Probability Theory

Daniele Zago

January 24, 2022

Contents

Lecture 0: Probability review

2021-10-11

References Çinlar (2011, §1-2)
Paolella (2007)

In this section we summarize a (hopefully useful) review of concepts which can serve as a basis for the following lectures.

0.1 Probability spaces

Let E be a set, we want to define some useful quantities to build the notion of a probability space, that is, a space onto which a probability measure can be defined.

Def. (Sigma-algebra)
A non-empty collection $\mathcal{E}$ of subsets of E is called a σ-algebra on E if

a) $E \in \mathcal{E}$

b) (Closure under ^{c}) $A \in \mathcal{E} \implies A^c \in \mathcal{E}$

c) (Closure under $\cap$) $A_1, A_2, \ldots \in \mathcal{E} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{E}$

Remarks

› Every σ-algebra on E includes E and $\emptyset$ at least, indeed $\mathcal{E} = \{\emptyset, \mathcal{E}\}$ is called the *trivial* σ-algebra.

› Conversely, the maximal sigma algebra on E is given by the power set of E denoted by $\mathscr{P}(E)$.

› A countable (or uncountable) intersection of σ-algebras on E is again a σ-algebra on E. Given a collection $\mathcal{C}$ of subsets of E, we define the σ-algebra *generated by* $\mathcal{C}$ as the intersection of all σ-algebras $\mathcal{E}$ on E which contain $\mathcal{C}$,

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{E} \,:\, \mathcal{C} \subseteq \mathcal{E}} \mathcal{E}.$$

› If E is a topological space, then the σ-algebra generated by the collection of all open subsets of E is called the *Borel σ-algebra* and is denoted by $\mathcal{B}(E)$. $B \in \mathcal{B}(E)$ is called a *Borel set*.

› Given two sets E and F with σ-algebras $\mathcal{E}$ and $\mathcal{F}$, we can define the *σ-algebra generated by the rectangles* on $E \times F$ as

$$\mathcal{E} \otimes \mathcal{F} = \sigma\big(\{A \times B : A \subseteq \mathcal{E}, B \subseteq \mathcal{F}\}\big).$$

Moreover, if $\mathcal{E}$ and $\mathcal{F}$ are the Borel σ-algebra on $\mathbb{R}$, we have

$$\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2).$$

With the above definition of a σ-algebra, we can now define the basic type of space onto which a probability measure can be constructed.

Def. (Measurable space)

A ***measurable space*** is a pair $(E, \mathcal{E})$ where E is a set and $\mathcal{E}$ a σ-algebra on E. Elements of $\mathcal{E}$ are accordingly called *measurable sets*.

Let E and F be sets. A *function* $f : E \longrightarrow F$ is a rule that assigns an element $f(x) \in F$ to each $x \in E$. We are interested in a particular class of functions, namely those which are related to the sigma algebra defined on the spaces E and F.

Def. (Measurable function)

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \longrightarrow F$ is said to be ***measurable*** wrt to $\mathcal{E}$ and $\mathcal{F}$ if for every $B \in \mathcal{F}$,

$$f^{-1}(B) \in \mathcal{E}.$$

Prop. 1 (Measurable functions of measurable functions are measurable)

If f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ and g is measurable relative to $\mathcal{F}$ and $\mathcal{G}$, then $g \circ f : E \longrightarrow G$ given by $g \circ f(x) = g\big(f(x)\big)$ is measurable relative to $\mathcal{E}$ and $\mathcal{G}$.

Proof.
For $C \in \mathcal{G}$, we have that $(g \circ f)^{-1}(C) = f^{-1}\big(g^{-1}(C)\big)$. Now, $g^{-1}(C) \in \mathcal{F}$ since g is measurable, and therefore $f^{-1}\big(g^{-1}(C)\big) \in \mathcal{E}$ by the measurability of f.

□

Remark If μ is a measure on $\mathcal{E}$ and $f : E \to F$ is measurable wrt to E and F, then f induces a measure $\widehat{\mu}$ on $\mathcal{F}$ given by

$$\widehat{\mu}(B) = \mu\big(f^{-1}(B)\big), \quad B \in \mathcal{F}.$$

A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a set (set of *outcomes*), $\mathcal{F}$ is a σ-algebra on Ω (set of *events*), and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Mathematically, a probability space is a measure space where the measure has a total mass of one.

The probability measure has the following properties, which are verified for all finite measures:

$$
\begin{aligned}
&(\textit{Norming}) && \mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1, \mathbb{P}(H) = 1 - \mathbb{P}(H^c) \\
&(\textit{Monotonicity}) && H \subset K \implies \mathbb{P}(H) \leq \mathbb{P}(K) \\
&(\textit{Finite additivity}) && H \cap K = \emptyset \implies \mathbb{P}(H \cup K) = \mathbb{P}(H) + \mathbb{P}(K) \\
&(\textit{Countable additivity}) && (H_n)_{n\in\mathbb{N}} \text{ disjoint} \implies \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} H_n\Big) = \sum_{n\in\mathbb{N}} \mathbb{P}(H_n) \\
&(\textit{Sequential continuity}) && H_n \nearrow H \implies \mathbb{P}(H_n) \nearrow \mathbb{P}(H) \\
& && H_n \searrow H \implies \mathbb{P}(H_n) \searrow \mathbb{P}(H) \\
&(\textit{Boole's inequality}) && \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} H_n\Big) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(H_n).
\end{aligned}
$$

0.2 Random variables

Def. (Random variable)
Let $(E, \mathcal{E})$ be a measurable space. A mapping $X : \Omega \longrightarrow E$ is called a ***random variable*** provided that it be measurable relative to $\mathcal{F}$ and $\mathcal{E}$, that is, if for every $A \in \mathcal{E}$,

$$X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}.$$

In general, we say that X is E-valued with the σ-algebra $\mathcal{E}$ that is understood from context.

Def. (Distribution of a random variable)
Let X be a random variable on $(E, \mathcal{E})$, then we define the ***distribution of X*** as the image of μ of $\mathbb{P}$ under X,

$$\mu(A) = \mathbb{P}\big(X^{-1}(A)\big) = \mathbb{P}\big(X \in A\big), \quad A \in \mathcal{E}.$$

Let X be a r.v. in $(E, \mathcal{E})$ and let $(F, \mathcal{F})$ be another measurable space. Let now $f : E \longrightarrow F$ a measurable function relative to $\mathcal{E}$ and $\mathcal{F}$, then the composition $Y = f \circ E$

$$Y(\omega) = f \circ X(\omega) = f\big(X(\omega)\big), \quad \omega \in \Omega$$

is a random variable taking values in $(F, \mathcal{F})$ (Prop 1). If μ is the distribution of X, then the distribution ν of Y is $\nu = \mu \circ f^{-1}$:

$$\nu(B) = \mathbb{P}\big(Y \in B\big) = \mathbb{P}\big(X \in f^{-1}(B)\big) = \mu\big(f^{-1}(B)\big), \quad B \in \mathcal{F}.$$

Def. (Joint distribution)
If X and Y are random variables on $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively, then $Z = (X, Y)$ is random variable on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ and the distribution of Z is called the ***joint distribution of X and Y***, which is fully specified by

$$\pi(A \times B) = \mathbb{P}\big(X \in A, Y \in B\big), \quad \text{for all } A \in \mathcal{E}, B \in \mathcal{F}.$$

Def. (Marginal distribution)
If $Z = (X, Y)$ is a r.v. on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ that has joint distribution π, then the ***marginal distributions*** of X and Y are, respectively,

$$\mu(A) = \pi(A \times F) \quad \text{and} \quad \nu(B) = \pi(E \times B).$$

Def. (Independence)
With the previous assumptions, X and Y are said to be ***independent*** if their joint distribution is

$$\mathbb{P}\big(X \in A, Y \in B\big) = \mathbb{P}\big(X \in A\big)\mathbb{P}\big(Y \in B\big), \quad A \in \mathcal{E}, B \in \mathcal{F}.$$

Remark An arbitrary collection (countable or uncountable) of random variables is said to be *independent* if every finite subcollection $(X_{i_1}, \ldots, X_{i_n})$ is independent.

If X is a random variable, then its integral w.r.t. the measure $\mathbb{P}$ makes sense to talk about, since by definition it is $\mathcal{F}$-measurable.

Def. (Expected value)
The integral of X w.r.t the measure $\mathbb{P}$ is called the ***expected value of X***,

$$\mathbb{E}\big[X\big] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X d\mathbb{P}.$$

If $\mathbb{E}\big[X\big] < \infty$ then X is said to be integrable.

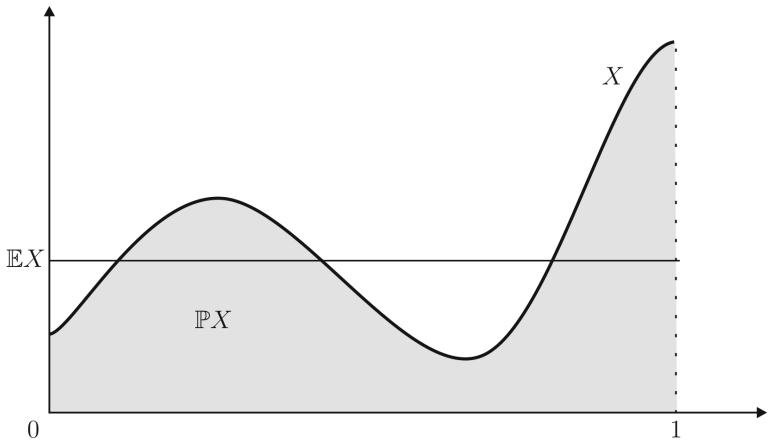

Figure 1: The integral $\mathbb{P}(X)$ is the area under X, the expected value $\mathbb{E}(X)$ is the constant "closest" to X.

Theorem 1 (Law of the unconscious statistician)
If X is a r.v. on $(E, \mathcal{E})$ and f is $\mathcal{E}$-measurable, then

$$\mathbb{E}\big[f(X)\big] = \int_\Omega f\big(X(\omega)\big)\mathbb{P}(d\omega)$$

Remark Choosing $f(X) = \mathbb{1}_A$, we find that $\mathbb{E}\big[\mathbb{1}_A(X)\big] = \mathbb{P}\big(X \in A\big)$.

0.3 L^p spaces

Def. (p-norm)
For $p \in [1, \infty)$ we define the ***p-norm of X*** to be

$$\|X\|_p = \mathbb{E}\big[|X|^p\big]^{1/p},$$

and for $p = \infty$ we define it as the ***essential supremum*** of X

$$\|X\|_\infty = \inf_{b \in \mathbb{R}^+} \{|X| \leq b \text{ almost surely}\}.$$

Remarks

- $\|X\|_p = 0 \implies X \equiv 0$ almost surely.
- $\|cX\|_p = c\|X\|_p$ for $x \geq 0$.

We have a very famous theorem which defines the relationship between different random variable norms.

Theorem 2 (Hölder's inequality)
For $p, q, r \in [1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$,

$$\|XY\|_r \leq \|X\|_p \|Y\|_q,$$

in particular for $r = 1, p = 2, q = 2$ we have Schwartz's inequality

$$\|XY\|_1 \leq \|X\|_2 \|Y\|_2.$$

Theorem 3 (Minkowski's inequality)
For $p \in [1, \infty]$,

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Lemma 1 (Jensen's inequality)
Let D be a convex subset of $\mathbb{R}^d$ and $f : D \longrightarrow \mathbb{R}$ be continuous and concave. If $X_1, \ldots, X_d$ are integrable r.v. and $(X_1, \ldots, X_d) \in D$ almost surely. Then,

$$\mathbb{E}\big[f(X_1, \ldots, X_d)\big] \leq f\big(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]\big).$$

0.4 Generating functions

References Paolella (2007, §1)

Various integrals of interest are obtained by choosing an appropriate function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of two variables, (t, X), and are usually viewed as a function of t after integration wrt to X,

$$\mathbb{E}\big[g(t, X)\big] = \int_{-\infty}^{\infty} g(t, x)\, \mathrm{d}F_X(x).$$

Some notable examples of these functions include the following:

- *n-th moment*: $g(n, x) = x^n \implies \mathbb{E}\big[X^n\big]$
- *n-th abs. moment*: $g(n, x) = |x|^n \implies \mathbb{E}\big[|X|^n\big]$
- *Probability-generating function*: $g(t, x) = t^x \implies G(t) = \mathbb{E}\big[t^X\big]$. This function is useful for discrete random variables, since
  - $p(k) = \mathbb{P}\big(X = k\big) = \frac{1}{k!} \cdot \frac{\partial}{\partial t} G(t)\Big|_{t=0}$
  - $G_X = G_Y \implies p_X = p_Y$.
  - The k^{th} factorial moment is

$$\mathbb{E}\left[\frac{X!}{(X-k)!}\right] = \frac{\partial}{\partial t} G(t)\Big|_{t=1^-}$$

– If $M_X(t)$ is the moment-generating function of X, then

$$G_X(e^t) = M_X(t).$$

– If $N \sim \mathbb{P}_N$ and $S_N = \sum_{i=1}^N X_i$, with $X_i \overset{\text{iid}}{\sim} \mathbb{P}_X$ and $N \perp\!\!\!\perp X_i$, then using the law of total expectation we have

$$G_{S_N}(t) = \mathbb{E}_{\mathbb{P}_N}\Big[\mathbb{E}_{\mathbb{P}_X}\big[t^{\sum_{i=1}^N X_i}|N\big]\Big] = \mathbb{E}_{\mathbb{P}_N}\big[G_X(t)^N\big] = G_N\big(G_X(t)\big).$$

0.4.1 Moment-generating function

Def. (Moment-generating function)
The ***moment-generating function*** (mgf) of a random variable X is the function $t \mapsto e^{tX}$ and is said to *exist* if there is an $h > 0$ such that

$$\text{For all } t \in (-h, h), \quad M_X(t) < \infty.$$

Remarks

› If $M_X(t)$ exists, then the *convergence strip* of $M_X(t)$ is the largest open interval such that $M_X(t) < \infty$,

$$\sup_h \big\{(-h, h) : M_X(t) < \infty \quad \forall t \in (-h, h)\big\}.$$

› For a location-scale family, if $Z = \mu + \sigma X$ we have that

$$M_Z(t) = \mathbb{E}\big[e^{t(\mu+\sigma X)}\big] = e^{\mu t} M_X(\sigma t).$$

› If $N \sim \mathbb{P}_N$ and $S_N = \sum_{i=1}^N X_i$, with $X_i \overset{\text{iid}}{\sim} \mathbb{P}_X$ and $N \perp\!\!\!\perp X_i$, then again by using the law of total expectation we have

$$M_{S_N}(t) = \mathbb{E}_{\mathbb{P}_N}\Big[\mathbb{E}_{\mathbb{P}_X}\big[e^{t\sum_{i=1}^N X_i}|N\big]\Big] = \mathbb{E}_{\mathbb{P}_N}\big[M_X(t)^N\big] = G_N\big(M_X(t)\big).$$

Theorem 4 (Existence of absoute moments)
If $M_X(t)$ exists, then for all $r \in (0, +\infty)$ we have that

$$\mathbb{E}\big[|X|^r\big] < \infty.$$

It can be shown that the derivative operator can be moved inside the expectation, and the moment-generating function can be used to compute the k^{th} moment of X.

Theorem 5 (Generation of moments)
If $M_X(t)$ exists, then we can write

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \mathbb{E}\big[e^{tX}\big] = \mathbb{E}\left[\frac{\partial}{\partial t} e^{tX}\right] = \mathbb{E}\big[X^j e^{tX}\big],$$

and therefore $\mathbb{E}\big[X^j\big] = \frac{\partial}{\partial t} M_X(t)\Big|_{t=0}$.

Example (mgf of DUnif(ϑ))
Let $X \sim \text{DUnif}(\vartheta)$, i.e. X is discrete with pmf

$$p_X(x;\vartheta) = \frac{1}{\vartheta} \mathbb{1}_{\{1,2,\dots,\vartheta\}}(x).$$

Then, the mgf of X is

$$M_X(t) = \mathbb{E}\big[e^{tX}\big] = \frac{1}{\vartheta} \sum_{i=1}^{\vartheta} e^{tj}.$$

From this, we can easily calculate $\mathbb{E}\big[X\big]$ simply by deriving wrt to t

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\vartheta} \frac{\partial}{\partial t} \sum_{j=1}^{\vartheta} e^{tj}\Big|_{t=0} \\
&= \frac{1}{\vartheta} \sum_{j=1}^{\vartheta} j e^{tj}\Big|_{t=0} \\
&= \frac{1}{\vartheta} \sum_{j=1}^{\vartheta} j \\
&= \frac{1}{\vartheta} \frac{\vartheta(\vartheta+1)}{2} \\
&= \frac{\vartheta+1}{2}.
\end{aligned}$$

Example (mgf of Unif(0, 1))
Let $X \sim \text{Unif}(0,1)$, then we find that the mgf of X is

$$M_X(t) = \int_0^1 e^{tx}\, \mathrm{d}x = \frac{1}{t}(e^t - 1),$$

which exists finite for all $t \in (0,1)$. Since the Taylor expansion of $M_X(t)$ around zero is

$$\frac{e^t - 1}{t} \overset{t\approx 0}{=} \frac{1}{t}\left(t + \frac{t^2}{2} + \frac{t^3}{6} + \frac{t^4}{24} + \dots\right) = 1 + \frac{t}{2} + \frac{t^2}{6} + \dots = \sum_{j=0}^{\infty} \frac{t^j}{(j+1)!},$$

we have that the r^{th} derivative has only the r^{th} term constantly equal to 1 in t at the numerator, and therefore

$$\mathbb{E}\big[X^r\big] = \frac{1}{r+1}.$$

For the multivariate case, we have a straightforward generalization of the mgf using vector notation.

Def. (Multivariate moment-generating function)
Let X be a multivariate r.v, then its ***moment-generating function*** is

$$M_X(t) = \mathbb{E}\big[e^{t^\top X}\big].$$

Theorem 6 (Sawa)
Let X_1, X_2 be r.v.s such that $\mathbb{P}\big(X_1 > 0\big) = 1$ with joint mgf $M_{X_1,X_2}(t_1, t_2)$ which exists for $t_1 < \varepsilon$ and $|t_2| < \varepsilon$, $\varepsilon > 0$. Then, we have that

$$\mathbb{E}\Big[\Big(\frac{X_2}{X_1}\Big)^k\Big] = \frac{1}{\Gamma(k)} \int_{-\infty}^{0} (-t_1)^{k-1} \Big[\frac{\partial^k}{\partial t_2^k} M_{x_1,x_2}(t_1, t_2)\Big]_{t_2=0} \mathrm{d}t_1.$$

0.4.2 Cumulant-generating function

Def. (Cumulant-generating function)
Let $M_X(t)$ be the moment-generating function of a r.v. X. Then, the ***cumulant-generating function*** $K_X(t)$ of X is

$$K_X(t) = \log M_X(t).$$

Remarks

› If $S_n = \sum_{i=1}^n X_i$ with X_i i.i.d, then

$$K_{S_n}(t) = nK_X(t).$$

› The j^{th} derivative of K_X evaluated at $t = 0$ is the j^{th} cumulant of X,

$$\kappa_j = \frac{\partial^j}{\partial t^j} K_X(t)\Big|_{t=0},$$

where if $\mu_j = \mathbb{E}\big[X^j\big]$, the first four cumulants are given by (Pace and Salvan, 1997):

$$\kappa_1 = \mu_1,$$

$$\kappa_2 = \mu_2 - \mu_1^2,$$

$$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^2,$$

$$\kappa_4 = \mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 + 6\mu_1^4.$$

Example (cgf of a $\mathcal{N}(\mu, \sigma^2)$)

For $X \sim \mathcal{N}(\mu, \sigma^2)$ we have that the moment-generating function is

$$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}} \implies K_X(t) = \log M_X(t) = \mu t + \sigma^2 \frac{t^2}{2}.$$

Therefore, the first two cumulants are

$$\begin{cases} \kappa_1 = \frac{\partial}{\partial t}\left(\mu t + \sigma^2 \frac{t^2}{2}\right)\Big|_{t=0} & = \mu, \\ \kappa_2 = \frac{\partial^2}{\partial t^2}\left(\mu t + \sigma^2 \frac{t^2}{2}\right)\Big|_{t=0} & = \sigma^2. \end{cases}$$

Other examples of cgf's can be found in (Paolella, 2007, pp. 8–10).

Lecture 1: Convergence and limit theorems

2021-10-14

References Gut (2009), first portion of the course

Email: stefano.pagliarani9@unibo.it

The course will be focussed on the stochastic processes portion of probability theory, after a brief reminder of limit theorems, conditional probability, and measure theory.

1.1 Convergence of random variables

Convergence of random variables is a little bit trickier than just real numbers.

Notation: AC is the set of absolutely continuous probability measures wrt the Lebesgue measure.

› *Absolute continuity*: if $\mu \in \text{AC}$ is absolutely continuous, we write

$$\mu(dx) = f(x)dx$$

› *Integration in measure spaces:* Let $X \sim \mu$, then by a theorem we have

$$\mathbb{E}\big[f(X)\big] = \int_{\mathbb{R}^d} f(x)\mu(dx), \tag{1}$$

and we can differentiate between two types of distribution:

a) μ discrete $\implies \mathbb{E}\big[X\big] = \sum_n xp(x)$

b) $\mu \in \text{AC} \implies \mathbb{E}\big[X\big] = \int_{\mathbb{R}^d} x \cdot f(x)dx$

Example (Intuition of convergence)
Consider $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$ for $n \in \mathbb{N}$, and it is absolutely continuous w.r.t. Lebesgue measure. This means that it admits a probability density which is defined by

$$\mu_n(dx) = \left(\begin{cases} n & \text{if } x \in [0, \frac{1}{n}] \\ 0 & \text{if } x \notin [0, \frac{1}{n}] \end{cases} \right) dx$$

It is intuitive to think that the measure is converging to a spike in zero, i.e.

$$\mu_n \xrightarrow{n\to\infty} \delta_0,$$

where δ_x denotes the Dirac delta distribution centered in x, such that $\delta_x(\{x\}) = 1$. We need to mathematically characterize this type of convergence in a more formal way than by intuition.

Maybe it could be that for any Borel set $A \subseteq \mathscr{B}(\mathbb{R})$,

$$\mu_n(A) \xrightarrow{n\to\infty} \delta_0(A),$$

but unfortunately this is wrong since we can see that, for $A = \{0\}$ and for all $n \in \mathbb{N}$:

$$\mu_n(\{0\}) = 0 \neq 1 = \delta_0(\{0\}).$$

So we can either throw out the idea that the uniform converges to a Dirac delta, or change the definition of convergence to accommodate for the behaviour in Figure 2.

Moreover, assume now that $X_n \sim \mu_n$ such that $\mu_n \xrightarrow{n\to\infty} \delta_0$, what can we say about the properties of X_n? In general (as we will see afterwards), this depends on the specific type of convergence that we assume.

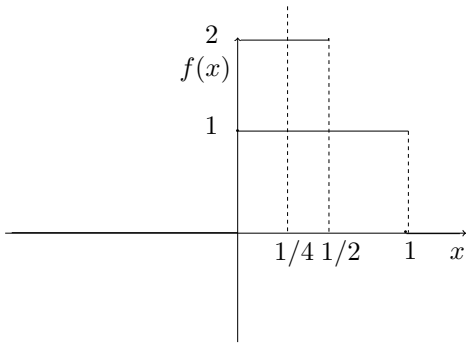

Figure 2: Convergence of the sequence of uniform distributions to the Dirac measure in zero.

Def. (Convergence in distribution)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $(\mathbb{R}^d, \mathscr{B})$. We say that μ_n ***converges in distribution*** to another distribution μ,

$$\mu_n \xrightarrow{d} \mu,$$

if, for any possible choice of *test function* $f \in C_b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} f(x)\mu(dx).$$

This convergence is in the sense of standard real analysis.

Notation: $C_b(\mathbb{R}^d)$ is the set of continuous bounded functions

Remark All test functions f define a measure when integrated wrt to $\mu_n(dx)$, and when all said measures are equal to those obtained by integrating against another distribution μ, then we obtain the convergence in distribution.

Example (Uniform distribution)

Consider $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$ and $\mu = \delta_0$, take any function $f \in C_b(\mathbb{R})$ and compute

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mu_n(dx) &= \int_0^{\frac{1}{n}} f(x) \cdot n \cdot dx \\ &= n \cdot \underbrace{\int_{[0,\frac{1}{n}]} f(x)dx}_{\approx \frac{1}{n}\cdot f(0)} \\ &\xrightarrow{n\to\infty} f(0).\end{aligned}$$

The last equality holds since f is continuous, and by the mean value theorem we can approximate it by the left extrema. However, by definition of the abstract integral wrt the Dirac delta function we have that

$$f(0) = \int_{\mathbb{R}} f(x)\delta_0(dx),$$

which proves that $\mu_n \xrightarrow{d} \mu$.

Remark If $A \in \mathscr{B}(\mathbb{R}^d)$ is an event and μ is a distribution, then

$$\mu(A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)dx,$$

where $\mathbb{1}_A$ is the indicator function such that

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Had we used $f \notin C_b(\mathbb{R}^d)$ instead, then we could have chosen $f = \mathbb{1}_{\{0\}}$ and convergence in distribution would not have been satisfied. The example below shows another case in which another type of convergence is useful in order to characterize a common-sense behaviour of random variables.

Example (Sequence of Dirac functions)
Consider $\mu_n = \delta_{1/n}$ and $\mu = \delta_0$, then it is clear that this is a discrete measure that in some intuitive sense converges to zero. If we choose $f(x) = \mathbb{1}_{\{0\}}$, then we find that

$$\int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} \mathbb{1}_{\{0\}}(x)\delta_{\frac{1}{n}}(dx) = \mathbb{1}_{\{0\}}(1/n) = 0 \quad \forall n,$$

and therefore does not converges to δ_0.

Recall: A random variable is such that the event $(X_n \in A) \in \mathcal{F}_n$, which means that the function is measurable.

Def. (Weak convergence of random variables)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables, $X_n : (\Omega_n, \mathcal{F}_n, \mathbb{P}_n) \longrightarrow (\mathbb{R}^d, \mathscr{B})$. Let now X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we say that X_n ***converges weakly/in distribution/in law***, $X_n \xrightarrow{d} X$, if their measures are such that

$$\mu_{X_n} \xrightarrow{d} \mu_X.$$

Remark By the definition of expected value in Equation (1), a family of random variables $(X_n)_{n\in\mathbb{N}}$ is such that, for any $f \in C_b(\mathbb{R}^d)$

$$X_n \xrightarrow{d} X \iff \mathbb{E}\big[f(X_n)\big] \xrightarrow{n\to\infty} \mathbb{E}\big[f(X)\big].$$

This is however the weakest type of convergence out of all those that we will consider, since in other cases the probability spaces might be different.

Def. (Stronger definitions of convergence)
$(X_n)_{n\in\mathbb{N}}$ sequence of random variables and X a r.v., all defined on the same probability space

$$X_n, X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (\mathbb{R}^d, \mathscr{B}).$$

Then we say that

a) X_n ***converges in L^p*** to X, and we denote it by $X_n \xrightarrow{L^p} X$ if X_n and X are random variables in $L^p = \{\text{r.v. on } (\Omega, \mathcal{F}, \mathbb{P}) : \mathbb{E}[|X|^p] < \infty\}$ and

$$\|X_n - X\|_{L^p} \xrightarrow{n\to\infty} 0,$$

where $\|X\|_{L^p} = \mathbb{E}\big[|X|^p\big]^{\frac{1}{p}}$.

b) X_n ***converges in probability*** to X, and we denote it by $X_n \xrightarrow{P} X$ if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = 0.$$

c) X_n ***converges almost surely*** to X, and we denote it by $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1,$$

where the event inside $\mathbb{P}$ is in the sense of real analysis,

$$\Big\{w \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\Big\},$$

which can be proven to be a measurable set and therefore a valid event.

Remark The L^p norm of the difference induces a *distance between functions* in the sense of functional analysis.

Example (Difference in interpretation)
Consider a Bernoulli game where we equally bet on an outcome ± 1. The second type of convergence does not tell us that almost surely our gain will converge to zero, but rather that we can set a small tolerance and find some n such that our gain will be smaller than that.

The following inequality is a basic tool for probability, which will be useful later on.

Theorem 7 (Markov's inequality)
Let X be a r.v. and $\lambda > 0$, then

$$\mathbb{P}\big(|X| > \lambda\big) \leq \frac{\mathbb{E}\big[|X|^p\big]}{\lambda^p}, \quad p \geq 0.$$

Proof.
If $\mathbb{E}\big[|X|^p\big] = \infty$, then there is nothing to prove. If instead $\mathbb{E}\big[|X|^p\big] < \infty$, then since $\mathbb{1}_A$ is either 1 or 0 we have

$$\begin{aligned}
\mathbb{E}\big[|X|^p\big] &\geq \mathbb{E}\big[|X|^p \cdot \mathbb{1}_{|X|>\lambda}\big] \\
&\geq \mathbb{E}\big[\lambda^p \cdot \mathbb{1}_{|X|>\lambda}\big] \qquad \text{(since } |X| \geq \lambda\text{)} \\
&= \lambda^p \cdot \mathbb{P}\big(|X| > \lambda\big).
\end{aligned}$$

□

Corollary 1 (Chebyshev's inequality)
By choosing $p = 2$ and considering the random variable $X - \mathbb{E}[X]$, Markov's inequality states that

$$\mathbb{P}\big[|X - \mathbb{E}[X]| > \lambda\big] \leq \frac{\mathbb{E}\big[|X - \mathbb{E}[X]|^2\big]}{\lambda^2} = \frac{\mathbb{V}[X]}{\lambda^2}.$$

Theorem 8
Under the according assumptions for X_n, X we have the following set of implications:

1. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X.$
2. $X_n \xrightarrow{P} X \implies$ *there is a subsequence* X_{k_n} *such that* $X_{k_n} \xrightarrow{a.s.} X.$
3. $X_n \xrightarrow{d} X \implies X_n \xrightarrow{P} X$ *iff* $\mu_X = \delta_{x_0}$
4. $X_n \xrightarrow{L^1} X \implies X_n \xrightarrow{P} X$
5. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{L^1} X$ *iff* $|X_n| \leq Y \in L^p$

Proof.

1. $\boxed{\text{a.s.} \implies \text{p}}$: $\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = \mathbb{E}\big[\mathbb{1}_{|X_n - X| \geq \varepsilon}\big]$ and the indicator function converges to zero as $n \to \infty$ by assumption. Since $\mathbb{1}_A$ is bounded, by the dominated convergence theorem the integral (expectation) also converges to zero.

4. $\boxed{L^p \implies p}$: Follows as a consequence of Markov's property, since we can majorize the probability by the expected value

$$\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) \overset{\text{Thm.7}}{\leq} \frac{\mathbb{E}\big[|X_n - X|^p\big]}{\varepsilon^p} = \frac{\|X_n - X\|_{L^p}^p}{\varepsilon^p} \xrightarrow{n\to\infty} 0.$$

where the convergence to 0 is a consequence of the L^p convergence assumption.

□

Example (A.s. does not imply L^p)

Let $m \in \mathbb{R}$ and $X_n = n^m \mathbb{1}_{[0,\frac{1}{n}]}$ on the probability space $\big([0,1], \mathscr{B}([0,1]), \lambda_{[0,1]}\big) \to \mathbb{R}$, and let's try to establish some convergence for the random variable X_n.

› If $\omega > 0$, then we can find some $\bar{n}$ such that X_n is equal to zero:

$$X_n(\omega) = n^m \mathbb{1}_{[0,\frac{1}{n}]}(\omega) \xrightarrow{n\to\infty} 0.$$

› If $\omega = 0$, then

$$X_n(0) = n^m \xrightarrow{n\to\infty} +\infty, \quad \text{for } m > 0,$$

however the event $\{0\}$ has null probability since we have a uniform distribution on $\big[0, \frac{1}{n}\big]$ at all steps of the limit, and as such we have

$$\mathbb{P}_{\mu_n}\big(\{0\}\big) = 0 \quad \text{for all } n \in \mathbb{N}.$$

Therefore, the set of limit elements for absolute convergence is

$$\Big\{\omega \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\Big\} = \Omega \setminus \{0\}.$$

Since $\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = \mathbb{P}\big(\Omega \setminus \{0\}\big) = 1$, we have that

$$X_n \xrightarrow{\text{a.s.}} X \equiv 0 \qquad (\implies X \xrightarrow{P} X).$$

On the other hand for L^p convergence we have that

$$\begin{aligned}
\mathbb{E}\big[|X_n - X|^p\big] &= \mathbb{E}\big[|X_n|^p\big] \\
&= \int_{[0,1]} n^{mp} \cdot \mathbb{1}_{[0,\frac{1}{n}]}(x) dx \\
&= n^{mp} \cdot \frac{1}{n} \\
&= n^{mp-1}.
\end{aligned}$$

We conclude that $X_n \xrightarrow{L^p} X \iff mp - 1 < 0 \iff m < 1/p$, but we always have almost-sure convergence for any $m > 0$.

Example (Gaussian distribution)
Consider $\mathcal{N}_{\mu,\sigma^2} = \varphi_{\mu,\sigma^2}(x)dx$, with

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

Consider now a sequence of real numbers $\mu_n \to \mu$ and a sequence of real numbers $\sigma_n \to 0$.

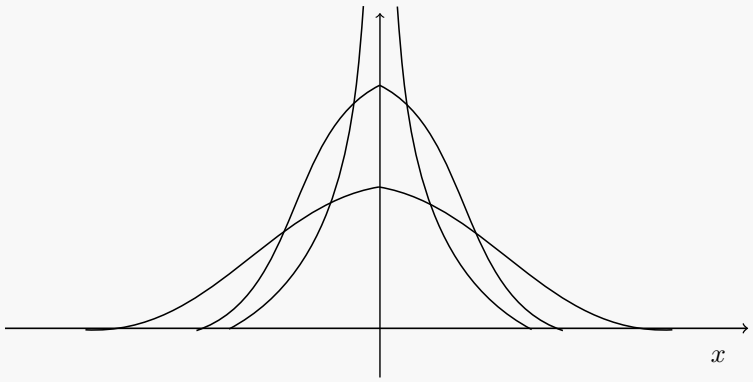

Figure 3: Convergence of the normal distribution to the Dirac delta function.

So we can expect that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$. As an exercise, prove this convergence (use a simple change of variables).

However, for the Gaussian case we can prove something stronger: if $X_n \sim \mathcal{N}_{\mu_n,\sigma_n}$ and $X \equiv \mu$ we can prove convergence in L^2. Using the triangle inequality, we can write

$$\mathbb{E}\big[|X_n - \mu|^2\big] \leq \mathbb{E}\big[|X_n - \mu_n|^2 + \underbrace{|\mu_n - \mu|^2}_{\to 0}\big],$$

and since $\mathbb{E}\big[|X_n - \mu_n|^2\big] = \mathbb{V}\big[X_n\big] = \sigma_n^2 \xrightarrow{n\to\infty} 0$, we also have L^2 convergence.

Exercise: prove that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$ if $\mu_n \to \mu$ and $\sigma_n \to 0$.

Proof.
Consider any test function $f \in \mathrm{C}_b(\mathbb{R})$, then if $\varphi(t)$ is the pdf of a $\mathcal{N}_{0,1}$ distribution we have that

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mathcal{N}_{\mu_n,\sigma_n}(dx) &= \int_{\mathbb{R}} f(x) \cdot \frac{1}{\sigma_n} \cdot \varphi\left(\frac{x-\mu_n}{\sigma_n}\right) dx && \text{(abs. continuity)} \\ &= \int_{\mathbb{R}} f\left(\sigma_n y + \mu_n\right) \cdot \cancel{\sigma_n} \frac{1}{\cancel{\sigma_n}} \varphi(y) dy && \text{(change of var.)}.\end{aligned}$$

Since both f and φ are bounded the function $t \mapsto f(t)\varphi(t)$ is bounded by $g(t) = \max_{t'} f(t') \cdot \varphi(t)$, which is Lebesgue integrable and the dominated convergence theorem can be therefore applied to

obtain the following equivalence

$$\lim_{n\to\infty}\int_{\mathbb{R}} f\left(\sigma_n y+\mu_n\right)\varphi(y)dy=\int_{\mathbb{R}}\lim_{n\to\infty} f(\sigma_n y+\mu_n)\varphi(y)dy=f(\mu)\int_{\mathbb{R}}\varphi(y)dy=f(\mu).$$

Therefore we have convergence in distribution to δ_μ by definition of the abstract integral wrt the Dirac measure.

□

Def. (C.d.f. of a distribution)
Given a distribution μ on $\mathbb{R}$, the **cdf** of μ is the function $F_\mu : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_\mu(x)=\mu\big((-\infty,x]\big).$$

Remark Among all known properties such as monotonicity, boundedness, etc, the most important for what follows is the property of *right-continuity*.

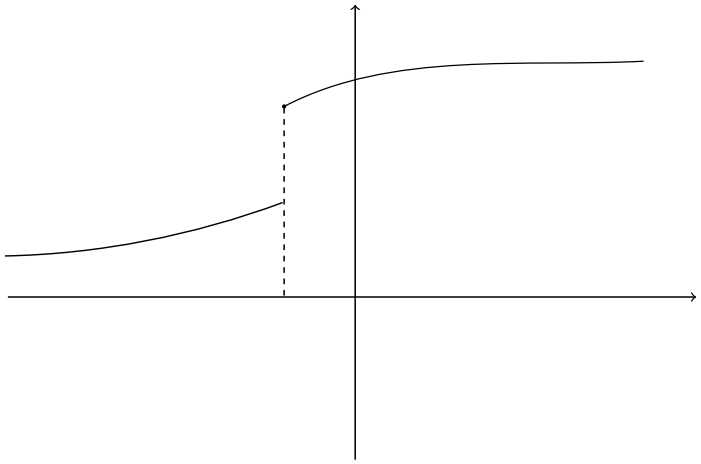

Figure 4: Right-continuity of the cumulative distribution function.

Def. (Cumulative distribution function)
Let X be a real-valued random variable, then the ***cumulative distribution function*** (CDF) of X is the function $F_X : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_X(x)=F_{\mu_X}(x)=\mathbb{P}\big(X\leq x\big)$$

Since the property of convergence in distribution is quite hard to prove for any bounded test function f, we want to characterize this property with respect to something else in order to make it easier to check it.

Example (Cdf of a uniform distribution)

Let $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$, then the cdf is

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ nx & \text{if } 0 < x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}$$

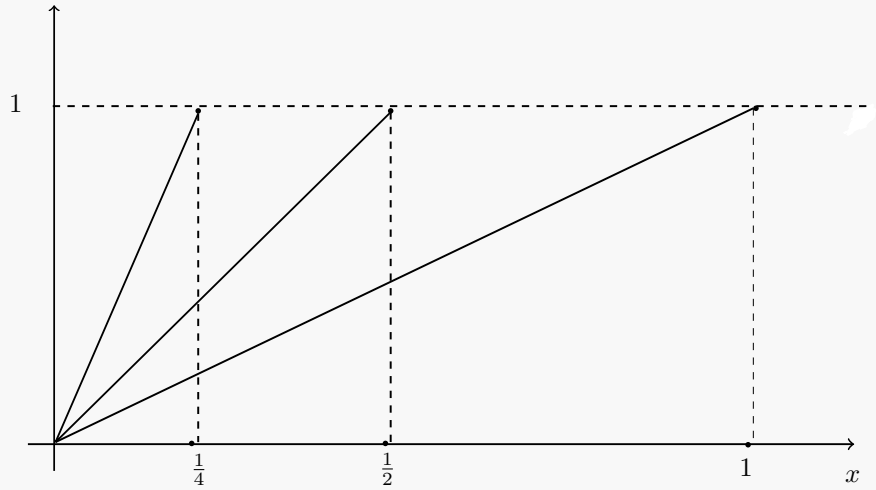

Figure 5: Convergence of the cdf of the uniform distribution to the unit step function.

The Dirac delta measure has a very simple cdf given by the unit step function,

$$F(x) = \mathbb{1}_{[0,\infty)}(x),$$

and in this example we have convergence of $F_n(x) \to F(x)$ in all points $x \in \mathbb{R}$ except for $x = 0$, since $F_n(0) = 0$ for all $n \in \mathbb{N}$.

Theorem 9 (Characterization of $\xrightarrow{d}$ using the cdf)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then we have that

$$\mu_n \xrightarrow{d} \mu \iff F_{\mu_n}(x) \xrightarrow{n\to\infty} F_\mu(x),$$

for all x that are points of continuity of F_μ.

Proof.
No.

□

Remark There can also be convergence in points of discontinuity, but it is not guaranteed in general.

Example (of convergence in the points of discontinuity)

$\mu_n = \delta_{-\frac{1}{n}}$, then it is clear that in this case also $\mu_n \to \delta_0$, and continuity is guaranteed for all points $x > 0$. However, in this case the cdf is such that

$$F_{\mu_n}(0) = F_{\delta_{-\frac{1}{n}}}(0) = 1 \quad \text{for all } n \in \mathbb{N},$$

therefore $\lim_{n\to\infty} F_{\mu_n}(0) = 1$ and convergence is satisfied both in the points of continuity as well as in the point of discontinuity of F.

Let us now discuss another important function when dealing with real-valued random variables, which also allows a convenient characterization of $\xrightarrow{d}$.

Def. (Characteristic function of a distribution)

Let μ be a distribution, then we say that the ***characteristic function*** (CHF) of μ is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} \mu(dx).$$

Def. (Characteristic function of a random variable)

Let X be a random variable with distribution μ on $\mathbb{R}^d$, then the ***characteristic function of X*** is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi_X(\eta) = \varphi_{\mu_X}(\eta) = \mathbb{E}\big[e^{i\langle X, \eta\rangle}\big].$$

Remark If $\mu \in$ AC has density f, then we can write it exactly as a Lebesgue integral and it equals to a scaled and "slowed" version of the Fourier transform,

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} f(x) dx.$$

Theorem 10 (Lévy, characterization of $\xrightarrow{d}$ using the CHF)

Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then

a) $\mu_n \xrightarrow{d} \mu \implies \varphi_n(\eta) \xrightarrow{n\to\infty} \varphi(\eta)$ *for any* $\eta \in \mathbb{R}^d$.

b) $\varphi \xrightarrow{n\to\infty} \varphi$ *everywhere, with* φ *continuous in* $\eta = 0$, *then* φ *is a CHF of a distribution* μ *and* $\mu_n \xrightarrow{d} \mu$.

Remark CHF's have some interesting properties, most notably

1. $\varphi(0) = 1$ since $\mathbb{E}\big[e^{i\langle 0, x\rangle}\big] = \mathbb{E}\big[1\big] = 1$.

2. φ_X is continuous in $\nu = 0$, which we can check by the limiting procedure

$$\lim_{\eta \to 0} \varphi_X(\eta) \overset{?}{=} \varphi_X(0) = 1.$$

Since $e^{i\vartheta} = \cos\vartheta + i\sin\vartheta$ is always equal in norm to 1 (Euler's formula), we can apply the dominated convergence theorem

$$\lim_{\eta \to 0} \mathbb{E}\big[e^{i\langle X,\eta\rangle}\big] \overset{\text{DCT}}{=} \mathbb{E}\big[\lim_{\eta \to 0} e^{i\langle X,\eta\rangle}\big] = \mathbb{E}\big[1\big] = 1.$$

1.2 Limit theorems

Notation: If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables, we define the partial sums and partial means by

$$S_n = X_1 + X_2 + \ldots + X_n,$$

$$M_n = S_n/n.$$

Theorem 11 (Law of large numbers)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables in $L^1(\Omega,\mathbb{P})$ that are i.i.d with mean $\mathbb{E}\big[X_n\big] = \mu$, then

- *(Weak L.L.N.) $M_n \xrightarrow{d} \mu$ and therefore $M_n \xrightarrow{P} \mu$ since μ is a constant.*
- *(Strong L.L.N.) $M_n \xrightarrow{a.s.} \mu$*

Proof.
We only prove the weak form since the strong one is very difficult. However, even for the weak form we would have to prove Lévy's theorem, which is also quite difficult. We will use the following lemma for proving the weak law of large numbers:

Lemma 2 (First derivative of the CHF)
For the CHF of a random variable X we can

$$\begin{aligned}\frac{\partial\varphi_X(\eta)}{\partial\eta} &= \frac{\partial}{\partial\eta}\mathbb{E}\big[e^{i\eta X}\big] \\ &= \mathbb{E}\big[\frac{\partial}{\partial\eta}e^{i\eta X}\big] \qquad (DCT\ since) \\ &= \mathbb{E}\big[iXe^{i\eta X}\big]\end{aligned}$$

And computing this value in $\eta = 0$, we have that

$$\frac{\partial}{\partial\eta}\varphi_X(\eta)\Big|_{\eta=0} = i\mathbb{E}\big[X\big].$$

We want to prove that the CHF of M_n converges to that of δ_μ and then use Lévy's theorem:

$$\lim_{n\to\infty} \varphi_{M_n}(\eta) \overset{?}{=} e^{i\eta\mu} = \mathbb{E}\big[e^{i\eta\mu}\big].$$

Start by explicitly writing the CHF of M_n:

$$\begin{aligned}\varphi_{M_n}(\eta) &= \mathbb{E}\big[e^{i\eta\frac{1}{n}\sum_{j=1}^n X_j}\big] \\ &= \mathbb{E}\big[\prod_{j=1}^n e^{i\frac{\eta}{n}X_j}\big] \\ &= \mathbb{E}\big[e^{i\frac{\eta}{n}X_1}\big]^n \qquad \text{(i.i.d)} \\ &= \varphi\big(\frac{\eta}{n}\big)^n.\end{aligned}$$

Using Lemma 2 we can apply a Taylor expansion of φ_{M_n} around $\eta = 0$:

$$\begin{aligned}\varphi_{M_n}(\eta) &= \Big(1 + \frac{\eta}{n}i\mu + o\big(\frac{1}{n}\big)\Big)^n \\ &= \Big(1 + \frac{\eta i\mu + \overbrace{n\cdot o\big(\frac{1}{n}\big)}^{\xrightarrow{n\to\infty}0}}{n}\Big)^n \\ &= e^{i\eta\mu} \qquad \text{(standard limit)}\end{aligned}$$

□

Remark Had we also assumed that $X_n \in L^2(\Omega,\mathbb{P})$ with $\mathbb{V}\big[X_n\big] = \sigma^2$, then this would've become a one-line proof since

$$\mathbb{P}\big(|M_n - \mu| > \varepsilon\big) \leq \frac{\mathbb{E}\big[\overbrace{|M_n-\mu|^2}^{L^2\text{converg.}}\big]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Using this, we have convergence in L^2 which implies $\xrightarrow{P}$ and $\xrightarrow{d}$. These inequalities are useful as a very basic estimate of the speed of convergence for Monte Carlo simulations and confidence regions, in order to provide error bounds. However, proper estimates are more refined and will be discussed later on.

Lecture 2: Central limit theorems

2021-10-21

One could already be satisfied with the LLN, which describes the behaviour of the empirical average M_n. However, this doesn't tell us what the distribution of M_n will look like as $n \longrightarrow \infty$.

Given the ways we saw in the examples, how does the law μ_{M_n} approach μ?

We can first compute some quantities related to M_n:

- $\mathbb{E}\big[M_n\big] = \mu$
- $\mathbb{V}\big[M_n\big] = \frac{\sigma^2}{n}$

We will try now to normalize the empirical average and see what we obtain as a result:

$$\tilde{M}_n = \frac{M_n - \mu}{\mathrm{sd}(M_n)} = \frac{\sqrt{n}(M_n - \mu)}{\sigma}.$$

Theorem 12 (Central limit theorem)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d r.v. in $L^2(\Omega, \mathbb{P})$, i.e. with finite variance, then we have that the normalized empirical average $\tilde{M}_n$ is such that

$$\tilde{M}_n \xrightarrow{d} \mathcal{N}_{0,1}.$$

Proof.
We use the following lemma for proving the central limit theorem:

Lemma 3 (Second derivative of the CHF)
We have that if $X \in L^2(\Omega, \mathbb{P})$,

$$\begin{aligned}\frac{\partial^2}{\partial\eta^2}\varphi_X(\eta) &= \frac{\partial}{\partial\eta}\mathbb{E}\big[iXe^{i\eta X}\big] \\ &= -\mathbb{E}\big[X^2 e^{i\eta X}\big] \qquad\qquad DCT \text{ if } \mathbb{E}[X^2] < \infty\end{aligned}$$

And by computing the derivative in $\eta = 0$,

$$\frac{\partial^2}{\partial\eta^2}\varphi_X(\eta)\Big|_{\eta=0} = -\mathbb{E}\big[X^2\big].$$

Consider $\mu = 0, \sigma^2 = 1$ which is not restrictive by the properties of the normal distribution.

$$\frac{M_n - \mu}{\sigma} = \frac{\frac{1}{n}\sum_{j=1}^n X_j - \mu}{\sigma} = \frac{1}{n}\sum_{j=1}^n \underbrace{\left(\frac{X_j - \mu}{\sigma}\right)}_{Z_j},$$

and the Z_j are such that $\mathbb{E}\big[Z_j\big] = 0, \mathbb{V}\big[Z_j\big] = 1$.

Now, the CHF of $\tilde{M}_n = S_n/\sqrt{n}$ can be written as

$$\begin{aligned}
\varphi_{\tilde{M}_n}(\eta) &= \varphi_{\frac{S_n}{\sqrt{n}}}(\eta) \\
&= \mathbb{E}\big[e^{i\eta \sum_{j=1}^{X_j}/\sqrt{n}}\big] \\
&= \mathbb{E}\big[e^{i\eta X_j/\sqrt{n}}\big]^n && \text{(i.i.d)} \\
&= \varphi_{X_1}(\eta/\sqrt{n})^n \\
&= \left(1 + \frac{1}{2}\frac{\eta^2}{n}\cdot(-1) + o\Big(\frac{1}{n}\Big)\right) && \text{(Taylor + Lemma 3)} \\
&= \left(1 + \frac{-\frac{1}{2}\eta + n\cdot o\big(\frac{1}{n}\big)}{n}\right)^n \\
&\xrightarrow{n\to\infty} e^{-\frac{\eta^2}{2}},
\end{aligned}$$

which is the characteristic function of a $\mathcal{N}_{0,1}$ random variable. The second-order expansion of φ_{X_1} does not contain the first term since the $\mathbb{E}\big[Z_j\big] = 0$, and we use Lemma 3 for the variance.

□

Remark We can think of the CLT as telling us that for large enough n,

$$\frac{\sqrt{n}(M_n - \mu)}{\sigma} \sim \mathcal{N}_{0,1} \implies M_n \sim \mathcal{N}_{\mu,\frac{\sigma^2}{n}} \xrightarrow{d} \delta_\mu.$$

We had already computed the expected value and variance,.r the CLT also tells us the shape of the distribution. Moreover, since $S_n = n \cdot M_n$ we also know that the partial summations behave as a normal distribution,

$$S_n \sim \mathcal{N}_{n\mu, n\sigma^2},$$

which however *does not* weakly converge to any probability distribution.

Example (Bernoulli game)
We consider a Bernoulli sequence of random variables: let $(E_n)_{n\in\mathbb{N}}$ be a sequence of independent events, such that $\mathbb{P}\big(E_n\big) = p$ for all n. Set $X_n := \mathbb{1}_{E_n}$ and consider the sequence of partial sums $S_n = \sum_{j=1}^n X_j \sim \text{Binom}(n,p)$.

Since $\mu = \mathbb{E}\big[X_n\big] = p$ and $\sigma^2 = \mathbb{V}\big[X_n\big] = p(1-p)$, the CLT tells us that the empirical average is such that

$$\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}_{0,1},$$

and therefore $S_n \xrightarrow{d} \mathcal{N}_{np,np(1-p)}$, which is called the De Moivre-Laplace approximation.

Example

Let $(Y_n)_{n\in\mathbb{N}}$ be a random sample of a random variable X, which means $Y_n \overset{\text{i.i.d}}{\sim} X$. We fix a real number $x \in \mathbb{R}$ and we consider the *empirical cumulative distribution function* of X,

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(Y_i).$$

Intuitively we expect that $F_n \xrightarrow{n\to\infty} F_X$, which is actually a consequence of the CLT. By defining

$$X_j = \mathbb{1}_{(-\infty,x]}(Y_j),$$

then we find that

- X_j are independent (transformation of i.i.d r.v.)
- $\mathbb{E}\big[X_j\big] = \mathbb{P}\big(Y_j \le x\big) = \mathbb{P}\big(X \le x\big) = F(x)$.

Therefore, we have that

i. LLN $\implies F_n(x) \xrightarrow{\text{a.s.}} F_X(x)$

ii. CLT $\implies \sqrt{n}\big(F_n(x) - F_X(x)\big) \xrightarrow{d} \mathcal{N}_{0,F_X(x)(1-F_X(x))}$.

However, we can also prove a convergence result which is stronger than the pointwise convergence.

Theorem 13 (Glivenko-Cantelli)

With the assumptions defined above, the empirical cdf of X is such that

$$\sup_x \|F_n(x) - F_X(x)\| \xrightarrow{a.s.} 0.$$

Proof.

No.

□

Unfortunately, with the above theorem we don't have an estimate for the number of observations needed for an asymptotical normal behaviour. However we can state the following result, which holds for *any* random variable X:

Theorem 14 (Berry-Essen)

If X_n is such that $\mathbb{E}\big[|X_n|^3\big] < \infty$, then if $\Phi(\cdot)$ is the cdf of a $\mathcal{N}_{0,1}$ random variable we have that

$$\sup_x |F_{\tilde{M}_n}(x) - \Phi(x)| \le c\frac{\mathbb{E}\big[|X|^3\big]}{\sigma^3\sqrt{n}},$$

with $c \approx 0.79\ldots$

Proof.
No.

□

Remark The result holds for all possible choice of distributions, and although this convergence can be considered slow – $o(n^{-1/2})$ – we usually observe a faster convergence behaviour when using common distributions.

Example (Counterexample when $\mathbb{E}(X_n)$ is not defined)
Let $\mu_{X_n}(dx) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} dx$. If we were in a convergence situation, then we would expect $\mu_{M_n} \to \delta_0$. However, this random variable is such that

$$\begin{aligned}
\varphi_{M_n}(\eta) &\overset{\text{iid}}{=} \varphi_{X_1}\left(\frac{\eta}{n}\right)^n \\
&= e^{-\left|\frac{\eta}{n}\right| \cdot n} \qquad \text{(CHF of Cauchy distrib.)} \\
&= e^{-|\eta|} \\
&= \varphi_{X_1}(\eta).
\end{aligned}$$

Therefore, we see that $M_n \sim X_1$ for all n and thus it does not converge to 0. This is a consequence of the fact that X does not have a finite integral, $\mathbb{E}\big[|X_1|\big] = +\infty$.

We now state some useful generalizations of the central limit theorem, which extend its applicability to the non-identically distributed case.

Theorem 15 (Lyapunov's CLT)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of r.v. such that

i. $\mu_n = \mathbb{E}\big[X_n\big]$, $\sigma_n^2 = \mathbb{V}\big[X_n\big] < \infty$

ii. X_n *are independent*

iii. There exists $\delta > 0$ *such that*

$$\lim_{n\to\infty} \frac{1}{\vartheta^{2+\delta}} \sum_{j=1}^{n} \mathbb{E}\big[|X_j - \mu_j|^{2+\delta}\big] = 0,$$

where $\vartheta_n^2 = \sum_{j=1}^n \sigma_j^2$.

Then, we have that

$$\frac{1}{\vartheta_n} \sum_{j=1}^{n} (X_j - \mu_j) \xrightarrow{d} \mathcal{N}_{0,1}.$$

Proof.
No.

□

Theorem 16 (Lindeberg's CLT)

Same as Lyapunov's CLT but with the third condition replaced by

iii′. For all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\vartheta_n^2} \sum_{j=1}^{n} \mathbb{E}\Big[(X_j - \mu_j)^2 \mathbb{1}_{[\varepsilon\vartheta_n,\infty)}\big(|X_j - \mu_j|\big)\Big] = 0$$

Proof.
No.

□

Exercises

1. Prove that Lindeberg's CLT $\implies$ Lyapunov's CLT.
2. Starting from p. 176 of Gut (2009): Ex. 2, 19, 21, 24, 32.

Lecture 3: Simulations and independence

2021-10-28

3.1 Monte Carlo simulation

In this lecture we start by considering some applications of the convergence theorems we discussed earlier, in particular under the context of *Monte Carlo simulation*.

Consider a sequence of i.i.d random variables $X_1, X_2, \ldots, X_n, \ldots$ of a given r.v. X. For the sake of simplicity we will assume that $X \in L^2(\Omega, \mathbb{P})$, i.e.

$$\begin{cases} \mathbb{E}[X] = \mu < \infty \\ \mathbb{V}[X] = \sigma^2 < \infty \end{cases}$$

The goal of Monte Carlo simulation is to use the observed sample to approximate the expected value μ using the LLN and/or CLT. In particular, we will use the fact that by the LLN,

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{\text{a.s.}} \mu.$$

Remark If we consider f measurable and such that $f(X) \in L^2(\Omega, \mathbb{P})$, then the transformed sequence $f(X_1), f(X_2), \ldots, f(X_n)$ is a sample from $f(X)$. Therefore,

$$M_n^{(f)} = \frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow{\text{a.s.}} f(\mu).$$

Therefore, an interesting question to pose is the following one:

What is a good choice of n in order to obtain a good accuracy for the simulation?

We will try to answer this question by considering two approaches. Firstly, using the fact that $X \in L^2(\Omega, \mathbb{P})$, we can apply Chebyshev's inequality (corollary 1) and assert that for any fixed tolerance $\varepsilon > 0$:

$$\mathbb{P}\big(|M_n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2} \implies \mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

We then find the minimum number of observations $\bar{n} \in \mathbb{N}$ such that, for some specified probability p, we have $\mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \geq p$ for all $n \geq \bar{n}$:

$$1 - \frac{\sigma^2}{n\varepsilon^2} \geq p \iff n \geq \frac{\sigma^2}{\varepsilon^2(1-p)} = \bar{n}. \tag{2}$$

Remark We have that the limit $\bar{n} = \bar{n}(\sigma^2, \varepsilon, p)$ is a function of three quantities, of which σ^2 is not known and is usually estimated either from the sampled data or from previous simulations.

Moreover, from Equation (2) we notice that the minimum number of samples is such that

$$\bar{n}(\sigma^2, \varepsilon, p) \longrightarrow +\infty \quad \text{if either} \quad \begin{cases} \sigma^2 & \longrightarrow \infty \\ \varepsilon & \longrightarrow 0^+ \\ p & \longrightarrow 1^- \end{cases}$$

The convergence however is quite slow and can be refined in terms of p by using the Central Limit Theorem. If n is large enough, we know by the CLT that

$$M_n - \mu \sim \mathcal{N}_{0, \frac{\sigma^2}{n}},$$

therefore we compute the approximate coverage probability

$$\mathbb{P}\big(|M_n - \mu| < \varepsilon\big) \overset{n \gg 1}{\approx} \mathbb{P}\Big(\Big| \underbrace{\frac{(M_n - \mu)\sqrt{n}}{\sigma}}_{\xrightarrow{d} \mathcal{N}_{0,1}} \Big| < \frac{\sqrt{n}\varepsilon}{\sigma}\Big) \overset{\text{sym.}}{=} 2\left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - \frac{1}{2}\right) = 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1,$$

where the last equalities come from the symmetry of the Gaussian density function. Now we want to solve the inequality

$$\begin{aligned} 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1 \geq p &\overset{\text{monot.}}{\Longrightarrow} \frac{\sqrt{n}\varepsilon}{\sigma} \geq \Phi^{-1}\left(\frac{1+p}{2}\right) \\ &\iff n \geq \frac{\sigma^2}{\varepsilon^2} \cdot \left[\Phi^{-1}\left(\frac{1-p}{2}\right)\right]^2. \end{aligned}$$

What we claim is that the factor is sharper than the previous result $\frac{1}{1-p}$ in Equation (2), hence it is what is used in practice when computing the confidence interval.

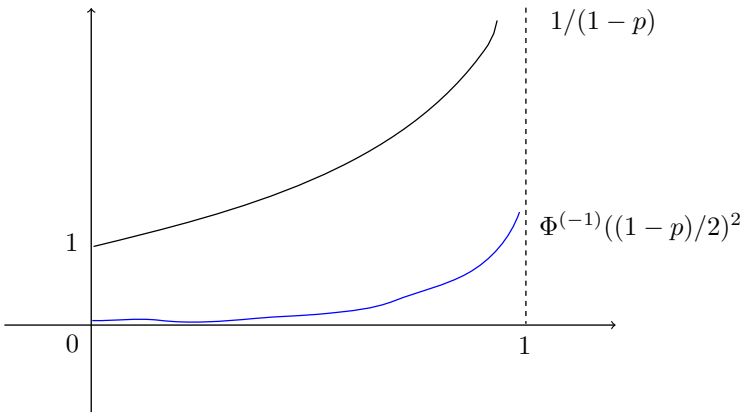

Figure 6: Sharpness of the bounds when using the two approximations.

3.2 Conditioning

Conditional probability and conditional random variables become an extremely hard topic when dealing with events that have probability zero, i.e. for continuous distribution and continuous-time stochastic processes.

Example (Dice roll)
Consider the rolling of two dice, we are interested in the outcome. We consider the sample space $\Omega = \{(i,j) : i,j = 1,\dots,6\}$. Since we have discrete events there is no problem in considering the σ-algebra given by the power set $\mathcal{F} = \mathscr{P}(\Omega)$. As for the probability measure on the measurable space we use the uniform probability $\mathbb{P} = \text{Unif}_\Omega$ on Ω:

$$\mathbb{P}\big(\{(i,j)\}\big) = \frac{1}{36}.$$

Define two variables X_1, X_2 such that X_j is the result of the j^{th} throw,

$$X_1(\omega) = X_1\big((\omega_1,\omega_2)\big) = \omega_1.$$

$$X_2(\omega) = X_2\big((\omega_1,\omega_2)\big) = \omega_2.$$

Consider the event $A =$ "the sum of the two die is smaller or equal than 6" and suppose that we win when this event occurs,

$$A = \{X_1 + X_2 \leq 6\} \implies Y = \mathbb{1}_A - \mathbb{1}_{A^c} \quad \text{is the expected win.}$$

Therefore, $\mathbb{E}[Y] = \mathbb{P}(A) - \mathbb{P}(A^c) = (15-21)/36 = -1/6$.

Assume now that the dice are instead thrown sequentially, i.e. we observe at $t = 1$ the outcome $X_1 = 5$. No one would think now that the chances of winning would be the same as before, so the observer should *update their belief* about their probabilities. Since now we can only win if the next throw is $X_2 = 1$, it's immediate to find that

$$\mathbb{P}\big(A|X_1 = 5\big) = \mathbb{P}\big(X_2 = 1\big) = \frac{1}{6}.$$

Remarks

› To calculate $\mathbb{P}(A|X_1 = 5)$ we assumed some sort of independence structure, i.e.

$$A \cap \{X_1 = 5\} = \{X_1 = 5\} \cap \{X_2 = 1\}.$$

› How do we update our belief if these random variables are not independent?

› What does it mean for two random variables to be independent in the first place?

To answer these questions we need a good definition of conditional probability, from which we will derive a notion of conditional expected value

$$\mathbb{E}[Y] \rightsquigarrow \mathbb{E}[Y|X_1 = 5] = \frac{1}{6} - \frac{5}{6}.$$

Def. (Conditional proability)
Let $A, B \in \mathcal{F}$ be events with $\mathbb{P}(B) > 0$, then we say that the ***conditional probability of A given B*** is given by

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

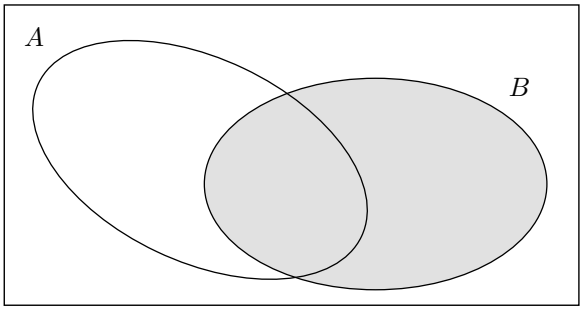

Figure 7: In some sense B takes place of the event space Ω when calculating the probability of the event $A|B$. We can interpret this by saying that the admissible σ-algebra for the new observation is *updated* upon observing the event B.

Remark Observe that the function that maps $A \mapsto \mathbb{P}\big(A|B\big) = \mathbb{P}_{|B}(A)$ *is a new probability measure* for any $Ain\mathcal{F}$, since it satisfies the Kolmogorov axioms:

$$\mathbb{P}_{|B}(\Omega) = 1$$

$$\mathbb{P}_{|B}(A^c) = 1 - \mathbb{P}_{|B}(A)$$

$$\mathbb{P}_{|B}\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \overset{\text{disj.}}{=} \sum_{n\in\mathbb{N}} \mathbb{P}_{|B}(A_n)$$

Def. (Conditional expectation)
For any $Y \in L^1(\Omega, \mathbb{P})$ we define the ***conditional expectation*** of Y given event B as the expected value w.r. to the newly-defined conditional probability measure,

$$\mathbb{E}\big[Y|B\big] = \mathbb{E}_{\mathbb{P}_{|B}}\big[Y\big]$$

Remark We can prove that $\mathbb{E}[Y|B] = \frac{1}{\mathbb{P}(B)}\mathbb{E}\big[Y \cdot \mathbb{1}_B\big]$, which yields a convenient way of calculating the conditional probability only by using the *a priori* probability measure $\mathbb{P}$.

In general, we can define any conditional quantity that we already defined for standard random variables, such as conditional variances, etc.

If $X \in L^2(\Omega, \mathbb{P})$ is a random vector, $X : \Omega \longrightarrow \mathbb{R}^n$ then its *covariance matrix* is

$$\mathrm{Cov}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

When conditioning w.r. to an event, we have the conditional covariance matrix

$$\mathrm{Cov}(X|B) = \mathbb{E}_{\mathbb{P}_{|B}}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right].$$

3.3 Independence

Since we have a satisfactory notion of conditional probability, by intuition we could define two events A and B to be independent if

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

Remark If $\mathbb{P}(A), \mathbb{P}(B) > 0$ then we have that by Bayes' formula,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

According to our intuitive definition, then we obtain that $\mathbb{P}(B) = \mathbb{P}(B|A)$. The main problem however is when $\mathbb{P}(B) = 0$, which is when the theory of probability diverges into different approaches.

3.3.1 Kolmogorov's approach

If we take for granted the definition of independence as $\mathbb{P}(A|B) = \mathbb{P}(A)$, then we obtain the following identity:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B).$$

Therefore, we can always go back in the other direction by using this as a definition of independence and re-discovering that $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Def. (Independence of events)
Two events $A, B \in \mathcal{F}$ are said to be ***independent events*** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Remark If $\mathbb{P}(A) = 0$ and/or $\mathbb{P}(B) = 0$, then we have that

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = 0 \overset{\text{moton.}}{=} \mathbb{P}(A \cap B).$$

Example
If $A \cap B = \emptyset$ with $\mathbb{P}(A) = 0$ then according to the definition that we gave this would mean that A and B are independent. However,

$$A \cap B = \emptyset \implies A, B \text{ ARE LOGICALLY DEPENDENT.}$$

With Kolmogorov's approach we can just say that we ignore these philosophical subtleties and work with events that are meaningful in practice.

Instead, in the approach of de Finetti we define $\mathbb{P}(A|B) = \mathbb{P}(A)$ and consider the logical coherence of the events, recovering as a *theorem* the relationship

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

With Kolmogorov's approach it is possible to define $\mathbb{P}(A|B)$ even for events that have probability zero.

Def. (Correlated events)
Two events $A, B \in \mathcal{F}$ are ***positively correlated*** if $\mathbb{P}(A|B) = \mathbb{P}(A)$, which can be seen to be true if and only if also $\mathbb{P}(B|A) > \mathbb{P}(B)$ (using Bayes' theorem).

Sometimes we observe one and only one event out of a set of events, i.e. we have a partition, and we would like to define how the probability measures get updated.

Example (Dice roll (cont.))
Consider all events $E_i = \{X_1 = i\}$, $i = 1, \dots, 6$, then the family of events

$$\mathcal{E} = (E_i)_{i=1,\dots,6}$$

is a partition and we can define the family of conditional measures given the partition of events whose conditional value is still unknown to us

$$\mathbb{P}(A|\mathcal{E}) = \sum_{i=1}^{6} \underbrace{\mathbb{P}(A|E_i)}_{\text{numbers}} \cdot \underbrace{\mathbb{1}_{E_i}}_{\text{r.v.}}$$

which is a *random measure*. Therefore, we can compute the conditional expected value given the partition $\mathcal{E}$

$$\mathbb{E}\big[Y|\mathcal{E}\big] = \sum_{i=1}^{6} \underbrace{\mathbb{E}\big[Y|E_i\big]}_{\text{numbers}} \cdot \underbrace{\mathbb{1}_{E_i}}_{\text{r.v.}}.$$

This type of structure is useful to model stochastic processes that evolve over time as the new partitions are observable and (eventually) observed.

Def. (Conditional probability w.r. to a partition)
Let $\mathcal{E}$ be a countable partition of events with positive probability,

- $\mathcal{E} = (E_n)_{n\in\mathbb{N}}, \quad \mathbb{P}(E_n) > 0$ for all $n \in \mathbb{N}$.
- $E_n \cap E_m = \emptyset$ for all $n \neq m$.
- $\bigcup_{n\in\mathbb{N}} E_n = \Omega$.

Given $A \in \mathcal{F}$, we define the ***conditional probability w.r. to the partition*** $\mathcal{E}$ as the random measure given by

$$\mathbb{P}(A|\mathcal{E}) = \sum_{n\in\mathbb{N}} \mathbb{P}(A|E_n) \cdot \mathbb{1}_{E_n}$$

Remark Consider the function $A \mapsto \mathbb{P}(A|\mathcal{E}) :=$, then this is a *random probability measure*, i.e. by letting A vary over all possible events we have a function

$$\mathbb{P}_{|\mathcal{E}} : \mathcal{F} \longrightarrow [0,1].$$

Def. (Conditional expectation w.r. to a partition)
For any $Y \in L^1(\Omega, \mathbb{P})$ we can define the ***conditional expectation given the partition*** as the expected value under the random probability measure $\mathbb{P}_{|\mathcal{E}}$,

$$\mathbb{E}\big[Y|\mathcal{E}\big] = \mathbb{E}_{\mathbb{P}_{|\mathcal{E}}}\big[Y\big] = \sum_{n\in\mathbb{N}} \mathbb{E}\big[Y|E_n\big] \cdot \mathbb{1}_{E_n}$$

This could be the end of the story, unless we also want to consider a) uncountable partitions and b) events with zero probabilities, which is the case for absolutely continuous probability measures and continuous-time stochastic processes.

Example (Dice rolls (cont. ii))
This time we consider two *continuous dice*, where the probability space is now $\Omega = [0,6] \times [0,6]$, $\mathcal{F} = \mathcal{B}$, $\mathbb{P} = \text{Unif}_\Omega = \text{Unif}_{[0,6]} \otimes \text{Unif}_{[0,6]}$. We consider the same variables,

$$X_1(\omega) = \omega_1$$

$$X_2(\omega) = \omega_2$$

$$A = \{X_1 + X_2 \leq 6\}$$

$$Y = \mathbb{1}_A - \mathbb{1}_{A^c}$$

Since we have a uniform distribution, $\mathbb{P}(A) = \frac{1}{2}$ and $\mathbb{E}[Y] = \frac{1}{2} - \frac{1}{2} = 0$. Let us now assume that we observe the event $\{X_1 = 5\}$, again we have the intuition to change our probabilities and expected value to

$$\mathbb{P}(A|\underbrace{X_1 = 5}_{\mathbb{P}(\cdot)=0}) = \mathbb{P}(X_2 \leq 1) = \frac{1}{6}.$$

$$\mathbb{E}\big[Y|X_1 = 5\big] = \frac{1}{6} - \frac{5}{6} = -\frac{2}{3}.$$

However both these quantities and the notion of independence are not defined by means of the previous definitions, since the conditioning event has probability zero.

In this case, $\mathcal{E} = (\{X_1 = k\})_{x\in[0,6]}$.

As it turns out, in order to obtain a formal definition of conditional probability we have to work the other way around: first by defining a good notion of $\mathbb{E}[Y|\mathcal{E}]$ and subsequently deduce a value for $\mathbb{P}(A|\mathcal{E})$.

Lecture 4: Conditional expectation

2021-11-04

Last lecture we considered a countable partition $\mathscr{P} = \{E_i : \bigcup_{i\in\mathbb{N}} E_i = \Omega, \mathbb{P}(E_i) > 0\}$ and we defined the conditional probability w.r. to $\mathscr{P}$:

$$\mathbb{P}(A|\mathscr{P}) = \sum_{i\in\mathbb{N}} \mathbb{P}(A|E_i)\cdot \mathbb{1}_{E_i}.$$

We discuss now the process of conditioning w.r. to more general (possibly uncountable) partitions and will start from the following observation.

Observation Consider a random variable $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then we have that the random variable defined by the expected value w.r. to the countable partition $\mathscr{P}$,

$$\mathbb{E}[X|\mathscr{P}] = \sum_{i\in\mathbb{N}} \mathbb{E}[X|E_i]\cdot \mathbb{1}_{E_i},$$

satisfies the following properties:

i. This r.v. is $\sigma(\mathscr{P})$-measurable (i.e. it is observable) since $\mathscr{P}$ is a countable partition and therefore we simply have that

$$\sigma(\mathscr{P}) = \{\text{all possible unions of elements of } \mathscr{P}\}.$$

ii. For any $A \in \sigma(\mathscr{P})$, we have that

$$\mathbb{E}[X|A] = \mathbb{E}\big[\underbrace{\mathbb{E}[X|\mathscr{P}]}_{\text{r.v.}}|A\big],$$

Proof.

$A \in \sigma(\mathscr{P}) \iff A = \bigcup_{j\in J} E_j$ with J countable, therefore we can write the following chain of equations

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X|\mathscr{P}]\big|A\big] &= \frac{1}{\mathbb{P}(A)}\int_A \mathbb{E}[X|\mathscr{P}]\,\mathrm{d}\mathbb{P} \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\big[\sum_{i\in\mathbb{N}} \mathbb{E}[X|E_i]\cdot \overbrace{\mathbb{1}_{E_i}\cdot\mathbb{1}_A}^{\mathbb{1}_{E_i\cap A}}\big] \qquad (A \text{ is union of } E_j's) \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\big[\sum_{j\in J} \overbrace{\mathbb{E}[X|E_j]}^{\text{constant}}\cdot\mathbb{1}_{E_j}\big] \\
&= \frac{1}{\mathbb{P}(A)}\cdot\sum_{j\in J}\frac{1}{\cancel{\mathbb{P}(E_j)}}\cdot\mathbb{E}[X|E_j]\cdot\cancel{\mathbb{P}(E_j)} \\
&= \frac{1}{\mathbb{P}(A)}\mathbb{E}\Big[\sum_{j\in J} X\cdot\mathbb{1}_{E_j}\Big] \\
&= \mathbb{E}[X|A].
\end{aligned}$$

□

4.1 General case

We start from defining the conditional expectation of X and work our way up to the definition of conditional probability. Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra of events.

Def. (Version of the conditional expectation of X given $\mathcal{G}$)
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, we say that a random variable Z is a ***version of the conditional expectation of X given $\mathcal{G}$*** if Z satisfies the following properties:

i. Z is $\mathcal{G}$-measurable

ii. For any $A \in \mathcal{G}$ such that $P(A) > 0$,

$$\mathbb{E}[X|A] = \mathbb{E}[Z|A] \iff \frac{1}{\mathbb{P}(A)}\mathbb{E}[X \cdot \mathbb{1}_A] = \frac{1}{\mathbb{P}(A)}\mathbb{E}[Z \cdot \mathbb{1}_A]$$

We denote by $\mathbb{E}[X|\mathcal{G}]$ the set of all such r.v.'s.

Remarks

› Sometimes Z is unique, but in general there might be equivalent random variables up to a set of measure zero, therefore $\mathbb{E}[X|A]$ defines an ***equivalence class***.

› Given $Z \in \mathbb{E}[X|\mathcal{G}]$ and $Z' \overset{\text{a.s.}}{=} Z$ it is not sufficient to guarantee that $Z' \in \mathbb{E}[X|\mathcal{G}]$ since Z' might not be measurable w.r. to $\mathcal{G}$. For the simplest example of such Z', consider if $\mathcal{G}^c$ is a set of measure zero and

$$Z' = \begin{cases} Z & \text{if } \omega \in \mathcal{G} \\ 1 & \text{if } \omega \in \mathcal{G}^c \end{cases}$$

Clearly, Z' is not measurable w.r. to $\mathcal{G}$ since $Z' = 1$ if $\omega \in \mathcal{G}^c$. However, since $\mathcal{G}^c$ has measure zero, we also have $Z \overset{\text{a.s.}}{=} Z'$.

Prop. 2 (Almost sure equality of versions)
If $Z, Z' \in \mathbb{E}[X|\mathcal{G}]$, then $Z' \overset{\text{a.s.}}{=} Z$.

Proof.
Consider for simplicity the case $X \in \mathbb{R}$, then we are going to show that for any choice of r.v.'s X, X' and $Z \in \mathbb{E}[X|\mathcal{G}]$ and $Z' \in \mathbb{E}[X'|\mathcal{G}]$, we have

$$X \overset{\text{a.s.}}{\leq} X' \implies Z \overset{\text{a.s.}}{\leq} Z'.$$

Then we will choose $X' = X$ and since the reverse holds because of symmetry we obtain a double inequality which implies strict equality.

By contradiction, assume instead that $\mathbb{P}(Z > Z') > 0$, which is the same of saying that $Z \not\overset{\text{a.s.}}{\leq} Z'$. Then, we would have that

$$\begin{aligned} 0 \overset{!!}{<} \mathbb{E}[(Z - Z') \cdot \mathbb{1}_{Z>Z'}] &= \mathbb{E}[Z \cdot \mathbb{1}_{Z>Z'}] - \mathbb{E}[Z' \cdot \mathbb{1}_{Z>Z'}] && (\{Z > Z'\} \in \mathcal{G}) \\ &= \mathbb{E}[X \cdot \mathbb{1}_{Z>Z'}] - \mathbb{E}[X' \cdot \mathbb{1}_{Z>Z'}] && (\text{def}) \\ &= \mathbb{E}[(X - X') \cdot \mathbb{1}_{Z>Z'}] \leq 0. && (X \overset{\text{a.s.}}{\leq} X' \text{ by Hp.}) \end{aligned}$$

Since we had assumed that $X \overset{\text{a.s.}}{\leq} X'$ we find a contradiction, and therefore we conclude that $\mathbb{P}(Z > Z') = 0$.

□

Corollary 2 (Conditional r.v.'s are equivalence classes)

The set of $\mathbb{E}[X|\mathcal{G}]$ is an equivalence class on the set of $\mathcal{G}$-measurable random variables w.r. to the "$\overset{a.s.}{=}$" operator.

Remarks

- This is important to remember, since in all textbooks and articles the equivalence class $\mathbb{E}[X|\mathcal{G}]$ is treated as a single random variable. All equalities and inequalities are interpreted as valid for a chosen single representative of the equivalence class.
- Any other Z that is $\mathcal{G}$-measurable has to be constant over the events E_i of the countable partition (otherwise Z^{-1} would not be a union of E_i's). If Z was a version of $\mathbb{E}[X|\mathcal{G}]$, then it would be constantly equal to $\mathbb{E}[X|E_i]$ on E_i, which is exactly the expected value defined by cases $\implies \mathbb{E}[X|\mathscr{P}]$ is unique.

Example (Trivial conditioning)

Consider $\mathcal{G} = \sigma(X)$, then clearly (i) and (ii) are trivially satisfied by taking $Z = X$. With an *abuse* of notation, we are going to write $X = \mathbb{E}[X|\mathcal{G}]$.

The following theorem guarantees existence of $\mathbb{E}[X|\mathcal{G}]$ for a particular subset of random variables X.

Theorem 17

Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then for any sub-$\sigma$-algebra $\mathcal{G}$ it holds that $\mathbb{E}[X|\mathcal{G}]$ is not empty.

Proof.

The proof is based on the Radon-Nikodym theorem between dominated probability measures.

□

Theorem 18 (Properties of $\mathbb{E}(X|\mathcal{G})$)
Let $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subset \mathcal{F}$. Then, the following properties hold:

1. *(**Linearity**) For any $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}[\alpha X + \beta Y|\mathcal{G}] = \alpha\mathbb{E}[X|\mathcal{G}] + \beta\mathbb{E}[Y|\mathcal{G}]$.*
2. *(**Monotonicity**) If $X \leq Y$ a.s. then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$.*
3. *If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X|\mathcal{G}] = X$.*
4. *If $\sigma(X) \perp \mathcal{G}$ then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.*
5. *(**Tower property**) If $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$.*
6. *If Y is $\mathcal{G}$-measurable and bounded, then $\mathbb{E}[Y \cdot X|\mathcal{G}] = Y \cdot \mathbb{E}[X|\mathcal{G}]$.*
7. *If Y is independent of X and $\mathcal{G}$, then $\mathbb{E}[Y \cdot X|\mathcal{G}] = \mathbb{E}[Y] \cdot \mathbb{E}[X|\mathcal{G}]$.*

Proof.

4 : Let $Z = \mathbb{E}[X]$, then clearly (i) is satisfied since a constant random variable is measurable w.r. to any σ-algebra. As for (ii), for any event $A \in \mathcal{G}$ such that $\mathbb{P}(A) > 0$ we can write

$$\begin{aligned}
\mathbb{E}[Z|A] &= \mathbb{E}\big[\mathbb{E}[X]|A\big] \\
&= \frac{1}{\mathbb{P}(A)} \cdot \mathbb{E}\big[\mathbb{E}[X] \cdot \mathbb{1}_A\big] \\
&= \frac{1}{\mathbb{P}(A)} \cdot \mathbb{E}[X] \cdot \mathbb{E}[\mathbb{1}_A] && (\mathbb{E}[X] \text{ is a constant}) \\
&= \frac{1}{\mathbb{P}(A)} \mathbb{E}[X \cdot \mathbb{1}_A] && (\text{indep.}) \\
&= \mathbb{E}[X|A]
\end{aligned}$$

5 : Let $Z \in \mathbb{E}[X|\mathcal{H}]$ and $Y \in \mathbb{E}[X|\mathcal{G}]$ with $\mathcal{H} \subset \mathcal{G}$, we want to prove that $Z \in \mathbb{E}[Y|\mathcal{H}]$. (i) is satisfied since $Z \in \mathbb{E}[X|\mathcal{H}]$ is $\mathcal{H}$-measurable by definition of version of $\mathbb{E}[X|\mathcal{H}]$. As for (ii), for any $A \in \mathcal{H}$ such that $\mathbb{P}(A) > 0$, we have that since $\mathcal{H} \subset \mathcal{G}$,

$$\mathbb{E}[Z|A] \overset{\text{def.}}{=} \mathbb{E}[X|A] \overset{A \in \mathcal{G}}{=} \mathbb{E}[Y|A].$$

6 : $(ii) \iff \mathbb{E}[X \cdot W] = \mathbb{E}[Z \cdot W]$ for any r.v. W that is $\mathcal{G}$-measurable and bounded.

□

Remarks

- $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ families of events are said to be independent if $A \perp B$ for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$.
- We say that a r.v. Y is independent of a r.v. X and a σ-algebra $\mathcal{G}$ if $\sigma(Y) \perp \sigma\big(\sigma(X) \cup \mathcal{G}\big)$.
- Property (5) means that reducing the information from $\mathcal{F} \to \mathcal{H}$ can be done by reducing multiple times, $\mathcal{F} \to \mathcal{G} \to \mathcal{H}$.

Example (Conditioning w.r. to a random variable)
In the particular case of $\mathcal{G} = \sigma(Y)$, then we can define the following random variable,

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$

Lemma 4 (Doob's theorem)
Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and Y r.v. that takes values in another measurable space, in this case assume $(\mathbb{R}^n, \mathcal{B})$. Then X is $\sigma(Y)$-measurable $\iff$ there exists a (possibly not unique) measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$X = \varphi(Y).$$

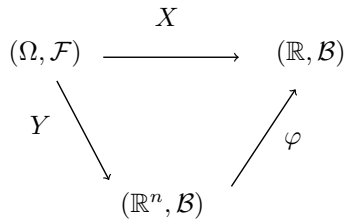

Figure 8: Schematization of Doob's theorem.

Corollary 3 (Existence of the regression function)
*There exists a measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^d$, which we call **regression function**, such that*

$$\varphi(Y) = \mathbb{E}[X|Y].$$

Proof.
We have that if the event $\{Y = y\}$ has non-negligible probability, then

$$\mathbb{E}[X|(Y=y)] \overset{\text{def}}{=} \mathbb{E}\big[\underbrace{\mathbb{E}[X|Y]}_{\text{Doob}\to\varphi(Y)} \big|(Y=y)\big] = \mathbb{E}[\varphi(Y)|(Y=y)] = \mathbb{E}[\varphi(y)] = \varphi(y).$$

□

Remarks

› φ might not be unique because (i) Doob's theorem does not guarantee unicity and (ii) it belongs to the equivalence class of conditional expectations.

› φ is however almost-surely unique w.r. to the law μ_Y of Y.

Notation The function φ is denoted as $\varphi(Y) = \mathbb{E}[X|Y = y]$, which is <u>not</u> the conditional expectation given the event $\{Y = y\}$ unless this set has positive probability.

Example
Consider the particular case in which $X = \mathbb{1}_{X\in H}$ for a Borel set $H \in \mathcal{B}$. Then, we can consider the following expected value

$$\mathbb{E}[\mathbb{1}_{X\in H}|\mathcal{G}] =: \mu_{X|\mathcal{G}}(H),$$

which we call the *conditional law of X given $\mathcal{G}$* calculated in the set H.

We can check that it is a probability measure

- If $H = \mathbb{R}^d$, we obtain $\mathbb{E}[\mathbb{1}_{X\in\mathbb{R}^d}|\mathcal{G}] = 1$.
- If $H = \emptyset$, then $\mathbb{E}[\mathbb{1}_{X\in\emptyset}|\mathcal{G}] = 0$.
- Given a sequence $(H_n)_{n\in\mathbb{N}}$ of disjoint sets, we have

$$\mathbb{E}[\mathbb{1}_{X\in\bigcup_{n\in\mathbb{N}} H_n}|\mathcal{G}] = \mathbb{E}\Big[\sum_{n\in\mathbb{N}} \mathbb{1}_{X\in H_n}|\mathcal{G}\Big] = \sum_{n\in\mathbb{N}} \mathbb{E}[\mathbb{1}_{X\in H_n}|\mathcal{G}].$$

The problem is that this holds in the almost-sure sense, i.e. everything is defined in terms of a representative of the equivalence class.

Theorem 19 (Regular conditional law)
Given $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $G \subset \mathcal{F}$ there always exist a family $\big(\mu_{X|\mathcal{G}}(\omega)\big)_{\omega\in\Omega}$ of probability measures on $\mathbb{R}^d$ such that for any Borel set H,

$$\mu_{X|\mathcal{G}}(H) = \mathbb{E}[\mathbb{1}_{X\in H}|\mathcal{G}].$$

*Such family is called the **regular version of conditional law of X given $\mathcal{G}$**.*

Proof.
No.

□

Using this definition of regular conditional law, there are many results that we can compute that will give the usual known results.

Theorem 20 (Conditional expectation)
If $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ and $f(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then we can write

$$\mathbb{E}[f(X)|\mathcal{G}] = \int_{\mathbb{R}^d} f(x)\mu_{X|\mathcal{G}}(dx).$$

Example (Conditional expected value)

If $f = \text{id}$, then this becomes the *conditional expected value of X given $\mathcal{G}$*,

$$\mathbb{E}[X|\mathcal{G}] = \int_{\mathbb{R}^d} x\mu_{X|\mathcal{G}}(dx)$$

Theorem 21 (Expectation of the conditional measure)
Following from the tower property of the conditional expectation, we have that

$$\mu_X(H) = \mathbb{E}[\mu_{X|\mathcal{G}}(H)]$$

Example
Choosing $\mathcal{G} = \sigma(Y)$ we obtain the conditional law of X given Y as $\mu_{X|Y} := \mu_{X|\sigma(Y)}$.

Theorem 22 (Joint distribution of two random variables)
Let X, Y be r.v. with values on $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively. Then, we have that for each $H \in \mathcal{B}(R^d)$ and $K \in \mathcal{B}(R^n)$,

$$\mu_{(X,Y)}(H \times K) = \mathbb{E}[\mu_{X|Y}(H) \cdot \mathbb{1}_{Y \in K}].$$

Remark This property can be used to prove that a joint random variable (X, Y) is absolutely continuous w.r. to Lebesgue measure, when starting from the conditional distribution $X|Y$ and the marginal distribution of Y, as long as we can apply Fubini's theorem.

Theorem 23 (Conditional law as a function of Y)
If X, Y are r.v.'s on $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively, then there exists a family $(\mu_{X|Y=y})_{y \in \mathbb{R}^n}$ of probability measures on $\mathbb{R}^d$ such that

i. For each $H \in \mathcal{B}$, the function $y \mapsto \mu_{X|Y=y}(H)$ is measurable.

ii. $(\mu_{X|Y=y})\big|_{y=Y} = \mu_{X|Y}$

Notation Sometimes we can find this written as $\mathbb{E}[\mathbb{1}_{X \in H}|Y = y]$ and the conditional probability function coincides with this quantity if $\mathbb{P}(Y = y) > 0$.

Theorem 24 (Conditional expected value of a function)
Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $f(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then the regression function

$$\mathbb{E}[f(X)|Y = y] = \int_{\mathbb{R}^d} f(x)\mu_{X|Y=y}(dx).$$

The last thing to study is what happens when X and Y have an absolutely continuous joint probability distribution, i.e. they admit a joint density.

If (X, Y) are jointly absolutely continuous, then X and Y are also absolutely continuous and

$$\gamma_Y(y) = \int_{\mathbb{R}^d} \gamma_{(X,Y)}(x, y)\, \mathrm{d}x.$$

Theorem 25 (Conditional density of two jointly a.c. random variables)
If (X, Y) are jointly absolutely continuous, then for any $y \in \mathbb{R}^n$ such that $\gamma_Y(y) > 0$ the conditional law function $\mu_{X|Y=y}$ is absolutely continuous with density given by

$$\gamma_{X|Y=y}(x) = \frac{\gamma_{(X,Y)}(x, y)}{\gamma_Y(y)}.$$

Remark For any Borel set H, $\mu_{X|Y=y}(H) = \int_H \gamma_{X|Y=y}(\mathrm{d}x)$

Corollary 4
For two jointly absolutely continuous random variables, we have that

$$\mathbb{E}[f(X)|Y = y] = \int_{\mathbb{R}^d} f(x)\gamma_{X|Y=y}(x)\, \mathrm{d}x,$$

whereas the unconditional expected value is

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} f(x)\gamma_{(X|Y=y)}(x)\gamma_Y(y)\, \mathrm{d}y\, \mathrm{d}x.$$

Exercises Given in the notes.

Lecture 5: Introduction to stochastic processes

2021-11-11

When we think about random variables, we can visualize them as random numbers which are determined by the result ω of some experiment. On the other hand, sometimes events unfold in time and the random number is replaced by a sequence of random variables $X_t(\omega)$ indexed by time t.

We replace the concept of random numbers with the concept of random trajectories, which can be thought as a random countable sequence (discrete time) or a random function (continuous time). Stochastic processes can be indeed be seen equivalently as

- A sequence of random variables.
- A random variable that takes values in the space of sequences.

Def. (Discrete time stochastic process)
A ***discrete time stochastic process*** is a family of random variables $X = (X_n)_{n\in I}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with $I \subseteq \mathbb{N}$.

Equivalent definition We could also equivalently define a stochastic process as a random variable on the space of sequences of dimension $|I|$, $X : \Omega \to (\mathbb{R}^d)^I$, where $(\mathbb{R}^d)^I = \{(x_n)_{n\in I} : x_i \in \mathbb{R}^d\}$.

Notation $(\mathbb{R}^d)^I$ is called the *trajectory space*. If $x \in (\mathbb{R}^d)^I$, then x is called a *trajectory*.

Example (Finite time stochastic process)
If $I = \{1, \dots, N\}$ then the set of trajectories (i.e. the *codomain* of the stochastic process) are

$$(\mathbb{R}^d)^I = \{(x_1, x_2, \dots, x_N) : x_i \in \mathbb{R}^d\}.$$

However we can think of X as a family $(X_1, X_2, \dots, X_N)$ such that X_i is a random variable with values on $\mathbb{R}^d$.

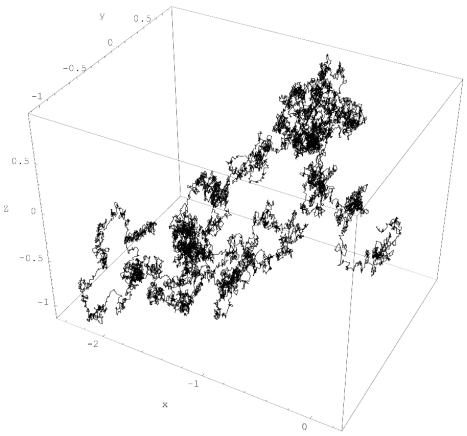

Figure 9: Trajectory of a stochastic process (Wiener process), where every random variable is a three-dimensional Gaussian distribution.

Although the definition above is intuitively correct, from a mathematical point of view we need to formalize the *measurability* of the second definition of the process. To do this, we need to equip the trajectory space $(\mathbb{R}^d)^I$ with a suitable σ-algebra such that the two definitions are equivalent. Such a σ-algebra (omitting proof) is the product σ-algebra generated by the rectangles and denoted by $\mathcal{B}^I$, which if $I \subseteq \mathbb{N}$ is finite then becomes the usual one,

$$\mathcal{B}^I = \bigotimes_{i \in I} \mathcal{B}_d, \qquad \text{if } |I| < \infty.$$

Example
If every X_n is defined on $\mathbb{R}$ and $I = \{1, 2, 3\}$, then the product σ-algebra is

$$\mathcal{B}^I = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^3).$$

Example (Finite coin tosses)
Consider for every $N \in \mathbb{N}$ the sample space $\Omega = \{0,1\}^N = \big\{(\omega_0, \omega_1, \omega_2, \dots, \omega_N), \omega_i \in \{0,1\}\big\}$. We choose the σ-algebra $\mathcal{F} = \big(\mathscr{P}(\{0,1\})\big)^N$ and we consider the product probability given by $\mathbb{P} = \bigotimes_{i=1}^N \text{Unif}_{\{0,1\}}$, which corresponds to a fair coin toss.

Remark Note that if the sequences were infinite, then $\mathcal{F} = \big(\mathscr{P}(\{0,1\})\big)^{\mathbb{N}}$ would not be a σ-algebra anymore, since this power set is uncountable and isomorphic to the power set of $[0,1]$, $\mathscr{P}([0,1])$. Therefore, one needs to consider a Lebesgue-like measure for infinite sequences.

As an example of a stochastic process, we could for instance define the following random variables,

$$X_n := \begin{cases} 1 & \text{if } w_n = 0 \\ -1 & \text{if } w_n = 1 \end{cases}$$

and we have that the sequence $X = (X_n)_{n \in \{1,\dots,N\}}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$.

If we define now $Y_n(\omega) := \sum_{i=1}^n X_i(\omega)$, this is again a random variable and thus the sequence $Y = (Y_n)_{n \in \{1,\dots,N\}}$ is again a stochastic process. Whereas X_n only depends on the n-th component of the outcome ω, Y_n depends instead on the events $(\omega_1, \dots, \omega_n)$ up to time n.

On the other hand, if we define now $Z_n := \sum_{i=n}^N X_i$, then $Z = (Z_n)_{n \in N}$ is a stochastic process that depends on $(\omega_n, \omega_{n+1}, \dots, \omega_N)$, which are the events from time n to time N. This means that at time n, Z_n ***cannot be observed***.

5.1 Filtrations

In the above stochastic model, there is a clear understanding of the type of events that we can observe at time n:

› Can we observe the event $\{\overbrace{(1,1,\dots,1)}^{N}\}$? Not at time $n < N$.

› However, at time n we can observe $\{Y_n = 5\}$, since Y depends only on ω_i for $i \leq n$.

In this case, if we define the following σ-algebra,

$$\mathcal{F}_n := \{\text{events observable up to time } n\},$$

then we conclude that the stochastic process Z is clearly different from both X and Y, in that

› X_n, Y_n are $\mathcal{F}_n$-measurable.

› Z_n is not $\mathcal{F}_n$-measurable.

We formalize this fundamental concept by the following definition, which serves as a basis for determining the measurability of a stochastic process.

Def. (Filtration)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, we say that a family $(\mathcal{F}_n)_{n\in I}$ of σ-algebras is a ***filtration*** if $(\mathcal{F}_n)_{n\in I}$ is such that

$$\mathcal{F}_n \subset \mathcal{F}_{n'} \subset \mathcal{F} \quad \text{for all } n < n'.$$

Remark $\mathcal{F}_n$ is an increasing family of σ-algebras, i.e. $\mathcal{F}_n \nearrow$, and this is useful to keep track of the *evolution of information* given by all the events which have been observed at time n.

Example (Previous)
In the previous example, $\{X_2 = 1\} \in \mathcal{F}_2$ but also $\{X_1 = 1\} \in \mathcal{F}_2$.

Def. (Adaptability)
A stochastic process $X = (X_n)_{n\in I}$ is ***adapted*** w.r. to a filtration $(\mathcal{F}_n)_{n\in I}$ if X_n is a $\mathcal{F}_n$-measurable function for any $n \in I$.

Example (Previous)
In the previous example, we have that the stochastic processes

› X and Y are adapted to $(\mathcal{F}_n)_{n\in I}$.

› Z is not adapted to $(\mathcal{F}_n)_{n\in I}$.

We now take another approach to the description of stochastic processes: we start by observing the values of the stochastic process and wonder which events are observable based on them.

Example (Finite dice rolls)
Let's consider a sample space $\Omega = \{1, \dots, 6\}^N = \big\{(\omega_1, \dots, \omega_N), \omega_i \in \{1, \dots, 6\}\big\}$, for $N \in \mathbb{N}$. Take again $\mathcal{F} = \mathscr{P}(\Omega)$ and the product probability $\mathbb{P}$.

Let's consider the stochastic process

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega_n \text{ is odd} \\ -1 & \text{if } \omega_n \text{ is even} \end{cases}$$

We now set $\mathcal{F}_n := \{\text{all set observable by observing the first } n \text{ rolls}\}$, from which X_n is clearly $\mathcal{F}_n$-measurable and X is adapted.

Another point of view Assume now that we can observe the random variables X_n but not the outcome of the experiment ω, i.e. we cannot observe $\mathcal{F}_n$.

The question now becomes: *Which events are observable once we observe X_n?* The set of events that we can observe by observing X_n turns out to be a filtration (no proof) and is the minimal class of events $(\mathcal{F}_n^X)_{n\in I}$ that make the process X adapted to it.

Def. (Natural filtration)
Given X stochastic process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we call the ***natural filtration*** of X the family $(\mathcal{F}_n^X)_{n\in I}$ given by

$$\mathcal{F}_n^X := \sigma\big(X_i^{-1}(A) : i \le n, A \in \mathcal{F}\big)$$

Remark $\mathcal{F}_n^X$ is the smallest filtration that makes the process X adapted and contains only the information related to the process itself.

5.2 Distribution of a stochastic process

We now turn to uniqueness of the stochastic processes, for which we have two versions (similar to random variables):

- *Strong* uniqueness (almost-sure equality)
- *Weak* uniqueness (equality in distribution)

Example (Why $\overset{\mathbf{d}}{=}$ is a tricky notion)
Consider $X \sim \mathcal{N}_{0,1}$ and $-X \sim \mathcal{N}_{0,1}$, then $X \overset{d}{=} -X$ but $X \overset{\text{a.s.}}{\neq} -X$.

Def. (Indistinguishable stochastic processes)
Let X, Y be stochastic processes. We say that X and Y are ***indistinguishable***, and we denote it by $X \overset{\text{a.s.}}{=} Y$, if

$$\mathbb{P}(X = Y) = \mathbb{P}\big(\{\omega \in \Omega : X_n(\omega) = Y_n(\omega)\ \forall n\}\big) = 1.$$

Now, to give an appropriate definition of $\overset{\text{d}}{=}$ for stochastic processes, we need to define what is the *law of a stochastic process*, which is something that becomes very technical. We instead use a shortcut which can be proven to be *equivalent* to the more technical construction.

Def. (Equality in distribution)
Let X, Y be stochastic processes, we say that X and Y are ***equal in law*** (or ***in distribution***), $X \overset{\text{d}}{=} Y$, if the marginal distribution of any finite collection of variables are equal, i.e. if for any finite choice of indices $n_1, n_2, \dots, n_k \in I$,

$$\mu_{(X_{n_1}, X_{n_2}, \dots, X_{n_k})} = \mu_{(Y_{n_1}, Y_{n_2}, \dots, Y_{n_k})},$$

which is equivalent to saying that for any such choice of indices,

$$(X_{n_1}, \dots, X_{n_k}) \overset{\text{d}}{=} (Y_{n_1}, \dots, Y_{n_k}).$$

Remark Requiring the processes to have the same law only at each time t is a very weak condition which is not enough to be a good definition of $\overset{\text{d}}{=}$. Indeed, $X_n \sim Y_n$ for all $n \not\Longrightarrow X \overset{\text{d}}{=} Y$ as we would intuitively mean it.

Example (Same marginal at each n but different law)
Let $X_n \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1 = \text{Unif}_{\{-1,1\}}$ for any n be a stochastic process that represents a balanced coin toss. Define now another stochastic process by

$$X'_n := X_1 \quad \forall n,$$

then we have that for any n, $X_n \sim X_1 \sim X'_n$. However, when seen as a whole process the two laws are completely different: if we consider the event of heads followed by tails, $H := \{(1, -1)\}$, then since $X'_n = X_1$ for all n it follows that

$$\mu_{(X_1, X_2)}(H) = \mathbb{P}(X_1 = 1, X_2 = -1) = \frac{1}{4},$$

$$\mu_{(X'_1, X'_2)}(H) = \mathbb{P}(X_1 = 1, X_1 = -1) = 0.$$

At each time the marginal distribution is the same, but if seen as a whole trajectory then the laws of the two stochastic processes are completely different.

Lecture 6: Martingales and Markov processes

2021-11-18

We introduce two important classes of stochastic processes which can be extended to the continuous time case.

6.1 Martingales

References Bass (2011, §3)

Martingales were well-known stochastic processes in economics which over the last decades became crucial in the theory of stochastic integration, from which we can construct continuous Markov processes (*diffusions*).

Def. (Discrete-time martingale)
A discrete stochastic process $X = (X_n)_n$ is called a ***martingale*** w.r. to a given filtration $(\mathcal{F}_n)_n$ if

i. $X_n \in L^1(\Omega, \mathbb{P})$ for all n.

ii. $\mathbb{E}[X_N|\mathcal{F}_n] = X_n$ for all $n \leq N$.

Adaptability There is no need to specify that X has to be adapted to $(\mathcal{F}_n)_n$, since $X_n = \mathbb{E}[X_N|\mathcal{F}_n]$ implies measurability w.r. to $\mathcal{F}_n$.

Expected value The second equality is a very strong property which tells us that if we condition the future process on the information at time n, then the expected value is equal to the value that we have observed. Using the tower property, we have that $\mathbb{E}[X_N] = \mathbb{E}[\mathbb{E}[X_N|\mathcal{F}_n]] = \mathbb{E}[X_n]$, therefore the expectation is *a priori* constant in time.

Example (Just $\mathbb{E}(X) = \mu$ is not enough)
Let $(X_n)_n$ be a family of independent random variables with $\mathbb{E}[X_n] = \mu$ for all n, and consider the natural filtration $(\mathcal{F}_n^X)_n$. The process $X = (X_n)_n$ is not a martingale for all possible distributions of X_n, since

$$\mathbb{E}[X_N|\mathcal{F}_n^X] \overset{\perp\!\!\!\perp}{=} \mathbb{E}[X_N] = \mu.$$

Therefore, this process is a martingale $\iff$ $\mathbb{E}[X_N] = \mu = X_n$ for all $n \leq N$, which is satisfied $\iff$ $X_n \equiv \mu$ almost surely.

Remark From the example above, independence is *orthogonal* to martingality, unless we choose a degenerate distribution $X_n \equiv \mu$.

Example (Martingale from independent variables)
Let us consider the process defined in the previous example, and define the stochastic process $Y_n = \sum_{k=1}^n X_k$. Clearly, Y_{n+1} and Y_n are marginally not independent, therefore the process

could be a martingale. Indeed, we have that

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n^X] = \mathbb{E}[Y_n + X_{n+1}|\mathcal{F}_n^X] = \underbrace{\mathbb{E}[Y_n|\mathcal{F}_n^X]}_{=Y_n} + \underbrace{\mathbb{E}[X_{n+1}|\mathcal{F}_n^X]}_{=\mathbb{E}[X_{n+1}]} = Y_n + \mu.$$

Therefore, we have that Y_n is a martingale $\iff \mu = 0$.

What can we say now about a martingale which is not defined w.r. to the filtration $\mathcal{F}_n^X$ but to a different filtration? For instance, what happens to the martingale property when enlarging to a bigger filtration?

Example (Adding events breaks martingality)
Let $X = (X_n)_n$ be a martingale w.r. to a filtration $(\mathcal{F}_n)_n$, and consider now a new filtration equal to all possible events $\mathcal{F}$ at all times, $(\mathcal{G}_n)_n = \mathcal{F}$. We now have that X is a martingale w.r. to $\mathcal{G}_n$ if

$$X_n = \mathbb{E}[X_N|\mathcal{G}_n] = \mathbb{E}[X_N|\mathcal{F}] = X_N,$$

therefore this means that X can again only be a constant process $X_n = \mu$ for all n.

In general When adding events we can't immediately conclude that the process is still a martingale.

Prop. 3 (Removing events does not break martingality)
Let $(X_n)_n$ be a martingale w.r. to a filtration $(\mathcal{F}_n)_n$. Let now $(\mathcal{G}_n)_n$ be another filtration such that

a) X is adapted to $(\mathcal{G}_n)_n$.

b) $\mathcal{G}_n \subset \mathcal{F}_n$ is a sub-filtration at all times.

Then, X is a martingale w.r. to $(\mathcal{G}_n)_n$.

Proof.
We use the tower property to prove the result, indeed since $\mathcal{G}_n \subset \mathcal{F}_n$ we can write

$$\mathbb{E}[X_N|\mathcal{G}_n] \overset{(b)}{=} \mathbb{E}[\overbrace{\mathbb{E}[X_N|\mathcal{F}_n]}^{=X_n}|\mathcal{G}_n]$$

$$\overset{(a)}{=} X_n$$

□

Corollary 5
If X is a martingale w.r. to any given filtration $(\mathcal{F}_n)_n$, then X is also a martingale w.r. to the natural filtration $(\mathcal{F}_n^X)_n$.

Proof.
Since $\sigma(X_n) \subseteq \mathcal{F}_n^X$ we can apply the tower property in order to show that

$$\mathbb{E}[X_N|X_n] = \mathbb{E}[\mathbb{E}[X_N|\mathcal{F}_n^X]|X_n] \overset{(a)}{=} \mathbb{E}[X_n|X_n] = X_n.$$

□

To sum up, the above properties show that if X is a martingale then for all $N \geq n$ we have that $\mathbb{E}[X_N|X_n] = X_n$.

Finally, we introduce two broader classes of stochastic processes whose intersection gives exactly the set of martingale processes.

Def. (Submartingale and supermartingale)
A process $X = (X_n)_n$ is called a ***submartingale*** (***supermartingale***) w.r. to a given filtration $(\mathcal{F}_n)_n$ if

i. $X_n \in L^1(\Omega, \mathbb{P})$ for all n.

ii. X is adapted to $(\mathcal{F}_n)_n$

iii. $X_n \overset{(\geq)}{\leq} \mathbb{E}[X_N|\mathcal{F}_n]$.

Expected value It's straightforward to check that, for a supermartingale (submartingale), the expected value is always increasing (decreasing), since

$$\mathbb{E}[X_N] = \mathbb{E}[\mathbb{E}[X_N|\mathcal{F}_n]] \underset{(\leq)}{\overset{\text{a.s.}}{\geq}} \mathbb{E}[X_n].$$

6.2 Stopping times

We now introduce a class of events which is extremely relevant to the analysis of stochastic process. Broadly speaking, this class of events is comprised by all events such that at time n we can tell whether they have occurred or not.

Def. (Stopping time)
Let $(\mathcal{F}_n)_n$ be a filtration. We say that a random variable $\tau : \Omega \longrightarrow [0, +\infty]$ is a ***stopping time*** if the event $\{\tau \leq n\}$ is such that

$$\{\tau \leq n\} \in \mathcal{F}_n \quad \text{for all } n.$$

Observability This is an observability condition for the random variable τ, i.e. at time n we must be able to tell whether the above event occurred or not based on the available information $\mathcal{F}_n$.

Remark Let τ be a stopping time and consider the event $\{\tau > n\}$. Then, the following events are also observable

$$\{\tau > n\} = \{\tau \leq n\}^c \in \mathcal{F}_n$$

$$\{\tau = n\} = \{\tau \leq n\} \setminus \underbrace{\{\tau \leq n-1\}}_{\in \mathcal{F}_{n-1} \subset \mathcal{F}_n} \in \mathcal{F}_n.$$

Example (Exit – or hitting – time)
Let X be a discrete-time stochastic process and consider a Borel set H. Let now I_H be the set of times at which X exits from H, i.e.

$$I_H := \{n : X_n \notin H\}.$$

Let now τ be the random variable that describes the time of first exit,

$$\tau := \begin{cases} \inf I_H & \text{if } I_n \neq \emptyset \\ +\infty & \text{if } I_H = \emptyset \end{cases}$$

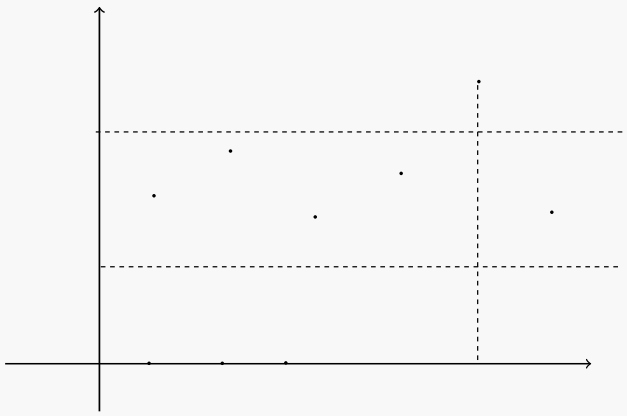

Figure 10: Example of a hitting time for a given set H.

This random variable is as a stopping time, since the event $\{\tau \leq n\}$ can be written as

$$\{\tau \leq n\} = \bigcup_{i \leq n} \underbrace{\{X_i \notin H\}}_{\mathcal{F}_i \subset \mathcal{F}_n} \in \mathcal{F}_n.$$

Continuous-time The previous example shows why this definition of a stopping time becomes problematic for continuous-time stochastic processes, due to the fact that a countable union of events is not guaranteed to belong to the σ-algebra $\mathcal{F}_n$.

6.3 Markov processes

Def. (Markov property)
A discrete-time stochastic process $X = (X_n)_n$ defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_n, \mathbb{P})$ has the ***Markov property*** if it is adapted and the following property holds true for any n:

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] = \mathbb{E}[\varphi(X_{n+1})|X_n], \tag{M}$$

for any φ $\mathcal{B}$-measurable and bounded.

Interpretation Expectation of future values conditional to all cumulated information is equal to the expectation given the value of the process at time n.

Regression function If X has the Markov property, then we can find a function g_n such that

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] = g_n(X_n),$$

where $g_n(x) = \mathbb{E}[\varphi(X_{n+1})|X_n = x]$ is the ***regression function*** (see Doob's Lemma 4).

In practice Assume now that $\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] = f_n(X_n)$ is a deterministic function of X_n, then by the tower property of $\mathbb{E}$ we can write

$$\mathbb{E}[\varphi(X_{n+1})|X_n] = \mathbb{E}[\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n]|X_n] = \mathbb{E}[f_n(X_n)|X_n] = f_n(X_n).$$

Therefore, if we can find that the expectation of X_{n+1} is a deterministic function of X_n, we can conclude that X has the Markov property.

Example (Independent r.v.'s form a Markov proces)
Let $X = (X_n)_n$ be a sequence of independent r.v.'s, then $\mathcal{F}_n = \mathcal{F}_n^X$ and

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n^X] \overset{\perp}{=} \mathbb{E}[\varphi(X_{n+1})].$$

Lemma 5 (Freezing)
If X, Y are random variables and $\mathcal{G}$ a σ-algebra such that Y and $\mathcal{G}$ are independent and X is $\mathcal{G}$-measurable, then we have that

$$\mathbb{E}[f(X,Y)|\mathcal{G}] = \mathbb{E}[f(x,Y)]\Big|_{x=X}$$

Proof.
No.

□

Interpretation Since Y is independent of the information, the randomness in X goes out of the conditioning operation.

Example (Cumulative sum is a Markov process)
Consider now the stochastic process $Y_n := \sum_{i=1}^n X_i$ for the process X defined in the previous example. Then, we have that

$$\begin{aligned}\mathbb{E}[\varphi(Y_{n+1})|\mathcal{F}_n^X] &\overset{\text{def}}{=} \mathbb{E}[\varphi(Y_n + X_{n+1})|\mathcal{F}_n^X] \\ &= \mathbb{E}[\varphi(y + X_{n+1})]\big|_{y=Y_n} \qquad \text{(Freezing Lemma 5)},\end{aligned}$$

which is a deterministic function of X_{n+1} and therefore makes Y a Markov process.

Prop. 4 (Characterization of Markov's property)
The Markov property (M) for a process X is equivalent to satisfying, for any $A \in \mathcal{B}$,

$$\underbrace{\mathbb{E}[\mathbb{1}_{X_{n+1}\in A}|\mathcal{F}_n]}_{\mathbb{P}(X_{n+1}\in A|\mathcal{F}_n)} = \underbrace{\mathbb{E}[\mathbb{1}_{X_{n+1}\in A}|X_n]}_{\mathbb{P}(X_{n+1}\in A|X_n)} \qquad (M')$$

Proof.
$(M) \implies (M')$: Since $\mathbb{1}_A$ is a bounded and $\mathcal{B}$-measurable function, it is valid by choosing $\varphi = \mathbb{1}_A$.

$(M') \implies (M)$: Let $(\varphi_k)_k$ be a sequence of simple functions of the type $\varphi_k = \sum_{j=1}^m c_{j,k}\mathbb{1}_{A_{j,k}}$, which are bounded and $\mathcal{B}$-measurable, and such that

$$\varphi_k \xrightarrow{k\to\infty} \varphi.$$

See for instance here for the standard construction of such a sequence of simple functions $(\varphi_k)_k$ when approximating a bounded function φ. With this approximation, we can chain the following equations:

$$\begin{aligned}\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] &\overset{\text{DCT}}{=} \lim_{k\to\infty} \mathbb{E}[\varphi_k(X_{n+1})|\mathcal{F}_n] \\ &= \lim_{k\to\infty} \sum_{j=1}^m c_{j,k}\mathbb{E}[\mathbb{1}_{A_{j,k}}(X_{n+1})|\mathcal{F}_n] \\ &= \lim_{k\to\infty} \sum_{j=1}^m c_{j,k}\mathbb{P}(X_{m+1} \in A_{j,k}|\mathcal{F}_n) \\ &= \lim_{k\to\infty} \sum_{j=1}^m c_{j,k}\mathbb{P}(X_{m+1} \in A_{j,k}|X_n) \qquad (M') \\ &= \text{(Do the steps backwards)} \\ &= \mathbb{E}[\varphi(X_{n+1})|X_n].\end{aligned}$$

□

Example (Enlarging the filtration breaks Markov)
Let $\mathcal{G}_n = \mathcal{F}$ be the maximal filtration for all $n \in \mathbb{N}$, then for a discrete Markov process X we have

$$\mathbb{E}[X_{n+1}|\mathcal{G}_n] = \mathbb{E}[X_{n+1}|\mathcal{F}] = X_{n+1} \neq \mathbb{E}[\varphi(X_{n+1})|X_n].$$

On the other hand, when we reduce the filtration we have a preservation result analogous to what we have seen with martingales (Prop. 3).

Prop. 5 (Reducing the filtration preserves Markov)
If X has (M) and $(\mathcal{G}_n)_n$ is a filtration such that

a) X is adapted to $(\mathcal{G}_n)_n$

b) $\mathcal{G}_n \subset \mathcal{F}_n$,

then X has (M) w.r. to $(\mathcal{G}_n)_n$.

Proof.
Similarly to Prop. 3, use the tower property of the conditional expected value.

□

Prop. 6 (Equivalent definition of Markov's property)
Property (M) for a process X is equivalent to satisfying, for each $N > n$,

$$\mathbb{P}(X_N \in A|\mathcal{F}_n) = \mathbb{P}(X_N \in A|X_n) \qquad (M'')$$

Proof.
Homework.

□

Validity All the properties we have discussed until now are expressed in their general form and are valid for any type of discrete-time stochastic process, i.e. whether each random variable X_n is characterized either by a continuous or discrete distribution. What we discuss below is a specialization of the properties in the case when X is a discrete-time process for which X_n takes discrete values.

6.4 Markov chains

References Brémaud (2020)

Def. (Discrete-time process)
A discrete-time process X is called a ***discrete process*** if X_n takes values on a countable state space E.

Example Some examples are $E = \mathbb{N}, \mathbb{N}^2, \mathbb{Z}, \mathbb{Z}^2, \dots$

Notation Following the notation of Brémaud (2020), we use i, j, k, h, l and $i_0, i_1, i_n, \dots$ to denote the elements of the countable space E.

Def. (Markov chain)
A discrete process X is called a ***Markov chain*** if it has the Markov property w.r. to the natural filtration $(\mathcal{F}_n^X)_n$.

Prop. 7 (Equivalent definition of Markov chain for discrete processes)
A discrete process X is a Markov chain if and only if for any n and for any $i_0, i_1, \dots, i_n, j \in E$

$$\mathbb{P}(X_{n+1} = j | X_n = i_n, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i_n), \qquad (*)$$

whenever this probability is valid, i.e. $\mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) > 0$.

Proof.
No.

□

Problem This definition works only for processes which are in discrete time and are defined on a countable state space. The more general definition (M) can be used instead for discrete-time continuous processes.

Def. (Homogeneous Markov chain)
We call a Markov chain X ***homogeneous*** (HMC) if the right-hand side of $(*)$ does not depend on n, i.e. if

$$\mathbb{P}(X_{n+1} = j | X_n = i_n, \dots, X_0 = i_0) = \mathbb{P}(X_1 = j | X_0 = i_n).$$

Example (HMC)
If X is a HMC then, for example

$$\mathbb{P}(X_3 = 4 | X_2 = 1, X_1 = 0, X_0 = -1) \overset{(*)}{=} \mathbb{P}(X_3 = 4 | X_2 = 1) \overset{\text{HMC}}{=} \mathbb{P}(X_1 = 4 | X_0 = 1).$$

A HMC is particularly important since we can define a transition matrix that describes the transition from one state to another regardless of the time.

Def. (Transition matrix of a HMC)
For a HMC X we define the ***transition matrix*** as the countable family of numbers

$$P = (p_{ij})_{i,j \in E}, \qquad p_{ij} = \mathbb{P}(X_1 = j | X_0 = i).$$

Properties of P For any $i \in E$, every row of P is a probability distribution and therefore P is a *stochastic matrix*, i.e.

$$\sum_{j \in E} p_{ij} = 1 \qquad \text{for all } i \in E.$$

Consider now the process of making two Markov chain transitions. In this case we have to use P two times in order to transition from $X_0 \to X_1$ and then from $X_1 \to X_2$. To compute these probabilities, we introduce a generalization of the matrix multiplication and addition operations in order to define the powers of an infinite-dimensional matrix $P^2, P^3, \text{etc}\ldots$

Algebraic operations Let $A = (a_{ij})_{i,j \in E}$ and $B = (b_{ij})_{i,j \in E}$ be two transition matrices, then we generalize the usual sum and product operations for standard matrices as

$$A + B = (a_{ij} + b_{ij})_{i,j \in E}$$

$$A \cdot B = \Big(\sum_{k \in E} a_{ik} b_{ik} \Big)_{i,j \in E}$$

Let now $\boldsymbol{x} = (x_i)_{i \in E}$ be a column vector, then

$$A\boldsymbol{x} = \Big(\sum_{k \in E} A_{ik} x_k \Big)_{i \in E}$$

$$\boldsymbol{x}^\top A = \Big(\sum_{k \in E} x_k A_{ki} \Big)_{i \in E}$$

Example (1D random walk)
Consider a r.v. X_0 with values in $E = \mathbb{Z}$. Let now $(Z_n)_{n \in \mathbb{N}}$ be i.i.d r.v.'s such that

$$Z_n \sim p\delta_1 + (1-p)\delta_{-1}, \quad p \in (0,1).$$

We set $X_{n+1} = X_n + Z_{n+1}$ and we consider the stochastic process $X = (X_n)_{n \in \mathbb{N}}$. We already know that X is a HMC, and this stochastic process increases by 1 with probability p and decreases by 1 with probability $1-p$. Therefore, its transition matrix is given by

$$p_{ij} = \begin{cases} p & \text{if } j = i+1 \\ 1-p & \text{if } j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

Exercises

1. Proof of the proposition
2. (Brémaud, 2020, p. 88) ex. 2.1.1 - 2.1.6, 2.2.1

Lecture 7: Homogeneous Markov chains

2021-11-25

Homogeneous Markov chains are the basis upon which more complicated Markov chains are studied. Their properties will be studied in detail in the following lectures, starting from the basic quantities of interest.

7.1 Initial distribution

The distribution of a homogeneous Markov chain X only depends on

1. The initial law π_0, $\pi_0(\{i\}) = \mathbb{P}(X_0 = i)$ for all $i \in E$.
2. The transition matrix P.

More precisely, for $i_0, i_1, \ldots, i_n \in E$ we have that the probability of the path from i_0 through $i_1, \ldots, i_n$ is equal to

$$\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) = \pi_0(i_0) \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-1} i_n}.$$

***n* steps ahead** Now we need to compute the conditional probability of transition for multiple time steps, $\mathbb{P}(X_n = i_n | X_0 = i_0)$. We consider the probability distribution at time n,

$$\begin{aligned}
\pi_n(j) &= \mathbb{P}(X_n = j) \\
&= \sum_{i \in E} \mathbb{P}(X_n = j, X_{n-1} = i) \\
&= \sum_{i \in E} p_{ij} \mathbb{P}(X_{n-1} = i) \\
&= \sum_{i \in E} \pi_{n-1}(i) \cdot p_{ij} \\
&= (\pi_{n-1} P)_j \qquad (\pi \text{ row vector})
\end{aligned}$$

Therefore, $\pi_n = \pi_{n-1} \cdot P$ and if we repeat this process n times we obtain the following equation

$$\boxed{\pi_n = \pi_0 P^n}$$

Notation We denote by P^n_{ij} the element (i, j) of P^n.

Example

Let $E = \{1, 2, 3, 4\}$ and consider the initial distribution $\pi_0(\{j\}) = \frac{1}{4}$ with the transition graph in Figure 11.

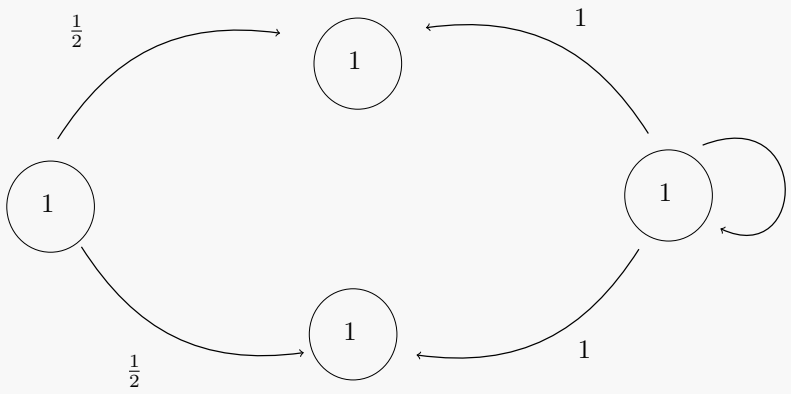

Figure 11: transitionGraph

For the above graph, $\mathbb{P}(X_3 = j|X_2 = 2)$ is not defined, whereas $\mathbb{P}(X_2 = j|X_1 = 2)$ is well-defined.

Future paths With Markov chains we can consider two general events,

$$\text{FUTURE} \quad A = (X_{n+1} = j_1) \cap \ldots \cap (X_{n+k} = j_k)$$

$$\text{PAST} \quad B = (X_{n-1} = j_{n-1}) \cap \ldots \cap (X_0 = j_0)$$

From the Markov property (M) we can prove (long and boring proof) that

$$(M) \iff \mathbb{P}(A|X_n = i_n, B) = \mathbb{P}(A|X_n = i_n). \tag{3}$$

This is a bit stronger than the single value at time $n + 1$, since we consider the whole trajectory from time $n + 1$ to $n + k$.

Conditional independence In (3) we can multiply by $\mathbb{P}(B|X_n = i_n)$ to get

$$\mathbb{P}(A|X_n = i_n, B)\frac{\mathbb{P}(B \cap \{X_n = i_n\})}{\mathbb{P}(\{X_n = i_n\})} = \frac{\mathbb{P}(A \cap B \cap \{X_n = i_n\})}{\cancel{\mathbb{P}(\{X_n = i_n\} \cap B)}}\cancel{\frac{\mathbb{P}(B \cap \{X_n = i_n\})}{\mathbb{P}(\{X_n = i_n\})}}$$

therefore

$$\boxed{\mathbb{P}(A \cap B|X_n = i_n) = \mathbb{P}(A|X_n = i_n)\mathbb{P}(B|X_n = i_n)}$$

From this we conclude that $A \perp\!\!\!\perp B$ conditional to $X_n = i_n$.

7.2 Canonical construction

Theorem 26 (Canonical representation)
Let $(Z_n)_n$ be a sequence of i.i.d r.v.'s with values on a measurable space $(G, \mathcal{G})$. Let $f : E \times G \longrightarrow E$ be a measurable function w.r. to the product σ-algebra $\mathscr{P}(E) \otimes \mathcal{G}$ and $X_0 \perp\!\!\!\perp (Z_n)_n$ be an initial r.v. with values on E. Then, if we define

$$X_{n+1} := f(X_n, Z_{n+1}), \quad \text{for all } n \in \mathbb{N}_0,$$

we have that the process $X = (X_n)_{n\in\mathbb{N}_0}$ is a homogeneous Markov chain with transition matrix

$$P_{ij} = \mathbb{P}\big(f(i, Z_1) = j\big).$$

Non-i.i.d If the Z_n are not identically distributed, then we would have a non-homogeneous Markov chain with transition matrix $P_{ij}(n) = \mathbb{P}\big(f(i, Z_n) = j\big)$

White noise The above representation is also called a Markov chain *driven by white noise*, since the sequence of $(Z_n)_n$ is of i.i.d random variables.

Proof.
We will prove this under the general Markov property (M). For a set $A \subset E$ we can write

$$\mathbb{P}(X_{n+1} \in A|\mathcal{F}_n^X) \overset{\text{def}}{=} \mathbb{E}(\mathbb{1}_{X_{n+1}\in A}|\mathcal{F}_n^X) = \mathbb{E}[\mathbb{1}_A(X_{n+1})|\mathcal{F}_n^X].$$

With this notation we can use the definition of X_{n+1} in the theorem and since Z_{n+1} is independent of $\mathcal{F}_n^X = \sigma(X_1, \ldots, X_n) = \sigma(Z_1, \ldots, Z_n)$ we can write

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_A(X_{n+1})|\mathcal{F}_n^X] &= \mathbb{E}[\mathbb{1}_A\big(f(X_n, Z_{n+1})\big)|\mathcal{F}_n^X] \\
&= \mathbb{E}[\mathbb{1}_A\big(f(i, Z_{n+1})\big)]\Big|_{i=X_n} && \text{(freezing lemma 5)} \\
&= \mathbb{E}[\mathbb{1}_A\big(f(i, Z_1)\big)]\Big|_{i=X_n} && (Z_n \text{ are i.i.d})
\end{aligned}$$

and since this is a deterministic function of X_n, we have the Markov property. Moreover, if we choose $A = \{i\}$ we have the transition probability

$$P_{ij} = \mathbb{P}\big(f(i, Z_1) = j\big).$$

□

Example (Random walk)
Setting $G = \{-1, 1\}$ and $f(i, z) = i + z$ yields the 1D random walk

$$X_{n+1} = X_n + Z_{n+1}, \quad Z = p\delta_1 + (1 - p)\delta_{-1}.$$

7.2.1 Reverse problem

The problem we now consider is the following: given π_0 and $P = (P_{ij})_{i,j\in E}$ transition matrix, we want to find a process $X = (X_n)_n$ such that

1. X is a HMC.
2. $X_0 = \pi_0$.
3. P is the transition matrix of X.

To do so, we generate a categorical variable with probabilities $P_i = (p_{i1}, \dots, p_{iE})$ and we transition from i to j if the categorical variable indicates the set j. In order to write the process for checking theorem 26 we define the following objects:

› $(Z_n)_n$ i.i.d with $Z_n \sim \text{Unif}_{[0,1]}$.

› $f : E \times [0,1] \longrightarrow E$ defined as

$$f(i,z) = \sum_{j\in E} j \cdot \mathbb{1}_{A_{ij}}(z),$$

where $A_{ij} := \left[\sum_{k=0}^{j-1} P_{ik}, \sum_{k=0}^{j} P_{ik}\right]$

We obtain the canonical construction by defining

$$X_{n+1} := f(X_n, Z_{n+1}),$$

which is a homogeneous Markov chain by theorem 26 with transition matrix P as required, since

$$\mathbb{P}\big(f(i, Z_1) = j\big) = \mathbb{P}(Z_1 \in A_{ij}) = P_{ij}.$$

Now, is the canonical representation unique or do we have situations like $X \sim \mathcal{N}(0,1)$ and $-X \overset{\text{d}}{=} X$? It turns out that the canonical representation is <u>not</u> unique, and it can be seen via the following counterexample.

Example (Urn of Ehrenfest)
We have urns A and B

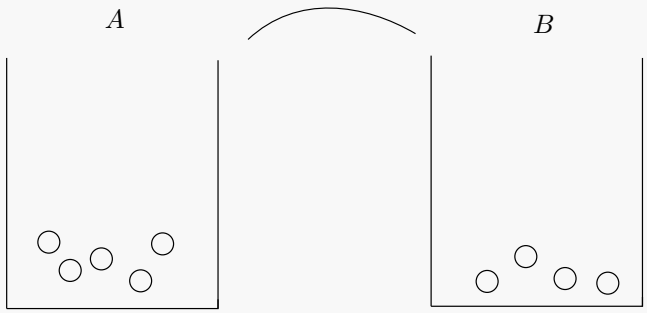

Figure 12: Urn of Ehrenfest with $N = 9$ total balls.

We have N total balls and everytime we choose a ball its urn is changed, therefore we can write this process

$$X_{n+1} = X_n + \underbrace{Z_{n+1}}_{\in\{-1,1\}} .$$

We observe that although the functional relationship is the same, the process is different from the random walk, since Z_{n+1} and X_n are not independent,

$$\mathbb{P}(Z_{n+1} = 1 | X_n = i) = \frac{N-i}{N}.$$

The point is that this process cannot be written as a canonical representation, however it's possible to write another process with the same law using the canonical representation.

Example (Gambler's ruin)

We consider a 1D random walk $X = (X_n)_{n\in\mathbb{N}_0}$ lower bound $c \in \mathbb{N}_0$ and with an initial value $X_0 \equiv a$. At each time we can either win or lose, and X_n is the cumulated gain at time n. Consider also the stopping time

$$\tau = \begin{cases} \inf\{n : X_n \in \{0, c\}\} \\ +\infty \qquad\qquad\qquad\qquad \text{if } X_n \notin \{0,c\} \text{ for all } n \end{cases}$$

We want to compute

1. $\mathbb{P}(\underbrace{\tau < \infty, X_\tau = c}_{F:=\text{winning game}} | X_0 = a)$.
2. $\mathbb{E}[\tau | X_0 = a]$.

Problem 1 – We set $u(i) = \mathbb{P}(F|X_0 = i)$ for all $i = 0, \dots, c$. We have some *boundary conditions* on u, since

$$\begin{cases} u(0) = 0 \\ u(c) = 1 \end{cases} .$$

Remark One can prove that, since by the Markov property (M) the future trajectory depends only on the present value, then

$$\mathbb{P}(F|X_0 = i) = u(i) = \mathbb{P}(F|X_n = i),$$

Therefore

$$\begin{aligned} \mathbb{P}(F|X_0 = i) &= \sum_{j\in E} \mathbb{P}(F, X_1 = j | X_0 = i) \\ &= \sum_{j\in E} \mathbb{P}(F|X_1 = j)\mathbb{P}(X_1 = j|X_0 = i) \\ &= \sum_{j\in E} u(j) P_{ij} \end{aligned}$$

In the case of the random walk we therefore have that P_{ij} is $\neq 0$ only for $j = i-1$ and $j = i+1$, and we get the following recursion for $i = 1, \dots, c-1$

$$\begin{cases} u(0) = 0 \\ u(i) = u(i+1)p + u(i-1)(1-p) \\ u(c) = 1. \end{cases}$$

We can compute this solution as

$$u(i) = \begin{cases} \frac{1-\left(\frac{1-p}{p}\right)^i}{1-\left(\frac{1-p}{p}\right)^c} & \text{if } p \neq \frac{1}{2} \\ \frac{i}{c} & \text{if } p = 2 \end{cases}$$

Remark Define now $F^c :=$ losing the game, then since $p = \frac{1}{2}$ the game is symmetric and

$$\mathbb{P}(F^c | X_0 = i) = \frac{c-i}{c}.$$

Moreover, this shows that

$$\mathbb{P}(\tau < \infty | X_0 = i) = \mathbb{P}(F | X_0 = i) + \mathbb{P}(F^c | X_0 = i) = \frac{i}{c} + \frac{c-i}{c} = 1.$$

Problem 2 – For a specific initial wealth i, we can apply the same one-step analysis we did before, this time in terms of expected values: set $m(i) := \mathbb{E}[\tau | X_0 = i]$

Remark From the Markov property (M) we can show that $\mathbb{E}[\tau | X_1 = i] = m(i) + 1$.
We partition w.r. to the only two events that can occur, which are

$$\begin{aligned} m(i) &= \mathbb{E}[\tau \cdot \mathbb{1}_{\{X_1 = i-1\}} | X_0 = i] + \mathbb{E}[\tau \cdot \mathbb{1}_{\{X_1 = i+1\}} | X_0 = i] \\ &= \mathbb{E}[\tau | X_1 = i+1] \cdot \mathbb{P}(X_1 = i+1 | X_0 = i) + \mathbb{E}[\tau | X_1 = i-1] \cdot \mathbb{P}(X_1 = i-1 | X_0 = i) \\ &= \big(m(i+1) + 1\big)p + \big(m(i-1) + 1\big)(1-p) \\ &= p \cdot m(i+1) + (1-p) \cdot m(i-1) + 1. \end{aligned}$$

This goes together with the two boundary conditions

$$\begin{cases} m(0) = 0 \\ m(i) = p \cdot m(i+1) + (1-p) \cdot m(i-1) + 1 \\ m(c) = 0 \end{cases}$$

Solving for $p = \frac{1}{2}$ we obtain $m(i) = i(c-i)$.

In general, for computing probabilities and expected values of *absorbing states* we can apply this type of *first-step analysis*.

7.3 Stationarity

We want to characterize the property of stationarity for a stochastic process, that is, Markov chains whose distribution does not change over time.

Def. (Stationary distribution)
We say that a probability distribution $\pi = (\pi_1, \pi_2, \dots, \pi_{|E|})^\top$ on E is ***stationary*** if

$$\pi^\top = \pi^\top \cdot P.$$

Remark A distribution is not universally stationary, but w.r. to a chosen transition matrix P.

Remark If π is stationary, then $\pi^\top = \pi^\top P^n$ for any $n \in \mathbb{N}$.

Def. (Stationary Markov chain)
A HMC is called ***stationary*** if

$$\mathbb{P}(X_n = i_0, X_{n+1} = i_{n+1} \dots, X_{n+k} = i_k) = \mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_k = i_k),$$

for any choice of $n \in \mathbb{N}_0, k \in \mathbb{N}$ and $i_0, \dots, i_k \in E$.

Prop. 8 (Stationarity of the HMC)
An HMC is stationary $\iff$ the initial distribution π_0 is stationary.

Proof.
$\implies$: Simple to prove.
$\impliedby$: We start by computing

$$\begin{aligned}
\mathbb{P}(X_n = i_0, \dots, X_{n+k} = i_k) &= \mathbb{P}(X_n = i_0) \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{k-1} i_k} \\
&= \pi_0^\top \cdot P^n \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{k-1} i_k} \\
&= \pi_0^\top \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{k-1} i_k} \qquad \text{(stationary)} \\
&= \mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_k = i_k).
\end{aligned}$$

□

Lecture 8: Global balance equation

2021-11-26

A useful tool in the analysis of homogeneous Markov chains is the global balance equation. The equation stems from the definition of stationary distribution,

$$\pi^\top P = \pi^\top \iff \pi(i) = \sum_{j \in E} \pi(j) P_{ji}, \quad \text{for all } i \in E.$$

Example

Let $E = \{1, 2\}$ and $P = \left(\begin{smallmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{smallmatrix}\right)$ for $\alpha, \beta \in (0, 1)$, therefore we can write the global balance equations:

$$\begin{cases} \pi(1) = \pi(1)(1-\alpha) + \pi(2)\beta \\ \pi(2) = \pi(1)\alpha + \pi(2)(1-\beta) \end{cases} \tag{4}$$

which can be solved as a linear system in order to find the solution in $\pi(1), \pi(2)$.

$$0 = \pi^\top (P - I) = \pi^\top \underbrace{\left(\begin{smallmatrix} -\alpha & \alpha \\ \beta & -\beta \end{smallmatrix}\right)}_{\det = 0}$$

, since the determinant is null all the solutions to (4) are found by solving the first equation, constrained to the fact that π is a probability distribution,

$$\begin{cases} 0 = -\alpha\pi(1) + \beta\pi(2) & \text{(global balance)} \\ \pi(1) = 1 - \pi(2) & \text{(probability distribution)} \end{cases}$$

$$\Updownarrow$$

$$\begin{cases} \pi(1) = \frac{\beta}{\alpha}\pi(2) = \frac{\beta}{\alpha+\beta} \\ \pi(2) = \frac{1}{1+\frac{\beta}{\alpha}} = \frac{\alpha}{\alpha+\beta} \end{cases}$$

Example (Urn of Ehrenfest (cont.))

We recall that for the urn of Ehrenfest we have

$$P_{ij} = \begin{cases} \frac{i}{N} & \text{if } j = i-1 \\ \frac{N-i}{N} & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases},$$

therefore the global balance equation can be computed as

$$\begin{cases} \pi(0) = \pi(1) \cdot \frac{1}{N} \\ \pi(N) = \pi(N-1) \cdot \frac{1}{N} \\ \pi(i) = \pi(i-1) \cdot \frac{N-i+1}{N} + \pi(i+1) \cdot \frac{i+1}{N} \quad i = 1, \dots, N-1 \end{cases}$$

We claim that $\pi(i) = \pi_0 \cdot \binom{N}{i+1}$ for any $i = 0, \dots, N$. From the first equation, we have

$\boxed{i = 1}$: $\pi(1) = \pi(0) \cdot N$

$\boxed{i \neq 1}$: By induction, assuming we isolate $\pi(i+1)$ as

$$\begin{aligned}
\pi(i+1) &= \frac{N}{i+1}\left(\pi(i) - \pi(i-1)\cdot\frac{N-i+1}{N}\right) \\
&\overset{\text{induct.}}{=} \frac{N}{i+1}\left(\pi(0)\binom{N}{i} - \pi(0)\binom{N}{i-1}\frac{N-i+1}{N}\right) \\
&= \pi(0)\cdot N!\left(\frac{N}{(i+1)i!(N-i)!} - \underbrace{\frac{N-i+1}{(i+1)(i-1)!(N-i+1)!}}_{=\frac{1}{(i+1)(i+1)!(N-i)!}}\right) \qquad \text{(collect } \pi(0), N!\text{)} \\
&= \pi(0)\cdot N!\cdot\left(\frac{N-i}{(i+1)!(N-i)!}\right) \\
&= \pi(0)\cdot N!\cdot\left(\frac{1}{(i+1)!(N-i+1)!}\right) \\
&= \pi(0)\cdot\binom{N}{i+1}
\end{aligned}$$

Now, the only unknown is $\pi(0)$, which we get by the constraint that π is a distribution, therefore

$$\begin{aligned}
\sum_{i=0}^{N}\pi(i) = 1 &\iff \pi(0)\cdot\sum_{i=0}^{N}\binom{N}{i} = 1 \\
&\iff \pi(0)\cdot 2^N = 1 \qquad \text{(binomial theorem)} \\
&\iff \pi(0) = \frac{1}{2^N}.
\end{aligned}$$

Therefore, $\pi(i) = \binom{N}{i}2^{-N} = \binom{N}{i}\frac{1}{2^i}\frac{1}{2^{N-i}}$ for $i = 0, \dots, N$, therefore $\pi = \text{Bin}(N, \frac{1}{2})$.

Example (1-D random walk)

Consider X a 1-D random walk with $p = \frac{1}{2}$, with transition matrix

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

then, using the global balance equation we have

$$\pi(i) = \frac{1}{2}\pi(i-1) + \frac{1}{2}\pi(i+1), \quad \text{for all } i \in \mathbb{Z}.$$

By induction,

$$\pi(i) = \pi(0) + i\big(\pi(i) - \pi(0)\big) \quad \text{for all } i \geq 2$$

A stationary distribution would imply $\pi(i) = \pi(0)$ for any $i \geq 0$, but since we have infinite values of i even having the condition

$$\pi(i) = \pi(0) \implies \sum_{i=0}^{\infty} \pi(0) = \infty.$$

Remark In general there might exist many stationary distributions for a given matrix P.

Example (Multiple stationary distributions)
We just have to find a distribution such that

$$\pi^\top = \pi^\top P,$$

therefore if $P = I$ this equation is satisfied for any initial distribution π.

8.1 Strong Markov property

Let τ be a stopping time w.r. to $(\mathcal{F}_n^X)_n$, then $\{\tau = n\} \in \mathcal{F}_n^X = \sigma(X_0, X_1, \ldots, X_n)$. However, by Doob's theorem the random variable $\mathbb{1}_{\{\tau=n\}}$ can be written as a function of $X_0, \ldots, X_n$,

$$\mathbb{1}_{\{\tau=n\}} = \psi(X_0, X_1, \ldots, X_n).$$

Example (Return time)
Define

$$T_i := \begin{cases} \min\{n \in \mathbb{N} : X_n = i\} & \text{if } \neq \emptyset \\ +\infty & \text{if } X_n \neq i \text{ for all } n \in \mathbb{N} \end{cases}$$

This is different from the hitting time, since in that case we had

$$\tau_i := \begin{cases} \min\{n \in \mathbb{N}_0 : X_n = i\} & \text{if } \neq \emptyset \\ +\infty & \text{if } X_n \neq i \text{ for all } n \in \mathbb{N} \end{cases}$$

The problem is that for $X_0 = i \not\Longrightarrow T < \infty$.

Example (Successive returns)
Fix $i \in E$ and let

$$\tau_1^{(i)} := T_i,$$

$$\tau_{n+1}^{(i)} := \begin{cases} \min\{m > \tau_n^{(i)} : X_m = i\} & \text{if } \neq \emptyset \\ +\infty & \text{if } X_m \neq i \text{ for all } m > \tau_n^{(i)} \end{cases}$$

We observe that the event

$$\{\tau_k^{(i)} = m\} = \underbrace{\{X_m = i\}}_{\in \mathcal{F}_m^X} \cap \underbrace{\Big\{ \sum_{n=1}^{m} \mathbb{1}_{\{X_n = i\}} = k \Big\}}_{\in \mathcal{F}_m^X}.$$

Prop. 9 (Stopping times)
If τ is a stopping time, then for all $n \in \mathbb{N}_0$ (negatives don't work),

- $\tau + n$ *is a stopping time.*
- $\tau \wedge n = \min\{\tau, n\}$ *is a stopping time.*

Proof.

$$\{\tau + n = m\} \overset{n \geq 0}{=} \{\tau = m - n\} \in \mathcal{F}_m$$

$$\{\min\{\tau, n\} = m\} = \{\tau < n\}$$

□

Let $X = (X_n)_n$ be a discrete-time stochastic process and τ a stopping time. We now consider the stochastic process given by the process observed only when τ occurs, i.e.

$$X_\tau(\omega) := \begin{cases} X_{\tau(\omega)}(\omega) & \text{if } \tau(\omega) < \infty \\ \Delta \notin E & \text{if } \tau(\omega) = \infty \end{cases}$$

We want to investigate the behaviour of

$$(X_{\tau+n})_n \quad \text{“process } X \text{ after } \tau\text{”}$$

$$(X_{\min\{\tau,n\}})_n \quad \text{“process } X \text{ stopped at } \tau\text{”}$$

Given the observations in Prop. 9, the above processes remain adapted to the filtration $\mathcal{F}_n^X$. However, there is more we can say about the processes which are observed in terms of the stopping times, which is summarized by the following theorem.

Theorem 27 (Strong Markov property)
Let X be a HMC and τ be a stopping time, then for any $i \in E$ we have

a) $\mathbb{P}(A | X_\tau = i, B) = \mathbb{P}(A | X_\tau = i)$ *for any A future trajectory and B any previous trajectory,*

$$A = \{X_{\tau+1} = j_1, \ldots, X_{\tau+k} = j_k\}$$

$$B = \{X_{0\wedge\tau} = i_0, \ldots, X_{n\wedge\tau} = i_n\}$$

for any $n, k \in \mathbb{N}$

b) The process $(X_{\tau+n})_n$ is a HMC with transition matrix P.

Proof.
No.

□

8.2 Topology of the transition matrix

We now turn our attention to studying some properties related to the transition matrix P, in particular accessible states and communicating states.

Def. (Accessible state)
A state $j \in E$ is called ***accessible*** from the state $i \in E$ if there exists $M \in \mathbb{N}_0$ such that

$$P_{ij}^M = \mathbb{P}(X_M = j | X_0 = i) > 0,$$

and we write $i \longrightarrow j$.

Def. (Communicating states)
Two states i, j ***communicate*** if $i \longrightarrow j$ and $j \longrightarrow i$, and we write $i \longleftrightarrow j$.

Prop. 10 (Communication is an equivalence relation)
For two states $i, j \in E$ we have that "$\longleftrightarrow$" is an equivalence relation, since

1. $i \longleftrightarrow i$
2. $i \longleftrightarrow j \implies j \longleftrightarrow i$
3. $\underbrace{i \longleftrightarrow j}_{P_{ij}^M > 0}$ *and* $\underbrace{j \longleftrightarrow k}_{P_{jk}^{M'} > 0} \implies \underbrace{i \longleftrightarrow k}_{P_{ik}^{M+M'}}$.

With this relationship we can define an equivalence class $E/\leftrightarrow$, for which every element $i \in E$ belongs to one and only equivalence class.

Def. (Closedness)
A state $i \in E$ is called ***closed*** if $p_{ii} = 1$. A set $C \subset E$ is called ***closed*** if $\sum_{j \in C} p_{ij} = 1$ for all $i \in C$.

Exercises

› Example 2.4, 3.2, 5.5, 5.6 one per group as a seminar

› Exercise 2.6.2, 2.7.1, 2.7.2 choose 2 per group, the last group does 2.2.3 and 2.5.3

Lecture 9: Recurrence of states

2021-12-02

9.1 Recurrence

Notation In the following we will denote the conditional probability and expectation with a subscript, i.e.

$$\mathbb{P}_i = \mathbb{P}(\cdot|X_0 = i), \quad \mathbb{E}_i = \mathbb{E}[\cdot|X_0 = i]$$

Def. (Recurrence and transience)
Let X be a HMC and $i \in E$, we say that i is ***recurrent*** if $\mathbb{P}_i(T_i < \infty) = 1$, where T_i is the event of first return at state i. Otherwise it is called ***transient***.

Def. (Positive recurrence)
If a state i is recurrent and in addition $\mathbb{E}[T_i] < \infty$, then i is called ***positive recurrent***. Otherwise, it is called ***null recurrent***.

Example (Recurrence conditions)
Let $E = \mathbb{N}_0$ and consider a transition graph represented in Figure 13 below. We assume that $p_n \in (0, 1)$ for every n, and we want to study whether state 0 is recurrent or not.

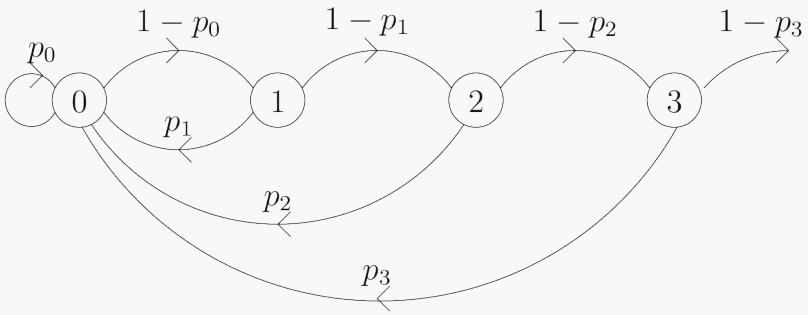

Figure 13: Transition graph of the success-runs chain.

Let $T_0 :=$ "first return time to 0", then we have that

$$\mathbb{P}(T_0 = n) = \begin{cases} \overbrace{(1-p_0)(1-p_1)\cdot\ldots\cdot(1-p_{n-2})}^{n-1 \text{ times different from } 0} p_{n-1} & \text{if } n \geq 2 \\ p_0 & \text{if } n = 1 \end{cases}$$

We can write the probability of T_0 being finite as the following limit,

$$\begin{aligned}
\mathbb{P}(T_0 < \infty) &= \mathbb{P}_0(\bigcup_{n\in\mathbb{N}} \{T_0 = n\}) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}_0(T_0 = i) && \text{(disjoint)} \\
&= \lim_{n\to\infty} \mathbb{P}_0(T_0 \leq n) \\
&= \lim_{n\to\infty} \big(1 - \mathbb{P}_0(T_0 > n)\big) \\
&= 1 - \lim_{n\to\infty} \prod_{i=0}^{n-1} (1 - p_i) && \text{(move right } n \text{ times)}
\end{aligned}$$

The above limit is zero $\overset{(!)}{\iff} \sum_{i=0}^{\infty} p_i = +\infty$, which is the condition for recurrence of the state.

9.2 Potential matrix criterion

We will now study two criteria for determining the recurrence of a stochastic process in terms of its transition matrix P.

Def. (Potential matrix)
We define the ***potential matrix*** of a Markov chain with transition matrix P as the matrix

$$G := \sum_{n\in\mathbb{N}_0} P^n$$

Finiteness The entries of the matrix G are not necessarily finite, think for example to the transition matrix $P = I$.

Returns to j If we define the random variable $N_j := \sum_{n\in\mathbb{N}_0} \mathbb{1}_{\{X_n=j\}}$, i.e. N_j counts the number of times that the chain returns to j, then

$$\mathbb{E}_i[N_j] = \sum_{n\in\mathbb{N}_0} \mathbb{E}_i[\mathbb{1}_{\{X_n=j\}}] = \sum_{n\in\mathbb{N}_0} \mathbb{P}_i(X_n = j) = \sum_{n\in\mathbb{N}_0} P^n_{ij} = G_{ij}.$$

Lemma 6 (Distribution of the return time)
Let $T_j :=$ "return time to state j" and $f_{ij} = \mathbb{P}_i(T_j < +\infty)$. Then, we have that the number of visit N_j to state j is such that

$$\mathbb{P}_i(N_j = r) = \begin{cases} f_{ij} f_{jj}^{r-1}(1 - f_{jj}) & \text{if } r \in \mathbb{N} \\ 1 - f_{ij} & \text{if } r = 0 \end{cases}$$

Interpretation For $r = 0$, it's the probability of never returning to j, i.e. $\mathbb{P}_i(T_j = +\infty)$. For $r > 0$ we go once from $i \rightsquigarrow j$, then $j \rightsquigarrow j$ happens $r - 1$ times and then we do not return there anymore for the rest of the chain.

Proof.

$\boxed{r = 0}$ Proof is simply $\{N_j = 0\} = \{T_j = +\infty\}$.

$\boxed{r > 0}$ We prove this property by induction, i.e. we assume the statement true for a given $r \in \mathbb{N}$. For $r + 1$ we have that

$$\begin{aligned}
\mathbb{P}_i(N_j > r) &= 1 - \mathbb{P}_i(N_j \leq r) \\
&= 1 - \sum_{k=0}^{r} \mathbb{P}_i(N_j = k) \\
&\overset{\text{Hp.}}{=} \cancel{1} - (\cancel{1} - f_{ij}) - \sum_{k=1}^{r} f_{ij} f_{jj}^{k-1}(1 - f_{jj}) \\
&= f_{ij} - f_{ij}(1 - f_{jj}) \sum_{k=1}^{r} f_{jj}^{k-1} \\
&= f_{ij} - f_{ij}\cancel{(1 - f_{jj})} \frac{1 - f_{jj}^r}{\cancel{1 - f_{jj}}} \\
&= f_{ij} - f_{ij}(1 - f_{jj}^r) \\
&= f_{ij} \cdot f_{jj}^r.
\end{aligned}$$

!! Long proof, cannot use the one below because induction assumes the property valid for r and not for $r + 1$!!.

$$\begin{aligned}
\mathbb{P}_i(N_j = r + 1) &= \mathbb{P}_i(N_j > r) - \mathbb{P}_i(N_j > r + 1) \\
&= f_{ij} \cdot f_{jj}^r - f_{ij} f_{jj}^{r+1} \\
&= f_{ij} f_{jj}^r (1 - f_{jj})
\end{aligned}$$

□

Theorem 28 (Potential matrix criterion)
The state $i \in E$ is recurrent $\iff$ $G_{ii} = +\infty$.

Proof.
The proof relies on the previous result (Lemma 6), which characterizes the distribution of the return time T:

$$\overbrace{\mathbb{P}_j(T_j < +\infty)}^{f_{jj}} = 1 \iff \mathbb{P}_j(N_j = +\infty) = 1.$$
$$\iff \mathbb{E}_j[N_j] = +\infty.$$

□

Example (1-D random walk is recurrent for $p = 1/2$)
Consider a 1-D random walk such that $X_n = X_{n-1} + Z_n$, with transition probabilities

$$p_{ij} = \begin{cases} p & \text{if } j = i+1 \\ 0 & \text{if } j = i = 1 \\ 1-p & \text{if } j = i-1 \end{cases}$$

One can show that for odd transitions, $P_{00}^{2n+1} = 0$, since we cannot go back to state 0 in an odd number of steps. On the other hand, for even transitions we have

$$P_{00}^{2n} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{n!n!} p^n (1-p)^n.$$

Using Stirling's approximation, $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$, we have

$$P_{00}^{2n} \sim \frac{\big(4 \cdot p(1-p)\big)^n}{\sqrt{2\pi n}}, \quad \text{for } n \gg 1.$$

Hence, the maximum of P_{00}^{2n} as a function of p is attained when

$$p = \frac{1}{2} \implies \big(4 \cdot p(1-p)\big)^n = 1^n = 1,$$

and so we have that $P_{00}^{2n} \overset{n \gg 1}{\approx} 1/\sqrt{2\pi n}$. Therefore, in order for the process to be recurrent we must have

$$\sum_{n \in \mathbb{N}} P_{00}^{2n} = +\infty \iff p = \frac{1}{2},$$

thus if the random walk is symmetric then all states are recurrent. Otherwise, every state is transient.

9.3 Structure of the transition matrix

A theoretical application of the potential matrix criterion is to the proof that recurrence is a communication class property. Indeed, we have the following remark in terms of communicating states:

Remark Assume $j \in E$ is recurrent and accessible from another state $i \in E$. Then,

$$G_{ij} = \mathbb{E}_i[N_j] = \infty.$$

Proof.
Exercise.

□

Theorem 29 (Recurrence of accessible states)
If $i \longleftrightarrow j$, then i is (positive) recurrent $\iff$ j is (positive) recurrent.

Proof.
By definition, i and j communicate if there are $M, N > 0$ such that $p_{ij}(M) > 0$ and $p_{ji}(N) > 0$. Consider now the probability of the following transition:

$$i \xrightarrow{M \text{ steps}} j \xrightarrow{n \text{ steps}} j \xrightarrow{N \text{ steps}} i \qquad p_{ij}(M)p_{jj}(n)p_{ji}(N)$$

$$i \xrightarrow{M+n+N \text{ steps}} i \qquad p_{ii}(M+n+N)$$

Since the first transition is contained in the second,

$$p_{ii}(M+n+N) \geq p_{ij}(M)p_{jj}(n)p_{ji}(N) = \alpha \cdot p_{jj}(n).$$

With the same argument,

$$p_{jj}(M+n+N) \geq p_{ji}(M)p_{ii}(n)p_{ij}(N) = \beta \cdot p_{ii}(n).$$

Therefore, this implies that the series $\sum_{n=0}^{\infty} p_{ii}(n)$ and $\sum_{n=0}^{\infty} p_{jj}(n)$ either both diverge or both converge.

□

Corollary 6 (Recurrent classes)
If a class $R \subset E$ is recurrent then it is closed.

Proof.
If $i \in R$ goes to $j \in E$, then since i is recurrent the chain must at some point return to i, which implies $P_{ji}^{M} > 0$ for some M and therefore $j \in R$.

□

Indeed, observe that we can rearrange in terms of recurrent classes (first) and transient classes (last):

$$\mathbf{P} = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & T \\ \hline R_1 & \mathbf{P}_1 & 0 & 0 & 0 \\ R_2 & 0 & \mathbf{P}_2 & 0 & 0 \\ R_3 & 0 & 0 & \mathbf{P}_3 & 0 \\ T & & & & \end{array}$$

Figure 14: Rearrangement of the communicating classes in terms of their type.

Def. (Irreducibility)
A HMC X is ***irreducible*** $\iff$ E is the only communicating class.

Notation For irreducible Markov chains, it makes sense to talk about recurrence *of the chain* in place of recurrence of a single state.

Corollary 7 (Irreducibility and recurrence)
If a HMC X is irreducible, then all states $i \in E$ will be either recurrent or transient.

Proof.
Follows from theorem 29.

□

Irreducible Markov chains are special, in the sense that we have another criterion which will tell us about positive recurrence of the chain.

Theorem 30 (Positive recurrence of an irreducible Markov chain)
An irreducible HMC is positive recurrent $\iff$ the HMC admits a stationary distribution π.

Homeworks Each group chooses one example and two exercises.

› Examples 3.2, 3.3, 1.3 pp. 98–108

› Problems 3.1.4, 3.1.6, 3.2.2, 3.3.1, 3.3.3

Lecture 10: Long-run behaviour

2021-12-09

10.1 Invariant measure of a HMC

Def. (Invariant measure)
Let P be a transition matrix for a HMC X. A sequence $(x_i)_{i\in E}$ is called an ***invariant measure*** on E w.r. to P if

1. $x_i \in [0, +\infty)$ for all $i \in E$.
2. $x^\top = x^\top P$, i.e.
$$x_i = \sum_{j\in E} x_j p_{ji}.$$

Remark We do not require the above sum to be finite, since x can be any type of measure on E.

Theorem 31 (Invariant measure and return times)
Let X be an irreducible recurrent HMC, $i_0 \in E$, and T_{i_0} as the first return time to i_0. Set now $x_i^0 = x_i^{i_0} := \mathbb{E}_i[\sum_{n\in\mathbb{N}} \mathbb{1}_{\{X_n=i\}} \cdot \mathbb{1}_{n\le T_{i_0}}]$ for all $i \in E$. $x_i^{i_0}$ is the number of visits at i prior to time T_{i_0}, then

a) $x_i^0 > 0$ *for all* $i \in E$

b) $x^0 = (x_i^0)_{i\in E}$ *is an invariant measure.*

Remark $x_{i_0}^0 = 1$ by definition of T_{i_0}

Remark

$$\sum_{i\in E} x_i^0 = \sum_{i\in E} \mathbb{E}_{i_0}[\sum_{n\in\mathbb{N}} \mathbb{1}_{\{X_n=i\}}\mathbb{1}_{n\le T_{i_0}}] = \mathbb{E}_{i_0}[\sum_{n\in\mathbb{N}}\sum_{i\in E} \mathbb{1}_{X_n=i}\mathbb{1}_{n\le T_{i_0}}] = \mathbb{E}_{i_0}[\sum_{n\in\mathbb{N}} \mathbb{1}_{n\le T_{i_0}}] = \mathbb{E}[T_{i_0}]$$

This relationship allows us to see a connection between invariant measures and recurrent Markov chains, in particular when the Markov chain is positive recurrent we have a finite total mass for the invariant measure.

Theorem 32 (Uniqueness of the invariant measure)
The invariant measure of a recurrent HMC is unique up to a multiplicative constant, furthermore the chain is positive recurrent $\iff \sum_{i\in E} x_i^0 < +\infty$.

Proof.
See definition of an invariant measure and observe that $x^\top = x^\top P$ is preserved via multiplication.
□

Theorem 33 (Recurrence and stationary distribution of an irreducible HMC)
An irreducible HMC is positive recurrent $\iff$ there exists a stationary distribution π, and in this case π is unique and componentwise strictly positive.

Proof.
$\boxed{\implies}$: Follows from previous results.

$\boxed{\impliedby}$: Recall that a stationary distribution is such that $\pi(i) = \sum_{i\in E} \pi(j)P_{ji}^n$ for all $n \in \mathbb{N}$. By contradiction, assume that the chain X is transient, then

$$\lim_{n\to\infty} P_{ji}^n = 0 \qquad \text{(potential matrix criterion)},$$

but since the global balance equation holds for every n, it also holds in the limit

$$\pi(i) = \lim_{n\to\infty} \sum_{i\in E} \pi(j)P_{ji}^n = 0.$$

□

Theorem 34 (Positive recurrence and expected value of return time)
If X is positive recurrent with stationary distribution π, then for any $i \in E$ we have

$$\pi(i) \cdot \mathbb{E}_i[T_i] = 1.$$

Proof.
We say that

$$\pi(i) = \frac{x_i^0}{\sum_{i\in E} x_i^0}$$

□

Intuitively, a finite number of states such that all of them communicate with each other leads to a positive probability of returning to each state. Hence, we have the following theorem:

Theorem 35 (Irreducibility of a finite chain)
If E is finite, any irreducible HMC is positive recurrent.

Proof.
Assume that X is transient, then by the potential matrix criterion

$$\sum_{n\in\mathbb{N}} P_{ij}^n < +\infty,$$

then since E is finite we have that

$$\sum_{j\in E}\sum_{n\in\mathbb{N}} P_{ij}^n < +\infty \implies \sum_{n\in\mathbb{N}} \underbrace{\sum_{j\in E} P_{ij}^n}_{=\,1} = \sum_{n\in\mathbb{N}} 1 = +\infty,$$

and this is a contradiction.

□

10.2 Ergodicity

Theorem 36 (Ergodic theorem)
Let X be a positive recurrent HMC and let $f : E \to \mathbb{R}$ such that f is summable w.r. to the stationary distribution, i.e.

$$\sum_{i\in E} |f(i)| \cdot \pi(i) < +\infty, \qquad \left(\int_E |f| \, \mathrm{d}\pi < \infty\right)$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(x_k) \overset{a.s.}{=} \sum_{i\in E} f(i)\pi(i). \qquad \left(\int_E f \, \mathrm{d}\pi\right)$$

Proof.
Consider $\tau_1, \tau_2, \ldots, \tau_n, \ldots$ the return times at i_0 and set $U_p := \sum_{n=\tau_p+1}^{\tau_{p+1}} f(X_n)$, it can be proved that $(U_p)_{p\in\mathbb{N}}$ is a sequence of i.i.d r.v.'s. Using the strong Markov property, we have

$$\begin{aligned}
\mathbb{E}[U_1] &\overset{\text{s.M.}}{=} \mathbb{E}_{i_0}\Big[\sum_{n=1}^{T} f(X_n)\Big] \\
&= \mathbb{E}_{i_0}\Big[\sum_{n=1}^{T}\sum_{i\in E} f(i) \cdot \mathbb{1}_{(X_n=i)}\Big] \\
&= \sum_{i\in E} f(i)\mathbb{E}_{i_0}\Big[\sum_{n=1}^{T} \mathbb{1}_{(X_n=i)}\Big] \\
&= \sum_{i\in E} f(i) \cdot \pi_i^{i_0}.
\end{aligned}$$

Now, using the strong law of large numbers the above equalities yield

$$\lim_{n\to\infty} \frac{1}{n} \sum_{p=1}^{n} U_p \overset{\text{a.s.}}{=} \sum_{i\in E} f(i)\pi_i^{i_0}.$$

By definition, the above quantity $\sum_{p=1}^{n} U_p = \sum_{k=T+1}^{\tau_n+1} f(X_k)$.

Note By definition, if $\mu(n) :=$ "number of visits to i_0 prior to n", the return time is such that $\tau_{\mu(n)} \le n \le \tau_{\mu(n)+1}$. From this, we can write (since f is positive)

$$\underbrace{\frac{\sum_{k=1}^{\tau_{\mu(n)}} f(X_k)}{\mu(n)}}_{\xrightarrow{n\to\infty} \sum_{i\in E} f(i)\pi_i^{i_0}} \le \frac{\sum_{k=1}^{n} f(X_k)}{\mu(n)} \le \underbrace{\frac{\sum_{k=1}^{\tau_{\mu(n)+1}} f(X_k)}{\mu(n)}}_{\xrightarrow{n\to\infty} \sum_{i\in E} f(i)\pi_i^{i_0}}$$

□

Remark In the statement there is no need to know in advance the value of the stationary distribution. Moreover, this result does not depend on the choice of the initial distribution ν_0.

Remark

› Setting $f \equiv \mathrm{id}$ yields $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \int_E x \, \mathrm{d}\pi$.

› Setting $f(i) = \mathbb{1}_{\{i_0\}}(i)$ yields $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_k=i_0} = \pi(i_0)$.

Lemma 7 (Ergodic theorem on the number of visits)
Let X be an irreducible recurrent HMC, and let $\mu(n)$ be the number of visits to i_0 prior to time n,

$$\mu(n) := \sum_{k=1}^{n} \mathbb{1}_{X_k=i_0}.$$

Let now $f : E \to \mathbb{R}$ measurable such that

$$\sum_{i\in E} |f(i)| \cdot x_i^{i_0} < \infty,$$

then

$$\lim_{n\to\infty} \frac{1}{\mu(n)} \sum_{k=1}^{n} f(X_n) = \sum_{i\in E} f(i) x_i^{i_0}$$

Proof.
Using the ergodic theorem, by the previous lemma we have

$$\lim_{n\to\infty} \frac{n}{\mu(n)} = \sum_{i\in E} x_i^{i_0},$$

but then

$$\lim_{n\to\infty} \frac{\mu(n)}{n} \cdot \frac{1}{\mu(n)} \sum_{k=1}^{n} f(x_n)$$

...

□

Corollary 8 (Convergence of a function of multiple states)
Let X be an irreducible positive recurrent HMC and let $g : E^{1+L} \longrightarrow \mathbb{R}$ with $L \in \mathbb{N}_0$, such that

$$\sum_{i_0,\ldots,i_L\in E} |g(i_0,\ldots,i_L)| \cdot \pi(i_0) P_{i_0 i_1} \ldots P_{i_{L-1} i_L} < \infty,$$

or equivalently $\mathbb{E}_\pi[|g(X_0,\ldots,X_L)|] < \infty$, then

$$\lim_{n\to\infty} \frac{1}{N} \sum_{k=1}^{N} g(X_k, X_{k+1}, \ldots, X_{k+L}) = \sum_{i_0,\ldots,i_L} g(i_0,\ldots,i_L)\pi(i_0) \cdot P_{i_0 i_1} \ldots P_{i_{L-1} i_L}.$$

Proof.
Consider the process $Y_n := (X_n, X_{n+1}, \dots, X_{n+L})$, then it can be proved that $Y = (Y_n)_n$ is a positive recurrent HMC with stationary distribution with stationary distribution

$$\pi(i_0) \cdot P_{i_0 i_1} \dots P_{i_{L-1}\, i_L},$$

and the conclusion follows from the ergodic theorem.

□

Lecture 11: Continuous-time stochastic processes

2021-12-10

In the previous lectures we only considered stochastic processes which are defined on both discrete state spaces and discrete time domains. Now we want to move towards continuous stochastic processes in both qualities, starting from the easier generalization, which is continuous-time discrete stochastic processes.

We can think about the easiest type of discrete continuous-time Markov processes, which are sequences of step functions where the jumps are performed at a random (countable) sequence of times $T_0, T_1, \ldots, T_n, \ldots$.

11.1 Continuous-time processes

Def. (Continuous-time stochastic process)
Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we say that a ***continuous-time stochastic process*** on $I \subset \mathbb{R}$ is a family $(X_t)_{t\in I}$ of random variables (with values on $\mathbb{R}^d$).

Measurability By definition, X_t is $\mathcal{F}$-measurable for any $t \in I$.

Typical choices Usually, $I = \mathbb{R}_{\geq 0}, I = [0, T], \ldots$

Equivalent formulation Similarly to the space of sequences of discrete stochastic processes, we can equivalently define a continuous-time stochastic process as

$$X : \Omega \longrightarrow (\mathbb{R}^d)^I = \{\text{functions } I \to \mathbb{R}^d\}$$

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

when equipped with a σ-algebra on $(\mathbb{R}^d)^I$.

Equivalent formulation II A second equivalent formulation of the continuous-time stochastic processes is the following function,

$$X : \Omega \times I \longrightarrow \mathbb{R}^d$$

$$(\omega, t) \longmapsto X_t(\omega)$$

with X a $(\mathcal{F} \otimes \mathcal{B})$-measurable function.

Def. (Natural filtration)
Given X continuous-time stochastic process, we define the ***natural filtration*** of X is given by

$$\mathcal{F}_t^X := \sigma(X_s, s \in I, s \leq t)$$

$$= \sigma\big((X_s \in H), H \in \mathcal{B}, s \in I, s \leq t\big).$$

Def. (Continuous-time martingale)
A continuous-time process X is called a ***martingale*** w.r. to a given filtration $(\mathcal{F}_t)_{t\in I}$ if

1. $X_t \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ for all $t \in I$
2. $\mathbb{E}[X_T|\mathcal{F}_t] = X_t$ for all $T \geq t$

Remark Recall that $\mathbb{E}[X_t]$ is constant w.r. to t and $\mathbb{E}[X_T|X_t] = X_t$ for all $T > t$.

The first differences between continuous-time and discrete-time stochastic processes lie in the uniqueness properties of the processes. For discrete-time, we saw two definitions of equality (in law and almost-sure), and the latter definition changes a bit.

Def. (Modifications)
X, Y continuous time stochastic processes are called ***modifications*** if they are equal at all fixed times,

$$X_t \overset{\text{a.s.}}{=} Y_t \quad \text{for all } t \in I.$$

Def. (Indistinguishability)
X, Y continuous time stochastic processes are called ***indistinguishable*** if equality holds in the sense of functions,

$$X_t \overset{\text{a.s.}}{=} Y_t,$$

i.e. if $\mathbb{P}(X_t = Y_t \text{ for all } t \in I) = 1$.

Remark Indistinguishable is stronger than modification. Indeed,

$$\begin{aligned}(X = Y) &= (X_t = Y_t \text{ for all } t \in I) \\ &= \underbrace{\bigcap_{t\in I}(X_t = Y_t)}_{\text{Not countable}}\end{aligned}$$

therefore, if the right hand side have $\mathbb{P}(\cdot) = 1$ does not imply that their intersection has probability 1.

Regularity The equivalence between modifications and indistinguishable can be restored with some regularity conditions. If X, Y are continuous then we can consider $\bigcap_{t\in I\cap\mathbb{Q}}$ and take the limit, in order to get a countable intersection. In this case, X, Y modifications $\implies$ X, Y indistinguishable.

Example

Consider the sample space $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}$, and $\mathbb{P} = \text{Unif}_{[0,1]}$. Let now

$$X_t(\omega) = \mathbb{1}_{\{t\}}(\omega) = \begin{cases} 1 & \text{if } \omega = t \\ 0 & \text{if } \omega \neq t \end{cases}$$

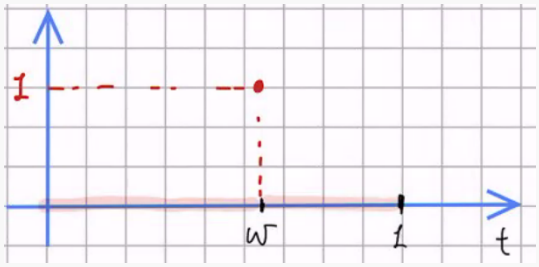

Figure 15: Example of a realization from the stochastic process X.

Consider now another trivial stochastic process $Y_t \equiv 0$ for all $t \in [0,1]$, then it is clear that for any $\omega \in [0,1]$, $X(\omega) \neq Y(\omega)$. On the other hand, for any $t \in [0,1]$ we have

$$\{X_t = Y_t\} = \Omega \setminus \{t\},$$

and therefore $\mathbb{P}(\{X_t = Y_t\}) = 1$. We conclude that X and Y are modifications, but not indistinguishable.

Def. (Stopping time)
A random time $\tau : \Omega \to [0, +\infty]$ is called a ***stopping time*** w.r. to a filtration $(\mathcal{F}_t)_{t \in I}$ if

$$\{\tau \leq t\} \in \mathcal{F}_t, \quad \text{for all } t \geq 0.$$

Interpretation The interpretation remains the usual, i.e. at each time t we must be able to know whether the event happened or not.

Remark If τ is a stopping time, then

$$\{\tau < t\} = \bigcup_{n \in \mathbb{N}} \underbrace{\left\{\tau \leq t - \frac{1}{n}\right\}}_{\in \mathcal{F}_{t-\frac{1}{n}} \subset \mathcal{F}_t} \in \mathcal{F}_t,$$

and also $\{\tau = t\} \in \mathcal{F}_t$.

The difference between discrete-time and continuous-time stochastic processes lies in hitting times.

Example (Hitting times)
Let X be a continuous-time stochastic process adapted to $(\mathcal{F}_t)_{t \geq 0}$ and let $H \subset \mathbb{R}^d$ Borel set.

Let now

$$\tau := \begin{cases} \inf\{t \geq 0 : X_t \in H\} & \text{if } \neq \emptyset \\ +\infty & \text{if } = \emptyset \end{cases}$$

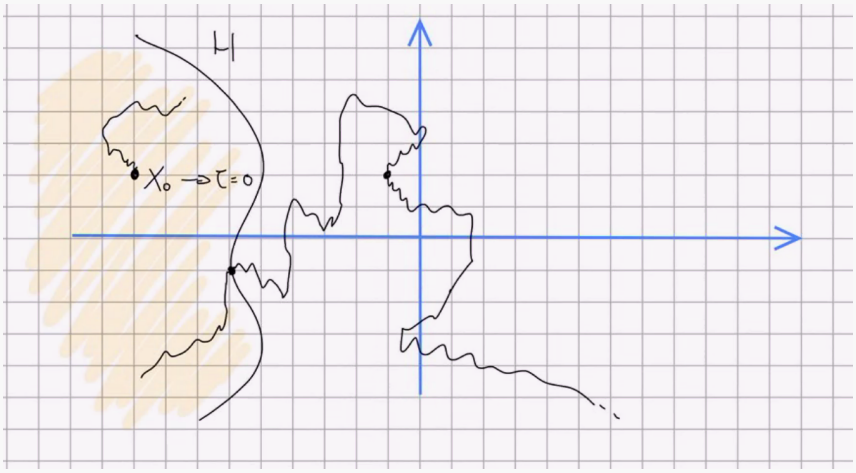

Figure 16: Example of a hitting time on a set H of a continuous-time stochastic process.

We are interested in finding out whether the time τ is a stopping time or not. We start by decomposing the event

$$\begin{aligned} \{\tau \leq t\} &= \{\tau = t\} \cup \{\tau < t\} \\ &= \{\tau = t\} \cup \Big(\bigcup_{s \in [0,t)} \{X_s \in H\} \Big) \\ &= \{\tau = t\} \cup \Big(\bigcup_{s \in [0,t) \cap \mathbb{Q}} \{X_s \in H\} \Big) \qquad \text{(continuous and } H \text{ open/closed)} \end{aligned}$$

The second set belongs to $\mathcal{F}_t$ if X has continuous paths and H is an open or closed set. On the other hand, $\{\tau = t\}$ must be considered in two cases:

1. H is a closed set, then

$$\{\tau = t\} = \underbrace{\{X_t \in H\}}_{\in \mathcal{F}_t} \cap \Big(\underbrace{\bigcap_{s \in [0,t) \cap \mathbb{Q}} \{X_s \notin H\}}_{\in \mathcal{F}_t} \Big),$$

therefore τ is a stopping time.

2. H is an open set, in this case $\{X_t \in H\} \notin \mathcal{F}_t$, and therefore we cannot use the same argument as before:

$$\{\tau = t\} = \bigcup_{\varepsilon>0} \underbrace{\{X_{t+\varepsilon} \in H\}}_{\notin \mathcal{F}_t} \cap \Big(\underbrace{\bigcap_{s\in[0,t)\cap\mathbb{Q}} \{X_s \notin H\}}_{\in \mathcal{F}_t} \Big),$$

therefore if H is an open set we must be able to see at least one step into the future in order to tell whether the event happened or not $\implies$ τ is <u>not</u> a stopping time.

Solution In order to overcome the above problem, we use a mathematical trick – that has no real probabilistic interpretation – and require the filtration $\mathcal{F}_t$ to be such ***continuous to the right***, i.e. for any $\varepsilon > 0$,

$$\bigcup_{\varepsilon>0} \mathcal{F}_{t+\varepsilon} = \mathcal{F}_t.$$

11.2 Point and counting processes

The goal here is to define a counting process whose trajectories are step functions, which will be a very useful tool for analyzing continuous-time processes.

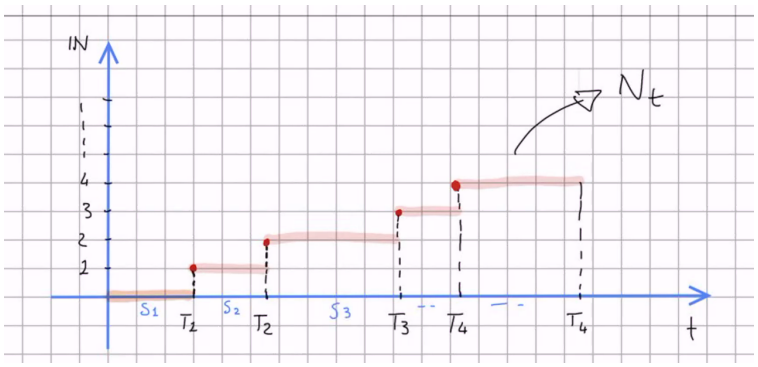

Figure 17: Example of a counting process.

The important thing is that the jump times of the process $T_1, T_2, \ldots, T_n, \ldots$ are random variables $T_i \in \mathbb{R}$, which requires the development of a more articulated theory of stochastic processes.

Notation We denote by $S_n = T_n - T_{n-1}$ the sequence of ***inter-arrival times*** of the process, and by

$$N(a, b] = \sum_{n\in\mathbb{N}} \mathbb{1}_{(a,b]} T_n$$

the ***number of events*** that occurred in the interval $(a, b]$.

Def. (Point process)
A stochastic ***point process*** is a discrete time stochastic process $T = (T_n)_{n\in\mathbb{N}_0}$ such that

i. $T_0 \equiv 0$

ii. $T_0 < T_1 < T_2 < \ldots < T_n < \ldots$

iii. $\lim_{n\to\infty} T_n = +\infty$

Remark Some relaxations have been considered in literature, especially properties (*ii*) and (*iii*) could be replaced by:

$\tilde{ii}$. $T_0 \leq T_1 \leq T_2 \leq \ldots \leq T_n \leq \ldots$ (*Multiple arrivals*)

$i\tilde{ii}$. Removed (*Explosion*)

Intuition The idea behind the point process is to give a set of indices at which something happens to the stochastic process.

Def. (Counting process)
If T is a point process, then the process $N = (N_t)_{t\geq 0}$ defined by

$$N_t := N(0,t] = \sum_{n\in\mathbb{N}_0} \mathbb{1}_{(0,t]}(T_n), \quad t \geq 0$$

is called the ***counting process*** of the point process T.

Properties

› $N_0 = 0$ by definition, since the indicator function excludes 0.

› $N_t \nearrow$ and $N_t \in \mathbb{N}_0$ since the set in the indicator function is increasing.

› The trajectories $t \longmapsto N_t$ are ***càdlag***, i.e. continuous from the right and have limit from the left (see Figure 17).

Def. (Homogeneous Poisson process)
A ***homogeneous Poisson process*** with intensity $\lambda > 0$ is a counting process $(N_t)_{t\geq 0}$ such that

a) For any $k \in \mathbb{N}$ and any finite selection $0 \leq t_1 \leq t_2 \leq \ldots \leq t_k$, increments are independent:

$$(N_{t_k} - N_{t_{k-1}}) \perp\!\!\!\perp (N_{t_{k-1}} - N_{t_{k-2}}) \perp\!\!\!\perp \ldots \perp\!\!\!\perp (N_{t_2} - N_{t_1})$$

b) For any $0 \leq t \leq T$, $N_T - N_t \sim \text{Pois}\big(\lambda(T-t)\big)$.

Remark Using (b) and $(N_0 = 0) \implies N_t \sim \text{Pois}(\lambda t)$ for any $t \in \mathbb{R}$.

Theorem 37 (Characterization of a Poisson process)
Let $N = (N_t)_{t\geq 0}$ be a counting process, then

$$N \text{ is a Poisson process with intensity } \lambda > 0 \iff (T_n - T_{n-1})_{n\in\mathbb{N}} \overset{iid}{\sim} Exp(\lambda).$$

Construction The above theorem lets us construct a Poisson process by starting with a sequence of i.i.d Exp(λ) waiting times.

Proof.
The proof relies strongly on the independence and the loss of memory of the exponential distribution. □

Lemma 8 (Distribution of the jump times)
If $(S_n)_{n\in\mathbb{N}}$ is exponentially distributed with parameter λ, then $T_n \in AC$ and with density

$$f_{T_n}(t) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}\mathbb{1}_{[0,\infty)}(t).$$

Proof.
For $n = 1$, by definition $T_1 = S_1$ and the result is trivial. By induction, we have

$$f_{T_{n+1}}(t) = f_{T_n+S_{n+1}}(t) \overset{\perp\!\!\!\perp}{=} \int_{\mathbb{R}} f_{T_n}(s) f_{S_{n+1}}(t-s)\,\mathrm{d}s = \ldots = \lambda e^{-\lambda t}\frac{(\lambda t)^n}{n}\mathbb{1}_{[0,\infty)}(t).$$

□

Now, we can use the above lemma to show that $N_t \sim \text{Pois}(\lambda t)$, since

$$\mathbb{P}(t \geq T_{n+1}) \overset{\text{L.8}}{=} \int_0^t \lambda e^{-\lambda s}\frac{(\lambda s)^n}{n!}\,\mathrm{d}s \overset{\text{parts}}{=} \mathbb{P}(t \geq T_n) - e^{-\lambda t}\frac{(\lambda t)^n}{n!},$$

we can observe from Figure 17 that in order to have $N_t = n$ we require:

$$\begin{aligned}\mathbb{P}(N_t = n) &= \{n^{\text{th}} \text{ arrival before } t \text{ and } (n+1)^{\text{th}} \text{ after } t\}\\ &= \mathbb{P}(T_n \leq t) - \mathbb{P}(T_{n+1} \leq t)\\ &= e^{-\lambda t}\frac{(\lambda t)^n}{n!}.\end{aligned}$$

Other important properties

1. For any fixed $t > 0$, N is continuous at t with probability 1 (consequence of Lemma 8).
2. $\mathbb{E}[T_n - T_{n-1}] = \mathbb{E}[S_n] = \frac{1}{\lambda}$, therefore λ controls both the expected waiting time and the average number of arrivals in the time unit ($\mathbb{E}[N_t] = \lambda t$).

Lecture 12: Continuous stochastic processes

2021-12-16

We discuss the compound Poisson process, which instead of jumping of a unitary value we have jumps at $T_1, T_2, \ldots, T_k, \ldots$ of random size $Z_1, Z_2, \ldots, Z_k, \ldots$

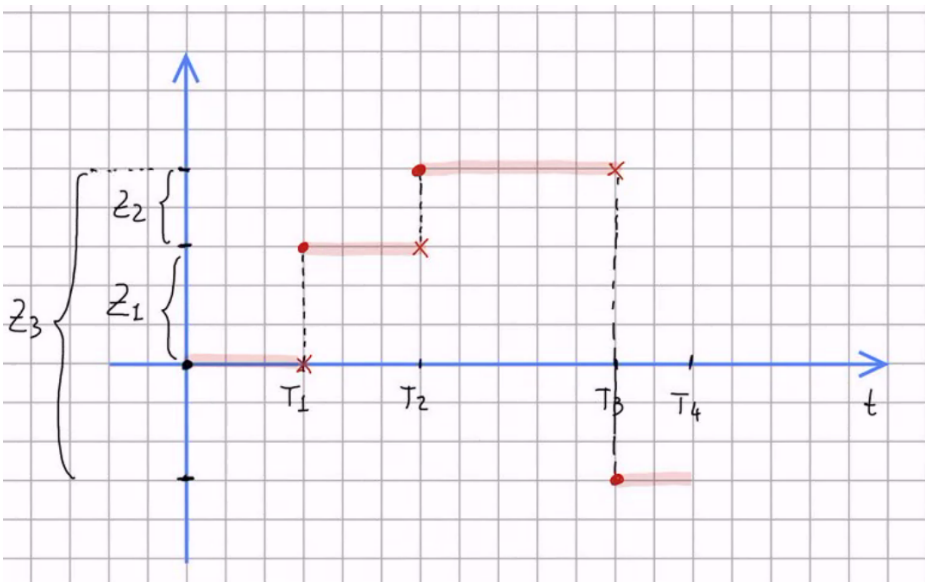

Figure 18: Example of a compound Poisson process.

Def. (Compound Poisson process)
Let $N = (N_t)_{t\geq 0}$ be a Poisson process with parameter λ, and let also $(Z_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d r.v.'s, and the whole sequence independent from N. The process set as

$$X_t = \sum_{k=1}^{N_t} Z_k, \quad t \geq 0$$

is called a ***compound Poisson process***.

Remark If $Z_k \sim \delta_1$, then X is a homogeneous Poisson process.

Markov property In the continuous-time case, the Markov property reads

$$\mathbb{E}[\varphi(X_T)|\mathcal{F}_t] = \mathbb{E}[\varphi(X_T)|X_t], \quad \text{for all } 0 \leq t \leq T \text{ and } \varphi \text{ bounded and measurable.}$$

Prop. 11 (Poisson process is Markov)
The Poisson process N has the Markov property w.r. to its natural filtration $(\mathcal{F}_t^N)_{t\geq 0}$.

Proof.

We start by writing $N_T = N_t + \overbrace{(N_T - N_t)}^{\text{increment}}$, and since N_t is measurable w.r. to $\mathcal{F}_t^N$ and $(N_T - N_t)$ is independent from the

$$\mathbb{E}[\varphi(N_T)|\mathcal{F}_t^N] \overset{\text{Freez.}}{=} \mathbb{E}[\varphi(n + N_T - N_t)]\big|_{n=N_t},$$

which is a function of N_t, and therefore has the Markov property. Moreover, we also have the regression function

$$\mathbb{E}[\varphi(N_T)|N_t = n] = \mathbb{E}[\varphi(n + Z)], \quad \text{with } Z \sim \text{Pois}\big(\lambda(T-t)\big).$$

□

Prop. 12 (Compound Poisson process is Markov)
A compound Poisson process X has the Markov property w.r. to its natural filtration $(\mathcal{F}_t^X)_{t\geq 0}$.

Proof.
Just like before, we can write $X_T = X_t + (X_T - X_t)$, and the increment is such that

$$X_T - X_t = \sum_{k=N_t+1}^{N_T} Z_k = \sum_{j=1}^{N_T - N_t} Z_{N_t+j},$$

and since Z_n are independent this quantity is independent of $\mathcal{F}_t^X$. Therefore, by the freezing lemma the process is again Markov:

$$\begin{aligned}\mathbb{E}[\varphi(X_T)|\mathcal{F}_t^X] &= \mathbb{E}[\varphi(x + \sum_{j=1}^{N_T-N_t} Z_{n+j})]\big|_{n=N_t, x=X_t} \\ &= \mathbb{E}[\varphi(x + \sum_{j=1}^{N_T-N_t})Z_j]\big|_{x=X_t} \qquad \text{(indep.)}\end{aligned}$$

□

12.1 General continuous-time Markov chains

We are going to extend the concept of transition matrix to the continuous case, by describing the evolution of the process in terms of some sort of derivative of the process.

Def. (Continuous-time discrete Markov chain)
A stochastic process $(X_t)_{t\geq 0}$ with values on E, $|E| = |\mathbb{N}|$, is a ***continuous-time discrete Markov chain*** if it has the Markov property w.r. to its natural filtration $(\mathcal{F}_t^X)_{t\geq}$.

Prop. 13

A stochastic process $(X_t)_{t\geq 0}$ with values on E is a continuous-time MC $\iff$

$$\mathbb{P}(X_{t+s} = j | X_s = i, X_{s_k} = i_k, \dots, X_{s_0} = i_0) = \mathbb{P}(X_{t+s} = j | X_s = i)$$

for all $t > 0$ and $0 \leq s_0 < s_1 < s_k < s$ and $i_0, \dots, i_k, i, j \in E$ such that the probability is well-defined.

Def. (Homogeneous Markov Chain)

if $\mathbb{P}\,()$ does not depend on s, then the MC X is called ***homogeneous*** (HMC).

Remark For any $t \geq 0$, let $P(t) := (P_{ij}(t))_{i,j\in E}$ as defined by

$$P_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i) \overset{\text{HMC}}{=} \mathbb{P}(X_{t+s} = j | X_s = i).$$

This object is usually called the ***transition semigroup*** of the Markov chain.

Properties

1. $P(0) = I_{|E|}$
2. For all $t > 0$, $\sum_{k\in E} P_{ik}(t) = 1$ for all $i \in E$.
3. $P(t+s) = P(t) \cdot P(s)$ (*Chapman-Kolmogorov*)

$$\begin{aligned}
P_{ij}(t+s) &= \mathbb{P}(X_{t+s} = j | X_0 = i) \\
&= \sum_{k\in E} \mathbb{P}(X_{t+s} = j, X_s = k | X_0 = i) \\
&= \mathbb{P}(X_{t+s} = j | X_s = j, X_0 = i) \cdot \mathbb{P}(X_s = k | X_0 = i) \\
&= \mathbb{P}(X_{t+s} = j | X_s = k) \cdot \mathbb{P}(X_s = k | X_0 = i) \qquad \text{(Markov)} \\
&= \big(P(t) \cdot P(s)\big)_{ij}
\end{aligned}$$

Remark Denote by $\nu(t)$ the distribution of X_t, then

1. $\nu(t)^\top = \nu(0)^\top \cdot P(t)$
2. $\mathbb{P}(X_0 = i_0, X_{t_1} = i_1, \dots, X_{t_k} = i_k) = \nu_{i_0}(0) \cdot P_{i_0 i_1}(t_1) \cdot \ldots \cdot P_{i_{k-1} i_k}(t_k - t_{k-1})$.

Example (H.P.P)

We can cast a homogeneous Poisson process $(N_t)_{t\geq 0}$ into this framework: if $j \geq i$, we have that the transition $i \to j$ has probability

$$\mathbb{P}(N_{t+h} = j | N_t = i) = \mathbb{P}(N_{t+h} - N_t = j - i | N_t = i),$$

and we know that the increments are $N_{t+h} - N_t \sim \text{Pois}(\lambda h)$, therefore

$$\mathbb{P}(N_{t+h} = j | N_t = i) = \begin{cases} 0 & \text{if j < i} \\ e^{-\lambda h} \frac{(\lambda h)^{j-i}}{(j-i)!} & \text{if } j \geq i \end{cases}$$

Example

Let N be a H.P.P set

$$X_t = (-1)^{N_t}, \quad E = \{-1, 1\}.$$

Then, we have that

$$\begin{aligned} \mathbb{P}(X_{t+s} = 1 | X_s = -1) &= \mathbb{P}\big((-1)^{N_s} \cdot (-1)^{N_{t+s} - N_s} = 1 | X_s = -1\big) \\ &= \mathbb{P}\big((-1)^{N_{t+s} - N_s} \cdot (-1) = 1\big) \\ &= \mathbb{P}(N_{t+s} - N_s \text{ is odd}) \\ &= \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{2n+1}}{(2n+1)!} \\ &= e^{-\lambda t} \cdot \frac{1}{2}(e^{\lambda t} - e^{-\lambda t}) \\ &= \frac{1}{2}(1 - e^{-2\lambda t}). \end{aligned}$$

Then, we have

$$P(t) = \frac{1}{2} \begin{pmatrix} 1 + e^{-2\lambda t} & 1 - e^{-2\lambda t} \\ 1 - e^{-2\lambda t} & 1 + e^{-2\lambda t} \end{pmatrix}.$$

When we have functions defined on intervals, one of the questions that arises is about the regularity of the function, i.e. continuity, derivatives, ...

12.2 Regularity of $P(t)$

Def. (Continuity)

Let $(P(t))_{t \geq 0}$ be a transition semigroup (satisfies prop. 1-2-3), then we say that P is ***continuous*** if it is continuous at $t = 0$,

$$\lim_{h \to 0^+} P(h) = P(0) = I.$$

Prop. 14 (Continuity of the semigroup)

If $P(t)$ is continuous at $t = 0$, then it is continuous for every $t \geq 0$.

Def.
The ***generator*** of $P(\cdot)$ is its derivative

$$A := \lim_{h\to 0^+} \frac{P(h) - P(0)}{h} = P'(h).$$

Remark If $A = (q_{ij})_{i,j\in E}$, then we can explicitly write

$$q_{ij} = \begin{cases} \lim_{h\to 0^+} \frac{P_{ij}(h)}{h} & \text{if } i \neq j \\ \lim_{h\to 0^+} \frac{P_{ii}(h) - 1}{h} & \text{if } i = j \end{cases}$$

Prop. 15 (Differentiability of $P(t)$)

Consider for $h > 0$ the right derivative at time t, we can write

$$\lim_{h\to 0^+} \frac{P(t+h) - P(t)}{h} = \lim_{h\to 0^+} \frac{P(t)P(h) - P(t)}{h} = P(t) \lim_{h\to 0^+} \frac{\overbrace{P(h) - I}^{\xrightarrow{h\to 0^+} A}}{h} = P(t) \cdot A.$$

As for the left derivative, $h < 0$, we can use the semigroup property to write

$$\begin{aligned} \lim_{h\to 0^-} \frac{P(t+h) - P(t)}{h} &= \lim_{h\to 0^-} \frac{P(t+h) - P(t+h)P(-h)}{h} = \lim_{h\to 0^-} \frac{\overbrace{P(t+h)}^{\xrightarrow{h\to 0^-} P(t)}}{} \frac{I - P(-h)}{h} \\ &= P(t) \lim_{-h\to 0^+} \frac{-I + P(-h)}{-h} \\ &= P(t) \cdot A. \end{aligned}$$

Theorem 38 (Kolmogorov's forward equation)
The above calculations show that P is such that the following ODE holds:

$$P'(t) = P(t) \cdot A,$$

*which is defined with an initial condition $P(0) = I$, and thus called a **forward** equation.*

Remark If E is finite, then A is a proper matrix and the ODE can be solved explicitly, and is the ***matrix exponential***

$$P(t) = e^{tA} = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!}.$$

12.3 Continuous Markov processes

We now briefly discuss continuous Markov processes and how we can construct them by integrating two special Markov processes: a Brownian motion and a compound Poisson process.

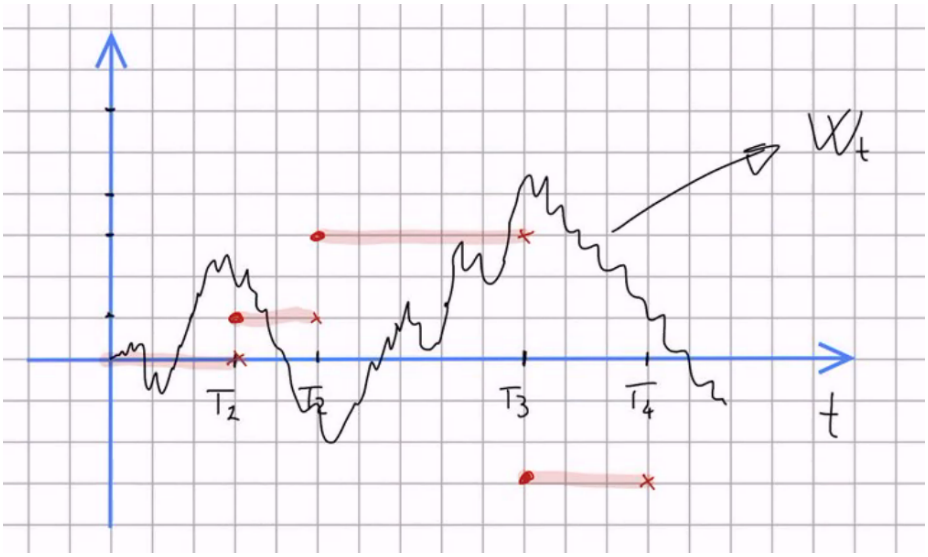

Figure 19: continuousTimeMarkovProcessExample

Consider $X_t = \sum_{k=1}^{N_t} Z_k$ compound Poisson process, and consider the following process:

$$Y_t := \int_0^t g(Y_s)\,\mathrm{d}X_s \iff \mathrm{d}Y_t = g(Y_t)\,\mathrm{d}X_t.$$

This defines a stochastic differential equation, and the integral has to be interpreted as a Riemann-Stieltjes integral.

Riemann integral $\int_0^t g(s)\,\mathrm{d}s = \lim_{h\to 0}\sum_{i=1}^{n} g(s_i)(s_{i+1}-s_i)$, where $(s_i)_{i\in\{1,\dots,n\}}$ is an equispaced grid of increment h.

Riemann-Stieltjes We replace the $\mathrm{d}s$ with a $\mathrm{d}F(x)$, which we interpret as

$$\lim_{h\to 0}\sum_{i=1}^{n} g(s_i)\big(F(s_{i+1}) - F(s_i)\big)$$

Since the trajectories of X_t are step functions, the result turns out to be the values of Y at the jump points, i.e. where $N_s - \lim_{u\to s^-} N_u \neq 0$. Therefore, we can construct the Riemann-Stieltjes integral as

$$Y(t) = \int_0^t Y(s)\,\mathrm{d}X_s = \sum_{\substack{s\in[0,t]:\\ N_s - N_{s^-}\neq 0}} g(Y_s)(X_s - X_{s^-}).$$

This process is more general but has still step-functions as realization paths. In order to do so, we have to replace the integrator X_s with a Brownian motion $X(s)$.

Def. (Brownian motion)
A process $(W_t)_{t\geq 0}$ with values on $\mathbb{R}$ defined on a is called a ***Brownian motion*** if

i. $W_0 = 0$

ii. W is continuous almost-surely.

iii. W is adapted to $(\mathcal{F}_t)_{t\geq 0}$

iv. $W_t - W_s \perp \mathcal{F}_s$ for all $0 \leq s < t$

v. $W_t - W_s \sim \mathcal{N}(0, t-s)$ for any $0 \leq s < t$.

Existence It's possible to construct a stochastic process such that the above definition is satisfied, but we don't show it explicitly.

Last property Since $\mathbb{E}[(W_t - W_s)^2] = t - s$, the last property is sometimes denoted by $\Delta W_t \sim \sqrt{\Delta t}$.

Remarks

› W is a martingale

› W has the Markov property (adapted + indep. increments + freezing).

› $W_t = W_t - W_0 \sim \mathcal{N}(0, t)$.

› $W_t - W_s \sim \mathcal{N}(0, t-s)$ and therefore we can write the ***transition density*** of the Brownian motion as

$$p(t, x; T, y) = \frac{1}{\sqrt{2\pi(T-t)}} e^{-\frac{1}{2}\frac{(y-x)^2}{T-t}}.$$

This is the density of a $\mathcal{N}(x, T-t)$ and gives the density of the conditional law of $\mu_{W_T|W_t=x}$.

› Fixing $(T, y) \in \mathbb{R} \times \mathbb{R}$, $p(\cdot, \cdot; T, y)$ solves the following partial differential equation

$$\partial_t p(t, x; T, y) + \frac{1}{2}\partial_{xx} p(t, x; T, y) = 0, \tag{5}$$

which is called the ***Kolmogorov backward equation*** and the operator $A := \frac{1}{2}\partial_{xx}$ is called the generator of W. For a given continuous function $\varphi \in C_b(\mathbb{R})$ we can set

$$u(t, x; T) := \mathbb{E}[\varphi(X_T)|X_t = x] = \int_{\mathbb{R}} \varphi(y) \underbrace{p(y, x; T, y)}_{\text{solves (5)}} \, \mathrm{d}y,$$

and that $\mathcal{N}(x, T-t) \xrightarrow{d} \delta_x$ for $t \to T$, then this density solves the backward Cauchy problem

$$\begin{cases} \partial_t u(t, x; T) + \frac{1}{2}\partial_{xx} u(t, x; T) = 0 \quad \text{for } t < T \\ u(T, x; T) = \varphi(x) \end{cases}$$

Problem The idea for constructing a continuous-time stochastic process is to obtain

$$X_t = X_0 + \int_0^t \sigma(X_s)\,\mathrm{d}W_s,$$

however for the Riemann-Stieltjes we require W to have bounded variation. For the Brownian motion, however, since $W(w)$ has unbounded total variation we cannot simply solve this integral in a direct way, which causes the partial sums of the Riemann-Stieltjes integral partial to not converge.

Prop. 16

The quadratic variation of the Brownian motion is bounded in L^2, i.e.

$$\sum_{i=1}^{N} |W_{t_i} - W_{t_{i-1}}|^2 \xrightarrow{|\Delta| \mapsto 0^+} t, \quad \text{in } L^2 \text{ for all } t.$$

One can therefore define a *stochastic integral* for any process $u_s \in \mathbb{L}^2_{\text{loc}}$, where

$$\mathbb{L}^2_{\text{loc}} = \{(u_s)_s \text{ stochastic processes s.t. } \int_0^t u_s^2\,\mathrm{d}s < \infty\},$$

as

$$\int_0^t u_s\,\mathrm{d}W_s, \quad t \in [0,T],$$

which is not however a Riemann-Stieltjes integral but a more complicated construction.

Once we have this defined, we can consider integral equations of the form (*diffusion process*)

$$X_t = X_0 + \underbrace{\int_0^t \mu(s, X_s)\,\mathrm{d}s}_{\text{Lebesgue integral}} + \underbrace{\int_0^t \sigma(s, X_s)\,\mathrm{d}W_s}_{\text{stochastic integral}}, \tag{$*$}$$

or equivalently in differential notation,

$$\mathrm{d}X_t = \mu(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t,$$

which yields the most general form of Markov process we can think of:

- Markov property
- Continuous trajectories

The link between stochastic differential equations, partial differential equations, and Markov processes is due to Ito's lemma.

Lemma 9 (Ito's lemma)

If X solves the SDE $(*)$, let $f \in C^{1,2}\big([0,T] \times \mathbb{R}\big)$, which means that $\partial_{tf}, \partial_{xf}, \partial_{xxf}$ are continuous. Then, the process obtained by composing f with X, is still a diffusion and is such that

$$\mathrm{d}f(t, X_t) = (\partial_t + A_t)f(t, X_t)\,\mathrm{d}t + \sigma(t, X_s) \cdot \partial_x f(t, X_t)\,\mathrm{d}W_t,$$

In the lemma above, the operator A_t is the generator of the process X_t and is given by

$$A_t = \mu(t,x)\partial_x + \frac{1}{2}\sigma^2(t,x)\partial_{xx}.$$

Example
If $\mu \equiv 0$ and $\sigma equiv 1$, then X_t is a Brownian motion and

$$\mathrm{d}X_t = \mathrm{d}W_t \implies A_t = \frac{1}{2}\partial_{xx},$$

which means that A_t kind of extends the generator of a Brownian motion.

Theorem 39
Under suitable assumptions, setting

$$u(t,x;T) = \mathbb{E}[\varphi(X_T)|X_t = x],$$

we have that u solves the following backward Cauchy problem

$$\begin{cases} (\partial_t + A_t)u(t,x;T) = 0 \quad \text{if } t < T \\ u(T,x;T) = \varphi(x) \end{cases}$$

Proof.
(Application of Ito's lemma.)

□

References

Bass, R. F. (2011). *Stochastic Processes*. Cambridge ; New York: Cambridge University Press.

Brémaud, P. (2020). *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Second. Springer Nature.

Çinlar, E. (2011). *Probability and Stochastics*. Vol. 261. Graduate Texts in Mathematics. New York.

Gut, A. (2009). *An Intermediate Course in Probability*. 2° edizione. Springer Nature.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Pub.

Paolella, M. S. (2007). *Intermediate Probability: A Computational Approach*. Wiley-Interscience.