MAA305 - Probability: Stochastic Processes

Quentin Cormier
September 14, 2026

Contents

1 Reminders on probability theory

Probability spaces, random variables

Definition 1.1. *Let X be a set. A σ-algebra $\mathcal{F}$ on X is a set of subsets of X satisfying*

(a) $\varnothing \in \mathcal{F}$

(b) $A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$

(c) $(A_n \in \mathcal{F}, n \in \mathbb{N}) \implies \bigcup_n A_n \in \mathcal{F}$.

The pair $(X, \mathcal{F})$ *is called a measurable space.*

Example 1.2. *Let* I *be a countable set. Then* $(I, \mathcal{P}(I))$ *is a measurable space.*

Example 1.3. *The smallest* σ*-algebra which contains all the open balls of* $\mathbb{R}^d$ *is called the Borel sigma-algebra. Its elements are called the Borel sets.*

Definition 1.4. *Let* $(X_1, \mathcal{F}_1)$ *and* $(X_2, \mathcal{F}_2)$ *be two measurable spaces. A function* $f : X_1 \to X_2$ *is measurable if* $f^{-1}(A) \in \mathcal{F}_1$ *for all* $A \in \mathcal{F}_2$.

Definition 1.5. *We say that* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space if* $(\Omega, \mathcal{F})$ *is a measurable space and* $\mathbb{P} : \mathcal{F} \to [0,1]$ *is a function such that:*

- $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\varnothing) = 0$.
- *Given a sequence* $(A_n)_n$ *of pairwise disjoint events,* $A_n \in \mathcal{F}$*, we have* $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.

The elements of $\mathcal{F}$ are called the *events* and $\mathbb{P}$ is a *probability measure*. We recall that for any sequence of events (A_n), it holds that

$$\mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n).$$

Definition 1.6. *Let* $(X, \mathcal{E})$ *be a measurable space. A function* $Y : \Omega \to X$ *which is measurable is called a random variable.*

Let $Y : \Omega \to X$ be a random variable. For $B \in \mathcal{E}$, we denote

$$\{Y \in B\} := Y^{-1}(B) = \{\omega \in \Omega : Y(\omega) \in B\}.$$

Finally, given $Y : \Omega \to X$ a random variable, its distribution (or law) is the probability measure on $(X, \mathcal{E})$ defined by

$$\mu(B) = \mathbb{P}(Y \in B), \quad \forall B \in \mathcal{E}.$$

Assume now that $X = \mathbb{R}$. To every non-negative or integrable real-valued random variable Y is associated an expectation $\mathbb{E}(Y)$, which is the integral of Y with respect to $\mathbb{P}$. Given $A \in \mathcal{F}$, we denote by $\mathbb{1}_A$ the following random variable:

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

We have $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$. We also recall that expectation is linear: $\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$.

Example 1.7. *Given a countable state space* I*, a random variable* $Y : \Omega \to I$ *models a random state with distribution:*

$$\lambda_i = \mathbb{P}(Y = i) = \mathbb{P}(\{\omega : Y(\omega) = i\}).$$

Let $f : I \to [0, \infty]$*. The expectation of* $Z = f(Y) = f \circ Y$ *is explicitly given by*

$$\mathbb{E}(f(Y)) = \sum_{i \in I} f(i)\lambda_i.$$

Independence

Definition 1.8. *Let $A_i \in \mathcal{F}, i \in I$ be a family of events. We say that they are independent if for any* ***finite*** *subset of indices $J \subset I$, it holds that*

$$\mathbb{P}(\bigcap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j).$$

Let $(X_i, \mathcal{E}_i), i \in I$ be a family of measurable spaces. For all $i \in I$, let $Y_i : \Omega \to X_i$ be a random variable. We say that the random variables $Y_i, i \in I$ are independent if for any ***finite*** *subset $J \subset I$, for all $B_j \in \mathcal{E}_j$, $j \in J$, we have*

$$\mathbb{P}(\bigcap_{j \in J} \{Y_j \in B_j\}) = \prod_{j \in J} \mathbb{P}(Y_j \in B_j).$$

Conditional probability

Finally, let $B \in \mathcal{F}$ be an event with $\mathbb{P}(B) > 0$. We recall that for any $A \in \mathcal{F}$, the probability of A given B is defined as:

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

We verify that $\mathcal{F} \ni A \mapsto \mathbb{P}(A \mid B)$ is a new probability measure on $(\Omega, \mathcal{F})$.

The strong law of large numbers

Finally, we recall the strong law of large numbers. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(X_n)_{n \geq 0}$, $X_n : \Omega \to \mathbb{R}$, be a collection of independent random variables. Assume that all the variables have the same distribution. Assume that $\mathbb{E}|X_1| < \infty$. Then

$$\mathbb{P}(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}(X_1)) = 1.$$

This means that there exists $A \in \mathcal{F}$ with $\mathbb{P}(A) = 1$ and

$$\forall \omega \in A, \quad \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \mathbb{E}(X_1).$$

We say that there is *almost sure* convergence.

2 Definition of Markov Chains and basic properties

Let I be a countable set. Each $i \in I$ is called a state. A function $\lambda : I \to [0, \infty)$ is a measure on I. If the total mass is equal to 1:

$$\sum_{i \in I} \lambda_i = 1,$$

then we call λ a distribution (or probability law). If X is a random variable taking values in I, its associated distribution is

$$\lambda_i = \mathbb{P}(X = i).$$

We say that a matrix $P : (P_{ij} : i, j \in I)$ is *stochastic* if every row is a distribution, that is:

$$\forall i \in I, \quad \sum_{j \in I} P_{ij} = 1.$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 2.1. *A discrete time stochastic process is a collection of random variables $(X_n)_{n \geq 0}$ defined on the same probability space.*

"A Markov chain is a stochastic process for which the future evolution only depends on the past through the current state."

Definition 2.2. *Let λ be a distribution on I and P a stochastic matrix. Let $(X_n)_{n \geq 0}$ be a sequence of random variables taking values in I. We say that (X_n) is a Markov chain with initial distribution λ and transition matrix P if, for all $n \geq 0$ and all $i_0, \cdots, i_n \in I$,*

$$\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n) = \lambda_{i_0} P_{i_0 i_1} \cdots P_{i_{n-1} i_n}.$$

For $n = 0$, the product of transition probabilities is empty and equals 1, so this identity reads $\mathbb{P}(X_0 = i_0) = \lambda_{i_0}$.

We write $\mathrm{MC}(\lambda, P)$ in short. The following theorem gives an equivalent formulation in terms of conditional probabilities.

Theorem 2.3 (Conditional Markov property)**.** *Let λ be a distribution on I and P a stochastic matrix. A sequence $(X_n)_{n \geq 0}$ of random variables taking values in I is an $MC(\lambda, P)$ if and only if the following two conditions hold:*

(a) $\mathbb{P}(X_0 = i) = \lambda_i$ *for every* $i \in I$.

(b) For every $n \geq 0$ and every $i_0, \cdots, i_n \in I$ such that

$$\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n) > 0,$$

we have, for every $j \in I$,

$$\mathbb{P}(X_{n+1} = j \mid X_0 = i_0, \cdots, X_n = i_n) = P_{i_n j}.$$

Proof. For a history $i_0, \cdots, i_n$, write $A_n = \{X_0 = i_0, \cdots, X_n = i_n\}$. Assume first that (X_n) is an $\mathrm{MC}(\lambda, P)$. Condition (a) is the case $n = 0$ of Definition 2.2. Applying the defining identity at times n and $n + 1$ gives

$$\mathbb{P}(A_n \cap \{X_{n+1} = j\}) = \mathbb{P}(A_n) P_{i_n j}.$$

Whenever $\mathbb{P}(A_n) > 0$, dividing by $\mathbb{P}(A_n)$ proves (b).

Conversely, assume (a) and (b). We prove the identity in Definition 2.2 by induction on n. For $n = 0$, it is precisely (a). Suppose it holds at time n and fix $i_0, \cdots, i_{n+1} \in I$. If $\mathbb{P}(A_n) > 0$, condition (b) gives

$$\mathbb{P}(A_{n+1}) = \mathbb{P}(A_n) \mathbb{P}(X_{n+1} = i_{n+1} \mid A_n) = \mathbb{P}(A_n) P_{i_n i_{n+1}}.$$

If $\mathbb{P}(A_n) = 0$, the identity $\mathbb{P}(A_{n+1}) = \mathbb{P}(A_n) P_{i_n i_{n+1}}$ still holds: both sides are zero, since $A_{n+1} \subset A_n$. In either case, the induction hypothesis yields

$$\mathbb{P}(A_{n+1}) = \lambda_{i_0} P_{i_0 i_1} \cdots P_{i_n i_{n+1}},$$

which completes the proof. $\square$

Condition (b) says that, given a history of positive probability, the distribution of the next state depends only on the current state.

Example 2.4. *Susie lives in London and has an umbrella and every day travels from home to office in the morning and from office to home in the evening. For each trip, she takes the umbrella if it is available and if it rains. If it does not rain, the umbrella stays where it is. At each trip, the probability of finding rain is* p.

- *If Susie and the umbrella are both at home or both at the office, then Susie takes the umbrella iff it rains.*
- *If Susie and the umbrella are at different places, she does not take the umbrella.*

Let $C_n = \{$ *it rains during* n*-th trip* $\}$; (C_n) *are independent events with* $\mathbb{P}(C_n) = p$. *Let*

$$X_n = \begin{cases} 1 & \textit{if Susie and the umbrella are at the same place at time } n \\ 0 & \textit{otherwise}\ . \end{cases}$$

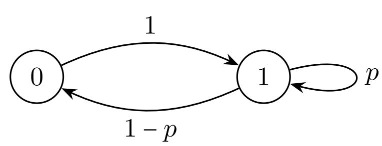
We have:

- $X_{n+1} = 1$ *if* $X_n = 0$.
- $X_{n+1} = 1$ *if* $X_n = 1$ *and* C_n.
- $X_{n+1} = 0$ *if* $X_n = 1$ *and it does not rain.*

Therefore, $\mathbb{P}(X_{n+1} = 1 \mid X_n = 0) = 1$, $\mathbb{P}(X_{n+1} = 1 \mid X_n = 1) = \mathbb{P}(C_n) = p$. *We deduce that* $(X_n)_n$ *is a Markov chain on* $I = \{0, 1\}$, *with transition matrix*

$$P = \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Graphical representation:



Assume that at time $n = 0$, *Susie and the umbrella are at the same place. It holds that*

$$\mathbb{P}(X_n = 0 \mid X_0 = 1) = (P^n)_{10}.$$

Let $S_n = \{$*it rains at trip* n *and Susie has no umbrella*$\}$. *Then* $\mathbb{P}(S_n) = \mathbb{P}(X_n = 0, C_n) = \mathbb{P}(X_n = 0)\mathbb{P}(C_n) = p(P^n)_{10}$.

For $i \in I$ we denote by δ_i the *Dirac* distribution defined as:

$$\delta_i(j) = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise.} \end{cases}$$

Theorem 2.5. *Let (X_n) be an $MC(\lambda, P)$. Then, conditional on the event $X_m = i$, $(X_{m+n})_{n\geq 0}$ is an $MC(\delta_i, P)$ and is independent of $(X_0, \cdots, X_m)$.*

Proof. Recall that $\mathbb{P}(\cdot \mid X_m = i)$ defines a new probability measure. The statement of the theorem says:

(a) For this new probability measure, $(X_{n+m})_n$ is an $MC(\delta_i, P)$. That is:

$$\mathbb{P}(X_m = i_m, \cdots, X_{m+n} = i_{m+n} \mid X_m = i) = \delta_i(i_m) P_{i_m, i_{m+1}} \cdots P_{i_{n+m-1} i_{n+m}}.$$

(b) Under this new probability measure, the two random variables $(X_{n+m})_{n\geq 0}$ and $(X_0, \cdots, X_m)$ are independent. That is:

$$\mathbb{P}((X_{n+m})_{n\geq 0} \in B_1, (X_0, \cdots, X_m) \in B_2 \mid X_m = i) = \\ \mathbb{P}((X_{n+m})_{n\geq 0} \in B_1 \mid X_m = i) \mathbb{P}((X_0, \cdots, X_m) \in B_2 \mid X_m = i)$$

Let A be the event

$$A = \{X_0 = i_0, \cdots, X_m = i_m\}.$$

We prove that for all $n \geq 0$,

$$\mathbb{P}(\{X_m = i_m, \cdots, X_{m+n} = i_{m+n}\} \cap A \mid X_m = i) = \delta_i(i_m) P_{i_m i_{m+1}} \cdots P_{i_{m+n-1} i_{m+n}} \mathbb{P}(A \mid X_m = i).$$

We admit that this implies the two points above. We have to show that

$$\frac{\mathbb{P}(X_0 = i_0, \cdots X_{m+n} = i_{m+n}, X_m = i)}{\mathbb{P}(X_m = i)} = \delta_i(i_m) P_{i_m i_{m+1}} \cdots P_{i_{m+n-1} i_{m+n}} \frac{\mathbb{P}(X_0 = i_0 \cdots X_m = i_m, X_m = i)}{\mathbb{P}(X_m = i)}.$$

But this is true by Definition 2.2. $\square$

Given P a stochastic matrix and λ a measure, let

$$(\lambda P)_j = \sum_{i \in I} \lambda_i P_{ij}, \quad (P^2)_{ij} = \sum_{k \in I} P_{ik} P_{kj}.$$

We verify that P^2 is a new stochastic matrix. We define by induction P^n. By convention, $(P^0)_{ij} = \delta_i(j)$ (identity matrix). In the case where $\lambda_i > 0$, we write:

$$\mathbb{P}_i(A) := \mathbb{P}(A \mid X_0 = i), \quad A \in \mathcal{F}.$$

By the Markov property at time 0, under $\mathbb{P}_i$, $(X_n)_{n\geq 0}$ is an $MC(\delta_i, P)$.

Theorem 2.6. *Let (X_n) be an $MC(\lambda, P)$. Then for all $n, m \geq 0$,*

(a) $\mathbb{P}(X_n = j) = (\lambda P^n)_j$

(b) $\mathbb{P}_i(X_n = j) = \mathbb{P}(X_{n+m} = j \mid X_m = i) = (P^n)_{ij}$.

Proof. By Definition 2.2, we have

$$\begin{aligned} \mathbb{P}(X_n = j) &= \sum_{i_0 \in I, \cdots i_{n-1} \in I} \mathbb{P}(X_0 = i_0, \cdots X_{n-1} = i_{n-1}, X_n = j) \\ &= \sum_{i_0 \in I, \cdots i_{n-1} \in I} \lambda_{i_0} P_{i_0 i_1} \cdots P_{i_{n-1} j} \\ &= (\lambda P^n)_j. \end{aligned}$$

Second, by the Markov property, conditional on $X_m = i$, $(X_{m+n})_{n\geq 0}$ is an $MC(\delta_i, P)$ so we take $\lambda = \delta_i$ in the first point to obtain the result. $\square$

3 Simulation of trajectories on a computer

It is easy to sequentially simulate a trajectory of a Markov chain on a computer. First draw a random variable $X_0 \sim \lambda$, this serves as the initial condition. Then, draw sequentially $X_{n+1} \sim (P_{X_n j}, j \in I)$. We will see during the tutorial classes algorithms to do that efficiently. Assume now that you have generated N independent trajectories $(X_n^p), p \in 1 \cdots N, n \in 0 \cdots n_{\max}$. Our goal is to evaluate the probability that some event occurs. For instance, say we want to evaluate $\mathbb{P}(X_{n_{\max}} = 1)$. By the strong law of large numbers, it holds that with probability 1,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{p=1}^{N} \mathbb{1}_{\{X^p_{n_{\max}} = 1\}} = \mathbb{E}\mathbb{1}_{\{X_{n_{\max}} = 1\}} = \mathbb{P}(X_{n_{\max}} = 1).$$

This shows that for N large enough,

$$\frac{1}{N} \sum_{p=1}^{N} \mathbb{1}_{\{X^p_{n_{\max}} = 1\}} \approx \mathbb{P}(X_{n_{\max}} = 1).$$

The error can be estimated using the central limit theorem. This powerful and simple method to estimate probabilities of various scenarios is known as the *Monte-Carlo* method. The alternative *deterministic* method to evaluate this probability is to use

$$\mathbb{P}(X_{n_{\max}} = 1) = (\lambda P^{n_{\max}})_1,$$

which requires computing P to the power $n_{\max}$.

4 Class structure

We say that $i \rightsquigarrow j$ (i leads to j) if

$$\mathbb{P}_i(\exists n : X_n = j) > 0.$$

We say that $i \leftrightarrow j$ (i communicates with j) if both $i \rightsquigarrow j$ and $j \rightsquigarrow i$.

Theorem 4.1. *Let $i, j \in I$, $i \neq j$. The following are equivalent:*

(a) $i \rightsquigarrow j$

(b) $\exists n, \exists i_0, \cdots, i_n$ *such that* $P_{i_0 i_1} \cdots P_{i_{n-1} i_n} > 0$ *with* $i_0 = i$ *and* $i_n = j$.

(c) $(P^n)_{ij} > 0$ *for some* $n \geq 0$.

Proof. We have:

$$(P^n)_{ij} = \mathbb{P}_i(X_n = j) \leq \mathbb{P}_i(\exists n : X_n = j) = \mathbb{P}_i(\bigcup_n \{X_n = j\}) \leq \sum_n \mathbb{P}_i(X_n = j) = \sum_n (P^n)_{ij}.$$

So (a) and (c) are equivalent. As $(P^n)_{ij} = \sum_{i_1, \cdots, i_{n-1}} P_{i, i_1} \cdots P_{i_{n-1} j}$, (b) and (c) are also equivalent. □

By the property (b), if $i \rightsquigarrow j$ and $j \rightsquigarrow k$ then $i \rightsquigarrow k$. In addition, $i \rightsquigarrow i$, so $\leftrightarrow$ is an equivalence relation on I. Therefore, we can consider the equivalence classes associated to this relation. We call them the *communicating classes*. We say that a class C is *closed* if

$$i \in C, i \rightsquigarrow j \implies j \in C.$$

A closed class is a class for which there is no escape. We say that a state i is *absorbing* if $\{i\}$ is a closed class. Finally

Definition 4.2. *We say that a Markov chain is irreducible if I is a single class.*

5 Hitting times, absorption probabilities

Let (X_n) be an $MC(\lambda, P)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let A be a subset of I. We define the random variable $H^A : \Omega \to \mathbb{N} \cup \{\infty\}$ as

$$H^A(\omega) = \inf\{n \geq 0 : X_n(\omega) \in A\},$$

with the convention that $\inf \varnothing = \infty$. We then consider:

$$h_i^A = \mathbb{P}_i(H^A < \infty).$$

Theorem 5.1. *The vector $(h_i^A, i \in I)$ is the minimal non-negative solution of the system*

$$\begin{cases} h_i^A = 1, & \text{for } i \in A \\ h_i^A = \sum_{j \in I} P_{ij} h_j^A & \text{for } i \notin A. \end{cases}$$

Proof. If $X_0 = i \in A$, then $H^A = 0$ and so $h_i^A = 1$. If $i \notin A$, we have, using the Markov property:

$$\begin{aligned} h_i^A &= \mathbb{P}_i(H^A < \infty) \\ &= \sum_{j \in I} \mathbb{P}_i(H^A < \infty \mid X_1 = j)\mathbb{P}_i(X_1 = j) \\ &= \sum_{j \in I} \mathbb{P}_j(H^A < \infty) P_{ij} = \sum_{j \in I} h_j^A P_{ij}. \end{aligned}$$

Let x be a non-negative solution of this system. We have $h_i^A = x_i = 1$ for $i \in A$. Let $i \notin A$. We have

$$\begin{aligned} x_i &= \sum_{j \in A} P_{ij} + \sum_{j \notin A} P_{ij} x_j \\ &= \sum_{j \in A} P_{ij} + \sum_{j \notin A, k \in A} P_{ij} P_{jk} + \sum_{j \notin A, k \notin A} P_{ij} P_{jk} x_k \\ &= \mathbb{P}_i(X_1 \in A) + \mathbb{P}_i(X_1 \notin A, X_2 \in A) + \sum_{j \notin A, k \notin A} P_{ij} P_{jk} x_k. \end{aligned}$$

By induction, we deduce that

$$\begin{aligned} x_i = {} & \mathbb{P}_i(X_1 \in A) + \mathbb{P}_i(X_1 \notin A, X_2 \in A) + \cdots + \mathbb{P}_i(X_1 \notin A, \cdots, X_{n-1} \notin A, X_n \in A) \\ & + \sum_{j_1, \ldots j_n \notin A} P_{ij_1} P_{j_1 j_2} \cdots P_{j_{n-1} j_n} x_{j_n}. \end{aligned}$$

By assumption, the last term is non-negative so $x_i \geq \mathbb{P}_i(H^A \leq n)$. We let $n \to \infty$ and obtain $x_i \geq \mathbb{P}_i(H^A < \infty) = h_i^A$. □

Example 5.2 (The gambler's ruin). *Let $p \in (0,1)$. Consider the Markov chain defined on $I = \mathbb{N} = \{0,1,2,\cdots\}$ with transition probabilities given by*

$$\forall i \in \{1,2,\cdots\}, \quad P_{i,i+1} = p, \quad P_{i,i-1} = 1-p, \quad \text{and} \quad P_{00} = 1.$$

Let $q = 1-p$. Set $h_i = \mathbb{P}_i(\text{hit } 0)$, the probability of losing everything, starting with an amount of money i. It is the minimal non-negative solution of

$$\begin{aligned} h_0 &= 1 \\ h_i &= ph_{i+1} + qh_{i-1} \quad \text{for } i \in \{1,2,\cdots\}. \end{aligned}$$

If $p \neq q$, any solution of the last equation has the form $h_i = A + B(q/p)^i$. If $p < q$, then $h_i \leq 1$ forces $B = 0$ and so $h_i = 1$. If $p > q$, since $h_0 = 1$, we find $h_i = (q/p)^i + A(1-(q/p)^i)$. The minimal non-negative solution is thus $h_i = (q/p)^i$. Finally, if $p = q = 1/2$, the general solution is $h_i = A + Bi$. We find that $B = 0, A = 1$ and so $h_i = 1$.

6 Stopping times and the Strong Markov property

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (X_n) be an $MC(\lambda, P)$ on this space.

Definition 6.1. *A random variable $T : \Omega \to \{0,1,2,\cdots\} \cup \{\infty\}$ is called a stopping time if for all $n \geq 0$, there exists a* ***deterministic*** *function $F_n : I^{n+1} \to \{0,1\}$ such that*

$$\forall n, \quad \mathbb{1}_{\{T=n\}} = F_n(X_0, X_1, \cdots, X_n).$$

Intuitively, that means that by watching the process up to time n, you know if the event occurs by time n or not.

Example 6.2. *The first hitting time $H^A = \inf\{n \geq 0 : X_n \in A\}$ is a stopping time because*

$$\{H^A = n\} = \{X_0 \notin A, X_1 \notin A, \cdots, X_{n-1} \notin A, X_n \in A\}.$$

Example 6.3. *The last exit time $L^A = \sup\{n \geq 0 : X_n \in A\}$ is not in general a stopping time, because the event $\{L^A = n\}$ depends on whether $(X_{n+m})_{m\geq 1}$ visits A or not.*

Theorem 6.4. *Let (X_n) be an $MC(\lambda, P)$. Let T be a stopping time for (X_n). Then, conditionally on $\{T < \infty, X_T = i\}$, it holds that $(X_{n+T})_{n\geq 0}$ is an $MC(\delta_i, P)$ and independent of $X_0, \cdots, X_T$.*

Proof. Recall that $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot \mid T < \infty, X_T = i)$ defines a new probability measure. The theorem says:

(a) For this new probability measure, $(X_{n+T})_n$ is an $MC(\delta_i, P)$. That is:

$$\tilde{\mathbb{P}}(X_T = i_0, \cdots, X_{T+n} = i_n) = \delta_i(i_0) P_{i_0 i_1} \cdots P_{i_{n-1} i_n}.$$

(b) Under $\tilde{\mathbb{P}}$, the two random variables $(X_{n+T})_{n\geq 0}$ and $(X_0, \cdots, X_T)$ are independent. That is:

$$\tilde{\mathbb{P}}((X_{n+T})_n \in B_1, (X_0, \cdots, X_T) \in B_2) = \tilde{\mathbb{P}}((X_{n+T})_n \in B_1) \tilde{\mathbb{P}}((X_0, \cdots, X_T) \in B_2).$$

We only prove the first point. We have to prove that

$$\underbrace{\mathbb{P}(X_T = i_0, \cdots, X_{T+n} = i_n, T < \infty, X_T = i)}_{=:p} = \mathbb{P}(X_T = i, T < \infty)\delta_i(i_0)P_{i_0 i_1}\cdots P_{i_{n-1}i_n}.$$

By the law of total probability,

$$p = \sum_{m \geq 0} \mathbb{P}(X_m = i_0, \cdots, X_{m+n} = i_n, X_m = i, T = m).$$

In addition, by assumption, $\{T = m\} = \{F_m(X_0, \cdots, X_m) = 1\}$. Therefore,

$$p = \sum_{m \geq 0} \mathbb{P}(X_m = i_0, \cdots, X_{m+n} = i_n, F_m(X_0, \cdots, X_m) = 1 \mid X_m = i)\mathbb{P}(X_m = i).$$

At this point, we use the Markov property (Theorem 2.5). We find that under $\mathbb{P}(\cdot \mid X_m = i)$, (X_{m+n}) is an $MC(\delta_i, P)$ and is independent of $(X_0, \cdots, X_m)$. Therefore,

$$p = \sum_{m \geq 0} \delta_i(i_0)P_{i_0 i_1}\cdots P_{i_{n-1}i_n}\mathbb{P}(F_m(X_0, \cdots, X_m) = 1 \mid X_m = i)\mathbb{P}(X_m = i).$$

Using that $\mathbb{P}(F_m(X_0, \cdots, X_m) = 1 \mid X_m = i)\mathbb{P}(X_m = i) = \mathbb{P}(T = m | X_m = i)\mathbb{P}(X_m = i) = \mathbb{P}(T = m, X_m = i)$, we find that

$$p = \delta_i(i_0)P_{i_0 i_1}\cdots P_{i_{n-1}i_n}\mathbb{P}(X_T = i, T < \infty),$$

as required. $\square$

Example 6.5. *Let's go back to the gambler's ruin problem. Recall that the Markov chain is defined on $I = \mathbb{N} = \{0, 1, 2, \cdots\}$ with transition probabilities given by*

$$\forall i \in \{1, 2, \cdots\}, \quad P_{i,i+1} = p, \quad P_{i,i-1} = 1 - p, \quad \text{and} \quad P_{00} = 1.$$

Again assume $0 < p < 1$ and write $q := 1 - p$. Set:

$$H_j = \inf\{n \geq 0, X_n = j\}.$$

We have already computed

$$\mathbb{P}_1(H_0 < \infty) = \begin{cases} 1 & \text{if } q \geq p \\ q/p & \text{if } q < p. \end{cases}$$

We now want to compute the complete distribution of H_0 under $\mathbb{P}_1$. To do so, we consider the generating function for $s \in [0, 1)$:

$$\phi(s) = \mathbb{E}_1(s^{H_0}) = \sum_{n < \infty} s^n \mathbb{P}_1(H_0 = n).$$

We use the conventions $s^\infty = 0$ and $s^0 = 1$, including when $s = 0$. Suppose first we start at 2. To reach 0, we must first reach 1. Define the remaining hitting time by

$$\tilde{H}_0 = \begin{cases} \inf\{n \geq 0 : X_{H_1 + n} = 0\}, & \text{if } H_1 < \infty, \\ 0, & \text{if } H_1 = \infty. \end{cases}$$

Then $H_0 = H_1 + \tilde{H}_0$ *under* $\mathbb{P}_2$*, also when* $H_1 = \infty$*. Let* $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}_2(\cdot \mid H_1 < \infty)$*. Since* $s^{H_0} = 0$ *on* $\{H_1 = \infty\}$*,*

$$\mathbb{E}_2 s^{H_0} = \tilde{\mathbb{E}}[s^{H_1} s^{\tilde{H}_0}]\mathbb{P}_2(H_1 < \infty).$$

Under $\tilde{\mathbb{P}}$*, we have*

$$H_1 = \sum_{n\geq 1} n \mathbb{1}_{X_0 \neq 1, \cdots, X_{n-1} \neq 1, X_n = 1}.$$

Thus H_1 *is determined by the stopped past. By the strong Markov property, under* $\tilde{\mathbb{P}}$*,* $(X_{n+H_1})_{n\geq 0}$ *is independent of* H_1*. Since* $\tilde{H}_0$ *depends only on this shifted process,* H_1 *and* $\tilde{H}_0$ *are independent under* $\tilde{\mathbb{P}}$*. Hence*

$$\mathbb{E}_2 s^{H_0} = \tilde{\mathbb{E}} s^{H_1} \tilde{\mathbb{E}} s^{\tilde{H}_0} \mathbb{P}_2(H_1 < \infty).$$

Under $\tilde{\mathbb{P}}$*,* (X_{n+H_1}) *is an* $MC(\delta_1, P)$*, so*

$$\tilde{\mathbb{E}} s^{\tilde{H}_0} = \mathbb{E}_1 s^{H_0} = \phi(s).$$

Finally, $\tilde{\mathbb{E}} s^{H_1} \mathbb{P}_2(H_1 < \infty) = \mathbb{E}_2 s^{H_1} = \phi(s)$*, exploiting the invariance by translation of the problem (reaching 0 from state 1 is the same as reaching 1 from state 2). Altogether, we have*

$$\boxed{\mathbb{E}_2 s^{H_0} = \phi^2(s).}$$

Assume now we start at 1*. We apply the Markov property at time* 1*. We find*

$$\begin{aligned}\phi(s) = \mathbb{E}_1 s^{H_0} &= \mathbb{E}_1[s^1 \mid X_1 = 0]\mathbb{P}_1(X_1 = 0) + \mathbb{E}_1[s^{H_0} \mid X_1 = 2]\mathbb{P}_1(X_1 = 2) \\ &= qs + ps\mathbb{E}_2 s^{H_0}.\end{aligned}$$

We used that $H_0 = \inf\{n \geq 0 : X_n = 0\} = 1 + \inf\{n \geq 0 : X_{n+1} = 0\}$*, and under* $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}_1(\cdot \mid X_1 = 2)$*,* $(X_{n+1})_n$ *is an* $MC(\delta_2, P)$*. Altogether, we find that*

$$\boxed{\phi(s) = sq + sp\phi^2(s)}.$$

For $0 < s < 1$*, the polynomial* $psu^2 - u + qs$ *has positive leading coefficient and takes the value* $s - 1 < 0$ *at* $u = 1$*, so its larger root is greater than* 1*. Since* $\phi(s) \leq 1$*, we obtain*

$$\phi(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps} \quad (0 < s < 1), \qquad \phi(0) = 0.$$

We expand ϕ *as a power series in* s*. We find:*

$$\phi(s) = qs + pq^2 s^3 + 2p^2 q^3 s^5 + 5p^3 q^4 s^7 + 14p^4 q^5 s^9 + O\left(s^{11}\right).$$

So, for instance:

$$\mathbb{P}_1(H_0 = 1) = q, \quad \mathbb{P}_1(H_0 = 3) = pq^2, \quad \mathbb{P}_1(H_0 = 5) = 2p^2 q^3, \quad \mathbb{P}_1(H_0 = 7) = 5p^3 q^4,$$

and so on.

7 Recurrence and transience

Definition 7.1. *Let (X_n) be a Markov chain with transition matrix P.*

(a) A state i is recurrent if

$$\mathbb{P}_i(X_n = i \text{ for infinitely many } n\ ) = 1.$$

(b) A state i is transient if

$$\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 0.$$

The first passage time to state i is

$$T_i = \inf\{n \geq 1 : X_n = i\}, \quad (\inf \varnothing = \infty).$$

We define similarly by induction the r-th passage time to state i by:

$$T_i^{(0)} = 0, \quad T_i^{(r+1)} = \inf\{n \geq 1 + T_i^{(r)} : X_n = i\}.$$

In addition, we consider the length of the r-th excursion

$$r \in \{1, 2, \cdots\}, \quad S_i^{(r)} = \begin{cases} T_i^{(r)} - T_i^{(r-1)} & \text{if } T_i^{(r-1)} < \infty \\ 0 \quad \text{otherwise.} \end{cases}$$

Lemma 7.2. *Conditional on $T_i^{(r-1)} < \infty$, $S_i^{(r)}$ is independent of $(X_1, \cdots, X_{T_i^{(r-1)}})$ and*

$$\mathbb{P}(S_i^{(r)} = n \mid T_i^{(r-1)} < \infty) = \mathbb{P}_i(T_i = n).$$

Proof. Let $T = T_i^{(r-1)}$. It is a stopping time. It holds that $X_T = i$ on $T < \infty$. Therefore, on $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot \mid T < \infty)$, by the Strong Markov property, it holds that under $\tilde{\mathbb{P}}$, $(X_{T+n})_n$ is an $MC(\delta_i, P)$ which is independent of $(X_0, X_1, \cdots X_T)$. As $S_i^{(r)} = \inf\{n \geq 1, X_{T+n} = i\}$, $S_i^{(r)}$ is independent of $(X_0, X_1, \cdots, X_T)$ and

$$\tilde{\mathbb{P}}(S_i^{(r)} = n) = \mathbb{P}_i(T_i = n).$$

□

Let V_i be the total number of visits to the state i:

$$V_i = \sum_{n \geq 0} \mathbb{1}_{\{X_n = i\}}.$$

We have

$$\mathbb{E}_i V_i = \sum_{n \geq 0} \mathbb{P}_i(X_n = i) = \sum_{n \geq 0} (P^n)_{ii}.$$

Lemma 7.3. *Let $f_i = \mathbb{P}_i(T_i < \infty)$. Then it holds that for all $r = 0, 1, \cdots$, we have*

$$\mathbb{P}_i(V_i > r) = (f_i)^r.$$

Proof. Assume that $X_0 = i$ then $\{V_i > r\} = \{T_i^{(r)} < \infty\}$. We prove the result by induction on r. We have using the previous lemma,

$$\begin{aligned}
\mathbb{P}_i(V_i > r+1) &= \mathbb{P}_i(T_i^{(r+1)} < \infty) \\
&= \mathbb{P}_i(T_i^{(r)} < \infty, S_i^{(r+1)} < \infty) \\
&= \mathbb{P}_i(S_i^{(r+1)} < \infty \mid T_i^{(r)} < \infty)\mathbb{P}_i(T_i^{(r)} < \infty) \\
&= \mathbb{P}_i(T_i < \infty)\mathbb{P}_i(V_i > r) = f_i(f_i)^r.
\end{aligned}$$

As the result is obviously true for $r = 0$, we are done. □

Recall that for any non-negative random variable V,

$$\begin{aligned}
\sum_{r=0}^{\infty} \mathbb{P}(V > r) &= \sum_{r=0}^{\infty} \sum_{n=r+1}^{\infty} \mathbb{P}(V = n) \\
&= \sum_{n=1}^{\infty} \sum_{r=0}^{n-1} \mathbb{P}(V = n) = \sum_{n=1}^{\infty} n\mathbb{P}(V = n) = \mathbb{E}V.
\end{aligned}$$

Theorem 7.4. *We have*

(a) if $\mathbb{P}_i(T_i < \infty) = 1$ then i is recurrent and $\sum_{n=0}^{\infty}(P^n)_{ii} = \infty$.

(b) if $\mathbb{P}_i(T_i < \infty) < 1$ then i is transient and $\sum_{n=0}^{\infty}(P^n)_{ii} < \infty$.

In particular, each state is either transient or recurrent.

Proof. Assume $\mathbb{P}_i(T_i < \infty) = 1$, then $\mathbb{P}_i(V_i = \infty) = \lim_{r\to\infty} \mathbb{P}_i(V_i > r) = 1$: i is recurrent. In addition,

$$\mathbb{E}_i(V_i) = \infty = \sum_{n=0}^{\infty}(P^n)_{ii}.$$

Suppose now that $f_i = \mathbb{P}_i(T_i < \infty) < 1$. Then $\mathbb{E}_i V_i = \sum_{r=0}^{\infty}(f_i)^r = \frac{1}{1-f_i} < \infty$. By Markov's inequality, $\mathbb{P}_i(V_i > r) \le \frac{1}{r}\mathbb{E}_i V_i$. Take the limit $r \to \infty$:

$$\mathbb{P}_i(V_i = \infty) = 0.$$

Therefore, i is transient. □

Lemma 7.5. *Let C be a communicating class. Then either all states in C are recurrent or transient.*

Proof. Assume that $i \in C$ is transient. Let $j \neq i \in C$. As $i \rightsquigarrow j$ and $j \rightsquigarrow i$, there exist $n, m \ge 0$ such that $(P^n)_{ij} > 0$ and $(P^m)_{ji} > 0$. Let $r \ge 0$, we have

$$(P^{r+n+m})_{ii} \ge (P^n)_{ij}(P^r)_{jj}(P^m)_{ji},$$

so

$$\sum_{r=0}^{\infty}(P^r)_{jj} \le \frac{1}{(P^n)_{ij}(P^m)_{ji}} \sum_{r=0}^{\infty}(P^{r+n+m})_{ii} < \infty,$$

so j is also transient. □

Definition 7.6. *A communicating class is recurrent if it contains a recurrent state. A communicating class is transient if it contains a transient state.*

Theorem 7.7. *Every recurrent class is closed.*

Proof. Let C be a class which is not closed. So there exists $i \in C$ and $j \notin C$ such that $i \rightsquigarrow j$. There exists $m \geq 1$ such that

$$\mathbb{P}_i(X_m = j) > 0.$$

Then,

$$\begin{aligned}\mathbb{P}_i(X_n = i \text{ for infinitely many } n) &\leq \mathbb{P}_i(X_n = i \text{ for infinitely many } n, X_m = j) \\ &\quad + \mathbb{P}_i(X_n = i \text{ for infinitely many } n, X_m \neq j) := A + B.\end{aligned}$$

It holds that $A \leq \mathbb{P}_i(X_m = j, \exists n : X_{n+m} = i) = \mathbb{P}_i(\exists n : X_{n+m} = i \mid X_m = j)\mathbb{P}_i(X_m = j) = \mathbb{P}_j(T_i < \infty)\mathbb{P}_i(X_m = j)$. As $j \notin C$, $\mathbb{P}_j(T_i < \infty) = 0$ and so $A = 0$. In addition, $B \leq \mathbb{P}_i(X_m \neq j) = 1 - \mathbb{P}_i(X_m = j) < 1$. So $A + B < 1$, therefore i (and so C) is transient. ☐

Theorem 7.8. *Every* ***finite*** *closed class is recurrent.*

Proof. As C is closed and finite, there exists at least one state i such that

$$p := \mathbb{P}(X_n = i \text{ for infinitely many } n) > 0.$$

Otherwise, we would have

$$\mathbb{P}(\exists N : \forall n \geq N, X_n \notin C) = 1,$$

but this contradicts the fact that C is closed... By the strong Markov property, $p = \mathbb{P}_i(X_n = i \text{ for infinitely many } n)\mathbb{P}_i(T_i < \infty)$. As $\mathbb{P}_i(X_n = i \text{ for infinitely many } n) > 0$, i is not transient. So i and C are recurrent. ☐

Theorem 7.9. *Let P be irreducible and recurrent. Then for all $j \in I$, we have $\mathbb{P}(T_j < \infty) = 1$.*

Proof. By the Markov property, $\mathbb{P}(T_j < \infty) = \sum_{i \in I} \mathbb{P}(X_0 = i)\mathbb{P}_i(T_j < \infty)$. We prove that $\mathbb{P}_i(T_j < \infty) = 1$ for all i. Let m such that $(P^m)_{ji} > 0$. We have

$$\begin{aligned}1 &= \mathbb{P}_j(X_n = j \text{ for infinitely many } n) \\ &= \mathbb{P}_j(X_n = j \text{ for some } n \ \geq m + 1) \\ &= \sum_{k \in I} \mathbb{P}_j(X_n = j \text{ for some } n \geq m + 1 \mid X_m = k)\mathbb{P}_j(X_m = k) \\ &= \sum_{k \in I} \mathbb{P}_k(T_j < \infty)(P^m)_{jk}.\end{aligned}$$

We used the Markov property to obtain the last line. Because, $\sum_{k \in I}(P^m)_{jk} = 1$, we must have $\mathbb{P}_i(T_j < \infty) = 1$. ☐

8 Invariant distributions

We recall that a measure λ is any row vector $(\lambda_i, i \in I)$ with $\lambda_i \geq 0$ for all $i \in I$.

Definition 8.1. *We say λ is invariant if $\lambda P = \lambda$.*

Theorem 8.2. *Let $(X_n)_n$ be an $MC(\lambda, P)$, and assume that λ is invariant for P. Then $\mathbb{P}(X_n = i) = \lambda_i$ for all n. In that case, the distribution of (X_n) is invariant with respect to time.*

Proof. We have $\mathbb{P}(X_m = i) = (\lambda P^m)_i = \lambda_i$. □

We start with a not very useful result; but it does explain the methodology we will use to construct invariant measures.

Theorem 8.3. *Assume I is finite. Assume that there exists $i \in I$ such that*

$$\forall j \in I, \quad \lim_{n\to\infty} (P^n)_{ij} = \pi_j.$$

Then $\pi = (\pi_j, j \in I)$ is an invariant distribution of P.

Proof. We have

$$\sum_{j\in I} \pi_j = \lim_{n\to\infty} \sum_{j\in I} P^n_{ij} = 1.$$

And:

$$\pi_j = \lim_{n\to\infty} (P^{n+1})_{ij} = \lim_{n\to\infty} \sum_{k\in I} (P^n)_{ik} P_{kj} = \sum_{k\in I} \pi_k P_{kj}.$$

We used that I is finite to exchange the limits and the sums. □

Example 8.4. *Let*

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

We look for an invariant distribution of the Markov chain π, such that $\pi = \pi P$, that is:

$$\begin{aligned} \pi_1 &= \frac{1}{2}\pi_3 \\ \pi_2 &= \pi_1 + \frac{1}{2}\pi_2 \\ \pi_3 &= \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3. \end{aligned}$$

In addition, we want $\pi_1 + \pi_2 + \pi_3 = 1$. We find

$$\pi = (1/5\ 2/5\ 2/5).$$

Fix a state k. Recall $T_k = \inf\{n \geq 1 : X_n = k\}$. We consider:

$$\gamma_i^k = \mathbb{E}_k \sum_{n=0}^{T_k - 1} \mathbb{1}_{\{X_n = i\}}.$$

This is the expected number of visits to state i before the return to state k.

Theorem 8.5. *Let P be irreducible and recurrent. Then*

(a) $\gamma_k^k = 1$*;*

(b) $\gamma^k = (\gamma_i^k, i \in I)$ *satisfies* $\gamma^k P = \gamma^k$*;*

(c) $0 < \gamma_i^k < \infty$.

Proof. (a) We are OK as under $\mathbb{P}_k$, $X_0 = k$ and $X_n \neq k$ for all $n \in \{1, \cdots, T_k - 1\}$.

(b) Note that $\{n \leq T_k\} = \{X_1 \neq k, X_2 \neq k, \cdots X_{n-1} \neq k\}$, and this event only depends on $\{X_1, \cdots, X_{n-1}\}$. We apply the Markov property at time $n-1$. Under $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}_k(\cdot \mid X_{n-1} = i)$, it holds that $(X_{m+(n-1)})$ is an $MC(\delta_i, P)$ which is independent of $(X_0, \cdots X_{n-1})$. So

$$\tilde{\mathbb{P}}(X_n = j, n \leq T_k) = \tilde{\mathbb{P}}(X_{(n-1)+1} = j, X_1 \neq k, \cdots X_{n-1} \neq k) = P_{ij}\tilde{\mathbb{P}}(n \leq T_k).$$

Therefore,

$$\begin{aligned}\mathbb{P}_k(X_n = j, n \leq T_k, X_{n-1} = i) &= \tilde{\mathbb{P}}(X_n = j, n \leq T_k)\mathbb{P}_k(X_{n-1} = i) \\ &= P_{ij}\mathbb{P}_k(n \leq T_k, X_{n-1} = i).\end{aligned}$$

P is recurrent, so under $\mathbb{P}_k$, $T_k < \infty$ and $X_{T_k} = k$ with probability one. So

$$\begin{aligned}\gamma_j^k &= \mathbb{E}_k \sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = j, n \leq T_k\}} \\ &= \sum_{i \in I} \sum_{n=1}^{\infty} \mathbb{P}_k(X_n = j, n \leq T_k, X_{n-1} = i) \\ &= \sum_{i \in I} \sum_{n=1}^{\infty} P_{ij}\mathbb{P}_k(X_{n-1} = i, n \leq T_k) \\ &= \sum_{i \in I} P_{ij}\mathbb{E}_k \sum_{m=0}^{T_k - 1} \mathbb{1}_{\{X_m = i\}} \\ &= \sum_{i \in I} P_{ij}\gamma_i^k.\end{aligned}$$

So $\gamma^k = \gamma^k P$ as claimed.

(c) Let $i \in I$. P is irreducible so there exists $n, m \geq 0$ such that $(P^n)_{ik} > 0$ and $(P^m)_{ki} > 0$. We have $\gamma^k = \gamma^k P^m$ so

$$\gamma_i^k = \sum_{j \in I} (P^m)_{ji}\gamma_j^k \geq (P^m)_{ki}\gamma_k^k > 0.$$

As $\gamma^k P^n = \gamma^k$ we also have

$$1 = \gamma_k^k = \sum_{j \in I} (P^n)_{jk}\gamma_j^k \geq (P^n)_{ik}\gamma_i^k,$$

so $\gamma_i^k \leq 1/(P^n)_{ik} < \infty$.

□

Theorem 8.6. *Let P be irreducible and let λ be an invariant measure for P with $\lambda_k = 1$. Then $\lambda \geq \gamma^k$. If in addition P is recurrent then $\lambda = \gamma^k$.*

Proof. Let $j \in I$, we have:

$$
\begin{aligned}
\lambda_j &= \sum_{i_1 \in I} \lambda_{i_1} P_{i_1 j} \\
&= P_{kj} + \sum_{i_1 \neq k} \lambda_{i_1} P_{i_1 j} \\
&= \sum_{i_1 \neq k, i_2 \neq k} \lambda_{i_2} P_{i_2 i_1} P_{i_1 j} + \left(P_{kj} + \sum_{i_1 \neq k} P_{k i_1} P_{i_1 j} \right) \\
&= \sum_{i_1 \neq k \cdots i_n \neq k} \lambda_{i_n} P_{i_n i_{n-1}} \cdots P_{i_1 j} + \left(P_{kj} + \sum_{i_1 \neq k} P_{k i_1} P_{i_1 j} + \cdots + \sum_{i_1 \neq k \cdots i_{n-1} \neq k} P_{k i_{n-1}} \cdots P_{i_2 i_1} P_{i_1 j} \right).
\end{aligned}
$$

So

$$\lambda_j \geq \underbrace{\mathbb{P}_k(X_1 = j, T_k \geq 1) + \mathbb{P}_k(X_2 = j, T_k \geq 2) + \cdots + \mathbb{P}_k(X_n = j, T_k \geq n)}_{= \mathbb{E}_k \sum_{i=1}^n \mathbb{1}_{\{X_i = j, T_k \geq i\}}}$$

That is:

$$\lambda_j \geq \mathbb{E}_k \sum_{i=1}^{\min(T_k, n)} \mathbb{1}_{\{X_i = j\}}.$$

We let $n \to \infty$ and find $\lambda_j \geq \gamma_j^k$. Assume now that P is recurrent. Then γ^k is also an invariant measure by the previous theorem. So $\mu = \lambda - \gamma^k$ is an invariant measure. Fix $i \in I$, as P is irreducible, there exists n such that $(P^n)_{ik} > 0$. Therefore,

$$0 = \mu_k = \sum_{j \in I} \mu_j (P^n)_{jk} \geq \mu_i (P^n)_{ik},$$

so $\mu_i = 0$. □

Recall that a state i is recurrent if

$$\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 1.$$

We have seen that this is equivalent to $\mathbb{P}_i(T_i < \infty) = 1$.

Definition 8.7. *We say that a state i is positive recurrent if the expected return time*

$$m_i = \mathbb{E}_i(T_i)$$

is finite. A recurrent state with $m_i = \infty$ is said to be null-recurrent.

Theorem 8.8. *Let P be irreducible. Then the following are equivalent:*

(a) All the states are positive recurrent.

(b) At least one state i is positive recurrent.

(c) P has an invariant distribution, denoted by π.

Under one of these three conditions, it holds that $m_i = 1/\pi_i$.

Proof. We first prove that (b) implies (c). As i is positive recurrent, i is recurrent, so P is recurrent. So γ^i is an invariant measure. Note that

$$\sum_{k\in I}\gamma_k^i = \mathbb{E}_i \sum_{n=0}^{T_i-1} \mathbb{1}_{\{X_n\in I\}} = \mathbb{E}_i T_i < \infty.$$

So $\pi_j = \frac{\gamma_j^i}{\mathbb{E}_i T_i}$ is an invariant distribution.

We now prove that (c) implies (a). Let $k \in I$. As P is irreducible and the π_i sum to 1, we have

$$\pi_k = \sum_{i\in I}\pi_i (P^n)_{ik} > 0.$$

So $\lambda_i = \pi_i/\pi_k$ satisfies $\lambda_k = 1$ and so for all i, $\lambda_i \geq \gamma_i^k$. Therefore,

$$m_k = \sum_{i\in I}\gamma_i^k \leq \sum_{i\in I}\frac{\pi_i}{\pi_k} = \frac{1}{\pi_k} < \infty,$$

so k is positive recurrent.

In fact, when one of the three conditions holds, P is recurrent, so we have $\lambda_i = \gamma_i^k$ and so

$$m_k = \frac{1}{\pi_k}.$$

□

9 Convergence to equilibrium

Our goal is to investigate the limit $(P^n)_{ij}$ as n goes to infinity. If the state space is finite and the limit exists for all j then we have seen that the limit should be the invariant distribution. But the limit does not always exist! The problem comes from periodic behaviors and lack of mixing. Consider for instance:

Example 9.1.

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then $P^{2n} = I$ and $P^{2n+1} = P$. So $(P^n)_{ij}$ does not converge as n goes to infinity.

This periodic effect is an obstacle to convergence and motivates the following definition.

Definition 9.2. *A state $i \in I$ is called aperiodic if $(P^n)_{ii} > 0$ for all sufficiently large n.*

Exercise 9.3. *Show that a state i is aperiodic if and only if*

$$gcd\{n \geq 0 : (P^n)_{ii} > 0\} = 1.$$

Hint: first note that if $(P^n)_{ii} > 0$ then $(P^{nk})_{ii} > 0$ for all integers $k \geq 0$.

Lemma 9.4. *Suppose that P is irreducible and has an aperiodic state i. Then for all states j, k, $(P^n)_{jk} > 0$ for sufficiently large n. In particular, all states are aperiodic.*

Proof. There exist $r, s \geq 0$ such that $(P^r)_{ji} > 0$ and $(P^s)_{ik} > 0$ so

$$P^{r+s+n}_{jk} \geq (P^r)_{ji}(P^n)_{ii}(P^s)_{ik} > 0,$$

for n large enough. □

Theorem 9.5. *Let P be irreducible and aperiodic. Suppose that P has an invariant distribution π. Let λ be any distribution and (X_n) be $MC(\lambda, P)$. Then,*

$$\mathbb{P}(X_n = j) \to \pi_j, \quad \text{as } n \to \infty \text{ for all } j.$$

In particular, $(P^n)_{ij} \to \pi_j$ for all $i, j \in I$.

Proof. Let (Y_n) be an $MC(\pi, P)$ independent of (X_n), defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $b \in I$ be fixed (this serves as a reference state) and let

$$T = \inf\{n \geq 1 : X_n = Y_n = b\}.$$

Step 1. We prove that $\mathbb{P}(T < \infty) = 1$. Let $W_n = (X_n, Y_n)$. Note that (W_n) is a Markov chain on $I \times I$ with transition probability

$$\tilde{P}_{(i,k),(j,l)} = P_{ij}P_{kl},$$

and initial condition $\mu_{(i,k)} = \lambda_i \pi_k$. As P is aperiodic, we have

$$\tilde{P}^n_{(i,k),(j,l)} = (P^n)_{ij}(P^n)_{kl} > 0,$$

for n large enough, so $\tilde{P}$ is irreducible. Also, clearly,

$$\tilde{\pi}_{(i,k)} = \pi_i \pi_k,$$

is an invariant distribution for $\tilde{P}$. So $\tilde{P}$ is positive recurrent, in particular recurrent. We deduce that $\mathbb{P}(T < \infty) = 1$, as T is the first passage time of the state (b, b) of (W_n).

Step 2. Set

$$Z_n = \begin{cases} X_n & \text{if } n < T \\ Y_n & \text{otherwise.} \end{cases}.$$

We show that (Z_n) is an $MC(\lambda, P)$. Indeed, as the event $\{T \leq n\}$ only depends on $\{(X_0, Y_0), \cdots, (X_n, Y_n)\}$. So, by the Markov property at time n applied to (W_n), we have

$$\begin{aligned}
&\mathbb{P}(Z_{n+1} = j \mid Z_n = i, \underbrace{Z_{n-1} = i_{n-1}, \cdots, Z_0 = i_0}_{A_n}) \\
&\quad = \mathbb{P}(Z_{n+1} = j \mid Z_n = i, A_n, T \leq n)\mathbb{P}(T \leq n) + \mathbb{P}(Z_{n+1} = j \mid Z_n = i, A_n, T > n)\mathbb{P}(T > n) \\
&\quad = \mathbb{P}(Y_{n+1} = j \mid Y_n = i, A_n, T \leq n)\mathbb{P}(T \leq n) + \mathbb{P}(X_{n+1} = j \mid X_n = i, A_n, T > n)\mathbb{P}(T > n) \\
&\quad = P_{ij}\mathbb{P}(T \leq n) + P_{ij}\mathbb{P}(T > n) = P_{ij},
\end{aligned}$$

hence (Z_n) is an $MC(\lambda, P)$. We used that

$$\begin{aligned}
\mathbb{P}(Y_{n+1} = j \mid Y_n = i, A_n, T \leq n) &= \sum_{j \in I} \mathbb{P}(Y_{n+1} = j \mid Y_n = i, X_n = k, A_n, T \leq n)\mathbb{P}(X_n = k) \\
&= \sum_{l,k \in I} \mathbb{P}_{(k,i)}(W_1 = (j,l))\mathbb{P}(Y_n = k) \\
&= P_{ij}.
\end{aligned}$$

Step 3. We have $\mathbb{P}(Z_n = j) = \mathbb{P}(X_n = j, n < T) + \mathbb{P}(Y_n = j, n \geq T)$ and

$$\pi_j = \mathbb{P}(Y_n = j) = \mathbb{P}(Y_n = j, n \geq T) + \mathbb{P}(Y_n = j, n < T).$$

So

$$\begin{aligned} |\mathbb{P}(X_n = j) - \pi_j| &= |\mathbb{P}(Z_n = j) - \pi_j| \\ &\leq |\mathbb{P}(X_n = j, T > n) + \mathbb{P}(Y_n = j, T > n)| \\ &\leq 2\mathbb{P}(T > n). \end{aligned}$$

As $\mathbb{P}(T > n) \to \mathbb{P}(T = \infty) = 0$, we deduce the result.

□

It is important to understand why the above proof fails when the chain is not aperiodic. For that, we go back to the Example 9.1 and note that with probability 1/2, (X_n, Y_n) never meet.

10 Time reversal

Theorem 10.1. *Let P be irreducible with an invariant distribution π. Suppose that $(X_n)_{0\leq n\leq N}$ is an $MC(\pi, P)$. Let $Y_n = X_{N-n}$. Then (Y_n) is an $MC(\pi, \hat{P})$, where:*

$$\pi_j \hat{P}_{ji} = \pi_i P_{ij}.$$

In addition, $\hat{P}$ is also irreducible, with invariant distribution π.

Proof. First, $\hat{P}$ is a stochastic matrix. Indeed, as π is invariant for P, we have

$$\sum_{i\in I} \hat{P}_{ji} = \sum_{i\in I} \frac{\pi_i}{\pi_j} P_{ij} = 1.$$

As P is a stochastic matrix,

$$\sum_{j\in I} \pi_j \hat{P}_{ji} = \sum_{j\in I} \pi_i P_{ij} = \pi_i.$$

We deduce that π is an invariant distribution for $\hat{P}$. In addition,

$$\begin{aligned} \mathbb{P}(Y_0 = i_0, Y_1 = i_1, \cdots, Y_N = i_N) &= \mathbb{P}(X_N = i_0, X_{N-1} = i_1, \cdots X_0 = i_N) \\ &= P_{i_1 i_0} P_{i_2 i_1} \cdots P_{i_N i_{N-1}} \pi_{i_N} \\ &= \frac{\pi_{i_0}}{\pi_{i_1}} \hat{P}_{i_0 i_1} \frac{\pi_{i_1}}{\pi_{i_2}} \hat{P}_{i_1 i_2} \cdots \frac{\pi_{i_{N-1}}}{\pi_{i_N}} \hat{P}_{i_{N-1} i_N} \pi_{i_N} \\ &= \pi_{i_0} \hat{P}_{i_0 i_1} \hat{P}_{i_1 i_2} \cdots \hat{P}_{i_{N-1} i_N}. \end{aligned}$$

We deduce that (Y_n) is an $MC(\pi, \hat{P})$. Finally, as P is irreducible, for each pair $i, j \in I$, there exists $i = i_0, i_1, \cdots, i_{n-1}, i_n = j$ such that

$$P_{i_0 i_1} \cdots P_{i_{n-1}, i_n} > 0.$$

So,

$$\hat{P}_{i_n, i_{n-1}} \cdots \hat{P}_{i_1, i_0} > 0,$$

as well. □

Definition 10.2. *A stochastic matrix P and a measure λ are said to be in detailed balance if:*

$$\lambda_i P_{ij} = \lambda_j P_{ji}, \quad \forall i, j \in I.$$

Lemma 10.3. *If P and λ are in detailed balance, then λ is invariant for P.*

Proof. We have

$$(\lambda P)_i = \sum_{j \in I} \lambda_j P_{ji} = \sum_{j \in I} \lambda_i P_{ij} = \lambda_i.$$

□

Let (X_n) be an $MC(\lambda, P)$ with P irreducible. We say that (X_n) is reversible if for all $N \geq 0$, (X_{N-n}) is also $MC(\lambda, P)$.

Theorem 10.4. *Let P be an irreducible stochastic matrix, and λ be a distribution. Let (X_n) be an $MC(\lambda, P)$. Then the two following are equivalent:*

(a) (X_n) is reversible

(b) P and λ are in detailed balance.

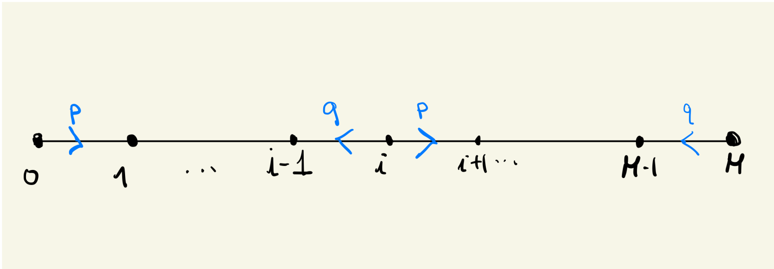
Proof. If (X_n) is reversible, λ is invariant for P and $Y_n = X_{N-n}$ is an MC with transition probability $\hat{P} = P$. We deduce that P and λ are in detailed balance. The converse is equally easy. □

Example 10.5. *Consider*

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

The invariant probability measure is $\pi = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$. So $\hat{P} = P^T$, the transpose of P, which is different from P: the chain is not reversible. In fact, in the long run, the chain moves clockwise. Under time-reversal, it would run backwards.

Example 10.6. *Let $p \in (0,1)$ and $q = 1 - p$. Consider the following Markov chain on $\{0, \cdots, M\}$:*



We try to compute the invariant probability measure by solving the detailed balance equation:

$$\lambda_i P_{ij} = \lambda_j P_{ji}.$$

That is:

$$\forall i \in \{0, \cdots, M-1\}, \quad \lambda_i p = \lambda_{i+1} q$$
$$\forall i \in \{1, \cdots, M\}, \quad \lambda_i q = \lambda_{i-1} p.$$

So $\lambda_i = \lambda_0 \left(\frac{p}{q}\right)^i$. *This can be normalized to give an invariant distribution. So the chain is reversible. To understand better what this means, assume that* p *is very close to* 1. *Then, the chain tends to move to the right; and its time reversal to the left. But, we start at equilibrium, which is concentrated near* M*: the chain spends most of its time near* M*, with occasional excursions to the left. This behavior is symmetrical in time.*

Example 10.7. *Let* $G = (V, E)$ *be a graph. The set* V *is the set of vertices. The set* $E \subset V \times V$ *is the set of edges. Two vertices* i, j *are joined by an edge if* $(i, j) \in E$. *The valency* v_i *of a vertex* i *is the number of edges at* i. *Assume that every vertex has finite valency. The random walk on* G *chooses edges with equal probability. Therefore, the transition probabilities are given by*

$$P_{ij} = \begin{cases} \frac{1}{v_i} & \text{if } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases} .$$

Assume that G *is connected, so* P *is irreducible. Then,* P *is in detailed balance with*

$$v = (v_i, i \in V).$$

If the total valency is finite:

$$\sigma = \sum_{i \in V} v_i < \infty$$

then $\pi = v/\sigma$ *is an invariant probability measure and* P *is reversible.*

Example 10.8. *A random knight on a chessboard makes each permissible move with equal probability. It starts in a corner. How long on average will it take to return to the same corner? This is a random walk on a graph. We have*

$$\mathbb{E}_c T_c = \frac{1}{\pi_c} = \sum_{i \in V} \frac{v_i}{v_c}.$$

The four corners have valency 2. *The eight adjacent squares have valency* 3. *There are* $4 \times 4 + 4$ *squares of valency* 4, *and* 4×4 *of valency* 6. *Finally the* 4×4 *central squares have valency* 8. *So*

$$\mathbb{E}_c T_c = \frac{1}{2}\left(4 * 2 + 8 * 3 + 20 * 4 + 16 * 6 + 16 * 8\right) = 168.$$

11 The ergodic theorem

We first recall a version of the strong law of large numbers.

Theorem 11.1. *Let* (Y_i) *be a sequence of i.i.d. non-negative random variables with* $\mathbb{E}Y_1 = \mu$. *Then*

$$\mathbb{P}\left(\frac{Y_1 + \cdots + Y_n}{n} \to_{n \to +\infty} \mu\right) = 1.$$

Proof. If $\mu < \infty$, this is the standard law of large numbers with finite first moment. If $\mu = \infty$, set $Y_i^N = \min(Y_i, N)$. Then

$$\frac{Y_1 + \cdots + Y_n}{n} \geq \frac{Y_1^N + \cdots + Y_n^N}{n} \to_{n \to \infty} \mathbb{E}\min(Y_1, N).$$

So, with probability 1,

$$\liminf_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} \geq \mathbb{E}\min(Y_1, N).$$

By the monotone convergence theorem, $\mathbb{E}\min(Y_1, N) \to_{N\to\infty} \mathbb{E}Y_1 = \infty$. □

Denote by $V_i(n)$ the number of visits in state i before time n:

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{1}_{\{X_k = i\}}.$$

Then $V_i(n)/n$ is the proportion of the time spent in state i.

Theorem 11.2. *Let P be irreducible and let λ be any distribution. Let (X_n) be an $MC(\lambda, P)$. Then*

$$\mathbb{P}(\lim_{n\to\infty} \frac{V_i(n)}{n} = \frac{1}{m_i}) = 1,$$

where $m_i = \mathbb{E}_i(T_i)$ is the expected return time to state i.

Proof. If P is transient, then the total number of visits in state i is almost surely finite:

$$\mathbb{P}(V_i < \infty) = 1.$$

So

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \to 0 = \frac{1}{m_i}.$$

Assume P is recurrent. For $T = T_i$, we have (X_{n+T}) is $MC(\delta_i, P)$ and $\mathbb{P}(T < \infty) = 1$. In addition, (X_{n+T}) is independent of $(X_0, \cdots, X_T)$. As the long-time proportion of time spent in i is the same for (X_{n+T}) and (X_n), we can assume without loss of generality that $\lambda = \delta_i$. Write $S_i^{(r)}$ for the length of the r-th excursion to state i:

$$T_i^{(0)} = 0, \quad T_i^{(r+1)} = \inf\{n \geq 1 + T_i^{(r)} : X_n = i\}.$$
$$r \in \{1, 2, \cdots\}, \quad S_i^{(r)} = T_i^{(r)} - T_i^{(r-1)}.$$

By the strong Markov property, the $(S_i^{(r)}, r \geq 1)$ are $i.i.d$ with mean

$$\mathbb{E}S_i^{(1)} = \mathbb{E}_i T_i = m_i.$$

We have

$$S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)} \leq n - 1$$
$$S_i^{(1)} + \cdots + S_i^{(V_i(n))} \geq n.$$

So:

$$\frac{S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{S_i^{(1)} + \cdots + S_i^{(V_i(n))}}{V_i(n)}.$$

As P is recurrent,

$$\mathbb{P}(V_i(n) \to \infty) = 1.$$

So, by the strong law of large numbers, with probability 1,

$$\frac{S_i^{(1)} + \cdots + S_i^{(V_i(n))}}{V_i(n)} \to_{n\to\infty} m_i.$$

Similarly,

$$\frac{S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)}}{V_i(n)} = \underbrace{\frac{S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)}}{V_i(n)-1}}_{\to m_i} \underbrace{\frac{V_i(n)-1}{V_i(n)}}_{\to 1},$$

we deduce that

$$\mathbb{P}\left(\frac{V_i(n)}{n} \to_{n\to+\infty} \frac{1}{m_i}\right) = 1.$$

□

Theorem 11.3. *Let P be irreducible and positive recurrent. Let λ be any distribution. Let (X_n) be an $MC(\lambda, P)$. Then, for any bounded function $f : I \to \mathbb{R}$, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \langle f, \pi\rangle\right) = 1,$$

where π is the unique invariant distribution and

$$\langle f, \pi\rangle = \sum_{i\in I} f_i \pi_i.$$

Proof. Without loss of generality, we assume that $|f|$ is bounded by 1. Let $J \subset I$ be finite. We have

$$\begin{aligned}
|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \langle f,\pi\rangle| &= |\sum_{i\in I}\left(\frac{V_i(n)}{n} - \pi_i\right) f_i| \\
&\le |\sum_{i\in J}\left(\frac{V_i(n)}{n} - \pi_i\right) f_i| + |\sum_{i\notin J}\left(\frac{V_i(n)}{n} - \pi_i\right) f_i| \\
&\le \sum_{i\in J}|\frac{V_i(n)}{n} - \pi_i| + \sum_{i\notin J}\left(\frac{V_i(n)}{n} + \pi_i\right) =: A + B.
\end{aligned}$$

As

$$\sum_{i\in J} V_i(n) + \sum_{i\notin J} V_i(n) = n,$$

we have

$$B = \sum_{i\in I}\pi_i - \sum_{i\in J}\frac{V_i(n)}{n} + \sum_{i\notin J}\pi_i.$$

Altogether,

$$A + B \le 2\sum_{i\in J}|\frac{V_i(n)}{n} - \pi_i| + 2\sum_{i\notin J}\pi_i.$$

Fix now $\epsilon > 0$. We choose J such that

$$\sum_{i\notin J}\pi_i \le \frac{\epsilon}{4}.$$

By the previous result, with probability 1, there exists $N = N(\omega)$ such that for all $n \geq N$:

$$\sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| \leq \frac{\epsilon}{4}.$$

We deduce that

$$\forall n \geq N(\omega), \quad \left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \langle f, \pi \rangle \right| \leq \epsilon.$$

□

12 Total variation distance

Let (X_n) be an $MC(\lambda, P)$ with P irreducible and aperiodic with invariant distribution π. Question: how fast does (X_n) converge to π? We need a distance on probability measures to quantify the rate of convergence.

Definition 12.1. *Denote by $\mathcal{P}(I)$ the set of all probability distributions on I. The total variation distance $d_{TV} : \mathcal{P}(I) \times \mathcal{P}(I) \to \mathbb{R}_+$ is defined by:*

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{i \in I} |\mu_i - \nu_i|.$$

Proposition 12.2. *The total variation distance satisfies*

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sup_f \left\{ \sum_{i \in I} f(i)\mu_i - \sum_{i \in I} f(i)\nu_i \ : \ \sup_{i \in I} |f(i)| \leq 1 \right\}.$$

Proof. If $\sup_{i \in I} |f(i)| \leq 1$, then

$$\frac{1}{2} \left| \sum_{i \in I} f(i)\mu_i - \sum_{i \in I} f(i)\nu_i \right| \leq \frac{1}{2} \sum_{i \in I} |\mu(i) - \nu(i)| = d_{TV}(\nu, \mu).$$

Conversely, define:

$$f(x) = \begin{cases} 1 & \text{if } \mu(x) \geq \nu(x) \\ -1 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \frac{1}{2} \left| \sum_{i \in I} f(i)\mu(i) - \sum_{i \in I} f(i)\nu(i) \right| &= \frac{1}{2} \sum_{i \in I} f(i)(\mu(i) - \nu(i)) \\ &= \frac{1}{2} \sum_{i \in I} |\mu(i) - \nu(i)| = d_{TV}(\nu, \mu). \end{aligned}$$

□

Lemma 12.3. *We also have*

$$d_{TV}(\nu, \mu) = \sum_{i \in I : \mu_i \geq \nu_i} (\mu_i - \nu_i).$$

Proof. Let $A = \sum_{i \in I: \mu_i \geq \nu_i} (\mu_i - \nu_i)$ and $B = \sum_{i \in I: \mu_i < \nu_i} (\nu_i - \mu_i)$. We have $A + B = 2d_{TV}(\nu, \mu)$ and $A - B = 0$ so $A = d_{TV}(\nu, \mu)$. □

Theorem 12.4. *The total variation distance satisfies:*

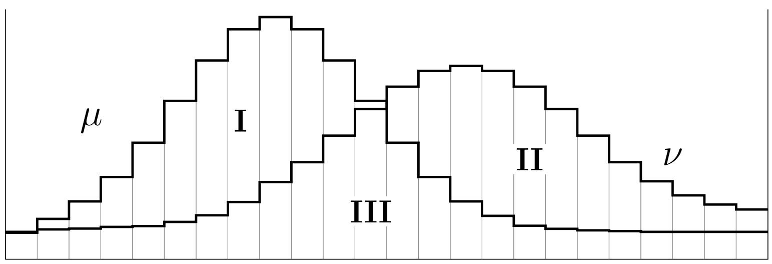
$$d_{TV}(\mu, \nu) = \inf_{\mathcal{L}(X)=\mu, \mathcal{L}(Y)=\nu} \mathbb{P}(X \neq Y).$$

The infimum is taken over all random variables (X, Y) defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the distribution of X is μ and the distribution of Y is ν.

Proof. Let f such that $\sup_{i \in I} |f(i)| \leq 1$. Let (X, Y) such that the distribution of X is μ and the distribution of Y is ν. We have

$$\sum_{i \in I} f(i)\mu_i - \sum_{i \in I} f(i)\nu_i = \mathbb{E}f(X) - \mathbb{E}f(Y) \leq \mathbb{E}|f(X) - f(Y)| \leq 2\mathbb{P}(X \neq Y).$$

And so $d_{TV}(\nu, \mu) \leq \inf_{\mathcal{L}(X)=\mu, \mathcal{L}(Y)=\nu} \mathbb{P}(X \neq Y)$. For the other side, it suffices to construct a coupling (X, Y) for which $\mathbb{P}(X \neq Y)$ is equal to the total variation distance. For that, we force X and Y to be equal as often as possible.



Consider the region III, bounded by $\min(\mu_i, \nu_i)$. The area of region I is $\sum_{i: \mu_i \geq \nu_i} (\mu_i - \nu_i) = d_{TV}(\nu, \mu)$. Similarly, the area of region II is $d_{TV}(\nu, \mu)$. As the area of $I + III = 1$ and the area of $II + III = 1$, we deduce that the area of III is $1 - d_{TV}(\nu, \mu)$. Our coupling works as follows. We sample a random variable uniformly in area I+III and X is set to be the abscissa of the selected point. If the point is in III we set $Y = X$. Otherwise, we sample a new point in the area II and set Y to be its abscissa. More formally, let

$$p = \sum_{i \in I} \min(\mu_i, \nu_i) = 1 - d_{TV}(\nu, \mu).$$

We flip a coin with probability of heads p.

(a) If the coin comes up heads, we choose value Z according to distribution

$$\gamma_{III}(i) = \min(\mu_i, \nu_i)/p,$$

and set $X = Y = Z$.

(b) Otherwise, we choose X according to distribution

$$\gamma_I(i) = \begin{cases} \frac{\mu(i) - \nu(i)}{d_{TV}(\nu, \mu)} & \text{if } \mu(i) > \nu(i) \\ 0 & \text{otherwise}. \end{cases}.$$

And we choose Y independently with distribution

$$\gamma_{II}(i) = \begin{cases} \frac{\nu(i)-\mu(i)}{d_{TV}(\nu,\mu)} & \text{if } \nu(i) > \mu(i) \\ 0 \quad \text{otherwise .} \end{cases} .$$

We have

$$p\gamma_{III}(i) + (1-p)\gamma_I(i) = \mu(i)$$
$$p\gamma_{III}(i) + (1-p)\gamma_{II}(i) = \nu(i).$$

So the distribution of X is μ and the distribution of Y is ν. With this construction, we have $\mathbb{P}(X = Y) = p$ and so $\mathbb{P}(X \neq Y) = 1 - p = d_{TV}(\nu, \mu)$. This ends the proof. □

13 Rate of convergence towards equilibrium

Theorem 13.1. *Assume that P is irreducible and assume that*

$$\beta(P) = \sum_{j\in I} \inf_{i\in I} P_{ij} > 0.$$

Then the Markov chain has a unique invariant distribution π and

$$d_{TV}((P^n)_{x\cdot}, (P^n)_{y\cdot}) \leq (1 - \beta(P))^n.$$

In addition, if (X_n) is an $MC(\lambda, P)$, we have:

$$d_{TV}(\mathcal{L}(X_n), \pi) \leq (1 - \beta(P))^n d_{TV}(\lambda, \pi).$$

Proof. We proceed in the following steps.

Step 1. Fix $x, y \in I$. We construct coupled Markov chains (that is, defined on the same probability space), such that (X_n^x) and (X_n^y) are respectively $MC(\delta_x, P)$ and $MC(\delta_y, P)$ and such that they become equal as quickly as possible. To construct the coupling, let (U_n) be a sequence of i.i.d. random variables which are uniformly distributed on $[0, 1]$. We also consider:

$$\nu(z) = \frac{\inf_{x\in I} P(x, z)}{\beta(P)}.$$

This is a probability measure on I and we have

$$\forall i \in I, \quad P_{ij} \geq \beta(P)\nu(j).$$

We set $X_0^x = x$ and $X_0^y = y$. We then proceed recursively as follows. On $\{U_1 \leq \beta(P)\}$, we choose $X_1^x = X_1^y \sim \nu(\mathrm{d}z)$. Moreover, on $\{U_1 > \beta(P)\}$, we choose

$$X_1^x \sim \frac{P(x, \mathrm{d}z) - \beta(P)\nu(\mathrm{d}z)}{1 - \beta(P)}.$$

And independently from this choice,

$$X_1^y \sim \frac{P(y, \mathrm{d}z) - \beta(P)\nu(\mathrm{d}z)}{1 - \beta(P)}.$$

We have:

$$\begin{aligned}\mathbb{P}(X_1^x = z) &= \mathbb{P}(X_1^x = z \mid U_1 \le \beta(P))\beta(P) + \mathbb{P}(X_1^x = z \mid U_1 > \beta(P))(1-\beta(P)) \\ &= \nu(z)\beta(P) + P(x,z) - \beta(P)\nu(z) = P(x,z).\end{aligned}$$

Similarly, the distribution of X_1^y is $P_{y,\cdot}$. If $X_1^x \neq X_1^y$, we continue in the same manner, using U_2 for the choice of X_2^x and X_2^y, and so on. We stop at the first time:

$$T_{\text{coupl}} = \inf\{n \ge 0 : X_n^x = X_n^y\}.$$

From that time on, we keep both trajectories together, evolving according to the transition probability matrix P. That is, for all $n \ge T_{\text{coupl}}$, we set $X_n^x = X_n^y$ in such a way that $(X_{n+T_{\text{coupl}}}^x)$ is an $MC(\delta_{X^x_{T_{\text{coupl}}}}, P)$, independently from the past up to T_{coupl}. By construction, we have

$$T_{\text{coupl}} \le \tau = \inf\{n \ge 1 : U_n \le \beta(P)\}.$$

Hence

$$d_{TV}(P^n_{x,\cdot}, P^n_{y,\cdot}) \le \mathbb{P}(X_n^x \neq X_n^y) \le \mathbb{P}(\tau > n) = (1-\beta(P))^n.$$

Step 2. We now consider arbitrary initial distributions $\nu, \mu \in \mathcal{P}(I)$. We consider (X_0, Y_0) an optimal coupling between ν, μ. That is: $\mathbb{P}(X_0 \neq Y_0) = d_{TV}(\nu, \mu)$. We have from the construction above:

$$\begin{aligned}d_{TV}(\mathcal{L}(X_n), \mathcal{L}(Y_n)) &\le \mathbb{P}(X_n \neq Y_n) \\ &\le \sum_{x,y\in I, x\neq y} \mathbb{P}(X_n \neq Y_n \mid X_0 = x, Y_0 = y)\mathbb{P}(X_0 = x, Y_0 = y) \\ &\le \sum_{x,y\in I, x\neq y} \mathbb{P}(X_n^x \neq X_n^y)\mathbb{P}(X_0 = x, Y_0 = y) \\ &\le \sum_{x,y\in I, x\neq y} (1-\beta(P))^n \mathbb{P}(X_0 = x, Y_0 = y) \\ &= (1-\beta(P))^n \mathbb{P}(X_0 \neq Y_0).\end{aligned}$$

Therefore, if (X_n) is an $MC(\nu, P)$ and (Y_n) a $MC(\mu, P)$ then

$$d_{TV}(\mathcal{L}(X_n), \mathcal{L}(Y_n)) \le (1-\beta(P))^n d_{TV}(\nu, \mu).$$

Step 3. The sequence $\mathcal{L}(X_n^x)$ is a Cauchy sequence in $\ell^1(I)$. As this space is complete, $\mathcal{L}(X_n^x)$ converges to some limit π. We verify easily that π is a probability measure on I and it is invariant for P.

□

14 Markov Chain Monte Carlo

Given an irreducible stochastic matrix P, we have given conditions such that P has a unique invariant distribution π. We now consider the inverse problem. Given a probability distribution π, can we find a stochastic matrix P such that π is the invariant distribution of the Markov chain driven by P?

Example 14.1. *Let $\{1,\cdots,q\}$ be a set of colors. Let $G=(V,E)$ be a finite graph. A proper q-coloring of G is a function $x:V\to\{1,\cdots,q\}$ such that*

$$(u,v)\in E \implies x(u)\neq x(v).$$

Question: how to sample uniformly in the set of proper q-colorings? It is difficult, if not impossible for large graphs to find the list of all valid q-colorings. Idea: construct a Markov chain taking values on $I=\{$ set of all valid q-colorings $\}$ such that the invariant measure of this Markov chain is the uniform measure on all proper q-colorings.

We shall see two different constructions: Metropolis and Gibbs. In both cases, one constructs a Markov chain (X_n), such that for large n, $\mathcal{L}(X_n)$ is close to the target distribution π.

Metropolis's algorithm

Let Ψ be any symmetric matrix on I. Therefore, Ψ is reversible with respect to the uniform measure on I. Goal: modify Ψ to obtain a chain with invariant distribution π. The chain works as follows: when at state x, consider a candidate move $y\sim\Psi(x,\cdot)$. Accept this new state with probability $a(x,y)$. With probability $1-a(x,y)$, reject and the new state is equal to the previous state x. Therefore, the new chain has transition probabilities given by

$$P(x,y)=\begin{cases}\Psi(x,y)a(x,y) & \text{if } x\neq y\\ 1-\sum_{z\in I,z\neq x}\Psi(x,z)a(x,z) & \text{if } x=y\end{cases}$$

This new chain has π as an invariant distribution if

$$\forall x,y\in I, x\neq y,\quad \pi(x)\Psi(x,y)a(x,y)=\pi(y)\Psi(y,x)a(y,x).$$

As Ψ is symmetric, this reduces to:

$$\forall x,y\in I, x\neq y,\quad \pi(x)a(x,y)=\pi(y)a(y,x).$$

We have $a(x,y)\le 1$ and $a(x,y)\le\frac{\pi(y)}{\pi(x)}$. We choose:

$$\boxed{a_*(x,y)=\min(1,\frac{\pi(y)}{\pi(x)}).}$$

We verify that $\pi(x)a_*(x,y)=\pi(y)a_*(y,x)$. Hence, π is the invariant probability measure of the Markov chain with transition probability P. Altogether, we have

$$P(x,y)=\begin{cases}\Psi(x,y)\left[1\wedge\frac{\pi(y)}{\pi(x)}\right] & \text{if } x\neq y\\ 1-\sum_{z\in I,z\neq x}\Psi(x,z)\left[1\wedge\frac{\pi(z)}{\pi(x)}\right] & \text{if } x=y\end{cases}$$

Remark 14.2. *Importantly, the chain only depends on the ratio $\frac{\pi(y)}{\pi(x)}$. In many cases of interest, π is only known up to a normalizing factor, i.e. $\pi(x)=\frac{h(x)}{Z}$ where Z is impossible to compute, as $Z=\sum_{j\in I}h(j)$ and $|I|$ is extremely large. We crucially don't need to compute Z to run the Metropolis chain.*

Example 14.3. *Consider a very large graph $G = (V, E)$. Let $f : V \to \mathbb{R}$ be a scalar function. Consider $\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z(\lambda)}$. As λ goes to infinity, π_λ puts more and more mass to the points which maximize f. Indeed, let $V^* = \{x \in V, \quad f(x) = f^* = \max_{y\in V} f(y)\}$. Then*

$$\lim_{\lambda\to\infty} \pi_\lambda(x) = \lim_{\lambda\to\infty} \frac{\lambda^{f(x)-f^*}}{|V^*| + \sum_{y\in V\setminus V^*} \lambda^{f(y)-f^*}} = \frac{\mathbb{1}_{\{x\in V^*\}}}{|V^*|}.$$

Our goal is to sample according to π_λ. For simplicity, assume that G is a regular graph. So the simple random walk on G has a symmetric transition matrix $\Psi(x, y)$. Then, the Metropolis algorithm works as follows. Assume that $X_n = x$. We choose a candidate state y uniformly among the neighbors of x ($y \sim \Psi(x, \cdot)$). If $f(y) \geq f(x)$ we accept y with probability 1. If $f(y) < f(x)$, we accept with probability $\lambda^{f(y)-f(x)}$.

General base chain. We can adapt the Metropolis's algorithm above when the initial transition matrix $\Psi(x, y)$ is not symmetric. Let Ψ be an irreducible probability matrix on I and let π be the target probability distribution. When at state x, generate $y \sim \Psi(x, \cdot)$. Move to y with probability

$$\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1.$$

Remain at x with the complementary probability. The transition matrix P is:

$$P(x,y) = \begin{cases} \Psi(x,y)\left[\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1\right] & \text{if } x \neq y \\ 1 - \sum_{z:z\neq x} \Psi(x,z)\left[\frac{\pi(z)\Psi(z,x)}{\pi(x)\Psi(x,z)} \wedge 1\right] & \text{if } x = y \end{cases}$$

We verify that P and π are in detailed balance. Indeed, let $x \neq y$. Without loss of generality, assume that

$$\pi(y)\Psi(y,x) \leq \pi(x)\Psi(x,y)$$

Then

$$\pi(x)P(x,y) = \pi(y)\Psi(y,x) = \pi(y)P(y,x),$$

as $\Psi(y, x) = \Psi(x, y)$.

Example 14.4. *Suppose you know neither the vertex set V nor the edge set E of a graph G. However, you are able to perform a simple random walk on G. (Many computer and social networks have this form; each vertex knows who its neighbors are, but not the global structure of the graph.) If the graph is not regular, then the stationary distribution is not uniform, so the distribution of the walk will not converge to the uniform distribution. We want to sample uniformly from V. We can use the Metropolis algorithm to modify the simple random walk and ensure a uniform stationary distribution. We have $\Psi(x, y) = \frac{\mathbb{1}_{\{(x,y)\in E\}}}{\deg(x)}$. Therefore, given a neighbor y of x chosen uniformly, we accept it with probability:*

$$\frac{\deg(x)}{\deg(y)} \wedge 1.$$

This biases the walk against moving to higher degree vertices, giving a uniform stationary distribution.

Gibbs sampler

In many examples in MCMC, the state space is of the form

$$I = \prod_{m \in \Lambda} S_m,$$

where Λ is a finite set (the sites) and S_m is a set of values for each site m. The elements of I are called the configurations. Given a probability distribution π on a space of configurations, we construct a Markov chain with invariant distribution equal to π. This chain is often called the Gibbs sampler, especially in statistical contexts. We start with an example.

Example 14.5. *A proper q-coloring of a graph $G = (V, E)$ is an element x of $\{1, 2, \cdots q\}^V$, the set of functions from V to $\{1, \cdots, q\}$ such that $x(v) \neq x(w)$ for all edges $(v, w) \in E$. We construct here a Markov chain on the set of proper q-colorings of G. Thus, here*

$$I = V^{\{1,\cdots,q\}}.$$

Given a configuration x and a vertex v, call a color j allowable at v if j is different from all colors assigned to neighbors of v. That is, a color is allowable at v if it does not belong to $\{x(w) \mid (v, w) \in E\}$. Given a proper coloring x, we can generate a new coloring by

1. *Select a vertex $v \in V$ at random.*
2. *Select at random a color j uniformly from the allowable colors at v.*
3. *re-color vertex v with color j.*

Proposition 14.6. *The resulting Markov chain has uniform stationary distribution over all the proper q-colorings.*

Proof. Transitions are only permitted between colorings differing at a single vertex. If two colorings x, y agree everywhere except at v, then the probability to move from x to y is

$$\frac{1}{|V|} \frac{1}{|A_v(x)|},$$

where $A_v(x)$ is the set of allowable colors at v in configuration x. We have $A_v(x) = A_v(y)$ therefore, $P(x, y) = P(y, x)$. So P and the uniform measure on all the proper q-coloring are in detailed balance. □

The Gibbs sampler is the following Markov chain. Let $x \in I$ be the current state. We first choose a site $m \in \Lambda$ at random (for instance, choose a vertex at random). Then the new state is chosen according to the measure π conditioned on the set of states equal to x at all sites different from m. That is, given $x \in I$ and $m \in \Lambda$, let

$$\mathcal{X}(x, m) = \{y \in I \mid y(w) = x(w) \text{ for all } w \neq m\},$$

be the set of all states agreeing with x everywhere except possibly at m. Define then

$$\pi^{x,m}(y) = \pi(y \mid \mathcal{X}(x, m)) = \begin{cases} \frac{\pi(y)}{\pi(\mathcal{X}(x,m))} & \text{if } y \in \mathcal{X}(x, m) \\ 0 \quad \text{otherwise.} \end{cases}$$

This is the distribution π conditioned on the set $\mathcal{X}(x,m)$. The rule for updating a configuration x is: pick a site m uniformly at random and choose a new configuration according to $\pi^{x,m}$. For $x \neq y$, the transition probability satisfies:

$$\pi(x)P(x,y) = \sum_{m\in\Lambda} \frac{1}{|\Lambda|} \frac{\pi(x)\pi(y)}{\pi(\mathcal{X}(x,m))} \mathbb{1}_{\{y\in\mathcal{X}(x,m)\}}.$$

As $y \in \mathcal{X}(x,m) \iff x \in \mathcal{X}(y,m)$, we have

$$\pi(y)P(y,x) = \sum_{m\in\Lambda} \frac{1}{|\Lambda|} \frac{\pi(x)\pi(y)}{\pi(\mathcal{X}(x,m))} \mathbb{1}_{\{y\in\mathcal{X}(x,m)\}} = \pi(x)P(x,y).$$

Hence, the chain is in detailed balance with π. The Gibbs sampler is particularly useful in Bayesian statistics. Instead of choosing the sites uniformly at random, we can also consider them all one after the other.

15 Application to Bayesian statistics

Assume that you observe the outcome of n i.i.d. random variables $Y = (Y_i)_{i\in\{1,\cdots n\}}$ such that

$$Y_i \sim \mathcal{N}(\mu, \tau^{-1}).$$

The objective is to estimate the unknown parameters $\mu \in \mathbb{R}$ and $\tau \in \mathbb{R}_+$ based on the observations Y. In Bayesian statistics, the main assumption is that the unknown parameters (μ,τ) are themselves random variables. The law of (μ,τ) is assumed to be known. This law is called the prior distribution. It encodes the prior knowledge on the parameters. For the sake of this example, assume that:

$$(\mu,\tau) \sim \mathcal{N}(\theta_0, \phi_0^{-1})\Gamma(\alpha_0, \beta_0),$$

where the four parameters $\theta_0, \phi_0^{-1}, \alpha_0, \beta_0$ are known. * That is, we have:

$$\mathbb{P}(\mu \in \mathrm{d}\mu, \tau \in \mathrm{d}\tau) \propto e^{-\phi_0 \frac{(\mu-\theta_0)^2}{2}} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau} \mathrm{d}\mu \mathrm{d}\tau.$$

We then wish to update our knowledge of (μ,τ) using the observation of the data Y. In other words, we wish to compute

$$\mathbb{P}(\mu, \tau \in \mathrm{d}\mu, \mathrm{d}\tau \mid Y = (y_1, \cdots, y_n)).$$

This new probability measure is called the posterior distribution. Typically, we would like to compute its mean and its variances/covariance. That gives the updated knowledge and uncertainty about the parameters (μ,τ). To compute the posterior distribution, one uses Bayes' rule, which states that

$$\mathbb{P}(\mu, \tau \in \mathrm{d}\mu, \mathrm{d}\tau \mid Y = y) = \frac{\mathbb{P}(Y \in \mathrm{d}y \mid \mu,\tau)\mathbb{P}(\mu,\tau \in \mathrm{d}\mu, \mathrm{d}\tau)}{\mathbb{P}(Y \in \mathrm{d}y)}.$$

Therefore, the posterior distribution is:

$$\mathbb{P}(\mu, \tau \in \mathrm{d}\mu, \mathrm{d}\tau \mid Y = y) \propto \tau^{n/2} \prod_{i=1}^{n} \exp(-\frac{\tau(\mu - y_i)^2}{2}) e^{-\phi_0 \frac{(\mu-\theta_0)^2}{2}} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau} \mathrm{d}\mu \mathrm{d}\tau.$$

*We recall that the Gamma distribution is $\Gamma(\alpha,\beta)(\mathrm{d}y) \propto y^{\alpha-1}e^{-\beta y}$.

This law is not of product form: for the posterior distribution, the two parameters μ, τ are not independent but rather depend in a complicated way on the observations. However, if we only compute the posterior distribution of a single marginal, then things are much easier. Indeed, we have:

$$\mathbb{P}(\tau \in \mathrm{d}\tau \mid Y = y, \mu) \sim \Gamma(\alpha_0 + n/2, \beta_0 + \sum_{i=1}^{n} \frac{(\mu - y_i)^2}{2}).$$

In addition, using the identity

$$a(\mu - y)^2 + b(\mu - z)^2 = (a + b)\left(\mu - \frac{ay + bz}{a + b}\right)^2 + \frac{ab}{a + b}(y - z)^2,$$

we find that

$$\mathbb{P}(\mu \in \mathrm{d}\mu \mid Y = y, \tau) \sim \mathcal{N}(\theta_n, \phi_n^{-1}),$$

where

$$\theta_n = \left(\phi_0\theta_0 + \tau \sum_{i=1}^{n} y_i\right) / (\phi_0 + n\tau)$$

and

$$\phi_n = \phi_0 + n\tau.$$

To sample $X = (\mu, \tau)$ according to the posterior distribution, the Gibbs sampler works as follows. The state space is:

$$I = \mathbb{R} \times \mathbb{R}_+ = S_1 \times S_2.$$

We first start with X_0 distributed, say from the prior distribution $X_0 \sim \mathcal{N}(\theta_0, \phi_0^{-1})\Gamma(\alpha_0, \beta_0)$. Then, at the k-th iteration, given $X_k = (\mu_k, \tau_k)$, we first simulate μ_{k+1} from $\mathbb{P}(\mu_{k+1} \in \mathrm{d}\mu \mid Y, \tau = \tau_k)$ (which is easy as it is a scalar normal distribution). Then we simulate τ_{k+1} with distribution $\mathbb{P}(\tau_{k+1} \in \mathrm{d}\tau \mid Y = y, \mu = \mu_{k+1})$, which is easy as it is a scalar Gamma distribution. Finally, we set:

$$X_{k+1} = (\mu_{k+1}, \tau_{k+1}).$$

Then (X_k) is a Markov chain with invariant probability measure equal to the posterior distribution. Therefore, we expect to have almost surely:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \int f(x)\mathbb{P}(\mu, \tau \in \mathrm{d}x \mid Y = y).$$

Therefore, by the ergodic theorem, we can compute the mean and the variance of the posterior distribution.

We have seen so far an example in dimension 2 (two unknown parameters). In practical applications, the number of unknowns can be very large. Assume for instance that you have $m \times n$ independent observations Y_{ij}, $i \in 1 \cdots n$, $j \in 1 \cdots m$ normally distributed with means μ_j and common variance τ^{-1}, with:

$$\mu_j \sim \mathcal{N}(\theta_0, \phi_0^{-1}), \quad \tau \sim \Gamma(\alpha_0, \beta_0).$$

Write $\mu = (\mu_1, \cdots, \mu_m)$. The prior density is

$$\pi(\mu, \tau) \propto \exp\left(-\frac{\phi_0}{2} \sum_{j=1}^{m} (\mu_j - \theta_0)^2\right) \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}.$$

The posterior distribution is:

$$\pi(\mu,\tau \mid y) \propto \exp\left(-\frac{\phi_0}{2}\sum_{j=1}^{m}(\mu_j-\theta_0)^2\right)\exp\left(-\tau\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\mu_j)^2/2\right)\tau^{\alpha_0-1+mn/2}e^{-\beta_0\tau}.$$

Therefore, one has:

$$\pi(\mu_j \mid y,\tau) \sim \mathcal{N}(\theta_{jn},\phi_n^{-1}), \quad \pi(\tau \mid y,\mu) \sim \Gamma(\alpha_n,\beta_n),$$

where

$$\theta_{jn} = \left(\phi_0\theta_0 + \tau\sum_{i=1}^{n} y_{ij}\right)/(\phi_0+n\tau), \quad \phi_n = \phi_0 + n\tau$$

$$\alpha_n = \alpha_0 + mn/2, \quad \beta_n = \beta_0 + \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\mu_j)^2/2.$$

The parameter space is:

$$I = \underbrace{\mathbb{R}\times\cdots\times\mathbb{R}}_{m \text{ times}}\times[0,\infty).$$

The Gibbs sampler works exactly as before. When m is large, the MCMC method is pretty much the only option to sample according to the posterior distribution.

Example 15.1. *This example is taken from the thesis of Leïla Haegel. The goal is to estimate the neutrino oscillation parameters, based on the T2K experiment. There are 6 oscillation parameters and 751 nuisance parameters to estimate. The prior distribution on the 6 oscillation parameters is chosen to be a product of the uniform measures to not induce a preference for certain values. The prior on the 751 other parameters is a Gaussian distribution. You can therefore look at her thesis to see a complicated, real world application of MCMC, to estimate the means and the standard deviations of the parameters, as well as the correlations.*