Master 2 Mathématiques
Parcours recherche

Advanced probability – 36h

Clément Erignoux and Lu Xu [*]

The course's key results and notions will be indicated by ★.

Contents

[*] For any typo/question, please contact us at *clement.erignoux@inria.fr* or *lu.xu@inria.fr*. The lecture notes will be put on the webpage, http://chercheurs.lille.inria.fr/cerignou/homepage.html in the "teaching" section.

– This version was last updated on November 30, 2022 –

Preliminary

A measurable space is a set Ω equipped with a σ-algebra $\mathscr{F}$. A probability measure on Ω is a *nonnegative, countably additive* function $P : \mathscr{F} \to [0,\infty)$ such that $P(\Omega) = 1$. The triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a *probability space*.

Notation: Given a topological space $(T, \mathcal{T})$, the Borel σ-algebra $\mathscr{B}(T, \mathcal{T})$ (often $\mathscr{B}(T)$ for short) is the σ-algebra generated by all open sets in $\mathcal{T}$. When we mention $[0,\infty)$, $[a,b]$ and R^d, they are always equipped with the Borel-σ-algebras.

A *random variable* X is a measurable function $X : (\Omega, \mathscr{F}) \to \mathbb{R}^d$. When $d > 1$, X is often called a *random vector*. The (push-forward) *distribution* of X is the probability measure P_X on $\mathbb{R}^d$ defined as

$$P_X(A) := \mathbb{P}(\{\omega; X(\omega) \in A\}), \quad \forall\, A \in \mathscr{B}(\mathbb{R}^d).$$

Given a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, the *expectation* is defined as

$$\mathbb{E}[f(X)] := \int_\Omega f(X) d\mathbb{P} = \int_{\mathbb{R}^d} f d\mathbb{P}_X.$$

For $p > 0$, we denote $X \in L^p(P)$ if $\mathbb{E}[|X|^p] < \infty$. In particular, $X \in L^1(P)$ is called integrable and $X \in L^2(P)$ is called square integrable.

Given random variables $\{X_n; n \geq 1\}$ and X, we say X_n converges to X *in probability* if

$$\lim_{n\to\infty} \mathbb{P}(\{\omega; |X_n(\omega) - X(\omega)| > \delta\}) = 0, \quad \forall\, \delta > 0.$$

We say X_n converges to X *almost surely* (a.s. for short) if

$$P\left(\left\{\omega; \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Suppose that X is integrable and $\mathscr{G} \subseteq \mathscr{F}$ is a sub-σ-algebra of $\mathscr{F}$. The *conditional expectation of X given $\mathscr{G}$* is a $\mathscr{G}$-measurable function Y such that

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A], \quad \forall A \in \mathscr{G}.$$

We denote $Y = \mathbb{E}[X\,|\,\mathscr{G}]$. Observe that if X is $\mathscr{G}$-measurable, then $\mathbb{E}[X\,|\,\mathscr{G}] = X$.

1 Stochastic process

Throughout this section, let $I = [0, \infty)$ or $[a, b]$ for some $0 \le a < b$.

Definition 1.1: Stochastic process ★

A stochastic process $X = \{X_t; t \in I\}$ is a family of random variables X_t defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

REMARK: *The element $t \in I$ is interpreted as* time*.*

REMARK: *The index set I could also be multi-dimensional. For example, for a d-dimensional manifold M, $X = \{X_x; x \in M\}$ is called a* random filed*.*

REMARK: *A stochastic process X can be viewed as the map*

$$I \times \Omega \ni (t, \omega) \mapsto X_t(\omega) \in \left(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d)\right).$$

X is called measurable *if this map is measurable with respect to the product σ-algebra $\mathscr{B}(I) \otimes \mathscr{F}$ on $I \times \Omega$. All the processes appearing in this note are measurable processes.*

1.1 Sample paths

Definition 1.2: Sample path ★

For any $\omega \in \Omega$, the function $I \ni t \mapsto X_t(\omega)$ is called a sample path (trajectory, realisation) of the process X.

REMARK: *We are interested in the properties of the sample paths, e.g., (usually P-a.s.) continuity and differentiability.*

EXAMPLE: Let T_1, T_2, ... be a sequence of independent, identically distributed (i.i.d.) random variables with exponential distribution $\mathbb{P}(T_i \in [0, t]) = 1 - e^{-t}$. Let ξ_1, ξ_2, ... be a sequence of i.i.d. random variables with Bernoulli distribution $\mathbb{P}(\xi_i = \pm 1) = \frac{1}{2}$, independent of T_i's. Define the stochastic process $\{X_t; t \in [0, \infty)\}$ by

$$X_0 = 0, \quad X_t = \sum_{i=1}^{n(t)} \xi_i, \quad n(t) := \sup_{n \ge 1} \left\{ \sum_{i=1}^{n} T_i \le t \right\}.$$

It is not hard to verify that the sample paths $t \mapsto X_t$ are right-continuous with left limit exists everywhere (denoted as *RCLL* or *càdlàg* paths for short).

T_i is interpreted as the time interval between the $(i-1)$-th and i-th ring of a random clock. A marker is placed at the origin of $\mathbb{Z}$ at time $t = 0$. Each time the clock rings, the marker is moved to its left or right neighbour according to the result of a fair coin tossing. The stochastic process $X = \{X_t; t \in [0, \infty)\}$ records the location of the marker and is called a *random walk*.

Suppose that $\{X_t, t \in I\}$ and $\{Y_t; t \in I\}$ are two stochastic processes defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition 1.3: Indistinguishable processes & modification

X and Y are called *indistinguishable* if they have a.s. the same sample paths:

$$\mathbb{P}(\{\omega \in \Omega; X_t(\omega) = Y_t(\omega), \forall\, t \in I\}) = 1; \tag{1.1}$$

Y is called a *modification* (or a *version*) of X if

$$\mathbb{P}(\{\omega \in \Omega; X_t(\omega) = Y_t(\omega)\}) = 1, \quad \forall\, t \in I.$$

If two stochastic processes are indistinguishable, they are apparently modifications of each other. The inverse is not true.

Example: Let T be a random variable with exponential distribution $\mathbb{P}(0 \le T < t) = 1 - e^{-t}$. Consider two stochastic processes:

$$\{X_t := 0; t \ge 0\} \quad \text{and} \quad \{Y_t := \mathbf{1}_{\{t=T\}}; t \ge 0\}.$$

Y is a modification of X since $\mathbb{P}(X_t = Y_t) = \mathbb{P}(T \ne t) = 1$ for all $t \in \mathbb{R}_+$. However, they are not indistinguishable: $\mathbb{P}(X_t = Y_t; \forall\, t \in \mathbb{R}_+) = \mathbb{P}(T = \infty) = 0$.

Example: Define $Y = \{Y_t; t \in [0, \infty)$ by replacing $n(t)$ in (1.1) with

$$n'(t) := \sup_{n \ge 1} \left\{ \sum_{i=1}^{n} T_i < t \right\}, \quad \forall\, t \ge 0.$$

Y is a modification of X, but with LCRL (left-continuous with right limit exists everywhere) sample paths.

Exercise 1

Suppose that Y is a modification of X and the sample paths of X and Y are a.s. right-continuous. Then X and Y are indistinguishable.

Observe that each sample path of X is an element of

$$(\mathbb{R}^d)^I := \left\{ \text{all vector-valued functions } x : I \to R^d \right\}.$$

A subset of $A \subseteq (\mathbb{R}^d)^I$ is called a cylinder set, if $\exists\, m \in \mathbb{N}$, t_1, ..., $t_n \in I$, $t_1 < \ldots < t_m$, and A_1, ..., $A_m \in \mathscr{B}(\mathbb{R}^d)$ such that

$$A = \{x \in (\mathbb{R}^d)^I; x(t_1) \in A_1, \ldots, x(t_m) \in A_m\}.$$

Equip $(\mathbb{R}^d)^I$ with the cylindrical σ-algebra $\mathscr{C}$, which is the smallest σ-algebra that contains all cylinder sets. Equivalently speaking, $\mathscr{C}$ is the smallest σ-algebra that makes all coordinate maps $\{\Pi_t; t \in I\}$ measurable, where

$$\Pi_t(x) := x(t), \quad \forall\, x \in (\mathbb{R}^d)^I.$$

A stochastic process X then can be viewed as a measurable map from $(\Omega, \mathscr{F})$ to $((\mathbb{R}^d)^I, \mathscr{C})$. The corresponding distribution $\mathbb{P}$ is called the law of the X.
Remark: *Given a distribution $\mathbb{P}$ on $((\mathbb{R}^d)^I, \mathscr{C})$, define*

$$(\Omega, \mathscr{F}, \mathbb{P}) := ((\mathbb{R}^d)^I, \mathscr{C}, \mathbb{P}), \quad X_t := \Pi_t, \quad \forall\, t \in I,$$

*then the distribution of the stochastic process $\{X_t; t \in I\}$ is $\mathbb{P}$ (*exercise*). Hence, we can mention the distribution of a process without specifying the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ where it is defined. Similarly, to construct a process, it suffices to construct a distribution on the sample path space.*

Let $\mathscr{I}$ be the collection of all finite, ordered subsets of I:

$$\mathscr{I} := \{\tilde{t} = (t_1, \ldots, t_m); m \in \mathbb{N}, \{t_1, \ldots, t_m\} \in I\}. \tag{1.2}$$

Definition 1.4: Finite-dimensional distributions ★

For any $\tilde{t} \in \mathscr{I}$, $\tilde{t} = (t_1, \ldots, t_m)$, let $\mathbb{P}_{\tilde{t}}$ be the distribution on $(\mathbb{R}^d)^m$ of the random vector $(X_{t_1}, \ldots, X_{t_m})$. We call the collection

$$\{\mathbb{P}_{\tilde{t}}; \tilde{t} \in \mathscr{I}\}$$

the family of *finite-dimensional distributions* (FDDs) associated to the stochastic process $\{X_t; t \in I\}$ (or equivalently, associated to the corresponding distribution $\mathbb{P}$).

Remark: *If Y is a modification of X, they have the same family of FDDs. The inverse fails to hold. Indeed, two processes defined on different probability spaces can have the same family of FDDs.*

The family of FDDs has the following properties:

(C1) for any permutation $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$ and A_1, ..., $A_m \in \mathscr{B}(\mathbb{R}^d)$,

$$\mathbb{P}_{(t_1,\ldots,t_m)}(A_1 \times \ldots \times A_m) = \mathbb{P}_{(t_{i_1},\ldots,t_{i_m})}(A_{i_1} \times \ldots \times A_{i_m});$$

(C2) for any $A \in \mathscr{B}((\mathbb{R}^d)^{m-1})$,

$$\mathbb{P}_{(t_1,\ldots,t_m)}\left(A \times \mathbb{R}^d\right) = \mathbb{P}_{(t_1,\ldots,t_{m-1})}(A).$$

Indeed, any collection of distributions that satisfies these properties turns out to be the family of FDDs associated to some stochastic process.

Definition 1.5: Consistency

For $\tilde{t} = (t_1, \ldots, t_m) \in \mathscr{I}$, let $\mathbb{Q}_{\tilde{t}}$ be a distribution on $(\mathbb{R}^d)^m$. The family $\{\mathbb{Q}_{\tilde{t}}; \tilde{t} \in \mathscr{I}\}$ is called *consistent* if (C1)–(C2) are satisfied.

Theorem 1.1: Daniell–Kolmogorov existence theorem

Let $\{\mathbb{Q}_{\tilde{t}}; t \in \mathscr{I}\}$ be a consistent family. Then there is a probability measure $\mathbb{Q}$ on $((\mathbb{R}^d)^I, \mathscr{C})$ (and thus a process $X = \{X_t; t \in I\}$) such that $\{\mathbb{Q}_{\tilde{t}}, \tilde{t} \in \mathscr{I}\}$ is the family of FDDs associated to $\mathbb{Q}$ (and X).

We sketch the brief idea of the proof.

Step 1. For any cylinder set $C = \{x; x(t_i) \in A_i, i = 1, \ldots, m\}$ where A_1, ..., $A_m \in \mathscr{B}(\mathbb{R}^d)$ and t_1, ..., $t_m \in I$, $t_1 < \ldots < t_m$, define

$$Q(C) := \mathbb{Q}_{(t_1,\ldots,t_m)}(A_1 \times \ldots \times A_m).$$

Step 2. Verify that Q is countably additive:

$$Q\left(\bigcup_{n\geq 1} C_n\right) = \sum_{n\geq 1} Q(C_n) \quad \text{for disjoint } \{C_n\}_{n=1}^{\infty}.$$

Step 3. Extend Q to a probability measure $\mathbb{Q} : \mathscr{C} \to [0, 1]$ (Carathéodory extension theorem). Such extension is unique.

End of lecture 1

1.2 Filtration

Definition 1.6: Filtration ★

Given $(\Omega, \mathscr{F})$, $\{\mathscr{F}_t; t \in I\}$ is called a *filtration* if it is a nondecreasing family of sub-σ-algebras of $\mathscr{F}$, i.e., each $\mathscr{F}_t$ is a σ-algebra on Ω and $\mathscr{F}_t \subseteq \mathscr{F}_{t'} \subseteq \mathscr{F}$ for all t, $t' \in I$ such that $t < t'$.

Example: Given a stochastic process $\{X_t; t \in [0, \infty)\}$, let[1]

$$\mathscr{F}_t^X := \sigma\{X_s; 0 \leq s \leq t\}, \qquad \forall\, t \in [0, \infty), \tag{1.3}$$

then $\{\mathscr{F}_t^X\}$ is a filtration. It is called the natural filtration (generated by X). Note that by time t, an observer of X knows any $A \in \mathscr{F}_t^X$ has occurred or not.

Example: Given a filtration $\{\mathscr{F}_t; t \in [0, \infty)\}$, define

$$\mathscr{F}_\infty := \sigma\Big(\bigcup_{t\geq 0} \mathscr{F}_t\Big), \qquad \mathscr{N} := \{A \in \mathscr{F}_\infty; \mathbb{P}(A) = 0\}.$$

Let $\bar{\mathscr{F}}_t := \sigma(\mathscr{F}_t \cup \mathscr{N})$, then $\{\bar{\mathscr{F}}_t; t \in [0, \infty)\}$ forms a filtration.

[1] Recall that $\sigma\{X_s; 0 \leq s \leq t\}$ is the σ-algebra generated by the subsets $X_s^{-1}(A)$ for $A \in \mathscr{B}(\mathbb{R}^d)$ and $s \in [0, t]$.

Definition 1.7: Complete filtration

A filtration is *complete* if $\mathscr{F}_t = \bar{\mathscr{F}}_t$ for all t (equivalently, if $\mathscr{N} \subseteq \mathscr{F}_0$).

Example: Given a filtration $\{\mathscr{F}_t; t \in I\}$, define

$$\mathscr{F}_{t+} := \bigcap_{s>t, s\in I} \mathscr{F}_s, \quad \mathscr{F}_{t-} := \sigma\Big(\bigcup_{s<t, s\in I} \mathscr{F}_s\Big).$$

Then $\{\mathscr{F}_{t\pm}; t \in I\}$ are also filtrations and $\mathscr{F}_{t-} \subseteq \mathscr{F}_t \subseteq \mathscr{F}_{t+}$ for all $t \in I$.

Definition 1.8: Continuity of filtration

A filtration is called *right-continuous* if $\mathscr{F}_t = \mathscr{F}_{t+}$ for all t. It is called *left-continuous* if $\mathscr{F}_t = \mathscr{F}_{t-}$ for all t.

Remark: *Observe that even a process has right-continuous (even continuous) sample paths, the corresponding natural filtration may fail to be right-continuous. Let Z be a real-valued random variable and*

$$X_t := \max\{t-1, 0\}Z, \quad \forall t \in [0, \infty).$$

X has continuous sample paths, while the corresponding natural filtration is not right-continuous if Z is not a constant: $\mathscr{F}_t^X = \{\emptyset, \Omega\}$ for $t \in [0,1]$ and $\mathscr{F}_t^X = \sigma(Z)$ for $t > 1$.

Remark: *It is sometimes convenient to consider a filtration which is right-continuous and complete. Such a filtration is said to satisfy the* usual conditions *and is called an* augmented filtration*.*

Given any filtration $\{\mathscr{F}_t; t \in I\}$, we can define

$$\tilde{\bar{\mathscr{F}}}_t := \bigcap_{s>t, s\in I} \sigma(\mathscr{N} \cup \mathscr{F}_t), \quad \forall\, t \in I,$$

then $\{\tilde{\bar{\mathscr{F}}}_t; t \in I\}$ becomes an augmented filtration.

Suppose that $\{X_t; t \in I\}$ is a stochastic process defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ equipped with a filtration $\{\mathscr{F}_t; t \in I\}$.

Definition 1.9: Adapted process ★

X is *adapted* to $\{\mathscr{F}_t; t \in I\}$ ($\{\mathscr{F}_t\}$-adapted for short) if X_t is $\mathscr{F}_t$-measurable for each $t \in I$. An adapted process is denoted by $\{X_t, \mathscr{F}_t; t \in I\}$.

Some simple facts:

1. Any process X is adapted to the natural filtration $\{\mathscr{F}_t^X\}$.

2. If $\{X_t; \mathscr{F}_t\}$ is an adapted process, so is $\{X_t; \mathscr{F}_{t+}\}$.

3. If $\{X_t; \mathscr{F}_t\}$ is an adapted process and Y is a modification of X, then $\{Y_t; \bar{\mathscr{F}}_t\}$ is an adapted process.

Recall that a stochastic process is measurable if the map $(t, \omega) \to X_t(\omega)$ is $\mathscr{B}(I) \otimes \mathscr{F}$-measurable.

Definition 1.10: Progressively measurable

An adapted process $\{X_t, \mathscr{F}_t; t \in I\}$ is called *progressively measurable* if for all $t \in I$, the following map is $\mathscr{B}([0,t] \cap I) \otimes \mathscr{F}_t$-measurable:

$$([0,t] \cap I) \times \Omega \ni (s, \omega) \mapsto X_s(\omega) \in \left(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d)\right).$$

Exercise 2

If $\{X_t, \mathscr{F}_t\}$ is an adapted process such that $t \mapsto X_t(\omega)$ is right-continuous for *all* $\omega \in \Omega$, then it is progressively measurable.

1.3 Stopping time

Suppose that $\{\mathscr{F}_t; t \in [0, \infty)\}$ is a filtration on $(\Omega, \mathscr{F}, \mathbb{P})$. A random time is a random variable $T : (\Omega, \mathscr{F}) \to [0, \infty]$ that is allowed to take infinite value[2].

Definition 1.11: Stopping & optional time ★

A random time T is called an $\{\mathscr{F}_t\}$-*stopping time* if

$$\{\omega \in \Omega; T(\omega) \le t\} \in \mathscr{F}_t, \qquad \forall\, t \in [0, \infty).$$

T is called an $\{\mathscr{F}_t\}$-*optional time* if

$$\{\omega \in \Omega; T(\omega) < t\} \in \mathscr{F}_t, \qquad \forall\, t \in [0, \infty).$$

Example: Every deterministic time t is a stopping time.

Example: Given an adapted process $\{X_t, \mathscr{F}_t; t \in [0, \infty)\}$ and $\Gamma \in \mathscr{B}(\mathbb{R}^d)$, define the *hitting time*

$$H_\Gamma(\omega) := \inf\{t \ge 0; X_t(\omega) \in \Gamma\}. \tag{1.4}$$

H_Γ is an optional time if X is right-continuous and Γ is open (*exercise*).

[2] There is an $\Omega_0 \in \mathscr{F}$ such that $T = \infty$ on Ω/Ω_0 and $T\mathbf{1}_{\Omega_0}$ is a random variable.

Proposition 1.2: Basic properties

Any stopping time is optional. A random time is an $\{\mathscr{F}_t\}$-optional time if and only if it is an $\{\mathscr{F}_{t+}\}$-stopping time. In particular, optional time and stopping time coincide for right-continuous filtration.

Proof: Suppose that T is an $\{\mathscr{F}_{t+}\}$-stopping time, then $\{T \le t-\varepsilon\} \in \mathscr{F}_{(t-\varepsilon)+} \subseteq \mathscr{F}_t$ for any $\varepsilon > 0$. Hence,

$$\{T < t\} = \bigcup_{n\ge 1} \left\{T \le t - \frac{1}{n}\right\} \in \mathscr{F}_t, \quad \forall\, t \in I,$$

so T is an $\{\mathscr{F}_t\}$-optional time.

Suppose that T is an $\{\mathscr{F}_t\}$-optional time, then

$$\left\{T \le t - \frac{1}{n}\right\} \in \mathscr{F}_{t+\frac{1}{n}} \subseteq \mathscr{F}_{t+\frac{1}{m}}, \quad \forall\, m \le n.$$

Therefore,

$$\{T \le t\} = \bigcap_{n\ge m} \left\{T < t + \frac{1}{n}\right\} \in \mathscr{F}_{t+\frac{1}{m}}, \quad \forall\, m \ge 1,$$

so $\{T \le t\} \in \cup_{m\ge1}\mathscr{F}_{t+\frac{1}{m}} = \mathscr{F}_{t+}$ for all $t \in I$. Hence, T is an $\{\mathscr{F}_{t+}\}$-stopping time. □

Exercise 3

Suppose that S and T are $\{\mathscr{F}_t\}$-stopping times. Show that $S \wedge T$, $S \vee T$ and $S + T$ are $\{\mathscr{F}_t\}$-stopping times.

We shall answer two fundamental questions: what is observable up to a stopping time T, and can a process literally be *stopped* by T?

Definition 1.12: Information prior to a stopping time ★

Given an $\{\mathscr{F}_t\}$-stopping time T, define

$$\mathscr{F}_T := \{A \in \mathscr{F}; A \cap \{\omega; T(\omega) \le t\} \in \mathscr{F}_t, \forall\, t \in [0,\infty)\}. \tag{1.5}$$

Exercise 4

Verify the following:

1. $\mathscr{F}_T$ forms a σ-algebra;
2. T is $\mathscr{F}_T$-measurable;
3. if $T \equiv t$ (deterministic time), then $\mathscr{F}_T = \mathscr{F}_t$.

Proposition 1.3: Monotonicity ★

For two $\{\mathscr{F}_t\}$-stopping times S and T, if $S(\omega) \le T(\omega)$, $\forall\, \omega \in \Omega$, then $\mathscr{F}_S \subseteq \mathscr{F}_T$. The same result holds if $S \le T$, P-a.s. and $\{\mathscr{F}_t\}$ is complete.

Proof: We shall prove a stronger result: $\forall\, A \in \mathscr{F}_S$, $A \cap \{S \le T\} \in \mathscr{F}_T$. Observe that for any $t \in [0,\infty)$,

$$(A \cap \{S \le T\}) \cap \{T \le t\} = \underbrace{A \cap \{S \le t\}}_{\in \mathscr{F}_t} \cap \underbrace{\{T \le t\}}_{\in \mathscr{F}_t} \cap \{S \wedge t \le T \wedge t\}.$$

It suffice to observe that both $S \wedge t$ and $T \wedge t$ are $\mathscr{F}_t$-measurable. □

Given a stochastic process X and a random time T, define

$$X_T(\omega) := X_{T(\omega)}(\omega), \quad \forall\, \omega \in \{T < \infty\}. \tag{1.6}$$

If T is finite, i.e., $\mathbb{P}(T < \infty) = 1$, X_T is almost everywhere defined on Ω.

Exercise 5

Is X_T a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$? (If $X = \{X_t\}$ is measurable)

Proposition 1.4

Suppose that $\{X_t, \mathscr{F}_t; t \in [0,\infty)\}$ is a progressively measurable process and T is an $\{\mathscr{F}_t\}$-stopping time.

1. If T is finite, i.e., $\mathbb{P}(T < \infty) = 1$, then X_T is $\mathscr{F}_T$-measurable.
2. The process $\{X_{T\wedge t}, \mathscr{F}_t; t \in [0,\infty)\}$ is progressively measurable.

Proof: We first show the second statement. Fix an arbitrary $t > 0$. The map $[0,t] \times \Omega \ni (s,\omega) \mapsto X_{T\wedge s}(\omega)$ is the composition of two maps:

$$\begin{aligned} \varphi_1 &: [0,t] \times \Omega \to [0,t] \times \Omega, \quad && (s,\omega) \mapsto (T(\omega) \wedge s, \omega); \\ \varphi_2 &: [0,t] \times \Omega \to \mathbb{R}^d, && (s,\omega) \mapsto X_s(\omega). \end{aligned}$$

Both φ_1 and φ_2 are measurable maps, so is $\varphi_2 \circ \varphi_1$.

For the first statement, it suffices to show for any $A \in \mathscr{B}(\mathbb{R}^d)$ that

$$\{X_T \in A\} \cap \{T \le t\} \in \mathscr{F}_t, \quad \forall\, t \in [0,\infty).$$

Since $\{X_T \in A\} \cap \{T \le t\} = \{X_{T\wedge t} \in A\} \cap \{T \le t\}$, it directly follows from the second statement. □

Definition 1.13: Stopped process ★

The process $\{X_{T\wedge t}, \mathscr{F}_t; t \in [0,\infty)\}$ defined above is called a *stopped process*.

Example: Suppose that $\{\mathscr{F}_t\}$ is right-continuous and X is an $\{\mathscr{F}_t\}$-adapted process with right-continuous sample paths. The hitting time

$$T := \inf\{t \geq 0; |X_t(\omega)| > M\} \tag{1.7}$$

is then a stopping time for any $M > 0$. Hence, $\{X_{T\wedge t}, \mathscr{F}_t\}$ is an adapted process with *uniformly bounded* sample paths. It is called a *localization* (or *cut-off*) of X.

——— *End of lecture 2* ———

1.4 Martingale

In this section, let $\{X_t, \mathscr{F}_t; t \in [0,\infty)\}$ be an adapted process taking real values: $M_t \in \mathbb{R}$, $\forall\, t \in [0,\infty)$.

Definition 1.14: Martingale ★

$\{X_t, \mathscr{F}_t\}$ is called a *martingale* if $\mathbb{E}[|X_t|] < \infty$, $\forall\, t \geq 0$ and

$$\mathbb{E}[X_t \,|\, \mathscr{F}_s] = X_s, \quad \forall\, 0 \leq s \leq t.$$

Example: Given a random variable X and a filtration $\{\mathscr{F}_t; t \in [0,\infty)\}$. If $\mathbb{E}[|X|] < \infty$, $M_t := \mathbb{E}[X \,|\, \mathscr{F}_t]$ is a martingale.

Definition 1.15: Sub-martingale & super-martingale ★

$\{X_t, \mathscr{F}_t\}$ is a *sub-martingale* (respectively, a *super-martingale*) if $\mathbb{E}[|X_t|] < \infty$ for each t and $\mathbb{E}[X_t \,|\, \mathscr{F}_s] \geq M_s$ (respectively, $\mathbb{E}[M_t \,|\, \mathscr{F}_s] \leq M_s$), P-a.s. for all $0 \leq s \leq t$.

Remark: *The map $t \mapsto \mathbb{E}[X_t]$ is nondecreasing (respectively, non-increasing) if X is a sub-martingale (respectively, super-martingale).*

Remark: *If $\{X_t, \mathscr{F}_t\}$ is a sub-martingale and $t_1 \leq t_2 \leq \ldots$ is a discrete sequence, then $\{\{X_{t_n}, \mathscr{F}_{t_n}; n = 1, 2, \ldots\}$ is a discrete-time sub-martingale. Similarly for super-martingale and martingale.*

Example: Let $\{T_i; i = 1, 2, \ldots\}$ be an i.i.d. sequence of exponential random times with intensity $\lambda > 0$: $\mathbb{P}(T_i \in [0,t]) = 1 - e^{-\lambda t}$, $\forall t \geq 0$. Define

$$S_0 := 0, \quad S_n := T_1 + \ldots + T_n. \tag{1.8}$$

S_n can be viewed as the time of the n-th happening of a random event. Consider the counting process

$$N_t := \max\{n \geq 0; S_n \leq t\}, \quad \forall\, t \geq 0.$$

N_t records the times of the events happened within $[0, t]$. Let $\mathscr{F}_t := \mathscr{F}_t^N$ be the natural filtration associated with N_t. The process $\{N_t; \mathscr{F}_t; t \in [0, \infty)\}$ is called a *Poisson process* with density (or rate) λ.

Exercise 6

Show that

1. $\mathbb{P}(S_{N_s+1} > t \,|\, \mathscr{F}_s) = e^{-\lambda(t-s)}$, $\forall\, 0 \leq s < t$;
2. $\mathbb{P}(N_t - N_s = n \,|\, \mathscr{F}_s) = e^{-\lambda(t-s)} \frac{[\lambda(t-s)]^n}{n!}$, i.e., $N_t - N_s$ is a Poisson random variable with density $\lambda(t-s)$ and is independent of $\mathscr{F}_s$, $\forall\, 0 \leq s < t$;
3. $\{N_t; \mathscr{F}_t\}$ is a sub-martingale and $\{N_t - \lambda t; \mathscr{F}_t\}$ is a martingale.

Hint: suppose that we know there are n happenings up to time s, then all events in $\mathscr{F}_s$ are determined by the values of T_1, T_2, ..., T_n. In mathematical language, $\forall\, A \in \mathscr{F}_s$ and $n \geq 1$, $\exists\, \tilde{A} \in \sigma(T_1, \ldots, T_n)$ such that $A \cap \{N_s = n\} = \tilde{A} \cap \{N_s = n\}$.

Example: Given a martingale $\{M_t; \mathscr{F}_t\}$ and a convex function $\Phi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[|\Phi(M_t)| < \infty]$, $\forall\, t \in [0, \infty)$. By Jensen's inequality, $\{\Phi(M_t); \mathscr{F}_t\}$ is a sub-martingale. In particular, if $\mathbb{E}[M_t^2] < \infty$, $\forall\, t \in [0, \infty)$, then $\{M_t^2; \mathscr{F}_t\}$ is a sub-martingale.

Theorem 1.5: Doob's optional sampling theorem ★

Let $\{X_t, \mathscr{F}_t; t \in [0, \infty)\}$ be a *right-continuous* sub-martingale and $S \leq T$ be two *bounded* stopping times: $\mathbb{P}(S \leq T \leq a) = 1$ for some $a \in \mathbb{R}$. Recall the random variables X_S, X_T defined in (1.6). Then $\mathbb{E}[|X_T|] < \infty$ and

$$\mathbb{E}[X_T \,|\, \mathscr{F}_S] \geq X_S, \quad P - a.s., \tag{1.9}$$

where $\mathscr{F}_S$ is defined in Definition 1.12.

Remark: *If the sub-martingale has a final element X_∞, i.e., an $\mathscr{F}_\infty$-measurable r. v. such that $X_t \leq \mathbb{E}[X_\infty|\mathscr{F}_t]$ for all t, then* (1.9) *remains true for stopping times which are not necessarily bounded.*

Proof: First, from Exercise 2 and Proposition 1.4, X_S is $\mathscr{F}_S$-measurable and X_T is $\mathscr{F}_T$-measurable. Also recall Proposition 1.3 that $\mathscr{F}_S \subseteq \mathscr{F}_T$. To prove (1.9), we exploit a discretization method as illustrated below.

Step 1. For $k \geq 1$, define $T_k(\omega) = n2^{-k}$ if $T(\omega) \in [(n-1)2^{-k}, n2^{-k})$. For every ω, $T_k(\omega) \geq T_{k+1}(\omega)$ and $\lim_{k\to\infty} T_k(\omega) = T(\omega)$. Similar definition applies to S.

Exercise 7

Each T_k is a bounded stopping time taking values from $\{n2^{-k}; n \in \mathbb{N}\}$.

Step 2. $\{X_{n2^{-k}}, \mathscr{F}_{n2^{-k}}; n \in \mathbb{N}\}$ is a discrete-time martingale. Apply the discrete-time optional sampling theorem to conclude

$$\mathbb{E}\left[X_{T_k} | \mathscr{F}_{S_k}\right] \geq X_{S_k}, \quad \forall k \geq 1. \tag{1.10}$$

Step 3. We need to take the limit $k \to \infty$ in (1.10) to obtain (1.9). Since $\{T_k\}$ is a nondecreasing sequence such that $T_k \to T$ as $k \to \infty$ and X has right-continuous path, $X_{T_k}(\omega) \to T(\omega)$ for each $\omega \in \Omega$ and similarly for X_{S_k}. In order to guarantee the same convergence in L^1, we need the following lemma.

Lemma 1.6: Uniform integrability ★

The family $\{X_{T_k}; k \geq 1\}$ is *uniformly integrable*, i.e.,

$$\lim_{\lambda \to \infty} \sup_{k \geq 1} \mathbb{E}\left[|X_{T_k}| \mathbf{1}_{\{|X_{T_k}| > \lambda\}}\right] = 0.$$

The same result holds for $\{X_{S_k}; k \geq 1\}$.

Indeed, suppose that Lemma 1.6 holds. Observe that

$$\left|X_{T_k} - X_T\right| \leq \left||X_{T_k}| \mathbf{1}_{\{|X_{T_k}| \leq \lambda\}} - |X_T| \mathbf{1}_{\{|X_T| \leq \lambda\}}\right| + |X_{T_k}| \mathbf{1}_{\{|X_{T_k}| > \lambda\}} + |X_T| \mathbf{1}_{\{|X_T| > \lambda\}}.$$

For any $\varepsilon > 0$, by Lemma 1.6, $\exists \lambda > 0$ such that $\mathbb{E}[|X_{T_k}| \mathbf{1}_{\{|X_{T_k}| > \lambda\}}] < \varepsilon$ for all $k \geq 1$. The monotonic convergence theorem then allows us to choose this λ sufficiently large such that $\mathbb{E}[|X_T| \mathbf{1}_{\{|X_T| > \lambda\}}] < \varepsilon$. For the first term,

$$\lim_{k \to \infty} \mathbb{E}\left[\left||X_{T_k}| \mathbf{1}_{\{|X_{T_k}| \leq \lambda\}} - |X_T| \mathbf{1}_{\{|X_T| \leq \lambda\}}\right|\right] = 0,$$

by the bounded convergence theorem. Therefore, $\lim_{k \to \infty} \mathbb{E}[|X_{T_k} - X_T|] = 0$ and similarly for X_{S_k} and X_S. For $A \in \mathscr{F}_S \subseteq \mathscr{F}_{S_k}$,

$$\mathbb{E}[X_T \mathbf{1}_A] = \lim_{k \to \infty} \mathbb{E}[X_{T_k} \mathbf{1}_A] \geq \lim_{k \to \infty} \mathbb{E}[X_{S_k} \mathbf{1}_A] = \mathbb{E}[X_S \mathbf{1}_A].$$

The inequality (1.9) then follows directly.

Step 4. We are left with the proof of Lemma 1.6. For $x \in \mathbb{R}$, let $x^+ = x \vee 0$ and $x^- = -(x \wedge 0)$. Since $T < \ldots \leq T_{k+1} \leq T_k \leq \ldots \leq T_1 \leq a + \frac{1}{2}$ and both X and X^+ are sub-martingales,

$$\mathbb{P}(|X_{T_k}| > \lambda) \leq \lambda^{-1} \mathbb{E}[|X_{T_k}|] = \lambda^{-1} \mathbb{E}\left[2X_{T_k}^+ - X_{T_k}\right] \leq \lambda^{-1} \mathbb{E}\left[2X_{T_1}^+ - X_T\right]$$

for any $\lambda > 0$. Therefore,

$$\lim_{\lambda \to \infty} \sup_{k \geq 1} \mathbb{P}(X_{T_k} > \lambda) = \lim_{\lambda \to \infty} \sup_{k \geq 1} \mathbb{P}(X_{T_k} < -\lambda) = 0. \tag{1.11}$$

As X^+ is a sub-martingale and T_k decreases in k,

$$\lim_{\lambda\to\infty}\sup_{k\geq 1}\mathbb{E}\left[X^+_{T_k}\mathbf{1}_{\{X^+_{T_k}>\lambda\}}\right] \leq \lim_{\lambda\to\infty}\sup_{k\geq 1}\mathbb{E}\left[X^+_{T_1}\mathbf{1}_{\{X^+_{T_k}>\lambda\}}\right] = 0, \tag{1.12}$$

where the limit follows from (1.11). On the other hand,

$$\begin{aligned}\mathbb{E}\left[X^-_{T_k}\mathbf{1}_{\{X^-_{T_k}>\lambda\}}\right] &= -\mathbb{E}[X_{T_k}] + \mathbb{E}\left[X_{T_k}\mathbf{1}_{\{X_{T_k}\geq-\lambda\}}\right]\\ &\leq -\mathbb{E}[X_{T_k}] + \mathbb{E}\left[X_{T_\ell}\mathbf{1}_{\{X_{T_k}\geq-\lambda\}}\right]\\ &= \mathbb{E}[X_{T_\ell}] - \mathbb{E}[X_{T_k}] - \mathbb{E}\left[X_{T_\ell}\mathbf{1}_{\{X_{T_k}<-\lambda\}}\right],\end{aligned}$$

for any $\ell < k$. Since $\{\mathbb{E}[X_{T_k}]; k \geq 1\}$ forms a Cauchy sequence, given an arbitrary $\varepsilon > 0$ one can choose some $\ell = \ell_\varepsilon$ such that

$$\lim_{\lambda\to\infty}\sup_{k>\ell}\mathbb{E}\left[X^-_{T_k}\mathbf{1}_{\{X^-_{T_k}>\lambda\}}\right] \leq \varepsilon + \lim_{\lambda\to\infty}\sup_{k>\ell}\mathbb{E}\left[|X_{T_1}|\mathbf{1}_{\{X_{T_k}<-\lambda\}}\right] = \varepsilon.$$

Hence, $\lim_{\lambda\to\infty}\sup_{k\geq1}\mathbb{E}[X^-_{T_k}\mathbf{1}_{\{X^-_{T_k}>\lambda\}}] \leq \varepsilon$ and since ε is arbitrary,

$$\lim_{\lambda\to\infty}\sup_{k\geq 1}\mathbb{E}\left[X^-_{T_k}\mathbf{1}_{\{X^-_{T_k}>\lambda\}}\right] = 0. \tag{1.13}$$

Summing up (1.12) and (1.13), the desired estimate follows. □

Corollary 1.7: stopped martingale ★

If $\{M_t, \mathscr{F}_t\}$ is a right-continuous martingale and S, T are two bounded stopping times such that $\mathbb{P}(S \leq T) = 1$, then $\mathbb{E}[M_T \,|\, \mathscr{F}_S] = M_S$, P-a.s.

Proof: Only to observe that $\pm M$ are both sub-martingales. □

End of lecture 3

Let $\{M_t, \mathscr{F}_t; t \in [0,\infty)$ be a martingale such that

1. each M_t is square integrable: $\mathbb{E}[M_t^2] < \infty$, $\forall\, t \in [0,\infty)$,
2. the sample paths $t \mapsto M_t(\omega)$ are *continuous* for all $\omega \in \Omega$.

Given $t > 0$, Π is called a *partition* of $[0,T]$ if $\Pi = \{t_0, t_1, \dots, t_m\}$ such that $0 = t_0 \leq t_1 \leq \cdots \leq t_m = T$. The norm of a partition Π is defined as

$$\|\Pi\| := \max\{|t_k - t_{k-1}|; k = 1, \dots, m\}.$$

For a continuous, square integrable martingale M, $T \geq 0$ and a partition $\Pi = \{t_0, t_1, \dots, t_m\}$ of $[0,T]$, let

$$V_{[0,T]}(M,\Pi) := \sum_{k=1}^{m}(M_{t_k} - M_{t_{k-1}})^2.$$

Observe that $V_{[0,T]}(M,\Pi)$ is an integrable random variable.

Exercise 8

Show that $\mathbb{E}[V_{[0,T]}(M,\Pi)] = \mathbb{E}[M_T^2 - M_0^2]$ for any partition Π.

Theorem 1.8

For continuous, square integrable martingale M and any $T \geq 0$, $V_{[0,T]}(M,\Pi)$ converges in probability as $\|\Pi\| \to 0$.

Definition 1.16: Quadratic variation ★

The *quadratic variation* of M is the process $\{\langle M\rangle_t, \mathscr{F}_t; t \in [0,\infty)\}$ given by

$$\langle M\rangle_t := \lim_{\|\Pi\|\to 0} V_{[0,t]}(M,\Pi) \quad \text{(in probability)}, \qquad \forall\, t \in [0,\infty). \tag{1.14}$$

We sketch the proof of Theorem 1.8 for the *bounded* case: $\exists\, K \in (0,\infty)$ such that $|M_t| \leq K$, $\forall\, t \in [0,\infty)$, P-a.s. In this case, the limit holds almost surely and in L^2. The general unbounded case is then solved by the localization argument (see (1.7)) and the optional sampling theorem.

Proof: Without loss of generality, we assume $M_0 = 0$ and fix some $T > 0$. Take any sequence of partitions $\{\Pi^{(n)}; n \geq 1\}$ such that $\|\Pi^{(n)}\| \to 0$, our aim is to show that $V_{[0,T]}(M,\Pi^{(n)})$ forms a Cauchy sequence in L^2, i.e.,

$$\lim_{n,n_*\to\infty} \mathbb{E}\left[\left|V_{[0,T]}(M,\Pi^{(n)}) - V_{[0,T]}(M,\Pi^{(n_*)})\right|^2\right] = 0.$$

Step 1. For any n, $n_* \geq 1$, let $\Pi := \Pi^{(n)} \cap \Pi^{(n_*)}$. Observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|V_{[0,T]}(M,\Pi^{(n)}) - V_{[0,T]}(M,\Pi^{(n_*)})\right|^2\right]\\
&\leq \mathbb{E}\left[\left|V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi^{(n)})\right|^2\right]\\
&\quad + \mathbb{E}\left[\left|V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi^{(n_*)})\right|^2\right].
\end{aligned}$$

Since $\Pi^{(n)}$, $\Pi^{(n_*)} \subseteq \Pi$, it suffices to prove that if $\Pi \supseteq \Pi'$,

$$\lim_{\|\Pi'\|\to 0} \mathbb{E}\left[\left|V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi')\right|^2\right] = 0,$$

Step 2. Fix any partition $\Pi = \{t_0, t_1, \ldots, t_m\}$ of $[0,T]$ and let

$$n = n(t;\Pi) := \max\{k; t_k \leq t\}, \qquad \forall\, t \in [0,T].$$

Define a stochastic process $\{v_t(M,\Pi); t \in [0,T]\}$ by

$$v_t(M,\Pi) := (M_t - M_{t_n})^2 + \sum_{k=1}^{n}(M_{t_k} - M_{t_{k-1}})^2,$$

where $n = n(t;\Pi)$. Observe that $v_0(M,\Pi) \equiv 0$, $v_T(M,\Pi) = V_{[0,T]}(M,\Pi)$.

Exercise 9

$\{M_t^2 - v_t(M,\Pi), \mathscr{F}_t; t \in [0,T]\}$ is a (bounded, continuous) martingale.

Given two partition Π and Π' of $[0,T]$, by Exercise 9, $\{X_t; t \in [0,T]\}$ is a (bounded, continuous) martingale, where

$$\begin{aligned} X_t &:= v_t(M,\Pi) - v_t(M,\Pi') \\ &= \left(M_t^2 - v_t(M,\Pi')\right) - \left(M_t^2 - v_t(M,\Pi)\right), \quad \forall\, t \in [0,T]. \end{aligned}$$

Recall that $X_T = V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi')$. Exercise 8 then yields that

$$\mathbb{E}\left[\left|V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi')\right|^2\right] = \mathbb{E}\left[X_T^2\right] = \mathbb{E}[V_{[0,T]}(X,\Pi)]. \tag{1.15}$$

Step 3. Suppose that $\Pi' \subseteq \Pi = \{0 = t_0 < \cdots < t_m = T\}$ and define

$$s_k := \max\{s \in \Pi'; s \le t_k\}, \quad \forall\, k = 1, \ldots, m.$$

Notice that $0 = s_0 \le \cdots \le s_m = T$ and $\{s_k; k = 0, \ldots, m\}$ gives all points in Π' with possible duplicates. Direct calculation then shows

$$V_{[0,T]}(X,\Pi) = 4\sum_{k=1}^{m}(M_{t_{k-1}} - M_{s_{k-1}})^2(M_{t_k} - M_{t_{k-1}})^2.$$

Using Cauchy–Schwarz inequality,

$$\begin{aligned} \mathbb{E}[V_{[0,T]}(X,\Pi)] &= 4\mathbb{E}\left[\sum_{k=1}^{m}(M_{t_{k-1}} - M_{s_{k-1}})^2(M_{t_k} - M_{t_{k-1}})^2\right] \\ &\le 4\mathbb{E}\left[\sup\nolimits_k \left\{(M_{t_k} - M_{s_k})^2\right\} \textstyle\sum_{k=1}^{m}(M_{t_k} - M_{t_{k-1}})^2\right] \\ &\le 4\sqrt{\mathbb{E}\left[\sup\nolimits_k \left\{(M_{t_k} - M_{s_k})^4\right\}\right]}\sqrt{\mathbb{E}\left[V_{[0,T]}^2(M,\Pi)\right]}. \end{aligned} \tag{1.16}$$

Our last task is to estimate the L^2-norm of $V_{[0,T]}^2(M,\Pi)$.

Exercise 10

Recall that we are in the bounded case: $\exists\, K \in (0,\infty)$ such that $|M_t| \le K$, P-a.s. Show that $\mathbb{E}[V_{[0,T]}^2(M,\Pi)] \le C_K$.

By (1.15), (1.16) and Exercise 10,

$$\begin{aligned} &\lim_{\|\Pi'\|\to 0} \mathbb{E}\left[\left|V_{[0,T]}(M,\Pi) - V_{[0,T]}(M,\Pi')\right|^2\right] \\ &\le C_K \limsup_{\|\Pi'\|\to 0} \sqrt{\mathbb{E}\left[\sup\nolimits_k \left\{(M_{t_k} - M_{s_k})^4\right\}\right]} \\ &\le C_K \limsup_{\sigma\to 0} \sqrt{\mathbb{E}\left[\sup\nolimits_{t,t'\in[0,T],|t-t'|\le\sigma} \left\{(M_t - M_{t'})^4\right\}\right]}. \end{aligned}$$

Recall that $\{M_t; t \in [0, T]\}$ has bounded and continuous sample paths, the last line is 0 due to bounded convergence theorem. □

Remark: *Observe that* $\langle M\rangle_0 \equiv 0$ *and* $\{\langle M\rangle_t; t \in [0, \infty)\}$ *has continuous, nondecreasing sample paths.*

The most important property of $\langle M\rangle$ is the following.

Proposition 1.9: Martingale property ★

Given a continuous, square integrable martingale $\{M_t, \mathscr{F}_t; t \in [0, \infty)\}$, $\{M_t^2 - \langle M\rangle_t, \mathscr{F}_t; t \in [0, \infty)\}$ is a martingale. In particular, $\mathbb{E}[M_t^2 - M_0^2] = \mathbb{E}[\langle M\rangle_t]$.

Proof: We still prove for the bounded case. Fix some $0 \le s < t$. For a partition $\Pi = \{0 = t_0 \le \cdots \le t_m = t\}$ of $[0, t]$, define $\Pi_s = (\Pi \cap [0, s)) \cup \{s\}$, i.e., $\Pi_s = \{t_0, \ldots, t_k, s\}$, where k is the largest subscript such that $t_k < s$. Then Π_s is a partition of $[0, s]$ and $\|\Pi_s\| \le \|\Pi\|$. From Exercise 9,

$$\mathbb{E}\left[M_t^2 - V_{[0,t]}(M, \Pi) \,\middle|\, \mathscr{F}_s\right] = M_s^2 - V_{[0,s]}(M, \Pi_s), \quad P\text{-a.s.}$$

Take the limit $\|\Pi_s\| \le \|\Pi\| \to 0$ and use the bounded convergence theorem (since we are in the bounded case). □

Remark: *The definition of quadratic variation can be extended to all* right-continuous, *square integrable martingales. Indeed, for such a martingale M, $\langle M\rangle$ is defined as a nondecreasing,* natural *process such that*

$$\langle M\rangle_t = 0 \textit{ and } M_t^2 - \langle M\rangle_t \textit{ is a right-continuous martingale.}$$

The existence of $\langle M\rangle_t$ is guaranteed by the Doob–Mayer decomposition *(see, e.g., I. Karatzas, S. Shreve:* Brownian motion and stochastic calculus, *Section 1.4). By the same theory, the choice of $\langle M\rangle$ is unique. Nevertheless, only for continuous M we have* (1.14)*, which justify the terminology "quadratic variation".*

Given two continuous, square integrable martingales $\{M_t, \mathscr{F}_t\}$ and $\{\tilde{M}, \mathscr{F}_t\}$, observe that for each t,

$$M_t\tilde{M}_t = \frac{1}{4}\left[(M_t + \tilde{M}_t)^2 - (M_t - \tilde{M}_t)^2\right].$$

The following definition becomes natural.

Definition 1.17: Cross variation

The *cross variation* of two continuous, square integrable martingales $\{M_t, \mathscr{F}_t\}$ and $\{\tilde{M}_t, \mathscr{F}_t\}$ is defined as the process $\langle M, \tilde{M}\rangle$ is given by

$$\langle M, \tilde{M}\rangle_t := \frac{1}{4}\Big(\langle M + \tilde{M}\rangle_t - \langle M - \tilde{M}\rangle_t\Big), \quad \forall\, t \in [0, \infty).$$

When $\langle M, \tilde{M}\rangle_t \equiv 0$, P-a.s., we say M and $\tilde{M}$ are *orthogonal*.

Exercise 11

Let M, $\tilde{M}$ and M' be continuous, square integrable martingales.

1. $\langle M, \alpha\tilde{M} + \beta M'\rangle = \alpha\langle M, \tilde{M}\rangle + \beta\langle M, M'\rangle$, $\forall\, \alpha, \beta \in \mathbb{R}$;
2. $\{M_t\tilde{M}_t - \langle M, \tilde{M}\rangle_t, \mathscr{F}_t; t \in [0, \infty)\}$ is a martingale;
3. $\langle M, \tilde{M}\rangle_t^2 \le \langle M\rangle_t\langle\tilde{M}\rangle_t$.

——— *End of lecture 4* ———

1.5 Solutions to some exercises

Exercise 2

Answer : For each $k = 1, 2, ...$, define the process $\{X_s^{(k)}; s \in [0, t]\}$ by

$$X_s^{(k)} := X_0\mathbf{1}\{s = 0\} + \sum_{n=1}^{2^k} X_{\frac{nt}{2^k}}\mathbf{1}\left\{s \in \left(\frac{(n-1)t}{2^k}, \frac{nt}{2^k}\right]\right\}, \quad \forall\, s \in [0, t].$$

The function $\varphi_k(s, \omega) := X_s^{(k)}(\omega)$ is a finite sum of $\mathscr{B}([0, t])\otimes\mathscr{F}_t$-measurable functions, hence also $\mathscr{B}([0, t]) \otimes \mathscr{F}_t$-measurable. On the other hand, since each sample path of X is right-continuous,

$$\lim_{k\to\infty} \varphi_k(s, \omega) = X_s(\omega), \quad \forall\, (s, \omega) \in [0, t] \times \Omega.$$

Therefore, $(s, \omega) \mapsto X_s(\omega)$ is $\mathscr{B}([0, t]) \otimes \mathscr{F}_t$-measurable. □

Exercise 9

Answer : The boundedness and the continuity are straightforward. We hereby prove the martingale property. Take $s < t$, it suffices to verify that

$$\mathbb{E}\left[M_t^2 - v_t(M, \Pi) - (M_s^2 - v_s(M, \Pi)) \,\middle|\, \mathscr{F}_s\right] = 0. \tag{1.17}$$

Denote $n = n(t; \Pi)$ and $n' = n(s; \Pi)$,

$$v_t(M, \Pi) - v_s(M, \Pi) = (M_t - M_{t_n})^2 + \sum_{k=n'+1}^{n} (M_{t_k} - M_{t_{k-1}})^2 - (M_s - M_{t_{n'}})^2.$$

For each $n' + 2 \ge k \ge n$, $\mathscr{F}_s \subseteq \mathscr{F}_{t_{k-1}}$ and

$$\begin{aligned}\mathbb{E}\left[(M_{t_k} - M_{t_{k-1}})^2 \,\middle|\, \mathscr{F}_s\right] &= \mathbb{E}\left[\mathbb{E}[(M_{t_k} - M_{t_{k-1}})^2|\mathscr{F}_{t_{k-1}}] \,\middle|\, \mathscr{F}_s\right] \\ &= \mathbb{E}\left[\mathbb{E}[M_{t_k}^2 - M_{t_{k-1}}^2|\mathscr{F}_{t_{k-1}}] \,\middle|\, \mathscr{F}_s\right] \\ &= \mathbb{E}\left[M_{t_k}^2 - M_{t_{k-1}}^2 \,\middle|\, \mathscr{F}_s\right].\end{aligned}$$

Similarly, $\mathbb{E}[(M_t - M_{t_n})^2|\mathscr{F}_s] = \mathbb{E}[M_t^2 - M_{t_n}^2|\mathscr{F}_s]$, so that

$$\begin{aligned}
&\mathbb{E}[v_t(M,\Pi) - v_s(M,\Pi) \,|\, \mathscr{F}_s] \\
=&\,\mathbb{E}\left[M_t^2 - M_{t_{n'+1}}^2 - (M_s - M_{t_{n'}})^2 + (M_{t_{n'+1}} - M_{t_{n'}})^2 \,\middle|\, \mathscr{F}_s\right] \\
=&\,\mathbb{E}\left[M_t^2 - M_s^2 + 2(M_s - M_{t_{n'+1}})M_{t_{n'}} \,\middle|\, \mathscr{F}_s\right].
\end{aligned}$$

Since $t_{n'} \leq s$, $\mathbb{E}[(M_s - M_{t_{n'+1}})M_{t_{n'}}|\mathscr{F}_s] = M_{t_{n'}}\mathbb{E}[(M_s - M_{t_{n'+1}})|\mathscr{F}_s] = 0$ and (1.17) then follows. □

Exercise 10

Answer : Without loss of generality, assume $M_0 = 0$. We begin with the following decomposition of $V_{[0,T]}^2(M,\Pi)$:

$$\sum_{k=1}^{m}(M_{t_k} - M_{t_{k-1}})^4 + 2\sum_{k=1}^{m-1}\sum_{\ell=k+1}^{m}(M_{t_k} - M_{t_{k-1}})^2(M_{t_\ell} - M_{t_{\ell-1}})^2.$$

We check the expectation of each term. For the first term, as $|M_{t_k}| \leq K$,

$$\mathbb{E}\left[\sum_{k=1}^{m}(M_{t_k} - M_{t_{k-1}})^4\right] \leq 4K^2\mathbb{E}\left[\sum_{k=1}^{m}(M_{t_k} - M_{t_{k-1}})^2\right] \leq 4K^4.$$

Notice that the last estimate follows from Exercise 8. Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\ell=k+1}^{m}(M_{t_k} - M_{t_{k-1}})^2(M_{t_\ell} - M_{t_{\ell-1}})^2\right] \\
=&\,\mathbb{E}\left[(M_{t_k} - M_{t_{k-1}})^2\sum_{\ell=k+1}^{m}\mathbb{E}\left[(M_{t_\ell} - M_{t_{\ell-1}})^2 \,\middle|\, \mathscr{F}_{t_{k-1}}\right]\right] \\
=&\,\mathbb{E}\left[(M_{t_k} - M_{t_{k-1}})^2(M_{t_m}^2 - M_{t_k}^2)\right] \leq 2K^2\mathbb{E}\left[(M_{t_k} - M_{t_{k-1}})^2\right].
\end{aligned}$$

Summing up for $k = 1, ..., m-1$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{m-1}\sum_{\ell=k+1}^{m}(M_{t_k} - M_{t_{k-1}})^2(M_{t_\ell} - M_{t_{\ell-1}})^2\right] \\
\leq&\,2K^2\mathbb{E}\left[\sum_{k=1}^{m-1}(M_{t_k} - M_{t_{k-1}})^2\right] \leq 2K^2\mathbb{E}\left[M_{t_{m-1}}^2\right] \leq 2K^4.
\end{aligned}$$

The upper bound then follows. □

2 Brownian motion

In this section, let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. All the stochastic processes mentioned below are defined on $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition 2.1: Brownian motion ★

A (one-dimensional, standard) *Brownian motion* is an adapted process $B = \{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ satisfying the following conditions:

(B1) for $s < t$, $B_t - B_s$ is independent of $\mathscr{F}_s$,

(B2) for $s < t$, $B_t - B_s \sim \mathcal{N}(0, t - s)$,

(B3) the sample paths $t \mapsto B_t(\omega)$ are continuous, P-a.s.

We usually also require that $\mathbb{P}(B_0 = 0) = 1$.

Remark: *With some abuse of notations, we shall also speak of a Brownian motion $\{B_t, \mathscr{F}_t; t \in [0, T]\}$ for $T > 0$. It is defined similarly.*

Remark: *A stochastic process $\{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ is called a Brownian motion starting from $x \in \mathbb{R}$ if $B_t - x$ is a Brownian motion.*

Remark: *It is possible that the filtration $\{\mathscr{F}_t\}$ in the definition of a Brownian motion B does not coincide with the natural filtration $\{\mathscr{F}_t^B\}$. Indeed, $\mathscr{F}_t$ can be larger than $\mathscr{F}_t^B$, provided that (B1) holds. Nevertheless, $\{B_t, \mathscr{F}_t^B; t \in [0, \infty)\}$ is still a Brownian motion, so a Brownian motion is sometimes mentioned without the filtration being specified.*

Remark: *(B1) is equivalent to the following statement*

(B1') For any $m \in \mathbb{N}_+$ and $t_1 < t_2 < \ldots < t_m$, $B_{t_2} - B_{t_1}$, $B_{t_3} - B_{t_2}$, ..., $B_{t_m} - B_{t_{m-1}}$ are mutually independent.

Remark: *(B1) and (B2) yield that Brownian motion has* stationary, independent increments*: for t_1, ..., $t_m \in [0, \infty)$ and $h > 0$, if $(t_i, t_i + h)$ do not overlap then $\{B_{t_i+h} - B_{t_i}; i = 1, \ldots, m\}$ forms an i.i.d. family of random variables. A stochastic process with stationary, independent increments and P-a.s.* càdlàg (RCLL) sample paths *is called a* Lévy process*.*

Exercise 12

Let B be a Brownian motion. Then

$$\mathbb{E}[B_t] = 0, \quad \mathbb{E}[B_s B_t] = s \wedge t, \quad \forall\, s, t \in [0, \infty). \tag{2.1}$$

Definition 2.2: Second definition of Brownian motion

An adapted process $\{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ is a Brownian motion if and only if all its finite-dimensional distributions are normal, (2.1) holds and its sample paths are continuous, P-a.s.

Proposition 2.1

Let $\{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ be a Brownian motion. Then the following processes are also Brownian motions:

1. $\{-B_t, \mathscr{F}_t; t \in [0, \infty)\}$,
2. $\{B_{h+t} - B_h, \mathscr{G}_t = \sigma(B_s; h \leq s \leq h+1); t \in [0, \infty)\}$ for any $h \geq 0$;
3. $\{c^{-\frac{1}{2}} B_{ct}, \mathscr{G}_t = \mathscr{F}_{ct}; t \in [0, \infty)\}$ for any $c > 0$;
4. $\{B_T - B_{T-t}; \mathscr{G}_t = \sigma(B_s, T - t \leq s \leq T); t \in [0, T]\}$ for any $T > 0$;
5. $\{X_t, \mathscr{G}_t; t \in [0, \infty)\}$, where

$$X_0 := \begin{cases} 0, & t = 0, \\ tB_{t^{-1}}, & t > 0, \end{cases} \qquad \mathscr{G}_t := \begin{cases} \{\emptyset, \Omega\}, & t = 0, \\ \sigma(B_s; s \geq t^{-1}), & t > 0. \end{cases} \tag{2.2}$$

Exercise 13

Show that $\{X_t, \mathscr{G}_t; t \in [0, \infty)\}$ defined by (2.2) is a Brownian motion. (*Notice that we need to show the continuity at* $t = 0$.)

Proposition 2.2: Martingale property ★

Suppose that $\{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ is a Brownian motion. Then

1. $\{B_t, \mathscr{F}_t\}$ is a martingale;
2. $\{B_t^2 - t, \mathscr{F}_t\}$ is a martingale.

Exercise 14

Show that the quadratic variation $\langle B \rangle_t = t$, $\forall\, t \in [0, \infty)$.

Answer : It suffices to show that

$$\lim_{\|\Pi\| \to 0} \sum_{k=1}^{m} (B_{t_k} - B_{t_{k-1}})^2 = t \quad \text{in probability,}$$

where Π is a partition $0 = t_0 < t_1 < \ldots < t_m = t$. Observe that

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{k=1}^{m}(B_{t_k} - B_{t_{k-1}})^2 - (t_k - t_{k-1})\right|^2\right] &= \sum_{k=1}^{m}\mathbb{E}\left[\left((B_{t_k} - B_{t_{k-1}})^2 - (t_k - t_{k-1})\right)^2\right] \\ &= \sum_{k=1}^{m}\mathbb{E}\left[(B_{t_k} - B_{t_{k-1}})^4\right] - (t_k - t_{k-1})^2 \\ &= 2\sum_{k=1}^{m}(t_k - t_{k-1})^2 \leq 2\|\Pi\|t \to 0,\end{aligned}$$

as the norm of the partition $\|\Pi\| \to 0$. □

From (B1) and (B2), one obtains the family of FDDs of a Brownian motion as follows.

Proposition 2.3: Finite-dimensional distributions ★

For $m \in \mathbb{N}_+$ and $0 \leq t_1 < \ldots < t_m$, the vector $(B_{t_1}, \ldots, B_{t_m}) \in \mathbb{R}^m$ has the joint density function given by

$$f_{t_1,\ldots,t_m}(\mathbf{x}) = \prod_{k=1}^{m} p(t_k - t_{k-1}; x_{k-1}, x_k), \quad \forall\, \mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m,$$

where we fix $t_0 = 0$, $x_0 = 0$ and

$$p(t; x, y) := \frac{1}{\sqrt{2\pi t}}\exp\left\{-\frac{(x-y)^2}{2t}\right\}, \quad \forall\, t > 0,\ x,\ y \in \mathbb{R}.$$

2.1 Construction of Brownian motion

We now prove the existence of Brownian motion by constructing a stochastic process that satisfies (B1)–(B3) in Definition 2.1. The gross idea starts from the family of FDDs obtained in Proposition 2.3. By virtue of Theorem 1.1, we can extend it to a probability measure on $(\mathbb{R}^{[0,\infty)}, \mathscr{C})$ (and thus a stochastic process X). Finally, we need to select a modification B of X with continuous sample paths.

First, recall the set of all finite, ordered subsets $\mathscr{I}$ of $[0, \infty)$ in (1.2). For $\tilde{t} = \{t_1, \ldots, t_m\} \in \mathscr{I}$ and $A \in \mathscr{B}(\mathbb{R}^m)$, define

$$\mathbb{P}_{\tilde{t}}(A) := \int_A f_{t_{n_1},\ldots,t_{n_m}}(\mathbf{x})d\mathbf{x},$$

where $\{n_1, \ldots, n_m\}$ is a rearrangement of $\{1, \ldots, m\}$ such that $0 \leq t_{n_1} < \ldots < t_{n_m}$.

Exercise 15

Show that $\{\mathbb{P}_{\tilde{t}}; \tilde{t} \in \mathscr{I}\}$ is a consistent family (Definition 1.5).

By Theorem 1.1, we obtain a probability measure $\mathbb{P}$ on $(\mathbb{R}^{[0,\infty)}, \mathscr{C})$, and thus a process X, such that (B1) and (B2) in Definition 2.1 are satisfied. Our construction would be completed, provided that X is continuous. A straightforward idea is to prove that $\mathbb{P}$ is concentrated on continuous sample path space. However, the next exercise shows that it fails to hold.

Exercise 16

Denote by $C([0,\infty))$ the subset of $\mathbb{R}^{[0,\infty)}$ that consists of all continuous functions. Prove that $C([0,\infty)) \notin \mathscr{C}$. Indeed, the only subset of $C([0,\infty))$ which belongs to $\mathscr{C}$ is the empty set.

Answer : Let $\mathscr{A}$ be the class of all subsets $\Omega_0 \in \mathbb{R}^{[0,\infty)}$ that satisfy the following condition: $\exists \{t_k; k \in \mathbb{N}_+\}$, such that

$$\{ f \in \mathbb{R}^{[0,\infty)}; f|_{\{t_k;k\in\mathbb{N}_+\}} = \omega|_{\{t_k;k\in\mathbb{N}_+\}} \text{ for some } \omega \in \Omega_0 \} \subseteq \Omega_0.$$

We can verify that:

1. any cylinder set $\{f; (f(s_1), \ldots, f(s_m)) \in A\} \in \mathscr{A}$;

2. $\mathscr{A}$ forms a σ-algebra.

Therefore, $\mathscr{C} \subseteq \mathscr{A}$. For any $\Omega_0 \in \mathscr{A}$, if $f \in \Omega_0$, $\omega := f + \mathbf{1}_{\{t_*\}}$ also belongs to Ω_0, provided that $t_* \notin \{t_k; k \in \mathbb{N}_+\}$. Since f and ω never be continuous simultaneously, Ω_0 is not a subset of $C([0,\infty))$. □

To solve this problem, we will modify the constructed process X_t in a proper way. Notice the following estimate.

Exercise 17

There is a constant $C = C_n$ such that

$$\mathbb{E}[|X_t - X_s|^{2n}] \le C|t-s|^n, \quad \forall\, t, s \in [0,\infty). \tag{2.3}$$

The next theorem helps to select a continuous modification.

Theorem 2.4: Kolmogorov–Čentsov continuity theorem

Suppose that $\exists\, \alpha > 0$, $\beta > 0$ and $C < \infty$, such that

$$E[|X_t - X_s|^\alpha] \le C|t-s|^{1+\beta}, \quad \forall\, s, t \in [0,T]. \tag{2.4}$$

Then, there is a modification Y of X, such that for each $\gamma \in (0, \beta/\alpha)$,

$$P\left\{\omega \in \Omega; |Y_t(\omega) - Y_s(\omega)| \le C_{\gamma,T}(\omega)|t-s|^\gamma, \forall\, s, t \in [0,T]\right\} = 1,$$

with some random variable $C_{\gamma,T} = C_{\gamma,T}(\omega)$. In particular, the sample paths of Y are uniformly continuous on $[0,T]$, P-a.s.

Remark: *It is important that β has to be strictly positive. The Poisson process ($\mathbb{E}[|N_t - N_s|] = \lambda|t-s|$) gives a counterexample for $\beta = 0$.*

Remark: *From (2.3) and Chebyshev's inequality,*

$$\lim_{t'\to t} \mathbb{P}(|X_{t'} - X_t| \geq \varepsilon) \leq C \lim_{n\to\infty} \varepsilon^{-\alpha}|t'-t|^{1+\beta} = 0, \quad \forall\, \varepsilon > 0,$$

i.e., at any fixed t, $\lim_{t'\to t} X_{t'} = X_t$ in probability. Note that this is insufficient to get the global continuity of sample paths.

Proof: Without loss of generality, we prove for $T = 1$. Fix some $\gamma \in (0, \beta/\alpha)$ and define for each $n \in \mathbb{N}_+$ that

$$E_n := \left\{\omega \in \Omega; \left|X_{\frac{k+1}{2^n}} - X_{\frac{k}{2^n}}\right| \geq 2^{-\gamma n}, \exists\, k = 0, \ldots, 2^n - 1\right\}.$$

From (2.3) and we Chebyshev's inequality,

$$\mathbb{P}(E_n) \leq \sum_{k=0}^{2^n-1} P\left\{\left|X_{\frac{k+1}{2^n}} - X_{\frac{k}{2^n}}\right| \geq 2^{-\gamma n}\right\} \leq C 2^n \frac{2^{-(1+\beta)n}}{2^{-\gamma\alpha n}} = \frac{C}{2^{(\beta-\gamma\alpha)n}}.$$

Hence, $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$. By Borel–Cantelli lemma,

$$\mathbb{P}(E) = 0, \quad \text{where } E = \limsup_{n\to\infty} E_n := \bigcap_{N\in\mathbb{N}} \bigcup_{n\geq N} E_n.$$

In other words, we can find a subset $\Omega^* \subseteq \Omega$ and a random integer $N = N(\omega)$ for each $\omega \in \Omega^*$, such that

1. $\mathbb{P}(\Omega^*) = 1$;
2. for each $\omega \in \Omega^*$ and $n > N(\omega)$,

$$\left|X_{\frac{k+1}{2^n}}(\omega) - X_{\frac{k}{2^n}}(\omega)\right| < 2^{-\gamma n}, \quad \forall\, k = 0, \ldots, 2^n - 1.$$

Let $D := \{k2^{-n}; n \in \mathbb{N}_+, k = 0, \ldots, 2^n\}$ be the set of dyadic rationals in $[0,1]$. Some triangle argument then yields that

$$|X_t(\omega) - X_s(\omega)| < C(\omega)|t-s|^{\gamma}, \quad \forall\, t, s \in D,$$

with $C = C(\omega)$ determined by $N(\omega)$ and γ.

Finally, we construct a process $Y = \{Y_t; t \in [0,1]\}$ with uniformly continuous sample paths by

$$Y_t(\omega) := \begin{cases} \lim_{t_n\in D, t_n\to t} X_t(\omega), & \text{if } \omega \in \Omega^*, \\ 0, & \text{if } \omega \notin \Omega^*, \end{cases} \quad \forall\, t \in [0,1].$$

For any t and $D \ni t_n \to t$, $X_{t_n} \to X_t$ in probability and $X_{t_n} \to Y_t$, P-a.s., so $\mathbb{P}(X_t = Y_t) = 1$. Hence, Y is a modification of X. □

From (2.3) and Theorem 2.4, for each $T > 0$ we define $\{Y_t^{(T)}; t \in [0,T]\}$ with uniformly continuous sample paths. The next exercise extends the definition to $[0,\infty)$ and completes the construction of Brownian motion. Indeed, it provides a stronger statement on the Höder continuity of the sample paths.

Exercise 18

A function $f : [0, \infty) \to \mathbb{R}$ is called *Hölder continuous* with exponent γ, or simply *γ-Hölder continuous*, if

$$|f(t) - f(s)| \le C|t - s|^{\gamma}, \quad \forall\, t, s \in [0, \infty).$$

It is called *locally γ-Hölder continuous*, if for any $[a, b] \subseteq [0, \infty)$,

$$|f(t) - f(s)| \le C_{a,b}|t - s|^{\gamma}, \quad \forall\, t, s \in [a, b].$$

Show that $\{X_t; t \in [0, \infty)\}$ has a modification $\{B_t; t \in [0, \infty)\}$ with locally γ-Hölder continuous sample paths for any $\gamma < \frac{1}{2}$.

Remark: *The Hölder exponent $(\frac{1}{2} - \varepsilon)$ is optimal for Brownian motion.*

Answer : Recall that for each $T > 0$, we obtained $\{Y_t^{(T)}; t \in [0, T]\}$ with γ-Hölder continuous sample paths, $\forall\, \gamma < \frac{1}{2}$. Since $Y^{(T)}$ is a modification of X,

$$\mathbb{P}(\Omega_T) = 1, \quad \Omega_T := \left\{\omega; Y_t^{(T)}(\omega) = X_t(\omega), \forall\, t \in [0, T] \cap \mathbb{Q}\right\}.$$

Hence, we obtain a set $\Omega_* := \cap_{T \in \mathbb{N}_+} \Omega_T$ such that $\mathbb{P}(\Omega_*) = 1$ and for any two positive integers T and T',

$$Y_t^{(T)}(\omega) = Y_t^{(T')}(\omega), \quad \forall\, \omega \in \Omega_*, \forall\, t \in [0, T \wedge T'] \cap \mathbb{Q}.$$

Noting that both $Y^{(T)}$ and $Y^{(T')}$ are continuous on $[0, T \wedge T']$, it leads to

$$Y_t^{(T)}(\omega) = Y_t^{(T')}(\omega), \quad \forall\, \omega \in \Omega_*, \forall\, t \in [0, T \wedge T'].$$

For $\omega \in \Omega_*$, define $B_t(\omega) := Y_t^{([t]+1)}(\omega)$ where $[t]$ stands for the greatest integer less than or equal to t. For $\omega \notin \Omega_*$, define $B_t(\omega) := 0$. □

2.2 The Brownian sample paths

Reading material: Probability: theory and examples, 5th edition, Sections 7.1 & 7.2, Cambridge University Press., by R. Durrett.

Suppose that $B = \{B_t, \mathscr{F}_t; t \in [0, \infty)\}$ is a Brownian motion. We call its sample path $t \mapsto B_t(\omega)$ a *Brownian sample path* or *Brownian paths*.

Proposition 2.5: Nowhere Lipschitz or differentiable ★

Brownian sample paths are not Lipschitz-continuous at any $t \in [0, \infty)$, P-a.s. Consequently, the Brownian sample paths are nowhere differentiable.

Proof: *Durrett, Theorem 7.1.6.* □

Recall the natural filtration $\{\mathscr{F}^B_t; t \in [0,\infty)\}$ generated by $\{B_t; t \in [0,\infty)\}$ and its right-continuous modification

$$\mathscr{F}^B_{t+} := \sigma\left(\bigcap_{s>t} \mathscr{F}^B_t\right), \quad \forall\, t \in [0,\infty).$$

Proposition 2.6: Blumenthal's 0-1 law

$\mathscr{F}^B_{0+}$ is degenerated: $\forall\, A \in \mathscr{F}^B_{0+}$, $\mathbb{P}(A) = 0$ or 1.

Proof: *Durrett, Theorem 7.2.1 and 7.2.2*. The proof exploits the *Markov property*, which will be introduced in the following sections. □

Proposition 2.7: 0-1 law for the tail field

Let $\mathscr{T}_t := \sigma(B_s; s \in [t,\infty))$ for $t \in [0,\infty)$. The *tail σ-filed* $\mathscr{T} := \cap_{t\in[0,\infty)} \mathscr{T}_t$ is also degenerated: $\forall\, A \in \mathscr{T}$, $\mathbb{P}(A) = 0$ or 1.

Proof: *Durrett, Theorem 7.2.7*. Direct consequence of Proposition 2.6 and the last assertion of Proposition 2.1. □

The 0-1 laws are widely used to prove some sample path properties. The following is a typical example.

Corollary 2.8

For P-a.s. Brownian sample paths,

$$\limsup_{t\to\infty} t^{-\frac{1}{2}} B_t = +\infty, \qquad \liminf_{t\to\infty} t^{-\frac{1}{2}} B_t = -\infty.$$

End of lecture 5

2.3 Donsker's theorem

Reading material: Probability: theory and examples, 5th edition, Sections 8.1 & 7.2, Cambridge University Press., by R. Durrett.

We built the Brownian motion on the sample path space $(\mathbb{R}^{[0,\infty)}, \mathscr{C})$, i.e., the space of all real-valued functions on $[0,\infty)$ with the cylinder σ-algebra. As the Brownian samples are continuous, the "canonical" space for Brownian motion should be $C_+ := C([0,\infty))$.

Recall that C_+ is a complete, separable metric space under the metric

$$d(f,g) := \sum_{n=1}^{\infty} \frac{1}{2^n} \sup_{t\in[0,n]} \{|f(t) - g(t)| \wedge 1\}, \quad \forall\, f,g \in C_+.$$

Let $\mathscr{B}(C_+)$ be the Borel σ-algebra generated by the open sets in the corresponding topology. $\mathscr{B}(C_+)$ coincides with the *cylinder σ-algebra* on C_+, i.e., the smallest σ-algebra containing all cylinder sets

$$\{\omega \in C_+; (\omega(t_1), \dots, \omega(t_m)) \in A\},$$

where $m \in \mathbb{N}_+$, t_1, ..., $t_m \in [0, \infty)$, $A \in \mathscr{B}(\mathbb{R}^m)$.

Let $\{X_i; i \in \mathbb{N}_+\}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = 1$. Define $S_0 = 0$, $S_n = S_{n-1} + X_n$ for $n \in \mathbb{N}_+$. The central limit theorem yields that $(\sqrt{n})^{-1} S_n$ converges weakly, as $n \to \infty$, to the standard normal distribution.

For each $n \in \mathbb{N}_+$, define

$$Y_t^n := \frac{1}{\sqrt{n}}(S_{[nt]} + (nt - [nt])X_{[nt]+1}), \quad \forall\, t \in [0, \infty),$$

where $[t]$ stands for the greatest integer less than or equal to t. Observing that $t \mapsto Y_t^n(\omega)$ is continuous, Y^n can be viewed as a function defined on Ω and taking values from C_+. Furthermore, the function is measurable:

$$Y^n : (\Omega, \mathscr{F}, \mathbb{P}) \to (C_+, \mathscr{B}(C_+)).$$

Denote by $\mathbb{Q}^n$ the probability measure of Y^n on $(C_+, \mathscr{B}(C_+))$.

Theorem 2.9: Donsker's invariance principle

$\{\mathbb{Q}^n, n \in \mathbb{N}_+\}$ converges weakly to a measure $\mathbb{Q}$ on $(C_+, \mathscr{B}(C_+))$, such that the coordinate mapping process $\{W_t, \mathscr{F}_t^W; t \in [0, \infty)\}$ given by

$$W_t(\omega) := \omega(t), \quad \forall\, \omega \in C_+,$$

is a Brownian motion.

Definition 2.3: Wiener measure

The probability measure $\mathbb{Q}$ on $(C+, \mathscr{B}(C_+))$ is called *Wiener measure*.

2.4 Gaussian process

Recall that a real-valued random variable X is called a Gaussian variable with center $\mu \in \mathbb{R}$ and covariance $\sigma^2 > 0$ if it has the density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad \forall\, x \in \mathbb{R}.$$

We simply $X \sim \mathcal{N}(\mu, \sigma^2)$. In particular, X is called a standard Gaussian variable when $\mu = 0$ and $\sigma^2 = 1$.

Proposition 2.10: Gaussian variable ★

Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_*, \sigma_*^2)$. Then

1. Normalisation $\sigma^{-1}(X - \mu) \sim \mathcal{N}(0, 1)$;
2. For each $n \in \mathbb{N}$, $\mathbb{E}[(X-\mu)^{2n+1}] = 0$, $\mathbb{E}[(X-\mu)^{2n}] = \frac{(2n)!}{2^n n!}\sigma^{2n}$;
3. The characteristic function $\varphi_X(r) := \mathbb{E}[e^{irX}] = \exp\{i\mu r - \frac{\sigma^2 r^2}{2}\}$;
4. X and Y are independent if and only if they are uncorrelated: $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0$;
5. If X and Y are independent, then $X + Y \sim \mathcal{N}(\mu + \mu_*, \sigma^2 + \sigma_*^2)$;
6. If $X \sim \mathcal{N}(0, 1)$, then $\mathbb{P}(X \geq a) \leq \frac{1}{\sqrt{2\pi a^2}} \exp\{-\frac{a^2}{2}\}$ for $a > 0$.

Similarly, a $\mathbb{R}^d$-valued random variable X is called a Gaussian vector (multi-dimensional Gaussian variable) if there is an independent family of standard Gaussian variables $\{Y_1, \dots, Y_n\}$, a deterministic vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$, and a matrix $\Sigma = (\sigma_{jk})_{d\times n}$, such that

$$X = \Sigma \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{d1} & \dots & \sigma_{dn} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}.$$

The following observation is straightforward:

$$\mathbb{E}[X] = \boldsymbol{\mu}, \quad \mathbb{E}\left[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T\right] = \Sigma\Sigma^T := \Gamma.$$

Hence, we call Γ the covariance matrix and denote $X \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma)$.

Proposition 2.11: Gaussian vector ★

Suppose that $X \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma)$. Then

1. Γ is a symmetric, positive semi-definite matrix;
2. When Γ is positive definite ($\det \Gamma \neq 0$), the density function of X is

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \Gamma}} \exp\left\{-\frac{1}{2}\left\langle \mathbf{x} - \boldsymbol{\mu}, \Gamma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\rangle\right\}, \quad \forall\, \mathbf{x} \in \mathbb{R}^d;$$

3. When Γ is singular ($\det \Gamma = 0$), X is called degenerated and does not have a density function.

Remark: *If constants are adopted as degenerated Gaussian variables ($\sigma = 0$), then $X = (X_1, \dots, X_d)$ is a Gaussian vector if and only if for all $(a_1, \dots, a_d) \in \mathbb{R}^d$, the linear combination $a_1X_1 + \dots + a_dX_d$ is a Gaussian variable.*

Definition 2.4: Gaussian process ★

A stochastic process $\{X_t; t \in I\}$ is called a Gaussian process if for all t_1, ..., $t_m \in I$, $(X_{t_1}, \dots, X_{t_m})$ is a Gaussian vector, i.e.,

$$a_1 X_{t_1} + a_2 X_{t_2} + \cdots + a_m X_{t_m}$$

is a Gaussian variable (or constant) for all $\mathbf{a} = (a_1, \dots, a_m) \in \mathbb{R}^m$.

Example: $X_t := \cos(t)\xi_1 + \sin(t)\xi_2$ with independent Gaussian variables ξ_1, ξ_2.

Example: The *Brownian bridge* $\{X_t := B_T - TB_1; t \in [0, T]\}$, where $\{B_t; t \in [0, T]\}$ is a Brownian motion. Observe that $X_0 = X_T = 0$.

Given a Gaussian process $\{X_t; t \in [0, \infty)\}$, we define the expectation $m : I \to \mathbb{R}^d$ and the auto-covariance $C : I^2 \to \mathbb{R}^{d \times d}$ by

$$m(t) := \mathbb{E}[X_t], \quad C(t, s) := \mathbb{E}\left[(X_t - \mu(t))(X_s - \mu(s))^T\right], \quad \forall\, t, s \in I.$$

Remark: *Note that the family of FDDs, and thus the law, of a Gaussian process X is completely determined by $m(\cdot)$ and $C(\cdot, \cdot)$.*

3 Introduction to stochastic integration

We give here an introduction to stochastic integration. We refer to the following book for a complete reference.

Reading material: Stochastic Integration and Differential Equations , by Philip E. Protter.

The latter is an exhaustive reference, we give here a simple introduction to stochastic integration using a more explicit, but limited, tool, namely the Wiener integral.

3.1 Wiener integral

For $T > 0$, let $f \in L^2(0,T)$. We construct the integration $\int_0^T f(t)dB_t$ where $(B_t, \mathscr{F}_t; t \in [0,\infty))$ is a Brownian motion by the following steps.

Step 1. For $f(t) = \mathbf{1}_{(a,b]}(t)$, $0 \le a < b \le T$, it is nature to define the integration as a random variable $I(f) : (\Omega, \mathscr{F}, \mathbb{P}) \to \mathbb{R}$ by

$$I(f) = \int_0^T f(t)dB_t := B_b - B_a, \quad \forall\, \omega \in \Omega.$$

Observe that $I(f)$ is "pathwisely" defined and

$$\mathbb{E}[I(f)] = 0, \quad \mathbb{E}[I^2(f)] = b - a = \int_0^T f^2(t)dt.$$

This definition can be easily extended to step functions with the form

$$f(t) = f_0 \mathbf{1}_{\{0\}}(t) + \sum_{k=1}^{m} f_k \mathbf{1}_{((t_{k-1},t_k]}(t), \quad \forall\, t \in [0,T], \tag{3.1}$$

where $m \in \mathbb{N}_+$, $0 = t_0 < \ldots < t_m = T$ and $f_i \in \mathbb{R}$. Define

$$I(f) = \int_0^T f(t)dB_t := \sum_{k=1}^{m} (B_{t_k} - B_{t_{k-1}}) f_i, \quad \forall\, \omega \in \Omega.$$

Denote by $\mathcal{E} = \mathcal{E}([0,T])$ the class of all step functions in (3.1). Recall that $L^2(P)$ stands for all square integrable random variables on $(\Omega, \mathscr{F}, \mathbb{P})$.

Proposition 3.1

$I(f)$ defined above satisfies the following conditions.

1. I is linear: $\forall\, \alpha, \beta \in \mathbb{R}$ and $f, g \in \mathcal{E}$,

$$I(\alpha f + \beta g) = \alpha I(f) + \beta I(g).$$

2. $I(f)$ is a Gaussian variable and $\mathbb{E}[I(f)] = 0$. Furthermore, $I : \mathcal{E} \to L^2(P)$

is an isometry if $\mathcal{E}$ is embedded into $L^2([0,T])$:

$$\|I(f)\|_{L^2(P)} = \|f\|_{L^2([0,T])}, \quad \forall f \in \mathcal{E}.$$

Proof: The first assertion is trivial. For the second one, only to see that

$$\mathbb{E}[I^2(f)] = \sum_{k=1}^{m} f_i^2 \mathbb{E}\left[(B_{t_k} - B_{t_{k-1}})^2\right] = \sum_{k=1}^{m} (t_k - t_{k-1}) f_i^2 = \int_0^T f^2(t)dt$$

for each step function f. □

Exercise 19

The class of step functions $\mathcal{E}$ is dense in $L^2([0,T])$.

Given an $f \in L^2([0,T])$, pick a sequence $f_n \in \mathcal{E}$ such that $\|f_n - f\|_{L^2([0,T])} \to 0$ as $n \to \infty$. The argument above leads to the natural definition

$$I(f) = \int_0^T f(t)dB_t := \lim_{n\to\infty} I(f_n),$$

where the limit is in $L^2(P)$. It is called the *Wiener integral* of f (with respect to the Brownian motion B).

Remark: *For each $f \in L^2([0,T])$, $I(f)$ is a Gaussian variable, $\mathbb{E}[I(f)] = 0$ and $I : L^2([0,T]) \to L^2(P)$ inherits the linearity and the isometric formula.*

Remark: *Using $\langle B\rangle_t = t$, the isometric formula can be written as*

$$\mathbb{E}[I^2(f)] = \int_0^T f^2(t)d\langle B\rangle_t.$$

Remark: *For $f \in L^2([0,T])$ and $t \in [0,T]$, define*

$$I_t(f) = \int_0^t f(s)dB_s := \int_0^T f(s)\mathbf{1}_{[0,t]}(s)ds.$$

$\{I_t(f), \mathscr{F}_t; t \in [0,T]\}$ *is a* continuous, square integrable martingale,

$$\langle I(f)\rangle_t = \int_0^t f^2(s)d\langle B\rangle_s, \quad \forall t \in [0,T].$$

It is a special case of the Itô isometry. *Similarly,*

$$\langle I(f), I(g)\rangle_t = \int_0^t f(s)g(s)d\langle B\rangle_s, \quad \forall t \in [0,T],$$

for f, $g \in L^2([0,T])$.

3.2 Examples of stochastic differential equation

Suppose that $\{B_t, \mathscr{F}_t; t \in [0,\infty)\}$ is a Brownian motion defined on $(\Omega, \mathscr{F}, \mathbb{P})$. With the definition of Wiener integral, we are allowed to consider some ordinary differential equations including dB_t.

Formally consider a stochastic differential equation

$$dX_t = b(t, X_t)dt + \sigma(t)dB_t, \quad X_0 = \xi, \tag{3.2}$$

where $b : \mathbb{R}^2 \to \mathbb{R}$, $\sigma : \mathbb{R} \to \mathbb{R}$ are some nice deterministic functions and ξ is a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$. The solution is interpreted as a stochastic process $\{X_t; t \geq 0\}$ such that

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s)dB_s, \quad \forall\, t \in [0,\infty).$$

Observe that the first integral in the right-hand side above is defined for each sample path $\omega \in \Omega$, while the second one should be viewed as the Wiener integral, provided that $\int_0^t \sigma^2(s)ds < \infty$ for all $t \in [0,\infty)$.

Theorem 3.2: Existence and uniqueness of solution

Assume that b is *uniformly Lipschitz continuous* and has *linear growth*: $\exists\, K > 0$, such that

$$|b(t,x) - b(t,y)| \leq K|x-y|, \quad |b(t,x)| \leq K(1+|x|),$$

for all $t \geq 0$ and x, $y \in \mathbb{R}$. Also let ξ be a square integrable random variable that is independent of $\mathscr{F}^B_\infty$. Then, (3.2) has a solution $\{X_t; t \geq 0\}$ that is P-a.s. continuous, $\mathbb{E}[\int_0^t |X_s|^2 ds] < \infty$ for all $t > 0$. Furthermore, X_t is $\{\mathscr{F}^{\xi,B}_t\}$-adapted, where

$$\mathscr{F}^{\xi,B}_t := \sigma(\xi, B_s; 0 \leq s \leq t), \quad \forall\, t \in [0,\infty).$$

The solution is unique in indistinguishable sense.

Example: For deterministic constants λ, $\sigma > 0$ and $\xi \in \mathbb{R}$,

$$dX_t = -\lambda X_t dt + \sigma dB_t, \quad X_0 = \xi,$$

has the explicit solution

$$X_t = e^{-\lambda t}\left(\xi + \sigma \int_0^t e^{\lambda s} dB_s\right), \quad \forall\, t \in [0,\infty).$$

Example: The Brownian bridge $\{X_t; t \in [0,1]\}$ is the solution to

$$dX_t = -\frac{X_t}{1-t}dt + dB_t, \quad X_0 = 0.$$

End of lecture 6

4 Markov processes

Reading material: Markov Processes, by James R. Kirkwood.

In this section, we introduce the notion of Markov processes, a very useful type of stochastic process that can be used to model a wide variety of objects in many different fields (e.g. biology, economics, physics). We first give a quick reminder on Markov chains, which are discrete time versions of Markov processes. Throughout this section, we fix a (finite or) countable set E, to be the *state space* for our random variables, and a fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$ on which they are all defined.

4.1 Discrete time Markov chains

4.1.1 Transition matrix, Markov chain

We start by a brief reminder on Markov chains.

Definition 4.1: Markov chain, transition matrix ★

A sequence of E-valued random variables $(X_n)_{n\in\mathbb{N}}$, is a *Markov chain* if for any integer $n \geq 0$, and for any $e_0, \dots, e_{n+1} \in E$, one has

$$\mathbb{P}(X_{n+1} = e_{n+1} \mid X_0 = e_0, \dots, X_n = e_n) = \mathbb{P}(X_{n+1} = e_{n+1} \mid X_n = e_n). \tag{4.1}$$

Property (4.1) is called *Markov property*. The set E is called the Markov chain's *state space*.
A Markov chain is called homogeneous if for any $n \in \mathbb{N}$, any $e, e' \in \mathbb{E}$

$$P_n(e, e') := \mathbb{P}(X_{n+1} = e' \mid X_n = e) = \mathbb{P}(X_1 = e' \mid X_0 = e)$$

does not depend on n. The matrix P is then called the Markov chain's *transition matrix*.

In other words, a Markov chain depends on the past *only through the present.* In the remainder of this subsection, $(X_n)_{n\in\mathbb{N}}$ designates a homogeneous Markov chain with transition matrix P.

Example: A discrete-time random walk on $\mathbb{Z}$ is defined by $X_0 = 0$, and for any $n \geq 0$,

$$X_{n+1} = \begin{cases} X_n - 1 & \text{with probability } 1/2 \\ X_n + 1 & \text{w.p. } 1/2 \end{cases}.$$

This process is a *Markov chain on the state space* $E := \mathbb{Z}$*. It's transition matrix is the doubly infinite matrix* P *with coordinates in* $\mathbb{Z}$ *given by*

$$\mathbb{P}(i, j) = \frac{1}{2}\mathbf{1}_{\{|i-j|=1\}}.$$

Such processes can model many things, from gaz diffusion to the stock market.

Proposition 4.1: Properties of the transition matrix

The coefficients of the transition matrix are in $[0, 1]$, and their sum over each row adds up to 1, i.e. $\forall e \in E$,

$$\sum_{e' \in E} \mathbb{P}(e, e') = 1.$$

An homogeneous Markov chain's transition Matrix is usually represented by a graph, with

- each vertice of the graph representing a state $e \in E$.
- An edge linking vertice $e \in \mathbb{E}$ with vertice $e' \in E$ is labelled with their transition probability $p := \mathbb{P}(e, e')$.

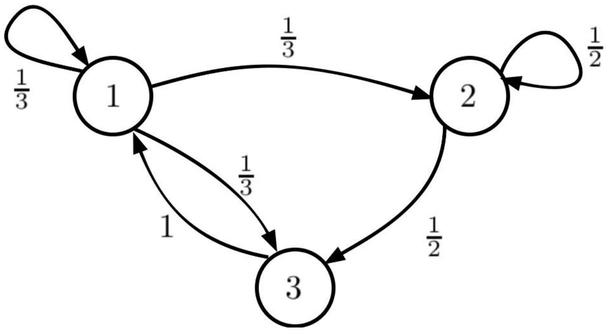
Exercise 20

Consider the Markov chain $(X_n)_{n\in\mathbb{N}}$ on $E = \{1, 2, 3\}$, with transition matrix

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$

Draw the graph representing this Markov chain.

Answer :



□

Proposition 4.2: Probability of a sample path ★

For any finite sequence of states $e_0, \dots, e_n \in E$,

$$\mathbb{P}(X_0 = e_0, \dots, X_n = e_n) = \mathbb{P}(X_0 = e_0) \prod_{k=1}^{n} \mathbb{P}(e_{k-1}, e_k).$$

For any $e \in E$, define $\mu_n(e) = \mathbb{P}(X_n = e)$, which allows to represent the distribution of X_n as a (row) vector $\mu_n = \{\mu_n(e),\ e \in E\}$.

Proposition 4.3: Matrix form for μ_n ★

For any time n, we have the matrix identities

$$\mu_{n+1} = \mu_n P \quad \text{and} \quad \mu_n = \mu_0 P^n.$$

In particular, P^k is the transition matrix of the Markov chain $(X_{nk})_{n\in\mathbb{N}}$

4.1.2 States of a Markov chain

Definition 4.2: Communicating states, irreducible chain ★

A state e' is *accessible* from another state e if there exists a finite sequence $e_0 := e,\ e_1,\ e_2,\ \dots, e_{n-1},\ e_n := e'$ such that $\forall i \in \{0, \dots n-1\}$, $\mathbb{P}(e_i, e_{i+1}) > 0$. We denote it by $e \to e'$. Note that by convention, for any e, $e \to e$, and that

$$e \to e' \quad \Leftrightarrow \quad \sup_{n\in\mathbb{N}}\{P^n(e, e')\} > 0.$$

If $e \to e'$ and $e' \to e$, we say that e and e' *communicate*, denoted by $e \leftrightarrow e'$. This is an *equivalence relation on E*, and defines a partition of E in *equivalence classes of communicating states*. If there is only on equivalence class, i.e. $\forall e, e' \in E$, $e \leftrightarrow e'$, the Markov chain is called *irreducible*.

Definition 4.3: Recurrent, transient, and aperiodic states ★

A state e is called *recurrent* if $\mathbb{P}(\exists n \geq 1 : X_n = e \mid X_0 = e) = 1$, and *transient* otherwise.
We call period of a state e the (maybe infinite) integer

$$d(e) = GCD\{k \geq 1 \ : \ P^k(e, e) > 0\}.$$

If $d(e) = 1$, we call e *aperiodic*.

Proposition 4.4: Recurrent and transient classes

Recurrence, transience, and period are class properties, meaning

- $\{e$ is recurrent and $e \leftrightarrow e'\} \ \Rightarrow\ e'$ is recurrent.
- $\{e$ is transient and $e \leftrightarrow e'\} \ \Rightarrow\ e'$ is transient.
- $\{d(e) = k$ and $e \leftrightarrow e'\} \ \Rightarrow\ d(e') = k$.

Furthermore, if the state space E is finite, there exists at least one recurrent state.

4.1.3 Invariant measures

Definition 4.4: Invariant measure ★

A probability distribution π on E seen as a row vector is called

- *invariant* with respect to P if $\pi P = \pi$,
- *reversible* w.r.t. P if for any $e,\ e' \in E$, $\pi(e)\mathbb{P}(e, e') = \pi(e')\mathbb{P}(e', e)$.

Any reversible probability measure is also invariant.

Proposition 4.5: Existence of invariant measures ★

An irreducible Markov chain has *at most one* invariant probability measure. If furthermore, its state space is finite, then it has *exactly* one invariant probability measure.

This notion is fundamental, because it gives the long-time distribution of the Markov chain. Indeed, assuming the existence of a limit,

$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \mu_{n+1} = \lim_{n\to\infty} \mu_n P,$$

therefore $\pi = \lim_{n\to\infty} \mu_n$ must be the invariant measure. More precisely:

Theorem 4.6: Convergence in law ★

Assume that $(X_n)_{n\in\mathbb{N}}$ is *irreducible* and *aperiodic*.

- If P has an invariant probability distribution π, then for any initial state e_0, and for any $e \in \mathbb{E}$,

  $$\mathbb{P}(X_n = e \mid X_0 = e_0) \underset{n\to\infty}{\longrightarrow} \pi(e).$$

- If P does not have an invariant probability distribution, then for any state e_0, and for any $e \in E$

  $$\mathbb{P}(X_n = e \mid X_0 = e_0) \underset{n\to\infty}{\longrightarrow} 0.$$

  Note that according to Proposition 4.3, $\mathbb{P}(X_n = e \mid X_0 = e_0) = [P^n](e_0, e)$.

Theorem 4.7: Ergodic theorem, weak law of large numbers

Assume that $(X_n)_{n\in\mathbb{N}}$ is *irreducible* and *aperiodic*, with invariant probability distribution π. For any function f integrable w.r.t. π (i.e. satisfying $\sum_{e\in E}|f(e)|\pi(e) < \infty$), we have

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=0}^{N-1} f(X_n) \xrightarrow[N\to\infty]{} \sum_{e\in E} f(e)\pi(e)\right) = 1.$$

4.2 General tools

Our goal is now to define Markov processes on countable sets. To do so, we will need a few key notions to keep their construction as simple as possible.

4.2.1 Exponential distribution

We first start by recalling the definition of the exponential distribution. It will play the key role in the construction of Markov processes, since all waiting time between successive jumps of the process will have exponential distribution.

Definition 4.5: Exponential variable ★

A random variable T taking values in $\mathbb{R}$ has *exponential distribution with parameter* $\lambda > 0$, noted $T \sim Exp(\lambda)$, if its density is given by

$$f_T(t) = \lambda e^{-\lambda t}\mathbf{1}_{[0,+\infty)}.$$

In particular, an exponential variable is characterized by its cumulative distribution function (fr: *fonction de répartition*)

$$F_T(t) = \mathbb{P}(T \le t) = 1 - e^{-\lambda t}$$

The expectation of T is then $\mathbb{E}(T) = 1/\lambda$.

Proposition 4.8: Multiplication by a constant

If $T \sim Exp(\lambda)$ has exponential distribution with parameter $\lambda > 0$, and if $\alpha > 0$, then αT has exponential distribution with parameter λ/α. In particular, λT is an $Exp(1)$ variable.

Proof: Given the characterization of distributions by their cumulative distribution function, this is immediate. □

Remark: *By convention, one can define exponential variables with parameter* 0 *by*

$$X \sim Exp(0) \quad \Leftrightarrow \quad \mathbb{P}(X = +\infty) = 1.$$

This is coherent with the definition and the previous property, since we can see X as the limit as $\lambda \searrow 0$ of $Exp(1)/\lambda$.

Proposition 4.9: Absence of memory ★

The exponential distribution is memoryless, in the sense that for any positive time t, we have

$$\{T - t \text{ conditioned on } T \geq t\} \overset{(d)}{=} T,$$

or in other words, for any $s \geq 0$

$$\mathbb{P}(T \geq s + t \mid T \geq t) \overset{(d)}{=} \mathbb{P}(T \geq s).$$

It is the only distribution on $[0, +\infty)$ that has this property for all $s, t > 0$.

Proof: Straightforward by definition of the conditional distribution. □

Proposition 4.10: Infimum of exponential variables ★

Fix summable sequence $\lambda_k > 0$, $\lambda := \sum_{k\in\mathbb{N}} \lambda_k < \infty$. Let $(T_k)_{k\in\mathbb{N}}$ be a sequence of independent, exponentially distributed variables with parameters λ_k. Then, $T := \inf_{k\in\mathbb{N}} T_k$, has exponential distribution with parameter λ,

$$\mathbb{P}(\exists! k \in \mathbb{N},\ T = T_k) = 1.$$

We denote by K the random corresponding index, namely $T = T_K$, for any $k \in \mathbb{N}$, we have

$$\mathbb{P}(K = k) = \frac{\lambda_k}{\lambda}.$$

Finally, the random variables T and K are independent.

Proof: One can straightforwardly compute

$$\begin{aligned}\mathbb{P}(T \geq t, K = k) = \mathbb{P}(T_j \geq T_k \geq t,\ \forall j \neq k) = \int_t^{+\infty} \mathbb{P}(T_j \geq s,\ \forall j \neq k)\lambda_k e^{-\lambda_j s} ds \\ = \int_t^{+\infty} \lambda_k e^{-\lambda s} ds = \frac{\lambda_k}{\lambda} e^{-\lambda t}.\end{aligned}$$

From this identity follow all the statements of the proposition. □

4.2.2 Intensity matrix and semi-group

For now, we assume for simplicity that the state-space E is finite. As we will see in Section 4.4, all notions given in this section can be extended to the case where E is countable. We now introduce the notion of intensity matrix, that plays for continuous-time, countable-state space Markov processes an analogous role to the transition matrix for Markov chains.

Definition 4.6: Intensity matrix ★

Assume that E is finite. An *intensity matrix* $L = [\ell_{e,e'}]_{e,e' \in E}$ is a matrix with coefficients in $\mathbb{R}$ defined on $E \times E$, satisfying

1. the diagonal coefficients are non-positive, $-\infty < \ell_{e,e} \leq 0$.
2. All other coefficients are non-negative, $0 \leq \ell_{e,e'} \leq +\infty$.
3. On each row, the diagonal coefficient is minus the sum of all other coefficients, $\sum_{e' \in E} \ell_{e,e'} = 0$.

Remark: *The intensity matrix can also be called* infinitesimal generator matrix*, or* transition rate matrix*.*

Proposition 4.11: Properties of the matrix exponential

Assume that E is finite. For any $t \geq 0$, and any matrix L on $E \times E$, we define the exponential matrix e^{tL} as the series

$$e^{tL} = \sum_{k=0}^{+\infty} \frac{(tL)^k}{k!} \tag{4.2}$$

We will admit that this matrix is well defined for any $t \geq 0$, and the this definition of the exponential shares the same properties as the real-valued one, i.e.

1. For any positive t, the convergence radius of this series is infinite.
2. For any integer n, $e^{nL} = (e^L)^n$.
3. If L_1 and L_2 commute, $e^{L_1+L_2} = e^{L_1}e^{L_2}$.

Proof: admitted. □

Exercise 21 : Computation of the semi-group

Compute $P_t[1,1]$ when L is the intensity matrix

$$L := \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 2 & 1 & -3 \end{pmatrix}$$

Answer : To compute the coefficients of the semi-group, the general strategy is to compute the eigenvalues $\alpha_1, \ldots, \alpha_3$ of the matrix L. Note that any intensity matrix will have 0 as eigenvalue, because

$$L\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

Furthermore, since P_t is a stochastic matrix, all af its eigenvalues must be non-positive.

Then, the eigenvalues of e^{tL} will be given by $e^{t\alpha_i}$, for $i = 1, \dots, 3$, so that $P_t[i, j]$ will be searched of the form

$$P_t[j,k] = \sum_{i=1}^{3} a^i_{j,k} e^{t\alpha_i}.$$

To identify the coefficients $a^i_{j,k}$, we use the fact that $P_0 = Id$, $P'_0 = L$, and $P''_0 = L^2$, and that although the eingenvalues may be complex, the semi-group's coefficients are real numbers. Note that this strategy becomes less and less applicable as the dimension of the matrix grows.

In this case, we find

$$\alpha_1 = -2, \quad \alpha_2 = -4, \quad \alpha_3 = 0,$$

si that we search $P_t[11] = a + be^{-2t} + ce^{-4t}$, with $P_0[1,1] = 1$, $P^1_0[1,1] = \ell_{1,1} = -2$, and $P^1_0[1,1] = L^2[1,1] = 7$. Solving this system yields the values of a, b and c.

□

Theorem 4.12: Semi-group ★

If E is finite, for any matrix L on $E \times E$, define the semi-group

$$\{P_t := e^{tL},\ t \geq 0\}.$$

Then, the matrix L is an intensity matrix, in the sense of Definition 4.6, if and only if for any $t \geq 0$, P_t is a stochastic matrix. (Recall that a matrix is called stochastic if is has non-negative entries, and its rows all sum to 1)

Proof: First assume that P_t is stochastic for any $t \geq 0$. Since tL and sL commute for any non-negative s, t, we have $P_{s+t} = P_s P_t$. In particular, by definition of the exponential matrix, as $t \searrow 0$, we can write

$$P_t = Id_E + tL + O(t^2), \tag{4.3}$$

so that since P_t is stochastic, all entries of L must be non-negative, except at the diagonal, and its entries on each row must sum to 0. This proves that L must be an intensity matrix.

Conversely, if L is an intensity matrix, for any fixed time t, we can write $P_t = P^n_{t/n}$. For t/n small enough, $P_{t/n}$ has positive entries, therefore so does P_t. Furthermore, one easily show by recurrence that L^n's rows also sum to 0, so that in particular, all rows of $\sum_{k=1}^{+\infty} tL^k/k!$ sum to 0. This concludes the proof. □

The identity $P_t = e^{tL}$ associates to any intensity matrix L a semi-group $(P_t)_{t\geq 0}$, with the binary operation $P_s \cdot P_t = P_{t+s}$. We now derive a few important properties to characterize the semi-group $(P_t)_{t\geq 0}$ given L.

Theorem 4.13: Kolmogorov equations ★

Assume that E is finite. Fix an intensity matrix L, the semi-group $(P_t)_{t\geq 0}$ is the unique solution to the Kolmogorov equations

1. $P'(t) = P_t L$ (forward equation),

2. $P'(t) = LP_t$ (backward equation),

It also satisfies $P_0^{(k)} = Q^k$.

Proof: The fact that P_t satisfies the equations is straightforward, given the power series formula for P_t and the fact it has infinite convergence radius. Regarding uniqueness, for any solution M_t to the first equation, one easily obtains that $M_t e^{-tQ}$ has null derivative, and is therefore constant equal to Id_E. The same proof holds for the second equation. □

4.3 Markov jump process on finite sets

4.3.1 Jump processes and measurability

Since we focus on Markov processes, we will not give a formal definition of general jump processes. Informally however, jump processes are continuous time processes with discrete jumps at random times $S_0 := 0 < S_1 < S_2 \ldots$, where the process changes its state to $Y_0, Y_1, Y_2, \ldots$. In what follows, we will refer to the set of visited states $(Y_k)_{k\in\mathbb{N}}$ as the process's *skeleton*, and to the time interval in-between jumps $\tau_k = S_k - S_{k-1}$ as the *holding times*.

For continuous-time processes, however, one needs to consider the question of measurability. Without assuming any regularity on a process X composed of a collection of measurable variables $(X_t)_{t\geq 0}$, some basic questions can remain unanswered. Given such a process X, for example, started from $X_0 = 0$, imagine that one wants to estimate the probability that it reaches value 1 before time t. We would therefore like to define the "event"

$$\left\{ \sup_{s\leq t} X_s \geq 1 \right\} = \bigcup_{s\leq t} \{X_s \in [1, +\infty)\}.$$

Unfortunately, because the segment $[0, t]$ is not countable, neither is the union above, so that $\{\sup_{s\leq t} X_s \geq 1\} \notin \mathscr{F}$ is not measurable, and does not have a probability.

For this reason, we add an additional assumption to our definition of jump processes, and assume that they are *right-continuous*, so that now the entire trajectory $(X_s)_{s\leq t}$ can be uniquely determined with a *countable* number of measurable events, which solves the previous problem. We now introduce continuous time Markov processes, which are a specific type of jump processes satisfying, like discrete time Markov chains, the Markov property ensuring that present jumps do not depend on the past of the process. Of course, we will make this statement precise later on, see Theorems 4.15 and 4.16.

4.3.2 First construction of Markov processes ★

Once again, we first tackle the case where E is finite. Fix an intensity matrix L, and an initial distribution μ on E. We now want to build a continuous time process $(X_t)_{t\geq 0}$, whose transition matrix $\mathbb{P}(X_t = \cdot \mid X_0 = \cdot)$ between times 0 and t is given by $P_t = e^{tL}$ for any t. Fix a time t, and assume we want to build the Markov process on the time-segment $[0, t]$. By the semi-group's property, we could start our process from μ, and then build the process as a Markov chain on n small time steps $dt = t/n$ with transition matrix

$$P_{t/n} = Id + \frac{t}{n}L + O(n^{-2})$$

and then take the limit as $n \to \infty$ and the time steps become infinitely short. Clearly, for any $k \leq n$, the distribution at time k/n will be given by $\mu P_{tk/n}$. Intuitively, as time-steps become infinitely small, this process should admit a well-defined limit, provided we give a sense to the construction between the discrete sampling times. In the remainder of this section, we will give meaning to this limit construction.

In practice, unless the intensity matrix and the state space are very simple, it is in general difficult to compute explicitly the transition matrix P_t given L. Fortunately, computing it is not necessary to actually build the limiting process described above. They are many ways to formulate the construction of Markov jump processes, but the one we present here has the advantage of needing the less notations and objects, and being somewhat intuitive. Recall that μ is the initial distribution for the process.

Starting from μ, the matrix L encodes all the information necessary to build $(X_t)_{t\geq 0}$. Unlike a Markov chain, that can have a positive *jump probability* from a state $e \in \mathbb{E}$ to itself, a Markov process is characterized by its *jump rates* $\ell_{e,e'}$ from e to $e' \neq e$. Those jump rates are the entries of the intensity matrix $L = [\ell_{e,e'}]$, and they encode the (non-negative) frequency at which the process tries to jump from e to e'. As is natural for a continuous time jump processes, a Markov process is composed of two parts

- its **skeleton**, which is the succession of different states the Markov process reaches, represented by a discrete time Markov chain $\overline{Y} := (Y_n)_{n\geq 0}$.
- its **holding times**, i.e. the time the Markov process waits between jumps, represented by a succession of exponential variables $\overline{\tau} := (\tau_n)_{n\geq 0}$.

In what follows, we will denote by

$$\lambda_e := -\ell_{e,e}$$

the rate at which state e is left. We will first build the skeleton, for which we choose $Y_0 = X_0 \sim \mu$. If $\lambda_{Y_0} = 0$, the Markov process cannot leave state Y_0, and we stop the construction. Otherwise, the next state Y_1 of the skeleton is chosen according to the distribution

$$\mathbb{P}(Y_1 = e \mid Y_0 = e_0) = \frac{\ell_{e_0,e}}{\sum_{e' \neq e_0} \ell_{e_0,e'}} = \frac{\ell_{e_0,e}}{\lambda_{e_0}}.$$

having chosen $Y_1 = e_1$, we then repeat the same procedure, and choose Y_2 according to the distribution

$$\mathbb{P}(Y_2 = e \mid Y_1 = e_1) = \frac{\ell_{e_1,e}}{\lambda_{e_1}}$$

if $\lambda_{e_1} > 0$, independently from the rest of the skeleton except Y_1, otherwise we stop the construction. We carry on with this construction, assuming $Y_n = e_n$ has been chosen, we similarly choose Y_{n+1} according to the distribution

$$\mathbb{P}(Y_{n+1} = e \mid Y_n = e_n) = \frac{\ell_{e_n,e}}{\lambda_{e_n}},$$

if $\lambda_{e_n} > 0$, otherwise we stop the construction. In other words, the skeleton (Y_n) can be seen as a Markov chain, with the same initial distribution μ as X, and with transition matrix

$$\Pi[e,e'] := \frac{\ell_{e,e'}}{\lambda_e}\mathbf{1}_{\{e' \neq e,\, \lambda_e > 0\}} + \mathbf{1}_{\{e = e',\, \lambda_e = 0\}}. \tag{4.4}$$

Now that the skeleton is built, we choose the holding times. Fix an i.i.d. sequence of $Exp(1)$ times $(T_k)_{k\in\mathbb{N}}$, we define the random variables

$$\tau_k = \frac{T_k}{\lambda_{Y_k}}, \qquad S_k = \sum_{m=0}^{k} \tau_m,$$

with the convention $\tau_k = +\infty$ if $\lambda_{Y_k} = 0$.

We are now ready to build our Markov process (X_t), by letting $X_s = Y_0$ on $[0, S_0)$, and for each discrete step k, letting $X_s = Y_k$ on $[S_{k-1}, S_k)$. Note in particular that if the skeleton visits a state with jump rate $\lambda_e = 0$, then the corresponding τ_k is infinite, and the Markov process remains there forever.

Definition 4.7: Absorbing state ★

We call *absorbing state* any state e of the Markov chain satisfying $\lambda_e = 0$.

The Markov process $(X_t)_{t\geq 0}$ described above is right continuous, and fully characterized by its intensity matrix L and its initial distribution μ. We denote a process following this construction, with intensity matrix L and initial state μ by $MP(\mu, L)$.

4.3.3 Second construction of Markov processes

We now present another, and equivalent construction of continuous time Markov processes. First, we define a Poisson clock.

Definition 4.8: Poisson clock

A *Poisson clock* with rate λ is a collection $\mathscr{S} = (S_k)_{k\in\mathbb{N}}$ of times $S_0 := 0 \leq S_1 \leq S_2 \leq \dots$ such that $(S_k - S_{k-1})_{k\geq 1}$ is a sequence of i.i.d. $Exp(\lambda)$ random variables. The S_k's will be referred to as *rings*. In other words, a Poisson clock initially waits an $Exp(\lambda)$ time to ring, and waits independent times $\sim Exp(\lambda)$ between consecutive rings.

One easily checks that a rate λ *Poisson clock* can be seen as the set of jump times of a rate λ *Poisson process*, and is therefore measurable w.r.t. the Poisson process's natural filtration.

Proposition 4.14: Markov property for Poisson clocks

For any stopping time τ w.r.t. the natural filtration of a rate λ Poisson process $(N_t)_{t\geq 0}$, the process $(N_{\tau+t} - N_\tau)_{t\geq 0}$ is a rate λ Poisson process independent from the information $\mathscr{F}_\tau$ prior to τ. In particular, a Poisson clock seen from a stopping time τ is still a Poisson clock.

We admit this proposition, and turn to an alternative construction for Markov processes: consider independently for any $e \neq e'$ a Poisson clock $\mathscr{S}^{e,e'}$ with rate $\ell_{e,e'}$, meaning that with each *transition* $e \to e'$ of the process is associated a Poisson clock that rings infinitely many times $S_0^{e,e'} := 0 \leq S_1^{e,e'} \leq S_2^{e,e'} \leq \dots$. We now build the Markov process as follows. We call e the *starting point* of the Poisson clock, and e' its *destination*.

- The process is initially in state $X_0 = Y_0 \sim \mu$.
- X_t remains in state Y_0 until one of the Poisson clock $(\mathscr{S}^{Y_0,e'})_{e'\in E}$ with starting point Y_0 rings, at time $S_0 = S_0^{Y_0,Y_1} = \min_{e'\neq e} S_0^{Y_0,e'}$. Then, the process performs the corresponding jumps , and remains in Y_1 until one of the Poisson clock $(\mathscr{S}^{Y_1,e'})_{e'\in E}$ with starting point Y_1 rings.
- We carry on with this construction, by waiting at each state Y_k until one of the Poisson clock with starting point Y_k rings at time

  $$S_k = S_{n_k}^{Y_k,Y_{k+1}} = \min_{\substack{e'\neq e\\ n\in\mathbb{N}}} \left\{ S_n^{Y_k,e'} \geq S_{k-1} \right\},$$

  at which point the process jumps to Y_k.

Note that if the process reaches a state Y_k such that $\lambda_{Y_k=0}$, none of the exponential clocks with starting point Y_k will ever ring, so that the system remains stuck in Y_k.

It is not hard to see that the two constructions presented above are equivalent: indeed, by the memoryless property of the exponential distribution, at any of the times S_k where the process jumps to Y_k, assuming that $\lambda_{Y_k} > 0$ the probability that the next clock to ring is associated with the destination e' is (see Propositions 4.14 and 4.10)

$$\mathbb{P}(Y_{k+1} = e') = \frac{\ell_{Y_k,e'}}{\lambda_{Y_k}},$$

and the corresponding holding time is the time needed for one of the Poisson clocks to ring

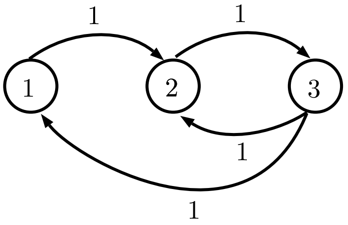
$$\tau_k = \min_{\substack{e' \neq e \\ n \in \mathbb{N}}} \left\{ S_n^{Y_k,e'} - S_{k-1}, \text{ for } S_n^{Y_k,e'} \geq S_{k-1} \right\} \overset{(d)}{=} Exp(\lambda_{Y_k})$$

Graphical representation of continuous time Markov Processes: as for discrete time Markov processes, continuous time processes can be represented by a graph, with set of vertices E, and with set of edges $\mathscr{E} = \{(e, e') \in E^2, \ell_{e,e'} > 0\}$. Each edge $(e, e') \in \mathscr{E}$ is then labeled with its corresponding jump rate $\ell_{e,e'}$. The only difference between graphs of for Markov chains or Markov process is that Markov processes' graphs do not have any self-edge (e, e), and edges' tags are in $[0, +\infty)$ instead of $[0, 1]$.

Example: The Markov process on $E := \{1, 2, 3\}$ with generator matrix

$$L = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

is represented by the graph



.

The corresponding Markov process jumps at rate 1 from 1 to 2, at rate 1 from 2 to 3, and at rate 2 from 3 to either 1 or 2, chosen w.p. 1/2, 1/2.

End of lecture 7

4.3.4 Fundamental properties of Markov processes

We now introduce some key properties of Markov processes. We start by the Markov property, already seen for discrete time Markov chains, that is a fundamental tool to study Markov processes.

Theorem 4.15: Markov property ★

Let X_t be a $MP(\mu, L)$. Fix $s \geq 0$ and $e \in E$. Then, the distribution of $(X_{t+s})_{t\geq 0}$ conditioned to $\{X_s = e\}$ is that of a $MP(\delta_e, L)$ independent of the past σ-algebra $\mathscr{F}_s^X = \sigma\{X_r, r \leq s\}$.

Proof: This result will be admitted. □

This result encodes the absence of memory of Markov processes, which is analogous to that of discrete time Markov chains. In the Markov property, the conditioning time s is fixed, but a stronger result can be stated if s is replaced by a stopping time T.

Theorem 4.16: Strong Markov property ★

Let T be a stopping time w.r.t. the natural filtration of a $MP(\mu, L)$ X_t. Then, the distribution of $(X_{t+T})_{t\geq 0}$ conditioned to $\{T < \infty, X_T = e\}$ is that of a $MP(\delta_e, L)$, independent of the past prior to the stopping time, i.e. independent of the sigma-algebra $\mathscr{F}_T^X$ introduced in Definition 1.12.

Proof: This result will be admitted. □

Proposition 4.17: Semi-group ★

The Markov process built in Section 4.3.2 has $P_t = e^{tL}$ for transition matrix between times 0 and t, i.e.

$$Q_t[e_0, e] := \mathbb{P}(X_t = e \mid X_0 = e_0) = P_t[e_0, e].$$

Proof: According to Theorem 4.13, it is enough to check that the transition matrix is solution to the Kolmogorov forward equation. More precisely, we consider the process at time t and $t + \varepsilon$, and we want to estimate

$$\lim_{\varepsilon \to 0} \frac{Q_{t+\varepsilon}[e_0, e] - Q_t[e_0, e]}{\varepsilon}. \tag{4.5}$$

To estimate the limit above, we first condition the first term on the value of X_t

$$Q_{t+\varepsilon}[e_0, e] = \sum_{e' \in E} \mathbb{P}(X_{t+\varepsilon} = e \mid X_t = e',\ X_0 = e_0) Q_t[e_0, e'].$$

By construction, $\mathbb{P}(X_{t+\varepsilon} = e \mid X_t = e',\ X_0 = e_0) = \mathbb{P}(X_{t+\varepsilon} = e \mid X_t = e')$ since once we know that we are at site e at time t, the past of the process does not affect the construction. Consider the event

$$J := \{X_t \text{ has jumped at least twice in the time interval } [t, t+\varepsilon)\}.$$

Since the holding times are exponential, we can write, by Markov property at time t, and strong Markov property at the next jump time, the crude bound over all possible holding times

$$\mathbb{P}(J) \leq \sup_{e,e' \in E} \mathbb{P}(\mathscr{E}_{\lambda_e} \leq \varepsilon, \mathscr{E}_{\lambda_{e'}} \leq \varepsilon) = \sup_{e,e' \in E} (1 - e^{\lambda_e \varepsilon})(1 - e^{\lambda_{e'} \varepsilon}) = O(\varepsilon^2),$$

where $\mathscr{E}_{\lambda_e}$, $\mathscr{E}_{\lambda_{e'}}$ represent two independent exponential holding times. Note that this bound is by no means sharp, since we only require that both times are less than ε, whereas their sum should be less than ε for J to occur. We can now rewrite the quantity inside the limit in 4.5 as

$$\frac{\sum_{e' \in E} \mathbb{P}(X_{t+\varepsilon} = e \text{ and } J^c \mid X_t = e') Q_t[e_0, e'] - Q_t[e_0, e]}{\varepsilon} + O(\varepsilon).$$

We first single out the term for $e' = e$, meaning that the process has not jump in ther time interval $[t, t + \varepsilon]$, which occurs with probability

$$\mathbb{P}(X_{t+\varepsilon} = e \text{ and } J^c \mid X_t = e) = e^{-\lambda_e \varepsilon} = 1 - \lambda_e \varepsilon + O(\varepsilon^2).$$

Here, we used the fact that exponential distribution is memoryless (cf. Proposition 4.9), since we do not know the time the process reached state e before time t. The other terms for $e \neq e'$ are easily computed as well, once again using the lack of memory of exponential variables, since they are the probabilities, starting from e' to jump to e before time ε,

$$\mathbb{P}(X_{t+\varepsilon} = e \text{ and } J^c \mid X_t = e') = \frac{\ell_{e',e}}{\lambda_{e'}}(1 - e^{-\lambda_{e'} \varepsilon}) = \ell_{e',e} \varepsilon + O(\varepsilon^2).$$

Putting all these identities together, we obtain

$$\frac{Q_{t+\varepsilon}[e_0, e] - Q_t[e_0, e]}{\varepsilon} = \sum_{e' \neq e} \ell_{e',e} Q_t[e_0, e'] - \lambda_e Q_t[e_0, e] + O(\varepsilon) = (Q_t L)[e_0, e] + O(\varepsilon).$$

Letting ε go to 0 proves that the matrix Q_t solves the Kolmogorov forward equation, and is therefore equal to $P_t = e^{tL}$ according to Theorem 4.13.

□

Definition 4.9: Communicating states, classes

States (e, e') *communicate* in a $MP(\mu, L)$ if both are accessible from the other, i.e. there exists $t, t' > 0$ such that $\mathbb{P}(X_t = e' \mid X_0 = e)$ and $\mathbb{P}(X_{t'} = e \mid X_0 = e')$ are both positive. The following characterization of accessible states are equivalent:

i) e' is accessible from e.

ii) e' is accessible from e in the skeleton (Y_k).

iii) There exists a sequence $e_1 = e,\ e_2, \ldots e_n = e'$ such that $\ell_{e_k, e_k+1} > 0\ \forall k \in \{1, \ldots, n - 1\}$.

iv) $P_t[e,e'] > 0$ forall $t > 0$.

v) $\exists t > 0$ such that $P_t[e,e'] > 0$.

In particular, the notion of communicating classes anf irreducibility for a Markov process coincide with those of its skeleton, see Definition 4.2.

Definition 4.10: Infinitesimal generator ★

Any vector F indexed by E can be seen as a function $f : E \to \mathbb{R}$. In the same way, the intensity matrix L associated with a Markov process X can be seen as an operator $\mathscr{L}$ mapping the set of functions $f : E \to \mathbb{R}$ into itself, by defining $\mathscr{L}f$ as the function on E

$$\mathscr{L}f(e) = \sum_{e' \neq e} \ell_{e,e'}\{f(e') - f(e)\}$$

The operator $\mathscr{L}$ is called the *infinitesimal generator* of the Markov process. In matrix form, the identity above rewrites LF as the vector whose e-entry is given by $\sum_{e' \in E} \ell_{e,e'} f(e')$.

Remark: *For general Markov processes, one can check that the infinitesimal generator can be seen as the derivative at time* 0 *of the semi-group:*

$$\mathscr{L}f(e) = \lim_{t \searrow 0} \frac{P_t F[e] - F[e]}{t}.$$

In reality, this definition of the infinitesimal generator is a general way to define it given the associated Markov semi-group. In what follows, we will at times characterize Markov processes either by their intensity matrix L *or their infinitesimal generator* $\mathscr{L}$*, but keep in mind that the two are perfectly equivalent for our use.*

4.4 Infinite but countable state spaces

Some precautions need to be taken in the case where the state space E is infinite but countable. Indeed consider for example the following Markov process N_t on $\mathbb{N}$, started from $N_0 = 0$, with generator operating on functions $f : \mathbb{N} \to \mathbb{R}$ as

$$\mathscr{L}f(k) = 2^k\{f(k+1) - f(k)\}.$$

In other words, once it reaches k, the process waits for a time $\tau_k \sim Exp(2^k)$ before jumping to $k+1$. Note that the total holding time before escaping to infinity,

$$\tau^* = \sum_{k \in \mathbb{N}} \tau_k$$

has finite expectation $\mathbb{E}(\tau^*) = 2$, so that in particular $\tau*$ is a.s. finite. This means that the chain jumps faster and faster and escapes to infinity in finite time a.s..

Definition 4.11: Explosive Markov process

Given a Markov process $X := (X_t)_{t\geq 0}$, we denote by $\tau^* = \sum_{k=0}^{\infty} \tau_k$ its total holding time. The process X is called *explosive* is τ^* is finite with positive probability, namely

$$\mathbb{P}(\tau^* < \infty) > 0.$$

We now give a criterion for non-explosiveness.

Theorem 4.18: Non-explosion criteria

If one of the following condition is satisfied, then X is non-explosive.

i) The state space E is finite.

ii) All states have bounded exit rate,

$$\sup_{e\in E} \lambda_e < \infty. \tag{4.6}$$

iii) There exists a state that is visited a.s. and is recurrent for the skeleton, i.e.

$$\exists e \in E, \mathbb{P}(\exists t \geq 0,\ X_t = e) = 1,$$

and e is recurrent for the Markov chain $(Y_k)_{k\in\mathbb{N}}$.

Note that these conditions are sufficient but not necessary. In the context of this course, we will focus on the case where condition *ii*) above is satisfied. In this case, the construction laid out in paragraph 4.3.2 holds verbatim, with the exception now that at each construction step k, if $\lambda_e > 0$, the distribution of the next state Y_{k+1} is chosen in a countable set instead of a finite one, with probability

$$\mathbb{P}(Y_{k+1} = e' \mid Y_k = e) = \frac{\ell_{e,e'}}{\lambda_e}.$$

When instead, the λ_e's are not bounded, or, even worse, when they can be infinite, subtler constructions are required, but those go much beyond the scope of this course.

The definition of the intensity matrix L in the case of a countable infinite state space remains the same, with the only difference being that L is a infinite square matrix, satisfying the same three properties 1-3. in Definition 4.6. It is not clear, however, that identity (4.2) gives a well-defined formulation for the exponential of L. This is not an issue, because in practice (4.2) is seldom used to identify the semi-group associated with a Markov process, one instead typically uses the Kolmogorov equations. The latter is guaranteed in the infinite countable case by the following result.

Theorem 4.19: Kolmogorov equations

Let E be a countable set, and let L be an intensity matrix according to Definition 4.6. Assume that L satisfies (4.6), the equation

$$P'_t = LP_t, \quad P_0 = Id_E,$$

has a unique non-negative solution, which forms a semi-group $P_sP_t = P_{s+t}$. Furthermore, P_t is also the unique solution to

$$P'_t = P_tL, \quad P_0 = Id_E.$$

Finally, a right-continuous process X_t is $MP(\mu, L)$ in the sense of paragraph 4.3.2 if and only if $X_0 \sim \mu$, and for any integer n, any $0 \leq t_0 \leq \cdots \leq t_n$ and $e_0, \ldots, e_n \in E$

$$\mathbb{P}(X_{t_n} = e_n \mid X_{t_0} = e_0, \ldots, X_{t_{n-1}} = e_{n-1}) = P_{t_n - t_{n-1}}[e_{n-1}, e_n].$$

For countable state spaces, however, the most convenient description for a given Markov process satisfying (4.6) is through its infinitesimal generator, operating on *bounded functions* f. Indeed, applying definition 4.10, one obtains the identity

$$\mathscr{L}f(e) = \sum_{e' \neq e} \ell_{e,e'}\{f(e') - f(e)\}$$

which is well defined if $\sum_{e'} \ell_{e,e'} \leq \infty$ and f is bounded.

Furthermore, under assumption 4.6, the Markov and strong Markov property (Theorems 4.15 and 4.16) both hold in the case of an infinite countable state space. Once again, this goes beyond the scope of our course, but as long as the underlying Markov process is well defined, the Markov property holds with great generality.

Corollary 4.20: Dynkin's formula ★

Let $(X_t)_{t\geq 0}$ be a $MP(\mu, L)$ on a countable set E satisfying (4.6). Then, for any bounded function $F : \mathbb{R}_+ \times E$

$$\frac{d}{dt}\mathbb{E}(F_t(X_t)) = \mathbb{E}((\mathscr{L}F_t)(X_t)) + \mathbb{E}((\partial_t F)(X_t)). \tag{4.7}$$

In particular, there exists a martingale M_t^F (w.r.t. $(X_t)_{t\geq 0}$'s natural filtration) such that

$$F_t(X_t) = F_0(X_0) + \int_0^t (\mathscr{L} + \partial_s)F_s(X_s)ds + M_t^F.$$

This identity is called *Dynkin's formula*.

Remark: *In the first identity, the first term corresponds to the variation due to the* *time-variation of (X_t), whereas the second corresponds to the time variation of F itself. In particular, if F does not depend on time,*

$$\frac{d}{dt}\mathbb{E}(F(X_t)) = \mathbb{E}((\mathscr{L}F)(X_t)).$$

One can show that the quadratic variation $\langle M^F\rangle_t$ is then given by the identity

$$\langle M^F\rangle_t = \int_0^t \left\{(\mathscr{L}F_s^2)(X_s) - 2F_s(\mathscr{L}F_s)(X_s)\right\} ds.$$

Dynkin's formula is a very useful way to compute the expectation of functionals of Markov processes. The estimate on its quadratic variation then yields some control on the fluctuations around this expectation.

Proof: We first prove the first identity. To do so, recall the Kolmogorov equations, $P_t' = LP_t$. Recall that the distribution of the process at time t is given by μP_t. In matrix form, $\mathbb{E}(F(X_t))$ rewrites as $\mu P_t F_t$, where we denote F_t for the vector with entries $[F_t(e)]_{e\in E}$. Then, we have in matrix form

$$\frac{d}{dt}\mathbb{E}(F_t(X_t)) = \mu P_t' F_t + \mu P_t(\partial_t F_t).$$

the second term can be rewritten as $\mathbb{E}(\partial_t F_t(X_t))$. By Kolmogorov's equation, the first term can be rewritten as $\mu P_t[LF_t] = \mathbb{E}(\mathscr{L}F_t(X_t))$, which proves the identity.

We now turn to the second identity, that we will only prove for F not depending on time, the adaptation for F time-dependent is straightforward. We only need to prove that

$$M_t^F := F(X_t) - F(X_0) - \int_0^t \mathscr{L}F(X_s)ds.$$

is an $(\mathscr{F}_t)$-martingale. Fix $r \leq t$, conditioning to $\mathscr{F}_r$, and splitting the integral in two parts, we obtain

$$\begin{aligned}\mathbb{E}(M_t^F \mid \mathscr{F}_r) = \mathbb{E}(F(X_t) \mid \mathscr{F}_r) - F(X_r) - \int_r^t \mathbb{E}\left[\mathscr{L}F(X_s) \mid \mathscr{F}_r\right] ds \\ + F(X_r) - F(X_0) - \int_0^r \mathscr{L}F(X_s)ds\end{aligned}$$

where we used repeatedly that X_s is $\mathscr{F}_r$-measurable for any $s \leq r$. The second line is exactly M_r^F, we now prove that the first line vanishes. By Markov property, conditionally to $X_r = e$, the process $\widetilde{X}^e := (X_{s+r})_{s\geq 0}$ is a $MP(\delta_e, L)$ independent from $\mathscr{F}_r$, so that the first line rewrites

$$\sum_{e\in E}\mathbb{P}(X_r = e)\left[\mathbb{E}(F(\widetilde{X}_{t-r}^e)) - F(e) - \int_0^{t-r} \mathbb{E}\left(\mathscr{L}F(\widetilde{X}_s^e)\right) ds\right].$$

By (4.7), for any fixed e, the quantity in brackets above vanishes, which proves as wanted that $\mathbb{E}(M^F(t) \mid \mathscr{F}_r) = M^F(r)$. □

The Dynkin formula is the reason the generator of the process is a key quantity to understand. Assume for example that you want to understand the evolution in time, for a given set $\Delta \subset E$, of $\mathbb{P}(X_t \in \Delta)$, one only has to apply the Dynkin's formula to the function $F = \mathbf{1}_{\{X_t \in \Delta\}}$, and apply the generator to F.

End of lecture 8

Exercise 22

We consider the evolution of a bacteria population. Initially, a single bacteria is born. All bacteria have the same behavior, independently for each bacteria: after a random time T_d after their birth, which has distribution $T_d \sim Exp(\lambda_d)$, the bacteria dies, and after a random time T_b^1 with distribution $T_b^1 \sim Exp(\lambda_b)$, the bacteria produces a new bacteria. Once a bacteria has produced a new bacteria, it wait once again a time $T_b^2 \sim Exp(\lambda_b)$ before producing another one, and so on until it finally dies. If $T_d < T_b^1$, the bacteria dies without having ever reproduced. We denote by $\mathbb{P}_1$ the distribution of the process started from 1 bacteria.

1) (i) Justify that the number N_t of bacteria in the system is a Markov process, whose only transitions with positive rates are $\ell_{n,n+1} = n\lambda_b$, $\ell_{n,n-1} = n\lambda_d$, and that 0 is an absorbing state.

(ii) Define $h(t) = \mathbb{P}_1(N_t = 0)$, show that

$$h(t) = \int_0^t e^{-(\lambda_b+\lambda_d)s}(\lambda_d + \lambda_b h(t-s)^2)ds.$$

Hint: denote τ_0 the first holding time, justify that $h(t) = \mathbb{P}_1(N_t = 0, \tau_0 < t)$, and exploit the Markov property at the time of the first jump.

(iii) Compute $h(t)$.

2) We now consider the same process started from a population with $n > 1$ bacteria. Compute $h^{(n)}(t) = \mathbb{P}_n(N_t = 0)$.

Answer :

1) (i) Consider the time to give birth to a new particle : given the number $N_t = n$ of particles, each one giving birth independently after an exponential time with parameter λ_b, the time to give birth to a bacteria is the min of n independent exponential variables, which follows an exponential distribution with parameter $n\lambda_b$ according to Proposition 4.10. The same is true for the death of bacteria, therefore when there are n bacteria, one is born at rate $n\lambda_b$ and one dies at rate $n\lambda_d$. Furthermore, w.p. 1, no two deaths or births can occur at the same time, therefore only the transitions $n \to n+1$ and $n \to n+1$ occur at positive rates. In particular, 0 is an absorbing state, since no transition occurs at positive rate at $N_t = 0$.

(ii) Recall that we define S_0 as the first time the process jumps. If $S_0 > t$, then $N_t = 1$. In particular, we have $h(t) = \mathbb{P}_1(N_t = 0, S_0 \leq t)$, so that projecting on the value of S_0,

$$h(t) = \mathbb{P}_1(N_t = 0, S_0 \le t) = \int_0^t ds\lambda e^{-\lambda s}\mathbb{P}_1(N_t = 0 \mid S_0 = s).$$

where we shortened $\lambda = \lambda_b + \lambda_d$ the total jump rate per bacteria. We can rewrite for any $s \le t$

$$\begin{aligned}\mathbb{P}_1(N_t = 0 \mid S_0 = s) =& \frac{\lambda_d}{\lambda}\mathbb{P}_1(N_t = 0 \mid S_0 = s, N_{S_0} = 0) + \frac{\lambda_b}{\lambda}\mathbb{P}_1(N_t = 0 \mid S_0 = s, N_{S_0} = 2)\\ =& \frac{\lambda_d}{\lambda} + \frac{\lambda_b}{\lambda}\mathbb{P}_2(N_{t-s} = 0)\end{aligned}$$

by strong Markov property. Furthermore, since the bacteria and their descendants evolve independently, the probability that the process started from 2 bacteria dies out is the square of the probability that the process started from one bacteria dies. We obtain as wanted

$$h(t) = \int_0^t e^{-\lambda s}\left(\lambda_d + \lambda_b h(t-s)^2\right) ds.$$

(iii) To compute the previous quantity, we first perform the change of variables $t - s \mapsto s$, to obtain

$$h(t) = \int_0^t e^{-\lambda(t-s)}\left(\lambda_d + \lambda_b h(s)^2\right) ds = e^{-\lambda t}\int_0^t e^{\lambda s}\left(\lambda_d + \lambda_b h(s)^2\right) ds.$$

Taking the time derivative yields

$$h'(t) = -\lambda h(t) + \lambda_d + \lambda_b h(t)^2.$$

We separate variables, to obtain

$$dt = \frac{dh}{\lambda_d + \lambda_b h^2 - \lambda h} = \frac{dh}{(1-h)(\lambda_d - \lambda_b h)} = \frac{adh}{1-h} + \frac{bdh}{\lambda_d - \lambda_b h},$$

with $a\lambda_d + b = 1$ and $-a\lambda_b - b = 0$, we obtain $a = 1/(\lambda_d - \lambda_b)$, $b = -\lambda_b/(\lambda_d - \lambda_b)$. This finally yields

$$t = -a\log(1-h) - \frac{b}{\lambda_b}\log\left(1 - \frac{\lambda_b}{\lambda_d}h\right),$$

so that

$$e^{(\lambda_d - \lambda_b)t} = \frac{1 - \frac{\lambda_b}{\lambda_d}h}{1-h} = \frac{1 - \frac{\lambda_b}{\lambda_d}}{1-h} + \frac{\lambda_b}{\lambda_d},$$

which in turn yields

$$h = \frac{e^{(\lambda_d - \lambda_b)t} - 1}{e^{(\lambda_d - \lambda_b)t} - \lambda_b/\lambda_d}.$$

Note in particular that if $\lambda_d > \lambda_b$, h goes to 1 as t goes to ∞, whereas if $\lambda_d < \lambda_b$, $h \to \lambda_d/\lambda_b$.

2) By the same argument as previously, descendants of bacteria evolve independently, to that $h^{(n)}(t) = h(t)^n$. □

4.5 Examples of Markov processes

We now give in more details a few key examples of Markov processes.

4.5.1 Poisson jump process ★

We start by the Poisson jump process, which has already been mentioned at several points throughout the course.

Definition 4.12: Poisson Process

A *Poisson jump process* with rate λ is a right-continuous, non decreasing process, with skeleton given by $Y_k = k \; \forall k \in \mathbb{N}$, and with independent holding times given by $\tau_k \sim Exp(\lambda)$. Note in particular that $S_k := \sum_{n=0}^{k-1} \tau_n$, $(S_k)_{k\in\mathbb{N}}$ is a rate λ *Poisson clock*.

In other words, a Poisson jump process with rate λ is a $MP(\delta_0, L)$ on $E = \mathbb{N}$, with intensity matrix given for $n, m \in \mathbb{N}$ by

$$L[n,m] = \ell_{n,m} := \lambda \left[\mathbf{1}_{\{m=n+1\}} - \mathbf{1}_{\{n=m\}} \right].$$

Its infinitesimal generator (see Def. 4.10) $\mathscr{L}_\lambda$ is characterized by its action on functions $f : \mathbb{N} \to \mathbb{R}$

$$\mathscr{L}_\lambda f(n) = \lambda \left[f(n+1) - f(n) \right].$$

Theorem 4.21: Markov property for Poisson jump processes

Let $(X_t)_{t\geq 0}$ be a Poisson jump process with parameter λ. For any fixed $s \geq 0$, $(X_{s+t} - X_s)_{t\geq 0}$ is also a Poisson Point process with parameter λ, and is independent from $\sigma(X_r, r \leq s)$. The same is true if s is replaced by a stopping time (Strong Markov property).

Proof: This is a direct consequence of the Markov property. □

Definition 4.13: Stationary and independent increments

A process $(X_t)_{t\geq 0}$ has *stationary increments* if the distribution of $X_{t+s} - X_s$ does not depend on s. It has *independent increments* if for any increasing family $t_0 = 0 < t_1 < \cdots < t_n$ the family $\{Y_k := X_{t_k} - X_{t_{k-1}}, 1 \leq k \leq n\}$ is independent.

Corollary 4.22

A càdlàg process $(X_t)_{t\geq 0}$ is a Poisson process with parameter λ if and only if it has stationary and independent increments, and if for any t, $X_t \sim Poi(\lambda t)$.

Proof: exercise. □

Theorem 4.23: Thinning and compounding of Poisson processes ★

Fix $\lambda > 0$, and consider a Poisson point process $(X_t)_{t\geq 0}$ with parameter λ. Fix a distribution $\pi = (p_i)_{1\leq i\leq n}$ on $\{1,\dots,n\}$, and a family of i.i.d. variables $(\xi_k)_k \in \mathbb{N} \sim \pi$, independent of X, to give each jump a label in $1,\dots,n$ (ξ_k is the label of the k-th jump in X). For $1 \leq i \leq n$, consider the processes

$$X_t^i = \sum_{k=1}^{X_t} \mathbf{1}_{\{\xi_k = i\}},$$

which jump by one when they encounter a jump in X_t with label i. Then, the $(X_t^i)_{t\geq 0}$ are *independent* Poisson processes with respective parameters $\lambda^i = \lambda p_i$.

Conversely , fix $\lambda_1,\dots,\lambda_n > 0$ and consider n independent Poisson processes $(X_t^i)_{t\geq 0}$ for $i = 1\dots n$ with respective parameter λ_i. Then, letting $\lambda = \sum_{i=1}^n \lambda_i$, the process $X_t = \sum_{i=1}^n X_t^i$ is a Poisson process with parameter λ.

Proof: exercise. □

4.5.2 Homogeneous Random walk on $\mathbb{Z}$

Definition 4.14: Symmetric Random walk

Given a sequence of *i.i.d.* Bernoulli random variables $(B_n)_{n\in\mathbb{N}}$, a discrete time random walk (started from the origin) is a Markov chain $Y_n := \sum_{k=1}^n (2B_k - 1)$. A *continuous time random walk* with rate λ has skeleton (Y_n), and independent holding times given by $\tau_k \sim Exp(\lambda)$.

In other words, a continuous time random walk is a $MP(\delta_0, L)$ on $E = \mathbb{Z}$, with intensity matrix given for $n, m \in \mathbb{N}$ by

$$L[n,m] = \ell_{n,m} := \lambda\left[\mathbf{1}_{\{m=n+1\}} + \mathbf{1}_{\{m=n-1\}} - 2\mathbf{1}_{\{n=m\}}\right].$$

Its infinitesimal generator (see Def. 4.10) $\mathscr{L}_\lambda$ is characterized by its action on functions $f : \mathbb{N} \to \mathbb{R}$

$$\begin{aligned}\mathscr{L}_\lambda f(n) =& \lambda\left[f(n+1) - f(n) - (f(n) - f(n-1))\right] \\ =& \lambda\left[f(n+1) + f(n-1) - 2f(n)\right], \end{aligned} \tag{4.8}$$

which is a discrete laplacian.

4.6 Recurrence, transience, invariant states

4.6.1 Hitting times

For any set of states $A \subset E$ and a $MP(\mu, L)$ on E, we denote by

$$T^A := \inf\{t \geq 0,\ X_t \in A\} \in [0, +\infty]$$

the hitting time of A.

Remark: *We can see with hitting times the importance for the Markov process to be right-continuous : indeed, if it was for example left continuous, we might have $\{T^{\{e\}} \leq t\} \notin \mathscr{F}_t$, and would not be a stopping time. In particular, we could not use the Markov property on it.*

Proposition 4.24: Hitting probabilities

The function

$$p_e^A = \mathbb{P}_e(T^A < \infty),$$

where $\mathbb{P}_e$ is the distribution of $MP(\delta_e, L)$ started from state e, is solution to

$$\begin{cases} p_e^A = 1 \text{ if } e \in A \\ \sum_{e' \in E} p_{e'}^A \ell_{e,e'} = 0 \text{ if } e \notin A. \end{cases}$$

Proof: The first identity is trivial, the second is a consequence of the strong Markov property applied at the time of the first jump:

$$\mathbb{P}_e(T^A < \infty) = \sum_{e' \neq e} \mathbb{P}_{e'}(T^A < \infty)\mathbb{P}(Y_1 = e') = \sum_{e' \neq e} \mathbb{P}_{e'}(T^A < \infty)\frac{\ell_{e,e'}}{-\ell_{e,e}},$$

which proves the second part. □

Remark: *Note that the second identity can be rewritten as $\mathscr{L} p_\cdot^A(e) = 0\ \forall e \in A$.*

Proposition 4.25: Hitting time

The function

$$q_e^A = \mathbb{E}_e(T^A)$$

is solution to

$$\begin{cases} q_e^A = 0 \text{ if } e \in A \\ \sum_{e' \in E} q_{e'}^A \ell_{e,e'} = 1 \text{ if } e \notin A. \end{cases}.$$

Proof: The proof is similar to the previous one, by applying the strong Markov property,

$$\mathbb{E}_e(T^A) = \mathbb{E}_e(\tau_0) + \mathbb{E}_e(T^A - \tau_0) = \mathbb{E}(\tau_0) + \sum_{e' \neq e} \mathbb{E}_{e'}(T^A)\mathbb{P}(Y_1 = e') = \frac{1}{\lambda_e} + \sum_{e' \neq e} \mathbb{E}_{e'}(T^A)\frac{\ell_{e,e'}}{\lambda_e},$$

which proves the identity. □

Exercise 23 : Hitting probability of random walks

We consider a rate $\lambda = 1$ symmetric random walk, with generator given by (4.8) above, we denote by $\mathbb{P}_k$ its distribution started from k. Choose $n \in \mathbb{Z}$, and define $g_k := \mathbb{P}_k(T^{\{0\}} > T^{\{n\}})$ the probability to reach n before 0. Find an equation satisfied by the function g, and compute g_1

Answer : Clearly $g_0 = 0$ and $g_n = 1$, and by Markov property applied at the first jump time, we can write for any $1 \leq k \leq n-1$

$$g_k = \frac{1}{2}\mathbb{P}_{k-1}(T^{\{0\}} > T^{\{n\}}) + \frac{1}{2}\mathbb{P}_{k+1}(T^{\{0\}} > T^{\{n\}}) = \frac{1}{2}(g_{k+1} + g_{k-1}),$$

therefore letting $\delta g_k := (1/2)(g_{k+1} - g_k)$, we have for any $0 \leq k \leq n-1$ that

$$\delta g_k = \delta g_{k+1}.$$

In particular, since $1 = g_n - g_0 = \sum_{k=0}^{n-1} \delta g_k = n\delta g_1$, so that $\delta g_k = 1/n \ \forall k \in \{0, \dots, n-1\}$. This yields

$$g_k = \sum_{m=0}^{k-1} \delta g_m = k/n.$$

□

End of lecture 9

4.6.2 Recurrence and transience

In this paragraph, we fix an intensity matrix L.

Definition 4.15: Recurrent state ★

A state e is called *recurrent* if

$$\mathbb{P}_e(\{t \geq 0, \ X_t = e\} \text{ is not a bounded set}) = 1,$$

where $\mathbb{P}_e$ is the distribution of a $MP(\delta_e, L)$. Otherwise, e is called *transient*.

Proposition 4.26: Recurrent and transient classes

A state e is transient iff

$$\mathbb{P}_e(\{t \geq 0, \ X_t = e\} \text{ is not a bounded set}) = 0,$$

i.e. the probability above can only be either 0 or 1. Furthermore, a state is recurrent (resp. transient) iff it is recurrent (resp. transient) for the skeleton

$(Y_k)_{k\in\mathbb{N}}$ of X.

In particular, as for Markov chains (cf. Proposition 4.4), recurrence and transience are classes properties : all states in a communicating class are either recurrent or transient.

Proof: Assume that e is transient, i.e.

$$\nu_e = \mathbb{P}_e(\{t \geq 0,\ X_t = e\} \text{ is not a bounded set}) < 1.$$

Then, we apply the Markov property to the first time the Markov chain gets back to e,

$$T_+^e = \inf\{t > \tau_0,\ X_t = e\} \in [0, +\infty],$$

which is a stopping time, to obtain

$$\nu_e = \mathbb{P}(T_+^e < \infty)\mathbb{P}_e(\{t \geq 0,\ X_t = e\} \text{ is not a bounded set}) = \mathbb{P}(T_+^e < \infty)\nu_e.$$

But since e is transient, $\mathbb{P}(T_+^e < \infty) < 1$, so that we must have $\nu_e = 0$.

The other properties are straightforward. □

We call *local time* at e the average time spent by the Markov process at site e. For example, assuming the chain starts from state e,

$$\mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right) = \int_0^{+\infty} \mathbb{P}_e(X_t = e)dt := \int_0^{+\infty} p_{e,e}(t)dt.$$

Proposition 4.27: Consequence on the local time

If $\lambda_e = 0$ or $\mathbb{P}_e(T_+^e < \infty) = 1$, then e is recurrent and $\int_0^{+\infty} p_{e,e}(t)dt = \infty$.

If $\lambda_e > 0$ and $p_e^+ := \mathbb{P}_e(T_+^e < \infty) < 1$, then e is transient and

$$\int_0^{+\infty} p_{e,e}(t)dt = 1/\lambda_e(1 - p_e^+) < \infty.$$

Proof: If $\lambda_e = 0$, then a.s., starting from e, we have $X_t = e\ \forall t > 0$. Otherwise, if $\mathbb{P}(T_+^e < \infty) = 1$, we can write

$$\int_0^{+\infty} p_{e,e}(t)dt = \int_0^{+\infty} \mathbb{E}_e(\mathbf{1}_{\{X_t=e\}})dt = \mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right).$$

By strong Markov property applied T_+^e, the right-hand side can be rewritten

$$\mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right) = \mathbb{E}_e(\tau_0) + \mathbb{P}(T_+^e < \infty)\mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right) = \frac{1}{\lambda_e} + \mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right).$$

This proves that $\mathbb{E}_e\left(\int_0^{+\infty} \mathbf{1}_{\{X_t=e\}}dt\right) = \infty$, because λ_e was assumed positive.

If $p_e^+ := \mathbb{P}(T_+^e < \infty) < 1$, we apply the same identity, to obtain that

$$\int_0^{+\infty} p_{e,e}(t)dt = \frac{1}{\lambda_e} + p_e^+ \int_0^{+\infty} p_{e,e}(t)dt,$$

which yields

$$\int_0^{+\infty} p_{e,e}(t)dt = \frac{1}{\lambda_e(1-p_e^+)}.$$

□

Remark: *the last identity is natural, since we want to compute the average time spent at site e starting from e. The process performs a number of excursions away from e, until the last excursion, during which e is never visited again. By Markov property, the excursions are independently and identically distributed, so that the number of excursion is distributed according to a geometric distribution with parameter $1-p_e^+$, whose average is $1/(1-p_e^+)$. Furthermore, during each excursion, the time spent in state e is the corresponding holding time, whose average is $1/\lambda_e$. Hence the result.*

Definition 4.16: Recurrent positive state ★

A recurrent state e is called *recurrent positive* iff

$$\mathbb{E}_e(T_+^e) < +\infty.$$

This is a class property : if e is recurrent positive, then any state e' which communicates with e is also recurrent positive.

4.6.3 Invariant measures

Definition 4.17: Invariant measures ★

A measure μ on E is *invariant* if $\mu L = 0$, in other words, if for any function F on E,

$$\mathbb{E}_\mu(\mathscr{L}F) = 0.$$

Remark: *The notion of invariant measure is justified by Theorem 4.19, since if X is a $MP(\mu, L)$, its semi-group P_t satisfies $P_t' = LP_t$, so that*

$$\frac{d}{dt}\mathbb{P}_\mu(X_t = e) = \mu L P_t \mathbf{1}_e = 0,$$

where $\mathbb{P}_\mu$ is the distribution of the process started from μ, so that in particular, $\mathbb{P}_\mu(X_t = e)$ is constant and equal to $\mu(e)$.

Proposition 4.28: Invariant measure and skeleton

A measure μ is invariant iff $\nu(e) = \lambda_e \mu(e)$ is invariant for the skeleton, in other words

$$\mu L = 0 \quad \Leftrightarrow \quad \nu \Pi = \nu,$$

where Π was defined by (4.4).

Proof: Immédiat. □

Theorem 4.29: Existence of invariant measures ★

Let L be an irreducible intensity matrix, and assume that it is non explosive. Then, it admits a unique invariant probability measure μ if and only if it is recurrent positive (in the sense that one/all of its state is/are, cf. Definition 4.16). In this case, we have

$$\mathbb{E}_e(T_+^e) = \frac{1}{\mu(e)\lambda_e}$$

Proof: We prove the two implications. Assume that μ is an invariant probability measure, we first prove that the chain must be recurrent. Then, if $X = MP(\mu, L)$, its distribution at all times is given by μ, so that in particular for any $t > 0$

$$\mathbb{E}_\mu\left(\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right) = \frac{1}{t}\int_0^t \mathbb{P}_\mu(X_t = e) = \mu(e).$$

If the the process is not recurrent, all states are transient, therefore according to Proposition 4.27, we must have for any state e

$$\mathbb{E}_\mu\left(\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right) \le \int_0^{+\infty} p_{e,e}(t)dt < +\infty.$$

Dividing the latter by t, and letting t go to infinity in both equations, we obtain that $\mu(e) = 0$ for any $e \in E$, which contradicts our assumption. The process is therefore recurrent.

Now assume that there is no recurrent positive state. Define $T^e := T^{\{e\}}$ the first time e is hit, as $t \to \infty$, by dominated convergence theorem

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{E}_\mu\left(\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right) &= \mathbb{E}_\mu\left(\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right)\\
&= \mathbb{E}_\mu\left(\mathbf{1}_{\{T^e<\infty\}}\lim_{t\to\infty}\frac{1}{t}\int_{T^e}^{t+T^e} \mathbf{1}_{\{X_s=e\}} ds\right)\\
&= \mathbb{E}_e\left(\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right)\mathbb{P}_\mu(T^e < \infty)\\
&= \mathbb{E}_e\left(\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}} ds\right),
\end{aligned}$$

because the process is recurrent. To establish the identities above, we used the strong Markov property at time T^e, and that for any fixed T, and any bounded function f vanishing on $[0,T]$, $\lim_{t\to\infty} t^{-1}\int_0^t f(s)ds = \lim_{t\to\infty} t^{-1}\int_T^{t+T} f(s)ds$.

Recall that under $\mathbb{P}_e$, we have $T_e < \infty$ a.s. because all states are recurrent. We define T_e^n the successive times the chain comes back to e, and τ_e^n the sucessive holding times at e, namely $T_1^e = T_+^e$,

$$\tau_n^e = \inf\{s > 0, X_{s+T_n^e} \neq e\} \quad \text{and} \quad T_{n+1}^e = \{\inf\{s > T_n^e + \tau_n^e,\ X_s = e\},$$

which are all finite a.s. by strong Markov property. Note that the τ_n^e are i.i.d. with distribution $Exp(\lambda_e)$, and the *excursion times* $\delta_n^e := T_{n+1}^e - T_n^e - \tau_n^e$, are also i.i.d. and independent from the τ_n^e's. The excursion times represent the time spent away from e after each visit.

Then,

$$\mathbb{E}_e\left(\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=e\}}ds\right) = \mathbb{E}_e\left(\lim_{n\to\infty}\frac{\sum_{k=1}^n \tau_k^e}{T_n^e}\right) = \mathbb{E}_e\left(\lim_{n\to\infty}\frac{\frac{1}{n}\sum_{k=1}^n \tau_k^e}{\frac{1}{n}\sum_{k=1}^n \delta_k^e + \tau_k^e}\right) = \frac{1}{\lambda_e\mathbb{E}_e(T_+^e)},$$

by the strong law of large numbers. Note that since the T_+^e are non negative, the law of large number holds regardless of whether $\mathbb{E}_e(T_+^e)$ is finite or not by monotonous convergence theorem, by applying it to $(\delta_k^e + \tau_k^e)\wedge M$, and letting M to ∞. All the above proves that

$$\mu(e) = \frac{1}{\lambda_e\mathbb{E}_e(T_+^e)} \qquad \forall e \in E,$$

therefore if μ is an invariant probability measure the process is recurrent positive, otherwise . □

Theorem 4.30: Long-time behavior

Let ν be a probability distribution on E, let L be an irreducible intensity matrix, and X a $MP(\nu, L)$. Then,

$$\exists \lim_{t\to\infty}\mathbb{P}(X_t = e) = \frac{1}{\lambda_e\mathbb{E}_e\left(T_+^e\right)}.$$

In particular, if L is recurrent positive,

$$\lim_{t\to\infty}\mathbb{P}(X_t = e) = \mu(e) > 0,$$

where μ is its unique invariant distribution, and otherwise

$$\lim_{t\to\infty}\mathbb{P}(X_t = e) = 0.$$

Proof: admitted. □

Definition 4.18: Reversible measure ★

A measure μ is called *reversible* w.r.t. an intensity matrix L if for any $e, e' \in E$,

$$\mu(e)\ell_{e,e'} = \mu(e')\ell_{e',e}.$$

Proposition 4.31: Invariant measures and semi-group

Any reversible measure w.r.t. L is invariant.

Proof: It is enough to write $\mu L(e') = \sum_{e\in E}\mu(e)\ell_{e,e'} = \sum_{e\in E}\mu(e')\ell_{e',e} = 0$. □

Remark: *Reversibility, as its name suggests, has to do with a process's time reversal. One can show in particular that given a Markov process (X_t) started from its invariant measure μ, for any $T > 0$, the process Y_t defined as the right-continuous version of $(X_{T-t})_{t\leq T}$ is also a Markov process, with invariant measure μ as well, and with jump rates*

$$\ell'_{e,e'} = \frac{\mu(e')\ell_{e',e}}{\mu(e)}.$$

In particular, if the measure μ is reversible, the time-reversed process started from its reversible state has the same jump rates as the original process, which means that the Markov process "looks" the same backward and forward in time.

4.6.4 Bonus: general Markov processes

When the state space is not countable, our construction of Markov processes is no longer valid, as Markov processes on continuous space are not necessarily jump processes. An important example is the Brownian motion, which has the defining property of Markov processes, namely the Markov property. We give here, the formal definition of general Markov processes.

Definition 4.19: Markov process

We say that $(X_t)_{t\geq 0}$ is a *Markov process* if it is càdlàg, and for any $s > t$, and any bounded function $f : E \to \mathbb{R}$,

$$\mathbb{E}(f(X_t) \mid X_s) = \mathbb{E}(f(X_t) \mid \mathscr{F}_s^X).$$

In other words, X_t depends on the past before time t, $(X_{t'})_{t'\leq s}$ only through X_s itself.
A Markov process is called *homogeneous* if the conditional distribution of X_t knowing $\mathscr{F}_t^X$ only dedends on $t - s$.

As is the case in coutable state spaces, general Markov processes can be characterized by their initial distribution μ and their infinitesimal generator $\mathscr{L}$. Then, one can define the semi-group $P_t = e^t L$ as the unique solution to the kolmogorov equation. Conversely, given the semi-group P_t of the Markov process, acting on bounded functions f as

$$P_t f(x) = \mathbb{E}(f(X_t) \mid X_0 = x),$$

the Markov infinitesimal generator can be defined as the operator

$$\mathscr{L} f(x) = \lim_{t \to 0} \frac{P_t f(x) - f(x)}{t}.$$

One can check that when the state space is countable, this definition is coherent with our construction of Markov processes. General construction of Markov processes, however, go beyond the scops of our course, so that we will not give more details on the subject.

5 Probabilistic concentration inequalities

Reading material: Concentration inequalities – a nonasymptotic theory of independence, by Stéphane Boucheron, Gábor Lugosi and Pascal Massart.

This section contains a selection of classical inequalities, mostly revolving around the law of large numbers. We start by recalling two elementary bounds, namely Markov's and Bienaymé-Chebychev's.

5.1 Reminder : classical inequalities

5.1.1 Markov's inequality

Markov's inequality allows one to estimate the probability that a non-negative random variable is large thanks to its expectation.

Proposition 5.1: Markov's inequality ★

Let X be a non-negative random variable, then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Remark: *As a consequence, for any increasing non-negative function* Φ*,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(\Phi(X))}{\Phi(a)}.$$

Proof: The proof is emmediate given the monotonicity of the expectation and the bound

$$a\mathbf{1}_{\{X \geq a\}} \leq X.$$

□

5.1.2 Chebychev's inequality

We now turn to (Bienaymé-)Chebychev's inequality , that gives one control on the distance of a random variable to its mean thanks to its variance.

Proposition 5.2: Chebychev's inequality

Let X be a square-integrable random variable, then for any $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{V(X)}{a^2}.$$

As a result, for any integer k,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq k\sigma(X)) \leq \frac{1}{k^2},$$

where $\sigma(X) = \sqrt{V(X)}$ is the standard deviation of X.

Proof: Chebychev's inequality is a direct consequence of Markov's, applied to the non-negative variable $(X - \mathbb{E}(X))^2$. □

5.1.3 Jensen's inequality

Jensen's inequality gives a very useful bound on $\mathbb{E}(\varphi(X))$ when φ is a convex function.

Proposition 5.3: Jensen's inequality ★

Let X be a random variable, φ a convex function, then

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)).$$

Remark: *To remember the direction of the inequality, consider the case of $\varphi(x) = x^2$, with a centered variable, $\mathbb{E}(X) = 0$. In this case, Jensen's inequality just states $\mathbb{E}(X^2) \geq 0$, which is trivially true.*

Proof: admitted. □

5.2 Concentration inequalities: the Gaussian case

By the law of large numbers, we have for an i.i.d. $(X_k)_{k\in\mathbb{N}}$

$$\frac{\sum_{k=1}^n X_k}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X_1).$$

One want to characterize this convergence, by finding an explicit function f such that

$$\mathbb{P}\left(\left|\sum_{k=1}^n [X_k - \mathbb{E}(X_k)]\right| > x\right) \leq f_n(x).$$

The CLT gives an *asymptotic* answer to this question, by choosing $f_n(x) = f(x/\sqrt{n})$, where f is the Gaussian error function, in which case the bound above is an equality in the limit $n \to \infty$.

In many cases, an asymptotic result is not enough, so that we want a bound for n fixed, called a *concentration inequality*. Such bounds are very useful to tackle a wide range of problems. We start by considering the case of i.i.d. Gaussian variables, in which case the empirical mean is also normally distributed and sharp explicit bounds can be obtained.

We now assume that $X_k \sim \mathcal{N}(0,1)$, in which case $S_n := \sum_{k=1}^n X_k \sim \mathcal{N}(0,n)$. Define

$$p_n(x) := \mathbb{P}(S_n \geq x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{t^2}{2n}\right) dt = \frac{1}{\sqrt{2\pi}} \int_{x/\sqrt{n}}^{+\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

The right-hand side is equal to $\frac{1}{2} erfc(x/\sqrt{2n})$, where *erfc* is called the complementary error function. A classical estimate on the error function yields the following sharp estimate.

Proposition 5.4

For any $n \geq 1$ and $x \geq 0$

$$\max\left\{0, \exp\left(-\frac{x^2}{2n}\right)\left(\frac{\sqrt{n}}{x} - \frac{\sqrt{n^3}}{x^3}\right)\right\} \leq p_n(x) \leq \exp\left(-\frac{x^2}{2n}\right)\frac{\sqrt{n}}{x}$$

Both the left-hand side and right-hand side are of the same order, therefore this bound is extremely sharp. In general, however, we do not have access to the exact distribution of S_n, so that in order to estimate the probability that S_n is far from its mean, we need to find other tools.

5.3 General case

5.3.1 Chernoff's inequality

We now consider a sequence of (non-necessarily identically distributed) independent variables $(X_n)_{n\in\mathbb{N}}$, we denote

$$S_n = \sum_{k=1}^n X_k$$

its partial sum, and

$$E_n = \mathbb{E}(S_n) = \sum_{k=1}^n \mathbb{E}(X_k) \quad \text{and} \quad V_n = \mathbb{E}((S_n - E_n)^2) = \sum_{k=1}^n V(X_k)$$

its expectation and variance.

Chernoff's inequality is a basic tool to exploit the X_k's independence to obtain a concentration bound.

Proposition 5.5: Chernoff's inequality

With the above notations,

$$\mathbb{P}(S_n \geq t) \leq \inf_{s\geq 0}\left\{e^{-st}\prod_{k=1}^n \mathbb{E}(e^{sX_k})\right\}.$$

Proof: Chernoff's bound is a straightforward consequence of Markov's inequality applied to the non-negative variables $e^{s\sum X_k}$. Note that if $\mathbb{E}(\exp(sX_k) = \infty$ for any s, Chernoff's bound is trivial, so that in order to obtain relevant bounds, one needs, at least in some segment $s \in [s_0, s_1]$, that sX_k have finite exponential moments. □

In some cases, the moments generating function $\mathbb{E}(\exp(sX_k))$ can be explicitly computed, and therefore optimizing over s yields a sharp bound.

5.3.2 Hoeffding's inequality

Hoeffding's inequality estimates the probability that the empirical mean S_n/n is far from its mean E_n/n for random variables taking values in a bounded domain.

Proposition 5.6: Hoeffding's inequality ★

Assume that there exists a constant $C > 0$ such that each X_k's state space has width at most C, in the sense that there exists two sequences $a_k \leq b_k$ such that

$$\mathbb{P}(a_k \leq X_k \leq b_k \ \forall k \in \mathbb{N}) = 1.$$

We denote $c_k = b_k - a_k \leq C$. Then, for any positive t,

$$\mathbb{P}(S_n - E_n \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \leq \exp\left(-\frac{2t^2}{nC^2}\right),$$

so that in particular

$$\mathbb{P}(|S_n - E_n| \geq t) \leq 2\exp\left(-\frac{2t^2}{nC^2}\right),$$

Proof: The proof is a consequence of *Hoeffding's Lemma*, that we will admit, that states that for any mean-0 variable X a.s. in $[a, b]$,

$$\mathbb{E}(e^{\lambda X}) \leq e^{\lambda^2(b-a)^2/8}. \qquad (5.1)$$

We will admit this bound, it is a consequence of optimizing a Jensen bound on $e^{\lambda X}$. We apply Hoeffding's Lemma and Chernoff's bound to $X = X_k - E(X_k)$, to obtain

$$\mathbb{P}(S_n - E_n \geq t) \leq e^{-st} \prod_{k=1}^n \mathbb{E}(e^{s[X_k - \mathbb{E}(X_k)]}) \leq e^{-st} e^{\sum_{k=1}^n s^2(b_k-a_k)^2/8}.$$

We optimize by taking $s = 4t/\left[\sum_{k=1}^n (b_k - a_k)^2\right]$, which proves the bound.

The estimate on the absolute value of $S_n - E_n$ is obtained by union bound. □

Remark: *By choosing $t = an$, Hoeffding's inequality gives a very strong convergence bound on the law of large numbers in the case of an i.i.d. sequence with bounded domain: letting $m = \mathbb{E}(S_n)/n = \mathbb{E}(X_1)$, the **strong law of large numbers** states that*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m,$$

and Hoeffding yields that this convergence is exponentially fast

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > a\right) \leq 2e^{-C'(a)n}$$

for $C'(a) = 2a^2/C^2 > 0$.

Further note that by CLT, in the case of i.i.d. variables for example, fluctuations of $S_n - E_n$ *around* 0 *are expected to be of the order* $\sqrt{n}$*, which is exactly the order Hoeffding's yields, since letting* $t \gg \sqrt{n}$ *yields a vanishing probability.*

Remark: *Hoeffding's inequality is a special case of a more general bound, Azuma's inequality, that also applies to more general types of martingales with bounded variations, sub-martingales and super-martingales for example.*

Remark: *The boundedness assumption may seem like it is extremely limiting, since most random variables one can think of are not bounded. This issue, however, can be solved when each* X_i*'s distribution tail decays exponentially, which will typically be the case since we want some large deviation type estimates (see last chapter). In this case,* $\mathbb{P}(|X_k| > M) = O(e^{-\alpha_k M})$*, so that we can apply Hoeffding's inequality to* $X_k^M := X_k \wedge M$*. By union bound, w.h.p* $1 - nO_M(\exp(-\min_k \alpha_k M))$*, all* X_k*'s and* X_k^M*'s are equal, so that Hoeffding's bound yields*

$$\mathbb{P}(S_n - E_n \geq t) \leq C_1 n \exp\left(-\min_k \alpha_k M\right) + \exp\left(-\frac{t^2}{2nM^2}\right)$$

where the first term represents the probability that one of the X_k^M *is different from* X_k*, and the second is Hoeffding's bound on the* X_k^M*'s. Setting* $M = O(\log n)$ *is generally sufficient to get rid of the first term, and does not loose too much on the second term. In the last chapter of this course, we will see some sharper bounds thanks to the theory of large deviations and Cramér's Theorem.*

5.3.3 First Bernstein inequality

One drawback of the Hoeffding inequality is that it becomes bad very fast as the support of the random variable increases, independently of the actual probability that the random variables are large. This issue is lifted in parts by the first Bernstein inequality.

Proposition 5.7: First Bernstein inequality

Under the same assumptions and hypothesis as in Hoeffding's inequality, for any positive t,

$$\mathbb{P}(|S_n - E_n| \geq t) \leq 2\exp\left(-\frac{2t^2}{V_n + Ct/3}\right).$$

Proof: admitted □

End of lecture 10

6 Large deviations principles

Reading material: Large deviations techniques and applications, by Amir Dembo, and Ofer Zeitouni.

6.1 Cramér's theorem

Chernoff's, Hoeffding's, and Bernstein's inequalities allow us to bound from above the probability that an empirical average deviates from its theoretical mean. In essence, these are large deviations bounds, which prove that an average deviation of $S_n - E_n$ of order ε occurs with a probability of order $\exp(-nf(\varepsilon))$. However, one can show, for random variables with sharply decaying tails, that the deviation probability is not only bounded from above, but also asymptotically equal to $\exp(-nf(\varepsilon))$. This is called a *large deviations principle*.

Although large deviations principles can be obtained for fairly general sequences of distributions, we will focus first on the distribution of a sequence of i.i.d. variables, through Cramér's theorem. We start by a key notion in the study of large deviations, namely Legendre's transform.

6.1.1 Legendre transform of convex functions

Definition 6.1: Legendre transform ★

Given a convex real-valued function Λ on $\mathbb{R}$, we define its *Legendre transform*, also called Cramér's transform in the context of large deviations, as the function

$$\Lambda^\star(x) = \sup_{t\in\mathbb{R}}\{xt - \Lambda(t)\} \in \mathbb{R} \cup \{+\infty\}.$$

This function, and a variable t_x at which the supremum is reached, is always well-defined if Λ is convex and $\Lambda^\star(x) < \infty$.

A graphical representation of the Legendre transform of a convex function Λ is represented in Figure 1. When the function Λ is not convex, this notion can be defined as well, and is then called *convex conjuguate*. In the context of this course, however, we will focus on the case where Λ is indeed convex.

Proposition 6.1: Properties of the Legendre transform ★

The legendre transform of a real-valued function is convex and lower semi-continuous, i.e. its level sets $\{\Lambda \le \alpha\}$ are closed for any α.

Proof: Both properties are quite elementary for real-valued functions. Fix $\alpha > 0$, and a sequence $(x_n)_{n\in\mathbb{N}}$ and assume that $\lim_{n\to\infty} x_n = x$ such that $\Lambda^\star(x_n) \le \alpha \ \forall n \in \mathbb{N}$, we will show that $\Lambda^\star(x) \le \alpha$. By the definition of the legendre transform, and our

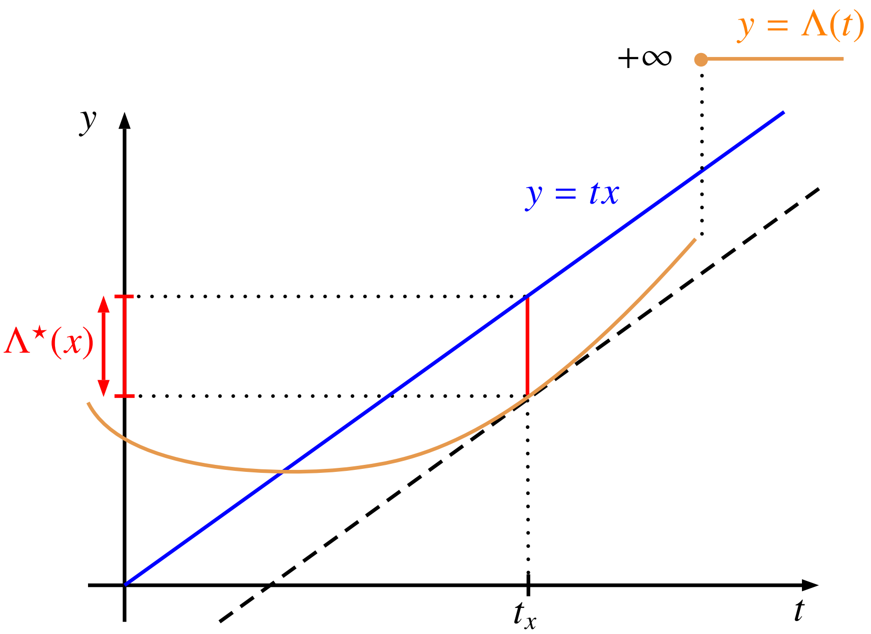


Figure 1: Representation of the convex tranform of Λ. One can see that $\frac{d}{dt}\Lambda(t_x) = x$, and conversely $\frac{d}{dx}\Lambda^\star(x) = t_x$.

assumption on the x_n's,

$$\sup_{\substack{t\in\mathbb{R}\\ n\in\mathbb{N}}}\{x_n t - \Lambda(t)\} \leq \alpha.$$

But for any fixed t, $tx - \Lambda(t) \leq \sup_n x_n t - \Lambda(t)$, so that taking the supremum over t proves $\Lambda^\star(x) \leq \alpha$, so that $\Lambda^\star$ is lower semi-continuous.

We now prove the convexity, assuming that Λ is convex. Fix $\theta \in [0,1]$, $x_1, x_2 \in \mathbb{R}$.

$$\begin{aligned}\Lambda^\star(\theta x_1 + (1-\theta)x_2) =& \sup_{t\in\mathbb{R}}\{\theta x_1 t + (1-\theta)x_2 t - \theta\Lambda(t) - (1-\theta)\Lambda(t)\}\\ \leq& \sup_{t_1,\, t_2\in\mathbb{R}}\{\theta x_1 t_1 + (1-\theta)x_2 t_2 - \theta\Lambda(t_1) - (1-\theta)\Lambda(t_2)\}\\ =&\theta\Lambda^\star(x_1) + (1-\theta)\Lambda^\star(x_2).\end{aligned}$$

□

Remark: *One can show that if Λ is convex and finite everywhere, then its Legendre transform $\Lambda^\star$ is C^∞ and strictly convex on the domain $(\Lambda^\star)^{-1}(\mathbb{R})$ where it is finite.*

Remark: *In general, for any convex function λ, for any x and t*

$$\Lambda(t) + \Lambda^\star(x) \geq tx, \tag{6.1}$$

with equality iff t and x are conjuguate of eachother.

We have the following result.

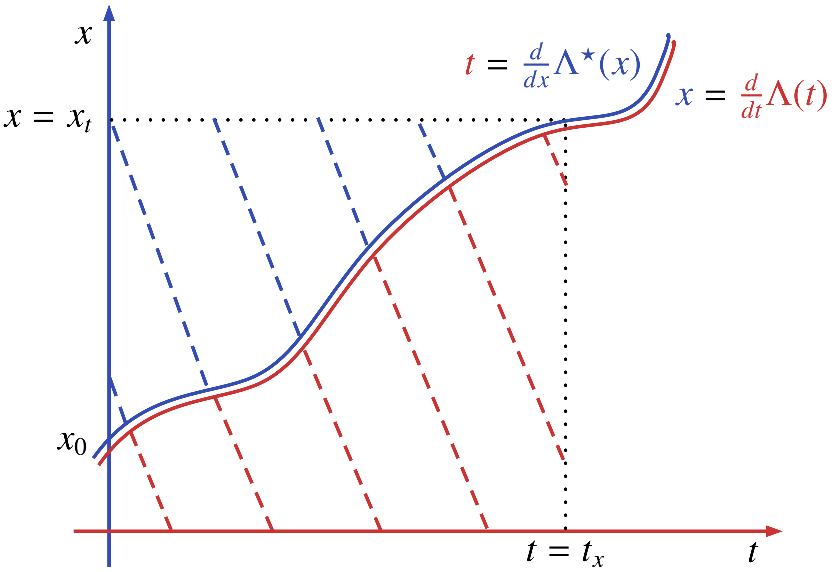

x
$x = x_t$
$t = \frac{d}{dx}\Lambda^\star(x)$
$x = \frac{d}{dt}\Lambda(t)$
x_0
$t = t_x$
t

Figure 2: Representation of the convex conjugation. Up to a constant, $\Lambda(t)$ is given by the red area, whereas $\Lambda^\star(x)$ is given by the blue area. In particular, up to constants, their sum adds up to tx.

Proposition 6.2: Convex conjugates ★

The Legendre transform is an involution on the set of convex functions, i.e. $(\Lambda^\star)^\star = \Lambda$ for any convex function Λ. Furthermore, if Λ is differentiable and strictly convex on $\mathbb{R}$, the derivatives of Λ and $\Lambda^\star$ are inverse of each other, i.e. $\forall t \in \mathbb{R}$

$$\frac{d}{dx}\Lambda^\star\left(\frac{d}{dt}\Lambda(t)\right) = t.$$

Furthermore, for any strictly convex function Λ, the variables satisfying the supremum are conjugate, in the sense that if

$$\Lambda^\star(x) = xt_x - \Lambda(t_x) \quad \text{and} \quad \Lambda(t) = tx_t - \Lambda^\star(x_t),$$

then $x_{t_x} = x$ and

$$x_t = \frac{d}{dt}\Lambda(t), \quad \text{and} \quad t_x = \frac{d}{dx}\Lambda^\star(x).$$

Proof: The last statement is proved in Figure 1. To prove that the variables (t, x) and the functions Λ, $\Lambda^\star$ are conjugate, we write as represented in Figure 2, that

$$tx = \int_0^t x_s ds + \int_{x_0}^x t_{x'} dx'.$$

The first term in the right-hand side is $\Lambda(t) - \Lambda(0)$, so that the equality case in

(6.1) identifies $\int_{x_0}^{x} t_{x'} dx'$ as $\Lambda^\star(x) - \Lambda(0)$, so that in particular

$$\frac{d}{dx}\Lambda^\star(x) = t_x = \left(\frac{d}{dt}\Lambda\right)^{-1}(x).$$

□

6.1.2 Cramér's transform

Consider a sequence $(X_k)_{k\in\mathbb{N}}$ of i.i.d. random variables.

Definition 6.2: logarithmic moments generating function (log-MGF) ★

We define the *logarithmic moment generating function* (log-MGF) Λ_X of a random variable X as the function

$$\Lambda_X(t) := \log \mathbb{E}(\exp(tX)) \in \mathbb{R} \cup \{+\infty\}.$$

It is also called the *cumulant generating function.*

One easily checks that

$$\Lambda_X(0) = 0, \qquad \frac{d}{dt}\Lambda_X(0) = \mathbb{E}(X).$$

Proposition 6.3: Convexity of the log-MGF

For any real-valued random variable X, its logarithmic moments generating function is convex, i.e. for any t_1, t_2, and $\theta \in [0,1]$,

$$\Lambda_X(\theta t_1 + (1-\theta)t_2) \le \theta\Lambda_X(t_1) + (1-\theta)\Lambda_X(t_2).$$

Proof: Recall that Holder's inequality yields, for $1/p + 1/q = 1$, that

$$\mathbb{E}(YZ) \le \mathbb{E}(Y^p)^{\frac{1}{p}}\mathbb{E}(Z^q)^{\frac{1}{q}}.$$

We apply it to $Y = e^{\theta t_1 X}$, $Z = e^{(1-\theta)t_2 X}$, $p = \theta^{-1}$, $q = (1-\theta)^{-1}$, to obtain that

$$\mathbb{E}(\exp(\theta t_1 X + (1-\theta)t_2 X)) \le \mathbb{E}(\exp(t_1 X))^\theta \mathbb{E}(\exp(t_2 X))^{1-\theta},$$

taking the log on both sides proves the convexity of Λ_X. □

Proposition 6.4: Legendre transform of the log-MGF

Fix a real-valued random variable X, denote by Λ_X its log-MGF. Assume that Λ_X is finite on an open set $]-\varepsilon,\varepsilon[$ containing the origin.

Then, the function $\Lambda_X^\star$ is convex and non-negative, reaches its minimum

$\Lambda_X^\star(x_0) = 0$ at $x_0 = \mathbb{E}(X)$, and is non-increasing on $(-\infty, \mathbb{E}(X)]$, and non-decreasing on $[\mathbb{E}(X), +\infty)$.

Proof: The convexity of $\Lambda_X^\star$ is a direct consequence of Propositions 6.1 and 6.3. It is non negative because $xt - \Lambda_X(t)$ vanishes at $t = 0$.

We now write that $\Lambda_X^\star(\mathbb{E}(X)) = \sup_t\{t\mathbb{E}(X) - \log(\mathbb{E}(e^{tX}))\}$. By Jensen's inequality, the second term $\log(\mathbb{E}(e^{tX}))$ is larger than $t\mathbb{E}(X)$ for any $t \in \mathbb{R}$, therefore $\Lambda_X^\star(\mathbb{E}(X)) = 0$. Since $\Lambda^\star$ is non-negative, $\mathbb{E}(X)$ realizes its minimum. Regarding the monotonicity, we write by convexity that for any $x > 0$, any $\theta \in [0, 1]$

$$\Lambda_X^\star(\theta\mathbb{E}(X) + (1-\theta)x) \leq (1-\theta)\Lambda_X^\star(x) \leq \Lambda_X^\star(x),$$

which proves that $\Lambda_X^\star$ is non-decreasing on $[\mathbb{E}(X), +\infty)$. The rest of the statement is proved analogously. □

6.1.3 Cramér's theorem

We now have all the tools needed to state Cramér' theorem, which explicitly identifies a large deviation principle for i.i.d. variables with finite log-MGF.

Theorem 6.5: Cramér's theorem ★

Consider an i.i.d. sequence $(X_k)_{k\in\mathbb{N}}$ of real-valued random variables, and assume that it has finite log-MGF on an open set containing 0, i.e.

$$\mathbb{E}(e^{tX_1}) < \infty \quad \forall t \in (-\varepsilon, \varepsilon).$$

Then, for any $x > \mathbb{E}(X_1)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq nx) = -\Lambda_{X_1}^\star(x). \tag{6.2}$$

More generally, for any for any $a < b \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n/n \in [a, b]) = -\inf_{x\in[a,b]} \Lambda_{X_1}^\star(x).$$

Remark: *In a completely symmetric way, Cramér's theorem also states that for any $x < \mathbb{E}(X_1)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \leq nx) = -\Lambda_{X_1}^\star(x).$$

Informally, Cramer's large deviation principle can be written

$$\mathbb{P}(S_n/n \simeq x) \sim \exp(-n\Lambda_{X_1}^\star(x)). \tag{6.3}$$

Note that when $x = \mathbb{E}(X_1)$, Cramér's theorem does not give any information, since the Probability that S_n/n is close to x is close to 1 and does not decay exponentially.

Remark 1: *Cramér's Theorem states, informally, that the way to create an unlikely large deviation is by creating the most likely among the possible unlikely deviations. More precisely, write informally that*

$$\mathbb{P}(S_n/n \in [a,b]) = \int_a^b \mathbb{P}(S_n/n \in [x, x+dx]).$$

However, for x, x' such that $\Lambda^\star_{X_1}(x) < \Lambda^\star_{X_1}(x')$

$$\mathbb{P}(S_n/n \in [x, x+dx]) \simeq e^{-n\Lambda^\star_{X_1}(x)} \gg e^{-n\Lambda^\star_{X_1}(x')} \simeq \mathbb{P}(S_n/n \in [x', x'+dx]),$$

so that the sum of these two probabilities is asymptotically equal to the larger of the two. Generalizing to all points $x \in [a,b]$, the remaining contribution is the largest, namely

$$\exp\left(-n \inf_{x\in[a,b]} \Lambda^\star_{X_1}(x)\right).$$

In particular, to create a deviation of at least *ε from the mean $\mathbb{E}(X)$, the most likely way is to create a deviation of* exactly *ε, so that*

$$\mathbb{P}(S_n/n \geq \mathbb{E}(X) + \varepsilon) \simeq \mathbb{P}(S_n/n \simeq \mathbb{E}(X) + \varepsilon) \simeq \exp\left(-n\Lambda^\star_{X_1}(\mathbb{E}(X)+\varepsilon)\right). \tag{6.4}$$

Proof: We give the proof in the case where X_1 has finite log-MGF everywhere, in order not to burden with technical details. We therefore assume that

$$\mathbb{E}(e^{tX_1}) < \infty \quad \forall t \in \mathbb{R}. \tag{6.5}$$

All large deviations principles are typically composed of a lower bound and an upper bound, which are proven separately.

We first give the upper bound, which is a straightforward consequence of Chernoff's inequality. The latter yields

$$\begin{aligned}\mathbb{P}(S_n \geq nx) \leq \inf_{s\geq 0}\left\{e^{-snx}\prod_{k=1}^n \mathbb{E}(e^{sX_k})\right\} &= \inf_{s\geq 0} \mathbb{E}(e^{sX_1})^n e^{-snx} \\ &= \inf_{s\geq 0} \exp\left(-n[sx - \Lambda_{X_1}(s)]\right) \leq \exp\left(-n\Lambda^\star_{X_1}(x)\right),\end{aligned}$$

which proves the upper bound. Note that the upper bound holds for any fixed n, not only in the limit $n \to \infty$.

Proving the lower bound always revolves around the same scheme : to estimate the probability of deviating to the value x, we first tilt the distribution of the X_i's to make this deviation typical, and then estimate the cost of doing so. First, we claim that to prove the lower bound, it is enough to show that for any $x \in \mathbb{R}$, any $\delta > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(nx \leq S_n \leq n(x+\delta)) \geq -\Lambda^\star_{X_1}(x). \tag{6.6}$$

Indeed, assume that the bound above holds for any x, δ. Then,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq nx) \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(nx \leq S_n \leq n(x+\delta)) \geq -\Lambda^\star_{X_1}(x),$$

which proves the lower bound we wanted.

Remark: *At a first glance, it may look like we are loosing much when writing the bound above, since we estimated $\mathbb{P}(S_n \geq nx)$ by $\mathbb{P}(nx \leq S_n \leq n(x+\delta))$, so that we have no hope of ultimately obtaining a sharp bound. This is not the case, however, thanks to* (6.4)*, since to create a deviation mean deviation* above *x, the most likely way is to create a deviation* exactly at *x (See Remark* 1 *above).*

We now fix once and for all $x^\star > \mathbb{E}(X)$, and prove (6.6) for $x = x^\star$. Fix $\delta > 0$, and let $t^\star \in \mathbb{R}$ such that

$$\Lambda^\star_{X_1}(x^\star) = t^\star x^\star - \Lambda_{X_1}(t^\star).$$

In other words, $t^\star$ solves the supremum in the definition of $\Lambda^\star_{X_1}(x^\star)$, and one can check (see Figure 1 and Proposition 6.2) that it is characterized by $t^\star = \frac{d}{dx}\Lambda^\star_{X_1}(x^\star)$, and conversely

$$x^\star = \frac{d}{dt}\Lambda_{X_1}(t^\star) = \frac{\mathbb{E}(X_1 e^{t^\star X_1})}{\mathbb{E}(e^{t^\star X_1})}.$$

Given the distribution ν of X_1, we define $\nu_{x^\star}$ as the tilted distribution

$$\nu_{x^\star}(dx) := \exp\left(t^\star x - \Lambda_{X_1}(t^\star)\right)\nu(dx) = \frac{e^{t^\star x}}{\mathbb{E}_\nu(e^{t^\star x})}\nu(dx). \tag{6.7}$$

We represent in Figures 3 and 3 the tilted log-MGF $\Lambda_{x^\star}$, its derivative, and its legendre transform.

Note in particular that

$$\int x\nu_{x^\star}(dx) = \int x \exp\left(t^\star x - \Lambda_{X_1}(t^\star)\right)\nu(dx) = x^\star,$$

so that $x^\star$ is the average value of the tilted distribution $\nu_{x^\star}$. We can now write for any $0 < \varepsilon < \delta$

$$\begin{aligned}
\mathbb{P}(nx^\star \leq S_n \leq n(x^\star+\delta)) \geq & \mathbb{P}(nx^\star \leq S_n \leq n(x^\star+\varepsilon)) \\
&= \int_{nx^\star \leq \sum x_k \leq n(x^\star+\varepsilon)} \nu(dx_1)\dots\nu(dx_n) \\
&\geq \int_{nx^\star \leq \sum x_k \leq n(x^\star+\varepsilon)} \exp\left\{n\Lambda_{X_1}(t^\star) - t^\star \sum_{k=1}^n x_k\right\}\nu_{x^\star}(dx_1)\dots\nu_{x^\star}(dx_n) \\
&\geq \exp\left\{n\Lambda_{X_1}(t^\star) - nt^\star(x^\star+\varepsilon)\right\}\mathbb{P}(0 \leq S^\star_n - nx^\star \leq n\varepsilon) \\
&= \exp\left\{-n\Lambda^\star_{X_1}(x^\star) - nt^\star\varepsilon\right\}\mathbb{P}\left(0 \leq \frac{S^\star_n - nx^\star}{\sqrt{n}} \leq \sqrt{n}\varepsilon\right),
\end{aligned}$$

where $S^\star_n$ is a sum of n independent random variables with distribution $\nu_{x^\star}$. By the CLT, the probability on the right-hand side converges as $n \to \infty$ to $1/2$, therefore

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(|S_n - nx^\star| \leq n\delta) \geq -\Lambda^\star_{X_1}(x^\star) - t^\star\varepsilon.$$

Letting $\varepsilon \to 0$ proves (6.6). □

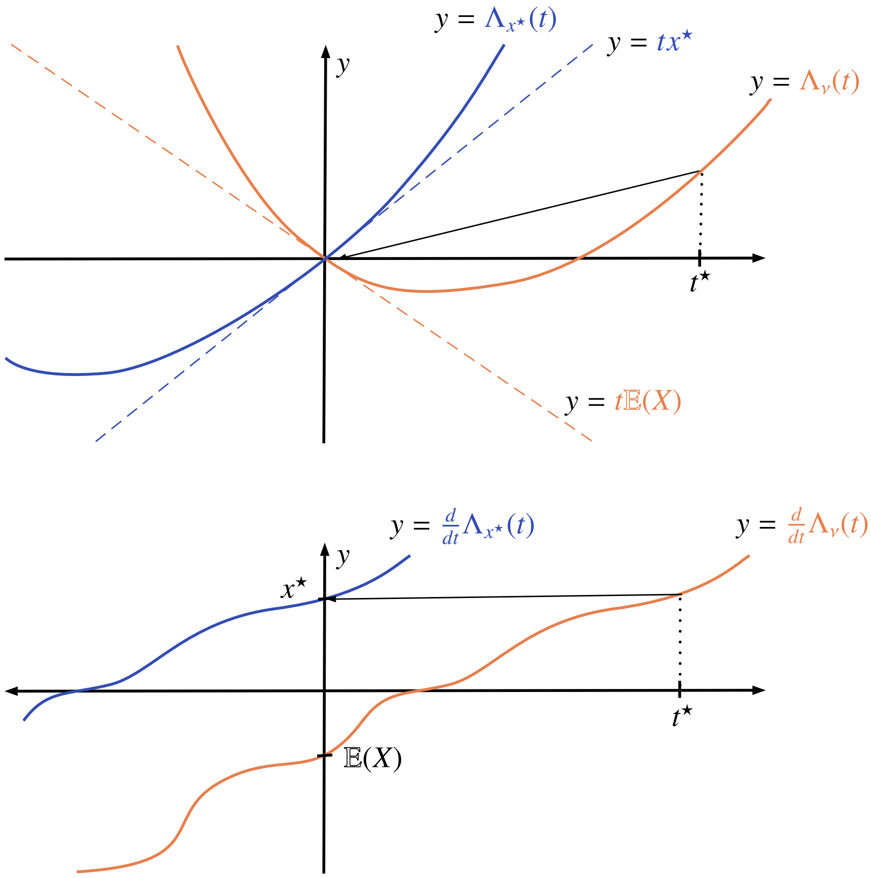


Figure 3: Representation of the log-MGF Λ_ν (orange) and the tilted log-MGF $\Lambda_{x^\star}(t)$ (blue) and their derivatives.

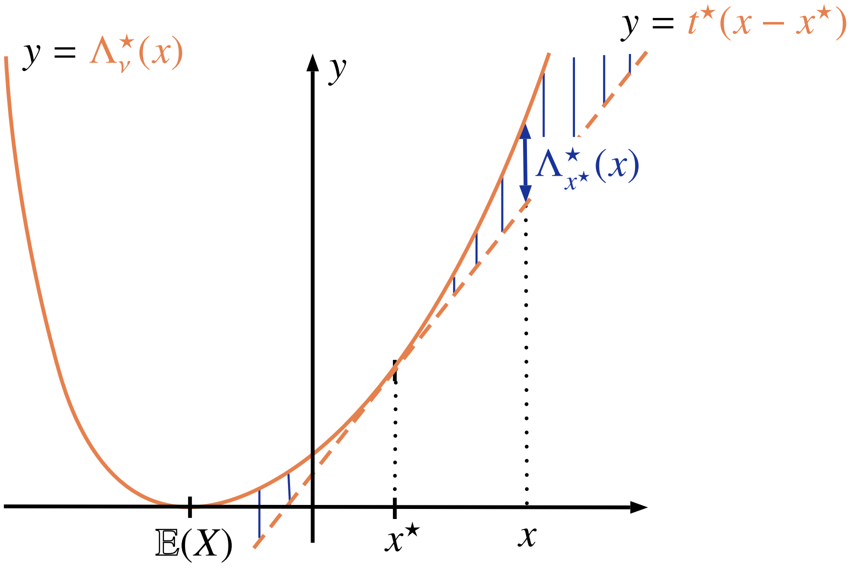

$y = \Lambda_\nu^\star(x)$
y
$y = t^\star(x - x^\star)$
$\Lambda_{x^\star}^\star(x)$
$\mathbb{E}(X)$
$x^\star$
x

Figure 4: Representation of the Legendre transform $\Lambda^\star(x)$ (orange) and the tilted legendre transform $\Lambda^\star_{x_\star}$ (blue).

Exercise 24 : Bernoulli random variables

Fix $p \in (0,1)$, and (X_k) an i.i.d. $Ber(p)$ variables. Recall that $S_n = \sum_{k=1}^n X_k$.

1) Compute the log-MGF of X_1.

2) Show that its Legendre transform is given by

$$\Lambda^\star_{X_1}(x) = \begin{cases} x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right) & \text{for } x \in [0,1] \\ +\infty \text{ for } x \notin [0,1] \end{cases} .$$

3) Compute $\mathbb{P}(S_n/n \geq 1)$, and show directly that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq n)$$

exist.

Answer :

1) we straightforwardly write

$$\mathbb{E}(e^t X_1) = pe^t + 1 - p,$$

so that

$$\Lambda_{X_1} = \log(pe^t + 1 - p).$$

2) One needs to find a solution to

$$\frac{d}{dt}\left\{xt - \log\left(pe^t + 1 - p\right)\right\} = 0 \quad \Rightarrow \quad t_x := \log\frac{(1-p)x}{(1-x)p} \quad \text{for } x \in [0,1].$$

One easily checks that for $x \notin [0,1]$, no solution exist. Since Λ is convex, so is $xt - \Lambda(t)$, therefore its minimum is given by the identity above. Computing $xt_x - \Lambda(t_x)$ yeilds the formula.

3) $\mathbb{P}(S_n/n \geq 1) = \mathbb{P}(S_n = n) = p^n$. The limit exist, and is equal to $\log p$. □

⚠

WARNING : The identity (6.2) *does not hold at the points of discontinuity of $\Lambda^\star_{X_1}$ if the inequality becomes strict. For example, as we have just seen for Bernoulli variables, $\Lambda^\star_{X_1}(1) = \log p$, but $\mathbb{P}(S_n > n) = 0$, so that*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > n) = -\infty \neq \Lambda^\star_{X_1}(1).$$

The reason for this discrepancy at the boundary will be made clear in the next section.

——— *End of lecture 11* ———

6.2 Large deviations principles

Cramér's theorem is a particular case of a more general notion called *Large deviation principles*. Cramér's theorem, more precisely, gives information on the distribution μ_n of the empirical average S_n of i.i.d. random variables with finite exponential moments. Large deviations principles are actually possible for general sequences of distributions. To introduce a little bit more generality, we will introduce our results for distributions on $\mathbb{R}^d$, but large deviations principles are actually available for very general topological spaces. For now, we consider $E = \mathbb{R}^d$.

Definition 6.3: (Good) rate function ★

A function $I : E \to [0, +\infty]$ is called a *rate function* if it is not identically equal to $+\infty$ and is lower semi-continuous, i.e. for any $a \in \mathbb{R}$, its lower level sets

$$D_a := \{x \in E, I(x) \leq a\}$$

are *closed sets*.
If instead they are *compact*, I is called a *good rate function*.

Proposition 6.6: infimum of a rate function

Let I be a good rate function, for any closed set $F \subset E$, there exists $x \in F$

$$I(x) = \inf_{y \in F} I(y).$$

Notation: in what follows, for any good rate function I, and any set $A \subset E$, we shorten

$$I(A) := \inf_{y \in A} I(y).$$

Definition 6.4: Large deviations principle ★

A sequence $(\mathbb{P}_n)_{n\in\mathbb{N}}$ of probability distributions on E is said to satisfy a *large deviations principle* with speed n and good rate function I, noted $LDP(n, I)$ if

i) I is a good rate function,

ii) for any closed set $F \subset E$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(F) \leq -I(F),$$

iii) for any open set $O \subset E$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(O) \geq -I(O).$$

For any set $A \subset E$, we denote by $\mathring{A}$ its interior, and $\bar{A}$ its closure. If $(\mathbb{P}_n)$ satisfies a $LDP(n, I)$, then for any A such that $I(\mathring{A}) = I(\bar{A})$,

$$\exists \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(A) = -I(A). \tag{6.8}$$

Remark: *as we have seen for the Cramér's theorem, proving a large deviations principles typically involves proving separately the lower and upper bound. Reformulated with the language of the previous definition, Cramér's theorem implies that given a sequence of i.i.d. random variables with finite exponential moments in a non-empty segment, the sequence $(\mathbb{P}_n)_{n\in\mathbb{N}}$ of distributions of the random variables S_n/n satisfies a large deviation principle with speed n, and whose rate function is given by the Legendre transform $\Lambda^\star_{X_1}$ of the X_i's* log*-MGF.*

Cramér's theorem, however, yields a stronger result than a standard large deviations principle, since it yields that for any closed set F,

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(F) = -I(F).$$

> *WARNING : a large deviations principles only yields upper and lower* ***bounds****, and the limit in the left hand side (6.8) can actually exist even when $I(\mathring{A}) \neq I(\bar{A})$. For example, consider the case of i.i.d. Bernoulli variables, and $A = [1, +\infty)$. Then as seen in Exercise 24, $I(\mathring{A}) = I((1, +\infty)) = +\infty$, and the large deviations lower bound becomes meaningless, although $\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(A) = \Lambda^\star_{X_1}(1) = \log p$.*

Remark: *Property iii) in the large deviations principle is equivalent to the following local property*

$$\text{“}\textit{for any } x \in E,\ r > 0, \quad \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(B_r(x)) \geq -I(x)\text{”},$$

where $B_r(x)$ is the Euclidean ball with radius r centered in x.

Proposition 6.7

If a sequence P_n satisfies a large deviation principle, the rate function is unique.

Proof: Assume by contradiction that two rate functions I_1 and I_2 satisfy $I_1(x) > I_2(x)$. Since I_1 is lower semi-continuous $I_1^{-1}((I_2(x), +\infty))$ is an open set, therefore there exists a ball $B_\delta(x)$ such that $I_1(y) > I_2(x)$ for any $y \in \bar{B}_\delta(x)$. Then, since we have both large deviations principles, we have

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(B_\delta(x)) \leq -I_1(\bar{B}_\delta(x)) < -I_2(x) = \lim_{\varepsilon\to 0} -I_2(B_\varepsilon(x))$$
$$\leq \limsup_{\varepsilon\to 0}\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(B_\varepsilon(x)) \leq \limsup_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n(B_\delta(x)),$$

which proves the contradiction by choosing a sequence δ_n realizing the $\limsup_{\varepsilon\to 0}$ above. □

Proposition 6.8

If a sequence $(\mathbb{P}_n)_{n\in\mathbb{N}}$ satisfies a large deviation principle there exists $x^\star$ such that $I(x^\star) = 0$, and if such a $x^\star$ is unique, then for any function bounded and continuous function f,

$$\int f(x)\mathbb{P}_n(dx) \underset{n\to\infty}{\to} f(x^\star).$$

Proof: admitted. □

6.3 Gärtner-Ellis theorem

The Gärtner-Ellis theorem is a general result to derive large deviations principles for general sequences of probability distributions. We consider random vectors Z_n taking value in $E = \mathbb{R}^d$, and denote $\mathbb{P}_n = \mathscr{L}(Z_n)$ the distribution of Z_n. Note that we **do not** assume that the Z_n take the form $n^{-1}\sum_{k=1}^n X_k$ for an i.i.d. sequence $(X_k)_k$.

Given a vector $\boldsymbol{t} = (t_1, \dots t_d)$, we define the cumulant generating function

$$\Lambda_n(\boldsymbol{t}) = \log \mathbb{E}(e^{\langle \boldsymbol{t}, Z_n\rangle}),$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product in $\mathbb{R}^d$. We assume that there exists a function $\Lambda : E \to \mathbb{R}\cup\{+\infty\}$ such that

$$\frac{\Lambda_n(n\boldsymbol{t})}{n} \underset{n\to\infty}{\longrightarrow} \Lambda(\boldsymbol{t}) \quad \forall \boldsymbol{t} \in \mathbb{R}^d. \tag{$\star$}$$

We define

$$D_\Lambda^\circ := \{\boldsymbol{t} \in E,\ \Lambda(\boldsymbol{t}) < +\infty\}$$

As in the one-dimensional case, the function Λ_n is convex (same proof).

Definition 6.5: Legendre transform and exposing hyperplanes

We define the *Legendre transform* of Λ as the function $\Lambda^\star$ on $\mathbb{R}^d$

$$\Lambda^\star(\boldsymbol{x}) = \sup_{t\in\mathbb{R}}\{\langle \boldsymbol{x}, \boldsymbol{t}\rangle - \Lambda(\boldsymbol{t})\} \in [0, +\infty].$$

This function is convex and lower semi-continuous.
A point $\boldsymbol{x}^\star \in \mathbb{R}^d$ is an *exposed point* of $\Lambda^\star$ if there exists $\boldsymbol{t} \in \mathbb{R}^d$ such that for any $\boldsymbol{x} \neq \boldsymbol{x} \in \mathbb{R}^d$

$$\Lambda^\star(\boldsymbol{x}) - \Lambda^\star(\boldsymbol{x}^\star) > \langle \boldsymbol{x} - \boldsymbol{x}^\star, \boldsymbol{t}\rangle,$$

in which case $\boldsymbol{t}$ is called an exposing hyperplane for $\boldsymbol{x}^\star$.

Remark: *The exposed points of $\Lambda^\star$ are those where $\Lambda^\star$ is strictly convex, i.e. points where it is strictly above one of its "tangents". The problem is that $\Lambda^\star$ is not necessarily differentiable at its exposed points, so that the tangent is not necessarily well defined. This is the reason why we do not give it as a definition.*

Definition 6.6: Essential smoothness

A convex function $\Lambda : \mathbb{R}^d \to (-\infty, +\infty]$ is *essentially smooth* if

i) D_Λ° is non-empty

ii) Λ is differentiable throughout D_Λ°

iii) For any sequence $\boldsymbol{t}_n$ converging to $\boldsymbol{t}$ in the boundary of D_Λ°, we have $\nabla\Lambda(\boldsymbol{t}_n) \to \infty$.

We can now state the Gärtner-Ellis theorem, which is the main result of this section. It gives a fairly general tool to prove that a sequence of distributions satisfies a large deviations principle, assuming that its sequence of log-moments generating function admits a well defined scaling limit.

Theorem 6.9: Gärtner-Ellis theorem ★

Assume that ($\star$) holds, then

1. for any closed set $F \subset E$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(F) \leq -I(F),$$

2. for any open set $O \subset E$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n(O) \geq -I(O \cap \mathscr{E}),$$

   where $\mathscr{E}$ is the set of exposed points of $\Lambda^\star$ whose exposing hyperplane $\boldsymbol{t}$ belongs to D_Λ°. Note that since $I(O\cap\mathscr{E})$ is a priori different from $I(O)$, the lower bound above does not guarantee a $LDP(n, \Lambda^\star)$.

3. If Λ is essentially smooth and lower-semicontinuous, then $(\mathbb{P}_n)$ satisfies a $LDP(n, \Lambda^\star)$.

We will not prove this result, however the strategy of its proof relies on roughly the same ingredients as Cramër's, in particular, in the case where $D_\Lambda^\circ = \mathbb{R}^d$. The upper bound relies on Chebychev's inequality, whereas the lower bound is proved using a change of reference measure to make to relevant unlikely event typical.

Remark: *It might seem unclear why the lower bound only holds for exposed hyperplanes. The reason for it is that once the tilting has been operated to derive* *the lower bound (see the proof of Cramér's Theorem 6.5, if it has been done at an exposed hyperplane at $x^\star$ (see Fig. 4, the resulting tilted rate function reaches a unique global minimum at $x^\star$, which yields that the second term in the lower bound vanishes.*

Exercise 25 : Exponential random variables

We want to show that in some cases, a LDP can be satisfied even when the Gärtner-Ellis theorem is not fully applicable. We consider $Z_n \sim Exp(n)$, and $\mathbb{P}_n$ their distribution.

1) (i) Compute the log-MGF Λ_n of Z_n.

(ii) Show that there exists a limiting function Λ satisfying ($\star$).

(iii) What is the density of Z_n ? Show, with direct computation, that $(\mathbb{P}_n)$ has a large deviations principle with a speed and a good rate function to be determined.

2) (i) Compute $\Lambda^\star$.

(ii) Show that $\mathscr{E} = \{0\}$. What is the lower bound given by the Gärtner-Ellis theorem ?

(iii) Conclude.

Answer :

1) (i) We compute

$$\Lambda_n(t) = \log \mathbb{E}(e^{tZ_n}) = \log\left(n\int_0^{+\infty} e^{tz}e^{-nz}dz\right) = \begin{cases} +\infty & \text{if } t \geq n \\ \log\frac{n}{n-t} & \text{if } t < n \end{cases}$$

(ii) One can check that

$$\Lambda(t) = \begin{cases} +\infty & \text{if } t \geq 1 \\ 0 & \text{if } t < 1 \end{cases}$$

satisfies ($\star$).

(iii) Direct computations yield that for any $a < b$,

$$\mathbb{P}(Z_n \in [a,b]) = \mathbb{P}(Z_n \in (a,b)) = e^{-na} - e^{-nb},$$

so that Z_n satisfies a LDP with good rate function

$$I(x) = \begin{cases} +\infty & \text{if } t \leq 0 \\ x & \text{if } t > 0. \end{cases}$$

2) (i) It is easily checked that $\Lambda^\star(x) = I(x)$.

(ii) The only point where $\Lambda^\star$ is strictly convex is $\{0\}$.

(iii) For any set not containing the origin, the Gärtner-Ellis lower bound is trivial, despite $\mathbb{P}_n$ satisfying a LDP. □

End of lecture 12

6.4 Bonus : Legendre transform and entropy

6.4.1 Relative entropy

The concept of large deviations is intimately linked to the notion of *relative entropy*. As we did for large deviations, we will present the topic in the context of real-valued variables, but everything is valid in a fairly general setting.

Definition 6.7: Relative entropy

Fix two measures μ and ν on $\mathbb{R}$, and assume that μ is absolutely continuous w.r.t. ν. We define the relative entropy of μ w.r.t. ν as

$$H(\mu \mid \nu) = \int \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) d\nu = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu.$$

Relative entropy is a useful concept, because it links the expectation of a function under μ to the one under ν, as guaranteed by the following proposition.

Proposition 6.10: Entropy inequality

Fix two measures μ and ν on $\mathbb{R}$, and assume that μ is absolutely continuous w.r.t. ν. Given a measurable bounded function f, we have

$$\mathbb{E}_\mu(f) \leq H(\mu \mid \nu) + \log \mathbb{E}_\nu(e^f).$$

Note that to estimate the expectation of f under μ, we loose something by switching to the reference measure ν, since now what we need to estimate $\log \mathbb{E}(e^f)$, which is much larger. This proposition actually stems from a second definition of the relative entropy:

Definition 6.8: Variational principle for the relative entropy

The relative entropy is also given by

$$H(\mu \mid \nu) = \sup_f \left\{\mathbb{E}_\mu(f) - \log \mathbb{E}_\nu(e^f)\right\}, \tag{6.9}$$

where the supremum is taken over bounded functions f.

Remark: *This second definition is much more general than the first, and is actually valid even if μ is not absolutely continuous w.r.t. ν, in which case the relative entropy is equal to $+\infty$.*

To see why Definition 6.8 holds, we claim that we can define, for any vector space E, and any function $\Lambda : E \to \mathbb{R}$, the Legendre transform $\Lambda^\star : E' \to \mathbb{R}$, where

E' denotes the dual space of E, as

$$\Lambda^\star(\varphi) = \sup_e \{\langle \varphi, e\rangle - \Lambda(e)\},$$

where φ is a linear form on E, and $\langle \varphi, e\rangle = \varphi(e)$ is the canonical pairing. Consider the vector space E of bounded functions on $\mathbb{R}$, its dual space E' can be seen (see the Riesz–Markov–Kakutani representation theorem) as the set $\mathcal{D}(\mathbb{R})$ of distributions on $\mathbb{R}$.

In particular, given a reference measure ν, Equation (6.9) identifies $H(\mu \mid \nu) = \Lambda^\star(\mu)$ as the Legendre transform of the functional $\Lambda : E \to \mathbb{R}$ defined as

$$\Lambda(f) := \log \mathbb{E}_\nu(e^f).$$

Admitting that this functional is strictly convex, as seen in Proposition 6.2, with any bounded function f can be associated a unique conjugate measure μ. Let us proceed by analogy between the two cases.

	$\Lambda : \mathbb{R} \to \mathbb{R}$	$\Lambda : E \to \mathbb{R}$
Variable	$t \in \mathbb{R}$	$f : \mathbb{R} \to \mathbb{R}$ bounded
Conjugate variable	$x \in \mathbb{R}$	μ distribution
Inner product	$\langle x, t\rangle = xt$	$\langle \mu, f\rangle := \int f d\mu$
Legendre transform	$\Lambda^\star(x)$	$\Lambda^\star(\mu)$
Conjugation relations	$\frac{d}{dt}\Lambda(t) = x$	$\frac{\delta\Lambda}{\delta f}[f; g] = \langle \mu, g\rangle$
	$\frac{d}{dx}\Lambda^\star(x) = t$	$\frac{\delta\Lambda^\star}{\delta\mu}[\mu; \pi] = \langle \pi, f\rangle$

In the identities above, the functional derivative $\frac{\delta\Lambda}{\delta f}$ at f is defined as

$$\frac{\delta\Lambda}{\delta f}[f; g] = \lim_{\varepsilon\to 0} \frac{\Lambda(f + \varepsilon g) - \Lambda(f)}{\varepsilon}.$$

In our case, according to (6.9), we are looking at the Legendre transform of the functional $\Lambda(f) = \log \mathbb{E}_\nu(e^f)$, whose functional derivative is given by

$$\frac{\delta\Lambda}{\delta f}[f; g] = \frac{\int g e^f d\nu}{\int e^f d\nu} = \int g d\mu,$$

according to the first conjugation relation. This identity must be true for every function g, we deduce that

$$e^f = \frac{d\mu}{d\nu} \qquad \Rightarrow \qquad f = \log\left(\frac{d\mu}{d\nu}\right), \tag{6.10}$$

where $d\mu/d\nu$ is the Radon derivative of μ w.r.t. ν. In particular,

$$H(\mu \mid \nu) := \Lambda^\star(\mu) = \langle \mu, \log\left(\frac{d\mu}{d\nu}\right)\rangle - \Lambda\left(\log\left(\frac{d\mu}{d\nu}\right)\right) = \int \frac{d\mu}{d\nu}\log\left(\frac{d\mu}{d\nu}\right)d\nu,$$

which proves that Definitions 6.7 and 6.8 are equivalent. Unfortunately, the identity above only holds if $d\mu/d\nu$ is bounded and bounded away from 0, otherwise the function f obtained through (6.10) is no longer bounded. Still, given a distribution μ, and f given by (6.10), we can define

$$f_\varepsilon = \mathbf{1}_{\{\varepsilon \le f \le \varepsilon^{-1}\}} f,$$

and prove both 6.7 and 6.8 by taking the limit $\varepsilon \to 0$, and replacing f by f_ε, which is bounded.

6.4.2 Large deviations and relative entropy

We now take a look at the proof of Cramér's theorem, in which for variables X with distribution ν, we defined in (6.7) the tilted measure

$$\nu_{x^\star}(dx) := \exp\left(t^\star x - \Lambda_{X_1}(t^\star)\right)\nu(dx) = \frac{e^{t^\star x}}{\mathbb{E}_\nu(e^{t^\star x})}\nu(dx),$$

a straightforward computation yields

$$H(\nu_{x^\star} \mid \nu) = \int \frac{e^{t^\star x}}{\mathbb{E}_\nu(e^{t^\star x})}(t^\star x - \Lambda_{X_1}(t^\star))d\nu(x) = t^\star x^\star - \Lambda_{X_1}(t^\star) = \Lambda^\star_{X_1}(x^\star).$$

In other words, the large deviations functional, evaluated at $x^\star$ associated with an i.i.d. sequence of random variables is the relative entropy between the $x^\star$-tilted measure $\nu_{x^\star}$ and the initial measure ν.

Consider now to illustrate the case of i.i.d. $Ber(p)$ variables (see Exercise 24), so that $\nu(X_1 = 1) = 1 - \nu(X_1 = 0) = p$. In the case of a discrete variable (e.g. on $\mathbb{N}$), the relative entropy between two distributions p_k and q_k is given by

$$H(q \mid p) = \sum_{k\in\mathbb{N}} q_k \log\frac{q_k}{p_k}.$$

Here, $\nu(\{1\}) = p_1 = p$, $\nu(\{0\}) = p_0 = 1 - p$, and furthermore

$$\nu_{x^\star}(\{1\}) = \frac{e^{t^\star}}{pe^{t^\star} + (1-p)} = x^\star, \qquad \nu_{x^\star}(\{0\}) = \frac{1-p}{pe^{t^\star} + (1-p)} = 1 - x^\star,$$

since (see Exercise 24, question 2)) $t^\star = \log\left((1-p)x^\star/(1-x^\star)p\right)$. In other words, unsurprisingly, the tilted measure (i.e. a $Ber(p) \sim \nu_p$ tilted to have mean $x^\star$) is simply $\nu_{x^\star}$. This is obvious for Bernoulli random variables, since the only random variable absolutely continuous w.r.t. ν_p, with mean x is ν_x. The large deviations functional is then immediately identified as

$$\Lambda^\star(x) = H(\nu_x \mid \nu_p) = x\log\frac{x}{p} + (1-p)\log\frac{1-x}{1-p}.$$

The big upside is that if one already knows what the tilted measure should look like, the large deviations functional is then immediately given by Definition 6.7.