Probabilistic Numerical Methods 2024–2025

# Lecture 2: Markov chains and the Monte Carlo method

julien.reygner@enpc.fr

## 1 Conditional expectation and distribution

We work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

### 1.1 Conditional expectation in the discrete space

Recall that for any $A, B \in \mathcal{A}$ such that $\mathbb{P}(B) > 0$, we define $\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B)$. Then $\mathbb{P}(\cdot|B)$ is a probability measure on $(\Omega, \mathcal{A})$ but also on $(B, \mathcal{A}_B)$ where the $\sigma$-field $\mathcal{A}_B := \{A \cap B : A \in \mathcal{A}\}$ is called the *trace* of $\mathcal{A}$ on $B$. This allows to see the conditional probability $\mathbb{P}(\cdot|B)$ as the *restriction* of $\mathbb{P}(\cdot)$ to $B$, with normalisation constant $\mathbb{P}(B)$ ensuring that it remains a probability measure.

Now fix $X \in \mathbf{L}^1(\mathbb{P})$ and define $\mathbb{E}[X|B] := \mathbb{E}[X\mathbb{1}_B]/\mathbb{P}(B)$ so that $\mathbb{P}(A|B) = \mathbb{E}[\mathbb{1}_A|B]$.

For a random variable $Z$ taking its values in some discrete space[1] $(F, \mathcal{F})$, set $F_Z := \{z \in F : \mathbb{P}(Z = z) > 0\}$. For any $z \in F_Z$, define $\varphi_X(z) := \mathbb{E}[X|Z = z]$. Then the random variable $\mathbb{E}[X|Z] := \varphi_X(Z)$ is well-defined, almost surely.

**Proposition 1.1** (Properties of conditional expectation)**.** *(i) 'Total expectation formula':* $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Z]]$.

*(ii) For any measurable*[2] *and bounded* $\psi : F \to \mathbb{R}$, $\mathbb{E}[\psi(Z)X|Z] = \psi(Z)\mathbb{E}[X|Z]$, *almost surely.*

The second point extends to all functions $\psi$ such that $\psi(Z)X \in \mathbf{L}^1(\mathbb{P})$.

**Exercise 1.2.** *Show that if* $X \in E$ *and* $Z \in F$ *are independent, then for any measurable* $g : E \times F \to \mathbb{R}$ *such that* $g(X, Z) \in \mathbf{L}^1(\mathbb{P})$, $\mathbb{E}[g(X, Z)|Z] = G(Z)$, *almost surely, where* $G(z) := \mathbb{E}[g(X, z)]$.

Combining the points (i) and (ii) of Proposition 1.1 yields the following statement: for any measurable and bounded function $\psi : F \to \mathbb{R}$,

$$\mathbb{E}[X\psi(Z)] = \mathbb{E}[\mathbb{E}[X|Z]\psi(Z)]. \tag{1}$$

If you think of $(X, Y) \mapsto \mathbb{E}[XY]$ as a scalar product in $\mathbf{L}^2(\mathbb{P})$, the identity above shows that $X - \mathbb{E}[X|Z]$ is orthogonal to the space of random variables of the form $\psi(Z)$, so that $\mathbb{E}[X|Z]$ is actually the orthogonal projection of $X$ on this space. This identity is the basis of the generalisation of the construction.

### 1.2 Conditional expectation in the general case

We now let $Z$ be a random variable in a measurable space $(F, \mathcal{F})$ which no longer needs to be discrete.

[1] This means that $F$ is finite or countably infinite and $\mathcal{F}$ is the power set of $F$.

[2] Note that since $F$ is discrete, all functions are measurable.

**Theorem 1.3** (Definition of conditional expectation)**.** *For any $X \in \mathbf{L}^1(\mathbb{P})$, there exists a measurable function $\varphi_X : F \to \mathbb{R}$ such that $\varphi_X(Z) \in \mathbf{L}^1(\mathbb{P})$ and for any measurable and bounded function $\psi : F \to \mathbb{R}$,*

$$\mathbb{E}[X\psi(Z)] = \mathbb{E}[\varphi_X(Z)\psi(Z)]. \tag{2}$$

*If there is another function $\widetilde{\varphi}_X$ with the same properties then $\varphi_X(Z) = \widetilde{\varphi}_X(Z)$, almost surely. This ensures that the random variable*

$$\mathbb{E}[X|Z] := \varphi_X(Z)$$

*is well-defined, almost surely.*

Notice that (2) is the same identity as (1). Moreover, the statements of Proposition 1.1 and Exercise 1.2 remain true with this general definition.

## 1.3 Conditional distribution

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces.

**Definition 1.4** (Markov kernel)**.** *A* Markov kernel *from $F$ to $E$ is a map $P : F \times \mathcal{E} \to [0, 1]$ such that:*

*(i) for any $z \in F$, $C \in \mathcal{E} \mapsto P(z, C)$ is a probability measure;*
*(ii) for any $C \in \mathcal{E}$, $z \in F \mapsto P(z, C)$ is measurable.*

**Definition 1.5** (Conditional distribution)**.** *Given random variables $X \in E$, $Z \in F$ and a Markov kernel $P$ from $F$ to $E$, $P(Z, \cdot)$ is a* conditional distribution of $X$ given $Z$ *if, for any measurable and bounded function $f : E \to \mathbb{R}$,*

$$\mathbb{E}[f(X)|Z] = \int_{x\in E} f(x)P(Z, \mathrm{d}x), \qquad \textit{almost surely.}$$

Equivalently, for any measurable and bounded function $g : E \times F \to \mathbb{R}$,

$$\mathbb{E}[g(X, Z)] = \int_{z\in F} \left( \int_{x\in E} g(x, z)P(z, \mathrm{d}x) \right) \mu_Z(\mathrm{d}z),$$

where $\mu_Z$ is the marginal distribution of $Z$. Denoting by $\mu_{(X,Z)}$ the joint law of the pair $(X, Z)$, we rewrite this identity in the short-hand notation

$$\mu_{(X,Z)}(\mathrm{d}x\mathrm{d}z) = \mu_Z(\mathrm{d}z)P(z, \mathrm{d}x),$$

which we wall a 'disintegration' formula.

As one may expect, the conditional expectation of $X \in \mathbf{L}^1(\mathbb{P})$ can be recovered from its conditional distribution from the formula

$$\mathbb{E}[X|Z] = \int_{x\in E} xP(Z, \mathrm{d}x),$$

where the identity holds almost surely.

**Theorem 1.6** (Existence of a conditional distribution[3])**.** *If $E$ is a Polish space[4] and $\mathcal{E}$ is its Borel $\sigma$-field, then $X$ always admits a conditional distribution $P(Z, \cdot)$, which is unique almost surely.*

[3] See Theorem 6.3 in Kallenberg, *Foundations of Modern Probability*, second edition.

[4] A Polish space is a topological space which is separable (it admits a dense and countable subset) and whose topology is induced by a metric making it complete.

From now on we will simply call any Markov kernel $P$ which satisfies the conclusion of Theorem 1.6 'the' conditional distribution of $X$ given $Z$.

We finally introduce a few notations: if $P$ is a Markov kernel from $F$ to $E$,

- for any measurable and bounded function $f : E \to \mathbb{R}$ we define the measurable and bounded function $Pf : F \to \mathbb{R}$ by $Pf(z) = \int_{x\in E} P(z, \mathrm{d}x)f(x)$;
- for any bounded measure $\mu$ on $F$, we define the bounded measure $\mu P$ on $E$ by $\mu P(C) = \int_{z\in F} \mu(\mathrm{d}z)P(z, C)$.

# 2 Markov property and stationary distribution

We fix a Polish space $E$ endowed with its Borel $\sigma$-field $\mathcal{E}$.

## 2.1 Markov property

**Definition 2.1** (Markov property)**.** *A sequence $(X_n)_{n\geq 0}$ of random variables in $E$ has the* Markov property *if for any $n \geq 0$, for any $A \in \mathcal{E}$,*

$$\mathbb{P}(X_{n+1} \in A|X_0, \dots, X_n) = \mathbb{P}(X_{n+1} \in A|X_n), \qquad \textit{almost surely.}$$

*A sequence with the Markov property is called a* Markov chain.

Denoting by $P_{n+1}$ the conditional distribution of $X_{n+1}$ given $X_n$, and by $\mu_{0:n}$ the joint distribution of $(X_0, \dots, X_n)$, the Markov property yields the disintegration formula

$$\mu_{0:n}(\mathrm{d}x_0 \cdots \mathrm{d}x_n) = \mu_{0:n-1}(\mathrm{d}x_0 \cdots \mathrm{d}x_{n-1})P_n(x_{n-1}, \mathrm{d}x_n).$$

Iterating this formula we get

$$\mu_{0:n}(\mathrm{d}x_0 \cdots \mathrm{d}x_n) = \mu_0(\mathrm{d}x_0)P_1(x_0, \mathrm{d}x_1) \cdots P_n(x_{n-1}, \mathrm{d}x_n),$$

where $\mu_0 := \mathrm{Law}(X_0)$, which shows that the law of $(X_0, \dots, X_n)$ is characterised by the law of the initial distribution $\mu_0$ and the sequence of Markov kernels $(P_n)_{n\geq 1}$, which are also called *transition kernels*. We also deduce the recursive identity $\mu_{n+1} = \mu_n P_{n+1}$ for the marginal distribution $\mu_n$ of $X_n$.

**Exercise 2.2** (Autoregressive model)**.** *Let $(\alpha_n)_{n\geq 1}$ and $(\sigma_n)_{n\geq 1}$ be two sequences of real numbers. Let $X_0$ be a random variable in $\mathbb{R}$ and $(G_n)_{n\geq 1}$ be a sequence of iid $\mathcal{N}(0, 1)$ random variables, independent from $X_0$. For any $n \geq 0$, define*

$$X_{n+1} = \alpha_{n+1}X_n + \sigma_{n+1}G_{n+1}.$$

*Show that $(X_n)_{n\geq 0}$ is a Markov chain and describe its transition kernels.*

## 2.2 Homogeneous chains and stationary distribution

A Markov chain is called *homogeneous* if its transition kernel does not depend on $n$.

**Definition 2.3** (Stationary distribution)**.** *A probability measure $\pi$ on $E$ is a* stationary distribution *for the homogeneous Markov chain $(X_n)_{n\geq 0}$ with transition kernel $P$ if it satisfies the identity $\pi = \pi P$.*

The notion of stationary distribution for a Markov chain only depends on its transition kernel and not on its initial distribution. However if a chain $(X_n)_{n\geq 0}$ with transition kernel $P$ has a stationary initial distribution $X_0 \sim \pi$, then $X_n \sim \pi$ for all $n \geq 0$.

**Exercise 2.4** (Autoregressive model, continued)**.** *In the setting of Exercise 2.2, assume that for all $n \geq 0$, $\alpha_n = \alpha \in (-1, 1)$ and $\sigma_n = \sigma \neq 0$. Find a stationary distribution for the chain $(X_n)_{n\geq 0}$.*

# 3 Ergodic theory of Markov chains in discrete spaces

From now on we assume that $E$ is discrete, and only consider homogeneous Markov chains. Then a Markov kernel $P$ can (and will) actually be described by a matrix $(P(x,y))_{x,y\in E}$ with nonnegative coefficients and such that for any $x \in E$, $\sum_{y\in E} P(x,y) = 1$, so that if $(X_n)_{n\geq 0}$ has transition matrix $P$, then $P(x,y) = \mathbb{P}(X_1 = y | X_0 = x)$.

Such a matrix is called *stochastic*. Its rows are probability measures on $E$, and therefore we will take the convention to consider measures on $E$ as row vectors (indexed by $E$), and functions on $E$ as column vectors. The notation introduced at the end of Section 1.3 now makes sense as matrix/vector products. In particular, for any $n \geq 1$, we then have $P^n(x,y) = \mathbb{P}(X_n = y | X_0 = x)$.

Representation of stochastic matrices as directed graphs. Example of the (asymmetric) random walk on the discrete torus, computation of the stationary distribution.

## 3.1 Existence of a stationary distribution

**Proposition 3.1** (Finite state space)**.** *If $E$ is finite, there always exists a stationary distribution.*

Proof by the 'Krylov–Bogoliubov' compactness argument.

When $E$ is countably infinite this no longer needs to be true, example of the (asymmetric) random walk on $\mathbb{Z}$.

**Definition 3.2** (Positive recurrence)**.** *For any $x \in E$, defined $\tau_x = \inf\{n \geq 1 : X_n = x\}$. The state $x$ is called* positive recurrent *if $\mathbb{E}_x[\tau_x] < +\infty$, where the notation $\mathbb{E}_x$ means that we are assuming that $X_0 = x$.*

**Proposition 3.3** (Positive recurrence and stationary distribution)**.** *If $x$ is positive recurrent, then the probability measure $\pi_x$ defined on $E$ by*

$$\forall y \in E, \qquad \pi_x(y) := \frac{\mathbb{E}_x\left[\sum_{n=0}^{\tau_x - 1} \mathbb{1}_{\{X_n = y\}}\right]}{\mathbb{E}_x[\tau_x]}$$

*is a stationary distribution for $(X_n)_{n\geq 0}$.*

## 3.2 Uniqueness of a stationary distribution

**Definition 3.4** (Irreducibility)**.** *The Markov chain $(X_n)_{n\geq 0}$ (or, equivalently, the stochastic matrix $P$) is* irreducible *if, for any $x, y \in E$, there exists $n \geq 1$ such that $P^n(x,y) > 0$.*

**Proposition 3.5** (Irreducibility implies uniqueness)**.** *If the Markov chain $(X_n)_{n\geq 0}$ is irreducible, then it admits at most one stationary distribution. Furthermore, if it has a positive recurrent state then all states are positive recurrent, and all probability measures $\pi_x$ introduced in Proposition 3.3 coincide with each other.*

If the chain is irreducible and has positive recurrent states, it is called *positive recurrent* itself.

## 3.3 Ergodic theorems

Proposition 3.5 shows that for an irreducible and positive recurrent chain, the stationary distribution $\pi(y)$ measures the average proportion of time spent in $y$ between two consecutive visits to

some arbitrary fixed state $x$. Since the Markov property implies that the excursions of the chain between such consecutive visits form independent and identically distributed excursions, one may apply the standard LLN to these excursions to obtain the following statement.

**Theorem 3.6** (Law of Large Numbers for Markov chains)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible and positive recurrent Markov chain, with unique stationary distribution $\pi$. For any $f \in \mathbf{L}^1(\pi)$,*

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) = \pi f = \sum_{x\in E} \pi(x)f(x), \qquad \textit{almost surely.}$$

Since the proof of Theorem 3.6 relies on the decomposition of the trajectory of $(X_n)_{n\geq 0}$ into iid excursions, the same argument may be expected to also yield a Central Limit Theorem. To infer the expression of the limiting variance in this statement, one may first try to compute $\lim_{n\to+\infty} \mathrm{Var}(\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} f(X_k))$. A formal computation leads to the formula

$$\sigma^2(f) := \mathrm{Var}_\pi(f(X_0)) + 2\sum_{n=1}^{+\infty} \mathrm{Cov}_\pi(f(X_0), f(X_n)),$$

where the notation $\mathrm{Var}_\pi$, $\mathrm{Cov}_\pi$ indicates that we take $X_0 \sim \pi$.

**Theorem 3.7** (Central Limit Theorem for Markov chains)**.** *Under the assumptions of Theorem 3.6, assume that $\sigma^2(f)$ is well-defined. Then*

$$\lim_{n\to+\infty} \sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \pi f\right) = \mathcal{N}(0, \sigma^2(f)), \qquad \textit{in distribution.}$$

From a numerical point of view, Theorem 3.6 indicates that, if one is interested in computing the expectation $\pi f$ (which is nothing but $\mathbb{E}[f(X)]$ when $X \sim \pi$) by the Monte Carlo method, but it is not possible to draw iid samples from $\pi$, then one may construct a Markov chain with stationary distribution $\pi$ and the empirical mean $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$ still converges to the correct limit, although the $X_k$ are no longer independent nor identically distributed. Theorem 3.7 then provides confidence intervals for this method. Therefore, this raises the following question: given a probability measure $\pi$ on $E$, can we design a Markov chain which admits $\pi$ as stationary distribution?

# 4 Markov chain Monte Carlo methods

## 4.1 Reversibility

**Definition 4.1** (Reversibility)**.** *A Markov chain $(X_n)_{n\geq 0}$ with transition matrix $P$ is reversible with respect to a probability measure $\pi$ on $E$ if it satisfies*

$$\forall x, y \in E, \qquad \pi(x)P(x,y) = \pi(y)P(y,x),$$

*which is called the* detailed balance equation*.*

The detailed balance equation precisely means that for any $x, y \in E$, $\mathbb{P}_\pi(X_0 = x, X_1 = y) = \mathbb{P}_\pi(X_0 = y, X_1 = x)$, that is to say that if $X_0 \sim \pi$, then the pairs $(X_0, X_1)$ and $(X_1, X_0)$ have the same law. In particular the first coordinates of each pair $X_0$ and $X_1$ have the same law, which means that if $(X_n)_{n\geq 0}$ is reversible with respect to $\pi$, then $\pi$ is stationary for $(X_n)_{n\geq 0}$.

## 4.2 The Metropolis algorithm

We write our target measure under the form

$$\pi_\beta(x) = \frac{1}{Z_\beta} \mathrm{e}^{-\beta V(x)}, \qquad Z_\beta := \sum_{x \in E} \mathrm{e}^{-\beta V(x)},$$

with $\beta \geq 0$ and $V : E \to \mathbb{R}$. Our computational assumptions are that, given $x \in E$, we are able to evaluate $V(x)$, but we cannot compute the sum $Z_\beta$. This prevents us from drawing iid samples from $\pi_\beta$ to apply the Monte Carlo method.

Given the target measure $\pi_\beta$, the basic ingredients of the Metropolis algorithm are:

- an irreducible stochastic matrix $Q$, called the *proposal* and such that $Q(x, y) > 0$ if and only if $Q(y, x) > 0$, under which we assume that we are able to draw transitions;
- a function $F : (0, +\infty) \to (0, 1]$, called the *acceptance* function, which satisfies the identity

$$\forall \rho > 0, \qquad F(\rho) = \rho F\left(\frac{1}{\rho}\right).$$

Examples of acceptance functions are the *Metropolis–Hastings rule* $F(\rho) = \min(1, \rho)$ and the *Barker rule* $F(\rho) = \frac{\rho}{1+\rho}$.

The algorithm then constructs a Markov chain $(X_n)_{n \geq 0}$: given $X_n = x$,

1. draw a state $y$ with probability $Q(x, y)$;
2. set $X_{n+1} = y$ with probability $F(\frac{\pi_\beta(y)Q(y,x)}{\pi_\beta(x)Q(x,y)})$, and $X_{n+1} = x$ otherwise.

Then it turns out that the Markov chain $(X_n)_{n \geq 0}$ is irreducible and reversible with respect to $\pi_\beta$, and can be simulated without the knowledge of $Z_\beta$ since the ratio $\pi_\beta(y)/\pi_\beta(x)$ does not depend on this quantity.