator} && \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} = C_{\text{X,X}} - C_{\text{X,Y}}\, C_{\text{Y,Y}}^{-1}\, C_{\text{Y,X}} \\ &\text{MMSE estimator} && \mathrm{Errvar}_{\widehat{x}_{\text{MMSE}}} = C_{\text{X,X}} - \mathrm{Var}\left(\mathrm{E}\left(\text{X} \mid \text{Y}\right)\right) \end{aligned} \tag{2.23}$$

➲ Thus, with respect to the MMSE predictor:

- the LMMSE estimator decreases the variance only if X and Y are correlated,
- the MMSE estimator decreases the variance only if $\mathrm{E}\left(\text{X} \mid \text{Y}\right)$ is not uniform.

17. The MMSE estimator is, by construction, the estimator with the lowest error variance. We can confirm that the LMMSE estimator error variance is indeed greater, since the Schur complement of $\mathrm{Var}\left(\mathrm{E}\left(\text{X} \mid \text{Y}\right)\right)$ in $\mathrm{Var}\left(\begin{bmatrix} \mathrm{E}\left(\text{X} \mid \text{Y}\right) \\ \text{Y} \end{bmatrix}\right)$ is positive definite (we use also the result of the exercise 19, page 89).

T144 **From prior to posterior**

The MMSE predictor, the LMMSE estimator and the MMSE estimator are all unbiased.
The error variances are sorted in decreasing order below:[17]

$$\begin{aligned}
&\text{MMSE predictor} && \mathrm{Errvar}_{\breve{x}_{\text{MMSE}}} = C_{\mathrm{X},\mathrm{X}} \\
&\text{LMMSE estimator} && \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} = C_{\mathrm{X},\mathrm{X}} - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{Y},\mathrm{X}} \\
&\text{MMSE estimator} && \mathrm{Errvar}_{\widehat{x}_{\text{MMSE}}} = C_{\mathrm{X},\mathrm{X}} - \mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)
\end{aligned} \tag{2.23}$$

➲ Thus, with respect to the MMSE predictor:

- the LMMSE estimator decreases the variance only if X and Y are correlated,
- the MMSE estimator decreases the variance only if $\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)$ is not uniform.

17. The MMSE estimator is, by construction, the estimator with the lowest error variance. We can confirm that the LMMSE estimator error variance is indeed greater, since the Schur complement of $\mathrm{Var}\left(\mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right)\right)$ in $\mathrm{Var}\left(\begin{bmatrix} \mathrm{E}\left(\mathrm{X} \mid \mathrm{Y}\right) \\ \mathrm{Y} \end{bmatrix}\right)$ is positive definite (we use also the result of the exercise 19, page 89).

T145 Linear model case

We assume that $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model (page 104); there exist a known matrix H and a r.v. W with known mean m_{W} and variance $C_{\mathrm{W,W}}$ such that:

$$\mathrm{Y} = H\,\mathrm{X} + \mathrm{W} \text{ with } \begin{cases} \mathrm{E}\left(\mathrm{W} \mid \mathrm{X}\right) & = m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{W} \mid \mathrm{X}\right) & = C_{\mathrm{W,W}} \end{cases} \tag{2.24}$$

We can derive the optimal linear estimators.

T145 # Linear model case

We assume that $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model (page 104); there exist a known matrix H and a r.v. W with known mean m_{W} and variance $C_{\mathrm{W},\mathrm{W}}$ such that:

$$\mathrm{Y} = H\,\mathrm{X} + \mathrm{W} \text{ with } \begin{cases} \mathrm{E}\left(\mathrm{W} \mid \mathrm{X}\right) & = m_{\mathrm{W}} \\ \mathrm{Var}\left(\mathrm{W} \mid \mathrm{X}\right) & = C_{\mathrm{W},\mathrm{W}} \end{cases} \tag{2.24}$$

We can derive the optimal linear estimators.

T146 The BLUE is written as:[P4]

$$\widehat{x}_{\text{BLUE}}(\mathrm{Y}) = (H^\intercal C_{\mathrm{W,W}}^{-1} H)^{-1} H^\intercal C_{\mathrm{W,W}}^{-1} (\mathrm{Y} - m_{\mathrm{W}}) \qquad \begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{BLUE}}}^{|\mathrm{p}}(\mathrm{X}) &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{BLUE}}}^{|\mathrm{p}}(\mathrm{X}) &= (H^\intercal C_{\mathrm{W,W}}^{-1} H)^{-1} \end{cases} \tag{2.25}$$

It minimizes the criterion $(\mathrm{Y} - H\,x)^\intercal C_{\mathrm{W,W}}^{-1} (\mathrm{Y} - H\,x)$ with respect to x.[18]

➔ If X has a mean m_{X} and a variance $C_{\mathrm{X,X}}$, the LMMSE estimator is written as[19]

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X,X}} H^\intercal (H\,C_{\mathrm{X,X}} H^\intercal + C_{\mathrm{W,W}})^{-1} (\mathrm{Y} - H\,m_{\mathrm{X}} - m_{\mathrm{W}})$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X,X}} - C_{\mathrm{X,X}} H^\intercal (H\,C_{\mathrm{X,X}} H^\intercal + C_{\mathrm{W,W}})^{-1} H\,C_{\mathrm{X,X}} \end{cases} \tag{2.26}$$

18. This is the Generalized Least Squares estimator (GLS). The Ordinary Least Squares estimator (OLS) is $\widehat{x}_{\text{OLS}}(\mathrm{Y}) = (H^\intercal H)^{-1} H^\intercal (\mathrm{Y} - m_{\mathrm{W}})$
19. This is a re-writing of (2.22), by means of the transform (1.84):

$$\begin{cases} m_{\mathrm{Y}} &= H\,m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y,Y}} &= H\,C_{\mathrm{X,X}} H^\intercal + C_{\mathrm{W,W}} \\ C_{\mathrm{X,Y}} &= C_{\mathrm{X,X}} H^\intercal \end{cases} \quad \text{and} \quad \begin{cases} \widehat{x}_{\text{LMMSE}}(\mathrm{Y}) &= m_{\mathrm{X}} + C_{\mathrm{X,Y}} C_{\mathrm{Y,Y}}^{-1} (\mathrm{Y} - m_{\mathrm{Y}}) \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X,X}} - C_{\mathrm{X,Y}} C_{\mathrm{Y,Y}}^{-1} C_{\mathrm{Y,X}} \end{cases}$$

T146 The BLUE is written as:[P4]

$$\widehat{x}_{\text{BLUE}}(\mathrm{Y}) = (H^{\mathsf{T}}\, C_{\mathrm{W,W}}^{-1}\, H)^{-1}\, H^{\mathsf{T}}\, C_{\mathrm{W,W}}^{-1}\, (\mathrm{Y} - m_{\mathrm{W}}) \qquad \begin{cases} \mathrm{Bias}^{|\mathrm{p}}_{\widehat{x}_{\text{BLUE}}}(\mathrm{X}) &= 0_{d_x} \\ \mathrm{Errvar}^{|\mathrm{p}}_{\widehat{x}_{\text{BLUE}}}(\mathrm{X}) &= (H^{\mathsf{T}}\, C_{\mathrm{W,W}}^{-1}\, H)^{-1} \end{cases} \tag{2.25}$$

It minimizes the criterion $(\mathrm{Y} - H\,x)^{\mathsf{T}}\, C_{\mathrm{W,W}}^{-1}\, (\mathrm{Y} - H\,x)$ with respect to x.[18]

If X has a mean m_{X} and a variance $C_{\mathrm{X,X}}$, the LMMSE estimator is written as[19]

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = m_{\mathrm{X}} + C_{\mathrm{X,X}}\, H^{\mathsf{T}}\, (H\, C_{\mathrm{X,X}}\, H^{\mathsf{T}} + C_{\mathrm{W,W}})^{-1}\, (\mathrm{Y} - H\, m_{\mathrm{X}} - m_{\mathrm{W}})$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X,X}} - C_{\mathrm{X,X}}\, H^{\mathsf{T}}\, (H\, C_{\mathrm{X,X}}\, H^{\mathsf{T}} + C_{\mathrm{W,W}})^{-1}\, H\, C_{\mathrm{X,X}} \end{cases} \tag{2.26}$$

18. This is the Generalized Least Squares estimator (GLS). The Ordinary Least Squares estimator (OLS) is $\widehat{x}_{\text{OLS}}(\mathrm{Y}) = (H^{\mathsf{T}}\, H)^{-1}\, H^{\mathsf{T}}\, (\mathrm{Y} - m_{\mathrm{W}})$
19. This is a re-writing of (2.22), by means of the transform (1.84):

$$\begin{cases} m_{\mathrm{Y}} &= H\, m_{\mathrm{X}} + m_{\mathrm{W}} \\ C_{\mathrm{Y,Y}} &= H\, C_{\mathrm{X,X}}\, H^{\mathsf{T}} + C_{\mathrm{W,W}} \\ C_{\mathrm{X,Y}} &= C_{\mathrm{X,X}}\, H^{\mathsf{T}} \end{cases} \quad \text{and} \quad \begin{cases} \widehat{x}_{\text{LMMSE}}(\mathrm{Y}) &= m_{\mathrm{X}} + C_{\mathrm{X,Y}}\, C_{\mathrm{Y,Y}}^{-1}\, (\mathrm{Y} - m_{\mathrm{Y}}) \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} &= C_{\mathrm{X,X}} - C_{\mathrm{X,Y}}\, C_{\mathrm{Y,Y}}^{-1}\, C_{\mathrm{Y,X}} \end{cases}$$

T146 The BLUE is written as:[P4]

$$\widehat{x}_{\text{BLUE}}(\text{Y}) = (H^{\intercal} C_{\text{W,W}}^{-1} H)^{-1} H^{\intercal} C_{\text{W,W}}^{-1} (\text{Y} - m_{\text{W}}) \qquad \begin{cases} \text{Bias}_{\widehat{x}_{\text{BLUE}}}^{|\text{p}}(\text{X}) & = 0_{d_x} \\ \text{Errvar}_{\widehat{x}_{\text{BLUE}}}^{|\text{p}}(\text{X}) & = (H^{\intercal} C_{\text{W,W}}^{-1} H)^{-1} \end{cases} \tag{2.25}$$

It minimizes the criterion $(\text{Y} - H\,x)^{\intercal} C_{\text{W,W}}^{-1} (\text{Y} - H\,x)$ with respect to x.[18]

- If X has a mean m_{X} and a variance $C_{\text{X,X}}$, the LMMSE estimator is written as[19]

$$\widehat{x}_{\text{LMMSE}}(\text{Y}) = m_{\text{X}} + C_{\text{X,X}} H^{\intercal} (H\, C_{\text{X,X}} H^{\intercal} + C_{\text{W,W}})^{-1} (\text{Y} - H\, m_{\text{X}} - m_{\text{W}})$$

$$\begin{cases} \text{Bias}_{\widehat{x}_{\text{LMMSE}}} & = 0_{d_x} \\ \text{Errvar}_{\widehat{x}_{\text{LMMSE}}} & = C_{\text{X,X}} - C_{\text{X,X}} H^{\intercal} (H\, C_{\text{X,X}} H^{\intercal} + C_{\text{W,W}})^{-1} H\, C_{\text{X,X}} \end{cases} \tag{2.26}$$

18. This is the Generalized Least Squares estimator (GLS). The Ordinary Least Squares estimator (OLS) is $\widehat{x}_{\text{OLS}}(\text{Y}) = (H^{\intercal} H)^{-1} H^{\intercal} (\text{Y} - m_{\text{W}})$
19. This is a re-writing of (2.22), by means of the transform (1.84):

$$\begin{cases} m_{\text{Y}} & = H\, m_{\text{X}} + m_{\text{W}} \\ C_{\text{Y,Y}} & = H\, C_{\text{X,X}} H^{\intercal} + C_{\text{W,W}} \\ C_{\text{X,Y}} & = \quad C_{\text{X,X}} H^{\intercal} \end{cases} \quad \text{and} \quad \begin{cases} \widehat{x}_{\text{LMMSE}}(\text{Y}) & = m_{\text{X}} + C_{\text{X,Y}} C_{\text{Y,Y}}^{-1} (\text{Y} - m_{\text{Y}}) \\ \text{Errvar}_{\widehat{x}_{\text{LMMSE}}} & = C_{\text{X,X}} - C_{\text{X,Y}} C_{\text{Y,Y}}^{-1} C_{\text{Y,X}} \end{cases}$$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{W},\mathrm{W}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\mathsf{T}}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1} \left[C_{\mathrm{X},\mathrm{X}}^{-1}\, m_{\mathrm{X}} + H^{\mathsf{T}}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, (\mathrm{Y} - m_{\mathrm{W}})\right]$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{LMMSE}}} & = 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{LMMSE}}} & = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\mathsf{T}}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1} \end{cases} \tag{2.27}$$

With a weak prior information, that is $C_{\mathrm{X},\mathrm{X}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{W},\mathrm{W}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\mathrm{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\mathrm{BLUE}}} = (H^{\mathsf{T}}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1}$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{W},\mathrm{W}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\intercal}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1} \left[C_{\mathrm{X},\mathrm{X}}^{-1}\, m_{\mathrm{X}} + H^{\intercal}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, (\mathrm{Y} - m_{\mathrm{W}})\right]$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{LMMSE}}} & = 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{LMMSE}}} & = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\intercal}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1} \end{cases} \tag{2.27}$$

With a weak prior information, that is $C_{\mathrm{X},\mathrm{X}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{W},\mathrm{W}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\mathrm{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\mathrm{BLUE}}} = (H^{\intercal}\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1}$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{W},\mathrm{W}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\top} C_{\mathrm{W},\mathrm{W}}^{-1} H)^{-1} [C_{\mathrm{X},\mathrm{X}}^{-1} m_{\mathrm{X}} + H^{\top} C_{\mathrm{W},\mathrm{W}}^{-1} (\mathrm{Y} - m_{\mathrm{W}})]$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{LMMSE}}} &= (C_{\mathrm{X},\mathrm{X}}^{-1} + H^{\top} C_{\mathrm{W},\mathrm{W}}^{-1} H)^{-1} \end{cases} \tag{2.27}$$

With a weak prior information, that is $C_{\mathrm{X},\mathrm{X}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{W},\mathrm{W}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\mathrm{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\mathrm{BLUE}}} = (H^{\top} C_{\mathrm{W},\mathrm{W}}^{-1} H)^{-1}$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{w,w}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{x,x}}^{-1} + H^\intercal C_{\mathrm{w,w}}^{-1} H)^{-1} \left[C_{\mathrm{x,x}}^{-1} m_{\mathrm{X}} + H^\intercal C_{\mathrm{w,w}}^{-1} (\mathrm{Y} - m_{\mathrm{w}})\right]$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{LMMSE}}} &= (C_{\mathrm{x,x}}^{-1} + H^\intercal C_{\mathrm{w,w}}^{-1} H)^{-1} \end{cases} \tag{2.27}$$

With a weak prior information, that is $C_{\mathrm{x,x}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{w,w}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\mathrm{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\mathrm{BLUE}}} = (H^\intercal C_{\mathrm{w,w}}^{-1} H)^{-1}$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{w,w}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\text{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{x,x}}^{-1} + H^{\intercal}\, C_{\mathrm{w,w}}^{-1}\, H)^{-1}\, [C_{\mathrm{x,x}}^{-1}\, m_{\mathrm{x}} + H^{\intercal}\, C_{\mathrm{w,w}}^{-1}\, (\mathrm{Y} - m_{\mathrm{w}})]$$
$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\text{LMMSE}}} & = 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\text{LMMSE}}} & = (C_{\mathrm{x,x}}^{-1} + H^{\intercal}\, C_{\mathrm{w,w}}^{-1}\, H)^{-1} \end{cases} \qquad (2.27)$$

With a weak prior information, that is $C_{\mathrm{x,x}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{w,w}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\text{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\text{BLUE}}} = (H^{\intercal}\, C_{\mathrm{w,w}}^{-1}\, H)^{-1}$

T147 Can we make a link between the numerous estimators?

- If W is normally distributed and independent of X, the BLUE is the MLE, and is efficient.[P5]
- If X and W are normally distributed and independent, the LMMSE estimator is the MMSE one.[20]
- If $C_{\mathrm{W},\mathrm{W}}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\widehat{x}_{\mathrm{LMMSE}}(\mathrm{Y}) = (C_{\mathrm{X},\mathrm{X}}^{-1} + H^\intercal\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1}\, [C_{\mathrm{X},\mathrm{X}}^{-1}\, m_{\mathrm{X}} + H^\intercal\, C_{\mathrm{W},\mathrm{W}}^{-1}\, (\mathrm{Y} - m_{\mathrm{W}})]$$

$$\begin{cases} \mathrm{Bias}_{\widehat{x}_{\mathrm{LMMSE}}} &= 0_{d_x} \\ \mathrm{Errvar}_{\widehat{x}_{\mathrm{LMMSE}}} &= (C_{\mathrm{X},\mathrm{X}}^{-1} + H^\intercal\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1} \end{cases} \qquad (2.27)$$

With a weak prior information, that is $C_{\mathrm{X},\mathrm{X}}^{-1}$ almost null, we get back the BLUE.[21]

If the observation is highly noisy, that is $C_{\mathrm{W},\mathrm{W}}^{-1}$ almost null, we get back the MMSE predictor.

20. More generally, it is sufficient that $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model.
21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\mathrm{Bias}_{\widehat{x}_{\mathrm{BLUE}}} = 0$ and $\mathrm{Errvar}_{\widehat{x}_{\mathrm{BLUE}}} = (H^\intercal\, C_{\mathrm{W},\mathrm{W}}^{-1}\, H)^{-1}$

T148 **Exercise 30.** Let's consider d_y thermometers. Each of them provides a measure Y_i of the actual temperature X. Given X, the errors are supposed zero-mean, independent, with the same variance σ^2.

a) write the BLUE and its variance.

b) What is the criterion which is minimized by this solution?

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{d_y} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} x + \begin{bmatrix} w_1 \\ \vdots \\ w_{d_y} \end{bmatrix}$$

$$Y = H \; X + W$$

$$H = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad m_w = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$C_{ww} = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

$$\hat{x}_{BLUE}(Y) = \left(H^T C_{ww}^{-1} H\right)^{-1} H^T C_{ww}^{-1} (Y - m_w)$$

$$\mathrm{ErrVar}(X) = \left(H^T C_{ww}^{-1} H\right)^{-1}$$

$$\hat{x}_{BLUE}(Y) = \frac{1}{d_y} \sum_{i=1}^{d_y} y_i$$

$$\mathrm{ErrVar}(K) = \frac{\sigma^2}{d_y}$$

T149 # Probability error analysis

We are interested in the estimation of a discrete valued parameter.
In an From an observation Y, an estimator provides an estimate $\widehat{x}(\mathrm{Y})$.

T149 # Probability error analysis

We are interested in the estimation of a discrete valued parameter.
In an From an observation Y, an estimator provides an estimate $\widehat{x}(\mathrm{Y})$.

- If $\widehat{x}(\mathrm{Y}) \neq \mathrm{X}$, we make an error.

The performances of an estimator can be measured by means of the error probability.

T150 **Given the parameter**, the error probability is the function from $\mathbb{X}$ to $[0,1]$ defined as:

$$\begin{aligned}\mathrm{Errprob}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x}) &= \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) \neq \mathrm{x} \mid \mathrm{x}\right)\\ &= 1 - \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) = \mathrm{x} \mid \mathrm{x}\right)\\ &= 1 - \int_{\{y\in\mathbb{Y} \mid \widehat{x}(y)=\mathrm{x}\}} p_{\mathrm{Y}|\mathrm{X}}(y)\;\mathrm{d}\,y\end{aligned} \tag{2.28}$$

For one value of the parameter, this probability is null (so minimal), if the integration domain is the support of the conditional distribution.

T150 **Given the parameter**, the error probability is the function from $\mathbb{X}$ to $[0,1]$ defined as:

$$\begin{aligned}\mathrm{Errprob}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{X}) &= \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) \neq \mathrm{X} \mid \mathrm{X}\right) \\ &= 1 - \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) = \mathrm{X} \mid \mathrm{X}\right) \\ &= 1 - \int_{\{y \in \mathbb{Y} \mid \widehat{x}(y) = \mathrm{X}\}} p_{\mathrm{Y}|\mathrm{X}}(y)\, \mathrm{d}\, y\end{aligned} \tag{2.28}$$

For one value of the parameter, this probability is null (so minimal), if the integration domain is the support of the conditional distribution.

- Thus, it is possible to nullify this error probability for all possible values of the parameter only if the supports $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X} = x\right)$, $x \in \mathbb{X}$ form a partition of $\mathrm{S}\left(\mathrm{Y}\right)$.
  The MLE provides the unique x such that $p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is not zero, and nullify the error probability.

  Nevertheless, in the general case, a compromise is necessary.

T150 **Given the parameter**, the error probability is the function from $\mathbb{X}$ to $[0,1]$ defined as:

$$\begin{aligned} \mathrm{Errprob}_{\widehat{x}}^{|\mathrm{p}}(\mathrm{x}) &= \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) \neq \mathrm{x} \mid \mathrm{x}\right) \\ &= 1 - \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) = \mathrm{x} \mid \mathrm{x}\right) \\ &= 1 - \int_{\{y \in \mathbb{Y} \mid \widehat{x}(y) = \mathrm{x}\}} p_{\mathrm{Y}|\mathrm{X}}(y)\, \mathrm{d}\, y \end{aligned} \tag{2.28}$$

For one value of the parameter, this probability is null (so minimal), if the integration domain is the support of the conditional distribution.

- Thus, it is possible to nullify this error probability for all possible values of the parameter only if the supports $\mathrm{S}\left(\mathrm{Y} \mid \mathrm{X} = x\right)$, $x \in \mathbb{X}$ form a partition of $\mathrm{S}\left(\mathrm{Y}\right)$.
  The MLE provides the unique x such that $p_{\mathrm{Y}|\mathrm{X}}(y, x)$ is not zero, and nullify the error probability.

  Nevertheless, in the general case, a compromise is necessary.

- **From a Bayesian point of view**, the error probability is a number in $[0,1]$ defined as:

$$\mathrm{Errprob}_{\widehat{x}} = \mathrm{Prob}\left(\widehat{x}(\mathrm{Y}) \neq \mathrm{X}\right) \tag{2.29}$$

  The MAP estimator minimizes the error probability.[P6][22]

22. In classification problems, this definition of the error probability is generally used.

T151 # Practical remarks

In numerous actual cases, the observation is a sample $\vec{\mathrm{Y}}_{n_\mathrm{r}} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_{n_\mathrm{r}})$ of a r.v. Y such that:

- the distribution of Y the parameter to estimate X has a known form;
- The sample is assumed i.i.d. given the unknown parameter.

The estimator corresponds to the definition of $\widehat{x}(\vec{\mathrm{Y}}_{n_\mathrm{r}})$.

- The likelihood is, for all $x \in \mathrm{S}(\mathrm{X})$:

$$p_{\vec{\mathrm{Y}}_{n_\mathrm{r}}|\mathrm{X}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}, x) = \prod_{k=1}^{n_\mathrm{r}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.30}$$

- Since the log function is increasing, the likelihood maximization is equivalent to the log-likelihood maximization:

$$\widehat{x}_{\mathrm{MLE}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}) = \arg\max_x \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) \quad \text{with} \quad \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) = \sum_{k=1}^{n_\mathrm{r}} \ln p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.31}$$

This trick simplifies the calculation in general.

T151 Practical remarks

In numerous actual cases, the observation is a sample $\vec{\mathrm{Y}}_{n_\mathrm{r}} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_{n_\mathrm{r}})$ of a r.v. Y such that:

- the distribution of Y the parameter to estimate X has a known form;
- The sample is assumed i.i.d. given the unknown parameter.

The estimator corresponds to the definition of $\widehat{x}(\vec{\mathrm{Y}}_{n_\mathrm{r}})$.

➔ The likelihood is, for all $x \in \mathrm{S}(\mathrm{X})$:

$$p_{\vec{\mathrm{Y}}_{n_\mathrm{r}}|\mathrm{X}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}, x) = \prod_{k=1}^{n_\mathrm{r}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.30}$$

➔ Since the log function is increasing, the likelihood maximization is equivalent to the log-likelihood maximization:

$$\widehat{x}_{\mathrm{MLE}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}) = \arg\max_x \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) \quad \text{with} \quad \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) = \sum_{k=1}^{n_\mathrm{r}} \ln p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.31}$$

This trick simplifies the calculation in general.

T151
Practical remarks

In numerous actual cases, the observation is a sample $\vec{\mathrm{Y}}_{n_\mathrm{r}} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_{n_\mathrm{r}})$ of a r.v. Y such that:

- the distribution of Y the parameter to estimate X has a known form;
- The sample is assumed i.i.d. given the unknown parameter.

The estimator corresponds to the definition of $\widehat{x}(\vec{\mathrm{Y}}_{n_\mathrm{r}})$.

- The likelihood is, for all $x \in \mathrm{S}(\mathrm{X})$:

$$p_{\vec{\mathrm{Y}}_{n_\mathrm{r}}|\mathrm{X}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}, x) = \prod_{k=1}^{n_\mathrm{r}} p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.30}$$

- Since the log function is increasing, the likelihood maximization is equivalent to the log-likelihood maximization:

$$\widehat{x}_{\mathrm{MLE}}(\vec{\mathrm{Y}}_{n_\mathrm{r}}) = \arg\max_x \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) \quad \text{with} \quad \ell_{\vec{\mathrm{Y}}_{n_\mathrm{r}}}(x) = \sum_{k=1}^{n_\mathrm{r}} \ln p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}_k, x) \tag{2.31}$$

This trick simplifies the calculation in general.

T152 ◁ ▷ **Exercise 31.** We want to check if a coin is correctly balanced. X is the probability to obtain "tail". We throw n_{r} times the coin. For n from 1 to n_{r}, Y_n takes the value 1 if we obtain "tail", and takes the value 0 if we obtain "head".

Write the MLE of X from a sample of Y with size n_{r}.

T153 ◁ ▷ **Exercise 32.** The r.v. Y takes the value $c \in \{1, \ldots, n_\mathrm{c}\}$ with the probability X_c. Write the MLE of $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_{n_\mathrm{c}})$ from a sample of Y with size n_r.

Beware! The problem has the constraint $\sum_{c=1}^{n_\mathrm{c}} \mathrm{X}_c = 1$, and can be easily solved using Lagrange multipliers technique.

T154 Summary

The estimation of an unknown parameter X from an observation Y is formalized by:

- the distribution of $\mathrm{Y} \mid \mathrm{X}$ in the classical estimation.
- the distribution of $\mathrm{X} \mid \mathrm{Y}$ in the Bayesian estimation (in practice, the distribution of (X, Y)).

This problem reduces to an optimization one (MLE, MAP) or an integration one (MMSE).

T154 # Summary

The estimation of an unknown parameter X from an observation Y is formalized by:

- the distribution of $\mathrm{Y} \mid \mathrm{X}$ in the classical estimation.
- the distribution of $\mathrm{X} \mid \mathrm{Y}$ in the Bayesian estimation (in practice, the distribution of (X, Y)).

This problem reduces to an optimization one (MLE, MAP) or an integration one (MMSE).

➔ The LMMSE estimator can be written by means of the second order properties m_{Y}, $C_{\mathrm{Y},\mathrm{Y}}$, m_{X}, $C_{\mathrm{X},\mathrm{X}}$, $C_{\mathrm{X},\mathrm{Y}}$.

T154 Summary

The estimation of an unknown parameter X from an observation Y is formalized by:

- the distribution of $\mathrm{Y} \mid \mathrm{X}$ in the classical estimation.
- the distribution of $\mathrm{X} \mid \mathrm{Y}$ in the Bayesian estimation (in practice, the distribution of (X, Y)).

This problem reduces to an optimization one (MLE, MAP) or an integration one (MMSE).

➔ The LMMSE estimator can be written by means of the second order properties m_{Y}, $C_{\mathrm{Y},\mathrm{Y}}$, m_{X}, $C_{\mathrm{X},\mathrm{X}}$, $C_{\mathrm{X},\mathrm{Y}}$.

➔ A linear model is an assumption on the conditional mean (assumed linear) and the conditional variance (assumed uniform).

if $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model, the LMMSE estimator is the MMSE one.
If $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model, we can write the BLUE and re-write the LMMSE estimator.

T154 # Summary

The estimation of an unknown parameter X from an observation Y is formalized by:

- the distribution of $\mathrm{Y} \mid \mathrm{X}$ in the classical estimation.
- the distribution of $\mathrm{X} \mid \mathrm{Y}$ in the Bayesian estimation (in practice, the distribution of (X, Y)).

This problem reduces to an optimization one (MLE, MAP) or an integration one (MMSE).

- The LMMSE estimator can be written by means of the second order properties m_{Y}, $C_{\mathrm{Y},\mathrm{Y}}$, m_{X}, $C_{\mathrm{X},\mathrm{X}}$, $C_{\mathrm{X},\mathrm{Y}}$.
- A linear model is an assumption on the conditional mean (assumed linear) and the conditional variance (assumed uniform).

  if $\mathrm{X} \mid \mathrm{Y}$ is driven by a linear model, the LMMSE estimator is the MMSE one.
  If $\mathrm{Y} \mid \mathrm{X}$ is driven by a linear model, we can write the BLUE and re-write the LMMSE estimator.
- The linear assumption is a school case; even if the formula $\mathrm{Y} = H\,\mathrm{X} + \mathrm{W}$ is fulfilled, the variance of W usually must be estimated. But, in general, there is no explicit solution. The optimization problem or the integration one has to be solved numerically.

T155 Chapter 3

Markov property

Chapter contents

T156

Stochastic processes: a short reminder

A **stochastic process** (or **random signal**) x is a function of time and chance.

- In the discrete time case, for all $n \in \mathbb{Z}$, $\mathrm{x}[n]$ is a r.v., a stochastic process is also called a **time series**
- In the continuous time case, for all $t \in \mathbb{R}$, $\mathrm{x}(t)$ is a r.v.

Every realization of a stochastic process is called a **trajectory**, or a **sample path**.

T156 Stochastic processes: a short reminder

A **stochastic process** (or **random signal**) x is a function of time and chance.

- In the discrete time case, for all $n \in \mathbb{Z}$, $\mathrm{x}[n]$ is a r.v., a stochastic process is also called a **time series**
- In the continuous time case, for all $t \in \mathbb{R}$, $\mathrm{x}(t)$ is a r.v.

very realization of a stochastic process is called a **trajectory**, or a **sample path**.

➔ The distribution of the random process corresponds to the distribution of all n_{t}-tuple $(\mathrm{x}(t_1), \ldots, \mathrm{x}(t_{n_{\mathrm{t}}}))$, for all number of times n_{t} and for all distinct times $(t_1, \ldots, t_{n_{\mathrm{t}}})$.

T156 Stochastic processes: a short reminder

A **stochastic process** (or **random signal**) x is a function of time and chance.

- In the discrete time case, for all $n \in \mathbb{Z}$, $\mathrm{x}[n]$ is a r.v., a stochastic process is also called a **time series**
- In the continuous time case, for all $t \in \mathbb{R}$, $\mathrm{x}(t)$ is a r.v.

very realization of a stochastic process is called a **trajectory**, or a **sample path**.

➔ The distribution of the random process corresponds to the distribution of all n_{t}-tuple $(\mathrm{x}(t_1), \ldots, \mathrm{x}(t_{n_{\mathrm{t}}}))$, for all number of times n_{t} and for all distinct times $(t_1, \ldots, t_{n_{\mathrm{t}}})$.

➔ A stochastic process is **independent** if the r.v. of all n_{t}-tuple are mutually independent.[1]

Such a process is completely unpredictable, since the knowledge of the trajectory in some times does not bring any information on the signal value at another time.

1. That is $p_{\mathrm{x}(t_1),\ldots,\mathrm{x}(t_{n_{\mathrm{t}}})} = \prod_{n=1}^{n_{\mathrm{t}}} p_{\mathrm{x}(t_n)}$.

T156 Stochastic processes: a short reminder

A **stochastic process** (or **random signal**) x is a function of time and chance.

- In the discrete time case, for all $n \in \mathbb{Z}$, $\mathrm{x}[n]$ is a r.v., a stochastic process is also called a **time series**
- In the continuous time case, for all $t \in \mathbb{R}$, $\mathrm{x}(t)$ is a r.v.

very realization of a stochastic process is called a **trajectory**, or a **sample path**.

- The distribution of the random process corresponds to the distribution of all n_t-tuple $(\mathrm{x}(t_1), \ldots, \mathrm{x}(t_{n_\mathrm{t}}))$, for all number of times n_t and for all distinct times $(t_1, \ldots, t_{n_\mathrm{t}})$.
- A stochastic process is **independent** if the r.v. of all n_t-tuple are mutually independent.[1]

  Such a process is completely unpredictable, since the knowledge of the trajectory in some times does not bring any information on the signal value at another time.
- A process is **independent and identically distributed** (i.i.d.) if it is independent and if the distribution of $\mathrm{x}(t)$ does not depend of t.

1. That is $p_{\mathrm{x}(t_1),\ldots,\mathrm{x}(t_{n_\mathrm{t}})} = \prod_{n=1}^{n_\mathrm{t}} p_{\mathrm{x}(t_n)}$.

T157 A process is **white** if it exists a positive semi-definite matrix Q such that, for all (t_1, t_2):

$$\mathrm{Cov}\left(\mathrm{x}(t_2), \mathrm{x}(t_1)\right) = Q\,\delta(t_2 - t_1)$$

where δ is the Dirac delta function (continuous time case) or the Kronecker delta (discrete time case).

Q is the **power spectral density**, or **power spectrum**.[2]

2. In general, the power spectrum is a function of frequency, constant for a white process.

T157 A process is **white** if it exists a positive semi-definite matrix Q such that, for all (t_1, t_2):

$$\mathrm{Cov}\left(\mathrm{x}(t_2), \mathrm{x}(t_1)\right) = Q\,\delta(t_2 - t_1)$$

where δ is the Dirac delta function (continuous time case) or the Kronecker delta (discrete time case).

Q is the **power spectral density**, or **power spectrum**.[2]

If x is a discrete time signal, its variance is Q.
If x is a continous time signal, its variance is infinite.

2. In general, the power spectrum is a function of frequency, constant for a white process.

T157 A process is **white** if it exists a positive semi-definite matrix Q such that, for all (t_1, t_2):

$$\operatorname{Cov}\left(\mathrm{x}(t_2), \mathrm{x}(t_1)\right) = Q\,\delta(t_2 - t_1)$$

where δ is the Dirac delta function (continuous time case) or the Kronecker delta (discrete time case).

Q is the **power spectral density**, or **power spectrum**.[2]

If x is a discrete time signal, its variance is Q.
If x is a continous time signal, its variance is infinite.

An i.i.d. process is white.

2. In general, the power spectrum is a function of frequency, constant for a white process.

T157 A process is **white** if it exists a positive semi-definite matrix Q such that, for all (t_1, t_2):

$$\mathrm{Cov}\,(\mathrm{x}(t_2), \mathrm{x}(t_1)) = Q\,\delta(t_2 - t_1)$$

where δ is the Dirac delta function (continuous time case) or the Kronecker delta (discrete time case).

Q is the **power spectral density**, or **power spectrum**.[2]

If x is a discrete time signal, its variance is Q.
If x is a continous time signal, its variance is infinite.

- An i.i.d. process is white.
- An independent stochastic process is a particular (and degenerate) Markov chain.

2. In general, the power spectrum is a function of frequency, constant for a white process.

T158 # Markov process

A stochastic process X is a **Markov process** (or a **Markov chain** in the discrete time case) if, given the process at the current time, past and future of the process are independent.

For all increasing sequence of times $(t_{-n_{\textsf{past}}}, \ldots, t_{-1}, t_0, t_1, \ldots, t_{n_{\textsf{future}}})$:

$$\underbrace{\big(\mathrm{X}(t_{-n_{\textsf{past}}}), \ldots, \mathrm{X}(t_{-1})\big)}_{\text{past}} \perp\!\!\!\perp \underbrace{\big(\mathrm{X}(t_1), \ldots, \mathrm{X}(t_{n_{\textsf{future}}})\big)}_{\text{future}} \mid \underbrace{\mathrm{X}(t_0)}_{\text{current}} \tag{3.1}$$

T158 Markov process

A stochastic process X is a **Markov process** (or a **Markov chain** in the discrete time case) if, given the process at the current time, past and future of the process are independent.

For all increasing sequence of times $(t_{-n_{\text{past}}}, \ldots, t_{-1}, t_0, t_1, \ldots, t_{n_{\text{future}}})$:

$$\underbrace{\left(\mathrm{X}(t_{-n_{\text{past}}}), \ldots, \mathrm{X}(t_{-1})\right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\left(\mathrm{X}(t_1), \ldots, \mathrm{X}(t_{n_{\text{future}}})\right)}_{\text{future}} \mid \underbrace{\mathrm{X}(t_0)}_{\text{current}} \tag{3.1}$$

- The r.v. $\mathrm{X}(t)$ (or the value it takes) is called the **state** of the process at time t.
  The current state trapped all the information about the process future contained in the process past.
  The path who led to this state does not matter.

T158 Markov process

A stochastic process X is a **Markov process** (or a **Markov chain** in the discrete time case) if, given the process at the current time, past and future of the process are independent.

For all increasing sequence of times $(t_{-n_{\text{past}}}, \ldots, t_{-1}, t_0, t_1, \ldots, t_{n_{\text{future}}})$:

$$\underbrace{\left(\mathrm{X}(t_{-n_{\text{past}}}), \ldots, \mathrm{X}(t_{-1})\right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\left(\mathrm{X}(t_1), \ldots, \mathrm{X}(t_{n_{\text{future}}})\right)}_{\text{future}} \mid \underbrace{\mathrm{X}(t_0)}_{\text{current}} \tag{3.1}$$

- The r.v. $\mathrm{X}(t)$ (or the value it takes) is called the **state** of the process at time t.
The current state trapped all the information about the process future contained in the process past.
The path who led to this state does not matter.
- It is sufficient that the property (3.1) holds for $n_{\text{future}} = 1$ to hold for any n_{future}; thus, X is Markovian if and only if, for all increasing sequence $(t_{-n_{\text{past}}}, \ldots, t_{-1}, t_0, t_1)$:[P1]

$$\underbrace{\left(\mathrm{X}(t_{-n_{\text{past}}}), \ldots, \mathrm{X}(t_{-1})\right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\mathrm{X}(t_1)}_{\text{future}} \mid \underbrace{\mathrm{X}(t_0)}_{\text{current}} \tag{3.2}$$

T159 Intuitively, a Markov process is completely characterized by:

- the distribution of $\mathrm{X}(t_{\min})$, where $t_{\min}$ is the initial time, from which we start to observe the process;
- the distribution of $\mathrm{X}(t_1) \mid \mathrm{X}(t_0)$ for all t_0, where t_1 is the close future of t_0.

T159 Intuitively, a Markov process is completely characterized by:

- the distribution of $\mathrm{X}(t_{\min})$, where $t_{\min}$ is the initial time, from which we start to observe the process;
- the distribution of $\mathrm{X}(t_1) \mid \mathrm{X}(t_0)$ for all t_0, where t_1 is the close future of t_0.

➲ In the discrete time case, we choose $t_{\min} = 1$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{X}[1]$;
- the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$.

T159 Intuitively, a Markov process is completely characterized by:

- the distribution of $\mathrm{X}(t_{\min})$, where $t_{\min}$ is the initial time, from which we start to observe the process;
- the distribution of $\mathrm{X}(t_1) \mid \mathrm{X}(t_0)$ for all t_0, where t_1 is the close future of t_0.

➲ In the discrete time case, we choose $t_{\min} = 1$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{X}[1]$;
- the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$.

➲ In the continuous time case,[3] we choose $t_{\min} = 0$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{X}(0)$;
- the **transition distribution**, that is the distribution of $\mathrm{X}(t+h) \mid \mathrm{X}(t)$ when $h \to 0$, for all $t \geqslant 0$.

3. If the state sample paths are continuous (in the usual mathematical meaning), the process is a **diffusion process**; if the state sample paths are piecewise constant, the process is a **jump process**.

T159 Intuitively, a Markov process is completely characterized by:

- the distribution of $\mathrm{x}(t_{\min})$, where $t_{\min}$ is the initial time, from which we start to observe the process;
- the distribution of $\mathrm{x}(t_1) \mid \mathrm{x}(t_0)$ for all t_0, where t_1 is the close future of t_0.

➲ In the discrete time case, we choose $t_{\min} = 1$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{x}[1]$;
- the **transition distribution**, that is the distribution of $\mathrm{x}[n+1] \mid \mathrm{x}[n]$, for all $n \geqslant 1$.

➲ In the continuous time case,[3] we choose $t_{\min} = 0$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{x}(0)$;
- the **transition distribution**, that is the distribution of $\mathrm{x}(t+h) \mid \mathrm{x}(t)$ when $h \to 0$, for all $t \geqslant 0$.

➲ For a continuous valued state, this can be interpreted as the expression of the PDF of the derivative $\dot{\mathrm{x}}(t)$. This interpretation will be used in this book, but we should have in mind that a strong extension of the notions of derivative and integration (for example, the Itô calculus) is needed for a satisfactory mathematical theory.

3. If the state sample paths are continuous (in the usual mathematical meaning), the process is a **diffusion process**; if the state sample paths are piecewise constant, the process is a **jump process**.

T159 Intuitively, a Markov process is completely characterized by:

- the distribution of $\mathrm{X}(t_{\min})$, where $t_{\min}$ is the initial time, from which we start to observe the process;
- the distribution of $\mathrm{X}(t_1) \mid \mathrm{X}(t_0)$ for all t_0, where t_1 is the close future of t_0.

➲ In the discrete time case, we choose $t_{\min} = 1$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{X}[1]$;
- the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$.

➲ In the continuous time case,[3] we choose $t_{\min} = 0$. The process is described by:

- the **initial distribution**, that is the distribution of $\mathrm{X}(0)$;
- the **transition distribution**, that is the distribution of $\mathrm{X}(t+h) \mid \mathrm{X}(t)$ when $h \to 0$, for all $t \geqslant 0$.

➲ For a continuous valued state, this can be interpreted as the expression of the PDF of the derivative $\dot{\mathrm{X}}(t)$. This interpretation will be used in this book, but we should have in mind that a strong extension of the notions of derivative and integration (for example, the Itô calculus) is needed for a satisfactory mathematical theory.

➲ If the transition distribution does not depend of n or t, the process is **homogeneous**.

3. If the state sample paths are continuous (in the usual mathematical meaning), the process is a **diffusion process**; if the state sample paths are piecewise constant, the process is a **jump process**.

T160

Examples

Random walk (discrete time, continuous or discrete state)

This model can be used, for example, for the search of lost bodies in the sea.

There exist an i.i.d. random sequence $\mathrm{V} = (\mathrm{V}[n])_{n \geqslant 1}$, independent of $\mathrm{X}[1]$, such that, for all $n \geqslant 1$:

$$\mathrm{X}[n+1] = \mathrm{X}[n] + \mathrm{V}[n] \tag{3.3}$$

V is called the **increments** sequence.

T160 # Examples

Random walk (discrete time, continuous or discrete state)

This model can be used, for example, for the search of lost bodies in the sea.

There exist an i.i.d. random sequence $\mathrm{V} = (\mathrm{V}[n])_{n \geqslant 1}$, independent of $\mathrm{X}[1]$, such that, for all $n \geqslant 1$:

$$\mathrm{X}[n+1] = \mathrm{X}[n] + \mathrm{V}[n] \tag{3.3}$$

V is called the **increments** sequence.

- X is a Markov chain for which the transition distribution is, for all $n \geqslant 1$, and for all (x, x^+):[P2]

$$p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+, x) = p_{\mathrm{V}[n]}(x^+ - x) \tag{3.4}$$

T161 A random walk corresponds to the cumulative sum of an independent sequence; for all $n \geqslant 1$:

$$\mathrm{x}[n] = \mathrm{x}[1] + \sum_{k=1}^{n-1} \mathrm{v}[k]$$

T161 A random walk corresponds to the cumulative sum of an independent sequence; for all $n \geqslant 1$:

$$\mathrm{x}[n] = \mathrm{x}[1] + \sum_{k=1}^{n-1} \mathrm{v}[k]$$

➔ With the complementary assumptions that:

- the initial state $\mathrm{x}[1]$ is deterministic and zero-valued,
- v is zero-mean, with variance Q,

then, for all (n, n'):

$$\mathrm{E}\left(\mathrm{x}[n]\right) = 0$$
$$\mathrm{Var}\left(\mathrm{x}[n]\right) = Q\,(n-1)$$
$$\mathrm{Cov}\left(\mathrm{x}[n], \mathrm{x}[n']\right) = Q\,\min(n-1, n'-1)$$

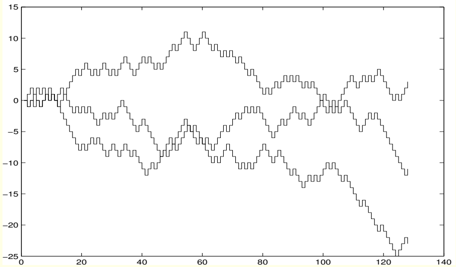


Figure 3.1: Three sample paths of a random walk

T162 The figure 3.1 represents three trajectories of a random walk whose jumps take the value ± 1 with equiprobability.

We can extend to the case where the initial state is not zero and where the jumps are not zero-mean, as in the **gambler's ruin** problem.

T163 **Wiener process (continuous time, continuous state)**

It is the continuous time version of the random walk.

It is also called **Brownian motion**, since the botanist Robert Brown described such phenomenons in the motion of particles in water.
It is used in mathematical finance.

T163 **Wiener process (continuous time, continuous state)**

It is the continuous time version of the random walk.

It is also called **Brownian motion**, since the botanist Robert Brown described such phenomenons in the motion of particles in water.
It is used in mathematical finance.

➔ There exist an i.i.d. random signal $\mathrm{v} : t \geqslant 0 \mapsto \mathrm{v}(t)$, independent of $\mathrm{x}(0)$, such that, for all $t \geqslant 0$:

$$\dot{\mathrm{x}}(t) = \mathrm{v}(t) \tag{3.5}$$

Thus, for all $t \geqslant 0$:

$$\mathrm{x}(t) = \mathrm{x}(0) + \int_0^t \mathrm{v}(\tau)\, \mathrm{d}\,\tau$$

T164 With the complementary assumptions that:

- the initial state $\mathrm{x}(0)$ is deterministic and zero-valued,
- v is zero-mean, with power spectral density Q,

then, for all (t, t'):[4]

$$\mathrm{E}\left(\mathrm{x}(t)\right) = 0$$
$$\mathrm{Var}\left(\mathrm{x}(t)\right) = Q\, t$$
$$\mathrm{Cov}\left(\mathrm{x}(t), \mathrm{x}(t')\right) = Q\, \min(t, t')$$

4. Let's define $\rho(x,t) = p_{\mathrm{x}(t)}(x)$. assumed Gaussian, $\rho(x,t) = \frac{1}{\sqrt{2\pi\, t\, Q}}\, \mathrm{e}^{-\frac{1}{2\,t}\, x^{\mathsf{T}} Q^{-1} x}$, which fulfills $\frac{\partial \rho}{\partial t} = \frac{1}{2}\, \mathrm{trace}\, Q\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$; if $Q = q\, I_{d_x}$ we obtain the diffusion equation $\frac{\partial \rho}{\partial t} = \frac{q}{2}\, \mathrm{trace}\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$ (where $\mathrm{trace}\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$)this trace is the Laplacian of ρ.

T164 With the complementary assumptions that:

- the initial state $\mathrm{x}(0)$ is deterministic and zero-valued,
- v is zero-mean, with power spectral density Q,

then, for all (t, t'):[4]

$$\mathrm{E}\left(\mathrm{x}(t)\right) = 0$$
$$\mathrm{Var}\left(\mathrm{x}(t)\right) = Q\, t$$
$$\mathrm{Cov}\left(\mathrm{x}(t), \mathrm{x}(t')\right) = Q\, \min(t, t')$$

Strictly speaking, this formula, together with the Gaussian assumption, should be used to define the Wiener process.
Thus, we can show that the sample paths are continuous, but they are not differentiable, at every time.
The definition $\dot{\mathrm{x}} = \mathrm{v}$ is easy to use, but implies a strong extension of the notion of derivative.

4. Let's define $\rho(x,t) = p_{\mathrm{x}(t)}(x)$. assumed Gaussian, $\rho(x,t) = \frac{1}{\sqrt{2\pi\, t\, Q}}\, \mathrm{e}^{-\frac{1}{2\,t}\, x^{\mathsf{T}} Q^{-1} x}$, which fulfills $\frac{\partial \rho}{\partial t} = \frac{1}{2}\, \mathrm{trace}\, Q\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$; if $Q = q\, I_{d_x}$ we obtain the diffusion equation $\frac{\partial \rho}{\partial t} = \frac{q}{2}\, \mathrm{trace}\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$ (where $\mathrm{trace}\, \frac{\partial^2 \rho}{\partial x \partial x^{\mathsf{T}}}$)this trace is the Laplacian of ρ.

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\operatorname{Prob}\Big(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\Big) = \begin{cases} 0 & \text{if } x^+ < x & \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 & \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 & \text{(}\lambda\text{ is the process intensity)} \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\operatorname{Prob}\left(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x & \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 & \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 & (\lambda \text{ is the process intensity}) \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\text{Prob}\left(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x & \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 & \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 & \text{(}\lambda\text{ is the process intensity)} \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\mathrm{Prob}\left(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x & \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 & \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 & \text{(}\lambda\text{ is the process intensity)} \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\mathrm{Prob}\left(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x \quad \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 \quad \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 \quad (\lambda \text{ is the process } \textbf{intensity}) \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

T165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

➔ The process takes its value in $\mathbb{N}$. The sample paths are initially zero valued, and increasing.
The transition PMF is, for $h > 0$ small enough, for all $t \geqslant 0$, for all (x^+, x):

$$\operatorname{Prob}\left(\mathrm{x}(t+h) = x^+ \mid \mathrm{x}(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x & \text{(the process is increasing)} \\ 0 & \text{if } x^+ > x+1 & \text{(no simultaneous arrivals))} \\ \lambda\, h & \text{if } x^+ = x+1 & (\lambda \text{ is the process } \textbf{intensity}) \\ 1 - \lambda\, h & \text{if } x^+ = x \end{cases} \tag{3.6}$$

➔ With these natural assumptions, $\mathrm{x}(t)$ is Poisson-distributed; for all $x \in \mathbb{N}$:

$$\operatorname{Prob}\left(\mathrm{x}(t) = x\right) = \exp(-\lambda\, t)\, \frac{(\lambda\, t)^x}{x!} \tag{3.7}$$

T166 ◁ ▷ **Exercise 33** (The assumption (3.6) implies (3.7)). Let X be a Poisson process.

Reminder: the solution of the differential equation $\dot{g}(t) = -\lambda\, g(t) + f(t)$ is $g(t) = \mathrm{e}^{-\lambda\, t}\, g(0) + \int_0^t \mathrm{e}^{-\lambda\,(t-\tau)}\, f(\tau)\,\mathrm{d}\,\tau$.

a) By means of the total probability law, write $p_{\mathrm{X}(t+h)}$ in function of $p_{\mathrm{X}(t)}$.

b) Let's note $g_x(t) = p_{\mathrm{X}(t)}(x)$. Write the differential equation which drives g_x in which g_{x-1} appears (take care of the particular case $x = 0$, by comparison with the general case $x > 0$).

c) Use a recursion to prove the formula (3.7).

T167 **DNA (discrete "time", discrete state)** A DNA strand is a sequence of 4 types of nucleotides: adenine, cytosine, guanine, thymine (A, C, G, T).

The sequential examination of a DNA strand can be considered as a trajectory of a Markov chain for which the state takes its value in the the **finite state space** $\{A, C, G, T\}$.

T167 **DNA (discrete "time", discrete state)** A DNA strand is a sequence of 4 types of nucleotides: adenine, cytosine, guanine, thymine (A, C, G, T).

The sequential examination of a DNA strand can be considered as a trajectory of a Markov chain for which the state takes its value in the the **finite state space** $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

- The transition distribution corresponds to $\mathrm{Prob}\left(\mathrm{x}[n+1] = x^+ \mid |\mathrm{x}[n] = x\right)$ for all x and x^+ in the state space, represented, by the **transition matrix** (4×4 in this DNA model).

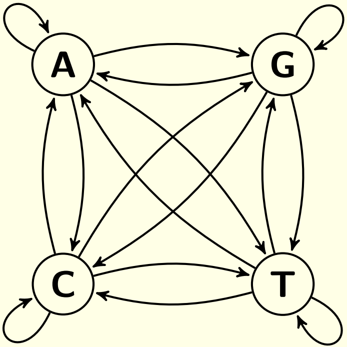


Figure 3.2: State graph of a finite state space Markov chain [ref]

T168 It can be represented by a state graph (figure 3.2) in which each arrow should be labelled with a probability (or suppressed if the probability is 0) [ref].

T169 Nevertheless, the homogeneity assumption of this Markov chain modeling the DNA strand is erroneous, since there are some pieces of this brand in which the dinucleotide CG is over-represented: the CpG islands. Another model with an homogeneous Markov chain uses the state space $\{\mathsf{A}_+, \mathsf{C}_+, \mathsf{G}_+, \mathsf{T}_+, \mathsf{A}_-, \mathsf{C}_-, \mathsf{G}_-, \mathsf{T}_-\}$ in which the $+$ index indicates that the nucleotide is inside a CpG island, and the $-$ index indicates that the nucleotide is not in such an island.

T169 Nevertheless, the homogeneity assumption of this Markov chain modeling the DNA strand is erroneous, since there are some pieces of this brand in which the dinucleotide CG is over-represented: the CpG islands. Another model with an homogeneous Markov chain uses the state space $\{A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-\}$ in which the $+$ index indicates that the nucleotide is inside a CpG island, and the $-$ index indicates that the nucleotide is not in such an island.

- The Markov chain takes its value in a 8 elements state space. Looking at a nucleotide along a DNA strand, we do not know if it belongs to a CpG island or not. The Markov chain is said to be "hidden".

T170 Hidden Markov models (discrete time case)

An **Hidden Markov Model** (HMM) corresponds to 2 stochastic processes:

- the state process $(\mathrm{X}[n])_{n\geqslant 1}$ which, in general, cannot be observed,
- the observation process $(\mathrm{Y}[n])_{n\geqslant 1}$, in practice the output of some sensors,

T170 Hidden Markov models (discrete time case)

An **Hidden Markov Model** (HMM) corresponds to 2 stochastic processes:

- the state process $(\mathrm{X}[n])_{n\geqslant 1}$ which, in general, cannot be observed,
- the observation process $(\mathrm{Y}[n])_{n\geqslant 1}$, in practice the output of some sensors,

➔ Given the present state, past and future are independent with the following convention:

$$\underbrace{\left(\begin{bmatrix}\mathrm{X}[1]\\ \mathrm{Y}[1]\end{bmatrix},\ldots,\begin{bmatrix}\mathrm{X}[n-1]\\ \mathrm{Y}[n-1]\end{bmatrix}\right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\left(\mathrm{Y}[n],\begin{bmatrix}\mathrm{X}[n+1]\\ \mathrm{Y}[n+1]\end{bmatrix},\ldots\right)}_{\text{future}} \mid \mathrm{X}[n] \tag{3.8}$$

T170 Hidden Markov models (discrete time case)

An **Hidden Markov Model** (HMM) corresponds to 2 stochastic processes:

- the state process $(\mathrm{X}[n])_{n\geqslant 1}$ which, in general, cannot be observed,
- the observation process $(\mathrm{Y}[n])_{n\geqslant 1}$, in practice the output of some sensors,

➔ Given the present state, past and future are independent with the following convention:

$$\underbrace{\left(\begin{bmatrix}\mathrm{X}[1]\\ \mathrm{Y}[1]\end{bmatrix},\dots,\begin{bmatrix}\mathrm{X}[n-1]\\ \mathrm{Y}[n-1]\end{bmatrix}\right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\left(\mathrm{Y}[n],\begin{bmatrix}\mathrm{X}[n+1]\\ \mathrm{Y}[n+1]\end{bmatrix},\dots\right)}_{\text{future}} \mid \mathrm{X}[n] \tag{3.8}$$

➔ It is necessary and sufficent to have, for all $n \geqslant 1$:

$$\underbrace{\left(\begin{bmatrix}\mathrm{X}[1]\\ \mathrm{Y}[1]\end{bmatrix},\dots,\begin{bmatrix}\mathrm{X}[n-1]\\ \mathrm{Y}[n-1]\end{bmatrix}\right)}_{\text{past}} \perp\!\!\!\perp (\mathrm{Y}[n],\mathrm{X}[n+1]) \mid \mathrm{X}[n] \tag{3.9}$$

T171 We can show that the state process is Markovian.[5]

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
- the state process is a Markov chain;
- given the state sequence, the observation sequence is independent;
- the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T171 We can show that the state process is Markovian.[5]

- In general, the future state and the present observation are assumed independent; for all $n \geqslant 1$:

$$\mathrm{X}[n+1] \perp\!\!\!\perp \mathrm{Y}[n] \mid \mathrm{X}[n] \tag{3.10}$$

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
   - the state process is a Markov chain;
   - given the state sequence, the observation sequence is independent;
   - the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T171 We can show that the state process is Markovian.[5]

- In general, the future state and the present observation are assumed independent; for all $n \geqslant 1$:

$$\mathrm{X}[n+1] \perp\!\!\!\perp \mathrm{Y}[n] \mid \mathrm{X}[n] \tag{3.10}$$

- Thus, the HMM is characterized by the following distributions:
  - the **initial distribution**, that is the distribution of the initial state $\mathrm{X}[1]$;
  - the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$;
  - the **emission distribution**, that is the distribution of $\mathrm{Y}[n] \mid \mathrm{X}[n]$, for all $n \geqslant 1$ (also known as the **local likelihood**).

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
   - the state process is a Markov chain;
   - given the state sequence, the observation sequence is independent;
   - the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T171 We can show that the state process is Markovian.[5]

- In general, the future state and the present observation are assumed independent; for all $n \geqslant 1$:

$$\mathrm{X}[n+1] \perp\!\!\!\perp \mathrm{Y}[n] \mid \mathrm{X}[n] \tag{3.10}$$

- Thus, the HMM is characterized by the following distributions:
  - the **initial distribution**, that is the distribution of the initial state $\mathrm{X}[1]$;
  - the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$;
  - the **emission distribution**, that is the distribution of $\mathrm{Y}[n] \mid \mathrm{X}[n]$, for all $n \geqslant 1$ (also known as the **local likelihood**).

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
   - the state process is a Markov chain;
   - given the state sequence, the observation sequence is independent;
   - the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T171 We can show that the state process is Markovian.[5]

- In general, the future state and the present observation are assumed independent; for all $n \geqslant 1$:

$$\mathrm{X}[n+1] \perp\!\!\!\perp \mathrm{Y}[n] \mid \mathrm{X}[n] \tag{3.10}$$

- Thus, the HMM is characterized by the following distributions:
  - the **initial distribution**, that is the distribution of the initial state $\mathrm{X}[1]$;
  - the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$;
  - the **emission distribution**, that is the distribution of $\mathrm{Y}[n] \mid \mathrm{X}[n]$, for all $n \geqslant 1$ (also known as the **local likelihood**).

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
   - the state process is a Markov chain;
   - given the state sequence, the observation sequence is independent;
   - the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T171 We can show that the state process is Markovian.[5]

- In general, the future state and the present observation are assumed independent; for all $n \geqslant 1$:

$$\mathrm{X}[n+1] \perp\!\!\!\perp \mathrm{Y}[n] \mid \mathrm{X}[n] \tag{3.10}$$

- Thus, the HMM is characterized by the following distributions:
  - the **initial distribution**, that is the distribution of the initial state $\mathrm{X}[1]$;
  - the **transition distribution**, that is the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n]$, for all $n \geqslant 1$;
  - the **emission distribution**, that is the distribution of $\mathrm{Y}[n] \mid \mathrm{X}[n]$, for all $n \geqslant 1$ (also known as the **local likelihood**).

5. The condition (3.9) is equivalent to the 3 hypotheses below [ref]:
   - the state process is a Markov chain;
   - given the state sequence, the observation sequence is independent;
   - the observation $\mathrm{Y}[n]$ at time n, given the state sequence, depend only of $(\mathrm{X}[n], \mathrm{X}[n+1])$.

T172 A path of the state process and of the observation process is generated with the induction on the figure 3.3.

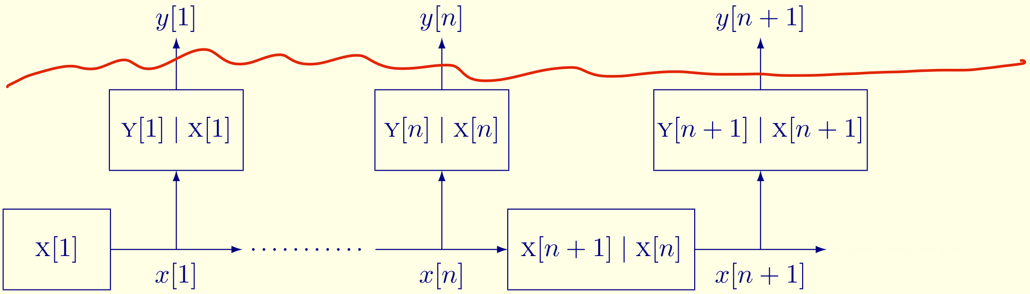


Figure 3.3: Hidden Markov Model

T173 The trajectography problems, that is the location of a moving target, are some typical examples of HMM. The state contains the coordinates of the target. The observation contains some goniometric measures. The transition distribution corresponds to the assumption on the target trajectory (for example, it almost follows a straight line). The emission distribution corresponds to the noise measurement distribution).

T173 The trajectography problems, that is the location of a moving target, are some typical examples of HMM. The state contains the coordinates of the target. The observation contains some goniometric measures. The transition distribution corresponds to the assumption on the target trajectory (for example, it almost follows a straight line). The emission distribution corresponds to the noise measurement distribution).

- The goal is to estimate the state path from the observation path.
  This can be:

A smoothing the future observation values are available to estimate the state at a given time;

A filtering: the current state has to be estimated "on the flight", from past data only, for an **online** implementation.

T173 The trajectography problems, that is the location of a moving target, are some typical examples of HMM. The state contains the coordinates of the target. The observation contains some goniometric measures. The transition distribution corresponds to the assumption on the target trajectory (for example, it almost follows a straight line). The emission distribution corresponds to the noise measurement distribution).

- The goal is to estimate the state path from the observation path.
  This can be:

A smoothing the future observation values are available to estimate the state at a given time;

A filtering: the current state has to be estimated "on the flight", from past data only, for an **online** implementation.

T173 The trajectography problems, that is the location of a moving target, are some typical examples of HMM. The state contains the coordinates of the target. The observation contains some goniometric measures. The transition distribution corresponds to the assumption on the target trajectory (for example, it almost follows a straight line). The emission distribution corresponds to the noise measurement distribution).

- The goal is to estimate the state path from the observation path.
  This can be:

A smoothing the future observation values are available to estimate the state at a given time;

A filtering: the current state has to be estimated "on the flight", from past data only, for an **online** implementation.

T174 Bayesian filtering

We observe $\mathrm{Y}[1:n]$, with growing n.

We observe $\mathrm{Y}[1:n]$, with growing n. The goal is to write the recursion on the probability distributions which will permit to estimate the current state, even to predict the future state and the future observation.

T174

Bayesian filtering

We observe $\mathrm{Y}[1:n]$, with growing n.

We observe $\mathrm{Y}[1:n]$, with growing n. The goal is to write the recursion on the probability distributions which will permit to estimate the current state, even to predict the future state and the future observation.

- For the sake of simplicity, we introduce the exponent $^{|n}$ which means "given the data $\mathrm{Y}[1:n]$".
The Bayesian filtering is a recursion on the PDFs below; for all (y, x, x^+):

$$p_{\mathrm{Y}[n]|\mathrm{Y}[1:n-1]}(y) \quad \text{noted} \quad p^{|n-1}_{\mathrm{Y}[n]}(y)$$
$$p_{\mathrm{X}[n]|\mathrm{Y}[1:n]}(x) \quad \text{noted} \quad p^{|n}_{\mathrm{X}[n]}(x)$$
$$p_{\mathrm{X}[n+1]|\mathrm{Y}[1:n]}(x^+) \quad \text{noted} \quad p^{|n}_{\mathrm{X}[n+1]}(x^+)$$

T175 The recursion is started with the PDF $p^{|0}_{\mathrm{x}[1]}$, simply the prior PDF of the initial state; for all x^+:[6]

$$\text{prior PDF of the first state} \qquad p^{|0}_{\mathrm{x}[1]}(x^+) = p_{\mathrm{x}[1]}(x^+) \tag{3.11}$$

6. In this book, the observation starts at time $n = 1$; thus, the exponent $^{|0}$ actually means "given nothing".

T175 The recursion is started with the PDF $p^{|0}_{\mathrm{X}[1]}$, simply the prior PDF of the initial state; for all x^+:[6]

$$\text{prior PDF of the first state} \qquad p^{|0}_{\mathrm{X}[1]}(x^+) = p_{\mathrm{X}[1]}(x^+) \tag{3.11}$$

- Then, for $n \geqslant 1$, we propagate the PDFs by means of the total probability law and the Bayes law; for all (y, x, x^+):

$$\text{prior PDF of the } n\text{th observation} \qquad p^{|n-1}_{\mathrm{Y}[n]}(y) = \int p_{\mathrm{Y}[n]|\mathrm{X}[n]}(y, x)\, p^{|n-1}_{\mathrm{X}[n]}(x)\,\mathrm{d}\,x \tag{3.12}$$

$$n\text{th observation} \qquad \mathrm{Y}[n] \leftarrow \text{sensors}$$

$$\text{posterior PDF of the } n\text{th state} \qquad \boxed{p^{|n}_{\mathrm{X}[n]}(x)} = \frac{p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], x)\, p^{|n-1}_{\mathrm{X}[n]}(x)}{p^{|n-1}_{\mathrm{Y}[n]}(\mathrm{Y}[n])} \tag{3.13}$$

$$\text{prior PDF of the } (n+1)\text{th state} \qquad p^{|n}_{\mathrm{X}[n+1]}(x^+) = \int p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+, x)\, p^{|n}_{\mathrm{X}[n]}(x)\,\mathrm{d}\,x \tag{3.14}$$

6. In this book, the observation starts at time $n = 1$; thus, the exponent $^{|0}$ actually means “given nothing”.

T175 The recursion is started with the PDF $p^{|0}_{\mathrm{X}[1]}$, simply the prior PDF of the initial state; for all x^+:[6]

$$\text{prior PDF of the first state} \qquad p^{|0}_{\mathrm{X}[1]}(x^+) = p_{\mathrm{X}[1]}(x^+) \tag{3.11}$$

- Then, for $n \geqslant 1$, we propagate the PDFs by means of the total probability law and the Bayes law; for all (y, x, x^+):

$$\text{prior PDF of the } n\text{th observation} \qquad p^{|n-1}_{\mathrm{Y}[n]}(y) = \int p^{\cancel{|n}}_{\mathrm{Y}[n]|\mathrm{X}[n]}(y, x)\, p^{|n-1}_{\mathrm{X}[n]}(x)\, \mathrm{d}\, x \tag{3.12}$$

$$n\text{th observation} \qquad \mathrm{Y}[n] \leftarrow \text{sensors}$$

$$\text{posterior PDF of the } n\text{th state} \qquad p^{|n}_{\mathrm{X}[n]}(x) = \frac{p^{\cancel{|n}}_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], x)\, p^{|n-1}_{\mathrm{X}[n]}(x)}{p^{|n-1}_{\mathrm{Y}[n]}(\mathrm{Y}[n])} \tag{3.13}$$

$$\text{prior PDF of the } (n+1)\text{th state} \qquad p^{|n}_{\mathrm{X}[n+1]}(x^+) = \int p^{\cancel{|n}}_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+, x)\, p^{|n}_{\mathrm{X}[n]}(x)\, \mathrm{d}\, x \tag{3.14}$$

- The formula (3.13) was obtained thanks to the notational trick $p^{|n-1}_{\mathrm{X}[n]|\mathrm{Y}[n]}(x, \mathrm{Y}[n]) = p^{|n}_{\mathrm{X}[n]}(x)$.

6. In this book, the observation starts at time $n = 1$; thus, the exponent $^{|0}$ actually means "given nothing".

T176 From these PDFs, we can deduce Bayesian estimates, for example the MMSE estimations below, together with their (co)variances given the data:[7]

nth observation prediction
$$\hat{\mathrm{Y}}^{|n-1}[n] = \mathsf{E}^{|n-1}\left(\mathrm{Y}[n]\right)$$
$$C_{\mathrm{Y},\mathrm{Y}}[n] = \mathrm{Var}^{|n-1}\left(\mathrm{Y}[n]\right)$$
$$C_{\mathrm{X},\mathrm{Y}}[n] = \mathrm{Cov}^{|n-1}\left(\mathrm{X}[n], \mathrm{Y}[n]\right)$$

nth state estimation
$$\boxed{\begin{aligned}\hat{\mathrm{X}}^{|n}[n] &= \mathsf{E}^{|n}\left(\mathrm{X}[n]\right)\\ P^{|n}[n] &= \mathrm{Var}^{|n}\left(\mathrm{X}[n]\right)\end{aligned}}$$

$(n+1)$th state prediction
$$\hat{\mathrm{X}}^{|n}[n+1] = \mathsf{E}^{|n}\left(\mathrm{X}[n+1]\right)$$
$$P^{|n}[n+1] = \mathrm{Var}^{|n}\left(\mathrm{X}[n+1]\right)$$

7. For example, $\mathsf{E}^{|n}\left(\mathrm{X}[n]\right) = \int x\, p^{|n}_{\mathrm{X}[n]}(x)\, \mathrm{d}\, x$.

T176 From these PDFs, we can deduce Bayesian estimates, for example the MMSE estimations below, together with their (co)variances given the data:[7]

$$
\begin{aligned}
&n\text{th observation prediction} & \hat{\mathrm{Y}}^{|n-1}[n] &= \mathsf{E}^{|n-1}\left(\mathrm{Y}[n]\right)\\
& & C_{\mathrm{Y},\mathrm{Y}}[n] &= \mathrm{Var}^{|n-1}\left(\mathrm{Y}[n]\right)\\
& & C_{\mathrm{X},\mathrm{Y}}[n] &= \mathrm{Cov}^{|n-1}\left(\mathrm{X}[n],\mathrm{Y}[n]\right)\\
&n\text{th state estimation} & \hat{\mathrm{X}}^{|n}[n] &= \mathsf{E}^{|n}\left(\mathrm{X}[n]\right)\\
& & P^{|n}[n] &= \mathrm{Var}^{|n}\left(\mathrm{X}[n]\right)\\
&(n+1)\text{th state prediction} & \hat{\mathrm{X}}^{|n}[n+1] &= \mathsf{E}^{|n}\left(\mathrm{X}[n+1]\right)\\
& & P^{|n}[n+1] &= \mathrm{Var}^{|n}\left(\mathrm{X}[n+1]\right)
\end{aligned}
$$

- In the linear Gaussian case, we obtain the **Kalman filter**, which propagates directly the quantities above, these quantities being necessary and sufficient to represent the Gaussian distributions.

7. For example, $\mathsf{E}^{|n}\left(\mathrm{X}[n]\right) = \int x\, p^{|n}_{\mathrm{X}[n]}(x)\, \mathrm{d}\,x$.

T177 Linear model and Kalman filtering

The linear model is written, for all $n \geqslant 1$, as:[8]

$$\begin{cases} \mathrm{Y}[n] & = H_n\, \mathrm{X}[n] + h_n + \mathrm{W}[n] \\ \mathrm{X}[n+1] & = F_n\, \mathrm{X}[n] + f_n + \mathrm{V}[n] \end{cases} \tag{3.15}$$

under the assumptions:

- the matrix and vector series $(F_n)_{n\geqslant 1}$, $(f_n)_{n\geqslant 1}$, $(H_n)_{n\geqslant 1}$, $(h_n)_{n\geqslant 1}$ are known;[9]
- the series $(\mathrm{V}[n])_{n\geqslant 1}$ is zero-mean, independent, with known variance Q_n;
- the series $(\mathrm{W}[n])_{n\geqslant 1}$ is zero-mean, independent, with known variance R_n;
- the mean and the variance of the initial state $\mathrm{X}[1]$ are known;
- the series $(\mathrm{V}[n])_{n\geqslant 1}$, $(\mathrm{W}[n])_{n\geqslant 1}$, and the initial state $\mathrm{X}[1]$ are mutually independent.

8. This is an HMM whose PDFs are:

$$\begin{aligned} & p_{\mathrm{X}[1]} && \text{with mean } m_{\mathrm{X}[1]} \text{ and variance } C_{\mathrm{X}[1],\mathrm{X}[1]} \\ & p_{\mathrm{Y}[n]|\mathrm{X}[n]}(y,x) = p_{\mathrm{W}[n]}(y - H_n\, x - h_n) && \text{where } \mathrm{W}[n] \text{ is zero-mean with variance } R_n \\ & p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+,x) = p_{\mathrm{V}[n]}(x^+ - F_n\, x - f_n) && \text{where } \mathrm{V}[n] \text{ is zero-mean with variance } Q_n \end{aligned}$$

9. They can be independent of time n. The index n is introduced for the sake of the generality.

T178 If the distributions of $\mathrm{x}[1]$, $\mathrm{v}[n]$, $\mathrm{w}[n]$ are Gaussian, then, the distributions involved in the Bayesian filtering (3.12), (3.13) and (3.14) remain Gaussian, the recursion is written in function of their mean and their variance. The means provide the MMSE estimates.

T178 If the distributions of $\mathrm{x}[1]$, $\mathrm{v}[n]$, $\mathrm{w}[n]$ are Gaussian, then, the distributions involved in the Bayesian filtering (3.12), (3.13) and (3.14) remain Gaussian, the recursion is written in function of their mean and their variance. The means provide the MMSE estimates.

- The Bayes filter becomes the Kalman filter, which is a recursive implementation of the MMSE applied to the Gaussian linear model.

T178 If the distributions of $\mathrm{x}[1]$, $\mathrm{v}[n]$, $\mathrm{w}[n]$ are Gaussian, then, the distributions involved in the Bayesian filtering (3.12), (3.13) and (3.14) remain Gaussian, the recursion is written in function of their mean and their variance. The means provide the MMSE estimates.

- The Bayes filter becomes the Kalman filter, which is a recursive implementation of the MMSE applied to the Gaussian linear model.
- If Gaussian assumption on the sequences v, w or the initial state is removed, we can show that the Kalman filter is nothing but an implementation of the LMMSE estimator.

T178 If the distributions of $\mathrm{X}[1]$, $\mathrm{V}[n]$, $\mathrm{W}[n]$ are Gaussian, then, the distributions involved in the Bayesian filtering (3.12), (3.13) and (3.14) remain Gaussian, the recursion is written in function of their mean and their variance. The means provide the MMSE estimates.

- The Bayes filter becomes the Kalman filter, which is a recursive implementation of the MMSE applied to the Gaussian linear model.
- If Gaussian assumption on the sequences V, W or the initial state is removed, we can show that the Kalman filter is nothing but an implementation of the LMMSE estimator.
- Thus, the variances below (which can be calculated in advance, since they do not depend of the observations) have the meaning of estimation error variance in the Bayesian point of view.

$$C_{\mathrm{Y},\mathrm{Y}}[n] = \mathrm{Var}\left(\hat{\mathrm{Y}}^{|n-1}[n] - \mathrm{Y}[n]\right) \qquad P^{|n}[n] = \mathrm{Var}\left(\hat{\mathrm{X}}^{|n}[n] - \mathrm{X}[n]\right) \qquad P^{|n}[n+1] = \mathrm{Var}\left(\hat{\mathrm{X}}^{|n}[n+1] - \mathrm{X}[n+1]\right.$$

T179 Thus, we recognize in the Kalman filter (algo. 3.1) the formulas of the LMMSE estimator applied to the linear model.

Algorithm 3.1 Kalman filter

Works on the model $H_n, h_n, F_n, f_n, R_n, Q_n, m_{\mathrm{X}[1]}, C_{\mathrm{X}[1],\mathrm{X}[1]}$

with $\begin{cases} \mathrm{Y}[n] &= H_n\,\mathrm{X}[n] + h_n + \mathrm{W}[n] \\ \mathrm{X}[n+1] &= F_n\,\mathrm{X}[n] + f_n + \mathrm{V}[n] \end{cases}$ and $\begin{cases} R_n = C_{\mathrm{W}[n],\mathrm{W}[n]} \\ Q_n = C_{\mathrm{V}[n],\mathrm{V}[n]} \end{cases}$

Initialization

prediction of $\mathrm{X}[1]$ $\downarrow$	$\hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$	provides $\hat{\mathrm{X}}^{\vert 0}[1]$
$\downarrow$	$P \leftarrow C_{\mathrm{X}[1],\mathrm{X}[1]}$	$P^{\vert 0}[1]$

Loop $(n \geqslant 1)$

prediction of $\mathrm{Y}[n]$	$\hat{\mathrm{Y}} \leftarrow H_n\,\hat{\mathrm{X}} + h_n$	$\hat{\mathrm{Y}}^{\vert n-1}[n]$
	$C_{\mathrm{X},\mathrm{Y}} \leftarrow P\,H_n^{\mathsf{T}}$	$C_{\mathrm{X},\mathrm{Y}}[n]$
	$C_{\mathrm{Y},\mathrm{Y}} \leftarrow H_n\,P\,H_n^{\mathsf{T}} + R_n$	$C_{\mathrm{Y},\mathrm{Y}}[n]$
$\rightarrow$ observation of $\mathrm{Y}[n]$	$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
estimation of $\mathrm{X}[n]$	$\hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\,C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - \hat{\mathrm{Y}})$	$\hat{\mathrm{X}}^{\vert n}[n]$ $\rightarrow$
	$P \leftarrow P - C_{\mathrm{X},\mathrm{Y}}\,C_{\mathrm{Y},\mathrm{Y}}^{-1}\,C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}$	$P^{\vert n}[n]$
prediction of $\mathrm{X}[n+1]$ $\downarrow$	$\hat{\mathrm{X}} \leftarrow F_n\,\hat{\mathrm{X}} + f_n$	$\hat{\mathrm{X}}^{\vert n}[n+1]$
$\downarrow$	$P \leftarrow F_n\,P\,F_n^{\mathsf{T}} + Q_n$	$P^{\vert n}[n+1]$

T180 The fundamental equation is the state estimation one, which corrects the prediction by means:

- of the **Kalman gain** $K[n] = C_{\mathrm{X},\mathrm{Y}}[n]\, C^{-1}_{\mathrm{Y},\mathrm{Y}}[n]$;
- of the **innovation** process $(\mathrm{Y}[n] - \hat{\mathrm{Y}}^{|n-1}[n])_{n\geqslant 1}$.[10]

$$\hat{\mathrm{X}}^{|n}[n] = \hat{\mathrm{X}}^{|n-1}[n] + K[n]\,(\mathrm{Y}[n] - \hat{\mathrm{Y}}^{|n-1}[n])$$

together with the error variance $P^{|n}[n]$ for which numerous formulations exist.[11]

10. Furthermore, the innovation sequence is uncorrelated.
11. With the Kalman gain $K[n]$, the variance update can also be written as:

$$P^{|n}[n] = (I - K[n]\,H_n)\,P^{|n-1}[n] \qquad \text{(the simplest)}$$

$$P^{|n}[n] = P^{|n-1}[n] - K[n]\,C_{\mathrm{Y},\mathrm{Y}}[n]\,K^{\mathsf{T}}[n]$$

$$P^{|n}[n] = (I - K[n]\,H_n)\,P^{|n-1}[n]\,(I - K[n]\,H_n)^{\mathsf{T}} + K[n]\,R_n\,K^{\mathsf{T}}[n] \qquad \text{(Joseph form, better conditioned [ref])}$$

$$(P^{|n}[n])^{-1} = (P^{|n-1}[n])^{-1} + H_n^{\mathsf{T}}\,R_n^{-1}\,H_n \qquad (\text{then } K[n] = P^{|n}[n]\,H_n^{\mathsf{T}}\,R_n^{-1})$$

T181 If the matrices F_n, H_n, Q_n, R_n are time-independent, the matrix $P^{|n-1}[n]$ tends to the limit $P[\infty]$ which fulfills the **discrete Riccati equation** obtained by grouping the formulas of variances update:

$$P[\infty] = F\,P[\infty]\,F^{\intercal} - F\,P[\infty]\,H^{\intercal}\,(H\,P[\infty]\,H^{\intercal} + R)^{-1}\,H\,P[\infty]\,F^{\intercal} + Q$$

T181 If the matrices F_n, H_n, Q_n, R_n are time-independent, the matrix $P^{|n-1}[n]$ tends to the limit $P[\infty]$ which fulfills the **discrete Riccati equation** obtained by grouping the formulas of variances update:

$$P[\infty] = F\,P[\infty]\,F^{\intercal} - F\,P[\infty]\,H^{\intercal}\,(H\,P[\infty]\,H^{\intercal} + R)^{-1}\,H\,P[\infty]\,F^{\intercal} + Q$$

- We obtains a highly simplified filter, by replacing the time-varying Kalman gain $K[n]$ by its limit $K[\infty]$:

$$K[\infty] = P[\infty]\,H^{\intercal}\,(H\,P[\infty]\,H^{\intercal} + R)^{-1}$$

T182 This filter asymptotically behaves like the Kalman filter, but is slower in the transience after the initialization (algo. 3.2).

Algorithm 3.2 Stationary Kalman filter

Works on the model $H, h_n, F, f_n, R, Q, m_{\mathrm{X}[1]}$

with $\begin{cases} \mathrm{Y}[n] &= H\,\mathrm{X}[n] + h_n + \mathrm{W}[n] \\ \mathrm{X}[n+1] &= F\,\mathrm{X}[n] + f_n + \mathrm{V}[n] \end{cases}$ and $\begin{cases} R = C_{\mathrm{W}[n],\mathrm{W}[n]} \\ Q = C_{\mathrm{V}[n],\mathrm{V}[n]} \end{cases}$

Preliminaries

Solve/P $\quad P = F\,P\,F^{\intercal} - F\,P\,H^{\intercal}\,(H\,P\,H^{\intercal} + R)^{-1}\,H\,P\,F^{\intercal} + Q$provides $P[\infty]$

Calculate $\quad K \leftarrow P\,H^{\intercal}\,(H\,P\,H^{\intercal} + R)^{-1}$ $\quad K[\infty]$

Initialization

prediction of $\mathrm{X}[1] \downarrow$ $\quad \hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$ $\quad \hat{\mathrm{X}}^{|0}[1]$

Loop $(n \geqslant 1)$

prediction of $\mathrm{Y}[n]$ $\quad \hat{\mathrm{Y}} \leftarrow H\,\hat{\mathrm{X}} + h_n$ $\quad \hat{\mathrm{Y}}^{|n-1}[n]$

$\rightarrow$ observation of $\mathrm{Y}[n]$ $\quad \mathrm{Y} \leftarrow$ sensors $\quad \mathrm{Y}[n]$

estimation of $\mathrm{X}[n]$ $\quad \hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + K\,(\mathrm{Y} - \hat{\mathrm{Y}})$ $\quad \hat{\mathrm{X}}^{|n}[n] \rightarrow$

prediction of $\mathrm{X}[n+1] \downarrow$ $\quad \hat{\mathrm{X}} \leftarrow F\,\hat{\mathrm{X}} + f_n$ $\quad \hat{\mathrm{X}}^{|n}[n+1]$

T183 # Adapting the Kalman filter to non-linear models

We suppose that the data generating process is, for all $n \geqslant 1$:

$$\begin{cases} \mathrm{Y}[n] & = h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] & = f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases} \tag{3.16}$$

under the following assumptions:

- the function series $(f_n)_{n \geqslant 1}$ and $(h_n)_{n \geqslant 1}$ are known;
- the series $(\mathrm{V}[n])_{n \geqslant 1}$ is zero-mean, independent, with known variance Q_n;
- the series $(\mathrm{W}[n])_{n \geqslant 1}$ is zero-mean, independent, with known variance R_n;
- the mean and the variance of the initial state $\mathrm{X}[1]$ are known;
- the series $(\mathrm{V}[n])_{n \geqslant 1}$, $(\mathrm{W}[n])_{n \geqslant 1}$, and the initial state $\mathrm{X}[1]$ are mutually independent.

T183

Adapting the Kalman filter to non-linear models

We suppose that the data generating process is, for all $n \geqslant 1$:

$$\begin{cases} \mathrm{Y}[n] & = h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] & = f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases} \tag{3.16}$$

under the following assumptions:

- the function series $(f_n)_{n\geqslant 1}$ and $(h_n)_{n\geqslant 1}$ are known;
- the series $(\mathrm{V}[n])_{n\geqslant 1}$ is zero-mean, independent, with known variance Q_n;
- the series $(\mathrm{W}[n])_{n\geqslant 1}$ is zero-mean, independent, with known variance R_n;
- the mean and the variance of the initial state $\mathrm{X}[1]$ are known;
- the series $(\mathrm{V}[n])_{n\geqslant 1}$, $(\mathrm{W}[n])_{n\geqslant 1}$, and the initial state $\mathrm{X}[1]$ are mutually independent.

➜ We do not derive the distributions involved in this HMM.[12]

We draw inspiration from the Kalman filter to derive an approximation of the Bayesian filter which propagates only some means and (co)variances.

12. The PDFs are $p_{\mathrm{Y}[n]|\mathrm{X}[n]}(y,x) = \int \delta(y - h_n(x,w))\, p_{\mathrm{W}[n]}(w)\,\mathrm{d}\,w$ et $p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+,x) = \int \delta(x^+ - f_n(x,v))\, p_{\mathrm{V}[n]}(v)\,\mathrm{d}\,v$.

T184 We maintain the LMMSE state update:

$$\begin{aligned}
&n\text{th observation prediction} && \hat{\mathrm{Y}}^{|n-1}[n] = \mathsf{E}^{|n-1}\left(h_n(\mathrm{X}[n], \mathrm{W}[n])\right) && (3.17)\\
& && C_{\mathrm{Y},\mathrm{Y}}[n] = \mathrm{Var}^{|n-1}\left(h_n(\mathrm{X}[n], \mathrm{W}[n])\right) &&\\
& && C_{\mathrm{X},\mathrm{Y}}[n] = \mathrm{Cov}^{|n-1}\left(\mathrm{X}[n], h_n(\mathrm{X}[n], \mathrm{W}[n])\right) &&\\
&n\text{th observation} && \mathrm{Y}[n] \leftarrow \text{sensors} &&\\
&n\text{th state estimation} && \hat{\mathrm{X}}^{|n}[n] \simeq \hat{\mathrm{X}}^{|n-1}[n] + C_{\mathrm{X},\mathrm{Y}}[n]\, C_{\mathrm{Y},\mathrm{Y}}^{-1}[n]\left(\mathrm{Y}[n] - \hat{\mathrm{Y}}^{|n-1}[n]\right) && (3.18)\\
& && P^{|n}[n] \simeq P^{|n-1}[n] - C_{\mathrm{X},\mathrm{Y}}[n]\, C_{\mathrm{Y},\mathrm{Y}}^{-1}[n]\, C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}[n] &&\\
&(n+1)\text{th state prediction} && \hat{\mathrm{X}}^{|n}[n+1] = \mathsf{E}^{|n}\left(f_n(\mathrm{X}[n], \mathrm{V}[n])\right) && (3.19)\\
& && P^{|n}[n+1] = \mathrm{Var}^{|n}\left(f_n(\mathrm{X}[n], \mathrm{V}[n])\right) &&
\end{aligned}$$

T184 We maintain the LMMSE state update:

nth observation prediction $\hat{\mathrm{Y}}^{|n-1}[n] = \mathsf{E}^{|n-1}\left(h_n(\mathrm{X}[n], \mathrm{W}[n])\right)$ (3.17)

$$C_{\mathrm{Y},\mathrm{Y}}[n] = \mathrm{Var}^{|n-1}\left(h_n(\mathrm{X}[n], \mathrm{W}[n])\right)$$

$$C_{\mathrm{X},\mathrm{Y}}[n] = \mathrm{Cov}^{|n-1}\left(\mathrm{X}[n], h_n(\mathrm{X}[n], \mathrm{W}[n])\right)$$

nth observation $\mathrm{Y}[n] \leftarrow$ sensors

nth state estimation $\hat{\mathrm{X}}^{|n}[n] \simeq \hat{\mathrm{X}}^{|n-1}[n] + C_{\mathrm{X},\mathrm{Y}}[n]\, C_{\mathrm{Y},\mathrm{Y}}^{-1}[n]\left(\mathrm{Y}[n] - \hat{\mathrm{Y}}^{|n-1}[n]\right)$ (3.18)

$$P^{|n}[n] \simeq P^{|n-1}[n] - C_{\mathrm{X},\mathrm{Y}}[n]\, C_{\mathrm{Y},\mathrm{Y}}^{-1}[n]\, C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}[n]$$

$(n+1)$th state prediction $\hat{\mathrm{X}}^{|n}[n+1] = \mathsf{E}^{|n}\left(f_n(\mathrm{X}[n], \mathrm{V}[n])\right)$ (3.19)

$$P^{|n}[n+1] = \mathrm{Var}^{|n}\left(f_n(\mathrm{X}[n], \mathrm{V}[n])\right)$$

➔ The formulas (3.17) and (3.19) are approximated through uncertainty propagation (page 116).

T185 If we propagate the uncertainty through linearization, we obtain the **Extended Kalman filter** (EKF, algo. 3.3).

Algorithm 3.3 Extended Kalman Filter (EKF)

Works on the model $h_n, f_n, R_n, Q_n, m_{\mathrm{X}[1]}, C_{\mathrm{X}[1],\mathrm{X}[1]}$

$$\text{with } \begin{cases} \mathrm{Y}[n] &= h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] &= f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases} \text{ and } \begin{cases} R_n &= C_{\mathrm{W}[n],\mathrm{W}[n]} \\ Q_n &= C_{\mathrm{V}[n],\mathrm{V}[n]} \end{cases}$$

Initialization

prediction of $\mathrm{X}[1] \downarrow$	$\hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$	provides $\hat{\mathrm{X}}^{\|0}[1]$
$\downarrow$	$P \leftarrow C_{\mathrm{X}[1],\mathrm{X}[1]}$	$P^{\|0}[1]$

Loop $(n \geqslant 1)$

Jacobian matrix of h_n	$H_x \leftarrow \dfrac{\partial h_n}{\partial x^{\intercal}}(\hat{\mathrm{X}}, 0)$ and $H_w \leftarrow \dfrac{\partial h_n}{\partial w^{\intercal}}(\hat{\mathrm{X}}, 0)$	
prediction of $\mathrm{Y}[n]$	$\hat{\mathrm{Y}} \leftarrow h_n(\hat{\mathrm{X}}, 0)$	$\hat{\mathrm{Y}}^{\|n-1}[n]$
	$C_{\mathrm{Y},\mathrm{Y}} \leftarrow H_x\, P\, H_x^{\intercal} + H_w\, R_n\, H_w^{\intercal}$	$C_{\mathrm{Y},\mathrm{Y}}[n]$
	$C_{\mathrm{X},\mathrm{Y}} \leftarrow P\, H_x^{\intercal}$	$C_{\mathrm{X},\mathrm{Y}}[n]$
$\rightarrow$ observation of $\mathrm{Y}[n]$	$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
estimation of $\mathrm{X}[n]$	$\hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, (\mathrm{Y} - \hat{\mathrm{Y}})$	$\hat{\mathrm{X}}^{\|n}[n] \rightarrow$
	$P \leftarrow P - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{X},\mathrm{Y}}^{\intercal}$	$P^{\|n}[n]$
Jacobian matrix of f_n	$F_x \leftarrow \dfrac{\partial f_n}{\partial x^{\intercal}}(\hat{\mathrm{X}}, 0)$ and $F_v \leftarrow \dfrac{\partial f_n}{\partial v^{\intercal}}(\hat{\mathrm{X}}, 0)$	
prediction of $\mathrm{X}[n+1] \downarrow$	$\hat{\mathrm{X}} \leftarrow f_n(\hat{\mathrm{X}}, 0)$	$\hat{\mathrm{X}}^{\|n}[n+1]$
$\downarrow$	$P \leftarrow F_x\, P\, F_x^{\intercal} + F_v\, Q_n\, F_v^{\intercal}$	$P^{\|n}[n+1]$

T186 The case of additive noises is provided in the algo. 3.4.

Algorithm 3.4 EKF, additive noise case

Works on the model $h_n, f_n, R_n, Q_n, m_{X[1]}, C_{X[1],X[1]}$

with $\begin{cases} Y[n] &= h_n(X[n]) + W[n] \\ X[n+1] &= f_n(X[n]) + V[n] \end{cases}$ and $\begin{cases} R_n = C_{W[n],W[n]} \\ Q_n = C_{V[n],V[n]} \end{cases}$

Initialization

prediction of $X[1]$ $\downarrow$	$\hat{X} \leftarrow m_{X[1]}$	provides $\hat{X}^{\|0}[1]$
$\downarrow$	$P \leftarrow C_{X[1],X[1]}$	$P^{\|0}[1]$

Loop $(n \geqslant 1)$

Jacobian matrix of h_n	$H \leftarrow \dfrac{\partial h_n}{\partial x^{\intercal}}(\hat{X})$	
prediction of $Y[n]$	$\hat{Y} \leftarrow h_n(\hat{X})$	$\hat{Y}^{\|n-1}[n]$
	$C_{X,Y} \leftarrow P\,H^{\intercal}$	$C_{X,Y}[n]$
	$C_{Y,Y} \leftarrow H\,P\,H^{\intercal} + R_n$	$C_{Y,Y}[n]$
$\rightarrow$ observation of $Y[n]$	$Y \leftarrow$ sensors	$Y[n]$
estimation of $X[n]$	$\hat{X} \leftarrow \hat{X} + C_{X,Y}\,C_{Y,Y}^{-1}\,(Y - \hat{Y})$	$\hat{X}^{\|n}[n]$ $\rightarrow$
	$P \leftarrow P - C_{X,Y}\,C_{Y,Y}^{-1}\,C_{X,Y}^{\intercal}$	$P^{\|n}[n]$
Jacobian matrix of f_n	$F \leftarrow \dfrac{\partial f_n}{\partial x^{\intercal}}(\hat{X})$	
prediction of $X[n+1]$ $\downarrow$	$\hat{X} \leftarrow f_n(\hat{X})$	$\hat{X}^{\|n}[n+1]$
$\downarrow$	$P \leftarrow F\,P\,F^{\intercal} + Q_n$	$P^{\|n}[n+1]$

T187 If we propagate the uncertainty through the unscented transform UT page 118,[13] we obtain the UKF (algo. 3.5), or the cubature Kalman filter (CKF) it the cubature is used.

13. if $\mathrm{Y} = h(\mathrm{X})$ then $(m_{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}) \simeq \mathrm{UT}\left(h, m_{\mathrm{X}}, C_{\mathrm{X},\mathrm{X}}\right)$

Algorithm 3.5 Unscented Kalman Filter (UKF)

Works on the model $h_n, f_n, R_n, Q_n, m_{\mathrm{X}[1]}, C_{\mathrm{X}[1],\mathrm{X}[1]}$

with $\begin{cases} \mathrm{Y}[n] &= h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] &= f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases}$ and $\begin{cases} R_n = C_{\mathrm{W}[n],\mathrm{W}[n]} \\ Q_n = C_{\mathrm{V}[n],\mathrm{V}[n]} \end{cases}$

Initialization

prediction of $\mathrm{X}[1] \downarrow$	$\hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$	provides $\hat{\mathrm{X}}^{\|0}[1]$
$\downarrow$	$P \leftarrow C_{\mathrm{X}[1],\mathrm{X}[1]}$	$P^{\|0}[1]$

Loop $(n \geqslant 1)$

prediction of $\mathrm{Y}[n]$	$(\hat{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}) \leftarrow \mathrm{UT}\left(h_n, \begin{bmatrix}\hat{\mathrm{X}} \\ 0\end{bmatrix}, \begin{bmatrix}P & 0 \\ 0 & R_n\end{bmatrix}\right)$	$\hat{\mathrm{Y}}^{\|n-1}[n], C_{\mathrm{Y},\mathrm{Y}}[n], C_{\mathrm{X},\mathrm{Y}}[n]$
$\rightarrow$ observation of $\mathrm{Y}[n]$	$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
estimation of $\mathrm{X}[n]$	$\hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, (\mathrm{Y} - \hat{\mathrm{Y}})$	$\hat{\mathrm{X}}^{\|n}[n] \rightarrow$
	$P \leftarrow P - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{X},\mathrm{Y}}^{\intercal}$	$P^{\|n}[n]$
prediction of $\mathrm{X}[n+1] \downarrow$	$(\hat{\mathrm{X}}, P) \leftarrow \mathrm{UT}\left(f_n, \begin{bmatrix}\hat{\mathrm{X}} \\ 0\end{bmatrix}, \begin{bmatrix}P & 0 \\ 0 & Q_n\end{bmatrix}\right)$	$\hat{\mathrm{X}}^{\|n}[n+1], P^{\|n}[n+1]$

T188 The UKF provides naturally some σ-points for $(\hat{\mathrm{X}}^{|n}[n+1], P^{|n}[n+1])$. In the case of an additive observation noise, we save one σ-points calculationas in the original algorithm [ref] (algo. 3.6):

$$\text{if } \mathrm{Y}[n] = h_n(\mathrm{X}[n]) + \mathrm{W}[n] \qquad \text{then} \qquad \mathrm{Var}^{|n-1}\left(h_n(\mathrm{X}[n]) + \mathrm{W}[n]\right) = \mathrm{Var}^{|n-1}\left(h_n(\mathrm{X}[n])\right) + R_n$$

Algorithm 3.6 UKF, additive observation noise case

Works on the model $h_n, f_n, R_n, Q_n, m_{\mathrm{X}[1]}, C_{\mathrm{X}[1],\mathrm{X}[1]}$

with $\begin{cases}\mathrm{Y}[n] &= h_n(\mathrm{X}[n]) + \mathrm{W}[n]\\ \mathrm{X}[n+1] &= f_n(\mathrm{X}[n], \mathrm{V}[n])\end{cases}$ and $\begin{cases}R_n = C_{\mathrm{W}[n],\mathrm{W}[n]}\\ Q_n = C_{\mathrm{V}[n],\mathrm{V}[n]}\end{cases}$

Initialization

prediction of $\mathrm{X}[1]$ $\downarrow$	$\hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$	provides $\hat{\mathrm{X}}^{\|0}[1]$
$\downarrow$	$P \leftarrow C_{\mathrm{X}[1],\mathrm{X}[1]}$	$P^{\|0}[1]$
prediction of $\mathrm{Y}[1]$ $\downarrow$	$(\hat{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}) \leftarrow \mathsf{UT}(h_1, \hat{\mathrm{X}}, P)$	$\hat{\mathrm{Y}}^{\|0}[1], C_{\mathrm{X},\mathrm{Y}}[1]$
$\downarrow$	$C_{\mathrm{Y},\mathrm{Y}} \leftarrow C_{\mathrm{Y},\mathrm{Y}} + R_1$	$C_{\mathrm{Y},\mathrm{Y}}[1]$

Loop $(n \geqslant 1)$

$\rightarrow$ observation of $\mathrm{Y}[n]$	$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
estimation of $\mathrm{X}[n]$	$\hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, (\mathrm{Y} - \hat{\mathrm{Y}})$	$\hat{\mathrm{X}}^{\|n}[n] \rightarrow$
	$P \leftarrow P - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}}$	$P^{\|n}[n]$
pred. $\mathrm{Y}[n+1]$ and $\mathrm{X}[n+1]$ $\downarrow$	$\left(\begin{bmatrix}\hat{\mathrm{X}}\\ \hat{\mathrm{Y}}\end{bmatrix}, \begin{bmatrix}P & C_{\mathrm{X},\mathrm{Y}}\\ C_{\mathrm{X},\mathrm{Y}}^{\mathsf{T}} & C_{\mathrm{Y},\mathrm{Y}}\end{bmatrix}\right) \leftarrow \mathsf{UT}\left(\begin{bmatrix}f_n\\ h_{n+1}\circ f_n\end{bmatrix}, \begin{bmatrix}\hat{\mathrm{X}}\\ 0\end{bmatrix}, \begin{bmatrix}P & 0\\ 0 & Q_n\end{bmatrix}\right)$	
		$\hat{\mathrm{X}}^{\|n}[n+1], \hat{\mathrm{Y}}^{\|n}[n+1], P^{\|n}[n+1], C_{\mathrm{X},\mathrm{Y}}[n+1]$
$\downarrow$	$C_{\mathrm{Y},\mathrm{Y}} \leftarrow C_{\mathrm{Y},\mathrm{Y}} + R_{n+1}$	$C_{\mathrm{Y},\mathrm{Y}}[n+1]$

T189 If the state noise is also additive:

$$\mathrm{X}[n+1] = f_n(\mathrm{X}[n]) + \mathrm{V}[n] \qquad \text{then} \qquad \mathrm{Var}^{|n}\left(f_n(\mathrm{X}[n]) + \mathrm{V}[n]\right) = \mathrm{Var}^{|n}\left(f_n(\mathrm{X}[n])\right) + Q_n$$

we can either use the algorithm 3.6, with one σ-points calculation in dimension $2\,d_x$, or take into account the formula above; this leads to two σ-points calculation in dimension d_x (algo. 3.7).

Algorithm 3.7 UKF, additive noise case

Works on the model $h_n, f_n, R_n, Q_n, m_{\mathrm{X}[1]}, C_{\mathrm{X}[1],\mathrm{X}[1]}$

with $\begin{cases} \mathrm{Y}[n] = h_n(\mathrm{X}[n]) + \mathrm{W}[n] \\ \mathrm{X}[n+1] = f_n(\mathrm{X}[n]) + \mathrm{V}[n] \end{cases}$ and $\begin{cases} R_n = C_{\mathrm{W}[n],\mathrm{W}[n]} \\ Q_n = C_{\mathrm{V}[n],\mathrm{V}[n]} \end{cases}$

Initialization

prediction of $\mathrm{X}[1]$ $\downarrow$	$\hat{\mathrm{X}} \leftarrow m_{\mathrm{X}[1]}$	provides $\hat{\mathrm{X}}^{|0}[1]$
$\downarrow$	$P \leftarrow C_{\mathrm{X}[1],\mathrm{X}[1]}$	$P^{|0}[1]$

Loop $(n \geqslant 1)$

prediction of $\mathrm{Y}[n]$	$(\hat{\mathrm{Y}}, C_{\mathrm{Y},\mathrm{Y}}, C_{\mathrm{X},\mathrm{Y}}) \leftarrow \mathsf{UT}(h_n, \hat{\mathrm{X}}, P)$	$\hat{\mathrm{Y}}^{|n-1}[n], C_{\mathrm{X},\mathrm{Y}}[n]$
	$C_{\mathrm{Y},\mathrm{Y}} \leftarrow C_{\mathrm{Y},\mathrm{Y}} + R_n$	$C_{\mathrm{Y},\mathrm{Y}}[n]$
$\rightarrow$ observation of $\mathrm{Y}[n]$	$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
estimation of $\mathrm{X}[n]$	$\hat{\mathrm{X}} \leftarrow \hat{\mathrm{X}} + C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\,(\mathrm{Y} - \hat{\mathrm{Y}})$	$\hat{\mathrm{X}}^{|n}[n] \rightarrow$
	$P \leftarrow P - C_{\mathrm{X},\mathrm{Y}}\, C_{\mathrm{Y},\mathrm{Y}}^{-1}\, C_{\mathrm{X},\mathrm{Y}}^{\intercal}$	$P^{|n}[n]$
prediction of $\mathrm{X}[n+1]$ $\downarrow$	$(\hat{\mathrm{X}}, P) \leftarrow \mathsf{UT}(f_n, \hat{\mathrm{X}}, P)$	$\hat{\mathrm{X}}^{|n}[n+1]$
$\downarrow$	$P \leftarrow P + Q_n$	$P^{|n}[n+1]$

T190 Both EKF and UKF are some adaptations of the standard Kalman filter to a non-linear model. If they are applied to a linear model, we retrieve exactly the Kalman filter.

In general, the matrices $C_{\mathrm{Y},\mathrm{Y}}[n]$, $P^{|n}[n]$, $P^{|n}[n+1]$ depend of the data and cannot be considered as estimation error variance in a bayesian meaning.

T190 Both EKF and UKF are some adaptations of the standard Kalman filter to a non-linear model. If they are applied to a linear model, we retrieve exactly the Kalman filter.

In general, the matrices $C_{\mathrm{Y},\mathrm{Y}}[n]$, $P^{|n}[n]$, $P^{|n}[n+1]$ depend of the data and cannot be considered as estimation error variance in a bayesian meaning.

- It is hard to give general theoretical results on the performances of such approximation-based algorithms. In practice, the UKF seems to provide less estimation error than the EKF. It does not need to compute the Jacobian matrices. We have to compare the numerical costs of the unscented transform and the Jacobian matrix computation.

T190 Both EKF and UKF are some adaptations of the standard Kalman filter to a non-linear model. If they are applied to a linear model, we retrieve exactly the Kalman filter.

In general, the matrices $C_{\mathrm{Y},\mathrm{Y}}[n]$, $P^{|n}[n]$, $P^{|n}[n+1]$ depend of the data and cannot be considered as estimation error variance in a bayesian meaning.

- It is hard to give general theoretical results on the performances of such approximation-based algorithms. In practice, the UKF seems to provide less estimation error than the EKF. It does not need to compute the Jacobian matrices. We have to compare the numerical costs of the unscented transform and the Jacobian matrix computation.
- If the EKF or the UKF do not provide suitable results, we can use **sequential Monte Carlo** methods such as the particle filter to approximate the Bayes filter.

T191 # Chapter 4

Stochastic simulation

Chapter contents

T192 Random sampling

We consider the universe of the students of the university.
The function which returns for each student his average mark is a r.v. X.
The r.v. $(\mathrm{X}_1, \ldots, \mathrm{X}_{n_\mathrm{r}})$ obtained in a repetition of this operation n_r times in the same experimental conditions is called a n_r-**sample** of the r.v. X.
This n_r-sample is an i.i.d. sequence, which is constituted from n_r independent copies of the r.v. X.
We performed a random **sampling**.

T192 # Random sampling

We consider the universe of the students of the university.
The function which returns for each student his average mark is a r.v. X.
The r.v. $(\mathrm{X}_1, \ldots, \mathrm{X}_{n_\mathrm{r}})$ obtained in a repetition of this operation n_r times in the same experimental conditions is called a n_r-**sample** of the r.v. X.
This n_r-sample is an i.i.d. sequence, which is constituted from n_r independent copies of the r.v. X.
We performed a random **sampling**.

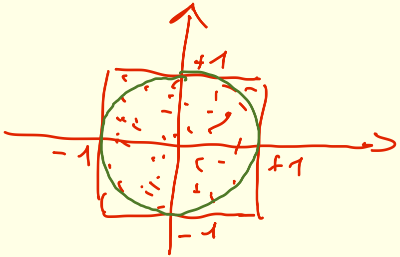
- The **stochastic simulation** consists to generate, by means of available calculation means, a sequence of numbers which can be considered as a realization of a n_r-sample of a r.v.

T193 For example, if we want to predict the weather conditions, we derive an algorithm based on a probabilistic model. If, on actual data, we obtain strange results, is this due to the model, to the prediction algorithm, or to a programming error?

We must test the algorithm on simulated data! If the prediction is correct on simulated data, but not on actual data, the model validity becomes questionable.

T193 For example, if we want to predict the weather conditions, we derive an algorithm based on a probabilistic model. If, on actual data, we obtain strange results, is this due to the model, to the prediction algorithm, or to a programming error?
We must test the algorithm on simulated data! If the prediction is correct on simulated data, but not on actual data, the model validity becomes questionable.

- Furthermore, with **Monte Carlo methods**, we perform a computation thanks to random sampling.
For example, let's try to estimate the value of π; we sample according to a uniform distribution over the square $[-1,1] \times [-1,1]$; the ~~number~~ proportion of realizations in the unit circle is an approximation of $\frac{\pi}{4}$.[1]



1. Matlab. To obtain an approximation of π (with a high n_{r}).
```
4*mean(abs([1 j]*rand(2,nr))<1)
```

T194 Pseudo-random numbers generators

Figure 4.1: White noise generator

It exists some analog devices which simulate the hazard in continuous time, the **generators** of **pseudo-random** numbers in this document are algorithm which create a sequence of numbers such that a bunch of statistical tests assesses that this sequence is a sample realization with a given probability distribution.

T194 # Pseudo-random numbers generators

Figure 4.1: White noise generator

It exists some analog devices which simulate the hazard in continuous time, the **generators** of **pseudo-random** numbers in this document are algorithm which create a sequence of numbers such that a bunch of statistical tests assesses that this sequence is a sample realization with a given probability distribution.

- An actual generator of random numbers does not exist (the prefix "pseudo" is important) since these algorithms are deterministic: we can always generate twice the same sequence, and the obtained sequences are periodic (with a very great period).

T195 For example, a **linear congruential generator** provide a sequence $(u_q)_{q\geqslant 1}$ by means of the recursion below, with s_q integer-valued:

$$u_q = \frac{s_q}{T}$$
$$s_{q+1} = (a\, s_q + b) \mod T$$

It must be initialized by s_1 (the **seed**).
The sequence period is at most T, some values of a and b provide a period T.
If $n_\mathrm{r} \ll T$, $(u_1, \ldots, u_{n_\mathrm{r}})$ may be viewed as a realization of a n_r-sample of a r.v. $\mathtt{U}$ uniformly distributed between 0 and 1.[2]

2. Matlab. To generate a n_r-sample uniformy distributed between 0 and 1 (the seed is processed by the function `rng`).
```
x = rand(1,nr);
```

T195 For example, a **linear congruential generator** provide a sequence $(u_q)_{q\geqslant 1}$ by means of the recursion below, with s_q integer-valued:

$$u_q = \frac{s_q}{T}$$
$$s_{q+1} = (a\, s_q + b) \mod T$$

It must be initialized by s_1 (the **seed**).
The sequence period is at most T, some values of a and b provide a period T.
If $n_\mathrm{r} \ll T$, $(u_1, \ldots, u_{n_\mathrm{r}})$ may be viewed as a realization of a n_r-sample of a r.v. U uniformly distributed between 0 and 1.[2]

- This algorithm behavior highly depends of the choice of a, b and T. An implementation which was widely used in the seventies was later proven to be a poor choice, so that som scientific results became doubtful [ref].

2. Matlab. To generate a n_r-sample uniformy distributed between 0 and 1 (the seed is processed by the function `rng`).
```
x = rand(1,nr);
```

T195 For example, a **linear congruential generator** provide a sequence $(u_q)_{q\geqslant 1}$ by means of the recursion below, with s_q integer-valued:

$$u_q = \frac{s_q}{T}$$
$$s_{q+1} = (a\, s_q + b) \mod T$$

It must be initialized by s_1 (the **seed**).
The sequence period is at most T, some values of a and b provide a period T.
If $n_\mathrm{r} \ll T$, $(u_1, \ldots, u_{n_\mathrm{r}})$ may be viewed as a realization of a n_r-sample of a r.v. U uniformly distributed between 0 and 1.[2]

- This algorithm behavior highly depends of the choice of a, b and T. An implementation which was widely used in the seventies was later proven to be a poor choice, so that som scientific results became doubtful [ref].

- Thereafter:
  - we will assume that the available calculation resources (computer, operating system, programming language with its libraries) provide a reliable uniformly distributed generator.
  - U will designate a r.v. we can simulate, but it will often be a uniformly distributed r.v.

2. Matlab. To generate a n_r-sample uniformy distributed between 0 and 1 (the seed is processed by the function `rng`).

```
x = rand(1,nr);
```

T196

Change of variable

If the r.v. X to be simulated can be obtained from the r.v. U with the transform $\mathrm{X} = h(\mathrm{U})$, we just have to simulate U and to apply the transform h to the result.

T196
Change of variable

If the r.v. X to be simulated can be obtained from the r.v. U with the transform $\mathrm{X} = h(\mathrm{U})$, we just have to simulate U and to apply the transform h to the result.

- In particular, in the dimension 1 case, in the CDF method, we use $h = F_{\mathrm{X}}^{-1}$ and U uniformly distributed between 0 and 1.[3]

3. If F_{X}^{-1} is available in the used programming language.

T196 # Change of variable

If the r.v. X to be simulated can be obtained from the r.v. U with the transform $\mathrm{X} = h(\mathrm{U})$, we just have to simulate U and to apply the transform h to the result.

- In particular, in the dimension 1 case, in the CDF method, we use $h = F_{\mathrm{X}}^{-1}$ and U uniformly distributed between 0 and 1.[3]
- The Box-Muller transform provides a sample driven by a standardized normal distribution..[4]

3. If F_{X}^{-1} is available in the used programming language.
4. Matlab. To generate a n_{r}-sample driven by a zero mean and unit variance Gaussian distribution in $\mathbb{R}^d$.
```
x = randn(d,nr);
```

T196

Change of variable

If the r.v. X to be simulated can be obtained from the r.v. U with the transform $\mathrm{X} = h(\mathrm{U})$, we just have to simulate U and to apply the transform h to the result.

- In particular, in the dimension 1 case, in the CDF method, we use $h = F_{\mathrm{X}}^{-1}$ and U uniformly distributed between 0 and 1.[3]
- The Box-Muller transform provides a sample driven by a standardized normal distribution..[4]
- To get a sample with a given mean and a given variance from a zero mean unit variance sample, we left multiply by a variance square root (the Cholesky decomposition for example), then we add the mean value.[5]

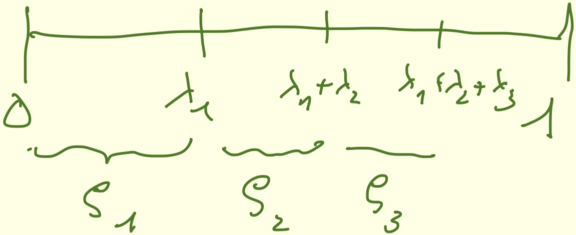
---

3. If F_{X}^{-1} is available in the used programming language.
4. Matlab. To generate a n_{r}-sample driven by a zero mean and unit variance Gaussian distribution in $\mathbb{R}^d$.
```
x = randn(d,nr);
```
5. Matlab. To generate a Gaussian n_{r}-sample, with mean m and variance C.
```
x = chol(C,'lower')*randn(length(C),nr) + m*ones(1,nr);
```

T197 Discrete distributions with finite support, mixture distributions

To simulate a r.v. z which takes its value in $\{\zeta_1, \ldots, \zeta_{n_c}\}$, with $\mathrm{Prob}\,(\mathrm{z} = z) = \lambda_z$:

- we simulate a r.v. uniformly distributed between 0 and 1:
- then we select ζ_c if the result belongs to the interval $[\sum_{\ell=1}^{c-1} \lambda_{\zeta_\ell}, \sum_{\ell=1}^{c} \lambda_{\zeta_\ell}]$. [6]



6. We choose ζ_{n_c} after $n_c - 1$ tests, and, for all $c < n_c$, we choose ζ_c after de c tests. The mean number of tests is $\sum_{c=1}^{n_c} c\,\lambda_c - \lambda_{n_c}$. We should sort the λ_c in decreasing order, in order to minimize this number.

T197 Discrete distributions with finite support, mixture distributions

To simulate a r.v. Z which takes its value in $\{\zeta_1, \ldots, \zeta_{n_c}\}$, with $\mathrm{Prob}(\mathrm{Z} = z) = \lambda_z$:

- we simulate a r.v. uniformly distributed between 0 and 1:
- then we select ζ_c if the result belongs to the interval $[\sum_{\ell=1}^{c-1} \lambda_{\zeta_\ell}, \sum_{\ell=1}^{c} \lambda_{\zeta_\ell}]$. [6]

➲ To simulate a r.v. X driven by a mixture distribution $p_{\mathrm{X}}(x) = \sum_{z \in \{\zeta_1, \ldots, \zeta_{n_c}\}} \lambda_z \, f_z(x)$ (page 113), we simulate a pair (X, Z) in two steps:

- we simulate Z according to the procedure above; let ζ_c be the result;
- we simulate X with the distribution of PDF f_{ζ_c}.

6. We choose ζ_{n_c} after $n_c - 1$ tests, and, for all $c < n_c$, we choose ζ_c after de c tests. The mean number of tests is $\sum_{c=1}^{n_c} c \, \lambda_c - \lambda_{n_c}$. We should sort the λ_c in decreasing order, in order to minimize this number.

T198 We simulate a n_r-sample $(\mathrm{Z}_q)_{1\leqslant q\leqslant n_\mathrm{r}}$ of Z.
Let K_c the number of elements which take the value ζ_c in this sample:

$$\mathrm{K}_c = \mathrm{Card}\{q \in \{1,\ldots,n_\mathrm{r}\} \mid \mathrm{Z}_q = \zeta_c\}$$

$\mathrm{K} = \underbrace{(\mathrm{K}_1,\ldots,\mathrm{K}_{n_\mathrm{c}})}_{\text{sum}=n_\mathrm{r}}$ is driven by a **multinomial distribution** with parameters $(n_\mathrm{r}, \underbrace{\lambda_{\zeta_1},\ldots,\lambda_{\zeta_{n_\mathrm{c}}}}_{\text{sum}=1})$. Its PMF is:

$$p_\mathrm{K}(k_1,\ldots,k_{n_\mathrm{c}}) = n_\mathrm{r}! \prod_{c=1}^{n_\mathrm{c}} \frac{\lambda_{\zeta_c}^{k_c}}{k_c!}$$

The proposed algorithm permits to perform a **multinomial sampling**, that is the simulation of a r.v. driven by a multinomial distribution.[7][8]

7. Matlab. To generate a n_r-sample x of a discrete r.v. Z which takes its value in $\{1,\ldots,n_\mathrm{c}\}$ such that $\mathrm{Prob}\,(\mathrm{Z}=c)=\lambda_c$, together with one realization k with n_c components ($\lambda = (\lambda_1,\ldots,\lambda_{n_\mathrm{c}})$), the cth component of k is the number of occurrences of the value c in the array z).
```
[k,z] = histc(rand(1,nr),[-Inf cumsum(lambda(1:end-1)) Inf]); k(end) = [];
```
8. There exist some approximation of the multinomial sampling, with less tests and less random trials (stratified sampling, Kitagawa sampling...) [ref]. For example, the systematic Kitagawa sampling needs only one sortition, whatever the length n_r is.

T199 # Rejection sampling

We want to simulate a r.v. X (without simple generator), with PDF p_{X}.

To do this, we simulate a pair of independent r.v. $(\mathring{\mathrm{X}}, \mathrm{U})$:

- $\mathring{\mathrm{X}}$ is an instrumental r.v. (for which we have a generator), with PDF $p_{\mathring{\mathrm{X}}}$, such that the coefficient ν below (obviously in the interval $[0,1]$ is strictly positive:

$$\nu = \inf_{x \in \mathrm{S}(\mathrm{X})} \frac{p_{\mathring{\mathrm{X}}}(x)}{p_{\mathrm{X}}(x)}$$

  and such that we can determine a value $a \in]0, \nu]$ (thus, for all x, $a\, p_{\mathrm{X}}(x) \leqslant p_{\mathring{\mathrm{X}}}(x)$);
- U is uniformly distributed between 0 and 1.

T200 Let T be the indicator function of the event $a\, p_{\mathrm{X}}(\mathring{\mathrm{X}}) \geqslant \mathrm{U}\, p_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})$; this test r.v. takes its value in $\{0, 1\}$.

Thus, the distribution of $\mathring{\mathrm{X}}$ given the event $\mathrm{T} = 1$ is the distribution of X, the acceptance ratio is a:[P1]

$$p_{\mathring{\mathrm{X}}|\mathrm{T}}(x, 1) = p_{\mathrm{X}}(x)$$

$$p_{\mathrm{T}}(1) = a$$

In practice, we repeat the simulation of $\mathring{\mathrm{X}}$ and U until the test is true.[9]

9. More generally, the target PDF p_{X} and the instrumental $p_{\mathring{\mathrm{X}}}$ must be known up to a multiplicative constant, that is $p_{\mathrm{X}} = \alpha\, \tilde{p}_{\mathrm{X}}$ and $p_{\mathring{\mathrm{X}}} = \beta\, \tilde{p}_{\mathring{\mathrm{X}}}$ with unknown α or β. We must determine $\tilde{a} > 0$ (beware! $\tilde{a}$ can be greater than 1) such that for all $\mathring{x}$, $\tilde{a}\, \tilde{p}_{\mathrm{X}}(\mathring{x}) \leqslant \tilde{p}_{\mathring{\mathrm{X}}}(\mathring{x})$, and the test becomes $\tilde{a}\, \tilde{p}_{\mathrm{X}}(\mathring{\mathrm{X}}) \geqslant \mathrm{U}\, \tilde{p}_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})$. Thus, $a = \frac{\beta}{\alpha}\, \tilde{a}$.

T200 Let T be the indicator function of the event $a\, p_{\mathrm{X}}(\mathring{\mathrm{X}}) \geqslant \mathrm{U}\, p_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})$; this test r.v. takes its value in $\{0, 1\}$.

Thus, the distribution of $\mathring{\mathrm{X}}$ given the event $\mathrm{T} = 1$ is the distribution of X, the acceptance ratio is a:[P1]

$$p_{\mathring{\mathrm{X}}|\mathrm{T}}(x, 1) = p_{\mathrm{X}}(x)$$

$$p_{\mathrm{T}}(1) = a$$

In practice, we repeat the simulation of $\mathring{\mathrm{X}}$ and U until the test is true.[9]

➔ For a fast algorithm, we must select an instrumental distribution such that:

- the acceptance ratio a is close to 1, to avoid a high number of lost trials;
- the instrumental r.v. is easily simulated;
- the test is easily calculated.

9. More generally, the target PDF p_{X} and the instrumental $p_{\mathring{\mathrm{X}}}$ must be known up to a multiplicative constant, that is $p_{\mathrm{X}} = \alpha\, \tilde{p}_{\mathrm{X}}$ and $p_{\mathring{\mathrm{X}}} = \beta\, \tilde{p}_{\mathring{\mathrm{X}}}$ with unknown α or β. We must determine $\tilde{a} > 0$ (beware! $\tilde{a}$ can be greater than 1) such that for all $\mathring{x}$, $\tilde{a}\, \tilde{p}_{\mathrm{X}}(\mathring{x}) \leqslant \tilde{p}_{\mathring{\mathrm{X}}}(\mathring{x})$, and the test becomes $\tilde{a}\, \tilde{p}_{\mathrm{X}}(\mathring{\mathrm{X}}) \geqslant \mathrm{U}\, \tilde{p}_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})$. Thus, $a = \frac{\beta}{\alpha}\, \tilde{a}$.

T201 The rejection method (algo. 4.1) is typically used to simulate a gamma distributed r.v. [ref].

Algorithm 4.1 Rejection method for stochastic simulation

Inputs

- Target PDF $\tilde{p}_{\mathrm{X}}$ (up to a multiplicative constant).
- Instrumental PDF $\tilde{p}_{\mathring{\mathrm{X}}}$ (up to a multiplicative constant) and associated generator.
- $\tilde{a} > 0$ such that for all $x \in \mathrm{S}\,(\mathrm{X})$, $\tilde{p}_{\mathring{\mathrm{X}}}(x) \geqslant \tilde{a}\,\tilde{p}_{\mathrm{X}}(x)$.

Returns a realization x of X.

Algorithm

- Repeat
  - Draw u uniformly distributed between 0 and 1
  - Draw $\mathring{x}$ according to the distribution with PDF $\propto \tilde{p}_{\mathring{\mathrm{X}}}$
- until $\tilde{a}\,\tilde{p}_{\mathrm{X}}(\mathring{x}) \geqslant u\,\tilde{p}_{\mathring{\mathrm{X}}}(\mathring{x})$
- $x = \mathring{x}$

T202 Chapter 5

Monte Carlo methods

We want to evaluate $\mathrm{E}\left(\phi(\mathrm{x})\right) = \displaystyle\int \phi(x)\, p_{\mathrm{x}}(x)\, \mathrm{d}\, x$ for a r.v. x and a function ϕ defined on the support of x such that this expectation exists.

Chapter contents

T203
Direct sampling

We can simulate X.
It is not necessary to express p_{X}.

T203 # Direct sampling

We can simulate X.
It is not necessary to express p_{X}.

- Let $(\mathrm{X}_q)_{1\leqslant q\leqslant n_{\mathrm{r}}}$ be a n_{r}-sample of X. The arithmetic mean $\sum_{q=1}^{n_{\mathrm{r}}} \frac{1}{n_{\mathrm{r}}}\,\phi(\mathrm{X}_q)$ is a r.v. with mean $\mathrm{E}\,(\phi(\mathrm{X}))$ (that is the quantity to evaluate), and variance $\frac{1}{n_{\mathrm{r}}}\,\mathrm{Var}\,(\phi(\mathrm{X}))$. Thus, we can write:

$$\boxed{\mathrm{E}\,(\phi(\mathrm{X})) \simeq \sum_{q=1}^{n_{\mathrm{r}}} \frac{1}{n_{\mathrm{r}}}\,\phi(\mathrm{X}_q)} \tag{5.1}$$

For every realization $(x_q)_{1\leqslant q\leqslant n_{\mathrm{r}}}$ of $(\mathrm{X}_q)_{1\leqslant q\leqslant n_{\mathrm{r}}}$, the empirical mean $\sum_{q=1}^{n_{\mathrm{r}}} \frac{1}{n_{\mathrm{r}}}\,\phi(x_q)$ is an approximation of $\mathrm{E}\,(\phi(\mathrm{X}))$.

T204

Importance sampling

We can simulate an instrumental r.v. $\mathring{\mathrm{x}}$ (the support of $\mathring{\mathrm{x}}$ must contain the support of x).
We can calculate the ratio $\frac{p_{\mathrm{x}}}{p_{\mathring{\mathrm{x}}}}$.

We easily check that the expectation $\mathrm{E}\left(\phi(\mathrm{x})\right)$ can be written as:

$$\mathrm{E}\left(\phi(\mathrm{x})\right) = \mathrm{E}\left(\frac{p_{\mathrm{x}}(\mathring{\mathrm{x}})}{p_{\mathring{\mathrm{x}}}(\mathring{\mathrm{x}})}\,\phi(\mathring{\mathrm{x}})\right)$$

$$\psi = \frac{p_x}{p_{\mathring{x}}}\phi$$

$$E(\phi(x)) = \int \phi(x)\, p_x(x)\, dx$$

$$E(\psi(\mathring{x})) = \int \psi(x)\, p_{\mathring{x}}(x)\, dx$$

$$= \int \frac{p_x(x)}{\cancel{p_{\mathring{x}}(x)}}\,\phi(x)\,\cancel{p_{\mathring{x}}(x)}\, dx$$

$$= E(\phi(x))$$

T204 # Importance sampling

We can simulate an instrumental r.v. $\mathring{\mathrm{X}}$ (the support of $\mathring{\mathrm{X}}$ must contain the support of X).
We can calculate the ratio $\frac{p_{\mathrm{X}}}{p_{\mathring{\mathrm{X}}}}$.

We easily check that the expectation $\mathrm{E}\left(\phi(\mathrm{X})\right)$ can be written as:

$$\mathrm{E}\left(\phi(\mathrm{X})\right) = \mathrm{E}\left(\frac{p_{\mathrm{X}}(\mathring{\mathrm{X}})}{p_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})}\,\phi(\mathring{\mathrm{X}})\right)$$

- Let $(\mathring{\mathrm{X}}_q)_{1\leqslant q\leqslant n_\mathrm{r}}$ be a n_r-sample of $\mathring{\mathrm{X}}$. We can write:

$$\mathrm{E}\left(\phi(\mathrm{X})\right) \simeq \sum_{q=1}^{n_\mathrm{r}} \underbrace{\frac{1}{n_\mathrm{r}}\,\frac{p_{\mathrm{X}}(\mathring{\mathrm{X}}_q)}{p_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}}_q)}}_{\omega_q(\mathring{\mathrm{X}}_q)}\,\phi(\mathring{\mathrm{X}}_q) \tag{5.2}$$

For every realization $(\mathring{x}_q)_{1\leqslant q\leqslant n_\mathrm{r}}$ of $(\mathring{\mathrm{X}}_q)_{1\leqslant q\leqslant n_\mathrm{r}}$, the weighted mean $\sum_{q=1}^{n_\mathrm{r}} \omega_q(\mathring{x}_q)\,\phi(\mathring{x}_q)$ is an approximation of $\mathrm{E}\left(\phi(\mathrm{X})\right)$. The Monte-Carlo method relies on a **weighted sampling**, also called **importance sampling**.[1]

1. If $\mathrm{Var}\left(\frac{p_{\mathrm{X}}(\mathring{\mathrm{X}})}{p_{\mathring{\mathrm{X}}}(\mathring{\mathrm{X}})}\,\phi(\mathring{\mathrm{X}})\right) < \mathrm{Var}\left(\phi(\mathrm{X})\right)$, we obtain a better approximation than with direct sampling; the variance reduction through importance sampling is beyond the scope of this document.

T205 Importance sampling with auxiliary variable

Typically, we use this method if x is driven by a mixture distribution; x is a marginal of the pair (x, z), with z the mixture component.

We can simulate the instrumental pair $(\mathring{\mathrm{x}}, \mathring{\mathrm{z}})$ (the support of $(\mathring{\mathrm{x}}, \mathring{\mathrm{z}})$ must contain the support of (x, z)) [ref].
We can calculate $\frac{p_{\mathrm{x},\mathrm{z}}}{p_{\mathring{\mathrm{x}},\mathring{\mathrm{z}}}}$.

We easily check the equality below, where $\mathring{\mathrm{z}}$ is the auxiliary variable:

$$\mathrm{E}\left(\phi(\mathrm{x})\right) = \mathrm{E}\left(\frac{p_{\mathrm{x},\mathrm{z}}(\mathring{\mathrm{x}}, \mathring{\mathrm{z}})}{p_{\mathring{\mathrm{x}},\mathring{\mathrm{z}}}(\mathring{\mathrm{x}}, \mathring{\mathrm{z}})}\,\phi(\mathring{\mathrm{x}})\right)$$

T205 Importance sampling with auxiliary variable

Typically, we use this method if X is driven by a mixture distribution; X is a marginal of the pair (X, Z), with Z the mixture component.

We can simulate the instrumental pair $(\mathring{\mathrm{X}}, \mathring{\mathrm{Z}})$ (the support of $(\mathring{\mathrm{X}}, \mathring{\mathrm{Z}})$ must contain the support of (X, Z)) [ref].
We can calculate $\frac{p_{\mathrm{X},\mathrm{Z}}}{p_{\mathring{\mathrm{X}},\mathring{\mathrm{Z}}}}$.

We easily check the equality below, where $\mathring{\mathrm{Z}}$ is the auxiliary variable:

$$\mathrm{E}\left(\phi(\mathrm{X})\right) = \mathrm{E}\left(\frac{p_{\mathrm{X},\mathrm{Z}}(\mathring{\mathrm{X}}, \mathring{\mathrm{Z}})}{p_{\mathring{\mathrm{X}},\mathring{\mathrm{Z}}}(\mathring{\mathrm{X}}, \mathring{\mathrm{Z}})}\, \phi(\mathring{\mathrm{X}})\right)$$

- Let $(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q)_{1 \leqslant q \leqslant n_{\mathrm{r}}}$ be a n_{r}-sample of $(\mathring{\mathrm{X}}, \mathring{\mathrm{Z}})$. We can write:

$$\mathrm{E}\left(\phi(\mathrm{X})\right) \simeq \sum_{q=1}^{n_{\mathrm{r}}} \underbrace{\frac{1}{n_{\mathrm{r}}}\, \frac{p_{\mathrm{X},\mathrm{Z}}(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q)}{p_{\mathring{\mathrm{X}},\mathring{\mathrm{Z}}}(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q)}}_{\omega(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q)}\, \phi(\mathring{\mathrm{X}}_q) \tag{5.3}$$

For every realization $(\mathring{x}_q, z_q)_{1 \leqslant q \leqslant n_{\mathrm{r}}}$ of $(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q)_{1 \leqslant q \leqslant n_{\mathrm{r}}}$, the weighted mean $\sum_{q=1}^{n_{\mathrm{r}}} \omega(\mathring{x}_q, \mathring{z}_q)\, \phi(\mathring{x}_q)$ is an approximation of $\mathrm{E}\left(\phi(\mathrm{X})\right)$.

T206 Particle approximation

The formulas (5.1), (5.2) ou (5.3) provide an approximation of $\mathrm{E}\left(\phi(\mathrm{X})\right)$ which can be re-written as:

$$\mathrm{E}\left(\phi(\mathrm{X})\right) \simeq \int \phi(x)\, \hat{p}_{\mathrm{X}}(x)\, \mathrm{d}\, x \quad \text{with} \quad \hat{p}_{\mathrm{X}}(x) = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q\, \delta(x - \mathring{\mathrm{X}}_q) \quad \text{and} \quad \omega_q = \begin{cases} \frac{1}{n_{\mathrm{r}}} & \text{direct} \\ \omega(\mathring{\mathrm{X}}_q) & \text{importance} \\ \omega(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q) & \text{aux. var.} \end{cases}$$

The pulses sum $\hat{p}_{\mathrm{X}}$ is a **particle** approximation of the actual PDF of X. It is random since it depends of the **particles** $(\mathring{\mathrm{X}}_q)_{1 \leqslant q \leqslant n_{\mathrm{r}}}$ and of the auxiliary variables $(\mathring{\mathrm{Z}}_q)_{1 \leqslant q \leqslant n_{\mathrm{r}}}$.

T206

Particle approximation

The formulas (5.1), (5.2) ou (5.3) provide an approximation of $\mathrm{E}\,(\phi(\mathrm{X}))$ which can be re-written as:

$$\mathrm{E}\,(\phi(\mathrm{X})) \simeq \int \phi(x)\,\hat{p}_{\mathrm{X}}(x)\;\mathrm{d}\,x \quad \text{with} \quad \hat{p}_{\mathrm{X}}(x) = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q\,\delta(x - \mathring{\mathrm{X}}_q) \quad \text{and} \quad \omega_q = \begin{cases} \frac{1}{n_{\mathrm{r}}} & \text{direct} \\ \omega(\mathring{\mathrm{X}}_q) & \text{importance} \\ \omega(\mathring{\mathrm{X}}_q, \mathring{\mathrm{Z}}_q) & \text{aux. var.} \end{cases}$$

The pulses sum $\hat{p}_{\mathrm{X}}$ is a **particle** approximation of the actual PDF of X. It is random since it depends of the **particles** $(\mathring{\mathrm{X}}_q)_{1\leqslant q\leqslant n_{\mathrm{r}}}$ and of the auxiliary variables $(\mathring{\mathrm{Z}}_q)_{1\leqslant q\leqslant n_{\mathrm{r}}}$.

⚠ In the importance sampling case, the weights are random, with mean $1/n_{\mathrm{r}}$; their sum is random, with mean 1. A realization of $\hat{p}_{\mathrm{X}}$ is not a PDF. To keep a PDF interpretation, we must divide the weights by their sum.

T207 The exercise below shows that the particle approximation by itself is an approximation of the PMF of a discrete r.v., but cannot be used as the PDF of a continuous r.v.

◁ **Exercise 34.** Show that for all x, $\mathrm{E}\left(\hat{p}_{\mathrm{X}}(x)\right) = p_{\mathrm{X}}(x)$, and that $\mathrm{Var}\left(\hat{p}_{\mathrm{X}}(x)\right)$ is finite if X is discrete, infinite if X is continuous.

T208 Application to Bayesian estimation

Let x be the parameter to estimate. ϕ is the identity function, $\hat{p}_{\mathrm{x}}$ a particle approximation of the *a priori* PDF for a prediction, of the *a posteriori* PDF for an estimation.

T208 Application to Bayesian estimation

Let x be the parameter to estimate. ϕ is the identity function, $\hat{p}_{\mathrm{x}}$ a particle approximation of the *a priori* PDF for a prediction, of the *a posteriori* PDF for an estimation.

- If x is numerically valued, the MMSE predictor approximation is the particles weighted mean:

$$\widehat{\mathrm{x}}_{\mathrm{MMSE}} = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q \, \mathring{\mathrm{x}}_q$$

T208 Application to Bayesian estimation

Let x be the parameter to estimate. ϕ is the identity function, $\hat{p}_{\mathrm{x}}$ a particle approximation of the *a priori* PDF for a prediction, of the *a posteriori* PDF for an estimation.

- If x is numerically valued, the MMSE predictor approximation is the particles weighted mean:

$$\widehat{\mathrm{x}}_{\mathrm{MMSE}} = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q \, \mathring{\mathrm{x}}_q$$

- If x is discrete, the MAP predictor approximation is the value for which the sum of associated weights is maximal (if the weights are equal, it is the majority vote):[2]

$$\widehat{\mathrm{x}}_{\mathrm{MAP}} = \arg\max_x \sum_{q \mid \mathring{\mathrm{x}}_q = x} \omega_q$$

2. If x has continuous and discrete components, that is $\mathrm{x} = (\mathrm{x}_{\mathrm{c}}, \mathrm{x}_{\mathrm{d}})$, we can use the marginal MAP for the discrete components, and the conditional MMSE for the continuous ones: $\hat{\mathrm{x}}_{\mathbf{d}} = \arg\max_{x_{\mathrm{d}}} \sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = x_{\mathbf{d}}} \omega_q$ and $\hat{\mathrm{x}}_{\mathbf{c}} = \frac{1}{\sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = \hat{\mathrm{x}}_{\mathbf{d}}} \omega_q} \sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = \hat{\mathrm{x}}_{\mathbf{d}}} \omega_q \, \mathring{\mathrm{x}}_{\mathbf{c},q}$

T208

Application to Bayesian estimation

Let x be the parameter to estimate. ϕ is the identity function, $\hat{p}_{\mathrm{x}}$ a particle approximation of the *a priori* PDF for a prediction, of the *a posteriori* PDF for an estimation.

- If x is numerically valued, the MMSE predictor approximation is the particles weighted mean:

$$\widehat{\mathrm{x}}_{\mathrm{MMSE}} = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q \, \mathring{\mathrm{x}}_q$$

- If x is discrete, the MAP predictor approximation is the value for which the sum of associated weights is maximal (if the weights are equal, it is the majority vote):[2]

$$\widehat{\mathrm{x}}_{\mathrm{MAP}} = \arg\max_{x} \sum_{q \mid \mathring{\mathrm{x}}_q = x} \omega_q$$

- The transposition to *a posteriori* estimators is straightforward.

we calculate $E(X|Y)$ using stochastic simulation

2. If x has continuous and discrete components, that is $\mathrm{x} = (\mathrm{x}_{\mathrm{c}}, \mathrm{x}_{\mathrm{d}})$, we can use the marginal MAP for the discrete components, and the conditional MMSE for the continuous ones: $\hat{\mathrm{x}}_{\mathbf{d}} = \arg\max_{x_{\mathrm{d}}} \sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = x_{\mathbf{d}}} \omega_q$ and $\hat{\mathrm{x}}_{\mathbf{c}} = \frac{1}{\sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = \hat{\mathrm{x}}_{\mathbf{d}}} \omega_q} \sum_{q \mid \mathring{\mathrm{x}}_{q,\mathbf{d}} = \hat{\mathrm{x}}_{\mathbf{d}}} \omega_q \, \mathring{\mathrm{x}}_{\mathbf{c},q}$

T209 # Chapter 6

Particle filter

We try to recursively estimate the state of an HMM. We implement a Monte Carlo based approximation of the Bayesian filter, that is a **Sequential Monte Carlo (SMC)** method.

Unlike the Kalman filter, we do not only transmit the mean and the variance, but a particle approximation of some probability distributions.

Chapter contents

T210 Principle

Lets's consider a Markov model with state $(\mathrm{X}[n])_{n\geqslant 1}$ and observation $(\mathrm{Y}[n])_{n\geqslant 1}$. To simplify the writing:

the initial PDF	$p_{\mathrm{X}[1]}(x)$	is noted	$\rho(x)$
the transition PDF	$p_{\mathrm{X}[n+1]\|\mathrm{X}[n]}(x^+, x)$	is noted	$\kappa_{n+1}(x^+, x)$

T210 # Principle

Lets's consider a Markov model with state $(\mathrm{X}[n])_{n\geqslant 1}$ and observation $(\mathrm{Y}[n])_{n\geqslant 1}$. To simplify the writing:

the initial PDF $p_{\mathrm{X}[1]}(x)$ is noted $\rho(x)$

the transition PDF $p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+, x)$ is noted $\kappa_{n+1}(x^+, x)$ and $p_{Y[n]|X[n]}$

- We remind the Bayes filtering (page 175) with the notations above ((3.12) and (3.13) are merged):

$$\text{initialization} \quad p^{|0}_{\mathrm{X}[1]}(x^+) = \rho(x^+) \tag{6.1}$$

$$\text{recursion } (n \geqslant 1) \quad p^{|n}_{\mathrm{X}[n]}(x) = \frac{p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], x)\, p^{|n-1}_{\mathrm{X}[n]}(x)}{\int p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], u)\, p^{|n-1}_{\mathrm{X}[n]}(u)\, \mathrm{d}\,u} \tag{6.2}$$

$$p^{|n}_{\mathrm{X}[n+1]}(x^+) = \int \kappa_{n+1}(x^+, x)\, p^{|n}_{\mathrm{X}[n]}(x)\, \mathrm{d}\,x \tag{6.3}$$

MMSE estimator $E(X[n] \mid Y[1], \dots, Y[n])$

$= \int x\, p^{|n}_{X[n]}(x)\, dx$

T210 # Principle

Lets's consider a Markov model with state $(\mathrm{X}[n])_{n\geqslant 1}$ and observation $(\mathrm{Y}[n])_{n\geqslant 1}$. To simplify the writing:

the initial PDF $p_{\mathrm{X}[1]}(x)$ is noted $\rho(x)$

the transition PDF $p_{\mathrm{X}[n+1]|\mathrm{X}[n]}(x^+, x)$ is noted $\kappa_{n+1}(x^+, x)$

- We remind the Bayes filtering (page 175) with the notations above ((3.12) and (3.13) are merged):

$$\text{initialization} \quad p^{|0}_{\mathrm{X}[1]}(x^+) = \rho(x^+) \tag{6.1}$$

$$\text{recursion } (n \geqslant 1) \quad p^{|n}_{\mathrm{X}[n]}(x) = \frac{p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], x)\, p^{|n-1}_{\mathrm{X}[n]}(x)}{\displaystyle\int p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], u)\, p^{|n-1}_{\mathrm{X}[n]}(u)\,\mathrm{d}\,u} \tag{6.2}$$

$$p^{|n}_{\mathrm{X}[n+1]}(x^+) = \int \kappa_{n+1}(x^+, x)\, p^{|n}_{\mathrm{X}[n]}(x)\,\mathrm{d}\,x \tag{6.3}$$

- The particle filter [ref] consists:
  - in a preliminary particle approximation of the initial distribution, through sampling;
  - in a propagation of this approximation with the Bayes filter;
  - a resampling step is then necessary.

T211 This **condensation** (“**cond**itional **dens**ity propag**ation**”) reminds the evolution theory (figure 6.1), in which the highly adapted individuals are selected.

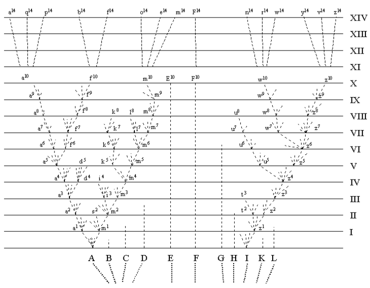


Figure 6.1: unique illustration in “Origin of Species” (Ch. Darwin, 1859)

T212 We use an instrumental stochastic process $\mathring{\mathrm{X}} = (\mathring{\mathrm{X}}[n])_{n\geqslant 1}$ which must be a Markov chain given the observations $\mathrm{Y} = (\mathrm{Y}[n])_{n\geqslant 1}$. To simplify the writing:

instrumental initial PDF	$p_{\mathring{\mathrm{X}}[1]\mid\mathrm{Y}}(x)$	is noted $\mathring{\rho}(x)$
instrumental transition PDF	$p_{\mathring{\mathrm{X}}[n+1]\mid\mathring{\mathrm{X}}[n],\mathrm{Y}}(x^+, x)$	is noted $\mathring{\kappa}_{n+1}(x^+, x)$

T213 **To start up**, we need a particle approximation of the initial distribution with PDF ρ. This is the initial sampling, based on the instrumental distribution with PDF $\mathring{\rho}$.

For all $q \in \{1, \ldots, n_\mathrm{r}\}$:

- an particle is sampled according to this distribution;
- its weight is calculated thanks to the importance sampling (5.2).

We obtain the particle approximation below:

$$\hat{p}^{|0}_{\mathrm{X}[1]}(x) = \sum_{q=1}^{n_\mathrm{r}} \widetilde{\omega}_q[1]\,\delta(x - \mathring{\mathrm{X}}_q[1]) \quad \text{with} \quad \widetilde{\omega}_q[1] = \frac{1}{n_\mathrm{r}}\,\frac{\rho(\mathring{\mathrm{X}}_q[1])}{\mathring{\rho}(\mathring{\mathrm{X}}_q[1])}$$

T214 **The recursion** of the Bayesian filtering is fed with a particle approximation:

$$\hat{p}_{\mathrm{X}[n]}^{|n-1}(x) \quad = \sum_{q=1}^{n_{\mathrm{r}}} \widetilde{\omega}_q[n]\, \delta(x - \mathring{\mathrm{X}}_q[n]) \tag{6.4}$$

We apply the formula (6.2); the particles are unchanged, their weight is multiplied by their local likelihood:

$$\hat{p}_{\mathrm{X}[n]}^{|n}(x) \quad = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q[n]\, \delta(x - \mathring{\mathrm{X}}_q[n]) \quad \text{with } \omega_q[n] = \frac{\widetilde{\omega}_q[n]\, p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], \mathring{\mathrm{X}}_q[n])}{\sum_{z=1}^{n_{\mathrm{r}}} \widetilde{\omega}_z[n]\, p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], \mathring{\mathrm{X}}_z[n])} \tag{6.5}$$

We apply the formula (6.3); we obtain a mixture distribution:

$$\hat{p}_{\mathrm{X}[n+1]}^{|n}(x^+) = \sum_{q=1}^{n_{\mathrm{r}}} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{X}}_q[n]) \tag{6.6}$$

T215 **To go ahead**, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.
This is the resampling.
We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \mathring{\omega}_q[n]\, \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.

T215 **To go ahead**, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$. This is the resampling.
We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \mathring{\omega}_q[n]\, \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.

- For all $q \in \{1, \ldots, n_\mathrm{r}\}$:
  - the parent particle index $\mathrm{z}_q[n+1]$ is selected according to the distribution $\mathrm{Prob}\,(\mathrm{z} = z) = \mathring{\omega}_z[n]$;[1]
  - the child particle $\mathring{\mathrm{x}}_q[n+1]$ mutates according to the distribution $x^+ \mapsto \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_{\mathrm{z}_q[n+1]}[n])$.
  - its weight is calculated thanks to the importance sampling with auxiliary variable (5.3).[2]

Thus, the formula (6.6) is replaced by:

$$\hat{p}^{|n}_{\mathrm{x}[n+1]}(x) = \sum_{q=1}^{n_\mathrm{r}} \widetilde{\omega}_q[n+1]\, \delta(x - \mathring{\mathrm{x}}_q[n+1]) \text{ with } \widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \left.\frac{\kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]}\right|_{z=\mathrm{z}_q[n+1]} \tag{6.7}$$

1. $z \in \{1, \ldots, n_\mathrm{r}\}$. See multinomial sampling, page 198. $\mathrm{z}_q[n+1]$ is the **auxiliary** variable.
2. We could use the standard importance sampling (5.2), at the cost of $2\,n_\mathrm{r}$ PDF calculation per particle: $\widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \frac{\sum_{z=1}^{n_\mathrm{r}} \kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\sum_{z=1}^{n_\mathrm{r}} \mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]}$

T215 **To go ahead**, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$. This is the resampling.
We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \mathring{\omega}_q[n]\, \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.

- For all $q \in \{1, \ldots, n_\mathrm{r}\}$:
  - the parent particle index $\mathrm{z}_q[n+1]$ is selected according to the distribution $\mathrm{Prob}\,(\mathrm{z} = z) = \mathring{\omega}_z[n]$;[1]
  - the child particle $\mathring{\mathrm{x}}_q[n+1]$ mutates according to the distribution $x^+ \mapsto \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_{\mathrm{z}_q[n+1]}[n])$.
  - its weight is calculated thanks to the importance sampling with auxiliary variable (5.3).[2]

Thus, the formula (6.6) is replaced by:

$$\hat{p}^{|n}_{\mathrm{x}[n+1]}(x) = \sum_{q=1}^{n_\mathrm{r}} \widetilde{\omega}_q[n+1]\, \delta(x - \mathring{\mathrm{x}}_q[n+1]) \text{ with } \widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \left. \frac{\kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]} \right|_{z=\mathrm{z}_q[n+1]} \tag{6.7}$$

1. $z \in \{1, \ldots, n_\mathrm{r}\}$. See multinomial sampling, page 198. $\mathrm{z}_q[n+1]$ is the **auxiliary** variable.
2. We could use the standard importance sampling (5.2), at the cost of $2\,n_\mathrm{r}$ PDF calculation per particle: $\widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \frac{\sum_{z=1}^{n_\mathrm{r}} \kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\sum_{z=1}^{n_\mathrm{r}} \mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]}$

T215 **To go ahead**, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_r} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$. This is the resampling.
We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_r} \mathring{\omega}_q[n]\, \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.

For all $q \in \{1, \ldots, n_r\}$:

- the parent particle index $\mathrm{z}_q[n+1]$ is selected according to the distribution $\mathrm{Prob}\,(\mathrm{z} = z) = \mathring{\omega}_z[n]$;[1]
- the child particle $\mathring{\mathrm{x}}_q[n+1]$ mutates according to the distribution $x^+ \mapsto \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_{\mathrm{z}_q[n+1]}[n])$.
- its weight is calculated thanks to the importance sampling with auxiliary variable (5.3).[2]

Thus, the formula (6.6) is replaced by:

$$\hat{p}^{|n}_{\mathrm{x}[n+1]}(x) = \sum_{q=1}^{n_r} \widetilde{\omega}_q[n+1]\, \delta(x - \mathring{\mathrm{x}}_q[n+1]) \text{ with } \widetilde{\omega}_q[n+1] = \frac{1}{n_r} \left. \frac{\kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]} \right|_{z=\mathrm{z}_q[n+1]} \tag{6.7}$$

1. $z \in \{1, \ldots, n_r\}$. See multinomial sampling, page 198. $\mathrm{z}_q[n+1]$ is the **auxiliary** variable.
2. We could use the standard importance sampling (5.2), at the cost of $2\,n_r$ PDF calculation per particle: $\widetilde{\omega}_q[n+1] = \frac{1}{n_r} \frac{\sum_{z=1}^{n_r} \kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\sum_{z=1}^{n_r} \mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]}$

T215 **To go ahead**, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \omega_q[n]\, \kappa_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.
This is the resampling.
We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_\mathrm{r}} \mathring{\omega}_q[n]\, \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_q[n])$.

- For all $q \in \{1, \ldots, n_\mathrm{r}\}$:
  - the parent particle index $\mathrm{z}_q[n+1]$ is selected according to the distribution $\mathrm{Prob}\,(\mathrm{z} = z) = \mathring{\omega}_z[n]$;[1]
  - the child particle $\mathring{\mathrm{x}}_q[n+1]$ mutates according to the distribution $x^+ \mapsto \mathring{\kappa}_{n+1}(x^+, \mathring{\mathrm{x}}_{\mathrm{z}_q[n+1]}[n])$.
  - its weight is calculated thanks to the importance sampling with auxiliary variable (5.3).[2]

Thus, the formula (6.6) is replaced by:

$$\hat{p}^{|n}_{\mathrm{x}[n+1]}(x) = \sum_{q=1}^{n_\mathrm{r}} \widetilde{\omega}_q[n+1]\, \delta(x - \mathring{\mathrm{x}}_q[n+1]) \text{ with } \widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \left. \frac{\kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]} \right|_{z=\mathrm{z}_q[n+1]} \tag{6.7}$$

1. $z \in \{1, \ldots, n_\mathrm{r}\}$. See multinomial sampling, page 198. $\mathrm{z}_q[n+1]$ is the **auxiliary** variable.
2. We could use the standard importance sampling (5.2), at the cost of $2\,n_\mathrm{r}$ PDF calculation per particle: $\widetilde{\omega}_q[n+1] = \frac{1}{n_\mathrm{r}} \frac{\sum_{z=1}^{n_\mathrm{r}} \kappa_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \omega_z[n]}{\sum_{z=1}^{n_\mathrm{r}} \mathring{\kappa}_{n+1}(\mathring{\mathrm{x}}_q[n+1], \mathring{\mathrm{x}}_z[n])\, \mathring{\omega}_z[n]}$

T216 For online processing, the instrumental chain, which is sampled and resampled, can rely on the current observation.

If $\mathring{\kappa}_{n+1}$ rely on $(\mathrm{Y}[k])_{1\leqslant k\leqslant n}$ (and $\mathring{\rho}$ on no observation), we can include a prediction step in the filter. This is the **bootstrap filter** case, in which the instrumental chain do not rely on any observation.

Sif $\mathring{\kappa}_{n+1}$ rely on $(\mathrm{Y}[k])_{1\leqslant k\leqslant n+1}$ (and $\mathring{\rho}$ on $\mathrm{Y}[1]$), this is meaningless, prediction and estimation steps must be merged.

T217 Bootstrap filter

The particle are sampled according to the initial distribution and the transition one of the model of the data, that is: $\mathring{\rho} = \rho$, $\mathring{\kappa}_{n+1} = \kappa_{n+1}$, and $\mathring{\omega}_q[n] = \omega_q[n]$ for all q.

The sampling of the distributions is then direct (no importance sampling).

T218 We get the "bootstrap" filter (algo. 6.1). The weights are proportional to the local likelihood.[34]

---

3. We can store the re-distributed paticles in the smae variable, as in the next note
4. Matlab. Resampling in the particle filter.
```
[~,z] = histc(rand(1,nr),[-Inf cumsum(omega(1:end-1)) Inf]); x = x(:,z);
```

Algorithm 6.1 Bootstrap filter

Can be applied to a model with state $(\mathrm{X}[n])_{n\geqslant 1}$, observation $(\mathrm{Y}[n])_{n\geqslant 1}$, **if we are able to**
- simulate the HMM state sequence,
- express $p_{\mathrm{Y}[n]|\mathrm{X}[n]}$.

Initialization

sampling ↓ $\forall q$	$\mathrm{X}_q \leftarrow$ sampling according to the distribution of $\mathrm{X}[1]$		provides $(\mathrm{X}_q[1])_{1\leqslant q\leqslant n_\mathrm{r}}$
prediction of $\mathrm{X}[1]$	$\hat{\mathrm{X}} \leftarrow \sum_{q=1}^{n_\mathrm{r}} \frac{1}{n_\mathrm{r}} \mathrm{X}_q$		$\hat{\mathrm{X}}^{\|0}[1]$

Loop $(n \geqslant 1)$

→ observation of $\mathrm{Y}[n]$		$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
weights update	$\forall q$	$\omega_q \leftarrow p_{\mathrm{Y}[n]\|\mathrm{X}[n]}(\mathrm{Y}, \mathrm{X}_q)$ (then normalization)	$(\omega_q[n])_{1\leqslant q\leqslant n_\mathrm{r}}$
estimation of $\mathrm{X}[n]$		$\hat{\mathrm{X}} \leftarrow \sum_{q=1}^{n_\mathrm{r}} \omega_q \mathrm{X}_q$	$\hat{\mathrm{X}}^{\|n}[n]$ →
resampling	$\forall q$	$\mathrm{Z}_q \leftarrow$ sampling according to the discrete distribution $(\omega_z)_{1\leqslant z\leqslant n_\mathrm{r}}$	$(\mathrm{Z}_q[n+1])_{1\leqslant q\leqslant n_\mathrm{r}}$
	$\forall q$	$\mathrm{X}_q^- \leftarrow \mathrm{X}_{\mathrm{Z}_q}$	$(\mathrm{X}_{\mathrm{Z}_q[n+1]}[n])_{1\leqslant q\leqslant n_\mathrm{r}}$
	↓ $\forall q$	$\mathrm{X}_q \leftarrow$ sampling according to the distribution of $\mathrm{X}[n+1] \mid \mathrm{X}[n] = \mathrm{X}_q^-$	$(\mathrm{X}_q[n+1])_{1\leqslant q\leqslant n_\mathrm{r}}$
prediction of $\mathrm{X}[n+1]$		$\hat{\mathrm{X}} \leftarrow \sum_{q=1}^{n_\mathrm{r}} \frac{1}{n_\mathrm{r}} \mathrm{X}_q$	$\hat{\mathrm{X}}^{\|n}[n+1]$

T219
Auxiliary particle filter

The bootstrap filter has few calculation per particle, but we need many particles because many of them are not selected.
It can be defective if the local likelihood of all the particles is 0. A solution is to sample the particle with a distribution based on the current observation (this is **adaptation** [ref]).

T220 The observation must precede the random sampling, a prediction is meaningless. We re-organize the general algorithm by merging the estimation stage and the prediction one (algo. 6.2).

We must know propose some instrumental distributions based on the current observation.

Algorithm 6.2 Auxiliary particle filter

Can be applied to a model with state $(\mathrm{X}[n])_{n\geqslant 1}$, observation $(\mathrm{Y}[n])_{n\geqslant 1}$, initial distribution ρ, transition κ_{n+1}, **if we are able to**

- simulate a chain $(\mathring{\mathrm{X}}[n])_{n\geqslant 1}$ with initial distribution $\mathring{\rho}$ and transition $\mathring{\kappa}_{n+1}$,
- express $\frac{\rho\, p_{\mathrm{Y}[1]|\mathrm{X}[1]}}{\mathring{\rho}}$ and $\frac{\kappa_n\, p_{\mathrm{Y}[n]|\mathrm{X}[n]}}{\mathring{\kappa}_n}$.

Initial time $(n = 1)$

$\rightarrow$ observation of $\mathrm{Y}[1]$ $\quad \mathrm{Y} \leftarrow$ sensors $\quad$ provides $\mathrm{Y}[1]$

sampling $\downarrow \forall q \quad \mathrm{X}_q \leftarrow$ sampling according to $\mathring{\rho}$ $\quad (\mathring{\mathrm{X}}_q[1])_{1\leqslant q\leqslant n_r}$

$\downarrow \forall q \quad \omega_q \leftarrow \dfrac{\rho(\mathrm{X}_q)}{\mathring{\rho}(\mathrm{X}_q)}\, p_{\mathrm{Y}[1]|\mathrm{X}[1]}(\mathrm{Y}, \mathrm{X}_q)$ (then normalization) $\quad (\omega_q[1])_{1\leqslant q\leqslant n_r}$

estimation of $\mathrm{X}[1]$ $\quad \hat{\mathrm{X}} \leftarrow \displaystyle\sum_{q=1}^{n_r} \omega_q\, \mathrm{X}_q$ $\quad \hat{\mathrm{X}}^{|1}[1] \rightarrow$

Loop $(n \geqslant 2)$

$\rightarrow$ observation of $\mathrm{Y}[n]$ $\quad \mathrm{Y} \leftarrow$ sensors $\quad \mathrm{Y}[n]$

resampling $\forall q \quad \mathrm{Z}_q \leftarrow$ sampling according to the discrete distribution $(\mathring{\omega}_z)_{1\leqslant z\leqslant n_r}$ $\quad (\mathrm{Z}_q[n])_{1\leqslant q\leqslant n_r}$

$\forall q \quad \mathrm{X}_q^- \leftarrow \mathrm{X}_{\mathrm{Z}_q}$ $\quad (\mathring{\mathrm{X}}_{\mathrm{Z}_q[n]}[n-1])_{1\leqslant q\leqslant n_r}$

$\forall q \quad \omega_q^- \leftarrow \omega_{\mathrm{Z}_q}$ $\quad (\omega_{\mathrm{Z}_q[n]}[n-1])_{1\leqslant q\leqslant n_r}$

$\forall q \quad \mathring{\omega}_q^- \leftarrow \mathring{\omega}_{\mathrm{Z}_q}$ $\quad (\mathring{\omega}_{\mathrm{Z}_q[n]}[n-1])_{1\leqslant q\leqslant n_r}$

$\downarrow \forall q \quad \mathrm{X}_q \leftarrow$ sampling according to $\mathring{\kappa}_n(., \mathrm{X}_{\mathrm{Z}_q}^-)$ $\quad (\mathring{\mathrm{X}}_q[n])_{1\leqslant q\leqslant n_r}$

$\downarrow \forall q \quad \omega_q \leftarrow \dfrac{\omega_q^-\, \kappa_n(\mathrm{X}_q, \mathrm{X}_q^-)}{\mathring{\omega}_q^-\, \mathring{\kappa}_n(\mathrm{X}_q, \mathrm{X}_q^-)}\, p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}, \mathrm{X}_q)$ (then normalization) $\quad (\omega_q[n])_{1\leqslant q\leqslant n_r}$

estimation of $\mathrm{X}[n]$ $\quad \hat{\mathrm{X}} \leftarrow \displaystyle\sum_{q=1}^{n_r} \omega_q\, \mathrm{X}_q$ $\quad \hat{\mathrm{X}}^{|n}[n] \rightarrow$

T221 # Fully adapted particle filter

For all time n, for all q, the weights $\omega_q[n]$ are some functions of $\mathring{\mathrm{X}}_{1:n_\mathrm{r}}[1:n]$, $\mathrm{Z}_{1:n_\mathrm{r}}[2:n]$, $\mathrm{Y}[1:n]$.

But they become deterministic and equal if we choose:[P1]

$$\mathring{\rho} = p_{\mathrm{X}[1]|\mathrm{Y}[1]} \tag{6.8}$$

$$\forall n \geqslant 2,\quad \mathring{\omega}_z[n-1] \propto \omega_z[n-1]\, p_{\mathrm{Y}[n]|\mathrm{X}[n-1]}(\mathrm{Y}[n], \mathring{\mathrm{X}}_z[n-1]) \tag{6.9}$$

$$\forall n \geqslant 2,\quad \mathring{\kappa}_n = p_{\mathrm{X}[n]|\mathrm{X}[n-1],\mathrm{Y}[n]} \tag{6.10}$$

T222 There are very few particles without child, the diversity is preserved during the selection (algo. 6.3).

Algorithm 6.3 Fully adapted particle filter

Can be applied to a model with state $(\mathrm{X}[n])_{n\geqslant 1}$, observation $(\mathrm{Y}[n])_{n\geqslant 1}$, **if we are able to**

- simulate $\mathrm{X}[1]$ given $\mathrm{Y}[1]$, and $\mathrm{X}[n]$ given $(\mathrm{X}[n-1], \mathrm{Y}[n])$, $n \geqslant 2$,
- express $p_{\mathrm{Y}[n]|\mathrm{X}[n-1]}$, $n \geqslant 2$.

Initial time $(n = 1)$

$\rightarrow$ observation of $\mathrm{Y}[1]$		$\mathrm{Y} \leftarrow$ sensors	provides $\mathrm{Y}[1]$
sampling $\downarrow$	$\forall q$	$\mathrm{X}_q \leftarrow$ sampling according to the distribution of $\mathrm{X}[1] \mid \mathrm{Y}[1] = \mathrm{Y}$	$(\mathring{\mathrm{X}}_q[1])_{1\leqslant q\leqslant n_\mathrm{r}}$
estimation of $\mathrm{X}[1]$		$\hat{\mathrm{X}} \leftarrow \sum_{q=1}^{n_\mathrm{r}} \frac{1}{n_\mathrm{r}} \mathrm{X}_q$	$\hat{\mathrm{X}}^{\vert 1}[1] \rightarrow$

Loop $(n \geqslant 2)$

$\rightarrow$ observation of $\mathrm{Y}[n]$		$\mathrm{Y} \leftarrow$ sensors	$\mathrm{Y}[n]$
selection distribution	$\forall z$	$\mathring{\omega}_z \leftarrow p_{\mathrm{Y}[n]\vert\mathrm{X}[n-1]}(\mathrm{Y}, \mathrm{X}_z)$ (then normalization)	$(\mathring{\omega}_z[n-1])_{1\leqslant z\leqslant n_\mathrm{r}}$
resampling	$\forall q$	$\mathrm{Z}_q \leftarrow$ sampling according to the discrete distribution $(\mathring{\omega}_z)_{1\leqslant z\leqslant n_\mathrm{r}}$	$(\mathrm{Z}_q[n])_{1\leqslant q\leqslant n_\mathrm{r}}$
	$\forall q$	$\mathrm{X}_q^- \leftarrow \mathrm{X}_{\mathrm{Z}_q}$	$(\mathring{\mathrm{X}}_{\mathrm{Z}_q[n]}[n-1])_{1\leqslant z\leqslant n_\mathrm{r}}$
$\downarrow$	$\forall q$	$\mathrm{X}_q \leftarrow$ sampling according to the distribution of $\mathrm{X}[n] \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q^-, \mathrm{Y}[n] = \mathrm{Y}$	$(\mathring{\mathrm{X}}_q[n])_{1\leqslant q\leqslant n_\mathrm{r}}$
estimation of $\mathrm{X}[n]$		$\hat{\mathrm{X}} \leftarrow \sum_{q=1}^{n_\mathrm{r}} \frac{1}{n_\mathrm{r}} \mathrm{X}_q$	$\hat{\mathrm{X}}^{\vert n}[n] \rightarrow$

T223 But it is rarely possible to sample according to the distributions (6.10). We sampled according to an "analogous" distribution, for example a normal distribution with suitable mean and variance. And it is rarely possible to express the conditional distribution of the formula (6.9). An approximate calculation is made, for example with a normal PDF with suitable mean and variance.

This is done in the **unscented particle filter**, in which we find a Kalman like algorithm for each particle!

T224 Unscented particle filter

Une solution consiste à choisir, pour tirer les n_{r} particules initiales $\mathring{\mathrm{X}}_q[1]$, la loi normale de moyenne et variance $\mathrm{E}(\mathrm{X}[1] \mid \mathrm{Y}[1])$ et $\mathrm{Var}(\mathrm{X}[1] \mid \mathrm{Y}[1])$, et de choisir, pour tirer les n_{r} particules $\mathring{\mathrm{X}}_q[n]$, $n \geqslant 2$, la loi normale de moyenne et variance $\mathrm{E}(\mathrm{X}[n] \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_{\mathrm{Z}_q[n]}[n-1], \mathrm{Y}[n])$ et $\mathrm{Var}(\mathrm{X}[n] \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_{\mathrm{Z}_q[n]}[n-1], \mathrm{Y}[n])$. Mais ces moyennes et variances sont en général impossible à calculer. On va les approcher, en utilisant les mêmes mécanismes de transformation de l'incertitude que ceux utilisés dans les filtres de Kalman étendu et sans parfum. On obtiendra donc, *pour chaque particule*, un calcul de moyenne et variance analogue au filtre de Kalman.

D'autre part, on va utiliser une partie de ces calculs pour adapter la sélection. Il faut donc faire ces calculs avant la sélection.

T225 On dispose du modèle non linéaire (3.16) avec les hypothèses associées page 183:

$$\begin{cases} \mathrm{Y}[n] & = h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] & = f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases}$$

On procède en 2 étapes, analogues aux étapes de prédiction et correction d'un filtre de Kalman, en partant la particule parente $\mathring{\mathrm{X}}_q[n-1]$.

On calcule d'abord, pour chaque particule, une approximation $\breve{\mathrm{X}}_q^{|n-1}[n]$ de la moyenne $\mathrm{E}\left(\mathrm{X}[n] \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1\right.$
c'est-à-dire:

$$\breve{\mathrm{X}}_q^{|n-1}[n] \simeq \mathrm{E}\left(f_{n-1}(\mathrm{X}[n-1], \mathrm{V}[n-1]) \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1]\right)$$

ainsi qu'une approximation $P_q^{|n-1}[n]$ de la variance associée, par transformation de l'incertitude par la fonction $f_{n-1}(\mathring{\mathrm{X}}_q[n-1], .)$, pour $\mathrm{V}[n-1]$ centré et de variance Q_{n-1} (on remarque que l'état est fixé, c'est le nuage de particules qui prend en compte sa variation).

T226 Puis, on corrige à l'aide de $\mathrm{Y}[n]$ pour obtenir une approximation $\breve{\mathrm{X}}_q^{|n}[n]$ de $\mathrm{E}\left(\mathrm{X}[n] \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1], \mathrm{Y}[n]\right)$ ainsi qu'une approximation $P_q^{|n}[n]$ de la variance associée, par LMMSE.

Les approximations suivantes:

$$\breve{\mathrm{Y}}_q[n] \simeq \mathrm{E}\left(h_n(\mathrm{X}[n], \mathrm{W}[n]) \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1]\right)$$
$$C_{\mathrm{X},\mathrm{Y};q}[n] \simeq \mathrm{Cov}\left(\mathrm{X}[n], h_n(\mathrm{X}[n], \mathrm{W}[n]) \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1]\right)$$
$$C_{\mathrm{Y},\mathrm{Y};q}[n] \simeq \mathrm{Var}\left(h_n(\mathrm{X}[n], \mathrm{W}[n]) \mid \mathrm{X}[n-1] = \mathring{\mathrm{X}}_q[n-1]\right)$$

sont obtenues par transformation de l'incertitude par la fonction h_n, pour $\mathrm{W}[n]$ centré de variance R_n, et $\mathrm{X}[n]$ de moyenne $\breve{\mathrm{X}}_q^{|n-1}[n]$ et variance $P_q^{|n-1}[n]$ calculées à la 1re étape.

La redistribution s'effectue à l'aide de la loi discrète suivante ($g(., P)$ désigne la densité de la loi normale centrée de variance P), obtenue en approchant la formule (6.9):

$$\mathring{\omega}_q[n-1] \propto \omega_q[n-1]\, g(\mathrm{Y}[n] - \breve{\mathrm{Y}}_q[n], C_{\mathrm{Y},\mathrm{Y};q}[n])$$

T227 On peut procéder par transformée sans parfum pour transformer l'incertitude (algo. 6.4), pour obtenir l'algorithme proposé dans [ref]), mais on pourrait aussi procéder par linéarisation.

Algorithm 6.4 Unscented particle filter

Can be applied to the model $\begin{cases} \mathrm{Y}[n] = h_n(\mathrm{X}[n], \mathrm{W}[n]) \\ \mathrm{X}[n+1] = f_n(\mathrm{X}[n], \mathrm{V}[n]) \end{cases}$ with $\begin{cases} C_{\mathrm{W}[n],\mathrm{W}[n]} = R_n \\ C_{\mathrm{V}[n],\mathrm{V}[n]} = Q_n \end{cases}$

Notation $g(., P)$ is the PDF of the zero-mean normal PDF with variance P.

Initial time $(n = 1)$

transform through h_1 $\quad \left(\breve{\mathrm{Y}}_1[1], C_{\mathrm{Y},\mathrm{Y};1}[1], C_{\mathrm{X},\mathrm{Y};1}[1]\right) = \mathrm{UT}\left(h_1, \begin{bmatrix} m_{\mathrm{X}[1]} \\ 0 \end{bmatrix}, \begin{bmatrix} C_{\mathrm{X}[1],\mathrm{X}[1]} & 0 \\ 0 & R_1 \end{bmatrix}\right)$

1st observation $\quad$ that is $\mathrm{Y}[1]$

distr. parameters $\quad \begin{cases} \breve{\mathrm{X}}_1^{|1}[1] = \breve{\mathrm{X}}_1^{|0}[1] + C_{\mathrm{X},\mathrm{Y};1}[1]\, C_{\mathrm{Y},\mathrm{Y};1}^{-1}[1]\, (\mathrm{Y}[1] - \breve{\mathrm{Y}}_1[1]) \\ P_1^{|1}[1] = P_1^{|0}[1] - C_{\mathrm{X},\mathrm{Y};1}[1]\, C_{\mathrm{Y},\mathrm{Y};1}^{-1}[1]\, C_{\mathrm{X},\mathrm{Y};1}^{\mathsf{T}}[1] \end{cases}$

sampling $(\forall q)$ $\quad$ sample $\mathring{\mathrm{X}}_q[1]$ according to the normal distribution with mean $\breve{\mathrm{X}}_1^{|1}[1]$ and variance $P_1^{|1}[1]$

weights calculation $(\forall q)$ $\quad \omega_q[1] \propto \dfrac{\rho(\mathring{\mathrm{X}}_q[1])\, p_{\mathrm{Y}[1]|\mathrm{X}[1]}(\mathrm{Y}[1], \mathring{\mathrm{X}}_q[1])}{g(\mathring{\mathrm{X}}_q[1] - \breve{\mathrm{X}}_1^{|1}[1], P_1^{|1}[1])}$ (then normalization)

estimation 1st state $\quad \hat{\mathrm{X}}^{|1}[1] = \displaystyle\sum_{q=1}^{n_{\mathrm{r}}} \omega_q[1]\, \mathring{\mathrm{X}}_q[1]$

Next times $(n \geqslant 2)$

transform through f_{n-1} $(\forall q)$ $\quad \left(\breve{\mathrm{X}}_q^{|n-1}[n], P_q^{|n-1}[n]\right) = \mathrm{UT}\left(f_{n-1}(\mathring{\mathrm{X}}_q[n-1], .), 0, Q_{n-1}\right)$

transform through h_n $(\forall q)$ $\quad \left(\breve{\mathrm{Y}}_q[n], C_{\mathrm{Y},\mathrm{Y};q}[n], C_{\mathrm{X},\mathrm{Y};q}[n]\right) = \mathrm{UT}\left(h_n, \begin{bmatrix} \breve{\mathrm{X}}_q^{|n-1}[n] \\ 0 \end{bmatrix}, \begin{bmatrix} P_q^{|n-1}[n] & 0 \\ 0 & R_n \end{bmatrix}\right)$

nth observation $\quad$ that is $\mathrm{Y}[n]$

distr. parameters $(\forall q)$ $\quad \begin{cases} \breve{\mathrm{X}}_q^{|n}[n] = \breve{\mathrm{X}}_q^{|n-1}[n] + C_{\mathrm{X},\mathrm{Y};q}[n]\, C_{\mathrm{Y},\mathrm{Y};q}^{-1}[n]\, (\mathrm{Y}[n] - \breve{\mathrm{Y}}_q[n]) \\ P_q^{|n}[n] = P_q^{|n-1}[n] - C_{\mathrm{X},\mathrm{Y};q}[n]\, C_{\mathrm{Y},\mathrm{Y};q}^{-1}[n]\, C_{\mathrm{X},\mathrm{Y};q}^{\mathsf{T}}[n] \end{cases}$

weights calculation $(\forall z)$ $\quad \mathring{\omega}_z[n-1] \propto \omega_z[n-1]\, g(\mathrm{Y}[n] - \breve{\mathrm{Y}}_z[n], C_{\mathrm{Y},\mathrm{Y};z}[n])$ (then normalization)

selection $(\forall q)$ $\quad$ sample $\mathrm{Z}_q[n]$ according to the discrete distribution $(\mathring{\omega}_z[n-1])_{1 \leqslant z \leqslant n_{\mathrm{r}}}$

mutation $(\forall q)$ $\quad$ sample $\mathring{\mathrm{X}}_q[n]$ according to the normal distr. with mean $\breve{\mathrm{X}}_{\mathrm{Z}_q[n]}^{|n}[n]$ and var. $\breve{P}_{\mathrm{Z}_q[n]}^{|n}[n]$

weights calculation $(\forall q)$ $\quad \omega_q[n] \propto \dfrac{\kappa_n(\mathring{\mathrm{X}}_q[n], \mathring{\mathrm{X}}_{\mathrm{Z}_q[n]}[n-1])\, p_{\mathrm{Y}[n]|\mathrm{X}[n]}(\mathrm{Y}[n], \mathring{\mathrm{X}}_q[n])}{g(\mathring{\mathrm{X}}_q[n] - \breve{\mathrm{X}}_{\mathrm{Z}_q[n]}^{|n}[n], P_{\mathrm{Z}_q[n]}^{|n}[n])}\, \dfrac{\omega_{\mathrm{Z}_q[n]}[n-1]}{\mathring{\omega}_{\mathrm{Z}_q[n]}[n-1]}$ (then normalization)

estimation nth state $\quad \hat{\mathrm{X}}^{|n}[n] = \displaystyle\sum_{q=1}^{n_{\mathrm{r}}} \omega_q[n]\, \mathring{\mathrm{X}}_q[n]$