ISFA, Université Lyon-I
Master 2

Processus stochastiques appliqués l'assurance et la finance
Stéphane Loisel

Review on conditional expectation, compound distributions in non-life insurance, discrete-time martingales, stopping-times, short introduction to continuous-time martingales, discrete-time Markov chains with countable state space.

Table des matières

1 Definition of Conditional Expectation

1.1 General definition

Recall the definition of conditional probability associated with Bayes' Rule

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

For a discrete random variable X we have

$$\mathbb{P}(A) = \sum_x \mathbb{P}(A, X = x) = \sum_x \mathbb{P}(A|X = x)\mathbb{P}(X = x)$$

and the resulting formula for conditional expectation

$$\begin{aligned} \mathbb{E}(Y|X = x) &= \int_\Omega Y(\omega)\mathbb{P}(dw|X = x) \\ &= \frac{\int_{X=x} Y(\omega)\mathbb{P}(dw)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{E}(Y\mathbf{1}_{(X=x)})}{\mathbb{P}(X = x)} \end{aligned}$$

We would like to extend this to handle more general situations where densities don't exist or we want to condition on very "complicated" sets.

Definition 1 *Given a random variable Y with $\mathbb{E}|Y| < \infty$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and some sub-σ-field $\mathcal{G} \subset \mathcal{A}$ we will define the* **conditional expectation** *as the almost surely unique random variable $\mathbb{E}(Y|\mathcal{G})$ which satisfies the following two conditions*

1. *$\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable*
2. *$\mathbb{E}(YZ) = \mathbb{E}(\mathbb{E}(Y|\mathcal{G})Z)$ for all Z which are bounded and $\mathcal{G}$-measurable*

Remark : one could replace 2. in the previous definition with :

$$\forall G \in \mathcal{G}, \quad \mathbb{E}(Y\mathbf{1}_G) = \mathbb{E}(\mathbb{E}(Y|\mathcal{G})\mathbf{1}_G).$$

Proof of existence and unicity

- **Existence** Using linearity, we need only consider $X \geq 0$. Define a measure Q on $\mathcal{F}$ by $Q(A) = \mathbb{E}[X\mathbf{1}_A]$ for $A \in \mathcal{F}$. This is trivially absolutely continuous with respect to $P_{|\mathcal{F}}$, the restriction of P to F. Let $\mathbb{E}[X|\mathcal{F}]$ be the Radon-Nikodym derivative of Q with respect to $P_{|\mathcal{F}}$. The Radon-Nikodym derivative is $\mathcal{F}$-measurable by construction and so provides the desired random variable.
- **Unicity :** If Y_1, Y_2 are two $\mathcal{F}$-measurable random variables with $\mathbb{E}[Y_1\mathbf{1}_A] = \mathbb{E}[Y_2\mathbf{1}_A]$ for all $A \in \mathcal{F}$, then $Y_1 = Y_2$, a.s., or conditional expectation is unique up to a.s. equivalence.

For $\mathcal{G} = \sigma(X)$ when X is a discrete variable, the space Ω is simply partitioned into disjoint sets $\Omega = \sqcup G_n$. Our definition for the discrete case gives

$$\begin{aligned}\mathbb{E}(Y|\sigma(X)) &= \mathbb{E}(Y|X) \\ &= \sum_n \frac{\mathbb{E}(Y\mathbf{1}_{X=x_n})}{\mathbb{P}(X=x_n)}\mathbf{1}_{X=x_n} \\ &= \sum_n \frac{\mathbb{E}(Y\mathbf{1}_{G_n})}{\mathbb{P}(G_n)}\mathbf{1}_{G_n}\end{aligned}$$

which is clearly $\mathcal{G}$-measurable. In general for $\mathcal{G} = \sigma(X)$:

Definition 2 Conditional expectation of Y given X

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ and X another random variable defined on $(\Omega, \mathcal{A}, P)$. Define then $E(Y \mid X)$ the conditional expectation of Y given X as $E(Y \mid \sigma(X))$.

Proposition 3 *Let $(\Omega, \mathcal{A})$ be a measurable space,*

$$Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$$

and X another real-valued random variable defined on $(\Omega, \mathcal{A}, P)$. As $X = f(Y)$, where f is measurable, real-valued function if and only if $\sigma(X) \subset \sigma(Y)$, we get that $E(Y \mid X)$ is a measurable function of X.

Proposition 4 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, and X and Y two independent random variables such that Y is P-integrable. Then $E(Y \mid X) = E(Y)$, P-almost surely.*

Do not mix this notion with the following :

1.2 Couples of random variables with p.d.f.

Proposition 5 *Let (X,Y) be a couple of real-valued random variables with p.d.f. $f_{X,Y}(x,y)$ w.r.t. the Lebesgue measure on $\mathbb{R}^2$. Denote the respective marginal p.d.f. of X and Y as $f_X(x)$ and $f_Y(y)$. Consider $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$. Then almost surely*

$$\forall C \in \mathcal{B}, P(X \in C \mid Y = y) = \int_C f_{X|Y}(x \mid y)dx.$$

If besides X is P-integrable, then

$$E(X \mid Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x \mid y)dx.$$

If $g : \mathbb{R}^2 \to \mathbb{R}$ is a measurable function such that $g(X,Y)$ is integrable, then

$$E(g(X,Y) \mid Y = y) = \int_{\mathbb{R}} g(x,y) f_{X|Y}(x \mid y)dx.$$

Remarks : As soon as $f_Y(y) > 0$, this defines the distribution of X given that $Y = y$, described by p.d.f $f_{X|Y}(x \mid y)$, which is nonnegative and of integral 1.
If X and Y are independent, $f_{X|Y} = f_X$ and $f_{Y|X} = f_Y$. To make the link with $\mathbb{E}[X|Y]$ would require to introduce the concept of regular conditional distribution.
Equation (5) may be useful to compute the mathematical expectation of $g(X,Y)$ as

$$E(g(X,Y)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x,y) f_{X|Y}(x \mid y)dx \right) f_Y(y)dy.$$

2 Properties of Conditional Expectation

2.1 Conditional expectation

$\mathbb{E}(\cdot|\mathcal{G})$ may be seen as an operator on random variables that transforms $\mathcal{A}$-measurable variables into $\mathcal{G}$-measurable ones.

Let us recall the basic properties of conditional expectation :

1. $\mathbb{E}(\cdot|\mathcal{G})$ is positive :
$$Y \geq 0 \to \mathbb{E}(Y|\mathcal{G}) \geq 0)$$

2. $\mathbb{E}(\cdot|\mathcal{G})$ is linear :
$$\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$$

3. $\mathbb{E}(\cdot|\mathcal{G})$ is a projection :
$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})$$

4. More generally, the "tower property". If $\mathcal{H} \subset \mathcal{G}$ then
$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H}) \mid \mathcal{G})$$
Proof : The right equality holds because $\mathbb{E}[X|\mathcal{H}]$ is $\mathcal{H}$- measurable, hence $\mathcal{G}$-measurable. To show the left equality, let $A \in \mathcal{H}$. Then since A is also in $\mathcal{G}$,
$$\mathbb{E}[\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{H}]\mathbf{1}_A].$$
Since both sides are $\mathcal{H}$- measurable, the equality follows.
5. $\mathbb{E}(\cdot|\mathcal{G})$ commutes with multiplication by $\mathcal{G}$-measurable variables :
$$\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})Y \text{ for } \mathbb{E}|XY| < \infty \text{ and } Y\mathcal{G}\text{measurable}$$
Proof : If $A \in \mathcal{G}$, then for any $B \in \mathcal{G}$,
$$\mathbb{E}[\mathbf{1}_A\mathbb{E}[X|\mathcal{G}]\mathbf{1}_B] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_{A\cap B}] = \mathbb{E}[X\mathbf{1}_{A\cap B}] = \mathbb{E}[(\mathbf{1}_AX)\mathbf{1}_B].$$
Since $\mathbf{1}_A\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, this shows that the required equality holds when $Y = \mathbf{1}_A$ and $A \in \mathcal{G}$. Using linearity and taking limits shows that the equality holds whenever Y is $\mathcal{G}$-measurable and X and XY are integrable.
6. $\mathbb{E}(\cdot|\mathcal{G})$ respects monotone convergence :
$$0 \leq X_n \uparrow X \implies \mathbb{E}(X_n|\mathcal{G}) \uparrow \mathbb{E}(X|\mathcal{G})$$

7. If ϕ is convex (in particular if $\phi(x) = x^2$) and $\mathbb{E}|\phi(X)| < \infty$ then a conditional form of Jensen's inequality holds :

$$\phi(\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(\phi(X)|\mathcal{G})$$

8. $\mathbb{E}(\cdot|\mathcal{G})$ is a continuous contraction of Ł^p for $p \geq 1$:

$$\|\mathbb{E}(X|\mathcal{G})\|_p \leq \|X\|_p$$

and

$$X_n \xrightarrow{Ł^2} X \text{ implies } \mathbb{E}(X_n|\mathcal{G}) \xrightarrow{Ł^2} \mathbb{E}(X|\mathcal{G})$$

9. Repeated Conditioning. For $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots$, $\mathcal{G}_\infty = \sigma(\cup\mathcal{G}_i)$, and $X \in Ł^p$ with $p \geq 1$ then

$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{a.s.} \mathbb{E}(X|\mathcal{G}_\infty)$$
$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{Ł^p} \mathbb{E}(X|\mathcal{G}_\infty)$$

10. Best approximation property :
Suppose that the random variable X is square-integrable, but not measurable with respect to $\mathcal{G}$. That is, the information in $\mathcal{G}$ does not completely determine the values of X. The conditional expectation, $Y = E[X \mid \mathcal{G}]$, has the property that it is the best approximation to X among functions measurable with respect to Y, in the least squares sense. That is, if $\tilde{Y}$ is $\mathcal{G}$-measurable, then

$$\mathbb{E}\left[(\tilde{Y} - X)^2\right] \geq \mathbb{E}\left[(Y - X)^2\right] \ .$$

It thus realizes the orthogonal projection of X onto a convex closed subset of a Hilbert space. This predicts the variance decomposition theorem that we shall see in a further section.

2.2 Conditional variance

Definition 6 *Let X be a square-integrable, real-valued random variable defined on a probability space $(\Omega, \mathcal{A}, P)$, and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$.*

Define the **conditional variance of** X **given** $\mathcal{F}$ *(denoted by* $Var(X \mid \mathcal{F})$*) as the random variable* $E((X - E(X \mid \mathcal{F}))^2 \mid \mathcal{F})$*.*
Define also the conditional variance of X *given a real-valued random variable* Y *defined on* $(\Omega, \mathcal{A}, P)$ *(denoted by* $Var(X \mid Y)$*) as the random variable* $E((X - E(X \mid Y))^2 \mid Y)$*.*

Proposition 7 $Var(X \mid \mathcal{F})$ *and* $Var(X \mid Y)$ *are well- defined, almost surely nonnegative and finite.*

$$Var(X \mid \mathcal{F}) = E(X^2 \mid \mathcal{F}) - E(X \mid \mathcal{F})^2,$$

and

$$Var(X \mid Y) = E(X^2 \mid Y) - E(X \mid Y)^2.$$

Proposition 8 Variance decomposition formula
Let (X, Y) *be a couple of random variables defined on a probability space* $(\Omega, \mathcal{A}, P)$*, such that* X *is square-integrable. Then*

$$Var(X) = E(Var(X \mid Y)) + Var(E(X \mid Y)).$$

This may be very useful in non-life insurance to find the variance of a compound distribution.

Proof :
- $\mathrm{Var}(X \mid Y) = E(X^2 \mid Y) - (E(X \mid Y))^2$.
- $E[\mathrm{Var}(X \mid Y)] = E[E(X^2 \mid Y)] - E[(E(X \mid Y))^2]$.
- $E[E(X^2 \mid Y)] = E[X^2]$.
- $E[\mathrm{Var}(X \mid Y)] = E[X^2] - E[(E(X \mid Y))^2]$.
- $\mathrm{Var}(E(X \mid Y)) = E[(E(X \mid Y))^2] - (E[E(X \mid Y)])^2$.
- $E[E(X \mid Y)] = E[X]$.
- Hence $\mathrm{Var}(E(X \mid Y)) = E[(E(X \mid Y))^2] - (E[X])^2$.

2.3 Compound distributions

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and
- $(X_n)_{n \in \mathbb{N}}$ a sequence of i.i.d., nonnegative random variables defined on $(\Omega, \mathcal{A}, P)$. X_n represents the severity of the n^{th} claim in the collective risk model.

– N an random variable defined on $(\Omega, \mathcal{A}, P)$ and taking values in $\mathbb{N}$, independent from the X_n. It represents the number of claims.
– Let $S_N = X_1 + ... + X_N$ represent the aggregate claim amount.

In many models we may know the mean and variance of N and X_1. How can we then get the mean and variance of S_N ? Simply by conditioning on the number of claims, and using conditional expectation and variance given N.

Proposition 9 *First,*

$$\mathbb{E}S_N = \mathbb{E}N.\mathbb{E}X_1$$

Moreover, thanks to the variance decomposition theorem, we may decompose $Var(S_N)$ into two parts : the first one represents the part due to variability in claim amounts ; the second one represents the part due to variability in the number of claims :

$$Var(S_N) = \mathbb{E}N.Var(X_1) + (\mathbb{E}X_1)^2.Var(N)$$

3 Discrete-time martingales

3.1 Definition and basic properties

Definition 10 Filtration

Let $(\Omega, \mathcal{A}, P)$ be a probability space. A **filtration** *is an increasing sequence $(\mathcal{F}_n)$, $n \in \mathbb{N}$ of sub-σ-algebras of $\mathcal{A}$. Denote the σ-algebra generated by all the $\mathcal{F}_n$, $n \in \mathbb{N}$ as $\mathcal{F}_\infty$.*

Definition 11 Adapted and predictable process

Let $(\mathcal{F}_n)$, $n \in \mathbb{N}$ be a filtration of a probability space $(\Omega, \mathcal{A}, P)$. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to be an **adapted process** *(to filtration $\mathcal{F}_n$) if $\forall n \in \mathbb{N}$, X_n is $\mathcal{F}_n$-measurable.*
If moreover $\forall n \in \mathbb{N}$, X_n is $\mathcal{F}_{n-1}$-measurable, X_n is said to be a **predictable process**.

Definition 12 martingale, (sub, super) martingale *Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $(\mathcal{F}_n)$, $n \in \mathbb{N}$ and let $(X_n)_{n\in\mathbb{N}}$ be an adapted process. If $\forall n \in \mathbb{N}$, X_n is P-integrable and if $E(X_{n+1} \mid \mathcal{F}_n) = \left(\begin{smallmatrix}\geq \\ \leq\end{smallmatrix}\right) X_n$, X_n is said to be a (sub, super) martingale w.r.t. $(\mathcal{F}_n)$.*

You may also say that $(X_n, \mathcal{F}_n)$ is a (sub, super) martingale.

Examples :

- Let X_n be a sequence of independent, integrable, centered random variables defined on a probability space $(\Omega, \mathcal{A}, P)$, and define the sequence of partial sums S_n as : $\forall n \in \mathbb{N}$, $S_n = \sum_{k\leqslant n} X_k$ and filtration $\mathcal{F}_n$, $n \in \mathbb{N}$ by $\forall n \in \mathbb{N}$, $\mathcal{F}_n = \sigma(X_k, k \leqslant n)$. S_n is a martingale (w.r.t. the filtration we have just defined).
- The special case where $X_0 = 0$ and $P(X_n = 1) = p = 1 - P(X_n = -1)$ is called a random walk on the integers. It is a (sub, super)martingale when $p = (\geq, \leq)\frac{1}{2}$.
- If the X_n are independent, centered, identically distributed, square-integrable random variables with variance σ^2, then $M_n = (S_n)^2 - n\sigma^2$ is also a martingale w.r.t. $(\mathcal{F}_n)$.
- Let X be an integrable random variable defined on a probability space $(\Omega, \mathcal{A}, P)$ with a filtration $(\mathcal{F}_n)$, $n \in \mathbb{N}$. Then $(X_n)_{n\in\mathbb{N}}$ defined as $X_n = E(X \mid \mathcal{F}_n)$ is a martingale (w.r.t. $(\mathcal{F}_n)$).

- Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $(\mathcal{F}_n)$, let X_n be a $(\mathcal{F}_n)$-martingale and let g be a convex function such that $\forall n \in \mathbb{N}$, $E \mid g(X_n) \mid < \infty$. Then $(g(X_n))_{n\in\mathbb{N}}$ is a sub-martingale.
  Think about $g(x) = e^x$ or $g(x) = x^p$ with $p \geq 1$ as important examples.
- Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $(\mathcal{F}_n)$, let X_n be a $(\mathcal{F}_n)$-martingale, and let V_n be $(\mathcal{F}_n)$-predictable and bounded. Then $H_0 = 0$ and

$$\forall n > 0, \quad H_n = \sum_{k=1}^{n} V_k(X_k - X_{k-1})$$

  defines another martingale, which may represent the sum of the gains up to time n if you have invested an amount V_k between times $k-1$ and k given the information you had at time $k-1$. It is also the discrete-time version of the stochastic integral of (V_k) w.r.t. the martingale X_k.

Theorem 13 Doob's decomposition theorem
Any $(\mathcal{F}_n)$-adapted process (X_n) in L^1 may be decomposed uniquely up to almost equality for all n into :

$$X_n = X_0 + M_n + A_n$$

where (M_n) is a martingale w.r.t. $(\mathcal{F}_n)$, and (A_n) is a predictable process w.r.t. $(\mathcal{F}_n)$, with $M_0 = A_0 = 0$.
Moreover (X_n) is a sub-martingale if and only if (A_n) is nonnegative and nondecreasing,
and if (X_n) is a nonnegative sub-martingale which is bounded in L^1, then A_n converges a.s. and in L^1 to an a.s. finite random variable A_∞.

Proof : Exercise.

Theorem 14 Doob maximal inequality
Let X_n be a nonnegative $(\mathcal{F}_n)$-sub-martingale. Then for $a > 0$,

$$P(\max_{1\leq k\leq n} X_k \geq a) \leq \frac{\mathbb{E}X_n}{a}.$$

Proof : Exercise.

Theorem 15 Doob L^p inequality
Let $(M_n)_{n\geq 0}$ be a $(\mathcal{F}_n)$-martingale such that all $X_n \in L^p$. Recall that $||x||_p = (\mathbb{E}X^p)^{1/p}$. Then

$$|| \max_{1\leq k\leq n} X_k||_p \geq \frac{p}{p-1}||X_n||_p.$$

The proof of this inequality is based on Holder's inequality. This equality will be useful to prove the L^p-convergence theorem 28.

3.2 Stopping Times

Stopping times correspond to a very simple, intuitive idea : it is a random time at which one can decide to stop playing without anticipating what will happen later. Assume you play a game at the casino : you can choose to stop after a deterministic number of games(say for example 100), just after you have won a certain amount of money (continue otherwise), or just after you have won a certain amount of money or lost everything. However you cannot choose to stop after the last game before the second time you lose. (You would need to know about the future to stop at that time.) Stopping times intervene in Doob's optional sampling theorem, and in general are useful to determine and even define the optimal strategy to adopt in various situations. (see for example gambler's ruin problem, or American option pricing.)

Definition 16 stopping time *Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $(\mathcal{F}_n)$, $n \in \mathbb{N}$. A* **stopping time** *is a random variable $T : \Omega \to \mathbb{N} \cup \{+\infty\}$ such that $P(T \in \mathbb{N}) = 1$ and $\forall\ n < \infty$, $\{T = n\} \in \mathcal{F}_n$.*

Remark : One could replace $T = n$ with $T \leq n$ in the previous definition.
Warning : In some books the author may not require $P(\tau < \infty) = 1$.
Properties :
- Constant integer-valued random variables are stopping times.
- If σ and τ are stopping times, so are $\inf(\tau, \sigma)$, $\sup(\tau, \sigma)$ and $\tau + \sigma$.
- If τ is a stopping time, so is τ^2.
- Let M_n be a $\mathcal{F}_n$-martingale with $M_0 = 0$ and $\alpha > 0$. Then $\forall N \in \mathbb{N}$, $\tau_N = N \wedge \inf\{n \in N, \quad |M_n| \geq \alpha\}$ is a stopping time , and more generally the first entrance time into Borelian sets.
- $\tau = \lim_{N \to \infty} \tau_N$ is a stopping time too.
- If τ is a stopping time, in general $\tau - 1$ is not a stopping time.
- in general the exit time of a Borelian set is not a stopping time (such as $\tau'_N = N \wedge \sup\{n \in N, \quad |M_n| \leq \alpha\}$).

Definition 17 *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let τ be a stopping time w.r.t. $(\mathcal{F}_n)_{n \in \mathbb{N}}$. Define the σ-fields of the events preceding τ as*

$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty, \quad \forall n \in \mathbb{N}, \quad A \cap \{\tau = n\} \in \mathcal{F}_n\}.$$

By construction, τ is $\mathcal{F}_\tau$-measurable.

Definition 18 *On the subset $\{\tau < +\infty\}$ one can define the random variable X_τ by $X_\tau(\omega) = \psi o \phi(w)$, where $\phi(\omega) = (T(\omega), \omega)$ and $\psi(n, \omega) = X_n(\omega)$. We complete the definition by $X_\tau = \beta$ on $\{\tau = +\infty\}$, where β is a constant.*

Proposition 19 *X_τ is thus $\mathcal{F}_\tau$-measurable, and moreover $X_\tau = X_n$ on every $\{\tau = n\}$.*

Exercise 20 *Show that :*
- *If τ and σ are two stopping times, the event $\{\tau \leq \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$*
- *$\tau \leq \sigma$ almost surely implies $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$.*
- *$\mathcal{F}_{\tau \wedge \sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.*

Note that by induction, one can show that $\forall n \in \mathbb{N}, \mathbb{E}M_n = \mathbb{E}M_0$ when (M_n) is a martingale. The following theorem asserts in particular that this remains true if one replaces n with any bounded stopping time.

3.3 Optional stopping theorem

Theorem 21 Doob's optional stopping theorem *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let τ and σ be two stopping times and (M_n) a martingale w.r.t. $(\mathcal{F}_n)_{n \in \mathbb{N}}$ such that :*

1. *$\mathbb{E}(\mid M_\sigma \mid) < +\infty$,*
2. *$\mathbb{E}(\mid M_\tau \mid) < +\infty$, and*
3. *$\liminf_{n \to \infty} \int_{\tau \geq n} \mid M_n \mid dP \quad < \infty$.*

Then

$$\mathbb{E}(M_\tau \mathbf{1}_{\tau \geq \sigma} \mid \mathcal{F}_\sigma) = (\geq) M_\sigma \mathbf{1}_{\tau \geq \sigma}.$$

If $\exists N \in \mathbb{N}$, such that $P(\tau \leq N) = 1$, then 2. and 3. are verified. If it is not the case, it may be helpful to apply the theorem for stopping times $\tau \wedge n$ (which are finite stopping times) and to try to use monotone or dominated convergence theorems to show the desired property for τ.
Exercise : Use the previous theorem to compute $\mathbb{E}(M_\tau)$ and $\mathbb{E}(M_\tau \mathbf{1}_{\tau > 0})$.

Theorem 22 (Wald)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d., integrable random variables defined on a probability space $(\Omega, \mathcal{A}, P)$, and let S_n be the partial sums $X_1 + \cdots + X_n$ and N be a stopping time w.r.t. the natural filtration of the X_i, with finite mean. Then

$$\mathbb{E}S_N = (\mathbb{E}N).\mathbb{E}X_1.$$

Proof :
Thanks to linearity of mathematical expectation, we can restrict ourselves to nonnegative X_i. Then

$$\begin{aligned}
\mathbb{E}S_N &= \sum_{n\geq 1} \mathbb{E}\left[\mathbf{1}_{\{N=n\}} \sum_{i=1}^{n} X_i\right] \\
&= \sum_{i\geq 1} \mathbb{E}\left[X_i \sum_{n\geq i} \mathbf{1}_{\{N=n\}}\right] \\
&= \sum_{i\geq 1} \mathbb{E}\left[X_i \mathbf{1}_{N\geq i}\right] \\
&= \sum_{i\geq 1} P(N \geq i)\mathbb{E}X_i
\end{aligned}$$

because $\{N \geq i\} = \overline{\{N < i\}} \in \sigma(X_1, \ldots, X_{i-1})$, which implies that each $\{N \geq i\}$ is independent from $\sigma(X_i)$. The last term is exactly $(\mathbb{E}X_1)(\mathbb{E}N)$. $\square$
Application : Typing monkey. Suppose that you have a monkey sitting in front of a keyboard with only 3 letter keys : O,T and L. Assume that the monkey exactly types one letter at random every second, and consider the first time in seconds N_1 when the word LOTO appears, and N_2 the first time when TOTO appears. For example, if the monkey types : OLTOTO-TOTLOTO, $N_1 = 13$ and $N_2 = 6$. Compute $\mathbb{E}N_1$ and $\mathbb{E}N_2$ with martingales. Interpret your result.

3.4 Quadratic variation of an adapted process

Definition 23 *Let (X_n) be a square-integrable martingale w.r.t. a filtration $(\mathcal{F}_n)$, (assume that all $X_n \in L^2$). The quadratic variation of X_n is the (adap-*

ted, nondecreasing) process $< X >_n$ *defined by* $< X >_0 = 0$ *and for* $n > 0$,

$$< X >_n = \sum_{i=1}^{n} (X_i - X_{i-1})^2$$

It is the unique process such that $< X >_0 = 0$ *and* $X_n^2 - < X >_n$ *is a martingale.*

Note the link with Doob decomposition theorem 13. This notion and its generalization in continuous time are very important for stochastic integration.

3.5 Martingale convergence

Definition 24 *The number of up-crossings of an interval* $[a, b]$ *is the number of times a process crosses from below* a *to above* b *: define*
$S_1 = \min\{k, \quad X_k \leq a\}$,
and $T_1 = \min\{k > S_1, \quad X_k \geq b\}$.
Define then by induction on i *:*

$$S_{i+1} = \min\{k > T_i, \quad X_k \leq a\},$$

$$T_{i+1} = \min\{k > S_{i+1}, \quad X_k \geq b\}.$$

The number of up-crossings U_n *before time* n *then corresponds to*

$$U_n = \max\{j, \quad T_j \leq n\}.$$

Lemma 25 (Up-crossing lemma)
If X_k *is a sub-martingale,*

$$\mathbb{E}U_n \leq \frac{1}{b-a}\mathbb{E}[(X_n - a)_+].$$

Proof : Exercise.

Theorem 26 *If* X_n *is a sub-martingale such that* $\sup_n \mathbb{E}[(X_n)_+] < \infty$, *then* X_n *converges a.s. as* $n \to +\infty$.

Proof : Let $U(a,b) = \lim_{n\to\infty} U_n$. For each $(a,b) \in \mathbb{Q}^2$, by monotone convergence and by Doob up-crossing lemma 25,

$$\mathbb{E}U(a,b) < +\infty.$$

Hence in particular $U(a,b) < +\infty$ a.s. Taking the union over all $(a,b) \in \mathbb{Q}^2$, we see that a.s. the sequence $X_n(\omega)$ cannot have $\limsup X_n > \liminf X_n$. (Otherwise take two rationals between them, the number of up-crossings between them would be infinite)
Therefore X_n converges a.s. in $\bar{\mathbb{R}}$. Let us show now that the limit cannot be infinite. Since X_n is a sub-martingale, $\mathbb{E}X_n \geq \mathbb{E}X_0$, and thus

$$\mathbb{E}|X_n| = \mathbb{E}[(X_n)_+] + \mathbb{E}[(X_n)_-] = 2\mathbb{E}[(X_n)_+] - \mathbb{E}X_n \leq 2\mathbb{E}[(X_n)_+] - \mathbb{E}X_0 < +\infty.$$

By Fatous lemma,

$$\mathbb{E}\left(\lim_{n\to+\infty} |X_n|\right) \leq \sup_{n\to+\infty} \mathbb{E}|X_n| < +\infty,$$

i.e. X_n converges a.s. to a finite limit. □

Corollary 27 *If X_n is a positive super-martingale or a martingale bounded above or below, X_n converges a.s.*

Proof : Immediate from theorem 26. □

Convergence in L^p is very different for $p = 1$ and for $p > 1$.

Theorem 28 (L^p convergence, $p > 1$)
If X_n is a martingale with $\sup_n \mathbb{E}|X_n|^p < \infty$ for some $p > 1$, then the convergence is in L^p as well as a.s. The result also holds for sub-martingales.

Proof : The proof is immediate from Doobs L^p-inequality (theorem 15) and Lebesgue's dominated convergence theorem. □
Convergence in L^1 requires a condition of uniform integrability. Recall the definition of uniform integrability :

Definition 29 (Uniform integrability)
A collection of integrable random variables $(X_i)_{i\in I}$ is uniformly integrable if

$$\sup_{i\in I} \int_{\{|X_i|>M\}} |X_i| d\mathbb{P} \to 0$$

as $M \to +\infty$.
X_n is a uniformly integrable martingale if the collection of random variables X_n, $n \in \mathbb{N}$ is uniformly integrable.

Recall also the standard integration theory result :

Lemma 30 *If X_n is a uniformly integrable sequence of random variables, which converges a.s., then it also converges in L^1.*

Theorem 31 *If X_n is a uniformly integrable martingale, then the convergence is in L^1.*

Proof : Thanks to theorem 26, X_n converges a.s. Because it is uniformly integrable, it also converges in L^1 thanks to lemma 30. $\square$

Theorem 32 *If X_n is a martingale such that $X_n \to X_\infty$ in L^p,* **for some** *$p \geq 1$, then*

$$\forall n \in \mathbb{N}, \quad X_n = \mathbb{E}[X_\infty | \mathcal{F}_n].$$

Proof : For all $j < n$, we have $X_j = \mathbb{E}[X_n | \mathcal{F}_j]$ by martingale definition and because $(\mathcal{F}_n)$ is a filtration.
For $A \in \mathcal{F}_j$,

$$\mathbb{E}[X_j \mathbf{1}_A] = \mathbb{E}[X_n \mathbf{1}_A] \to \mathbb{E}[X_\infty \mathbf{1}_A]$$

by the L^1-convergence of X_n to X_∞. Since this is true for all $A \in \mathcal{F}_j$,

$$X_j = \mathbb{E}[X_\infty | \mathcal{F}_j]. \quad \square$$

4 Continuous-time martingales

Definition 33 *A filtration in the continuous case is an increasing sequence $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ of sub-σ-fields of $\mathcal{A}$.*

Definition 34 *A continuous-time filtration is said to be* **right-continuous** *if*

$$\forall t > 0, \quad \mathcal{F}_t = \bigcap_{s > t} \mathcal{F}_s.$$

We will only deal with right-continuous, complete (i.e. $P(N) = 0 \Rightarrow N \in \mathcal{F}_0$) filtrations here, and with processes with right-continuous trajectories with left-hand limits (càdlàg) :

Definition 35 $(X_t)_{t\in\mathbb{R}^+}$ *is said to have right-continuous trajectories with right-hand limits (càdlàg) if almost-surely in* ω*,*

$$t \to X_t(\omega)$$

is right-continuous with right-hand limits (càdlàg).

We will only deal with this kind of continuous-time processes, because this enables to keep defining their properties only from their values at rational times, and thus for example $M_t = \sup_{0\leq s\leq t} X_s$ remains a random variable. You can rewrite $M_t = \min_n \left(\sup_{s\in[0,t+\frac{1}{n}]\cap\mathbb{Q}}\{X_s\}\right)$. Thanks to Kolmogorov extension theorem we can construct such processes. The fact that M_t is a random variable will be important for example in finance : think of X_t as the price of an asset at time t, and M_t as the highest record of the price between times 0 and t. To price barrier options (financial contracts which give you the right, but not the obligation to buy an asset at a fixed price in a certain amount of time, on condition that the price of the asset never exceeds a certain amount), you will need to compute $P(S_t \leq x, \quad M_t \leq y)$. If the option is a European option, you can only use your option at expiry date. If it is an American option, you may use it anytime before the expiry date. The time at which you choose to use your option will be determined thanks to stopping time theory. Most theorems true for discrete-time remain valid for right-continuous processes in continuous time.
Do not forget that in this section, all the processes that we consider are assumed to be càdlàg, and the filtrations are assumed to be complete and right-continuous.

Definition 36 stopping time *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space with a filtration* $(\mathcal{F}_t)$*,* $t \in \mathbb{R}^+$*. A* **stopping time** *is a random variable* $T : \Omega \to \mathbb{R}^+ \cup \{+\infty\}$ *such that* $P(T \in \mathbb{R}^+) = 1$ *and* $\forall\ t < \infty$*,* $\{T \leq t\} \in \mathcal{F}_t$*.*

Remark : One could replace $T \leq t$ with $T < t$ in the previous definition.
Warning : In some books the author may not require $P(\tau < \infty) = 1$.
Properties :

- Constant nonnegative random variables are stopping times.
- If σ and τ are stopping times, so are $\inf(\tau,\sigma)$, $\sup(\tau,\sigma)$ and $\tau+\sigma$.
- Let M_t be a $\mathcal{F}_t$-martingale with $M_0 = 0$ and $\alpha > 0$. Then $\forall N \in \mathbb{N}$, $\tau_N = N \wedge \inf\{t \in \mathbb{R}^+, \quad |M_t| \geq \alpha\}$ is a stopping time , and more generally the first entrance time into Borelian sets.

– $\tau = \lim_{N\to\infty} \tau_N$ is a stopping time too.

Definition 37 *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let τ be a stopping time w.r.t. $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$. Define the σ-fields of the events preceding τ as*

$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty, \quad \forall t \in \mathbb{R}^+, \quad A \cap \{\tau \leq t\} \in \mathcal{F}_t\}.$$

By construction, τ is $\mathcal{F}_\tau$-measurable.

Definition 38 *On the subset $\{\tau < +\infty\}$ one can define the random variable X_τ by $X_\tau(\omega) = \psi o\phi(\omega)$, where $\phi(\omega) = (T(\omega), \omega)$ and $\psi(t, \omega) = X_t(\omega)$. We complete the definition by $X_\tau = \beta$ on $\{\tau = +\infty\}$, where β is a constant.*

Proposition 39 *X_τ is thus $\mathcal{F}_\tau$-measurable, and moreover $X_\tau = X_t$ on every $\{\tau = t\}$.*

Proposition 40 *As in discrete time :*
- *If τ and σ are two stopping times, the event $\{\tau \leq \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$*
- *$\tau \leq \sigma$ almost surely implies $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$.*
- *$\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.*

Definition 41 continuous-time martingale, (sub, super) martingale *Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $(\mathcal{F}_t)$, $t \in \mathbb{R}^+$ and let $(X_t)_{t\in\mathbb{R}^+}$ be an adapted process. If $\forall t \in \mathbb{R}^+$, X_t is P-integrable and if $E(X_t \mid \mathcal{F}_s) = \begin{pmatrix}\geq\\ \leq\end{pmatrix} X_s$ for $0 < s \leq t$, X_t is said to be a (sub, super) martingale w.r.t. $(\mathcal{F}_t)$.*

Theorem 42 Doob's optional stopping theorem (continuous-time version) *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let τ and σ be two stopping times and (M_t) a martingale w.r.t. $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ such that :*
1. *$\mathbb{E}(\mid M_\sigma \mid) < +\infty$,*
2. *$\mathbb{E}(\mid M_\tau \mid) < +\infty$, and*
3. *$\lim \inf\limits_{t\to\infty} \int_{\tau\geq t} \mid M_t \mid dP \quad < \infty$.*

Then

$$\mathbb{E}(M_\tau \mathbf{1}_{\tau\geq\sigma} \mid \mathcal{F}_\sigma) = (\geq) M_\sigma \mathbf{1}_{\tau\geq\sigma}.$$

We still have existence and uniqueness of the quadratic variation of a continuous-time martingale, which is defined as the unique process $< X >_t$ such that $< X >_0 = 0$ and $X_t^2 - < X >_t$ is a martingale. The quadratic variation of X_t is a nondecreasing, adapted process.

5 Markov chains

5.1 General discrete-time Markov chains

Suppose S is a set with some topological structure that we will use as our state space. Think of S as being $\mathbb{R}^d$ or the positive integers, for example. A sequence of random variables $(X_0, X_1, \dots)$ from Ω to S is a Markov chain if

$$\mathbb{P}(X_{n+1} \in A | X_0, \dots, X_n) = \mathbb{P}(X_{n+1} \in A | X_n) \tag{1}$$

for all n and all measurable sets A. The definition of Markov chain has this information : to predict the probability that X_{n+1} is in any set, we only need to know where we currently are ; how we got there gives no new additional intuition. In other words the future depends from the past only through the present. The distribution of X_{n+1} given X_0, ..., X_n is just the same as the distribution of X_{n+1} given X_n.

Although there is quite a theory developed for Markov chains with arbitrary state spaces, we will confine our attention to the case where either S is finite, in which case we will usually suppose $S = 1, 2, \dots, n$, or countable and discrete, in which case we will usually suppose S is the set of positive integers. We are going to further restrict our attention to Markov chains where

$$\mathbb{P}(X_{n+1} \in A | X_n = x) = \mathbb{P}(X_1 \in A | X_0 = x), \tag{2}$$

that is, where the probabilities do not depend on n. Such Markov chains are said to have **stationary transition probabilities**. In life-insurance, one uses non-stationary Markov chains to model life expectation, because the probability to die next year increases with the age (at least after a certain age). Here we focus on the properties of stationary Markov chains. If the chain is not stationary, then very often (but not in life insurance), the only way of studying its behaviour is to use simulation.

5.2 Stationary, discrete-time Markov chains with countable state space

Define the initial distribution of a Markov chain with stationary transition probabilities by $\mu(i) = \mathbb{P}(X_0 = i)$. Define the transition probabilities by $p(i, j) = \mathbb{P}(X_{n+1} = j | X_n = i)$. Since the transition probabilities are stationary, $p(i, j)$ does not depend on n. In this case we can use the definition of conditional probability. If $\mathbb{P}(X_n = i) = 0$ for all n, that means we never visit i and we could drop the point i from the state space.

Proposition 43 *Let X be a Markov chain with initial distribution μ and transition probabilities $p(i,j)$. then*

$$\mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) = \mu(i_0)p(i_0,i_1)\dots p(i_{n-1},i_n) \tag{3}$$

Proof : We use induction on n. It is clearly true for $n = 0$ by the definition of $\mu(i)$. Suppose it holds for n ; we need to show it holds for $n+1$. For simplicity, we will do the case $n = 2$. Then

$$\begin{aligned}
& \mathbb{P}(X_3 = i_3, X_2 = i_2, X_1 = i_1, X_0 = i_0) \\
= \; & \mathbb{E}[\mathbb{P}(X_3 = i_3 | X_0 = i_0, X_1 = i_1, X_2 = i_2)\mathbf{1}_{X_2=i_2, X_1=i_1, X_0=i_0}] \\
= \; & \mathbb{E}[\mathbb{P}(X_3 = i_3 | X_2 = i_2)\mathbf{1}_{X_2=i_2, X_1=i_1, X_0=i_0}] \\
= \; & p(i_2,i_3)\mathbb{P}(X_2 = i_2, X_1 = i_1, X_0 = i_0).
\end{aligned}$$

Now by the induction hypothesis,

$$\mathbb{P}(X_2 = i_2, X_1 = i_1, X_0 = i_0) = p(i_1,i_2)p(i_0,i_1)\mu(i_0).$$

Substituting establishes the claim for $n = 3$. □
The above proposition says that the law of the Markov chain is determined by the $\mu(i)$ and $p(i,j)$. The formula (3) also gives a prescription for constructing a Markov chain given the $\mu(i)$ and $p(i,j)$, thanks to Kolmogorov extension theorem. This is an opportunity for us to state this theorem which is the justification of everything a lot of us just do without even thinking of the problem. It fixes the good framework for the study of discrete-time stochastic processes and basically enables us to consider only their finite-dimensional marginal distributions to characterize them. The two following theorems and their proofs may be omitted at first reading.

Theorem 44 Kolmogorov extension theorem
For all $n \geq 1$, let ψ_n be the projection from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ and π_n be the projection from $\mathbb{R}^{\mathbb{N}}$ to $\mathbb{R}^n$ respectively defined by : for all $x_1, x_2, \dots,$

$$\psi_n(x_1, \dots, x_{n+1}) = (x_1, \dots, x_n)$$

and

$$\pi_n(x_1, \dots, x_n, \dots) = (x_1, \dots, x_n).$$

Let $(\mathbb{P}_n)_{n \geq 1}$ be a sequence of probability measures such that

– *For all* $n \geq 1$, $\mathbb{P}_n$ *is a probability measure on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$,
– *and the image of* $\mathbb{P}_{n+1}$ *by* ψ_n *is* $\mathbb{P}_n$ *:*

$$\forall B \in \mathcal{B}(\mathbb{R}^n), \quad \mathbb{P}_{n+1}(\psi_n^{-1}(B)) = \mathbb{P}_n(B).$$

Then there exists a unique probability measure $\mathbb{P}$ *on* $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$*, such that for all* $n \geq 1$*, the image of* $\mathbb{P}$ *by* π_n *is* $\mathbb{P}_n$ *:*

$$\forall B \in \mathcal{B}(\mathbb{R}^n), \quad \mathbb{P}(\pi_n^{-1}(B)) = \mathbb{P}_n(B).$$

Proposition 45 *Suppose* $\mu(i)$ *is a sequence of nonnegative numbers such that* $\sum_i \mu(i) = 1$ *and for each* i *the sequence* $p(i,j)$ *is nonnegative and sums to* 1*. Then there exists a Markov chain with* $\mu(i)$ *as its initial distribution and* $p(i,j)$ *as the transition probabilities.*

Proof : Define $\Omega = \mathcal{S}^\infty$. Let F be the σ-field generated by the collection of sets $\{(i_0, i_1, \ldots, i_n) : n > 0, i_j \in \mathcal{S}\}$. An element ω of Ω is a sequence $(i_0, i_1, \ldots)$. Define $X_j(\omega) = i_j$ if $\omega = (i_0, i_1, \ldots)$. Define $\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n)$ by (3). Using the Kolmogorov extension theorem, one can show that $\mathbb{P}$ can be extended to a probability on Ω.

The above framework is rather abstract, but it is clear that under $\mathbb{P}$ the sequence X_n has initial distribution $\mu(i)$; what we need to show is that X_n is a Markov chain and that

$$\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \ldots X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) = p(i_n, i_n{+}1). \tag{4}$$

By the definition of conditional probability, the left hand side of (4) is

$$\begin{aligned} \mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n) &= \frac{\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n, \ldots, X_0 = i_0)}{\mathbb{P}(X_n = i_n, \ldots, X_0 = i_0)} \\ &= \frac{\mu(i_0) \ldots p(i_{n-1}, i_n) p(i_n, i_n + 1)}{\mu(i_0) \ldots p(i_{n-1}, i_n)} \\ &= p(i_n, i_{n+1}) \end{aligned} \tag{5}$$

as desired. To complete the proof we need to show

$$\frac{\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n)}{\mathbb{P}(X_n = i_n)} = p(i_n, i_{n+1}),$$

or

$$\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n) = p(i_n, i_{n+1}) \mathbb{P}(X_n = i_n). \tag{6}$$

Now

$$\begin{aligned}\mathbb{P}(X_n = i_n) &= \sum_{i_0,\dots,i_{n-1}} \mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)\\ &= \sum_{i_0,\dots,i_{n-1}} \mu(i_0)\dots p(i_{n-1}, i_n)\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n) &= \sum_{i_0,\dots,i_{n-1}} \mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)\\ &= p(i_n, i_{n+1}) \sum_{i_0,\dots,i_{n-1}} \mu(i_0)\dots p(i_{n-1}, i_n). \qquad (7)\end{aligned}$$

Equation (6) now follows. □

Note in this construction that the X_n sequence is fixed and does not depend on μ or p. Let $p(i, j)$ be fixed. The probability we constructed above is often denoted $\mathbb{P}^\mu$. If μ is point mass at a point i or x, it is denoted $\mathbb{P}^i$ or $\mathbb{P}^x$. So we have one probability space, one sequence X_n, but a whole family of probabilities $\mathbb{P}_\mu$. Later on we will see that this framework allows one to express the Markov property and strong Markov property in a convenient way. As part of the preparation for doing this, we define the shift operators $\theta_k : \Omega \to \Omega$ by

$$\theta_k(i_0, i_1, \dots) = (i_k, i_{k+1}, \dots).$$

Then $X_j \circ \theta_k = X_{j+k}$. To see this, if $\omega = (i_0, i_1, \dots)$, then

$$X_j \circ \theta_k(\omega) = X_j(i_k, i_{k+1}, \dots) = i_{j+k} = X_{j+k}(\omega).$$

Examples :

1. *Random walks on the integers* that we already introduced in martingale theory (let Y_i be an i.i.d. sequence of r.v.s, with $p = \mathbb{P}(Y_i = 1)$ and $1 - p = \mathbb{P}(Y_i = -1)$. Let $X_n = X_0 + \sum_{i=1}^n Y_i$). X_n is a martingale if $p = 0.5$, a sub-martingale if $p > 0.5$ and a super-martingale if $p < 0.5$. The X_n may also be viewed as a Markov chain with $p(i, i+1) = p$, and $p(i, i-1) = 1 - p$, and $p(i, j) = 0$ if $|j - i| \neq 1$. More general random walks on the integers also fit into this framework. To check that the random walk is Markov,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n)\\ =\ &\mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n | X_0 = i_0, \dots, X_n = i_n)\\ =\ &\mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n), \qquad (8)\end{aligned}$$

using the independence, while

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1}|X_n = i_n) &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n|X_n = i_n) \\ &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n).\end{aligned}$$

2. *Gambler's ruin* :
This may be viewed as a random walk on the integers with only $a+b+1$ states : $\{0, 1, \ldots, a+b\}$starting from a and stopped when the gambler gets ruined or has wealth $a+b$. (0 and $a+b$ will be called absorbing states because we define in this case $p(0,0) = p(a+b, a+b) = 1$).

3. *Random walks on graphs* : Suppose we have n points, and from each point there is some probability of going to another point. For example, suppose there are 5 points and we have $p(1,2) = \frac{1}{2}, p(1,3) = \frac{1}{2}, p(2,1) = \frac{1}{4}, p(2,3) = \frac{1}{2}, p(2,5) = \frac{1}{4}, p(3,1) = \frac{1}{4}, p(3,2) = \frac{1}{8}, p(3,3) = \frac{1}{2}, p(4,1) = 1, p(5,1) = \frac{1}{2}, p(5,5) = \frac{1}{2}$. The $p(i,j)$ are often arranged into a matrix :

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

Note the rows must sum to 1 since

$$\sum_{j=1}^{5} p(i,j) = \sum_{j=1}^{5} \mathbb{P}(X_1 = j|X_0 = i) = \mathbb{P}(X_1 \in \mathcal{S}|X_0 = i) = 1$$

.

4. *Very basic disablement and death model :*
We may use a 3-state or a 4-state, discrete-time Markov chain to represent the conditions of an insured individual. Of course these are simplified models with stationary transition probabilities, but in life insurance you will tackle the general case. Here we only have 3 states :
– State 0 : Alive and active
– State 1 : Alive but disabled
– State 2 Dead

Transitions may occur at the end of each year and the one-year transition probabilities are given by the matrix

$$P = \begin{pmatrix} 0.85 & 0.1 & 0.05 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{pmatrix}$$

To price insurance contracts, even if actuaries do not use in general stationary models, they would be interested in answering the following questions :

(a) What is the probability that an active and active person dies in exactly 2 years ?

(b) in less than 4 years ?

(c) Which state(s) is/are recurrent, transient, positive/null recurrent ?

(d) How many classes of irreducibility are there in this model ?

(e) If an individual is currently alive and active, what is the probability that he will ever become disabled ?

(f) If an individual is currently disabled, what is the probability that he will ever return to active life ?

(g) For an active person what is the remaining life expectation ?

(h) For a disabled person what is the remaining life expectation ?

(i) What is the expected length of a period of disablement ?

(j) What is the expected number of disablement for someone who is active today ?

(k) What is the expected total time of disablement until death for an active person today ?

The formalism we are introducing will enable us to answer these questions quite easily.

5. *Renewal processes* : Let Y_i be i.i.d. with $\mathbb{P}(Y_i = k) = a_k$ and the a_k are nonnegative and sum to 1. Let $T_0 = i_0$ and $T_n = T_0 + \sum_{i=1}^{n} Y_i$. We think of the Y_i as the time between the $(i-1)^{th}$ and the i^{th} accident in the collective risk model in non-life insurance and T_n the time when the n^{th} accident happens. Let $X_n = \min\{m-n : T_i = m \text{for some } i\}$. So X_n is the amount of time after time n until the next accident occurs.

If $X_n = j$ and $j > 0$, then $T_i = n + j$ for some i but T_i does not equal $n, n+1, \ldots, n+j-1$ for any i. So $T_i = (n+1)+(j-1)$ for some i and T_i does not equal $(n+1), (n+1)+1, \ldots, (n+1)+(j-2)$ for any i. Therefore $X_{n+1} = j-1$. So $p(i, i-1) = 1$ if $i \geq 1$. If $X_n = 0$, then an accident occurred at time n and X_{n+1} is 0 if another accident occurred immediately and $j-1$ if the next accident only occurs after a time j. The probability of this is a_j . So $p(0, j) = a_{j+1}$. All the other $p(i, j)$s are 0.

6. *Branching processes* : Consider k particles. At the next time interval, some of them die, and some of them split into several particles. The probability that a given particle will split into j particles is given by a_j , $j = 0, 1, \ldots,$ where the a_j are nonnegative and sum to 1. The behavior of each particle is independent of the behavior of all the other particles. If X_n is the number of particles at time n, then X_n is a Markov chain. Let Y_i be i.i.d. random variables with $\mathbb{P}(Y_i = j) = a_j$. The $p(i, j)$ for X_n are somewhat complicated, and can be defined by $p(i, j) = \mathbb{P}(\sum_{m=1}^{i} Y_m = j)$.

7. *Queues* We will discuss briefly the $M/G/1$ queue. The M refers to the fact that the customers arrive according to a Poisson process. So the probability that the number of customers arriving in a time interval of length t is k is given by $e^{-\lambda t}(\lambda t)^k/k!$ The G refers to the fact that the length of time it takes to serve a customer is given by a distribution that is not necessarily exponential. The 1 refers to the fact that there is 1 server. Suppose the length of time to serve one customer has distribution function F with density f. The probability that k customers arrive during the time it takes to serve one customer is

$$a_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} f(t) dt.$$

Let the Y_i be i.i.d. with $\mathbb{P}(Y_i = k-1) = a_k$. So Y_i is the number of customers arriving during the time it takes to serve one customer. Let $X_{n+1} = (X_n + Y_{n+1})^+$ be the number of customers waiting. Then X_n is a Markov chain with $p(0,0) = a_0 + a_1$ and $p(i, j-1+k) = a_k$ if $j \geq 1, k > 1$. Queuing theory is very much linked to ruin theory, and we are going to study the Poisson process and the compound Poisson process during the third week.

8. *Ehrenfest urns* : Suppose we have two urns with a total of r balls, k in one and $r-k$ in the other. Pick one of the r balls at random and move it to the other urn. Let X_n be the number of balls in the first urn. X_n is a Markov chain with $p(k, k+1) = (r-k)/r$, $p(k, k-1) = k/r$, and $p(i,j) = 0$ otherwise. One model for this is to consider two containers of air with a thin tube connecting them. Suppose a few molecules of a foreign substance are introduced. Then the number of molecules in the first container is like an Ehrenfest urn. We shall see that all states in this model are recurrent, so infinitely often all the molecules of the foreign substance will be in the first urn. Yet there is a tendency towards equilibrium, so on average there will be about the same number of molecules in each container for all large times.
9. *Birth and death processes* : Suppose there are i particles, and the probability of a birth is a_i, the probability of a death is b_i, where $a_i, b_i \geq 0, ai + bi \leq 1$. Setting X_n equal to the number of particles, then X_n is a Markov chain with $p(i, i+1) = a_i, p(i, i-1) = b_i$, and $p(i,i) = 1 - a_i - b_i$.

5.3 Markov properties

Let us look at

$$\mathbb{P}^x(X_{n+1} = j | X_n = i) = \mathbb{P}^x(X_1 = j | X_0 = i).$$

Let $g(y) = P^y(X_1 = j)$. We have

$$\mathbb{P}^x(X_1 = j, X_0 = k) = \begin{cases} \mathbb{P}^k(X_1 = j) & \text{if} \quad x = k, \\ 0 & \text{if} \quad x \neq k, \end{cases}$$

while

$$\begin{aligned} \mathbb{E}^x[g(X_0)\mathbf{1}_{X_0=k}] &= \mathbb{E}^x[g(k)\mathbf{1}_{X_0=k}] = \mathbb{P}^k(X_1 = j)\mathbb{P}^x(X_0 = k) \\ &= \begin{cases} \mathbb{P}^k(X_1 = j) & \text{if} \quad x = k \\ 0 & \text{if} \quad x \neq k \end{cases} \end{aligned}$$

It follows that $\mathbb{P}^x(X_1 = j | X_0) = g(X_0)$, and so

$$\mathbb{P}^x(X_1 = j | X_0 = i) = \mathbb{P}^i(X_1 = j).$$

The following are equivalent ways of writing this :

$$\begin{aligned}
\mathbb{E}^x[\mathbf{1}_{\{j\}}(X_{n+1})|X_n = i] &= \mathbb{E}^i[\mathbf{1}_{\{j\}}(X_1)];\\
\mathbb{E}^x[\mathbf{1}_{\{j\}}(X_1)\circ\theta_n|X_n = i] &= \mathbb{E}^i[\mathbf{1}_{\{}j\}(X_1)];\\
\mathbb{E}^x[Y\circ\theta_n|X_n = i] &= \mathbb{E}^i[Y], \quad Y = \mathbf{1}_{\{j\}}(X_1);\\
\mathbb{E}^x[Y\circ\theta_n|X_n] &= \mathbb{E}^{X_n}[Y];\\
\mathbb{E}^x[Y\circ\theta_n|F_n] &= \mathbb{E}^{X_n}[Y].
\end{aligned}$$

The last one is a complicated way of writing the first. We generalize this.

Theorem 46 Markov property
If Y is bounded and measurable, then

$$\mathbb{E}^x[Y\circ\theta_n|F_n] = \mathbb{E}^{X_n}[Y], \quad a.s. \tag{9}$$

for each n and x.

Proof : If we can prove this for $Y = f_1(X_1)\dots f_m(X_m)$, then we will have it for all Y by linearity and taking limits. We use induction on m. The case $m = 1$ is just the string of equivalences above. Suppose the result holds for m and we want to show it holds for $m+1$. We have

$$\begin{aligned}
&\mathbb{E}^x[f_1(X_{n+1})\dots f_{m+1}(X_{n+m+1})|F_n]\\
=\ &\mathbb{E}^x[\mathbb{E}^x[f_{m+1}(X_{n+m+1})|F_{n+m}]f_1(X_{n+1})\dots f_m(X_{n+m})|F_n]\\
=\ &\mathbb{E}^x[\mathbb{E}^{X_{n+m}}[f_{m+1}(X_1)]f_1(X_{n+1})\dots f_m(X_{n+m})|F_n]\\
=\ &\mathbb{E}^x[f_1(X_{n+1})\dots f_{m-1}(X_{n+m-1})h(X_{n+m})|F_n].
\end{aligned}$$

Here we used the result for $m = 1$ and we defined $h(y) = f_{n+m}(y)g(y)$, where $g(y) = \mathbb{E}^y[f_{m+1}(X_1)]$. Using the induction hypothesis, this is equal to

$$\begin{aligned}
\mathbb{E}^{X_n}[f_1(X_1)\dots f_{m-1}(X_{m-1})h(X_m)] &= \mathbb{E}^{X_n}[f_1(X_1)\dots f_m(X_m)\mathbb{E}^{X_m}f_{m+1}(X_1)]\\
&= \mathbb{E}^{X_n}[f_1(X_1)\dots f_m(X_m)\mathbb{E}[f_{m+1}(X_{m+1})|F_m]]\\
&= \mathbb{E}^{X_n}[f_1(X_1)\dots f_{m+1}(X_{m+1})], \qquad (10)
\end{aligned}$$

which is what we needed. □

Define $\theta_N(\omega) = (\theta_{N(\omega)})(\omega)$. The strong Markov property is the same as the Markov property, but where the fixed time n is replaced by a stopping time N.

Theorem 47 *If Y is bounded and measurable and N is a finite stopping time, then*

$$\mathbb{E}^x[Y \circ \theta_N | F_N] = \mathbb{E}^{X_N}[Y]. \tag{11}$$

Proof : We will show

$$\mathbb{P}^x(X_{N+1} = j | F_N) = \mathbb{P}^{X_N}(X_1 = j).$$

Once we have this, we can proceed as in the proof of the Theorem 46 to obtain our result. To show the above equality, we need to show that if $B \in \mathcal{F}_N$, then

$$\mathbb{P}^x(X_{N+1} = j, B) = \mathbb{E}^x[\mathbb{P}^{X_N}(X_1 = j)\mathbf{1}_B].$$

Recall that since $B \in \mathcal{F}_N$, then $B \cap (N = k) \in \mathcal{F}_k$. We have

$$\begin{aligned} \mathbb{P}^x(X_{N+1} = j, B, N = k) &= \mathbb{P}^x(X_{k+1} = j, B, N = k) \\ &= \mathbb{E}^x[\mathbb{P}^x(X_{k+1} = j | \mathcal{F}_k)\mathbf{1}_{B \cap (N=k)}] \\ &= \mathbb{E}^x[\mathbb{P}^{X_k}(X_1 = j)\mathbf{1}_{B \cap (N=k)}] \\ &= \mathbb{E}^x[\mathbb{P}^{X_N}(X_1 = j)\mathbf{1}_{B \cap (N=k)}]. \end{aligned}$$

Now sum over k ; since N is finite, we obtain our desired result.□

Another way of expressing the Markov property is through the

Proposition 48 Chapman-Kolmogorov equations
Let $p^n(i,j) = \mathbb{P}(X_n = j | X_0 = i)$. For all i, j, m, n we have

$$p^{n+m}(i,j) = \sum_{k \in \mathcal{S}} p^n(i,k) p^m(k,j).$$

Proof : We write

$$\begin{aligned} \mathbb{P}(X_{n+m} = j, X_0 = i) &= \sum_k \mathbb{P}(X_{n+m} = j, X_n = k, X_0 = i) \\ &= \sum_k \mathbb{P}(X_{n+m} = j | X_n = k, X_0 = i)\mathbb{P}(X_n = k | X_0 = i) P(X_0 = i) \\ &= \sum_k \mathbb{P}(X_{n+m} = j | X_n = k) p^n(i,k) \mathbb{P}(X_0 = i) \\ &= \sum_k p^m(k,j) p^n(i,k) \mathbb{P}(X_0 = i). \end{aligned}$$

If we divide both sides by $P(X_0 = i)$, we have our result. $\Box$

Note the resemblance to matrix multiplication. It is clear if P is the matrix made up of the $p(i,j)$, then P^n will be the matrix whose (i,j) entry is $p^n(i,j)$.

5.4 Recurrence and transience

Let
$$T_y = \min\{i > 0 : X_i = y\}.$$
This is the first time that X_i hits the point y. Even if $X_0 = y$ we would have $T_y > 0$. We let T_y^k be the k-th time that the Markov chain hits y and we set
$$r(x,y) = P^x(T_y < \infty),$$
the probability starting at x that the Markov chain ever hits y. If $r(x,y) > 0$, y is said to be **accessible** from x. Two states are said **communicating states** if they are accessible from each other (i.e. $\exists n \geq 1, p^n(x,y) > 0$ and $\exists m \geq 1, p^m(y,x) > 0$).

Proposition 49
$$\mathbb{P}^x(T_y^k < \infty) = r(x,y)r(y,y)^{k-1}.$$

Proof : The case $k = 1$ is just the definition, so suppose $k > 1$. Using the strong Markov property,
$$\begin{aligned}
\mathbb{P}^x(T_y < \infty) &= \mathbb{P}^x(T_y \circ \theta_{T_y^{k-1}} < \infty, T_y^{k-1} < \infty) \\
&= \mathbb{E}^x[\mathbb{P}^x(T_y \circ \theta_{T_y^{k-1}} < \infty | \mathcal{F}_{T_y^{k-1}})\mathbf{1}_{T_y^{k-1}} < \infty] \\
&= \mathbb{E}^x[\mathbb{P}^{X(T_y^{k-1})}(T_y < \infty)\mathbf{1}_{T_y^{k-1}}] \\
&= \mathbb{E}^x[\mathbb{P}^y(T_y < \infty)\mathbf{1}_{T_y^{k-1}} < \infty] \\
&= r(y,y)\mathbb{P}^x(T_y^{k-1} < \infty).
\end{aligned}$$
We used here the fact that at time T_y^{k-1} the Markov chain must be at the point y. Repeating this argument $k-2$ times yields the result.$\Box$

Definition 50 *We say that y is* **recurrent** *if $r(y,y) = 1$; otherwise we say y is* **transient**.

Proposition 51 *Let*

$$N(y) = \sum_{n=1}^{\infty} \mathbf{1}_{(X_n=y)}.$$

y is recurrent if and only if $\mathbb{E}^y N(y) = +\infty$.

Proof : Note

$$\mathbb{E}^y N(y) = \sum_{k=1}^{\infty} \mathbb{P}^y(N(y) \geq k) = \sum_{k=1}^{\infty} \mathbb{P}^y(T_y^k < \infty) = \sum_{k=1}^{\infty} r(y,y)^k.$$

We used the fact that $N(y)$ is the number of visits to y and the number of visits being larger than k is the same as the time of the k^{th} visit being finite. Since $r(y,y) \leq 1$, the left hand side will be finite if and only if $r(y,y) < 1$. □

Observe that

$$\mathbb{E}^y N(y) = \sum_n \mathbb{P}^y(X_n = y) = \sum_n p^n(y,y).$$

If we consider simple symmetric random walk on the integers, then $p^n(0,0)$ is 0 if n is odd and equal to $C_n^{n/2} 2^{-n}$ if n is even. This is because in order to be at 0 after n steps, the walk must have had $n/2$ positive steps and $n/2$ negative steps ; the probability of this is given by the binomial distribution. Using Stirlings approximation, we see that $p^n(0,0) \sim c/\sqrt{n}$ for n even, which diverges, and so simple random walk in one dimension is recurrent. A projection on the main bisectrices of the plane shows that simple symmetric random walk is also recurrent in 2 dimensions. However it is transient in 3 or more dimensions. Conclusion (humoristic of course) : If you consider random walk as the walk of someone who is completely drunk, it means you'd better be drunk driving a car than driving a plane.

Proposition 52 *If x is recurrent and $r(x,y) > 0$, then y is recurrent and $r(y,x) = 1$.*

Proof : First we show $r(y,x) = 1$. Suppose not. Since $r(x,y) > 0$, there is a smallest n and $y_1, \ldots, y_{n-1}$ such that $p(x,y_1)p(y_1,y_2)\ldots p(y_{n-1},y) > 0$. Since this is the smallest n, none of the y_i can equal x. Then

$$\mathbb{P}^x(T_x = \infty) \geq p(x,y_1)\ldots p(y_{n-1},y)(1 - r(y,x)) > 0,$$

a contradiction to x being recurrent.
Next we show that y is recurrent. Since $r(y,x) > 0$, there exists L such that $p^L(y,x) > 0$. Then

$$p^{L+n+K}(y,y) \geq p^L(y,x)p^n(x,x)p^K(x,y).$$

Summing over n,

$$\sum_n p^{L+n+K}(y,y) \geq p^L(y,x)p^K(x,y)\sum_n p^n(x,x) = \infty$$

.□

Definition 53 *We say a subset C of $\mathcal{S}$ is* **closed** *if $x \in C$ and $r(x,y) > 0$ implies $y \in C$. A subset D is* **irreducible** *if $x, y \in D$ implies $r(x,y) > 0$.*

Proposition 54 *Let C be finite and closed. Then C contains a recurrent state.*

Remark : From the preceding proposition, if C is irreducible, then all states will be recurrent.
Proof : If not, for all y we have $r(y,y) < 1$ and

$$\mathbb{E}^x N(y) = \sum_{k=1}^{\infty} r(x,y)r(y,y)^{k-1} = \frac{r(x,y)}{1-r(y,y)} < \infty.$$

Since C is finite, then $\sum_y \mathbb{E}^x N(y) < \infty$. But that is a contradiction since

$$\sum_y \mathbb{E}^x N(y) = \sum_y \sum_n p^n(x,y) = \sum_n \sum_y p^n(x,y) = \sum_n \mathbb{P}^x(X_n \in C) = \sum_n 1 = \infty.$$

□

Theorem 55 *Let $R = \{x : r(x,x) = 1\}$, the set of recurrent states. Then $R = \bigcup_{i=1}^{\infty} R_i$, where each R_i is closed and irreducible.*

Proof : Say $x \sim y$ if $r(x, y) > 0$. Since every state is recurrent, $x \sim x$ and if $x \sim y$, then $y \sim x$. If $x \sim y$ and $y \sim z$, then $p^n(x, y) > 0$ and $p^m(y, z) > 0$ for some n and m. Then $p^{n+m}(x, z) > 0$ or $x \sim z$. $\square$
Remark : Therefore we have an equivalence relation and we let the R_i be the equivalence classes.

Recurrence and transience in our examples : Looking at our examples, it is easy to see that in the Ehrenfest urn model all states are recurrent. For the branching process model, suppose $p(x, 0) > 0$ for all x. Then 0 is recurrent and all the other states are transient. In the renewal chain, there are two cases. If $\{k : a_k > 0\}$ is unbounded, all states are recurrent. If $K = \max\{k : a_k > 0\}$, then $\{0, 1, \ldots, K-1\}$ are recurrent states and the rest are transient. For the queueing model, let $\mu = \sum k a_k$, the expected number of people arriving during one customers service time. We may view this as a branching process by letting all the customers arriving during one persons service time be considered the progeny of that customer. It turns out that if $\mu \leq 1$, 0 is recurrent and all other states are also. If $\mu > 1$ all states are transient.

5.5 How to compute the expected number of visits to a state and $r(x, y)$

If the chain is finite and irreducible, then all states are recurrent and you will probably have to use stationary distribution and proposition 60.
To compute the (total) mean time spent in a transient state, or $r(x, y)$ when x and y are transient, you can use the potential matrix :

Define the matrix (eventually infinite, and with values in $\bar{\mathbb{N}}$) :

$$R(x, y) = \mathbb{E}^x(N(y))$$

R is called the potential matrix of the chain. One can show quite easily that

$$R = \sum_{n \geq 0} P^n.$$

For transient states only , take P_T sub-matrix of P (only lines and columns for transient states). Then you can easily compute it as :

$$R_T = (I - P_T)^{-1}$$

You can then deduce the $\mathbb{E}^x(N(y)) = R_T(x,y)$ and $r(x,y)$ thanks to :

$$P^x(T_y < \infty) = r(x,y) = \frac{R_T(x,y)}{R_T(y,y)} = \frac{\mathbb{E}^x(N(y))}{\mathbb{E}^y(N(y))}$$

5.6 Stationary measures

Definition 56 *A probability measure μ is a* **stationary distribution** *if*

$$\sum_x \mu(x)p(x,y) = \mu(y). \tag{12}$$

Important remark : In matrix notation this is $\mu P = \mu$, or μ is the left eigenvector corresponding to the eigenvalue 1. In the case of a stationary distribution, $\mathbb{P}^\mu(X_1 = y) = \mu(y)$, which implies that $X_1, X_2, \ldots$ all have the same distribution. We can use equation (12) when μ is a measure rather than a probability, in which case it is called a stationary measure.

In our examples : If we have a random walk on the integers, $\mu(x) = 1$ for all x serves as a stationary measure. In the case of an asymmetric random walk : $p(i, i+1) = p, p(i, i-1) = q = 1-p$ and $p \neq q$, setting $\mu(x) = (p/q)^x$ also works. In the Ehrenfest urn model, $\mu(x) = 2^{-r}C_x^r$ works. One way to see this is that μ is the distribution one gets if one flips r coins and puts a coin in the first urn when the coin is heads. A transition corresponds to picking a coin at random and turning it over.

Proposition 57 *Let x be recurrent and let $T = T_x$. Set*

$$\mu(y) = \mathbb{E}^x \sum_{n=0}^{T-1} \mathbf{1}_{(X_n = y)}.$$

Then μ is a stationary measure.

Proof : The idea of the proof is that $\mu(y)$ is the expected number of visits to y by the sequence $X_0, \ldots, X_{T-1}$ while μP is the expected number of visits to y by $X_1, \ldots, X_T$. These should be the same because $X_T = X_0 = x$.

We next turn to uniqueness of the stationary distribution.

Proposition 58 *If the Markov chain is irreducible and all states are recurrent, then the stationary measure is unique up to a constant multiple.*

Proposition 59 *If a stationary distribution exists, then $\mu(y) > 0$ implies y is recurrent.*

Proposition 60 *Recall that T_x is the first time to hit x. If the Markov chain is irreducible and has stationary distribution μ, then*

$$\mu(x) = \frac{1}{E^x T_x}.$$

Proof : $\mu(x) > 0$ for some x. If $y \in S$, then $r(x,y) > 0$ and so $p^n(x,y) > 0$ for some n. Hence $\mu(y) = \sum_x \mu(x) p^n(x,y) > 0$. Hence by Proposition 59, all states are recurrent. By the uniqueness of the stationary distribution, $\mu(x)$ is a constant multiple of

$$\mu_x(y) = \sum_{n=0}^{\infty} \mathbb{P}^x(X_n = y, T_x > n).$$

Note

$$\begin{aligned}
\sum_y \mu_x(y) &= \sum_y \sum_{n=0}^{\infty} \mathbb{P}^x(X_n = y, T_x > n) \\
&= \sum_n \sum_y \mathbb{P}^x(X_n = y, T_x > n) \\
&= \sum_n \mathbb{P}^x(T_x > n) = \mathbb{E}^x T_x.
\end{aligned}$$

Therefore, since

$$\sum_y \mu(y) = 1,$$

we have

$$\mu(x) = \frac{\mu_x(x)}{\sum_y \mu_x(y)} = \frac{1}{\mathbb{E}^x T_x}. \quad \square$$

Definition 61 *We make the following distinction for recurrent states. If $\mathbb{E}^x T_x < +\infty$, then x is said to be* **positive recurrent**. *If x is recurrent but $\mathbb{E}^x T_x = +\infty$, x is* **null recurrent**.

Proposition 62 *Suppose a chain is irreducible.*

1. *If there exists a positive recurrent state, then there is a stationary distribution.*
2. *If there is a stationary distribution, all states are recurrent.*
3. *If there exists a transient state, all states are transient.*
4. *If there exists a null recurrent state, all states are null recurrent.*

Proof : To show 1., if x is positive recurrent, then there exists a stationary measure with $\mu(x) = 1$. Then

$$\tilde{\mu}(y) = \frac{\mu(y)}{\mathbb{E}^x T_x}$$

will be a stationary distribution.
For 2., suppose $\mu(x) > 0$ for some x. We showed this implies $\mu(y) > 0$ for all y. Then $0 < \mu(y) = 1/\mathbb{E}^y T_y$, which implies $\mathbb{E}^y T_y < \infty$.
We showed that if x is recurrent and $r(x, y) > 0$, then y is recurrent. So 3. follows.
Suppose there exists a null recurrent state. If there exists a positive recurrent or transient state as well, then by 1. and 2. or by 3. all states are positive recurrent or transient, which is a contradiction, and 4. follows. □

5.7 Convergence of Markov chains

Our goal is to show that under certain conditions $p^n(x, y) \to \pi(y)$, where $\pi(y)$ is the stationary distribution. (In the null recurrent case $p^n(x, y) \to 0$.) Consider a random walk on the set $\{0, 1\}$, where with probability one on each step the chain moves to the other state. Then $p^n(x, y) = 0$ if $x \neq y$ and n is even. A less trivial case is the simple random walk on the integers. We need to eliminate this periodicity.

Definition 63 *Suppose x is recurrent, let $I_x = \{n \geq 1 : p^n(x, x) > 0\}$, and let d_x be the g.c.d. (greatest common divisor) of I_x. d_x is called the* **period** *of x.*

Proposition 64 *If $r(x, y) > 0$, then $d_y = d_x$.*

Proof : Since x is recurrent, $r(y,x) > 0$. Choose K and L such that $p^K(x,y), p^L(y,x) > 0$.

$$p^{K+L+n}(y,y) \geq p^L(y,x)p^n(x,x)p^K(x,y),$$

so taking $n = 0$, we have $p^{K+L}(y,y) > 0$, or d_y divides $K + L$. So d_y divides n if $p^n(x,x) > 0$, or d_y is a divisor of I_x. Hence d_y divides d_x. By symmetry d_x divides d_y. □

Proposition 65 *If $d_x = 1$, there exists m_0 such that $p^m(x,x) > 0$ whenever $m \geq m_0$.*

Proof : First of all, I_x is closed under addition : if $m, n \in Ix$,

$$p^{m+n}(x,x) \geq p^m(x,x)p^n(x,x) > 0.$$

Secondly, if there exists N such that $N, N+1 \in I_x$, let $m_0 = N^2$. If $m \geq m_0$, then $m - N^2 = kN + r$ for some $r < N$ and

$$m = r + N^2 + kN = r(N+1) + (N - r + k)N \in I_x. \tag{13}$$

Third, pick $n_0 \in I_x$ and $k > 0$ such that $n_0 + k \in I_x$. If $k = 1$, we are done. Since $d_x = 1$, there exists $n_1 \in I_x$ such that k does not divide n_1. We have $n_1 = mk + r$ for some $0 < r < k$. Note $(m+1)(n_0 + k) \in I_x$ and $(m+1)n_0 + n_1 \in I_x$. The difference between these two numbers is $(m+1)k - n_1 = k - r < k$. So now we have two numbers in I_k differing by less than or equal to $k - 1$. Repeating at most k times, we get two numbers in I_x differing by at most 1, and we are done. □

Definition 66 *We write d for d_x. A chain is* **aperiodic** *if $d = 1$. If $d > 1$, we say $x \sim y$ if $p^{kd}(x,y) > 0$ for some $k > 0$ We divide $\mathcal{S}$ into equivalence classes $\mathcal{S}_1, \ldots \mathcal{S}_d$.*

Every d steps the chain started in $\mathcal{S}_i$ is back in $\mathcal{S}_i$. So we look at $p' = p^d$ on $\mathcal{S}_i$.

Theorem 67 *Suppose the chain is irreducible, aperiodic, and has a stationary distribution π. Then $p^n(x,y) \to \pi(y)$ as $n \to \infty$.*

Proof : The idea is to take two copies of the chain with different starting distributions, let them run independently until they couple, i.e., hit each other, and then have them move together. So define

$$q((x_1, y_1), (x_2, y_2)) = \begin{cases} (p(x_1, x_2)p(y_1, y_2) & \text{if} x_1 \neq y_1, \\ p(x_1, x_2) & \text{if} x_1 = y_1, x_2 = y_2, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Let $Z_n = (X_n, Y_n)$ and $T = \min\{i : X_i = Y_i\}$. We have

$$\begin{aligned} \mathbb{P}(X_n = y) &= \mathbb{P}(X_n = y, T \leq n) + \mathbb{P}(X_n = y, T > n) \\ &= \mathbb{P}(Y_n = y, T \leq n) + \mathbb{P}(X_n = y, T > n), \end{aligned}$$

while

$$\mathbb{P}(Y_n = y) = \mathbb{P}(Y_n = y, T \leq n) + \mathbb{P}(Y_n = y, T > n).$$

Subtracting,

$$\begin{aligned} \mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y) &\leq \mathbb{P}(X_n = y, T > n) - \mathbb{P}(Y_n = y, T > n) \\ &\leq \mathbb{P}(X_n = y, T > n) \leq P(T > n). \end{aligned}$$

Using symmetry,

$$|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \leq \mathbb{P}(T > n).$$

Suppose we let Y_0 have distribution π and $X_0 = x$. Then

$$|p^n(x, y) - \pi(y)| \leq \mathbb{P}(T > n).$$

It remains to show $\mathbb{P}(T > n) \to 0$.
To do this, consider another chain $Z'_n = (X_n, Y_n)$, where now we take X_n, Y_n independent. Define

$$r((x_1, y_1), (x_2, y_2)) = p(x_1, x_2)p(y_1, y_2).$$

The chain under the transition probabilities r is irreducible.
To see this, there exist K and L such that $p^K(x_1, x_2) > 0$ and $p^L(y_1, y_2) > 0$. If M is large, $p^{L+M}(x_2, x_2) > 0$ and $p^{K+M}(y_2, y_2) > 0$. So $p^{K+L+M}(x_1, x_2) > 0$ and $p^{K+L+M}(y_1, y_2) > 0$, and hence we have $r^{K+L+M}((x_1, x_2), (y_1, y_2)) > 0$. It is easy to check that $\pi'(a, b) = \pi(a)\pi(b)$ is a stationary distribution for Z'. Hence Z'_n is recurrent, and hence it will hit (x, x), hence the time to hit the diagonal $\{(y, y) : y \in \mathcal{S}\}$ is finite. However the distribution of the time to hit the diagonal is the same as T. $\square$