Statistical modeling for stochastic processes
spring 2024

SDS

Contents

1 Markov Chains

1.1 Overview

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Then, a **stochastic process** is a collection

$$\{X(t,\omega),\ t \in T\}$$

of random variables

$$X : T \times \mathscr{F} \longrightarrow S$$

where:

- T is the index set (typically interpreted as time);

Example 1.1

$T := \mathbb{Z}_+ = \{0, 1, 2, \ldots\}$: T is a *discrete time* ($\{X_n,\ n \in \mathbb{Z}_+\}$).

Example 1.2

$T := \mathbb{R}_+$:T is a *continuous time.*

- S is the state space. S can be countable: we can take $S \subseteq \mathbb{Z}$ and denote the elements of S as $i, j, k, \ldots$ or $i_1, i_2, i_3, \ldots$.

If $X_n = i$ we say that X is in state i or that X visits i at time n. Typically we drop the argument ω in $X(t, \omega)$ and just write $X(t)$. $X(\cdot, \omega)$ for a fixed ω is a function of t and is called trajectory of X.

More generally one could have:

- $T \subset \mathbb{R}^d$: **random fields** (e.g. geographical coordinates: X is a pair of latitude and longitude data and the state is the height data → $T \subset \mathbb{R}^2$);
- $S \subset \mathbb{R}$;
- $S \subset \mathbb{R}^d$: **multivariate processes**;

Definition 1.1

A **Markov Chain** (MC) is a discrete-time stochastic process $\{X_n, n \in \mathbb{Z}_+\}$ taking values in a countable space S, such that:

$$\begin{aligned} &\mathbb{P}(X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) && (1)\\ =&\mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}) \qquad \forall n \geqslant 1,\ \forall i_0, \ldots, i_n \in S. && (2)\end{aligned}$$

This property is called **Markov property.**

Markov property defines a very broad class of Markov process es. Given the current state X_{n-1}, the next X_n does not depend on the state previous to $n-1$. We could express these concepts in terms of σ-algebras and filtrations: if

$$\mathscr{F}_n^x = \sigma(x_k, k \leqslant n)$$

and

$$\{\mathscr{F}_n^x, n \in \mathbb{Z}_+\}$$

is the natural filtration, the previous reads:

$$\mathbb{P}(X_n = i_n | \mathscr{F}_{n-1}^x) = \mathbb{P}(X_n = i_n | X_{n-1}).$$

$\mathscr{F}_{n-1}^x$ renders useless the information previous to $n-1$.

From what we studied in our earlier courses, the main assumption in classical inference was the fat that the samples were i.i.d. But at a certain point we need a more sophisticated model that gives up this assumption. If the data are not i.i.d. we need to make a choice regarding the dependence between the data. Markov processes are highly studied because they offer a way of mathematically studying the dependence between samples (another common way is recurring to *Bayesian inference*).

Remark

The Markoviality is one way of departing from i.i.d. assumptions and all results that follow come from this property.

The dyamics of X are specified therefore by the right hand side of the Markov Property equation 1

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

which is called **transition probability**.
We will assume that these are temporary homogeneous, i.e. they do not depend on time. So we can denote

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

for any time t. We call j the *arrival state* ad i the *starting state*.

Remark

Spatial homogeneity is a much stronger assumption, not made here in general. In this case, p_{ij} only depends on j: $p_{ij} = p_{0,j-1}$. For example, if we analyzed the probability of moving between the tiles of a grid, spatial homogeneity would imply that moving from one tile to another is the same regardless to the starting tile.

We tipically collect the p_{ij} in a **transition matrix**:

$$P = (p_{ij})\ i, j \in S$$

Example 1.3

For $S = \{0, 1, \dots, k\}$ we have:

$$\begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0k} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{k0} & p_{k1} & \cdots & p_{kk} \end{bmatrix}$$

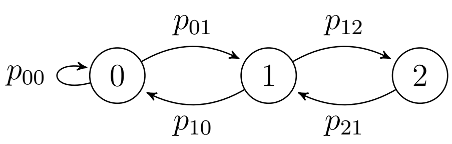
P is a stochastic matrix since:

- $p_{ij} \geqslant 0 \qquad \forall\, i, j \in S$;
- $\forall\, i \in S: \qquad \sum_{j \in S} p_{ij} = 1$.

Now i is then the conditional distribution of X_{n+1} given $X_n = i$.
We typically use a graph to represent a MC.

Example 1.4

$$P = \begin{bmatrix} p_{00} & p_{01} & 0 \\ p_{10} & 0 & p_{12} \\ 0 & p_{21} & 0 \end{bmatrix}$$



Is the transition matrix enough to derive everything about the Markov process? In other words, does the transition matrix fully characterize the distribution and the features of the process? The answer is yes but only if we include the **initial distribution** of the chain.

Definition 1.2

Define the **initial distribution** of the chain

$$\lambda_i = \mathbb{P}(X_0 = i) \qquad i \in S$$

as the law of the starting state.

If

$$\lambda_i = \delta_{ij}{}^{1} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

then $X_0 = j$ almost surely.

Proposition 1.1

Let X be a MC on S countable with initial distribution $\lambda = i \in S$ and transition matrix P. Then λ and P jointly fully charcterize the law of the chain.

Proof

What do we need to prove? We need to examine the joint distribution of a Markov Chain, which is an infinite sequence of random variables that are not i.i.d.: we can't therefore write the joint distribution as the product of the single distributions.
We need to show that (λ, P) allows us to compute all joint distributions for the chain, i.e. $\forall$ choices of $0 \leqslant j_1 < j_2 < \ldots < j_k$ the law $\mathbb{P}(X_{j_1} = i_{j_1}, \ldots, X_{j_k} = i_{j_k})^{a}$, which has k states, can be computed with (λ, P) only. We can use the previous as the marginal of the joint distribution for times $0, \ldots, j_k$:

$$\underbrace{0, 1, \ldots, j_{1-1}}_{\text{marginalize}}, j_1, \underbrace{j_{1+1}, \ldots,}_{\text{marginalize}} j_2, \underbrace{\ldots}_{\text{marginalize}}, j_k$$

i.e. it is

$$\sum_{\underbrace{i_h \in S | h \neq j_1, \ldots, j_k}_{\text{indices of times not chosen: } h \in \{o, \ldots, j_k\}}} \mathbb{P}(\underbrace{X_0 = i_0, X_1 = i_1, \ldots, X_{j_1} = i_{j_k}}_{j_k + 1 \text{ states}})$$

So it is enough to find $\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n)$, but by the chain rule this equals to

$$\mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \cdot \mathbb{P}(X_n = i_n | X_0, \ldots, X_{n-1})$$

and thanks to Markov Property this is equal to

$$\begin{aligned} &\mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \cdot p_{i_{n-1}, i_n} \\ = \ldots = &\underbrace{\mathbb{P}(X_0 = i_0)}_{\lambda_{i_0}} \cdot p_{i_0, i_1} \cdot \ldots \cdot p_{i_{n-1}, i_n} \end{aligned}$$

So we only need λ_{i_0} and the P matrix. □

[a] This is called a projection: we are projecting an infinite sequence on a finite subset.

[1] Kronecker's delta.

We can use a compact notation (took from Norris textbook)

$$X \sim \text{Markov}(\lambda, P).$$

Definition 1.3

We define the ***k*-step transition probabilities** as

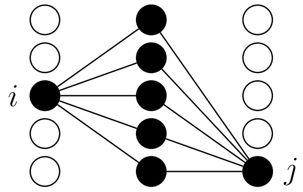
$$p_{ij}^{(k)} := \mathbb{P}(X_{n+k} = j | X_n = i), \qquad k \in \mathbb{N}$$
$$p_{ij}^{(0)} := \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$$

Proposition 1.2

Chapman-Kolmogorov equations: for all e, $m \in \mathbb{Z}_+$

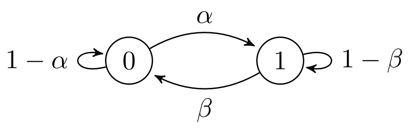
$$p_{ij}^{(e+m} = \sum_{k<j} p_{ik}^{(e)} \cdot p_{kj}^{(m)} \tag{3}$$

The proof is left as exercise (use graphical intuition below: marginalize out unwanted states and use Markov property).



If you take the transition matrix to the power of k, P^k is still a stochastic matrix and the entries are $p_{ij}^(k)$.

Example 1.5



$$P^2 = P \cdot P = \ldots = \begin{bmatrix} p_{00}^{(2)} & p_{01}^{(2)} \\ p_{10}^{(2)} & p_{11}^{(2)} \end{bmatrix}$$

In this case, the intermediate states don't matter. For instance we have $p_{00}^{(2)} = p_{00}p_{00} + p_{01}p_{10}$

1.2 Notable Markov processes

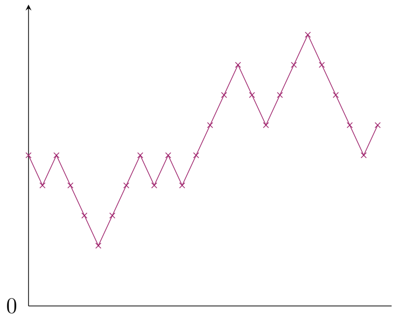
Random Walk

Given X_0, the simple random walk is defined as

$$X_n = X_{n-1} + Y_n$$

Y_ns are i.i.d.:

$$Y_n = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$



if $p = \frac{1}{2}$ the the random walk is symmetric.
Random walks have application, for instance, in:

- approximation of Brownian motion and other diffusion processes useful in:
  - physics;
  - math finance;
  - math biology;
- random explorations of space:
  - Monte-Carlo integration (Monte-Carlo Markov Chain methods);
  - stochastic optimization
  - integral approximation

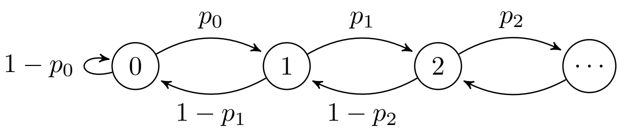
Some possible extensions are:

- changing the state space dimension: imagine a symmetric random walk in $\mathbb{Z}^d$, $i, j \in \mathbb{Z}^d$ with $i = (i_1, \dots, i_d)$:
$$p_{ij} = \begin{cases} \frac{1}{2d} & \text{if} \quad \sum_{k=1}^{q} |i_k - j_k| = 1 \\ 0 & \text{else;} \end{cases}$$
- change law of Y_n : $\mathbb{P}(Y_n = 1) = a_i, \ \in \mathbb{Z}$ (the case of the distribution being heavy tailed is particularly interesting);
- change the topology of S:
  - Random walk on manifold sphere, on torus...;
  - Random walk on graphs; for example choose uniformly with probability q a node in the graph (Google PageRank Algorithm).

Birth-and-death chains

We define the transition probabilities as

$$p_{ij} = \begin{cases} p_i & j = i+1 \\ 1-p_i & j = (i-1)^+ \\ 0 & \text{else} \end{cases}$$



- Time-varying size of a population (animal, viruses, numbero of request to a CPU...)

- size of a queue at a server
- dimension of a multivariate distribution used for estimation
- dimension k of a mixture model

$$\text{i.e.} \qquad \sum_{i=k}^{k} w_i f_i$$

with $\sum_{w_i} = 1$ and f_i being a density.

Plya urns

An urn contains W_0 white balls and B_0 black a ball. We draw a ball, check its colour, put it back and add to the urn another ball of the same colour: this behaviour is called reinforcement and it is the opposite of drawing without placement. Let W_n be the number of white balls after n draws (steps):

$$\mathbb{P}(W_{n+1} = j|W_0, \dots, W_n) = \begin{cases} \frac{W_n}{Wn+Bn} \rightarrow \text{"draw white"} & j = W_n + 1 \rightarrow \text{"add white"} \\ \frac{B_n}{Wn+Bn} \rightarrow \text{"draw black"} & j = W_n \rightarrow \text{"same number of whites"} \\ 0 & \text{else} \end{cases}$$

$$\begin{aligned} W_n + B_n &= W_0 + B_{0+n} \\ B_n &= W_0 + B_{0+n} - W_n \end{aligned}$$

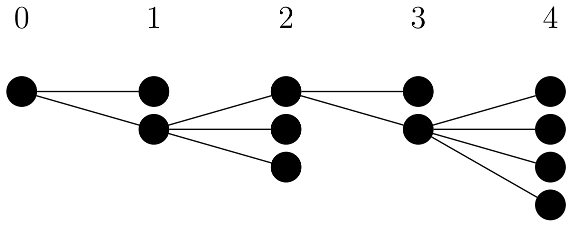
The probability only depends on W_n and the previous Ws are irrelevant, so the process is a Markov Chain These are applied in Bayesian inference, since they generate exchangeable sequences $((X_1, X_2) \overset{d}{=} (X_2, X_1))$. A classical result which is often used is:

$$\frac{W_n}{W_n + B_n} \xrightarrow{a.s} \theta \sim Beta(W_0, B_0)$$

Branching Processes

Branching processes are a simple model for evolving populations with reproduction. The basic formulation (Galton-Watson) is:

- individuals live on period;
- individuals generate clones independently (the original motivation was the survival of family names);
- the next generation is formed by the offspring of the previous generation.



Useful applications:

- epidemiology (virus contagion);
- population genetics;
- physics (random number of neutron produced at collisions).

Interesting extensions:

- extend to k types;
- add immigration.

Let X_n be the number of individuals in generation n and Y_i be the number of clones/offspring of individual i. We have $Y_i \overset{i.i.d.}{\sim} p_Y$ on $\mathbb{Z}_+$.

$$X_n = Y_1 + Y_2 + \ldots + Y_{X_{n-1}} = \sum_{i=1}^{X_{n-1}} Y_i \perp\!\!\!\perp X_{n-2}, X_{n-3}, \ldots$$

So the process is a Markov chain.

Proposition 1.3

Branching Property: Denote by $X^{(i)} = \{X_n, n \in \mathbb{Z}_+ | X_0 = 1\}$, characterizing the chain by a fixed starting point, the branching process started at i. Let $\tilde{X}, \hat{X}$ be independent copies of X. Then

$$X^{(i+j)} \overset{d}{=} \tilde{X}^{(i)} + \hat{X}^{(j)} \tag{4}$$

This is called **branching property**.

Proof

Let

$$g_z(s) = \mathbb{E}\Big[e^{sZ}\Big]$$

be the moment generating function of Z. We are interested in the moment generating function of $X_1 | X_0 = i$.

$$g_{X_1^{(i)}} = g_{\sum_{j=1}^{i} Y_j} = (g_Y)^i$$

This moment generating function characterizes the one step transition probability of getting to X_1, that is the population size at $n = 1$, from the state $X_0 = i$. Now

$$\begin{aligned} g_{\{\tilde{X}_1^{(i)} + \hat{X}_1^{(j)}\}} &= g_{\tilde{X}_1^{(i)}} \cdot g_{\hat{X}_1^{(j)}} = \\ = (g_Y)^i \cdot (g_Y)^j &= (g_Y)^{i+j} = g_{X_1^{(i+j)}} \end{aligned}$$

We have now proved for one step that, since $\tilde{X}_1^{(i)} + \hat{X}_1^{(j)}$ and $X_1^{(i+j)}$ share the same moment generating function, they have the same distribution. To extend this result to every step we can use Markov property, so that $X_1^{(i+j)} \overset{d}{=} \tilde{X}_1^{(i)} + \hat{X}_1^{(j)}$ characterizes the law of the entire chain. □

An interesting extension of the model consists in the behaviour of a population during extinction. This ultimately depends on:

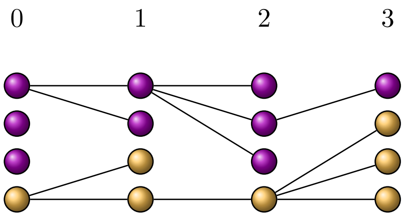
$$\mathbb{E}\Big[Y\Big] = \begin{cases} < 1 & X \text{ is subcritical} \\ = 1 & X \text{ is critical} \\ > 1 & X \text{ is supercritical} \end{cases}$$

An interesting case is given by introducing immigration to the subcritical case. This is called **Galton-Watson** branching process.

Wright-Fisher models

These model the time-varying frequency of 2 types (in general k or ∞ many types) in an evolving population of constant size. The original motivation was modelling the allelic type frequency of a gene locus. These models have the following characteristics:

- individuals live 1 period;
- individuals can be of type 0 or 1;
- population size is N $\forall n \geqslant \mathbb{Z}_+$;



We have now constrained the population size, so each parent can generate from 0 to N offspring and therefore birth events are no longer independent. A useful approach to model this behaviour is to imagine that during next generation individuals choose their parent at random from the previous generation, with the condition that it must be of the same type.
Let X_n be the number of type-0 individuals at time n. The individuals at generation n are:

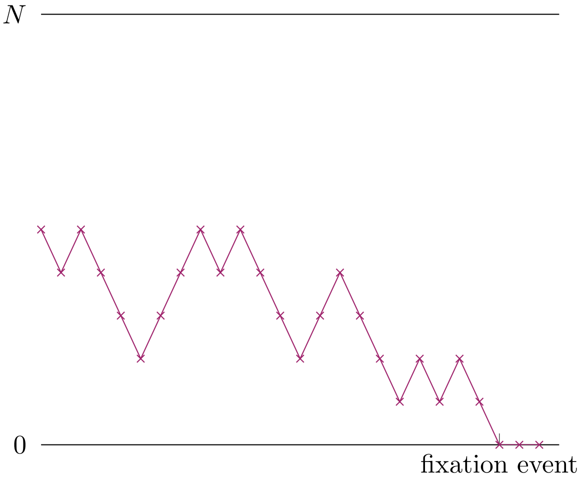
- of type-0 with probability $\frac{X_n}{N}$;
- of type-1 with probability $1 - \frac{X_n}{N}$;

So every member of a generation chooses its parent by a Bernoulli trial independently. Since we have N Bernoulli trial with parameter $p = \frac{X_n}{N}$, we have that:

$$X_n|X_{n-1} \sim Binom(N, \frac{X_{n-1}}{N}).$$

Note that

$$\begin{aligned} &\text{if} \quad X_{n-1} = 0 \implies X_n = 0 \qquad \text{a.s} \\ &\text{if} \quad X_{n-1} = N \implies X_n = N \qquad \text{a.s.} \end{aligned}$$



Some useful extension may include:

- mutations;
- $k \leqslant \infty$ types;
- uncountably many types.

For example, set mutations such that they only occur to the other type:

$$\alpha = \mathbb{P}(1 \to 0) \qquad \beta = \mathbb{P}(0 \to 1).$$

Mutations take place before reproduction (that is, before the binomial resampling method). We have:

- a $Bernoulli(\beta)$ trial on each of i individual of type 0. On average:
    - $i\beta$ individuals become of type 1;

- $i(1-\beta)$ individuals remain of type 0;

- a $Bernoulli(\alpha)$ trial on each of $N-i$ individual of type 1. On average:
  - $(N-i)\alpha$ individuals become of type 0;
  - $(N-i)(1-\alpha)$ individuals remain of type 1.

After mutation, the expected proportions are:

$$\begin{aligned} \text{type 0:} \qquad \tilde{p}_i &= \frac{1}{N}(i(1-\beta)+(N-i)\alpha) = \\ &= \frac{i}{N}(1-\beta)+(N-i)\alpha = \\ &= p_i(1-\beta)+(1-p_i)\alpha. \\ \\ \text{type 1:} \qquad 1-\tilde{p}_i &= p_i\beta+(1-p_i)(1-\alpha). \end{aligned}$$

We can interpret $\tilde{p}_i$ as the percentage of type 0 individuals (in expectation) before resampling. The transition probability has become:

$$= 0p_{ij} = \mathbb{P}(X_{n+1}=j|X_n=i) = \binom{N}{j}\tilde{p}_i^j(i-\tilde{p}_i)^{N-j}$$

Now, if $i=0 \implies p_i = \frac{i}{N} = 0 \implies \tilde{p}_i = \alpha \implies \mathbb{P}(X_{n+1}>0|X_n=0) = 1-(1-\alpha)^N > 0$.
If X_n is at the boundary $X_n \in \{o, N\}$, now there is positive probability of going back to the interior of the space state.

Transformation of Branching Processes

Intuitively, there are two branching processes embedded in the Wright-Fisher trajectory. We can formalize this connection:

Proposition 1.4

Let X and W be two independent branching processes with the same Poisson offspring distribution. Then, conditional on the total population size being constant and equal to N, X and $W = N - X$ are Wright-Fisher chains.

Proof

$$\begin{aligned} X_n &= \text{n of type 0 individuals} \\ W_n &= \text{n of type 1 individuals} \end{aligned}$$

With per capita offspring

$$Y \sim Pois(\lambda_j) \qquad j = 0, 1$$

where j is specific to the first and to the second branching process. Now, since the sum of Poisson variables is still a Poisson variable,

$$\begin{aligned} X_n &= \sum_{i=1}^{X_{n-1}} Y_i \sim Pois(X_{n-1}\lambda_0) \\ W_n &= \sum_{i=1}^{W_{n-1}} Y_i \sim Pois(W_{n-1}\lambda_1) \end{aligned}$$

We know, in general, that

$$Z_j \sim Pois(\gamma_j),\ j=0,1 \implies Z_0|Z_0+Z_1=N \sim Binom\left(N, \frac{\gamma_0}{\gamma_0+\gamma_1}\right)$$

Let's apply this property:

$$X_n|X_n+W_n=N, X_{n-1}, W_{n-1} \sim Binom(N, q_{n-1})$$

What we expect is for the parameter q_{n-1} to only depend on state $n-1$. We know that:

$$q_{n-1}=\frac{X_{n-1}\lambda_0}{X_{n-1}\lambda_0+\underbrace{W_{n-1}}_{N-X_{n-1}\text{ because we are conditioning on the whole population}}\lambda_1}=\frac{X_{n-1}\lambda_0}{X_{n-1}(\lambda_0-\lambda_1)+N\lambda_1}$$

But we said that the offspring distribution is the *same* Poisson distribution: so

$$\lambda_0=\lambda_1 \implies q_{n-1}=\frac{X_{n-1}}{N}$$
$$\implies X_n|X_{N-1}, X_n+W_n=N \sim Binom\left(N, \frac{X_{n-1}}{N}\right)$$

which is the transition probability of the Wright-Fisher chains, which characterizes the law of the entire chain. We are therefore claiming that the random variables whose state space is the ∞ trajectories of the chain is governed by one law. □

1.3 Review of Markov Property

The property we gave that the beginning can be seen as a regeneration property:

$$X \sim Markov(\lambda, P) \implies \{X_{n+k}, k \geqslant 0|X_n=i\} \sim Markov(\delta_i, P).$$

Here n is fixed and from n onward the chain starts afresh with the same properties. What if n is a random variable T, indicating a random time?

Definition 1.4

strong Markov property: We are interested in establishing the **strong Markov property**:

$$\mathbb{P}(X_{T+1}=j|X_0,\dots,X_T=i)=\mathbb{P}(X_{T+1}=j|X_T=i) \tag{5}$$

Not all random times are suitable: let $\{\mathscr{F}\}_{n\geqslant 0}$ be the natural filtration generated by X, where $\mathscr{F}_n=\sigma(X_u, 0\leqslant u\leqslant n)$ which is interpreted as the *flow of information* generated by the X trajectory up to time n.
A random variable $T:\Omega \to \mathbb{N}\cup\{\infty\}$ is a **stopping time** for X if

$$\{\omega \in \Omega : T(\omega)=n \in \mathscr{F}_n\}$$

i.e. $\{T=n\}$ is $\mathscr{F}_n$-measurable: we can express $\{T=n\}$ in terms of $X_0, X_1, \dots, X_n$. If we are interested in other relations:

$$\{T=n\}\in\mathscr{F}_n \implies \begin{cases}\{T\leqslant n\}=\bigcup_{i\leqslant n}\{T=n\}\in\mathscr{F}_n\\ \{T\neq n\}=\{T=n\}^c\in\mathscr{F}_n\\ \dots\end{cases}$$

Example 1.6

First passage time/first visit to i:

$$T = \inf\{n \geqslant 1 : X_n = i\}.$$

Is T a stopping time? The event $\{T = n\} = \{X_0 \neq i, X_1 \neq i, \dots, X_{n-1} \neq i, X_n = i\}$ by definition belongs to $\mathscr{F}_n$, so it is a stopping time.

Example 1.7

Last exit time from A ⊂ S

$$T = \sup\{n \geqslant 0 : X_n \in A\}.$$

Is T a stopping time? The event depends on $\{X_{n+m}, m \geqslant 1\}$. $\{T = m\}$ does not belong to $\mathscr{F}_n$ because it is anticipating using future information: T is not a stopping time.

Theorem 1.1

Strong Markov Property: Let $X \sim Markov(\lambda, P)$ and T be a stopping time for X. Then, conditional on $T < \infty$ and $X_T = i$,

$$\{X_{T+n}, n \geqslant 0\} \sim Markov(\delta_i, P)$$

is **independent** on the chain before time T.

If T is a stopping time (meaning that it depends on the past only) the the chain regenerates from T with the same properties.

Example 1.8

Let T be the first visit to j and $T' = T - 1$ the time prior to entering j for the 1st time. $\{T' = n\}$ depends on X_{n+1} which implies that $\{T' = n\} \notin \mathscr{F}_n$. Conditional to $X_{T'} = i$ the process is not Markov:

$$\{X_{T'+n}, n \geqslant 0\} \nsim Markov(\delta_i, P)$$

Why? The problem is that if $n = 1$ then the first step of the chain is deterministic, since we are imposing it to be j instead of following the transition matrix:

$$\mathbb{P}(X_{T'+1} = j | X_{T'} = i) = 1 \neq p_{ij} \qquad \text{since} \quad X_T = X_{T'+1} = j$$

so the chain does not start afresh with the same distribution properties.

Example 1.9

Suppose we observe X only when $X_n \in A \subset S$ (for instance, imagine we have an instrument only capable of measuring above a certain threshold A). $T^{(m)} = inf\{n > T^{(m)} : X_n \in A\}$ is the time of the n-th visit to A and $Y_m := X_{T^{(m)}}$.

Assume $\mathbb{P}(T^{(m)} < \infty) = 1 \quad \forall m \geqslant 1$:

$$\begin{aligned}
&\mathbb{P}(Y_{m+1} = i_{m+1} | Y_1 = i_1, \ldots, Y_m = i_m) = \\
=&\mathbb{P}(X_{T^{(m+1)}} = i_{m+1} | X_{T^{(1)}} = i_1, \ldots, X_{T^{(m)}} = i_m) = \\
&\text{(by strong Markov property)} \\
=&\mathbb{P}(X_{T^{(m+1)}} = i_{m+1} | X_{T^{(m)}} = i_m)
\end{aligned}$$

which means that Y is a Markov Chain on A. If I have a Markov chain only observable on a subset then I can "skip" intermediate steps and still have a Markov chain. This is pretty cool.

1.4 Properties of Markov chains

The goal of this section is to find the conditions on P to claim certain properties of Markov Chains, especially concerning long-run behaviour.

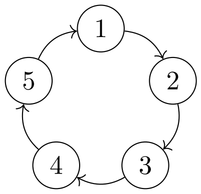
1.4.1 Communication classes

Definition 1.5

A state is **accessible** from i if $p_{ij}^{(n)} > 0$ for some n. Two states reciprocally accessible are said to **communicate**.

If a state is accessible from another state it means that the chain can go there in a finite number of steps.

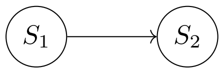
Example 1.10



In this case all the states communicate, since it is possible to go from one to any other with at most $5 < \infty$ steps.

Communicating states form an **equivalence class**: the relation is reflective, symmetric and transitive. This can be proved using the Chapman-Kolmogorov equation (Eq. 3). S can be split in two or more classes of communicating states, thus obtaining a **macroscopic description** of the chain dynamics.

Example 1.11



$S_1, S_2 \subset S$

In this case, once the chain leaves S_1 it can't go back. S_2 can be accessed by S_1 but not vice-versa. On the long run we can say that we should focus on S_2, since S_1 will be left for good sooner or later.

1.4.2 Irreducibility

Definition 1.6

A Markov chain with a single equivalence class generated by communicating states is said to be **irreducible**.

For instance, in the previous example the chain is not irreducible but once the chain leaves S_1 it becomes irreducible in S_2.
What about the main Markov processes?

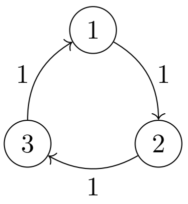
- The simple random walk is irreducible: we can always go in every state;
- in branching processes, if $X_n = 0$ then the process stops and $X_{n+1} = 0$. This means that $\{1, 2, \dots\}$ are not accessible from 0 and the chain is therefore not irreducible;
- in Wright-Fisher process with no mutations we face a similar situation: $X_n = 0 \implies X_{n+1} = 0$ and $X_n = N \implies X_{n+1} = N$: $\{1, \dots, N-1\}$ are not accessible from $\{0, N\}$ and the chain is therefore not irreducible. The introduction of mutations, though, would make it irreducible.

1.4.3 Periodicity

Definition 1.7

We define the **period** $d(i)$ of $i \in S$ the greatest common divisor of all $n \geqslant 1$ such that $p_{ii}^{(n)} > 0$. If $p_{ii}^{(n)} > 0$ for all sufficiently large n (or if $d(i) = 1$) we say that i is **aperiodic**.

Example 1.12



$\forall i = 1, 2, 3:$

$$p_{ii}^{(3n)} = 1 \qquad \forall n + 1$$
$$p_{ii}^{(3n+1)} = 0 \qquad \forall n \geqslant 1$$
$$p_{ii}^{(3n+2)} = 0 \qquad \forall n \geqslant 1$$
$$d(i) = 3 \qquad i \in S$$

Periodicity is a **class property**: if i, j communicate then they have the same period. This means that the period of the communication class S_1 is enough to study the period of a single $i \in S_1$.

Example 1.13

The simple random walk is irreducible: $S = \mathbb{Z}$ is a single communicating class. For $i > 0$:

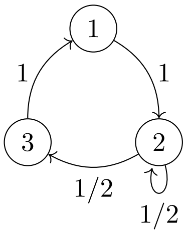
$$p_{00}^{(2n)} > 0 \qquad p_{00}^{(2n+1)} = 0 \qquad n \geqslant 0$$

so the period is 2.

Proposition 1.5

If X is irreducible and $p_{ii} > 0$ for some $i \in S$ then X is aperiodic.

Example 1.14



The chain is aperiodic: I can get from 3 back again to 3 in 3 steps of in 4,5,6,... if I cycle in 2.

$$p_{33}^{(n)} > 0 \qquad n \geqslant 3.$$

The greatest common denominator is 1: $d(3) = 1$

We can exploit this property: if P is periodic, define

$$P' = \varepsilon P + (1-\varepsilon)I, \qquad\qquad \varepsilon \in (0,1)$$

This is called the "lazy chain" because the chain is not going to move from its state in n with probability ε and P' is aperiodic. We will see that this modification essentially does not alter the distributional properties of the chain.

1.4.4 Recurrence

Recall now the first visit to i

$$T_i = \inf\{n \geqslant 1 : X_n = i\}$$

Definition 1.8

A state i is said do be **recurrent** if

$$\mathbb{P}(T_i < \infty | X_0 = i) = 1$$

, or equivalently

$$\mathbb{P}(T_i < \infty \quad \text{for some} \quad n | X_0 = i) = 1$$

which means that the return time is finite almost surely. The state i is otherwise said to be **transient**.

A transient i is such that

$$\mathbb{P}(T_i < \infty | X_0 = i) < 1.$$

It is worth noting that a transient state still has a positive probability of coming back on ∞.

Proposition 1.6

Let $p_{ii}^{(n)}$ be the return probability to i in n steps. Then $i \in S$ is recurrent if and only if

$$\sum_{n \geqslant 1} p_{ii}^{(n)} = \infty$$

and it is transient otherwise.

Let $I_n = \mathbb{1}(X_n = i)$:

$$\begin{aligned}\sum_{n\geqslant 1} p_{ii}^{(n)} &= \sum_{n\geqslant 1} \mathbb{P}(X_n|X_0 = i)\\ &= \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}\Big[I_n|X_0 = i\Big]\\ &= \lim_{N\to\infty} \mathbb{E}\Big[\underbrace{\sum_{n=1}^{N} I_n}_{f_N}|X_0 = i\Big]\\ &= \mathbb{E}\Big[\sum_{n=1}^{\infty} I_n|X_0 = i\Big]\\ &\text{(by monotone convergence theorem)}\end{aligned}$$

So i is recurrent if in expectation it is visited ∞ many times over the whole time horizon. We can define

$$G := \sum_{n\geqslant 0} p^n$$

which is sometimes called potential matrix.

Proposition 1.7

A state i is visited almost surely:

- infinitely often if recurrent;
- finitely often if transient

Intuition: $T_i^{(1)}$ is the first passage time to i. If i is recurrent then by definition

$$\mathbb{P}(T_i^{(1)} < \infty|X_0 = i) = 1.$$

Since $X_{T_i^{(1)}} = i$, from the strong Markov property we get:

$$X' = X_{T_i^{(1)}+n} \sim Markov(\delta_i, P)$$

so $X_{T_i^{(2)}}$ for X is the first passage time for X', which implies

$$T_i^{(1)} \overset{d}{=} T_i^{(2)} \implies \mathbb{P}(T_i^{(2)} < \infty|X_{T_i^{(1)}} = i) = 1.$$

So over an infinite time horizon we have infinitely many visits with probability 1.

Recurrence is a class property, so if a chain is irreducible it is enough to check one state. If we checked that i is recurrent all states that communicate with i will be visited infinitely many times.

Proposition 1.8

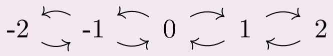
The simple random walk on $\mathbb{Z}$ is recurrent if and only if $p = \frac{1}{2}$, that is if the walk is symmetric.

Proof

To check that the chain is irreducible, it is enough to check that $i = 0$ is irreducible:

$$p_{00}^{(2n)} > 0 \qquad\qquad p_{00}^{(2n+1)} = 0 \qquad n \geqslant 1$$

so it is enough to check the even number of steps:



I sample right and left steps n times each:

$$p_{00}^{(2n)} = \binom{2n}{n} p^n (1-p)^n$$

Verify using Stirling's approximation that

$$n! \approx \text{const} \times n^{n+\frac{1}{2}} e^{-n} \implies p_{00}^{(2n)} \approx \frac{[4p(1-p)]^n}{\sqrt{n}}$$

so

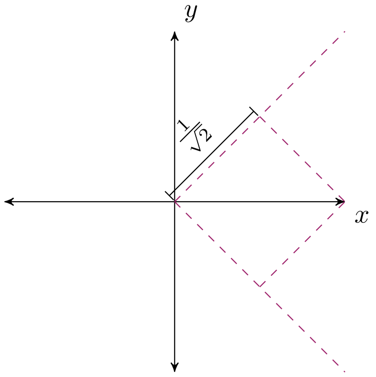
$$\sum_{n\geqslant 0} p_{00}^{(n)} \approx \sum_{n\geqslant 0} \frac{[4p(1-p)]^n}{\sqrt{n}}$$

$$\to \begin{cases} < \infty & 4p(1-p) < 1 = p(1-p) < \frac{1}{4} \text{ which means } p \neq \frac{1}{2} \\ = \infty & p = \frac{1}{2} \end{cases}$$

which means that the series is infinite when $p = 1/2$. □

What about a random walk on $\mathbb{Z}^d$?

$$d = 2, \; i, j \in \mathbb{Z}^2 \qquad p_{ij} = \begin{cases} \frac{1}{4} & \text{if } \;\; |i_1 - j_1| + |i_2 - j_2| \\ 0 & \text{else.} \end{cases}$$



The random walk on $\mathbb{Z}^2$ can be seen as 2 independent random walk on the diagonals of the cartesian plane.

$$p_{(0,0),(0,0)}^{(2n+1)} = 0$$

$$p_{(0,0),(0,0)}^{(2n)} = \Bigg[\underbrace{\binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n}_{\approx\frac{1}{\sqrt{n}}}\Bigg]^2$$

$$\sum_{n\geqslant 0} p_{(0,0),(0,0)}^{(n)} = \infty \implies \text{the chain is recurrent.}$$

Consider now a symmetric random walk on $\mathbb{Z}^3$:

$$\sum_{n\geqslant 0} p_{(0,0,0),(0,0,0)}^{(n)} = \Bigg[\underbrace{\binom{2n}{n}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^n}_{\approx\frac{1}{\sqrt{n}}}\Bigg]^3 \approx \frac{1}{n^{\frac{3}{2}}}$$

This is a convergent series, which means that from 3 dimensions onward the random walk *is transient*.

In the transient case we have:

$$\sum_{n\geqslant 1} p_{ii}^{(n)} < \infty \implies p_{ii}^{(n)} \xrightarrow[n\to\infty]{} 0$$

It can be proved that $\forall j \in S$, $p_{ii}^{(n)} \xrightarrow[n\to\infty]{} 0$. In the recurrent case, where we have $\sum p_{ii}^{(n)} = \infty$:

- $p_{ii}^{(n)} \to c > 0$, which causes the *divergence* of the series;
- $p_{ii}^{(n)} \to 0$ but slowly, not faster than $\frac{1}{n}$, causing divergence in this case as well.

But is this dichotomy useful? define

$$m_i = \mathbb{E}\Big[T_i | X_0 = i\Big]$$

as the **mean return time to i**. If i is transient, then

$$\mathbb{P}(T_i < \infty | X_0 = i) < 1 \implies \mathbb{P}(T_i = \infty | X_0 = i) > 0 \implies m_i = \infty$$

.

Definition 1.9

A recurrent state i is called:

- **positive** if $m_i < \infty$
- **null** if $m_i = \infty$

and these are *class properties*.

Later we will show that a null recurrent state i is such that $p_{ji}^{(n)} \to 0 \ \forall j \in S$, so transient and null/positive recurrence can be understood in terms of the tail of the distribution of the return times:

- positive probability at ∞ → transience
- probability mass is on $\mathbb{N}$ but it is not integrable (”heavy tailed”) → null recurrence
- probability mass is on $\mathbb{N}$ and it is integrable → positive recurrence

Example 1.15

The symmetric random walk on $\mathbb{Z}$ and $\mathbb{Z}^2$ are the only recurrent cases:

$$p_{00}^{(2n)} \propto \frac{1}{\sqrt{n}} \qquad\qquad p_{(0,0),(0,0)}^{2n} \propto \frac{1}{n}$$

Both tend to 0 as n tends to ∞, which means that both the states are null recurrent.

We are now interested in establishing how a chain can be classified as positive recurrent.

Proposition 1.9

On a finite S, an irreducible chain is positive recurrent.

Proof

$$S = \{0, \dots, N\}$$

$$\forall n \geqslant 1: \qquad \sum_{j=0}^{N} p_{ij}^{(n)} = 1.$$

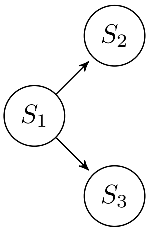
This means that we cannot have $p_{ji}^{(n)} \to 0 \; \forall j \in S$. So there exist a state $i \in S$ that is both positive and irreducible and therefore X must be positive □

Example 1.16

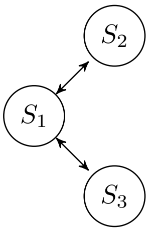
Consider a Wright-Fisher chain with $S = 0, \dots, N$. It has:

$$X_n = 0 \implies X_{n+1} = 0$$
$$X_n = \implies X_{n+1}$$

so S is finite but X is not irreducible.



Case without off mutations



Case with odd mutations;

More generally, the task can be difficult. Often, so-called *Lypaunov methods* are useful as a sufficient conidition.

Proposition 1.10

Let P be irreducible and let $h : S \to \mathbb{R}$ be such that $h(i) \geqslant 0 \; \forall i \in S$ and:

- $\sum_{k \in S} p_{ik} \cdot h(k) < \infty \qquad \forall i \in S_0$
- $\sum_{k \in S} p_{ik} \cdot h(k) \leqslant h(i) \cdot \varepsilon \qquad \forall i \notin S_0$

for a finite set $S_0 \subset S$ and some $\varepsilon > 0$. Then P is positive recurrent.

The role of $h(\cdot)$ is similar to the one of Lyopunov function used for the stability of ODEs. Here

$$\sum_k p_{ik} \cdot h(k) - \mathbb{E}\Big[h(x_{n+1} | X_n = 1)\Big]$$

The second condition says that $h(\cdot)$ decreases in expectation outside S_0: S_0 is **attractive**.

Example 1.17

$S + \mathbb{Z}_+, \qquad h(i) = i$. The condition requires

$$\mathbb{E}\Big[X_{n+1} - X_n | X_n = 1\Big] < 0, \qquad i > i_0$$

so the chain is atracted to the set $\{i : i \leqslant i_0\}$, giving stochastic stability.

1.4.5 Stationarity

Note:

$$\begin{aligned}\mathbb{P}(X_n = j) &= \overbrace{\sum_{i\in S}\mathbb{P}(X_0 = i)\mathbb{P}(X_n = j|P_0 = i)}^{\text{disintegrating the joint}}\\ &= \sum_{i\in S}\lambda_i p_{ij}^{(n)} = (\lambda P^n)_j\end{aligned}$$

$(\lambda P^n)_j$ can be seen as the marginal distribution of X at time n: in other words

$$X_0 \sim \lambda \implies X_n \sim \lambda P^n$$

Definition 1.10

A non negative (row) vector $\pi = (\pi_i, i \in S)$ is said to be an **invariant measure** for P if $\pi P = \pi$, called **global balance equation**. Namely:

$$\sum_{i\in S}\pi_i p_{oj} = \pi_j$$

i.e. marginalizing out the initial state with respect to π, the marginal measure after one step is preserved:

$$X_0 \sim \pi \implies X_i \sim \pi P = \pi$$

Furthermore,

$$\pi P^2 = \underbrace{\pi P}_{\pi} = \pi P = \pi$$

Iterativity yields that $\pi P^n = \pi$, therefore

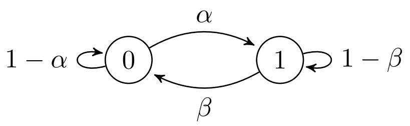
$$X_0 \sim \pi \implies X_n \sim \pi \qquad \forall n \geqslant 0$$

We have extended the one-step invariance to the n-step invariance. If we can normalize π:

$$\tilde{\pi}_i = \frac{\pi_i}{\sum_{j\in S}\pi_j}\tilde{\pi}$$

and we call $\tilde{\pi}$ **stationary distribution**.

Example 1.18



with $\pi = (\pi_0, \pi_1)$, $\pi_i > 0$. We need to try and solve the global balance equation:

$$\pi P = \begin{bmatrix}\pi_0 & \pi_1\end{bmatrix}\begin{bmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{bmatrix} = \begin{bmatrix}\pi_0(1-\alpha)+\pi_1\beta & \pi_0\alpha + \pi_1(1-\beta)\end{bmatrix}.$$

Impose $\pi P = \pi = (\pi_0, \pi_1)$:

$$\pi_0(1-\alpha) + \pi_1\beta = \pi_0 \implies \pi_0 = \pi_1\frac{\beta}{\alpha}.$$

Substituting,

$$\pi_1\frac{\beta}{\cancel{\alpha}}\cancel{\alpha} + \pi_1\beta = \pi_1 \implies \pi_1 = \pi_1 \quad \checkmark$$

with normalization we get

$$\begin{aligned}\pi_0 + \pi_1 = 1 &\longrightarrow \pi_1 \frac{\beta}{\alpha} + \pi_1 = 1 \\ &\longrightarrow \pi_1 = \frac{\beta}{\alpha+\beta}, \quad \pi_0 = \frac{\alpha}{\alpha+\beta}.\end{aligned}$$

Example 1.19

If P is stationary with respect to π then its "lazy" version

$$P' = \varepsilon P + (1-\varepsilon) I, \quad \varepsilon \in [0,1]$$

is still stationary.

Example 1.20

Consider the random walk on $\mathbb{Z}$. The global balance equation is:

$$\begin{aligned}\sum_{i \in S} \pi_i p_{ij} &= \pi_{j-1} p_{j-1,j} + \pi_{j+1} p_{j+1,j} \\ &= \pi_{j-1} p + \pi_{j+1}(1-p) \overset{?}{=} \pi_j\end{aligned}$$

Set $\pi_i = c \geqslant 0$:

$$cp + c(1-p) = c$$

So:

- the uniform measure on $\mathbb{Z}$ is invariant;
- the invariant measures are not necessarily unique;
- $p \neq \frac{1}{2}$ is the transient case, $p = \frac{1}{2}$ is the recurrent case.

So there exists an invariant measure that doesn't imply recurrence.

$$\sum_{i \in \mathbb{Z}} \pi_i = \begin{cases} \infty & c > 0 \\ 0 & c = 0 \end{cases}$$

We cannot normalize in this case (π is only σ-finite, not finite) so we have invariant measures but we don't have any stationary distribution: the chain is non stationary.

Exercise 1.1

Find the stationary distribution for $RW(p, 1-p)$ and $p_{00} = 1-p$ under correct restrictions on p.

Exercise 1.2

Do the same for $B\&D(p, 1-p)$ and $p_{00} = 1-p$ under the correct restrictions on p.

Exercise 1.3

Show that in $B\&D(p, 1-p)$:

$$p_i = \frac{b}{b+i}, \quad b > 0$$
$$\implies \pi_i = \frac{b+i}{2b} \sim Poiss(i, b)$$

(called size-biased Poisson).

We are interested in conditions that provide stationarity:

Proposition 1.11

If for some $i \in S$

$$p_{ij}^{(n)} \xrightarrow[n \to \infty]{} \pi_j \implies \pi \text{ is invariant.}$$

Proof

$$S = \{0, \dots, N\}$$
$$\forall n \geqslant 1 : \qquad \sum_{j=0}^{N} p_{ij}^{(n)} = 1.$$

This means that we cannot have $p_{ji}^{(n)} \to 0 \ \forall j \in S$. So there exist a state $i \in S$ that is both positive and irreducible and therefore X must be positive □

Example 1.21

Consider the random walk on $\mathbb{Z}$:

$$p_{ij}^{(n)} \to 0 \implies \pi = (\pi_i, i \in \mathbb{Z}) \quad \pi_i = 0$$

Which is invariant, as found above, with $c = 0$.

Example 1.22

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \qquad \alpha, \beta \neq 1$$

$$p_{00}^{(n+1)} \underset{\substack{\uparrow \\ \text{C.K.}}}{=} \sum_{k \in S} p_{0k}^{(n)} p_{k0} = p_{00}^{(n)} \underbrace{(1-\alpha}_{p_{00}} + \underbrace{pp01^{(n)}}_{1-p_{00}^{(n)}} \underbrace{\beta}_{p_{10}}$$

$$p_{00}^{(n+1)} = \dots = \beta + p_{00}^{(n)}(1-\alpha-\beta)$$

The recurrence equation brings to solution:

$$p_{00}^{(n)} = \frac{\beta}{\alpha+\beta} + (1-\alpha-\beta)^n \frac{\alpha}{\alpha+\beta} \to \frac{\beta}{\alpha+\beta}$$

as found earlier.

In general we do not want to assume that P^n converges.

Theorem 1.2

An irreducible Markov Chain has invariant distribution π if and only if it is positive recurrent, in which case π is unique and

$$\pi_i = \frac{1}{m_i}$$

where

$$m_i = \mathbb{E}\Big[T_i | X_0 = i\Big]$$

is the **expected return time**.

Example 1.23

Considering $P = I$, we found out that the invariant distribution is not unique. This is a contradiction, due to the fact that irreducibility is violated.

Example 1.24

Consider a random walk on $\mathbb{Z}$ for $p \in (0, 1)$ which is transient and null recurrent: This implies that $m_i = \infty$. Indeed $\pi_i = \frac{1}{m_i} = 0$ is invariant for the random walk.

Example 1.25

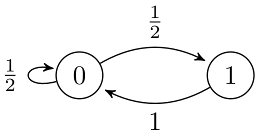
Consider a 2-state chain with

$$(\pi_0, \pi_1) = (\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta}).$$

If we interpret the statement literally, then

$$(m_0, m_1) = (\frac{\alpha + \beta}{\beta}, \frac{\alpha + \beta}{\alpha})$$

Let's choose, for example, $\alpha = 0.5$ and $\beta = 1$:



with $m_0 = \frac{3}{2}$ and $m_1 = 3$ We can double check:

- Return to 0: $\begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{2} \end{cases} \implies 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2} = m_0$

- Return to 1: $\begin{cases} 2 & \text{with probability } \frac{1}{2} \\ 3 & \text{with probability } \frac{1}{2^2} \\ \vdots & \\ (1+k) & \text{with probability } \frac{1}{2^k} \\ \vdots & \end{cases}$ so

$$m_1 = \sum_{k\geq 1}(1+k)\frac{1}{2^k} = \ldots = \sum_{k\geq 1}\frac{1}{2^k} + \sum_{k\geq 1}\frac{k}{2^k} = 1+2 = 3.$$

Definition 1.11

An irreducible Markov chain is said to be **reversible with respect to** π if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \forall i,j \in S$$

which is called **detailed balance equation**.

Proposition 1.12

If P and π are in detailed balance (that is, they satisfy the detailed balance equation, then π is invariant for P.

Proof

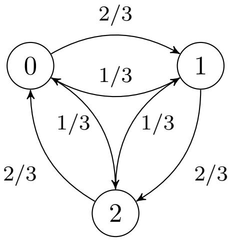
Integrate both sides of the equation with respect to i:

$$\begin{aligned} \sum_i \pi_i p_{ij} &= \sum_i \pi_j p_{ij} \\ &= \pi_j \underbrace{\sum_i p_{ij}}_{=1} = \pi_j. \end{aligned}$$

□

Detailed balance is a sort of local criterion on the chain propensity to go from $i \to j$ and $j \to i$, while the global balance is a global criterion on the chain propensity to go from $i \to j$ and $j \to i$.

Example 1.26



Check first that the uniform distribution is invariant:

$$\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}).$$

Nevertheless,

$$\pi_0 p_{01} = \frac{1}{3}\frac{2}{3} \neq \frac{1}{3}\frac{1}{3} = \pi_1 p_{10}$$

Example 1.27

Consider a Birth and Death process of parameters $(p, 1-p)$ with

$$p_{00} = 1 - p$$

The detailed balance equation is

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \forall i, j$$

Take $j = i + 1$: we get

$$\begin{aligned} \pi_i p_{i,i+1} &= \pi_{i+1} p_{i+1,i} \qquad q := 1 - p \\ \pi_i p &= \pi_{i+1} q \\ i = 0 &\implies \pi_1 = \pi_0 \frac{p}{q} \\ i = 1 &\implies \pi_2 = \pi_1 \frac{p}{q} = \pi_0 \left(\frac{p}{q}\right)^2 \\ &\implies \pi_k = \pi_0 \left(\frac{p}{q}\right)^k \end{aligned}$$

So, π_k is proportional to ρ^k, where $\rho = \frac{p}{q}$.
We now integrate π_k:

$$\sum_{k \geq 0} \pi_k < \infty \iff p < \frac{1}{2} \qquad (\rho < 1) + \dots$$

which implies that

$$\tilde{\pi} = \frac{\pi_k}{\sum_j \pi_j} \sim \mathrm{Geom}(\rho).$$

So, using detailed balance equations, we get a result faster than what we would obtain through global balance equations.

Let

$$< x, y > := \sum_{i \in S} x_i y_i \pi_i.$$

Definition 1.12

We define

$$\ell_2(\pi) := \{x \in \mathbb{R}^S : < x, x > < \infty\}.$$

A matrix P is **self-adjoint with respect to π** if

$$< Px, y > = < x, Py >, \qquad x, y \in \ell_2(\pi).$$

Proposition 1.13

Markov chain with transition matrix P is reversible with respect to π if and only if P is self-adjoint with respect to π.

It can be proved that if P is reversible with respect to π, then the **time reversal**

$$Y_n := X_{N-n}, \qquad 0 \leqslant n \leqslant N$$

has the same distribution as X if both are started in equilibrium.

1.4.6 Convergence

We can interpret time reversal as the possibility of reversing the sense of time. If the chain X starts from an arbitrary initial distribution, not necessarily at equilibrium, is it going to converge to the stationary distribution?

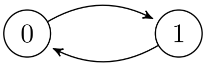
$$\lambda P^n \xrightarrow[?]{n\to\infty} \pi$$

The notion of convergence convergence has to be made more precise.

Example 1.28

Consider the matrix P:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$



The chain is

- irreducible
- positive-recurrent (the mean return time, which is the only return time, equals 2)
- The stationary distribution
$$\pi = (\frac{1}{2}, \frac{1}{2})$$
is invariant ✓

But,

- $P^{2n+1} = P$, $P^{2n} = I$, so $p_{ij}^{(n)} \not\to \pi_j$. Previously, we assumed convergence and said that a limiting distribution is invariant
- $\lambda P^{2n} = \lambda \qquad \lambda P^{2n+1} = 1 - \lambda$

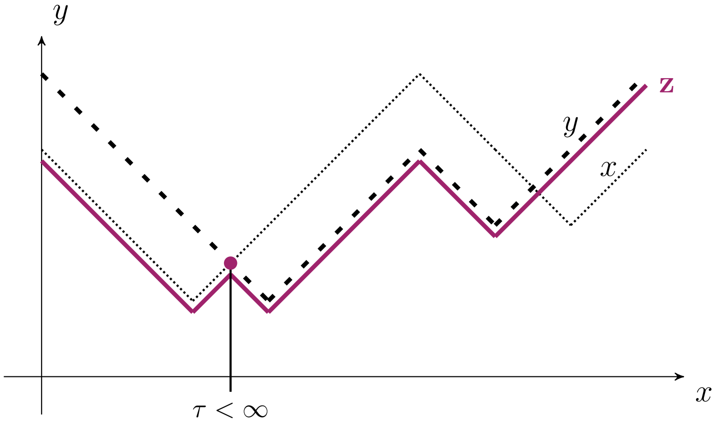
Why? The periodicity is preventing the chain to converge. Periodicity prevents convergence by making the dependence on the initial state too strong: if the chain starts from 0 the all even steps will bring to 0 and all odd steps will bring to 1.

Definition 1.13

An irreducible, aperiodic, positive recurrent chain is called **ergodic**.

Under these assumptions, we want to show that an ergodic chain converges to equilibrium from any initial distribution. To do so, we are going to use *coupling*. Consider

- π invariant for P
- $X \sim$ Markov (λ, P). λ is arbitrary and the so distributed X is the chain of interest.
- $Y \sim$ Markov (π, P) is an auxiliary chain
$$\implies Y_n \sim \pi, \forall n \geq 0$$
- Assume X and Y meet in finite time.



We obtain a new chain Z_n so defined:

$$Z_n = \begin{cases} X_n & n < \tau \\ Y_n & n \geq \tau \end{cases}$$

Now, if we show that $Z \sim$ Markov (λ, P) (reaching quote 2 assumptions, together with meeting in finite time), then

$$X \overset{d}{=} Z \qquad \text{and} \qquad Z_n \sim \pi \quad \forall n \geq \tau$$

which would imply that

$$X_0 \sim \lambda, X_n \xrightarrow{n\to\infty} \pi$$

Definition 1.14

Two Markov chains Z and Y on S are said to **couple** if there exists an almost surely finite stopping time τ, called **coupling time**, such that

$$Z_n = Y_n \qquad \forall n \geq \tau$$

This relates to the **total variation distance** between the marginals:

$$V \sim \lambda,\ W \sim \mu \text{ on } S$$

The total variation distance is

$$\begin{aligned} d_{TV}(\lambda, \mu) &= \sup_{A \subset S} |\mathbb{P}(V \in A) - \mathbb{P}(W \in A)| \\ &= \frac{1}{2} \sum_{i \in S} |\lambda_i - \mu_i| \end{aligned}$$

Proposition 1.14

Let $Z \sim \text{Markov}(\lambda, P)$ and let $Y \sim \text{Markov}(\mu, P)$;
assume that a coupling time exists. Then,

$$d_{TV}(\lambda P^n, \mu P^n) \xrightarrow{n\to\infty} 0$$

Proof

Consider $A \subset S$ and

$$\begin{aligned}\mathbb{P}(Z_n \in A) - \mathbb{P}(Y_n \in A) &= \mathbb{P}(Z_n \in A, n < \tau) + \mathbb{P}(Z_n \in A, n \geq \tau) + \\ &\quad - \mathbb{P}(Y_n \in A, n < \tau) - \mathbb{P}(Y_n \in A, n \geq \tau) \\ &\overset{Z_n = Y_n, \forall n \geq \tau}{=} \mathbb{P}(Z_n \in A, n < \tau) - \underbrace{\mathbb{P}(Y_n \in A, n < \tau)}_{\geq 0} \\ &\leq \mathbb{P}(\underbrace{Z_n \in A, n < \tau}_{\subset \{n < \tau\}}) \\ &\leq \mathbb{P}(n < \tau)\end{aligned}$$

Hence, by the arbitrarity of A,

$$\sup_{A \subset S} |\mathbb{P}(Z_n \in A) - \mathbb{P}(Y_n \in A)| \leq \mathbb{P}(\tau > n)$$

but $\tau < \infty$ almost surely, so

$$\mathbb{P}(\tau > n) \xrightarrow{n \to \infty} 0$$

□

So, we are sure that in a finite time they meet and hence Z is going to have the stationary distribution π. Also,

$$\exists \tau < \infty \implies d_{TV} \to 0$$

It is enough to show that such τ exists for ergodic chains.

Theorem 1.3

Let

- P be ergodic
- $X \sim \text{Markov}(\lambda, P)$ and $Y \sim \text{Markov}(\mu, P)$ be independent

Then, the stopping time

$$\tau = \inf\{n \geq 0 : X_n = Y_n\}$$

is almost surely finite, and the chain

$$Z_n = \begin{cases} X_n & n < \tau \\ Y_n & n \geq \tau \end{cases}$$

is $\text{Markov}(\lambda, P)$.

So, this yields a coupling time between $Z \sim \text{Markov}(\lambda, P)$ and $Y \sim \text{Markov}\ (\mu, P)$, which implies that

$$Z_n = Y_n \sim \mu P^n, \forall n \geq \tau$$

So, by the Proposition 1.4.6,

$$d_{TV}(\lambda P^n, \mu P^n) \xrightarrow{n \to \infty} 0$$

If we now let $\mu = \pi$, then $\mu P^n = \pi$ and hence

$$d_{TV}(\lambda P^n, \pi) \to 0$$

Remark

Imposing $\mu = \pi$ is not cheating since μ is auxiliary and we can equal it to what we need. The result is that all the marginals converge to the stationary.

Since $X \overset{d}{=} Z$, then

$$\mathbb{P}(X_n = j) \xrightarrow{n\to\infty} \pi_j, \qquad \text{for every initial distribution } \lambda$$

If $\lambda = \delta_i$:

$$\begin{aligned} d_{TV}(\lambda P^n, \pi) &= \frac{1}{2}\sum_{j\in S} |(\lambda P^n)_j - \pi_j| \\ &= \frac{1}{2}\sum_{j\in S} |\sum_{h} \lambda_h p_{hj}^{(n)} - \pi_j| \\ &= \frac{1}{2}\sum_{j\in S} |1 \cdot p_{ij}^{(n)} - \pi_j| \xrightarrow{n\to\infty} 0 \end{aligned}$$

Remark

In the first line, we take the supremum over a discrete space, so in the worst case we integrate the differences.

Hence, all the transition probabilities

$$p_{ij}^{(n)} \xrightarrow{n\to\infty} \pi_j, \forall i \in S$$

converge to π_j for every starting point (state) i.
In summary:

Theorem 1.4

Let P be ergodic with invariant distribution π, and let $X \sim \text{Markov}(\lambda, P)$. Then,

$$d_{TV}(\lambda P^n, \pi) \xrightarrow{n\to\infty} 0$$

and

$$p_{ij}^{(n)} \xrightarrow{n\to\infty} \pi_j, \qquad \forall i, j \in S.$$

Remark

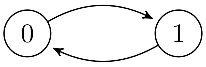
Ergodicity ensures that the two chains meet on the diagonal. λ is orthogonal to μ and there exists $B \in S$ such that

$$\sum_{i\in B} \lambda_i = 1$$
$$\sum_{i\in B^C} \mu_i = 1$$

Example 1.29

Consider the matrix P:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$



We know that

$$\pi = (\frac{1}{2}, \frac{1}{2}) \qquad \text{and} \qquad P^{(2n)} = I, P^{(2n+1)} = P$$

Consider also X, Y as characterized in Theorem 1.4.6, with $\lambda = \delta_0, \mu = \pi$. This means that X starts at 0 with probability 1:

$$X_0 = 0 \text{ a.s.}$$
$$Y_0 = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}$$

So, either they meet at $n = 0$ or they never meet.

It's useful to remind that are talking about almost sure meeting, not with probability 1. Which factor breaks the conclusion of meeting almost surely? The chain is:

- positive recurrent;
- irreducible;
- not aperiodic.

Theorem 1.5

Ergodic theorem. Let $X \sim Markov(\lambda, P)$ be irreducible and let $m_i = \mathbb{E}\Big[T_i | X_0 = i\Big]$. Then, almost surely,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(X_n = j) \xrightarrow[N \to \infty]{} \Pi_j = \frac{1}{m_j}.$$

Moreover, if P is positive recurrent with unique stationary distribution π and $f : S \to \mathbb{R}$ with respect to π, then

$$\frac{1}{N} \sum_{n=1}^{N} f(X_n) \xrightarrow[N \to \infty]{} \sum_{j \in S} f(j) \Pi_j.$$

The quantity $\frac{1}{N} \sum_{n=1}^{N} f(X_n)$ is the **ergodic average** and it is taken along the sample path of the chain. The right-hand side shows an object whose form recalls an expectation.
So, this is a way of relaxing the assumption of an i.i.d. sample. We are substituting it with the correlation associated to the Markovian structure, so that the convergence holds.

$$\sum_{j \in S} f(j) \Pi_j = \mathbb{E}\Big[f(X_n)\Big] \qquad \text{at equilibrium.}$$

So we have an equivalent form of the Strong Law of Large numbers for Markov Chains.

Remark

In the one-dimensional case, we have convergence of order $\frac{1}{\sqrt{10}}$ to 0, which is not enough to make $m_j < \infty$. Hence $m_j = \infty \implies \frac{1}{m_i} \to 0$, which is consistent with what we know.

The first claim holds for all irreducible chains, for example the Random Walk whereby:

$$p_{00}^{(n)} \to 0 \qquad \text{as } \frac{1}{\sqrt{n}} \text{ and } m_i = \infty$$

implies

$$\frac{1}{N}\sum_n \mathbb{1}(X_n = j) \to 0.$$

What about the speed of convergence?

Example 1.30

Consider

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

With

$$p_{00}^{(n)} = \frac{\beta}{\alpha+\beta} + (1-\alpha-\beta)^n \frac{\alpha}{\alpha+\beta}.$$

Set $\gamma_1 = 1$ and $\gamma_2 = 1-\alpha-\beta$. We can prove that

$$P^n = \frac{1}{\alpha+\beta}\begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} = \frac{1-\alpha-\beta}{\alpha+\beta}\begin{bmatrix} & \end{bmatrix}$$

$$\pi = \mathbf{1}^T = \begin{bmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{bmatrix}$$

$$P^n - \pi = \gamma_2^n B$$

Theorem 1.6

Let P be a $k \times k$ irreducible and aperiodic transition matrix. Denote the distinct eigenvalues as

$$1 = \gamma_1 > |\gamma_2| > |\gamma_3| > \ldots > |\gamma_k|.$$

Then

$$P^n = \pi + O(n^{m-1}|\gamma_2|^n)$$

where m is the algebraic multiplicity of γ_2.

Definition 1.15

A statement like

$$d_{TV}(\lambda P^n, \pi) \leqslant C(\lambda)\rho^n$$

is called **geometric ergodicity**, with $C(\lambda) \in \mathbb{R}$ depending on λ.

1.4.7 Quasi-stationary distributions

Let's now take into account **quasi-stationary distributions**.
Let $a \in S$ be an absorbing state with $\mathbb{P}(T_a < \infty) = 1$ visited almost surely in finite time. if X is irreducible then there is no stationarity.
Example: the Galton-Watson branching process with mean offspring $m = \mathbb{E}\Big[g\Big] < 1$. It could be of interest to study the behaviour before absorption.
Let λ be supported by $S_a = S/\{a\}$ and denote

$$\mathbb{P}_\lambda(\cdot) := \mathbb{P}(\cdot | X_0 \sim \lambda).$$

Then, given $A \subset S_a$, we are interested in

$$\mathbb{P}_\lambda(X_n \in A | T_a > n) = \frac{\mathbb{P}_\lambda(X_n \in A, T_a > n)}{\mathbb{P}_\lambda(T_a > n)} = \frac{\lambda P^n|_a}{\lambda P^n|_{S_a}}.$$

Definition 1.16

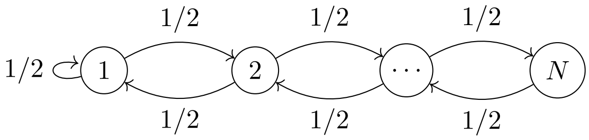
We say that π is a **quasi-stationary distribution** if

$$P_\pi(X_n = i | T_a > n) = \Pi_i \qquad \forall n \geqslant 0.$$

This process preserves the marginal, conditional on not getting observed.

Example 1.31

With $S = \{0, \dots, N\}$, let Y be a symmetric random walk on $S_0 = \{1, \dots, N\}$:



whose invariant π is the uniform on on S_0. Let τ be a finite random time on $\mathbb{N}$, independent of Y, such that:

$$X_n = \begin{cases} Y_n & n < \tau \\ 0 & n > \tau \end{cases}.$$

So, in τ, X jumps to 0 and there remains. The transition probabilities are:

$$i \in S_0 :\ p_{ij}^{(n)} = \underbrace{\begin{cases} \mathbb{P}(\tau = n) & j = 0 \\ \frac{1}{2}(1 - \mathbb{P}(\tau - n)) & j = i \pm 1 \end{cases}}_{\text{there are called } \textit{modulo boundaries}}$$

Since τ is independent, we have

$$\begin{aligned} \mathbb{P}_\pi(X_n = i | \tau > n) &= \mathbb{P}_\pi(X_n = i | X_n \neq 0) \\ &= \mathbb{P}_\pi(Y_n = i) = \Pi_i \end{aligned}$$

1.5 Hidden Markov Chains

Hidden Markov Chains are a widely applied statistical framework, useful whenever it is necessary to estimate the current status of a system based on noisy or incomplete observations.

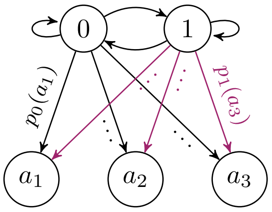
Examples:

- X_n: position of a moving object;
  Y_n: noisy observations of the position (by means, for instance, of a radar or a sonar);
- X_n: state of a productive system (that can be good or bad);
  Y_n: conditions of the output (that can be perfect or faulty);
- X_n: latent level of volatility in a financial market (high or low);
  Y_n: amount of financial products exchanged;
- X_n: number of alleles of type 0 in a population;
  Y_n: number of alleles of type 0 in a sample;

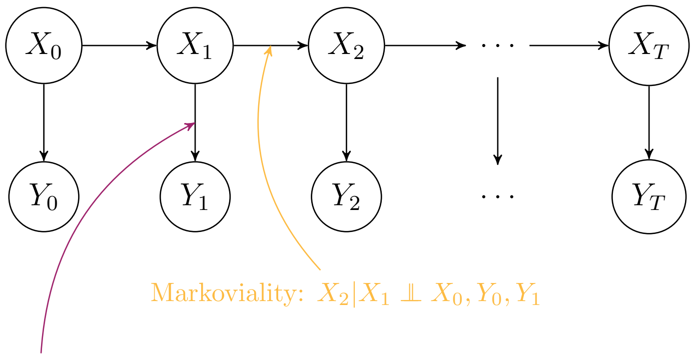
Let $X \sim Markov(\lambda, P)$ be unobserved (or hidden). This chain is sometimes called *signal*. Every time X enters a sate, an observation Y (that we assume discrete on $\{a_1, a_2, \dots\}$) is emitted with probability that depends on the current state of X only. The model is:

$$\diamond\, \mathbb{P}(X_n = j|X_{n-1}+i) = p_{ij}$$
$$\diamond\, \mathbb{P}(\underbrace{Y_n = y_n}_{\text{generic realization of } Y}|X_n = j) = p_j(y_n)$$

$p_j(y_n)$ is the **emission distribution**, that is the likelihood of observing Y_n if X is in j. If $S+\{0,1\}$:



An alternative depiction is:



Markoviality: $X_2|X_1 \perp\!\!\!\perp X_0, Y_0, Y_1$

These arrows represent the conditional independence of $Y_i|X_i$, which is independent from everything else

The goal is to establish

$$\begin{aligned} &\mathbb{P}(X_{0:T}|Y_{0:T}) \\ \text{with}\quad &X_{0:T} = (X_0, \dots, X_T) \\ &Y_{0:T} = (Y_0, \dots, Y_T) \end{aligned}$$

as the probability of a certain sequence of states of the system, given a sequence of observations. We can use the chain rule to write:

$$\mathbb{P}(X_T|Y_{0:T})\mathbb{P}(X_{T-1}|X_T, Y_{0:T}) \cdot \ldots \cdot \mathbb{P}(X_0|X_{1:T}, Y_{0:T})$$

as the backwards decomposition. The generic factor is:

$$\mathbb{P}(X_n|X_{n+1:T}, Y_{0:T}) = \mathbb{P}(\underbrace{X_n|Y_{0:T}}_{Z}, \underbrace{X_{n+1}}_{W})$$

by virtue of the fact that $X_n|X_{n+1:T} \perp X_{n+2:T}, Y_{n+1:T}$. By Bayes' theorem:

$$\begin{aligned} \mathbb{P}(Z|W) &\propto \mathbb{P}(Z)\mathbb{P}(W|Z) \\ &\propto \mathbb{P}(X_n|Y_{0:n})\mathbb{P}(X_{n+1}|X_n, Y_{0:n}) \\ &= \mathbb{P}(X_n|Y_{0:n})\mathbb{P}(X_{n+1}|X_n) \end{aligned}$$

Since $X_{n+1}|X_n$ is independent from everything. $\mathbb{P}(X_n|Y_{0:n})$ is the **filtering distribution**, the conditional law of X given the information of past and present data, while $\mathbb{P}(X_{n+1}|X_n)$ is the transition probability of X.
Define

$$F_n(j) = \mathbb{P}(X_n = j, Y_{0:n})$$

so that

$$\mathbb{P}(X_n|Y_{0:n}) = \frac{F_n(j)}{\sum_i F_n(i)}.$$

Then

$$\begin{aligned}
F_n(j) &= \mathbb{P}(Y_{0:n-1}, X_n = j, Y_n = y_n)\\
&= \sum_i \mathbb{P}(\underbrace{Y_{0:n-1}, X_{n-1} = i}_{F_{n-1}(i)}, X_n = j, Y_n = y_n)\\
&= \sum_i F_{n-1}(i)\,\underbrace{\mathbb{P}(X_n = j, Y_n = y_n|X_{n-1} = i, Y_{0:n-1})}_{\mathbb{P}(X_n|X_{n-1})\mathbb{P}(Y_n|X_n)}\\
&= \sum_i F_{n-1}(i) p_{ij} p_j(y_n)\\
&\implies F_n(j) + \Big(\sum_i F_{n-1}(i) p_{ij}\Big) p_j(y_n)
\end{aligned}$$

So that these can be computed recursively.

$$\begin{aligned}
F_0(j) &= \mathbb{P}(X_0 = j, Y_0) = \lambda_j p_j(y_0)\\
F_1(j) &= \sum_i F_0(i) p_{ij} p_j(y_1)\\
&= p_j(y_1) \sum_i \lambda_i p_i(y_0) p_{ij}
\end{aligned}$$

If S is finite, we can compute these quantities. Normalizing, we get the filtering distribution

$$\mathbb{P}(X_n = j|Y_{0:n}) = \frac{F_n(j)}{\sum_i F_n(i)}.$$

Moreover, if we consider the **marginal smoothing distribution**

$$\begin{aligned}
\mathbb{P}(X_n|y_{0:n}) &= \mathbb{P}(\underbrace{X_n|Y_{0:n}}_{Z}, \underbrace{Y_{n+1:T}}_{W})\\
&\propto \mathbb{P}(Z)\mathbb{P}(W|Z) \qquad \text{since } Y_{n+1:T}|X_n \perp\!\!\!\perp Y_{0:n}\\
&= \underbrace{\mathbb{P}(X_n|Y_{0:n})}_{\text{filtering distribution}} \overbrace{\mathbb{P}(Y_{n+1:T}|X_n)}^{B_n(\cdot)}.
\end{aligned}$$

$B_n(\cdot)$ is the **cost-to-go** function: it measures the likelihood of the future observations given a state of X:

$$B_n(j) + \mathbb{P}(Y_{n+1:T}|X_n = j)$$

with

$$\mathbb{P}(X_n|Y_{0:T}) \propto F_n(j) B_n(j).$$

Also, B_n can be computed recursively as

$$B_n(j) = \sum_i p_{ij} p_j(y_{n+1}) B_{n+1}(j)$$

starting from T and proceeding backwards. Finally, with these we have

$$\mathbb{P}(Y_{0:T}) = \sum_j F_n(j) B_n(j)$$

which represents the likelihood of the observations, which in principle can be maximized.

1.6 General state space

Assume for simplicity that $S \subset \mathbb{R}$ or $\mathbb{R}^k$ (uncountable). Define a time-homogeneous transition probability from

$$\mathbb{P}(x, A) := \mathbb{P}(X_{n+1} \in A|X_n = x), \qquad A \in \mathscr{B}(S).$$

If this has a density with respect to some dominant measure ν, we call

$$p = \frac{dP}{d\nu}$$

the **transition density**.
For example, if ν is the Lebesgue measure,

$$\mathbb{P}(x, A) = \int_A \mathbb{P}(x.y)dy$$

.

Definition 1.17

A Markov Chain o S uncountable is said to be **φ-irreducible** if there exists a σ-finite measure φ on S such that $\forall A \in \mathscr{B}(S)$ with $\varphi(A) > 0$ and for all $x \in S \; \exists \geqslant 1, \; n = n(x, A)$ such that $p^{(n)}(x, A) > 0$.

Definition 1.18

A φ-irreducible Markov Chain is said to be **Harris Recurrent** if $\forall A \in \mathscr{B}(S)$ such that $\varphi(A) > 0$ then $\mathbb{P}(X_n \in A \text{ i.o.}) = 1$. It is called *positive* if it admits an invariant distribution π.

π is invariant if $\int_A \pi(dx)p(x, dy) = \pi(A)$.

Theorem 1.7

Let X be aperiodic and positive Harris recurrent. Then

$$d_{TV}(\lambda P^n, \pi) \xrightarrow[n\to\infty]{} 0$$

where

- $\lambda P^n(A) = \int\limits_S P^n(x, A)\lambda(dx)$
- $d_{TV}(\lambda, \mu) = \sup\limits_{A\in\mathscr{B}} |\mathbb{P}(V \in A) - \mathbb{P}(W \in A)|$ if $V \sim \lambda$ and $W \sim \mu$.

If, in addition, $f : S \to \mathbb{R}$ is π-integrable

$$\frac{1}{N}\sum_{i=1}^{N} f(X_n) \xrightarrow[N\to\infty]{} \int_S f(x)\pi(dx)$$

with probability 1.

Definition 1.19

Let X be aperiodic and positive Harris recurrent. Then

$$d_{TV}(\lambda P^n, \pi) \xrightarrow[n\to\infty]{} 0$$

where

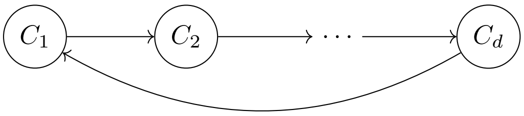
- $\lambda P^n(A) = \int\limits_S P^n(x, A)\lambda(dx)$

- $d_{TV}(\lambda, \mu) = \sup_{A\in\mathcal{B}} |\mathbb{P}(V \in A) - \mathbb{P}(W \in A)|$ if $V \sim \lambda$ and $W \sim \mu$.

If, in addition, $f : S \to \mathbb{R}$ is π-integrable

$$\frac{1}{N}\sum_{i=1}^{N} f(X_n) \xrightarrow[N\to\infty]{} \int_S f(x)\pi(dx)$$

with probability 1.



2 Monte-Carlo integration

We are interested in evaluating

$$\mu_f := \mathbb{E}_\pi[f(X)] = \int_S f(x)\pi(dx)$$

as an integration problem. If S is discrete this takes a more familiar form $\sum_{i\in S} f(i)\pi_i$
The problems can arise from the fact thatπ can be multidimensional, and possibly unmormalized.
Some example of these situations may be:

- $f(x) = 1 \implies \mu_f$ is normalizing constant.
- If f is the identity we are calculating the mean of π.
- If $f(x) = \mathbb{1}_A(x)$ then tail probability and confidence intervals.

We can also be interested in finding

$$argmax_{x\in S}\pi(x)$$

and this is an optimization problem, like for instance the mode of a distribution of interest.
Often an analytical computation is unfeasible and thus we need to use an approximation. Deterministic approximation include:

- Riemann integration: given a partition of S through points $x_1, \dots, x_n$ then

  $$\sum_{i=1}^{n} f(x_i)\pi(x_i)(x_i + x_{i-1}) \to \int f d\pi$$

  as $sup_i|x_i + x_{i-1}| \to 0$.picture
  potentially ineficient especially in high dimention.
- Laplace approximation: use a Gaussian kernel to approximate a function of interest centered at the mode.

Example 2.1

Let f be unimodal:

$$\int f(x)dx = \int e^{h(x)}dx$$

take $h(x) = logf(x)$ and take a taylor of h around the mode of f

$$h(x) \approx h(x_0) + \underbrace{h'(x_0)(x - x0)}_{f'/f|_{x_0}=0} + 1/2h''(x_0)(x - x0)manca$$

> negative because we are at the mode
>
> $$\int f(x)dx \approx e^{h(x_0)} \int e^{1/2|h''(x_0)|(x-x0)}dx =$$
>
> multiply and divide by $c := \sqrt{2\pi|h''(x_0)|^{-1}}$ and we have
>
> $$= f(x_0)c \int \mathcal{N}(x; x_0, |h''(x_0)|^{-1})dx$$
>
> this approximation is only good around the mode, but it loses a great part of its accuracy when we depart from the centre of the distribution: this makes it particularly bad for approximating, for instance, tails of a distribution. Moreover, if we work in high dimension more problems arise. f needs to be unimodal, else this approach can fail (e.g. mixture models that came up a lot in statistics)

Deterministic approximations typically do not exploit information about the shape of the distribution of interest and this naturally leads to **stochastic approximation**.

2.1 The Monte Carlo principle

the main idea consists in exploiting information about π and concentrate resources where they are most useful. From a general point of view simulate from π and compute an approximation of the functional of interest.

Algorithm 1

General strategy:

- sample $X_1, \dots, X_n \overset{iid}{\sim} \pi$.
  (N = MC sample size)
- $\mu_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$

Note that if $\pi_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$ then we have just

$$\mu_N = \int f(x)\pi_N(dx) = \frac{1}{N}\sum_{i=1}^{N} \int f(x)\delta_{X_i}(dx) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

now

$$\mathbb{E}_\pi\left[\mu_N\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\pi\left[f(x_i)\right] = \int f(x)\pi(dx) = \mu_f$$

so μ_N is unbiased. Moreover, if $\mathbb{E}_\pi(f) < \infty$,

$$\mu_N \xrightarrow{a.s.} \mu_f \qquad \text{by SLLN.}$$

If $\mathbb{E}_\pi(f^2) < \infty$, set

$$\sigma_f^2 := Var[f(X)] = \mathbb{E}_\pi[(f(X) - \mu_f)^2]$$

then μ_N has variance

$$\sigma_N^2 = Var(\mu_N) = \frac{\sigma_f^2}{N}$$

so μ_N is a constant estimator of μ_f. We also have a central limit theorem

$$\sqrt{N}(\mu_N - \mu_f) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2)$$

to be used, e.g., for constructing asymptotic confidence intervals. Some potential issues are:

- it may be computationally expensive;

- sampling directly from π could be unfeasible and/or too difficult (e.g. high-dimensional distribution, distributions whose normalizing constant is unknown...).

example. $Z \in \mathcal{N}(0,1)$. We are interested in

$$\mu = \mathbb{P}(Z > 5) = \int_{\mathbb{R}} \underbrace{\mathbb{1}(x > 5)}_{f(x)} \varphi(z) dz$$

The value for μ is around 2.87×10^{-7}. Now draw $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

$$\mu_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(x_i > 5) \to \mu$$

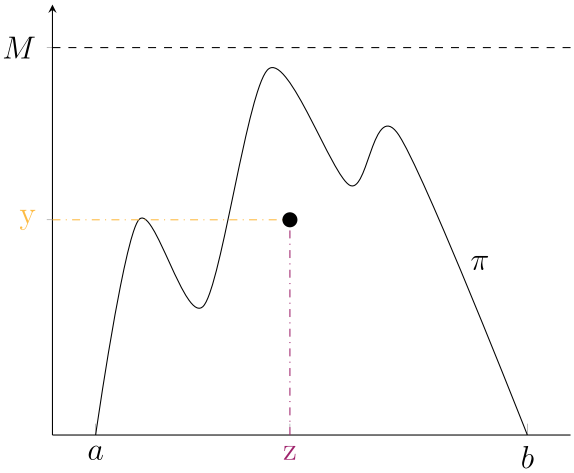
This means that as long as we draw $N < \mathcal{O}(10^7)$ we expect to see a positive indicator... 0 times. This is an example of situation where simulating π is computationally infeasible. To face this limitation, we turn to a very important idea: the **rejection sampling**.

2.2 Rejection Sampling

Algorithm 2

until N points are saved repeat:

- draw independently $Z \sim Unif(a,b)$ and $Y \sim Unif(0,M)$;
- if $y \leq \pi(z)$, set $X = z$.



Only points under π are kept, and these are draws from π. Cdf of accepted points:

$$\begin{aligned}
\mathbb{P}(Z \leq x | Y \leq \pi(Z)) &= \frac{\mathbb{P}(Z \leq x, Y \leq \pi(Z))}{\mathbb{P}(Y \leq \pi(Z))} \\
&= \frac{\mathbb{P}(Z \leq x, Y \leq \pi(Z))}{\mathbb{P}(Z \leq b, Y \leq \pi(Z))} \\
&= \frac{\int_a^x \int_0^{\pi(Z)} \frac{1}{M} dy \frac{1}{b-a} dz}{\int_a^b \int_0^{\pi(Z)} \frac{1}{M} dy \frac{1}{b-a} dz} \\
&= \int_a^x \pi(Z) dz
\end{aligned}$$

so now we can compute $\mu_N = \frac{1}{N} \sum_{i=1}^{N} f(X_i)$.

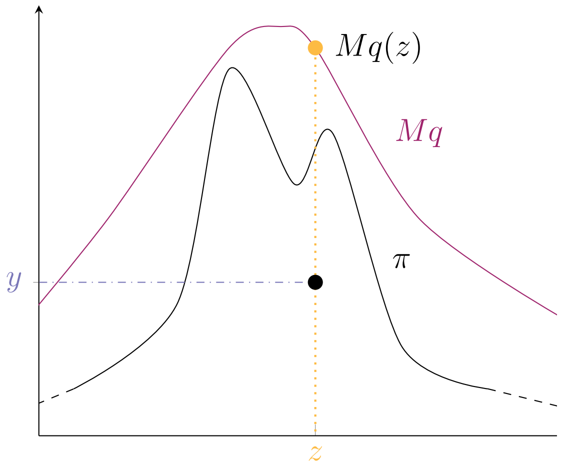
If the support of π is unbounded, we cannot draw from $Unif(a,b)$, the idea is that we are gonna use a non uniform bound and an auxiliary distribution.

Assume:

- $\exists M > 0$ and a density q such that
$$\pi(x) \leq Mq(x)$$
- we can draw from q, meaning that we should choose a q we can simulate relatively easily.

Algorithm 3

```
until $N$ points are saved:
    • draw $Z \sim q$, with $Y|Z = z \sim Unif(0, Mq(z))$;
    • if $Z \leqslant \pi(z)$, set $x = Z$.
```

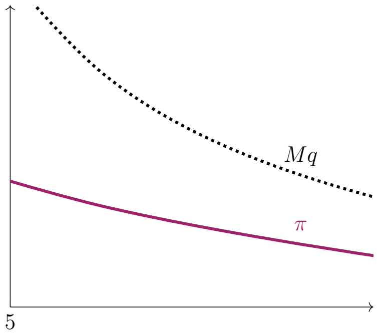


Verify $(-\infty \leqslant a < b \leqslant \infty)$=

$$= \frac{\displaystyle\int_{-\infty}^{x} \int_{0}^{\pi(z)} \frac{1}{\cancel{Mq(z)}} dy \, \cancel{q(z)} dz}{\displaystyle\int_{-\infty}^{\infty} \int_{0}^{\pi(z)} \frac{1}{\cancel{Mq(z)}} dy \, \cancel{q(z)} dz} = \mathbb{P}(X \leqslant x)$$

Example 2.2

take $\mu = \mathbb{P}(z > 5) = 2.87 \cdot 10^{-7}$. We need a non-uniform bound:



Choose $q(x) = e^{-(x-5)}\mathbb{1}(x > 5)$ with $M = \phi(5)$. We know that

$$\mu_N = \% \text{ of accepted points} \times \underbrace{\text{area of graph where we generate values}}_{\int_5^\infty Mq(x)dx = M\int_5^\infty e^{-(x-5)}d = M}$$

This algorithm is generally more efficient than naive Monte-Carlo where we would have to generate $x_i \sim N(0,1)$ variables and keep $x_i > 5$. It can be, anyway, costly: the acceptance probability over an interval (c,d) is

$$\begin{aligned}\mathbb{P}(c \leqslant Z \leqslant d, Y \leqslant \pi(z)) &= \int_c^d \int_0^{\pi(z)} \frac{dy}{Mq(z)} q(z)dz \\ &= \frac{1}{M}\int_c^d \pi(z)dz.\end{aligned}$$

In general, over the support, the acceptance probability is $\frac{1}{N}$. So, in order to have N points, we need to generate a number of points equal to

$$N' \sim Neg - Bin\Big(N, \frac{1}{M}\Big).$$

$$\frac{1}{M}\int_5^\infty \pi(x)dx = \frac{2.87 \cdot 10^{-7}}{1.49 \cdot 10^{-6}} = 0.19$$

2.3 Importance Sampling

This idea starts with a little analytical trick, in general terms the idea is to use all generated samples. Unlike the accept-reject method.

$$\begin{aligned}\mu = \mathbb{E}_\pi[f(x)] = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx \\ = \mathbb{E}_q[f(x)\frac{\pi(x)}{q(x)}]\end{aligned}$$

where q is a density whose support include that of π.
(i.e. $q(x) = 0 \implies \pi(x) = 0$)
Then we can:

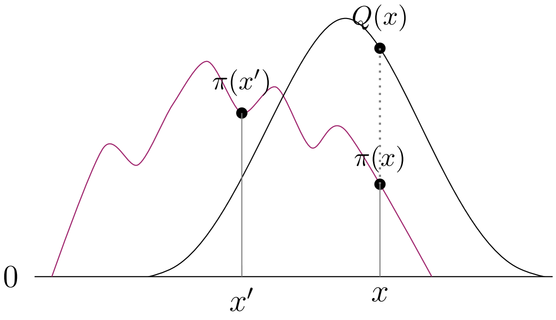
- draw $X_i \overset{\text{i.i.d.}}{\sim} q$
- assign weight $w(x_i) := \frac{\pi(X_i)}{q(X_i)}$ to X_i. (w importance weight)
- set
$$\mu_N = \frac{1}{N}\sum_{i=1}^N f(X_i)w(X_i)$$

Then

$$\mathbb{E}_q[\mu_N] = \mathbb{E}_q[f(x)w(x)] \overset{above}{=} \mu$$

moreover

$$\mu_N \xrightarrow{a.s.} \mu \qquad \text{as } n \to +\infty$$



- down-weight importance of x $\pi(x) < q(x) \implies w(x) = \frac{\pi(x)}{q(x)} < 1$.
- $w(x') > 1$ up-weight of x'.

draws in relevant regions are up-weighed automatically and no draws are wasted. The choice of q is more flexible than in the AR method.
Toy Example. $\mathbb{P}(Z > 5) = \mu = 2.87 \times 10^{-7}$
Use

$$q(x) = e^{-(x-5)}\mathbb{1}_{\{x>5\}},$$

and so

$$\mu = \int_5^\infty \phi(x)dx = \int_{\mathbb{R}} \mathbb{1}_{\{x>5\}}\phi(x)dx = \int_{\mathbb{R}} \underbrace{\mathbb{1}_{\{x>5\}}}_{=f(x)} \underbrace{\frac{\phi(x)}{q(x)}}_{=w(x)} q(x)dx$$

so $X_i \overset{\text{i.i.d.}}{\sim} q$.

$$\mu_N = \frac{1}{N}\sum_{i=1}^N f(X_i)w(X_i) = \frac{1}{N}\sum_{i=1}^N \underbrace{\mathbb{1}_{X_i>5}}_{\text{since } X_i \sim q} \frac{\phi(x_i)}{q(x_i)} = \frac{1}{N}\sum_{i=1}^N \frac{\phi(x_i)}{q(x_i)}$$

Additional upside:
if $\pi = \frac{\tilde{\pi}}{z}$ with z unknown. Then

$$\mu = \int f(x)\pi(x)dx = \int f(x)\frac{\tilde{\pi(x)}}{zq(x)}q(x)dx$$

$$\to \quad z = \int \tilde{\pi(x)}dx = \int \frac{\tilde{\pi(x)}}{q(x)}q(x)dx$$

so setting $\tilde{w}(x) = \frac{\tilde{\pi(x)}}{q(x)}$ we have

$$\mu = \underbrace{\frac{\int f(x)\tilde{w}(x)q(x)dx}{\int \tilde{w}(x)q(x)dx}}_{\text{self-normalizing estimate}}$$

Hence for $X_i \overset{\text{i.i.d.}}{\sim} q$ we now have

$$\tilde{\mu}_N = \frac{\frac{1}{N}\sum_{i=1}^N f(X_i)\tilde{w}(X_i)}{\frac{1}{N}\sum_{j=1}^N \tilde{w}(X_j)} = \sum_{i=1}^N f(X_i)\hat{w}(X_i)$$

with $\hat{w}(X_i) = \frac{\tilde{w}(X_i)}{\sum_{j=1}^N \tilde{w}(X_j)}$. Strong law of large numbers still applies.

2.4 Markov-Chain Monte-Carlo

When:

- Tails of q versus π can be very different so that $w(x)$ possibly unbounded
- q may be difficult to identify in high-dimensions

Idea: we can generate draws from π using Markov Chains, obtaining what is called **Markov Chains Monte Carlo**.
Given a target distribution π, we want to construct an ergodic Markov chain X with stationary distribution π, and use its trajectory to get draws from π.
Assume for now P is ergodic.
Given an intial state X_0, we can in principal simulate trajectory :
$\forall\ n \geq 0$, if $X_n = i$, draw $X_{n+1} = j$ with probability p_{ij} from ergodicity we know

$$\exists n_0 \in \mathbb{N} : X_n \sim \pi \quad \forall n \geq n_0$$

Such n_0 is called Mixing time, meant as an order of magnitude, not as a step. (e.g. $n_0 = O(10^4)$)
Studying mixing times is typically difficult. Usually one uses convergence test or diagnostics. (there is literature, Gelman-Rubin criterion).
Suppose we know n_0. Then

- simulate $\{X_n^{(i)}, n \leq n_0\}$ for $i = 1, \dots, N$.
- set $Y_i = X_{n_0+i}$ (sample path endpoint).

Then $(Y_i, \dots, Y_N)$ is the MCMC sample. These are i.i.d $\sim \pi$ (assume $X_0^{(i)}$ independent).
$\implies$ we are back t0 the Monte-Carlo scenario with

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} f(Y_i) \xrightarrow{a.s.} \mu_f$$

where

$$Var(\mu_N) = \frac{\sigma_f^2}{N}, \quad \sigma_f^2 = Var_\pi(f(Y)).$$

However we are discarding $N = n_0$ sample. If, for example, $N = 10^4$, $n_0 = 10^5 \implies 10^9$, which is expensive.
alternatively, we can think of using a single chain .

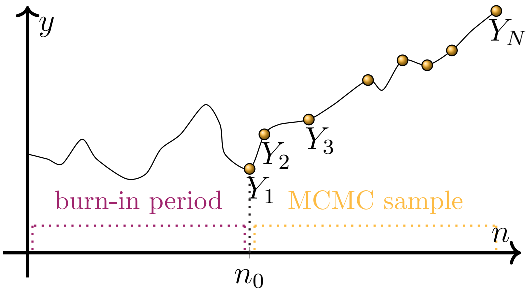
Algorithm 4

Given N:

- simulate $\{X_n, n\}$
- for n_0 large enough, set

$$\hat{Y}_i = X_{n_0+i}, \quad i = 1, \dots, N$$

$$\implies \mu_N = \frac{1}{N}\sum_{i=1}^{N} f(\hat{Y}_i) \approx \mu_f$$



Only n_0 values wasted. The MCMC samples are now correlated and the MCMC estimator is now no longer a generic sample average but it is an ergodic average. In terms of theoretical result we have the ergodic theorem that guarantees

$$\frac{1}{N}\sum_{i=1}^{N} f(\hat{Y}_i) \xrightarrow{a.s.} \mu_f = \mathbb{E}_\pi[f(Y)]$$

for any initial distribution.
Understand what we are losing, consider the variance of $\hat{\mu}_N$

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} f(\hat{y}_i)\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left[Var\Big(f(\hat{y}_i)\Big) + 2\sum_{k=1}^{N-1} Cov\Big(\underbrace{f(\hat{y}_i, f(\hat{y}_{i+k})}_{\text{autocovariance function}}\Big)\right]$$

at equilibrium, Var and Cov do not depend on time but only on the lag. In particular $Var(f(\hat{Y}_i)) = \sigma_f^2$.

$$\frac{1}{N^2}\sum_{i=1}^{N}\left[\sigma_f^2 + 2\sigma_f^2 \underbrace{\sum_{k=1}^{N-i}\frac{Cov(f(\hat{Y}_i), f(\hat{Y}_{i+k})}{\sigma_f^2}}_{\gamma_k = Corr(f(\hat{y}_i, f(\hat{y}_{i+k}))}\right] \approx \frac{\sigma_f^2}{N}(1 + \sum_{k\geq 1}\gamma_k)$$

This is higher the the variance of the MCMC estimator with N chains (in fact it is a MC estimator). The quantity

$$\tilde{N} = \frac{N}{1 + 2\sum_{k\geq 1} \gamma_k}$$

is called **effective sample size**. Indeed

$$\begin{aligned} \frac{Var(\hat{\mu}_N)}{Var(\mu_N)} &= \frac{\sigma_f^2/\tilde{N}}{\sigma_f^2/N} \\ \implies ESS &= N\frac{Var(\hat{\mu}_N)}{Var(\mu_N)} \in [0, N] \end{aligned}$$

so the ESS reppresent the size of an iid smaple with the same variance of the MCMC sample, giving a measure of the loss of efficiency determined by using a correlated sample.
Here we are hoping γ_k decays fast. Other wise the common practice is to perform what is called thinning, that is setting

$$\hat{Y}_i = X_{n_0+ih} \qquad i = 1, \dots, N$$

for a chosen $h \in \mathbb{N}$. This lowering the correlation among succesive samples.
In fact debated practice

- now we discard a higher number of samples.
- it is belived that keeping all sample after the burn-in yield's a better approximation of μ_f.

Observation even if are correlated are pieces of information so don't waste them, on the other hand maybe you want to remove correlation.

2.5 Metropolis-Hastings algorithm

The **Metropolis-Hastings algorithm** has been selected among the 10 most important algorithms of the 20th century.

The idea consists in running a chain with arbitrary transition matrix, but sometimes suppressing transition (in a ”right way”). Given a target π on S, let:

- Q be an aribtrary, irreducible transition matrix called *proposal matrix*;
- A be a matrix with entries $a_{ij} \in [0, 1]$ called *acceptance matrix*.

Algorithm 5

For $n \geqslant 1$, if $X_{n-1} = i$:

- draw j with probability q_{ij};
- set $X_n = j$ with probability a_{ij}; else set $X_n = X_{n-1}$

Q provides proposal states: from row i draw state j, A tells us if we should go in j or stay where we are. The resulting transitions are:

$$p_{ij} = \begin{cases} q_{ij}a_{ij} & j \neq i \\ 1 - \sum_{k\in S} q_{ik}a_{ik} & j = i. \end{cases}$$

→ $Q = (q_{ij})_{i,j} \in S$ is aribtrary (which means we can choose a distribution easy to simulate from);

→ $P = (p_{ij})_{i,j\in S}$ is irreducible since Q is aperiodic and since $X_n = X_{n-1}$ with positive probability.

So if we show P has invariant distribution π, then it is positive recurrent and the ergodic theorem applies: π is therefore also the equilibrium distribution.
Idea: occasionally suppress transitions to state that are less likely ”with respect to π”, so to speak.

Proposition 2.1

The Metropolis Hastings algirthm with proposal matrix Q and acceptance probabilities

$$a_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

generates a **reversible chain** with respect to π.

Proof

We need to show that $\pi_i p_{ij} = \pi_j p_{ji}$ (detailed balance for the resulting chain).

$$\begin{aligned} \pi_i p_{ij} &= \pi_i q_{ij} a_{ij} \\ &= \pi_i q_{ij} \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\} = \\ &= min\Bigg\{\pi_i q_{ij}, \pi_j q_{ji}\Bigg\}. \end{aligned}$$

But since minimum is symmetric and

$$min\{\pi_j q_{ji}, \pi_i q_{ij}\} = min\{\pi_i q_{ij}, \pi_j q_{j1}\}.$$

starting from $\pi_j q_{ji}$ yields the same term, so $\pi_i p_{ij} = \pi_j p_{ji}$. □

Algorithm 6

Metropolis-Hastings Algorithm For $n \geqslant 1$, if $X_{n-1} = i$:

- draw $J = j$ with probability q_{ij};
- draw $U \sim Unif(0,1)$;
- if $U < \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$ set $X_n = j$; else $X_n = X_{n-1}$

This algorithm has important features:

- the proposal distribution is fully arbitrary, with the sole condition of being irreducible. This leaves us a lot of freedom.
- it applies even if the normalizing constant of pi is unknown, since

$$\frac{\pi_j}{\pi_i} = \frac{\tilde{\pi}_j/\cancel{z}}{\tilde{\pi}_i/\cancel{z}} = \frac{\tilde{\pi}_j}{\tilde{\pi}_i}$$

Example 2.3

$$\pi(\cdot) = cPois(\cdot;\lambda_1) + (1-c)Pois(\cdot;\lambda_2)$$

is a mixture of Poisson distributions. We want to reconstruct this target using ergodic average. Propose states with a simple MC: we will use a $B\&D\left(\frac{1}{2},\frac{1}{2}\right)$ (i.e. a symmetric random walk on $\mathbb{Z}_+$).

$$\begin{aligned} &\implies q_{ij} = \frac{1}{2} \qquad \text{for } j = (i \pm 1)^+ \\ & a_{ij} = \min\{1, \alpha_{ij}\} \qquad \alpha_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \\ &\implies \alpha_{ij} = \frac{cPois(j;\lambda_1) + (1-c)Pois(j;\lambda_2)}{cPois(i;\lambda_1) + (1-c)Pois(i;\lambda_2)} \cdot \frac{\cancel{1/2}}{\cancel{1/2}} \end{aligned}$$

This last equality is true for $j = (i \pm 1)^+$ and arbitrary for all other js, which are not going to be proposed wince we use a random variable. The code can be found in appendix. The random walk is not positive recurrent, but Metropolis-Hastings "corrects" the trajectory so that it basically becomes recurrent.

The previous example refers to a typical class of Metropolis-Hastings algorithms called *random walk Metropolis-Hastings*, where

$$X_n = X_{n-1} + Z_n$$

Where Z_n has symmetric distribution around zero. The example provided, in particular, is a special case called **Metropolis** algorithm, where the proposal is symmetric so that

$$q_{ij} = q_{ji} \implies a_{ij} = \min\left\{1, \frac{\pi_j}{\pi_i}\right\}.$$

So:

- if $\pi_j < \pi_i$ we accept the step with probability < 0;
- if $\pi_j > \pi_i$ we accept the step with probability 1.

We can interpret this consequence of the algorithm as the fact that if the density of the new step increases, we accept the step and we continue exploring that direction; otherwise, we still have a chance to accept the new steps even if it leads us to a zone with lower density. Of course, if we only accepted steps that increase the density we would basically have an optimization algorithm that maximizes probability density (which is not what we want).

Example 2.4

$$\pi(\cdot) = cN(\cdot;\mu_1,\sigma_1^2) + (1-c)N(\cdot;\mu_2,\sigma_2^2).$$

We can use a random walk Metropolis-Hastings that proposes

$$Y = X + Z \qquad Z \sim N(0, \sigma_0^2)$$

if $X_{n-1} = X$, and α_{ij} is now

$$\alpha(x,y) = \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}$$

where $q(y|x) = N(y;x,\sigma_0^2)$. They are symmetric, so

$$\alpha(x,y) = \frac{\pi(y)}{\pi(x)}$$

Another special case is the *independent Metropolis-Hastings*, where the proposal does not depend on the current state, i.e.

$$q_{ij} = q_j \implies a_{ij} = \min\left\{1, \frac{\pi_j q_i}{\pi_i q_j}\right\} = \min\left\{1, \frac{\pi_j/q_j}{\pi_i/q_i}\right\}.$$

Proposals are accepted with probability 1 if

$$w_j = \frac{\pi_j}{\pi_j} > \frac{\pi_i}{\pi_i} = w_i \qquad \text{(it's a reweighing)}$$

Analogies:

- with *importance sampling*: $w(x) = \frac{\pi(x)}{q(x)}$ is the importance weight, i.e. we accept with probability 1 if $w_j > w_i$ (which means there has been an improvement in importance weights).
- with *rejection sampling*: propose $z = j$ with probability q_j and accept if $y \leqslant \pi_j$ where $y|z = j \sim Unif(0, Mq)$ which is the same as

$$U < \frac{\pi_j}{Mq_j}, \qquad U \sim Unif(0,1)$$

that here becomes $U < \frac{\pi_j/q_j}{\pi_i/q_i}$ instead of $\frac{\pi_j}{Mq_j}$.

2.6 Gibbs sampling

Gibbs sampling is formally a special case of Metropolis-Hastings but it found wider applications. It is useful with:

- multivariate π;
- models with latent variables;
- models specified using conditionals.

Gibbs sampling needs at least a bivariate space. Let $\pi = \pi_{X,Y}$ density on $S \times S$. We cannot sample directly from π but we can sample from the conditionals $\pi_{X|Y}$ and $\pi_{Y|X}$.

Algorithm 7

```
Given $(X_{n-1}, Y_{n-1}) = (x, y)$ as our current state,
    - draw $X' \sim \pi_{X|Y}(\cdot|Y)$
    - draw $Y' \sim \pi_{X|Y}(\cdot|X)$
    - set $(X_n, Y_n) = (X', Y')$
```

Why is this useful? If $(X, Y) \sim \pi_{x,y}$ this is equivalent to say that

$$Y \sim \pi_y \text{ (marginal distribution)}, \quad X|Y \sim \pi_{X|Y} \text{ (chain rule)}$$

but if $X' \sim \pi_{X|Y}$ then

$$(X', Y) \sim \pi_{X,Y}$$

which is obvious, since we drew it that way. The same holds for Y', so these transitions preserve the joint distribution $\pi_{X,Y}$ which is therefore invariant.

Proof

The first step of the algorithm can be seen as a Metropolis-Hastings step with proposal on $S^2 = S \times S$:

$$q = (X', Y'|X, Y) = \pi_{X'|Y'}(X'|y)\mathbb{1}(Y' = y).$$

In practice, we fix Y and then we update X. The acceptance rate is $\min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$. If we factorize in marginals and conditional we get:

$$\frac{\pi(X'Y')q(X, Y|X', Y')}{\pi(X, Y)q(X', Y'|X, Y)} = \frac{\pi(Y')\pi(X'|Y')\pi(X|Y')\mathbb{1}(Y = Y')}{\pi(Y)\pi(X|Y)\pi(X'|Y)\mathbb{1}(Y = Y')}$$

but since we imposed $Y = Y'$, all these factors cancel out:

$$= \frac{\cancel{\pi(Y')\pi(X'|Y')}\,\cancel{\pi(X|Y')}\,\cancel{\mathbb{1}(Y = Y')}}{\cancel{\pi(Y)\pi(X|Y)}\,\cancel{\pi(X'|Y)}\,\cancel{\mathbb{1}(Y = Y')}} = 1.$$

□

So every step of the Gibbs samples is made of 2 mini Metropolis-Hastings steps with probability of acceptance 1: we never reject any data. If we are only interested in X univeriates we can sometimes augment the state space to $S \times S'$ through can auxiliary variable $Y \in S'$ paradoxally introduced to ease the computation. The transition for X (marginal) can be written, integrating Y out, as

$$p_X(x'|X) = \int_{S'} \pi_{Y|X}(y|x)\pi_{X|Y}(x'|y)dy.$$

Stationarity requires, by the global balance conditions:

$$\int_S \pi_X(x)p_X(x'|X)dx = \pi_X(x').$$

We have

$$\int_S \pi_X(x)p_X(x'|X)dx = \int_S \pi_X(x)\left(\int_{S'} \pi_{Y|X}(y|X)\pi_{X|Y}(x'|Y))dx\right) dy.$$

Using Fubini's theorem

$$\int_{S'} \pi_{X|Y}(x'|Y) \int_S \underbrace{\pi_{Y|X}(y|X)\pi_X(X)}_{\pi_{X,Y}(X'Y)} dxdy$$

$$\int_S \underbrace{\pi_{X|Y}(x'|Y)\pi_Y(y)}_{\pi_{X,Y}(X',Y)} dy = \pi_X(x')$$

A Gibbs sampler on (X_n, Y_n) yields, marginally, a stationary X_n chain with respect to π_X. If we are interested in X, Y can be sometimes an auxiliary variable:

- we run the Gibbs sampler on (X_n, Y_n)
- discard Y_n

$$\Longrightarrow \quad X_n(\text{ at stationarity })$$
$$\text{is from } \pi_x = \int \pi_{X,Y}(x,y)dy$$

Example 2.5

In a popular Bayesian model we have

$$\pi_X(x) \propto x^{\alpha+k-1}e^{-\beta x}\frac{\Gamma(x)}{\Gamma(x+n)}, \qquad x \geqslant 0.$$

Normalizing and simulating from π is not trivial, due to the presence of gamma functions. However,

$$\frac{\Gamma(x)\Gamma(n)}{\Gamma(x+n)} = \int_0^1 y^{x-1}(1-y)^{n-1}\,\mathrm{d}y$$

so we can formulate a joint model on the augmented state space $\mathbb{R}_+ \times [0,1]$.

$$\pi_{X,Y}(x,y) \propto x^{\alpha+k-1}e^{-\beta x}\overbrace{y^{x-1}(1-y)^{n-1}}^{\text{Beta kernel}}$$

$$\Longrightarrow \int_0^1 \pi_{X,Y}(x,y)\,\mathrm{d}y = \pi_X(x)$$

So we have

- for fixed x, $\pi(y|x) = Beta(x,n)$
- for fixed y,

$$\pi(x|y) \propto x^{\alpha+k-1}x^{\alpha+k-1}e^{-\beta x}\underbrace{e^{\ln(y^x)}}_{=y^x}$$

$$= \underbrace{x^{\alpha+k-1}e^{-(\beta x-\ln y)x}.}_{x^a e^{-bx}\text{: Gamma distribution, once we fix } y}$$

> So the Gibbs sampler runs
>
> - $X|Y \sim Gamma(\alpha + k, \beta - \ln y)$
> - $Y|X \sim Beta(x, n)$
>
> and the just discards discard the value of Y.

More generally, let $\pi(x) = \pi(x_1, \dots, x_d)$ be a density on S^d, $d \geqslant 2$.
Denote $x_{(-i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$. Assume we know how to sample from the full conditional distributions

$$\pi(x_i|x_{(-i)})$$

The i^{th} component Gibbs sampler updates X_i and leaves the other coordinates unchanged, so

$$\left.\begin{array}{l}(X_1, \dots, X_i, \dots, X_d) \sim \pi \\ X_i' \sim \pi(x_i|x_{(-i)})\end{array}\right\} \implies (X_i, \dots, X_i', \dots, X_d) \sim \pi.$$

So π is invariant and this is a Metropolis-Hastings step with proposal

$$q(x'|x) = \pi((x_i|x_{(-i)})\mathbb{1}(x'_{(-i)} = x_{(-i)})$$

and the acceptance probability is 1.

The full Gibbs Sampler typically reads:

Algorithm 8

Given $X_n = x \in S^d$:

- draw i with probability p_i ($\sum^d p_i = 1$);
- draw $X_i' \sim \pi_{X_i|X_{(-i)}}(\cdot|x_{(-i)})$;
- set $X_{n+1} = (x_1, \dots, x_{i-1}, x_{i-1}, x_{i+1}, \dots, x_d)$.

If $p_i > 0$ for all $i = 1, \dots, d$, this algorithm produces a reversible chain with respect to π. This Gibbs sampling is so powerful that the cases in which we should *not* use it are very specific.

2.7 Slice sampler

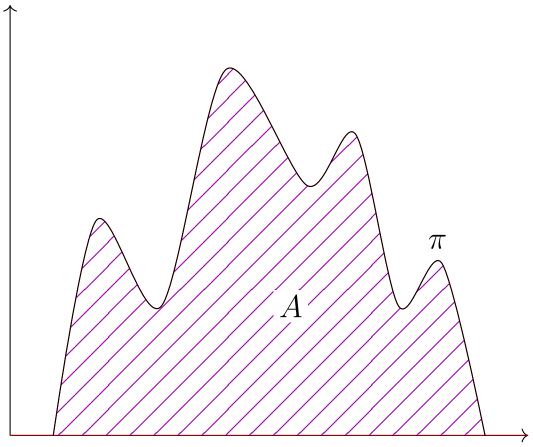
Let π_X be a density on S.

Lemma 2.1

Let A be the area under π_x, i.e.

$$A = \{(x, y) \in S \times \mathbb{R}_+ : 0 \leqslant y \leqslant \pi_X(x)\}$$

If (X, Y) has uniform distribution on A, then $X \sim \pi_x$



Proof

If (X,Y) is uniform on A we have $\pi_{X,Y}(x,y) \propto \mathbb{1}_{0 \leqslant y \leqslant \pi(x)}$ whose normalizing constant is

$$\int_S \int_0^\infty \mathbb{1}_{0 \leqslant y \leqslant \pi(x)} \, \mathrm{d}y \, \mathrm{d}x = \ldots = 1.$$

Then the marginal of X is

$$\begin{aligned} \int_0^\infty \pi_{X,Y}(x,y) \, \mathrm{d}y &= \int_o^\infty \mathbb{1}_{0 \leqslant y \leqslant \pi(x)} \, \mathrm{d}y \\ &= \int_0^{\pi_X(x)} \mathrm{d}y = \pi_X(x) \end{aligned}$$

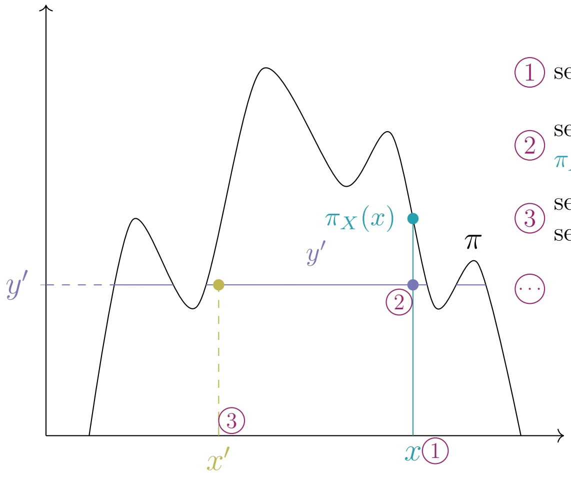
□

AR methods are based on this result. How does this connect with Markov-Chain Monte-Carlo? The main idea is that we can use a Markov Chain whose equilibrium distribution is a uniform on A, then discarded Y and keep X. The target is

$$\pi_{X|Y}(x,y) \alpha 1_{0 \leqslant y \leqslant \pi(x)}$$

and we need to construct a Markov Chain that is ergodic with respect to this target. We can think a Gibbs sampler that alternate the draws from the two condition:

- $Y'|X \sim \pi_{Y|X}(y|x) \propto \mathbb{1}_{0 \leqslant y \leqslant \pi(x)}$ (x is fixed: this step sets the height of the *vertical slice*);
- $X'|Y \sim \pi_{X|Y}(x|y) \propto \mathbb{1}_{0 \leqslant y \leqslant \pi(x)}$ (y is fixed: this step sets the length of the *horizontal slice*).



① select an x;

② select a y' ranging from 0 to $\pi_X(x)$, as long as it is *below* π;

③ select a x' ranging from the segments of the y' slice that are *below* π;

(⋯)

So these are reversible step with respect to $\pi_{X,Y}$ and at equilibrium we have uniform sample on A: then using the lemma 2.7, X is from π_X.

Example 2.6

$S = \mathbb{R}_+, \qquad \pi(x) = \frac{1}{2}e^{-\sqrt{x}}, x > 0.$

$$\pi^{-1}(y) = (\log 2y)^2$$

so the slice sampler is

- $Y'|x \sim$ Unif $(0, \frac{1}{2}e^{-\sqrt{x}})$
- $X'|y' \sim$ Unif $(0, (\log 2y))^2$

$\pi(x) \propto e^{-\sqrt{x}}, x \in \mathbb{Z}_+$ unnormalized.

$$\pi^{-1}(y) = (\log(y))^2$$

- $Y'|x \sim$ Unif $(0, e^{-\sqrt{x}})$
- $X'|y' \sim$ Unif$(\{0, 1, \dots, \lfloor(\log(y))^2\rfloor\})$

Remember that

$$\lceil x \rceil = \max\{n \geqslant 0 : n \leqslant x\}$$

If π is d-dimensional, it is typically difficult to identify

$$A_y = \{x : y \leqslant \pi(x)\}$$

If we can write

$$\pi(x) \propto \prod_{i=1}^{d} \pi_i(x)$$

we can use d auxiliary variables $y_1, \dots, y_d$, so that

$$\pi_i(x) = \int_0^{\pi_i(x)} \mathrm{d}y_i = \int \mathbb{1}_{0 \leqslant y_i \leqslant \pi_i(x)} \,\mathrm{d}y_i$$

so the augmented target is

$$\pi(x, y) = \pi(x_1, \dots, x_d, y_1, \dots, y_d) \propto \prod_{i=1}^{d} \mathbb{1}_{0 \leqslant y_i \leqslant \pi_i(x)}$$

In the following graphs, the chain oscillates but stays stable around a certain level. Even if it seems not to converge, the oscillations are actually pretty stable and do not change throughout time. So the chain is convergent in the end.

Sometimes convergence of the MCMC is deceptive, as one run exhibits convergence but multiple runs reveal differently.
In these cases we can combine different strategies.
For example, if the transition matrices P' and P'' both have invariant π, the convex linear combination

$$P = wP' + (1 - w)P'', \qquad w \in (0, 1)$$

has invariant π (exercise).
This is called a **mixture transition**. This means that with probability w we use the matrix P' and with compementary probability we use the matrix P''.
For example, one could choose

- P' as a RW-MH to explore locally

- P'' as an independent MH (that is, the proposal is independent on the current space: it is fixed) to explore globally

so with w close to 1, the chain once in a while takes a jump. This is useful to explore distribution with particularly low density areas that the chain will cross very seldom.
Example: local exploration and sometimes the chain takes a jump. More generally

$$P = \sum_{i=1}^{k} w_i P : i, \qquad \sum_{i=1}^{k} w_i = 1, \pi P_i = \pi$$

Another idea is to use a **cycle transition**.

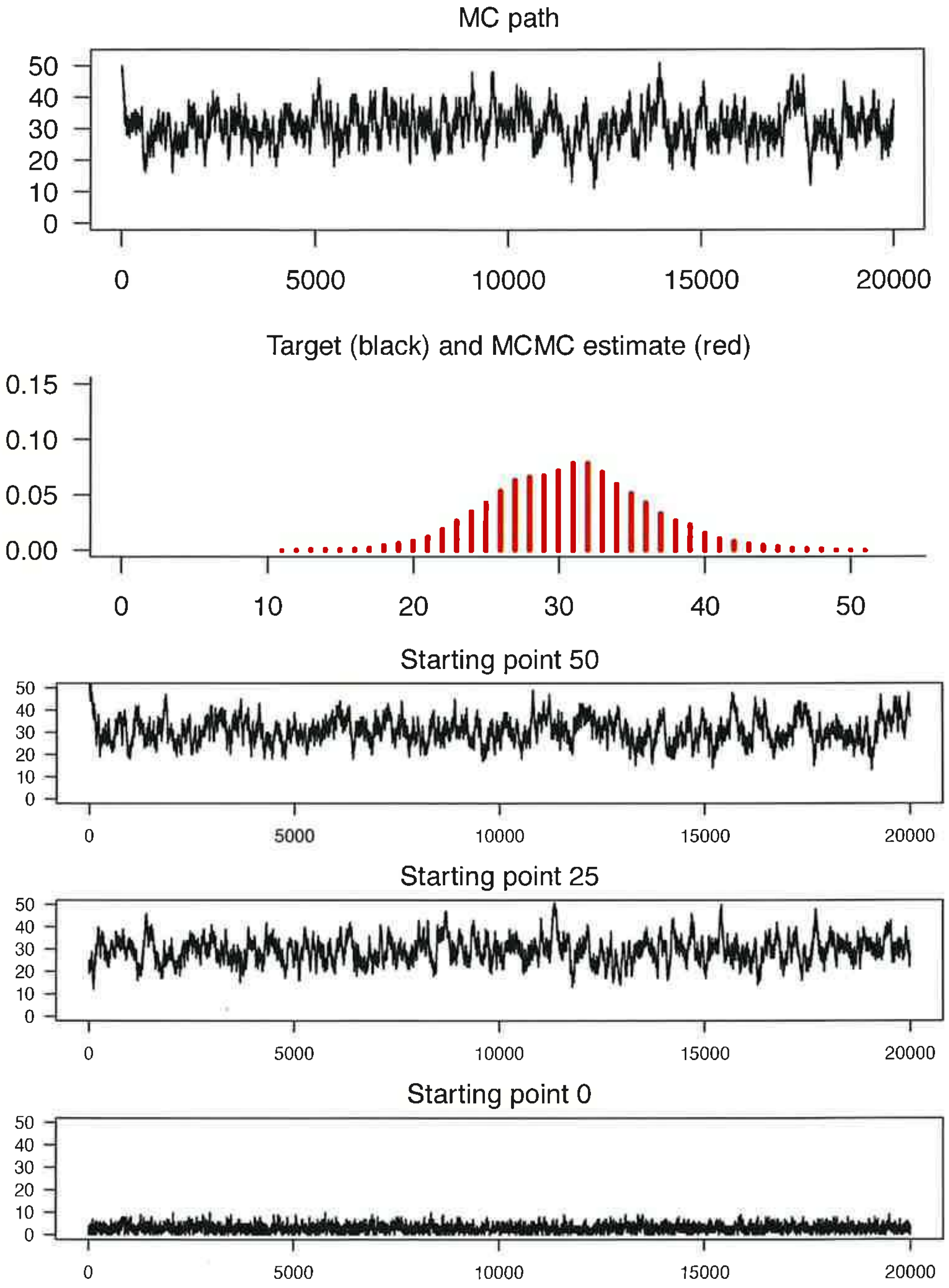
$$P = P' P''$$

which leaves π invariant (exercise), but not reversible in general.
More generally,

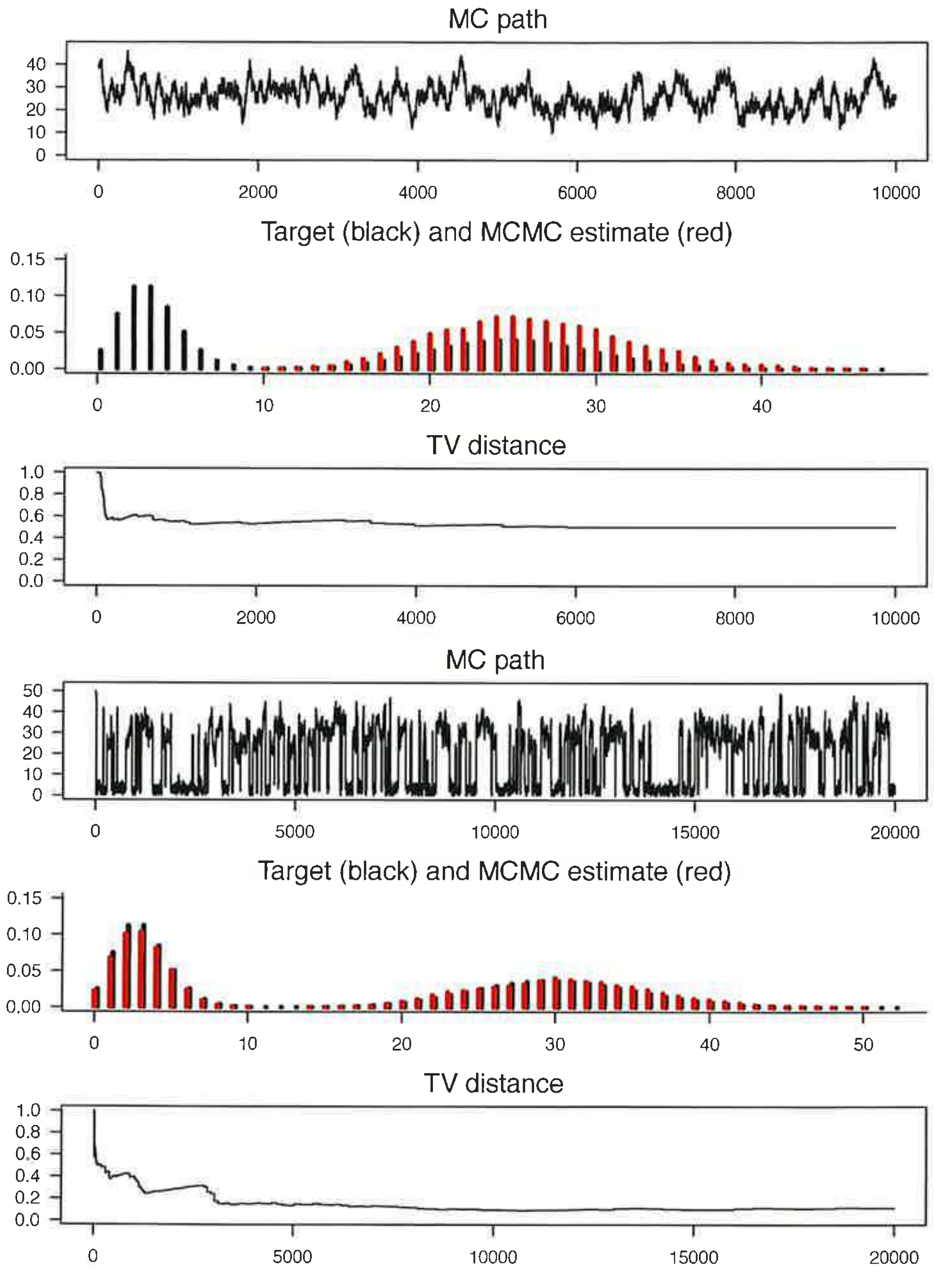
$$P = P_1 P_2 \dots P_k$$

For example, the multicomponent Gibbs sampler (with deterministic visits to coordinates).

This method aims to update k coordinates; when one of them is difficult to update it uses a Metropolis step to do it.



Top: the RW-MH seems to have converged. Bottom: a comparison of chains with different starting points reveals it has not.



Top: RW-MH trapped near a local maximum and unable to efficiently explore the state space (note the possible interpretation as a *quasi stationary distribution*). Bottom: a mixed kernel $P = cP' + (1-c)P''$ with P' a RW-MH and P'' an independent MH with a fixed Geometric proposal allows the chain (on average once every $1/(1-c)$ steps) to jump farther and overcome valleys between local maxima. The different *mixing*, i.e. the different way of exploring the space, is evident from the MC path.

Extensions and other methods:

- **Reversible jump MCMC**: it moves among spaces of different dimensions.
- **Langevin algorithms** or **gradient-based MCMC**: is it inspired by the Langevin diffusion

$$dX_t = \frac{1}{2}\frac{d}{dX_t}\log \pi(X_t)dt + dB_t$$

It uses information on the gradient of π to move towards regions of higher density.

There is a principle (Goldilocks's principle of MCMC): the variance of the steps has to be not too large, not too small.

3 Continuous time Markov chains

Definition 3.1

A **continuous time Markov chain (CTMC)** is a stochastic process $\{X(t), t \in T\}$ with index set $T = [0, +\infty)$ taking values in a countable set S, s.t.

$$\mathbb{P}(X(t_n) = i_n | X(t_0) = i_0, \dots, X(t_{n-1}) = x(t_{n-1}))$$

i.e. the Markov property holds for all $i_0, \dots, i_n \in S$ and $0 \leqslant t_0 < \cdots < t_n$.
We can define the transition probability

$$p_{ij}(s, s+t) := \mathbb{P}(X(s+t) = j | X(s) = i)$$

for $t > 0$, which we assume are time homogeneoCus, i.e. depend on t not on s. Hence we can define

$$p_{ij}(t) := p_{ij}(0, t)$$

since $p_{ij}(s, s+t) = p_{ij}(0, t)$.
Set

$$p_{ij}(0) := \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

Example 3.1

The Poisson process with intensity $\lambda > 0$ is defined by $X(0) = 0$ a.s. and

$$p_{ij}(t) = \frac{(\lambda t)^{j-1} e^{-\lambda t}}{(j-i)!}, \qquad j \geqslant 1$$

and 0 elsewhere. We see it is time-homogeneous on $S = \mathbb{Z}_+$. The increments are s.t.

$$X(s+t) - X(s) \sim Poisson(\lambda t)$$

and if we set $s = 0$, then $X(t) \sim Poisson(\lambda t)$.

Exercise 3.1

Show that the transition function of a MTMC, together with an initial distribution, determines all joint distribution

$$\mathbb{P}(X(t_0) = i_0, \dots, X(t_n) = x(t_n))$$

for all choice of $i_0, \dots, i_n$ and $t_o < \cdots < t_n$.

Denote P_t the square matrix with entries $\{p_{ij}(t)_i, j \in S$.
the family of $\{P_t\}_{t\geqslant 0}$ is called **transition semigroup**, since:

- $\sum_{j\in S} p_{ij} = 1$: row sum
- $P_0 = I$ (i.e. $p_{ij}(0) := \delta_{ij}$)
- $P_{t+s} = P_t P_s$: **semigroup property**, namely the Chapman-Kolmogorov's equations

$$p_{ij}(s+t) = \sum_{k\in S} p_{ik}(s)p_{kj}(t)$$

We will assume P_t is continuous at the origin (called *standard*)

$$p_{ij}(t) \to \delta_{ij} \qquad t \to 0$$

which implies continuity at all $t > 0$.
P_t is parametrized by time, and ideally we would like on object similar to the transition matrix for DTMC's, , relative to the temporal unit. We look at infinitesimal intervals.

Definition 3.2

The matrix Q given by

$$Q = \frac{d}{dt}P_t|_{t=0}$$

is called infinitesimal generator of the chain, and its entries $(q_{ij})_{i,j\in S}$ are called **infinitesimal rates**.

Proposition 3.1

Let P_t be a continuous transition semigroup on S countable. For all i, j the quantity $q_{ij} := p'_{ij}(0)$ exists and

$$q_{ij} \in \begin{cases} [0,\infty) & j \neq i \\ [-\infty, 0] & j = i \end{cases}$$

For $j \neq i$

$$q_[ij] = \lim_{h\to 0} \frac{p_{ij}(h) - \overbrace{p_{ij}(0)}^{0}}{h}$$

- $\implies p_{ij}(h) = q_{ij}h + o(h)$
- $\implies X$ goes to j from i with probability approximately $q_{ij}h$.
- $\implies o(h)$ collects the error of approximation and all other events (for example, X) goes to k before j).

So the chain has at most one jump at each instant. When $j = i$, since $p_{ii}(0) = 1$, the above calculation gives $q_{ii} \leqslant 0$.
Sometimes we will denote $q_i = -q_{ii}$.

Definition 3.3

A state i is called

- **absorbing** if $q_i = 0$

- **stable** if $0 < q_i < \infty$
- **instantaneous** if $q_i = \infty$

If $\sup_{i \in S} q_i < \infty, Q$ is **stable**.
If $q_i = \sum_{j \neq i} q_{ij}, \quad \forall i \in S, Q$ is called **conservative**.

Remark

The absolute value of the diagonal entry equals the sum of the off-diagonal entries.

$$1 = \sum_{j \in S} p_{ij}(h) = \sum_{j \neq i} p_{ij}(h) + p_{ii}(h)$$

which leads to

$$\frac{1 - p_{ii}(h)}{h} = \sum_{j \neq i} \frac{p_{ij}(h)}{h}$$

Taking the limit as h decreases to 0, we get

$$q_i = \lim_{h \downarrow 0} \sum_{j \neq i} \frac{p_{ij}(h)}{h}$$

If limit and sum exchange, we get

$$q_i = \sum_{j \neq i} q_{ij} \implies \sum_{j \in S} q_{ij} = 0$$

Remark

Conservativity comes from the fact that transition probabilities conserve the mass.

We assumed we had the transition probabilities are available but this is rarely the case. In many cases, when we move on to general and complex spaces, probabilities are specified through the generator (instead of the transition semigroup). We are going to argument later that this gives complete information about the process too, but if we have the generator, how can we check that the implied process is well defined? We need criteria to be checked on the generator. Often, only Q is available, while P_t is unknown. So, this identity relates to the conservation of probability mass by the transition semigroup. So it is enough to check that Q row sums are null (to verify this condition).
If Q is stable and conservative, q_i is interpreted as the rate of leaving the state i. So

$$p_{ii}(h) = 1 - q_i h + o(h)$$

If

- $q_i = 0$, then X leaves i at rate 0, so i is absorbing.
- $q_i = \infty$, then x leaves the state i instantaneously.

Proposition 3.2

A CTMC with finite state space is always stable and conservative.

So we can reverse the statement, saying that trouble scenario happen when the chain is not stable or not conservative.

Example 3.2

Consider X on $\mathbb{Z}_+$ with

$$q_{i,i+1} = \lambda \qquad q_{i,i-1} = i \qquad q_{ii} = -(\lambda + i)$$

and 0 elsewhere. Q is conservative (P_t would be well defined, if we knew it) but

$$\sup_{i \in S} q_i = \sup_{i \geqslant 0} \lambda + i = \infty$$

so it is not stable.
The meaning of this distinction will be treated later.

Definition 3.4

A general class of stable and conservative CTMC's is given by **regular jump processes** such that

- their paths are piecewise constant: for almost all ω and all $t \geqslant 0, \exists \varepsilon = \varepsilon(t, \omega)$ such that
$$X(t+s, \omega) = X(t, \omega) \qquad \forall s \in [0, \varepsilon]$$
This means that the chain remains where it is for a positive amount of time.
- they have finitely many points of discontinuity in every bounded time interval.

The typical behaviour is the one represented in Figure 1. For example, **explosive chains** are not regular in the above sense (see Figure 2): these chains have closer and closer jumps until there are infinitely many around $x = s$.

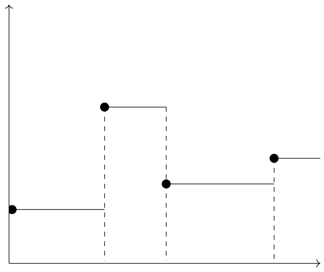


Figure 1: Regualar jumps process

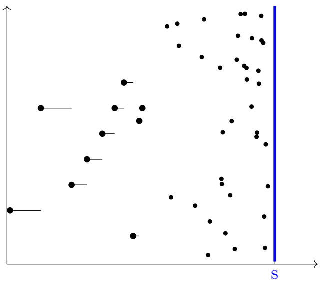


Figure 2: explosive chain

Example 3.3

Poisson process

$$\begin{aligned} p_{ij}(t) &= \frac{(\lambda t)^{j-i} e^{-\lambda t}}{(j-i)!} \qquad j \geqslant 1 \\ \implies p_{ii}(t) &= e^{-\lambda t} \\ p_{i,i+1}(t) &= \lambda t e^{-\lambda t} \\ p_{i,i+k}(t) &= \frac{(\lambda t)^k e^{-\lambda t}}{k!} \\ \implies q_{ii} &= -\lambda, \quad q_{i,i+1} = \lambda, \qquad \text{and 0 elsewhere.} \end{aligned}$$

Interarrival times: T_i waiting time in i.

$$\begin{aligned} \mathbb{P}(T_0 > t) &= \mathbb{P}(X(t) = 0 | X(0) = 0) = e^{-\lambda t} \\ \implies \mathbb{P}(T_0 < t) &= 1 - e^{-\lambda t} \implies T_0 \sim \text{Exp}(\lambda) \end{aligned}$$

So $\mathbb{P}(T_i > t | T_0 = t_0, \ldots, T_{i-1} = t_{i-1}) =$

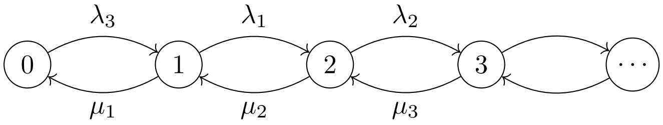
$$\begin{aligned} &s = t_0 + \ldots + t_{i-1} \\ &= \mathbb{P}(x(s+t) = i | x(s) = i) = e^{-\lambda t} \\ &\implies T_i \sim \text{Exp}(\lambda) \end{aligned}$$

Example 3.4

(Birth and Death processes)
Denote B&D(λ_i, μ_i) such that

$$q_{ij} = \begin{cases} \lambda_i & j = i+1 \quad i \geqslant 0 \\ \mu_i & j = i-1 \quad i > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$q_{ii} = -(\lambda_i + \mu_i)$$



whose invariant π is the uniform on on S_0. Let τ be a finite random time on $\mathbb{N}$, independent of Y, such that:

Special cases:

- $\mu_i = 0 \ \forall i, \lambda_i > 0$ **Birth process** which can be
  - ⋆ Poisson
  - ⋆ $\lambda_i = \underbrace{\lambda}_{\text{per capita birth rate}} \cdot i$ (**Yule process**)

  $\lambda_i = \lambda \cdot i + c$, adding immigration rate c.
- $\lambda_i = 0 \forall i, \mu_i > 0$ (**Pure death process**)

For example, if

$$\lambda_i = \lambda \cdot i, \qquad \mu_i = \mu \cdot i$$

the transition probabilities are very complicated to find, like in the following case: the process has $i = 1$

$$p_{1j}(t) = (1 - \mu\gamma)(1 - \lambda\gamma)(\lambda\gamma)^{j-i}$$

where

$$\gamma = \frac{e^{(\lambda-\mu)t} - 1}{\lambda e^{(\lambda-\mu)t} - \mu}$$

If X is specified only through the generator Q, a general task is to be able to describe the transition semigroup (which is difficult in general).
A first tool to do so is given by the so called **Kolmogorov's equations**.

Theorem 3.1

Let Q be stable and conservative. Then, P_t and Q satisfy:

- The **backward equation**

$$P_t' = QP_t \qquad p_{ij}'(t) = \sum_{k\in S} q_{ik}p_{kj}(t)$$

- If

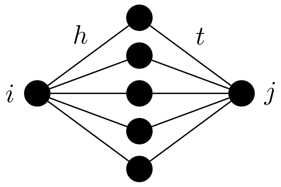
$$\sum_{k\in S} p_{ik}(t)q_k < \infty$$

  the **forward equation**

$$P_t' = P_tQ \qquad p_{ij}'(t) = \sum_{k\in S} p_{ik}(t)q_{kj}$$

Exercise 3.2

Prove the forward equation for S finite using CK and linearization.
The intuition consists in thinking about the backwards equation.



To obtain an intuition for the forward equation, let the second interval go to 0.

In principle (which boils down to simple cases), one could solve the Kolmogorov's equations to find the semigroup.

Exercise 3.3

Consider the generator of the Poisson process:

$$q_{i,i+1} = \lambda \qquad q_{ii} = -\lambda$$

Derive the transition probabilities from Kolmogorov's equations.

First of all, we are going to analyze an example to show how Kolmogorov equations can be used and also their complexity. Let's recall the Kolmogorov Backward Equation

$$p_{ij}'(t) = \sum_{k\in S} q_{ik}p_{kj}(t)$$

and the Kolmogorov Forward Equation

$$p_{ij}'(t) = \sum_{k\in S} p_{ik}(t)q_{kj}$$

which provide a relationship between P_t and Q. In simple cases they can be solved.

Example 3.5

Poisson Process

$$q_{i,i+1} = \lambda \qquad q_{ii} = -\lambda \text{ and } 0 \text{ elsewhere}$$

Let's consider the K.F.E. and let's consider

- $j = i$:

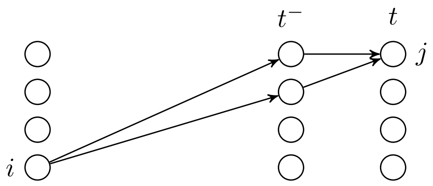
  $$p'_{ii}(t) = p_{ii}(t)q_{ii} = -\lambda p_{ii}(t)$$

  $k = i$ necessarily. This yields

  $$\begin{cases} p_{ii}(t) & = ce^{-\lambda t} \\ p_{ii}(0) & = 1 \implies c = 1 \end{cases}$$

  We find the constant using the boundary condition $p_{ii}(0) = 1$.

- $j > i$



$$\begin{aligned} p'_{ij}(t) &= p_{i,j-1}(t)q_{j-1,j} + p_{ij}(t)q_{jj} \\ &= \lambda p_{i,j-1}(t) - \lambda p_{ij}(t) \end{aligned}$$

Now, multiply by $e^{\lambda t}$ and get from

$$p'_{ij}(t) = \lambda p_{i,j-1}(t) - \lambda p_{ij}(t)$$

the following

$$\underbrace{e^{\lambda t}p'_{ij}(t) + \lambda e^{\lambda t}}_{\frac{\mathrm{d}}{\mathrm{d}t}\{e^{\lambda t}p_{ij}(t)\}} = \lambda e^{\lambda t}p_{i,j-1}(t)$$

assume $p_{i,j-1}(t) = \frac{(\lambda t)^{j-i-1}e^{-\lambda t}}{(j-i-1)!}$ and integrate to get

$$\frac{d}{dt}(e^{\lambda t}p_{ij}(t)) = \lambda \cancel{e^{\lambda t}} \frac{(\lambda t)^{j-i-1}\cancel{e^{-\lambda t}}}{(j-i-1)!} = \frac{\lambda^{j-i}}{(j-i-1)!}t^{j-i-1}$$

Which leads to

$$e^{\lambda t}p_{ij}(t) = \frac{\lambda^{j-i}}{(j-i-1)!}\frac{t^{j-i-1}}{j-i} + c'$$

which yields

$$\begin{cases} p_{ij}(t) & = \frac{(\lambda t)^{j-i}e^{-\lambda t}}{(j-i)!} + c' \\ p_{ij}(0) = 0 \implies c' = 0 & \end{cases}$$

What are the general solutions of the Kolmogorov equations?

Theorem 3.2

Let S be finite. Then

$$P_t = e^{tQ}, \qquad P_0 = I$$

is a stochastic semigroup and the unique solution of the K.E.'s.

Remark

The matrix exponential

$$e^{tQ} := \sum_{n\geqslant 0} \frac{(tQ)^n}{n!}$$

is convergent component-wise since $e^{tq_{ij}} < \infty, \forall t > 0$.

If A, B commute,

$$e^{A+B} = e^A e^B$$

which yields

$$P_{t+s} = e^{t+s}Q = e^{tQ}e^{sQ} = P_t P_s$$

Intuition: consider

$$\begin{aligned} p_{ij}(t) &= \delta_{ij} + q_{ij}t + o(t) \qquad \forall i, j \\ P_t &= I + tQ + o(t) \end{aligned}$$

Use the semigroup property to split $[0, t]$ into n subintervals, so

$$P_t = \underbrace{P_{\frac{t}{n}} \dots P_{\frac{t}{n}}}_{n \text{ times}} = (I + tQ + o(t))^n$$

which generate the exponential e^{tQ} as $n \to +\infty$. The generator also gives important information about the trajectories of the chain.

Definition 3.5

Define the Holding Times (or waiting).

$$T_i = \inf\{t \geqslant 0 : X(s+t) \neq i | X(s) = i\}$$

in state $i \in S$, with $T_i = \infty$ if i absorbing.

Proposition 3.3

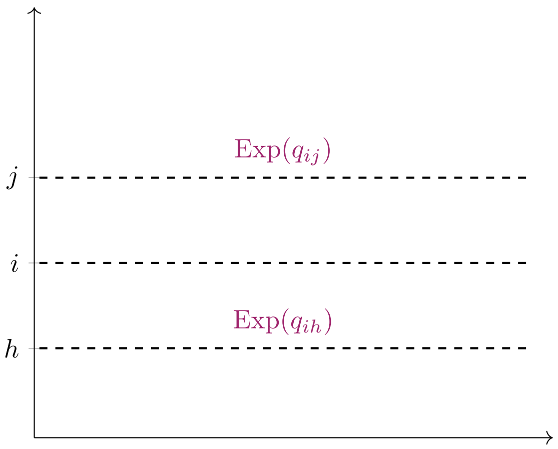
Let X be a regular jump CTCM with generator Q, and let T_i be the first holding time given $X(0) = i$. Then

- $T_i \sim Exp(q_i)$
- for $j \neq i$

  $$\mathbb{P}(X(T_i) = j | X(0) = i) = \frac{q_{ij}}{q_i}$$

  and these are independent.

Interpretation:



We can immagine that at every state there is a poisson process that rings at every time, exponential (?????).
There is a Poisson prosses with rate q_{ij} each $j \neq i$ acting as a clock (all independent). Competing clocks: the first ringing sets the new state.🕒
We know

$$\begin{aligned} &Z_k \overset{ind}{\sim} Exp(\lambda_k) \\ &\quad\implies \min_k Z_k \sim Exp(\sum_k \lambda_k) \qquad \text{shortest interval} \\ &\mathbb{P}(\min_{k \neq i} Z_k = Z_j) = \frac{\lambda_j}{\sum_h \lambda_h} hspace 0.5cm \text{ prop of j-th is the fist to ring} \end{aligned}$$

So $T_i \sim Exp(\sum_{j \neq i} q_{ij}) = Exp(q_i)$ and given T_i, the probability of going to j is

$$\frac{q_{ij}}{\sum_{j \neq i} q_{ij}} = \frac{q_{ij}}{q_i}$$

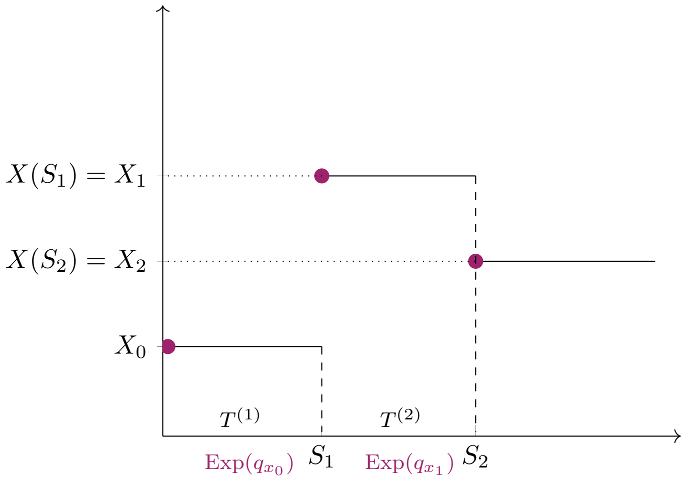
We can extend the above to the entire trajectory of the CTMC through the strong Markov property:

Theorem 3.3

A regular jump CTMC X with generator Q is a strong markov process, i.e. if τ is a stopping time w.r.t. to $\mathcal{F}_t^X = \sigma(X_u, u \leqslant t)$, conditional on $\tau < \infty$ and $X(\tau) = i$, $\{X(\tau + t), t \geqslant 0\}$ is a CTMC with initial distribution δ_i and generato Q, independent of $\{X(u), u \leqslant t\}$

If we now denote

- $T^{(1)}, T^{(2)}, \dots$ successive holding times
- $S_n = \sum_{i=1}^{n} T^{(i)}$ jumps times (Stopping times)
- $X_n := X(S_n)$



X_n is a DTMC called **embedded chain** with transition probabilities

$$p_{ij} = \begin{cases} \frac{q_{ij}}{q_i} & j \neq i \\ 0 & j = i \end{cases}$$

Note X_n is not allowed self transitions. Given $(X_n)_{n \geqslant 1}$

$$T^{(n)} \overset{\text{ind}}{\sim} Exp(q_{X_{n-1}})$$

We have factorised the state set of the chain.

Example 3.6

Birth and Death (λ_i, μ_i):

$$q_{ij} = \begin{cases} \lambda_i & j = i+1, i \geqslant 0 \\ \mu_i & j = i-1, i > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$q_{ii} = -(\lambda_i + \mu_i), \qquad q_i = \lambda_i + \mu_i \tag{6}$$

$\Longrightarrow$ holding times in i are

$$T_i \sim \text{ Exp}(q_i) = \text{ Exp } (\lambda_i + \mu_i)$$

$\Longrightarrow$ embedded chain X_n has transitions

$$p_{ij} \begin{cases} \frac{\lambda_i}{\lambda_i + \mu_i} & j = i+1 \\ \frac{\mu_i}{\lambda_i + \mu_i} & j = i-1 \\ 0 & \text{elsewhere} \end{cases}$$

A corollary of the above argument is the following:

Proposition 3.4

Two regular jump CTMC's woth the same Q, and initial distributions, have the same transition semigroup.

This is important since we know we can proceed even without the semigroup. Embedded chain and holding times are fully determined by th egenerator Q, which therefore provides an equivalent characterization of the CTMC. The above suggests an imemdiate strategy to simulate the chain's trajectory:

- Draw X_0 from the initial distribution
- for $n \geqslant 1$,if $X_{n-1} = i$
  - draw $T \sim \text{Exp}(q_i)$
  - given $T = t$, set
    $$X(t) = j \text{ with probability } \frac{q_{ij}}{q_i}$$

  called **gillespie** algorithm.

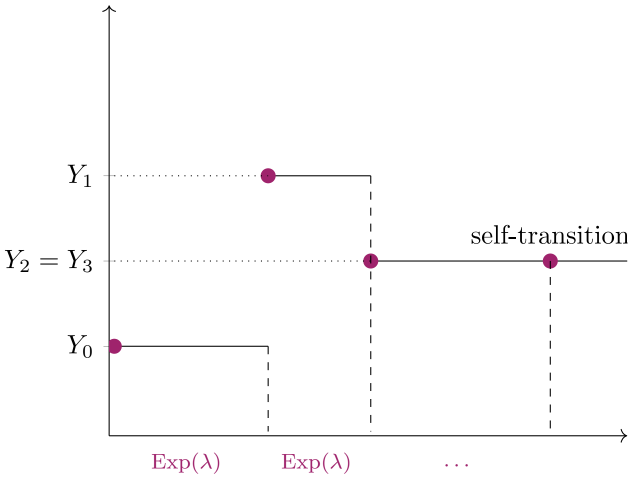
So, given a CTMC and generator - for every state - we draw the exponential with parameter represented by the first state and then we draw the normalized probabilities, iterating this process. Another subclass of CTMC which allows for an explicit description of P_t is the one defined below.

Definition 3.6

Let Y_n be a CTMC with transition matrix $\hat{P}$, and $N(t)$ a Poisson process with rate λ, independent of Y_n. The process

$$X(t) := Y_{N(t)}$$

is called **uniform chain with jump matrix** $\hat{P}$.



We have

$$X(t) = Y_n \qquad \forall t \text{such that } N(t) = n$$

Such $N(t)$ is called **subordinator** and $Y_{N(t)}$ is called **subordiated** chain.

	Holding times	**Self transitions**
Embedded chains	$T^{(n)} \overset{\text{ind}}{\sim} \text{Exp}(q_{x-1})$	not allowed by construction
Uniform chains	$T^{(n)} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$	if $\hat{p}_{ii} > 0$

Why are they interesting? Uniform chains allow an explicit form for the semigroup.

Proposition 3.5

Let X be a uniform chain with jump matrix $\hat{P}$ and rate λ Poisson subordinator. Then, the generator is written as

$$Q = \lambda(\hat{P} - I), \qquad P_t = \sum_{n \geqslant 0} \frac{(\lambda t)^n e^{-\lambda t}}{n!} \hat{P}^n$$

Proof

P_t: write $p_{ij}(t)$ and condition on $N(t) = n$ Q: differentiate P_t. □

The above representation for P_t is consistent with e^{tQ} (having S finite):

$$\begin{aligned} AB = BA \implies e^{A+B} &= e^A e^B \\ \implies e^{tQ} &= e^{\lambda t(\hat{P}-I)} \\ &= e^{\lambda t \hat{P}} \underbrace{e^{\lambda t I}}_{e^{\lambda t} I \text{ since } I^n = I} \end{aligned}$$

so

$$e^{tQ} = e^{-\lambda t} \sum_{n \geqslant 0} \frac{(\lambda t)^n}{n!} \hat{P}^n = \sum_{n \geqslant 0} (\lambda t)^n \frac{e^{\lambda t}}{n!} \hat{P}^n.$$

A regular jump CTMC $X(t)$ can be represented fully by means of:

- an embedded chain $X_n := X(S_n)$
$$S_n = \sum_{i=1}^{n} T^{(i)}$$
sequence of visited states
- given $(X_n)_{n \geq 1}$
$$T^{(i)} \overset{\text{i.i.d.}}{\sim} \text{Exp}(q_{X_{i-1}})$$
independently of $T^{(i-1)}$

We are going to introduce the **uniform chains**

$$X(t) = Y_{N(t)}$$

where

- Y_n is a CTMC with transition matrix $\hat{P}$ such that $\hat{p}_{ii} \geq 0$
- $N(t) \sim \text{Poisson}(\lambda t)$

$\implies$ $Q = \lambda(\hat{P} - I)$. Now, the question is: *Can we reparameterize a generic CTMC which we assume to be regular jump with generator Q as a uniform chain (for which we can describe the semigroup)?* Hint:
$Q = \lambda(\hat{P} - I)$ generator of uniform chain. We manupulate it as follows:

$$\lambda^{-1} Q = \hat{P} - I \implies \hat{P} = I + \lambda^{-1} Q$$

Can we identify this matrix $\hat{P}$ given Q?

The answer is affirmative, under the specific constraint:

$$\sup_i q_i < \infty$$

If this condition holds (so we have a stable Q), take:

- any number
$$\nu \geq \sup_i q_i = \sup_i \sum_{j \neq i} q_{ij}$$
if the chain is conservative
- holding times
$$T^{(n)} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\nu)$$
(they are picked from the universal clock 🕒)

- jump probabilities
$$\hat{p}_{ij} = \frac{q_{ij}}{\nu}$$
that could leave us some mass when the state is left.

$$\begin{aligned}\hat{p}_{ii} &= 1 - \sum_{j \neq i} \hat{p}_{ij} = 1 - \sum_{j \neq i} \frac{q_{ij}}{\nu} \\ &= 1 - \frac{q_i}{\nu} = 1 + \frac{q_{ii}}{\nu}\end{aligned}$$

$\implies \forall i, j$
$$\hat{p}_{ii} = \delta_{ij} + \frac{q_{ij}}{\nu}$$
which is equivalent, in matrix form, to saying
$$\hat{P} = I + \nu^{-1} Q$$
as hinted above.

(We picked ν, which is a rate faster than all the other rates of the universal clock.🕒) $\implies$
$$P_t = \sum_{n \geq 0} \frac{(\nu t)^n e^{-\nu t}}{n!} \hat{P}^n$$
is the semigroup. Let's now analyze some examples:

Example 3.7

We consider a two state chain, represented in the following Figure. The generator is Q = $\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$

$$\begin{aligned} T_0^{(n)} &\overset{\text{i.i.d.}}{\sim} \text{Exp}(q_0) = \text{Exp}(\lambda) \text{ in state } 0 \\ T_1^{(n)} &\overset{\text{i.i.d.}}{\sim} \text{Exp}(q_1) = \text{Exp}(\mu) \text{ in state } 1) \end{aligned}$$

- embedded chain has transition $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Let's make it uniform:

- $\nu = \lambda + \mu \implies$
$$\hat{T}_i^{(n)} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\nu) = \text{Exp}(\lambda + \mu), \qquad i = q_1$$
- jump matrix
$$\hat{P} = I + \nu^{-1} Q$$

$$\begin{aligned} \hat{p}_{01} &= \frac{q_{01}}{\nu} = \frac{\lambda}{\lambda + \mu} & \hat{p}_{00} &= 1 - \frac{\lambda}{\lambda + \mu} = \frac{\mu}{\lambda + \mu} \\ \hat{p}_{10} &= \frac{q_{10}}{\nu} = \frac{\mu}{\lambda + \mu} & \hat{p}_{11} &= \frac{\lambda}{\lambda + \mu} \end{aligned}$$

So now, since
$$\hat{P}^n = \hat{P} \text{ (check)}$$

and the chain is uniform:

$$\begin{aligned}
p_{00}(t) &= \sum_{n\geq 0} \frac{(\nu t)^n e^{-\nu t}}{n!} \hat{p}_{00}^{(n)} \\
&= e^{-\nu t} + \sum_{n\geq 1} \frac{(\nu t)^n e^{-\nu t}}{n!} \hat{p}_{00} \\
&= e^{-\nu t} + \hat{p}_{00}(1 - e^{-\nu t}) \\
&= \hat{p}_{00} + e^{-\nu t}(1 - \hat{p}_{00}) \\
&= \frac{\mu}{\lambda + \mu} + e^{-(\lambda+\mu)t} \frac{\lambda}{\lambda + \mu}
\end{aligned}$$

4 Stationarity

Irreducibility and recurrence are inherited from the embedded chain.

Definition 4.1

π (row vector) is an **invariant measure** for X if

$$\pi P_t = \pi \quad \forall t \geq 0.$$

This is called **global balance condition**.

$$\sum_{i\in S} \pi_i p_{ij}(t) = \pi_j \quad \forall i, j$$

If P_t is unknown, the above condition is impossible to be checked. Hence, we need a new criterion for Q.

Proposition 4.1

Let X be irreducible and recurrent. Then, there exists an invariant measure π, which is unique up to a multiplicative constant. Furthermore, we have that

$$\pi P_t = \pi \quad \forall t \geq 0 \iff \pi Q = \underline{0}$$

where $\underline{0} = (0, \ldots, 0)$

Proof

S finite:

$$\begin{aligned}
\frac{d}{dt}(\pi P_t)_j &= \frac{d}{dt} \sum_i \pi_i p_{ij}(t) \\
&= \sum_i \pi_i p'_{ij}(t) \overset{\text{by KBE}}{=} \sum_i \pi_i \sum_k q_{ik} p_{kj}(t) \\
&= \sum_k \underbrace{\left(\sum_i \pi_i q_{ik}\right)}_{(\pi Q)_k} p_{kj}(t)
\end{aligned}$$

Since X is irreducible,

$$p_{kj}(t) > 0 \quad \forall k, j$$

and therefore

$$(\pi Q)_k = 0 \quad \forall k$$

if and only if the R.H.S. is null
if and only if

$$(\pi P_t)_j$$

is constant with respect to t, in which case

$$(\pi P_t)_j = (\pi P_0)_j = \sum_i \pi_i \delta_{ij} = \pi_j$$

which is global balance. □

So $\pi Q = \underline{0} \implies \pi$ is stationary.

Example 4.1

Birth and Death(λ_i, μ_i) The criterion

$$\sum_i \pi_i q_{ij} = 0 \quad \forall j$$

reads

- $j = 0$

$$\pi_1 \mu_1 - \pi_0 \lambda_0 = 0$$

$$\pi_1 = \pi_0 \frac{\lambda_0}{\mu_1}$$

- $j \geq 1$:

$$\pi_{j-1}\lambda_{j-1} + \pi_{j+1}\mu_{j+1} - \pi_j(\lambda_j + \mu_j) = 0$$

- $j = 1$

$$\begin{aligned} \pi_2 \mu_2 &= \pi_1(\lambda_1 + \mu_1) - \pi_0 \lambda_0 \\ &= \pi_0 \frac{\lambda_0}{\mu_1}\lambda_1 + \pi_0 \frac{\lambda_0}{\mu_1}\mu_1 - \pi_0 \lambda_0 \end{aligned}$$

$$\pi_2 = \pi_0 \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2}$$

By induction, we get

$$\pi_k = \pi_0 \frac{\lambda_0 \ldots \lambda_{k-1}}{\mu_1 \ldots \mu_k}$$

if

$$C = \pi_0 \sum_{k \geq 1} \frac{\lambda_0 \ldots \lambda_{k-1}}{\mu_1 \ldots \mu_k} + \pi_0 < \infty$$

We can normalize to get the stationary distribution π.

· For example,

$$\lambda_i = \lambda \quad \mu_i = \mu, \lambda < \mu$$

$$C = \pi_0 \sum_{k \geq 0} (\frac{\lambda}{\mu})^n \qquad \rho = \frac{\lambda}{\mu}$$
$$= \pi_0 (1-\rho)^{-1}$$

$\implies$

$$\pi_k = (1-\rho)\rho^k$$

(compare DTMC)

- Consider

$$\lambda_i = \lambda, \quad \mu_i = \mu \cdot i$$

$$\pi_k \propto \frac{\lambda_0 \dots \lambda_{k-1}}{\mu_1 \dots \mu_k} = \frac{\lambda \dots \lambda}{\mu(2\mu)\dots(k\mu)} = \frac{(\frac{\lambda}{\mu})^k}{k!}$$

$\implies$

$$\pi_k = \text{Poisson}(k, \frac{\lambda}{\mu})$$

Can we make this uniform?

$$\sup_i q_i = \sup_i (\lambda + \mu_i) = \infty$$

We can't and this suggests that we cannot identify a DTMC with a Poisson stationary.

Proposition 4.2

Consider X is irreducible and uniform with jump matrix $\hat{P}$. If

$$\pi \hat{P} = \pi$$

then π is invariant for X.

When X cannot be made uniform, what is the relationship between its stationary and its embedded chain?

Proposition 4.3

Let X have generator Q and let $\tilde{P}$ be the transition matrix of its embedded chain. Then

$$\tilde{\pi}\tilde{P} = \tilde{\pi} \iff \pi Q = \underline{0}$$

with

$$\pi_j = \frac{\tilde{\pi}_j}{Q_j}$$

Intuition: $\tilde{\pi}$ is the stationary for the embedded

- $\tilde{\pi}_j$ represents the long-run percentage of jumps to state j
- $\frac{1}{q_j}$ represents the average time spent in j (since $T_j \sim$ Exp (q_j))
- $\frac{\tilde{\pi}_j}{q_j}$ is the long run percentage of time spent in j.

Definition 4.2

We say a CTMC is **ergodic** if it is irreducible and positive recurrent.

Note that even if the embedded chain is periodic, the CTMC is not since

$$p_{ii}(t) > 0 \quad \forall t \geq 0$$

Proposition 4.4

Consider X as an ergodic regular jump CTMC. $\forall i, j \in S$

$$p_{ij}(t) \to \pi_j \quad t \to \infty$$

with π is the unique stationary distribution.

Example 4.2

Consider $S = \{0, 1\}$

$$p_{00}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t} \to \frac{\mu}{\lambda + \mu} = \pi_0$$

Theorem 4.1

Consider X ergodic with stationary distribution Pi. For any initial distribution μ and π-integrable f (real-valued)

$$\frac{1}{t}\int_0^t f(X(s))ds \xrightarrow{t\to\infty} \sum_i f(i)\pi_i$$

a.s. - P^*_μ (i.e. with respect to the probability measyre induced on the space of trajectories by P_t and initial distribution μ).

If $\tilde{\pi}$ is the stationary of the embedded chain, the convergence is to

$$\sum_i f(i)\frac{1}{\tilde{m}_i q_i}$$

If we can make the chain uniform and we know about the subortinated chain (the one about which we specify or derive the transition matrix)

$$d_{TV}(\mu P^n, \pi) \leq C\alpha^n \quad C > 0, \alpha \in (0, 1)$$

geometrically ergodic.

Then, write

$$\underbrace{\mu P_t}_{\text{marginal in } t} - \pi = \sum_{n\geq 0} \frac{(\lambda t)^n e^{-\lambda t}}{n!}(\mu P^n - \pi)$$

we find

$$\begin{aligned} d_{TV}(\mu P_t, \pi) &\leq \sum_{n\geq 0} \frac{(\lambda t)^n e^{-\lambda t}}{n!}(\mu P^n - \pi) d_{TV}(\mu P^n, \pi) \\ &\leq Ce^{-\lambda t}\sum_{n\geq 0}\frac{(\lambda t)^n e^{-\lambda t}}{n!} \\ &= Ce^{-\lambda t(1-\alpha)} \end{aligned}$$

we have found a quantitative bound for the speed of convergence of our CTMC. Brownian motion from the random walk, it's like if we are zooming out. Zooming out, the process becomes more and more a Brownian Motion oscillating in the interval $(0, 1)$.

5 Scaling limits

If you think about social network, you may want to approximate the network, otherwise computations are impossible.
We are interested in

$$X^{(N)} \xrightarrow{d} X \qquad \text{as } N \to \infty$$

where

- $$X^{(N)} \text{ is a CTMC } \{X^{(N)}(t), t \geq 0\}$$

  indexed by $N \geq 1$.
  For example N is:

  - the population size
  - the parameter in the transition probability

- $\{X^{(N)}, N \geq 1\}$ is a sequence of CTMC's.

- X is a limit process

Let X be any stochastic process on $S \subset \mathbb{R}$. Indexed by $T \subset [0, +\infty)$

Definition 5.1

We call

- **projections of** X the vectors $(X(t_1), \dots, X(t_n)) \qquad t_i \in T, n \geq 1$
- **finite-dimensional distributions of** X the lows of its projections, i.e. the family

  $$\mathcal{M} = \{Q_{t_1,\dots,t_n} : t_i \in T, n \geq 1\}$$

  such that

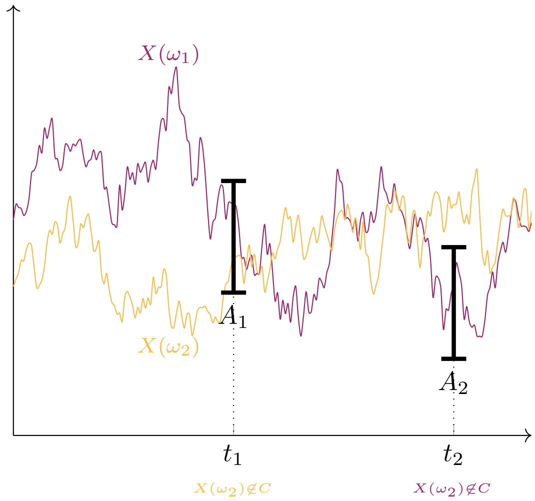
  $$Q_{t_1,\dots,t_n}(A_1 \times \cdots \times A_n) = \mathbb{P}(X(t_1) \in A_1, \dots X(t_n) \in A_n)$$

  for $A_i \in \mathcal{B}(S)$

The finite-dimensional distributions give the probability assigned to **cylinder sets**.

$$C = \{X \in \{f : [0, \infty) \to \mathbb{R}\} : X(t_1) \in A_1, \dots, X(t_n) \in A_n\}$$

And the intuition here is:



Theorem 5.1

The finite-dimensional distributions of a stochastic process satisfy two properties, called **Kolmogorov consistency conditions**.

(C1) The first condition is the following

$$Q_{t_1,\dots,t_n}(A_1 \times \cdots \times A_{n-1} \times \mathbb{R}) = Q_{t_1,\dots,t_{n-1}}(A_1 \times \cdots \times A_{n-1})$$

and it is a marginalization property.

(C2) for all permutations π of $\{1, 2, \dots, n\}$ with $\pi(i)$ the new position of i

$$Q_{t_1,\dots,t_n}(A_1 \times \dots \times A_n) = Q_{t_{\pi(1)},\dots,t_{\pi(n)}}(A_{\pi(1)} \times \dots \times A_{\pi(n)})$$

joint permutation of indexes and arguments.

If the queues are the finite-dimensional distributions of the chain; the converse is not trivial.

Theorem 5.2

Kolmogorov extension theorem: Let $\mathcal{M}$ be the family (as above) of probability measures that satisfy $C1$-$C2$. Then, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process X on such space such that the elements of $\mathcal{M}$ are the finite-dimensional distributions of X.

6 Weak convergence of C.Á.D.L.Á.G processes

Let' s consider r.v.'s $Z^{(N)} \sim \nu_n$ and $Z \sim \nu$ so

$$Z^{(N)} \xrightarrow{x} Z \text{ as } N \to \infty \iff$$
$$\nu_N \to \nu \text{ i.e. } \int f d\nu_N \to \int f d\nu$$

for $f \in \mathcal{B}(S)$.
$Z^{(N)}, Z$

- can be defined on different probability spaces $(\Omega_N, \mathcal{F}_N, \mathbb{P}_N)$
- take values on different spaces S_N
- can have different continuity structure e.g. ν_N discrete for all N and ν continuous.

When the Z are stochastic processes, there are some complications since we also need to take care of the nature of the trajectories (sample paths).

Definition 6.1

Z is a C.Á.D.L.Á.G.[a] process if its trajectory are right- continuous with left limits.

[a] continue droite, limite gauche

Example 6.1

Denote D_S the space of Cadlag functions from $[0, \infty)$ to S.
Take

- $X, X^{(N)}$ Cadlag processes with values in S, S_N respectively assuming

$$\lim_N S_N$$

is dense in S.
So these are random variables taking values in D_S and D_{S_N} respectively.

If $X^{(N)} \xrightarrow{d} X$, then this implies convergence of all finite-dimensional distributions, i.e.

$$\left[X^{(N)}_{t_1}, \dots, X^{(N)}_{t_n}\right] \xrightarrow{d} [X(t_1), \dots, X(t_n)]$$

The converse is true with additional requirements (such as tightness). Suppose we denote by C_S the space of continuous functions from $[0, \infty)$ to S. Then, it is allowed to have all $X^{(N)} \in D_S$ (all the trajectories of the entire sequence are discontinuous) and $X \in C_S$. So the limit process has continuous trajectories, but none of the other terms of the sequence, denoted as $X^{(N)}$.
We are interested in when this limit process is a diffusion process. Let $X(t)$ be Cadlag on $S \subset \mathbb{R}$. Define its increments

$$\Delta_h X(t) := X(t+h) - X(t)$$

and

$$\mathbb{E}_x[\Delta_h X(t)] := \mathbb{E}[\Delta_h X(t) | X(t) = x]$$

Definition 6.2

X is a **diffusion** if the following three conditions hold:

- $\mathbb{E}_x[\Delta_h X(t)] = \underbrace{\mu(x)}_{\text{drift of } X \text{ or infinitesimal mean}} h + o(h)$
- $\mathbb{E}_x[(\Delta_h X(t))^2] = \underbrace{\sigma^2(x)}_{\text{diffusion coefficient, infinitesimal variance}} h + o(h)$
- $\mathbb{E}_x[|\Delta_h X(t)|^p] = o(h)$ for some $p > 2$ which implies the continuity of the trajectories via Dynkin's condition.

If these hold, X is the solution of the **stochastic differential equation (SDE)**

$$dX(t) = \mu(X(t))dt + \sigma(X(t))d \underbrace{B(t)}_{\text{standard B.M.}}$$

Remark

Usually, the third condition is the most difficult to check so we are going to control only the validity of the first two.

Interpretation:
if $X(t) = x$ then

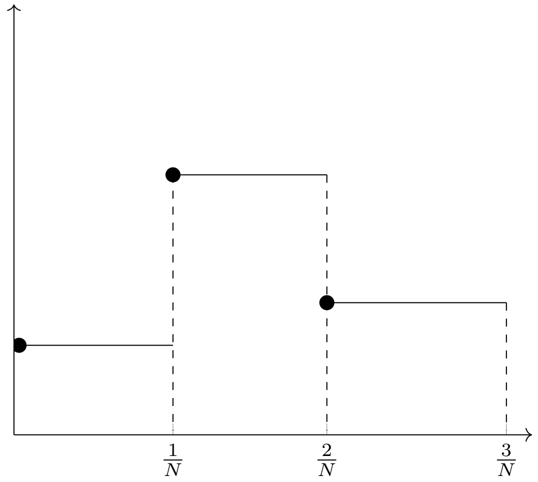
$$\Delta_h X(t) \approx \mu(x)h + \sigma(x) \underbrace{\Delta_h B(t)}_{\text{increment of SNM } \sim \mathcal{N}(0,h)}$$

This suggests a numerical way of simulating the trajectories: c.p. **Euler-Maruyama** scheme for approximating numerically on SDE. Let now $\{Y^{(N)}\}_{N \geq 1}$ be a sequence of DTMCs on S_N countable subset of S with limit dense in S.
Let $\{h_N, N \geq 1\} \subset \mathbb{R}_+$ such that $h_N \to 0$.
Define now the continuous time process

$$X^{(N)}(t) := Y^N \lfloor \frac{t}{h_N} \rfloor$$

where $\lfloor z \rfloor$ is the floor function.

$$\lfloor z \rfloor = \sup\{n \in \mathbb{N}, n \leq z\}$$



Example

$h_N = \frac{1}{N}$

$$X^{(N)}(t) = \begin{cases} Y_0^{(N)} & 0 \leq t < \frac{1}{N} \\ Y_1^{(N)} & \frac{1}{N} \leq t < \frac{2}{N} \\ \dots \end{cases}$$

a unit interval for $X^{(N)}$ corresponds to N steps of $Y^{(N)}$.

Define

$$\Delta X^{(N)}(t) = \Delta_{h_N} X^{(N)}(t) = X^{(N)}(t + h_N) - X^{(N)}(t)$$

and denote

$$\mathbb{E}_x[\cdot] := \mathbb{E}_x[\cdot | X^{(N)}(t) = x]$$

If we can show:

- $$\mathbb{E}_x[\Delta X^{(N)}(t)] = \mu(x) h_N + o(h_N)$$
$$\rightarrow \lim_{N \to 0} \frac{1}{h_N} \mathbb{E}_x[\Delta X^{(N)}(t)] = \mu(x)$$

- $$\mathbb{E}_x[(\Delta X^{(N)}(t))^2] = \sigma^2(x) h_N + o(h_N)$$

  $\implies$

$$\lim_{N \to 0} \frac{1}{h_N} \mathbb{E}_x[(\Delta X^{(N)}(t))^2] = \sigma^2(x)$$

- $$\mathbb{E}[(\Delta X^{(N)}(t))^4] = o(h_N)$$

Then with additional tecnical condition that we are not gonna consider, then we can claim that

$$X^{(N)} \overbrace{\rightarrow}^{d} X \quad \text{as } N \to \infty$$

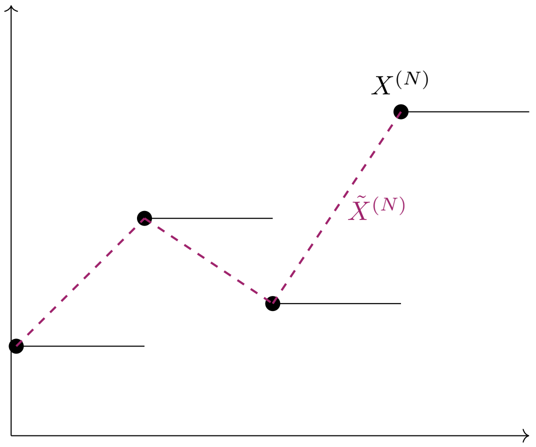
with X the solution of the SDR.

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t)$$

So X approximates $X^{(N)}$ for large N.

Comments:

- The Above $X^{(N)}$ are Cadlag and discontinuous.



We could interpolate to have $\tilde{X}^{(N)}$ continuous (not necessary). It is not even necessary to put DTMC in continuous time.

- Time rescaling:
We have used deterministic h_N intervals. Otherwise we can define a uniform chain with jump chain $Y^{(N)}$ and h_N^{-1}-rate Poisson process $\implies T^{(i)} \overset{\text{i.i.d.}}{\sim} \text{Exp}(h_N^{-1})$.

$h_N = \frac{1}{N}$ Exp (N)
$T^{(i)} \xrightarrow{\text{in mean square}} 0$

- Space rescaling: above we have implicitly assumed the states of $Y^{(N)}$ to get closer and closer as N increases. In general one way need to rescale space too e.g. For example, take

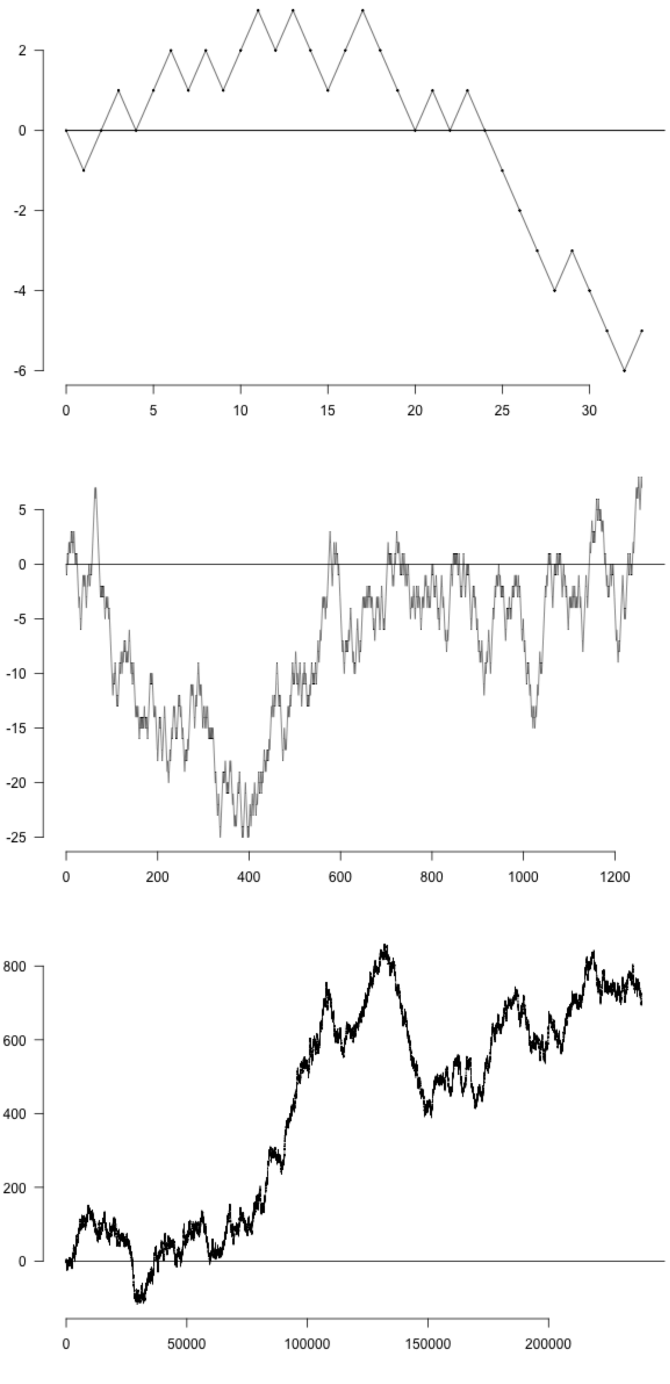
$$\frac{Z}{N} = \{0, +-\frac{1}{N} + -\frac{2}{N}, ..\}$$

So, in general we are interested in conditions for convergence of

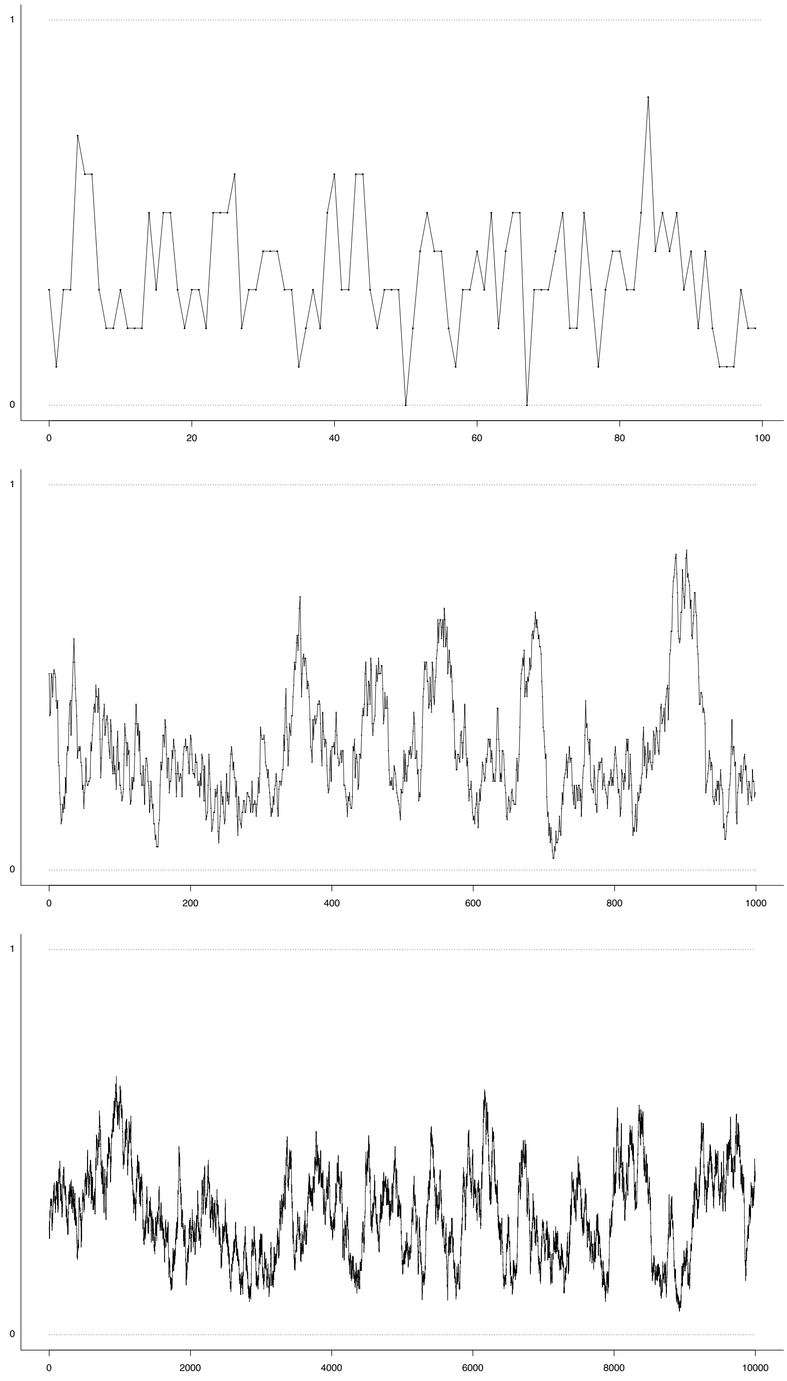
$$X^{(N)}(t) = \frac{Y^{(N)}_{\lfloor t/h_N \rfloor} - a_N}{b_N} \xrightarrow{d} X(t)$$

Here we have

- centering a_N;
- time rescaling h_N;
- space rescaling b_N.



RWs $X^{(N)}$ for three values of N.



WF chains $X^{(N)}$ for three values of N.

We are trying to analyse scaling limits of diffusions. Conditions for convergence of objects like the one we determined above:

$$X^{(N)}(t) = \frac{Y^{(N)}_{\lfloor t/h_N \rfloor} - a_N}{b_N} \xrightarrow{d} X(t)$$

where $X(t)$ is a diffusion process.

1. **Asymmetric Random Walk**: $Y^{(N)}$ a rw on $\mathbb{Z}$ with

$$p_{i,i+1} = \frac{1}{2} + \frac{\mu}{2\sqrt{N}} \quad p_{i,i-1} = \frac{1}{2} - \frac{\mu}{2\sqrt{N}}$$

for N large enough so they are in $[0.1]$.
Define

$$X^{(N)}(t) := \frac{Y^{(N)}_{\lfloor Nt \rfloor}}{\sqrt{N}} \tag{7}$$

which implies that

- we have time rescaling by $h_N = \frac{1}{N}$
- space rescaling by $b_N = \sqrt{N}$
- no centering

From item 2, we have that

$$Y^{(N)}_{\lfloor Nt \rfloor} = i \in \mathbb{Z} \qquad \implies X^{(N)}(t) = \frac{1}{\sqrt{N}}$$

which implies that

$$\Delta X^{(N)}(t) = \frac{1}{\sqrt{N}} \left(Y^{(N)}_{\lfloor Nt \rfloor + 1} - Y_{\lfloor Nt \rfloor} \right)$$

For $p \in N$, consider

$$\mathbb{E}_x[(\Delta x^{(N)}(t))^p] = \frac{1}{N^{p/2}} \left[(+1)^p (\frac{1}{2} + \frac{\mu}{2\sqrt{N}}) + (-1)^p (\frac{1}{2} - \frac{\mu}{2\sqrt{N}}) \right]$$

If $p = 1$,

$$\mathbb{E}_x[\Delta X^{(N)}(t)] = \frac{1}{N^{\frac{1}{2}}} \left[\frac{\cancel{1}}{\cancel{2}} + \frac{\mu}{2\sqrt{N}} - \frac{\cancel{1}}{\cancel{2}} + \frac{\mu}{2\sqrt{N}} \right] = \underbrace{\frac{1}{N}}_{h_N} \underbrace{\mu}_{\mu(x) \equiv \mu \in \mathbb{R}}$$

$$(\mathbb{E}_x(\Delta X^{(N)}(t)) = \mu(x) h + \sigma^2)$$

If $p = 2$,

$$\mathbb{E}_x[(\Delta X^{(N)}(t))^2] = \frac{1}{N^{\frac{1}{2}}} \left[\frac{1}{2} + \frac{\mu}{2\sqrt{N}} + \frac{1}{2} - \frac{\mu}{2\sqrt{N}} \right] = \underbrace{\frac{1}{N}}_{h_N}$$

that implies $\sigma^2(x) \equiv 1$.
If $p = 4$

$$\mathbb{E}_x((\Delta X^{(N)}(t))^4) = \frac{1}{N^2} = o(hN)$$

Then,

$$X^{(N)} \xrightarrow{d} X$$

where X solves

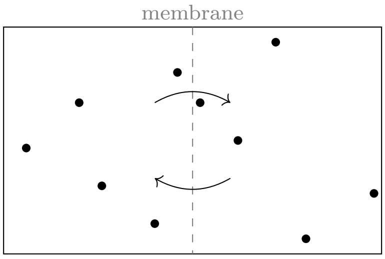
$$dX(t) = \mu dt + dB(t), \quad X(t) \in \mathbb{R}$$

2. **Ehrenfest urn**[2]

We have a total of $2N$ balls separated by a membrane into an urn. 1 ball is selected at random and moved to the other space. Let

$Y^{(N)}$ be the number of balls in the first space
$$S = \{0, \dots, 2N\}$$

$$p_{i,i-1} = \frac{i}{2N} = \frac{1}{2} - \frac{i}{2N} \qquad p_{i,i+1} = 1 - \frac{i}{2N} = \frac{1}{2} + \frac{i}{2N}$$



The difference with respect to the previous example is that now the process is a spatially inhomogeneous (but time homogeneuos) RW on S finite.
Define

$$X^{(N)}(t) := \frac{Y^{(N)}_{\lfloor Nt \rfloor} - N}{\sqrt{N}}$$

which implies

$$Y^{(N)} = i \quad \implies \quad i = x\sqrt{N} + N$$

when $X^{(N)}(t) = x$.
Let's rewrite the probabilities in terms of the x:

$$\begin{aligned} p_x(\Delta X^{(N)}(t) = \pm \frac{1}{\sqrt{N}}) &= \mathbb{P}(\Delta Y^{(N)}_{\lfloor t/h_N \rfloor} = \pm 1 | i = x\sqrt{N} + N) \\ &= \frac{1}{2} \pm \frac{N - (x\sqrt{N} + N}{2N} \\ &= \frac{1}{2} \pm \frac{x}{2\sqrt{N}} \end{aligned}$$

spatial inhomogeneity.
As we can see, we got a result very similar to the previous one with x in place of μ.

$$\begin{aligned} \mathbb{E}(\Delta X^{(N)}(t)) &= \frac{1}{\sqrt{N}}(\frac{1}{2} - \frac{x}{2\sqrt{N}}) - \frac{1}{\sqrt{N}}(\frac{1}{2} + \frac{x}{2\sqrt{N}}) \\ &= -\frac{x}{2N} - \frac{x}{2N} = -\frac{x}{N} \qquad (h_N = \frac{1}{N};\ \mu(x) = -x) \end{aligned}$$

$$\mathbb{E}_x[(\Delta X^{(N)}(t))^2] = \frac{1}{N}(\frac{1}{2} - \cancel{\frac{x}{2\sqrt{N}}}) + \frac{1}{N}(\frac{1}{2} + \cancel{\frac{x}{2\sqrt{N}}}) = \frac{1}{N}$$

So, again, $\sigma^2(x) \equiv 1$. It is immediate to verify that the moment of order 4:

$$\mathbb{E}_x[(\Delta X^{(N)}(t))^4] = O(\frac{1}{N^2}) = o(\frac{1}{N}) = o(h_N)$$

which implies that

$$X^{(N)} \xrightarrow{d} X$$

such that

$$dX(t) = -X(t)dt + dB(t), \quad X(t) \in \mathbb{R}$$

called **Ornstein-Uhlenbeck diffusion**, stationary with respect to

$$N(0, \frac{1}{2})$$

It has applications in math finance and in biology. Mean reversion: when it is the positive it pushes back to zero and when it is negative the process is pushed up to return to the 0 that is the long mean of the process. This is particular because is a process that is stationary (BM is not stationary and this property is useful in many application mathfinance, bioology).
We sent by N, we rescale and we determine a transformation of the previous one. We plot

[2] He thinks it is a model for gasses.

it and we find its Gaussian distribution.
It can be seen as a continuous-time analogue of AR(1).
Discretize over Δt interval to get

$$\underbrace{\Delta X_k}_{X_{k+1}-X_k} = -X_k \Delta t + \sqrt{\Delta t} \cdot \varepsilon_k$$

where

$$\varepsilon_k \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{2})$$

which implies that

$$\begin{aligned} \mathbb{E}(\Delta X_k)) &= -X_k \Delta t \\ \mathcal{V}ar(\Delta X_k) &= \frac{\Delta t}{2} \end{aligned}$$

and we write

$$X_{k+1} = X_k - X_k \Delta t + \sqrt{\Delta t}\epsilon_k = \underbrace{1 - \Delta t}_{a} X_k + \underbrace{\sqrt{\Delta t}}_{b} \epsilon_k$$

3. **Branching Processes** Consider

$$Y_n^{(N)} \text{ a BP } Y_n^{(N)} = \sum_{i=1}^{Y_{n-1}^{(N)}} Z_i^{(N)}$$

where n refers to the step, while N the parameterisation and
$Z_i^{(N)} \overset{i.i.d.}{\sim}$ with mean $\mu^{(N)} \in \mathbb{R}$ and variance $\sigma^2 > 0$.

$$\mathbb{E}[Y_n^{(N)} | Y_{n-1}^{(N)} = y] = \mu^{(N)} y \qquad \Longrightarrow \mathbb{E}_y[\Delta Y_n^{(N)}] = \mu^{(N)} y - y = y(\mu^{(N)} - 1)$$

Now observe:

$$\text{Var}_y(Y_n^{(N)}) = \sigma^2 y$$

Define

$$X^{(N)} := \frac{Y^{(N)\lfloor Nt \rfloor}}{N}$$

which implies that

$$X^{(N)}(t) = x \implies y = Nx$$

$$\mathbb{E}_x[\Delta X^{(N)}(t)] = \frac{1}{N}\mathbb{E}_y[\Delta Y_{\lfloor Nt \rfloor}^{(N)}] = \frac{1}{N} y(\mu^{(N)} - 1)$$

$$\begin{aligned} \mathbb{E}_x[(\Delta X^{(N)}(t))^2] &= \text{Var}_x(\Delta X^{(N)}(t)) + (\mathbb{E}_x[\Delta X^{(N)}(t)])^2 \\ &= \frac{1}{N^2} \underbrace{\text{Var}_y(\Delta Y_n)}_{\text{Var}_y(Y_{n+1})+0} + \frac{1}{N^2} Y^2 (\mu^{(N)} - 1)^2 \\ &= \frac{1}{N^2}\sigma^2 y + \frac{1}{N^2} y^2 (\mu^{(N)} - 1)^2 \end{aligned}$$

if we let $\mu^{(N)}$ be a perturbation of 1, i.e. $\mu^{(N)} = 1\frac{m}{N}$ so that $\mu^{(N)} - 1 = \frac{m}{N}$, then

$$\mathbb{E}_x[\Delta X] = \frac{1}{N} y \frac{m}{N} = \frac{my}{N^2} = \frac{mx}{N}$$

since $y = Nx$

$$\mathbb{E}_x[(\Delta X)^2] = \frac{\sigma^2 x}{N} + \underbrace{\frac{1}{N^2} N^2 x^2 \frac{m^2}{N^2}}_{o(\frac{1}{N})}$$

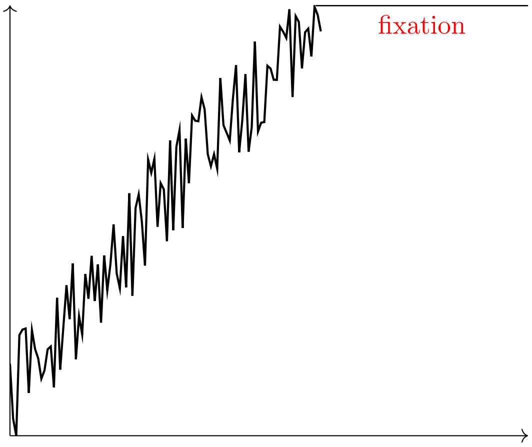
$h_N = \frac{1}{N}$, $\mu(x) = mx$ and $\sigma^2 = x$.

$$\implies \qquad X^{(N)} \xrightarrow{d} X \qquad \text{s.t.}$$

such that

$$dX(t) = mX(t)dt + \sigma\sqrt{X(t)}dB(t)$$

This process is called



- **Cox-Ingersoll-Ross diffusion** in math finance
- **Continuous state Branching Process** in math biology

So, if we touch 0 we are going to stay there forever. This is the situation we gained talking about Markov Chains and extinction of a certain population: once reached, the population stays extinguished.
0 is an *absorbing state*.
This reasoning has also a discretised counterpart.

4. **Wright-Fisher processes**
The transitions

$$Y_{n+1}^{(N)}|Y_n = i \sim \text{Binom}(N, \frac{i}{N})$$

are space inhomogeneous and indexed by N.

$$X^{(N)}(t) := \frac{Y_{\lfloor Nt \rfloor}^{(N)}}{N}$$

the percentage of type 0.
For brevity of notation,

$$Z_x := (X^{(N)}(t + \frac{1}{N})|X^{(N)}(t) = \underbrace{x}_{i/N}) \sim \frac{1}{N}\text{Binom}(N, x)$$

Let's now comput the first two moments, starting with the first one:

$$\mathbb{E}_x[\Delta X^{(N)}(t)] = \mathbb{E}(Zx) - x = \frac{1}{N}Nx = 0$$

and moving on to the second moment:

$$\begin{aligned}\mathbb{E}_x[(\Delta X^{(N)}(t)^2] &= \text{Var}_x[\underbrace{\Delta X(t)}_{z_x - x}] + (\underbrace{\mathbb{E}_x[\Delta X(t)]}_{0})^2 \\ &= \frac{1}{N^2}Nx(1-x) = \underbrace{\frac{1}{N}}_{h_N}\underbrace{x(1-x)}_{\sigma^2(x)}\end{aligned}$$

So

$$X^{(N)} \xrightarrow{d} X$$

such that

$$dX(t) = \sqrt{X(t)(1 - X(t))}dB(t)$$

If all offspring are of type 0, we are going to have always type 1. Here it is the same way of thinking.
Add mutations:

$$Y_{n+1}|Y_n = i \sim Binom(N, p_i)$$

where $p_i = \alpha(1-\frac{i}{N}) + (1-\beta)\frac{i}{N}$ and, in turn,

$$\alpha = \mathbb{P}(1 \to 0)$$
$$\beta = \mathbb{P}(0 \to 1)$$

$$\tilde{Z}_x := (X^{(N)}(t+\frac{1}{N})|X^N(t) = x) \sim \frac{1}{N} Binom(N, p_x)$$

where $x = \frac{i}{N}$ and $p_x = \alpha(x) + (1-\beta)x$.
If we compute the increment of X, we have

$$\mathbb{E}_x[\Delta X(t)] = \mathbb{E}_x[\tilde{Z}_x] = x = \frac{1}{N} Np_x - x = \ldots = \frac{1}{N}\underbrace{[\alpha(1-x) - \beta x]}_{\mu(x)}$$

Exercise 6.1

As exercise, show that

$$\mathbb{E}_x[(\Delta X(t))^2] = \cdots = \frac{1}{N}\underbrace{x(1-x)}_{\sigma^2(x)} + o(\frac{1}{N})$$

which implies that

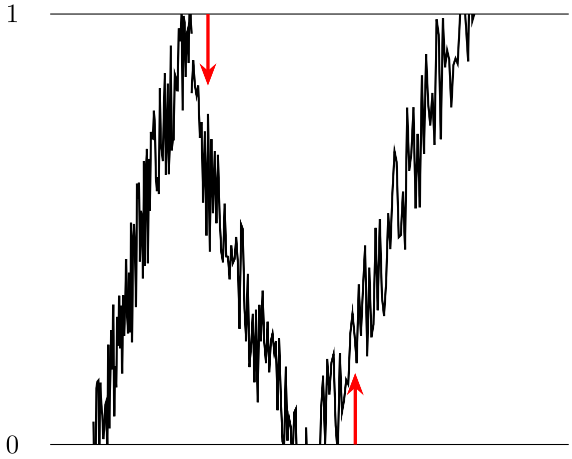
$$X^{(N)} \xrightarrow{d} X$$

such that

$$dX(t) = [2(1-X(t)) - \beta X(t)]dt + \sqrt{X(t)(1-X(t))}dB(t) = 0$$

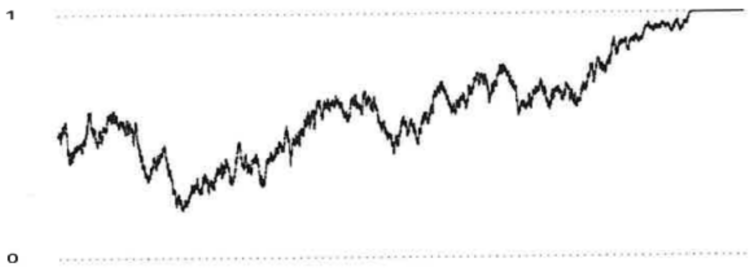
if $X(t) = 0, 1$.

- at $X(t) = 0$ we have $dX(t) = \alpha dt$
- at $X(t) = 1$ we have $dX(t) = -\beta dt$

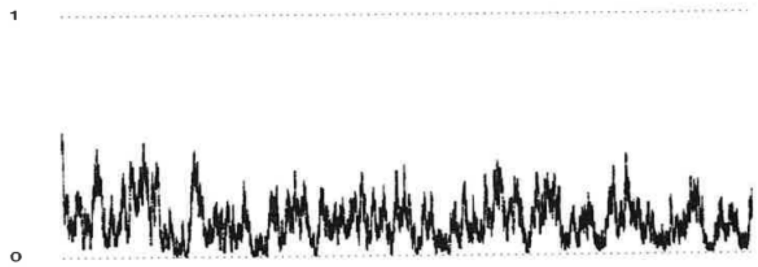


An application of the scaling limits is suggested by what we just wrote (?).

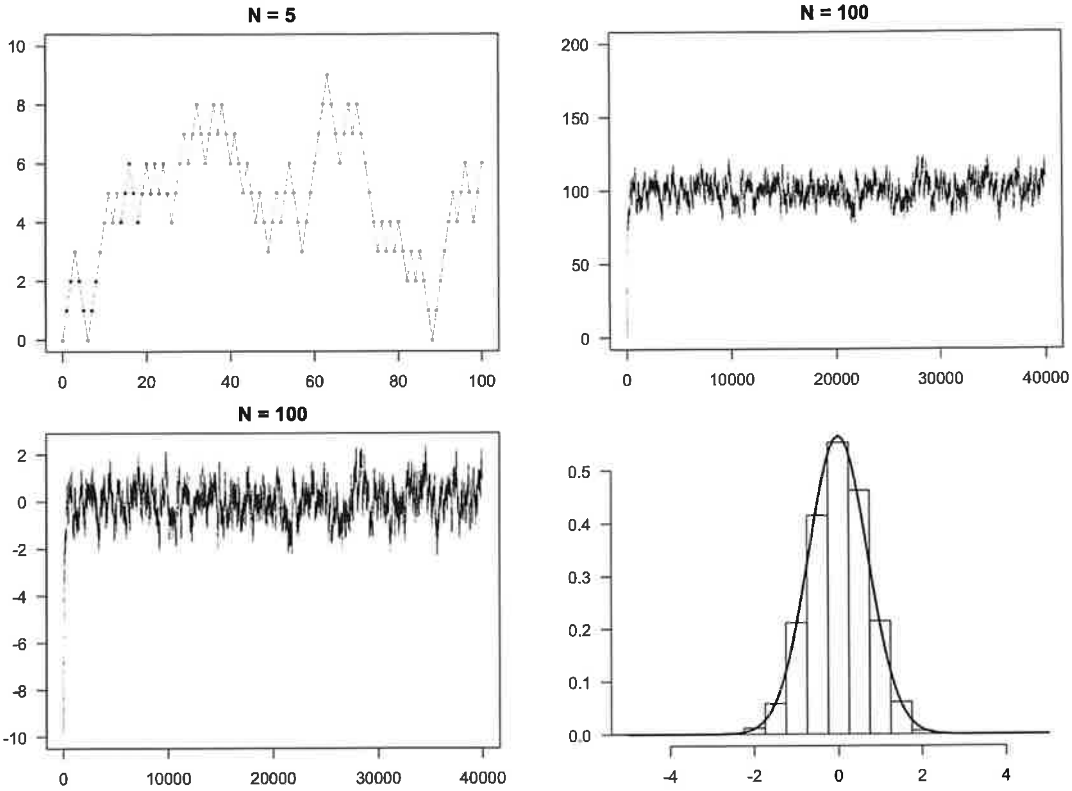
- we can avoid the computation of the conditional probability, which is very tricky.
- if the variance is small we cannot conclude that much. If it is too big, the process moves too much. So we study the limit diffusions under the problematic of rescaling.



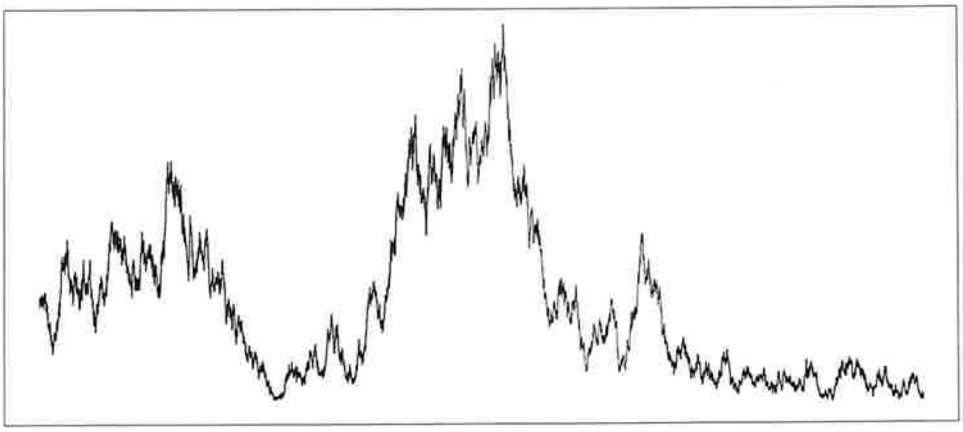
Sample path of a WF diffusion *without* mutation, exhibiting fixation at 1.



Sample path of a WF diffusion *with* mutation, with $a = 1$ and $b = 6$.

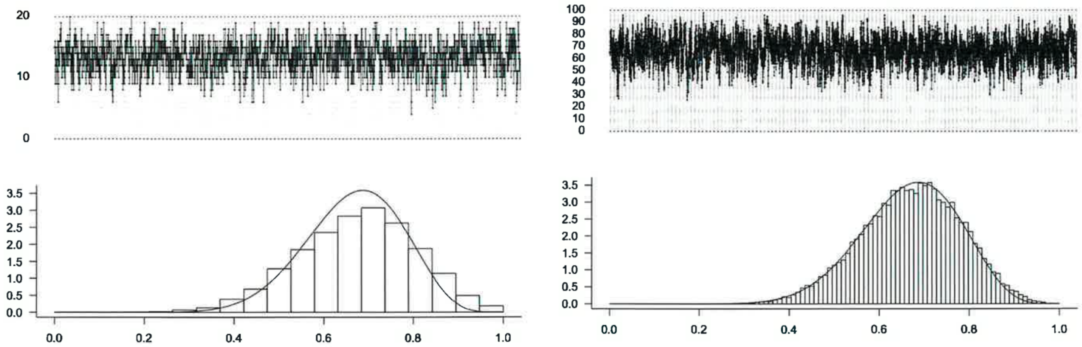


Top: two paths of the Ehrenfest urn for $N = 5, 100$.
Bottom: centered and rescaled process (left); ergodic frequencies vs. $N(0, 1/2)$ (right).

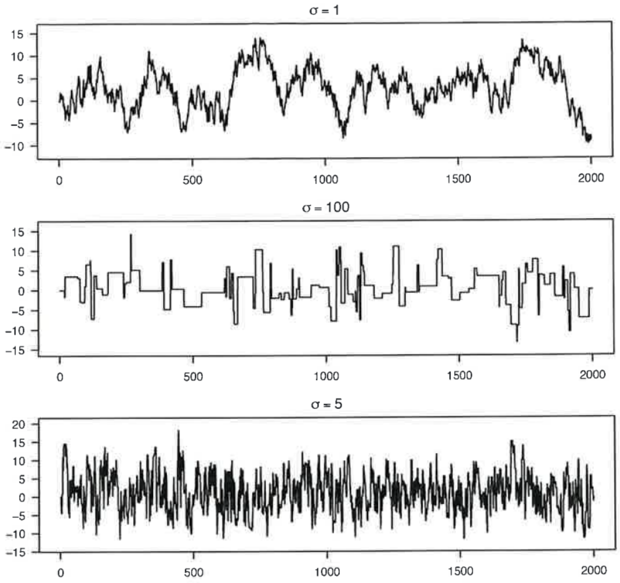


Sample path of a CIR diffusion on $\mathbb{R}_+$.

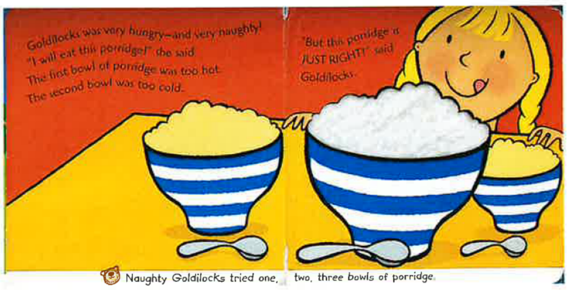
A possible application of these concepts is in the approximation of stationary distributions:



Another application may be the tuning of the variance in a Random Walk Metropolis-Hastings algorithm (trough the "Goldilocks" method) using scaling limits.

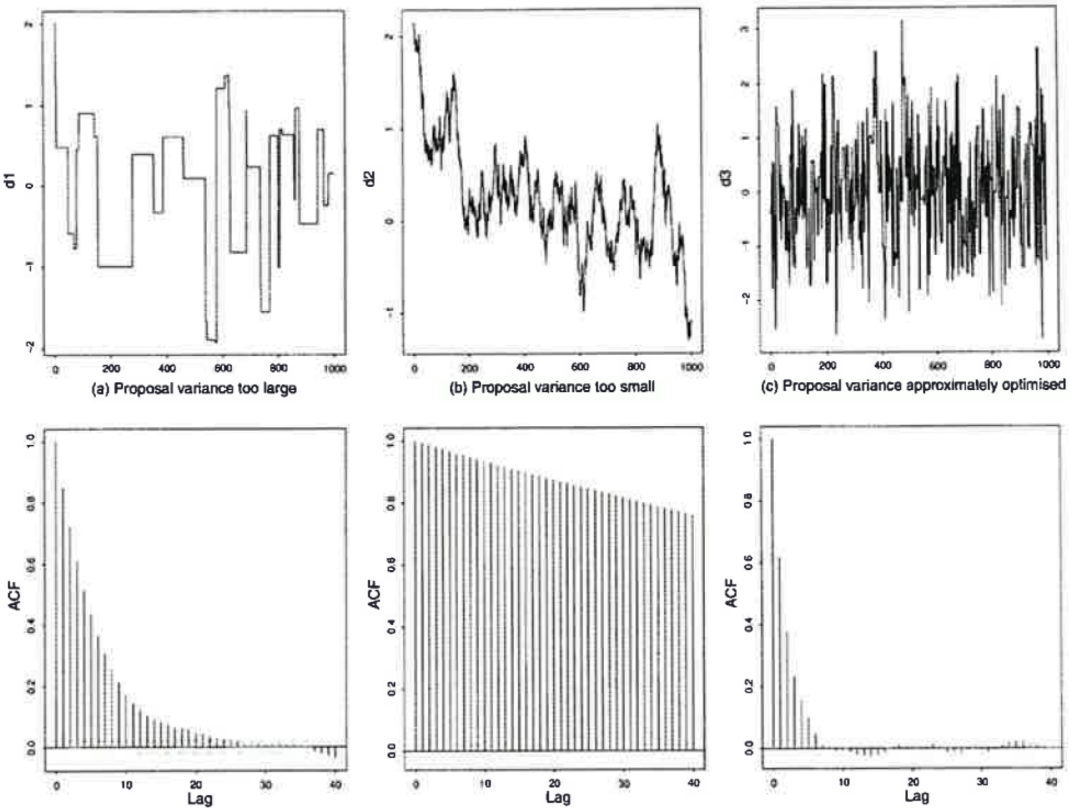


If the proposal variance is *too small*, the space is not explored efficiently (top); if the proposal variance is *too large*, the algorithm gets stuck in the same state for long periods (middle); if the proposal variance is *just right*, the space is explored efficiently. The right tuning can be investigated through the scaling limit of the MCMC (more at the end of the course).


Goldilocks was very hungry–and very naughty!
"I will eat this porridge!" she said.
The first bowl of porridge was too hot.
The second bowl was too cold.
"But this porridge is JUST RIGHT!" said Goldilocks.
Naughty Goldilocks tried one, two, three bowls of porridge.

This has been called the *Goldilocks principle* for RW-MH, terminology which recalls the famous fairy tale for kids.

We can also study different rescalings of a class of Random Walk Metropolis-Hastings algorithms and how to tune the variance of the proposal distribution accordingly:



(a) Proposal variance too large

(b) Proposal variance too small

(c) Proposal variance approximately optimised

Paths of a RW-MH algorithm for different choices of the scaling σ. Figure from ROBERTS, G.O. and ROSENTHAL, J.S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.*, **16**, 351–367.

7 Lévy processes

Lévy processes were introduced in the 1930's by Paul Lévy and Bruno De Finetti, one of the fathers of Bayesian Statistics. These processes have important applications to mathematical finance, mathematical biology and nonparametric Bayesian inference.

Definition 7.1

A random variable Y is said to be **infinitely divisible** (I.D.) if $\forall n \in \mathbb{N}$ there are $Y_i^{(n)}$ *i.i.d.* such that

$$Y \overset{d}{=} Y_1^{(n)} + \ldots + Y_n^{(n)}$$

i.e. Y can be rewritten as sum of an arbitrary number n of *i.i.d.* random of variables.

Example 7.1

- if

$$Y \sim \text{Pois}(\lambda) \qquad : \qquad Y_i^{(n)} \overset{i.i.d.}{\sim} \text{Pois}\left(\frac{\lambda}{n}\right)$$

by properties of the Poisson distribution, we know that

$$\sum_{i=1}^{n} Y_i^{(n)} \sim \text{Pois}\left(\sum_{i=1}^{n} \frac{\lambda}{n}\right) = \text{Pois}(\lambda);$$

- If

$$Y \sim \mathbb{N}(m, s^2) \qquad : \qquad Y_i^{(n)} \overset{i.i.d.}{\sim} N\left(\frac{m}{n}, \frac{s^2}{n}\right)$$

  then $\sum_{i=1}^{n} Y_i^{(n)} \sim N(n, s^2)$;

- if

$$Y \sim \Gamma(\alpha, \beta) \qquad : \qquad Y_i^{(n)} \overset{i.i.d.}{\sim} \Gamma\left(\frac{\alpha}{n}, \beta\right)$$

  then $\sum_{i=1}^{n} Y_i^{(n)} \sim \Gamma(\alpha, \beta)$.

A single way to establish I.D. is using $\mathbb{E}\left[\exp^{iuk}\right]$. Take, for example, $Y \sim \text{Pois}(\lambda)$.

$$\begin{aligned}
\mathbb{E}\left[e^{iuY}\right] &= \sum_{k \geqslant 0} \exp^{iuk} \lambda^k \frac{e^{-\lambda}}{k!} \\
&= e^{-\lambda} \sum_{k \geqslant 0} \frac{\left(\lambda \exp^{iu}\right)^k}{k!} \\
&= e^{-\lambda} e^{\lambda e^{iu}} \\
&= e^{-\lambda(1-e^{iu})}
\end{aligned}$$

So $Y_i^{(n)} \overset{i.i.d.}{\sim} \text{Pois}(\frac{\lambda}{n})$. Consider now

$$\begin{aligned}
\mathbb{E}\left[\exp^{iu \sum_i^n Y_i^{(n)}}\right] &= \left[\exp^{-\frac{\lambda}{n}(1-\exp^{iu})}\right]^n \qquad \text{using } i.i.d. \text{ properties} \\
&= e^{-\lambda(1-e^{iu})}.
\end{aligned}$$

This suggests a strategy: we could define the **characteristic exponent** (c.e.) function of Y to be:

$$\psi(u) = -\log \mathbb{E}\left[e^{iuY}\right]$$

Then Y is the I.D. if $\exists$ a random variable with c.e. $\psi^{(n)}$ such that:

$$\psi(u) = n\psi^{(n)}(u).$$

Exercise 7.1

Check this fact for $N(m, s^2)$.

Example 7.2

Consider $Y \sim \Gamma(\alpha, \beta)$. We have

$$\mathbb{E}\left[(\right] e^{iuY}) = \frac{1}{\left(1 - \frac{iu}{\beta}\right)^{\alpha}} = \left(\frac{1}{\left(1 - \frac{iu}{\beta}\right)^{\frac{\alpha}{n}}}\right)^n$$

which implies that

$$Y_i^{(n)} \overset{i.i.d.}{\sim} \Gamma(\frac{\alpha}{n}, \beta)$$

Definition 7.2

Let N $\sim$Poiss(λ) and $Z_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}$ with μ_0 mean at 0. The random variable

$$Y \sum_{i=1}^{N} Z_i \qquad (Y = 0 if N = 0)$$

is said to have **compound Poisson distribution**.

Its characteristic function is:

$$\begin{aligned}
\mathbb{E}\Big[e^{iu\sum_{i=1}^{N} Z_i}\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[e^{iu\sum_{i=1}^{k} Z_i} | N = k\Big]\Big] \\
&= \sum_{k\geqslant 0} \frac{\lambda^k e^{-\lambda}}{k!} \mathbb{E}\Big[e^{iu\sum_{i=1}^{k} Z_i}\Big] \\
&= \sum_{k\geqslant 0} \frac{\lambda^k e^{-\lambda}}{k!} \Big[\underbrace{\mathbb{E}\Big[e^{iuZ_1}\Big]}_{\int_{\mathbb{R}} e^{iux} F(\mathrm{d}x)} \Big]^k \\
&= \exp^{-\lambda} e^{\lambda \int_{\mathbb{R}} e^{iux} F(\mathrm{d}x)} \\
&= \exp^{-\lambda \int_{\mathbb{R}} (1-e^{iux}) F(\mathrm{d}x)} .
\end{aligned}$$

Now,

$$\begin{aligned}
\psi(u) &= \lambda \int_{\mathbb{R}} (1 - e^{iux} F(\mathrm{d}x)) \cdot \frac{n}{n} \\
&= n \underbrace{\frac{\lambda}{n} \int_{\mathbb{R}} (1 - e^{iux}) F(\mathrm{d}x)}_{\psi^{(n)}(u)} .
\end{aligned}$$

So Y is I.D. such that

$$Y = \sum_{i=1}^{N} Z_i \overset{d}{=} \sum_{j=1}^{n} Y_j^{(n)}$$

and we have

$$Y_j^{(n)} = \sum_{i=1}^{N_j^{(n)}} Z - i, \qquad Z_i \overset{i.i.d}{\sim} F \text{ and } N_j^{(n)} \overset{i.i.d.}{\sim} \text{Poiss}\left(\frac{\lambda}{n}\right)$$

Which is still a compounded Poisson with Poisson rate/mean $\frac{\lambda}{n}$.

Theorem 7.1

Lévy-Khintchine formula. A random variable on $\mathbb{R}$ with c.e. $\psi(u)$ is I.D. if and only if there exists:

- $\mu \in \mathbb{R}$;
- $\sigma \geqslant 0$;
- a measure π on $\mathbb{R} \setminus \{0\}$ satisfying:

  $$\int_{\mathbb{R}} \underbrace{(1 \wedge x^2)}_{\min\{1,x^2\}} \pi(\mathrm{d}x) < \infty$$

  such that, $\forall u \in \mathbb{R}$,

  $$\psi(u) = i\mu u + \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}} \left(1 - e^{iux} + iux\mathbb{1}_{(|x|<1)}\right) \pi(\mathrm{d}x).$$

This characterizes I.D. distributions:

- π is called **Lévy intensity**;
- $(\mu, \sigma, \pi) = (-m, s, 0)$ is called **characteristic triplet**.

Example 7.3

- $(mu, \sigma, \pi) = (-m, s, 0)$, i.e. $\pi \equiv 0$:

$$\psi(u) = -imu + \frac{1}{2}s^2u^2$$

  which is the c.e. of $N(m, s^2)$.

- $(\mu, \sigma, \pi) = (0, 0, \lambda\delta_1)$:

$$\begin{aligned}\psi(u) &= \lambda \int_{\mathbb{R}} (1 - e^{iux} + iux\,\underbrace{\mathbb{1}_{(|x|<1)}}_{=0})\delta_1(\mathrm{d}x) \\ &= \lambda \int_{\mathbb{R}} (1 - e^{iux})\delta_1(\mathrm{d}x) \\ &= \lambda(1 - e^{iu})\end{aligned}$$

  which means that $Y \sim \text{Pois}(\lambda)$.

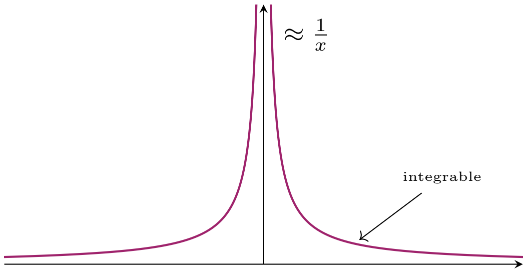
The requirement

$$\int_{\mathbb{R}} \min\{1, x^2\}\pi(\mathrm{d}x) < \infty$$

implies:

- $\int_{|x|\geqslant 1} \pi(\mathrm{d}x) < \infty$: the mean must be finite in both tails of π;
- $\int_{|x|\geqslant 1} x^2\pi(\mathrm{d}x) < \infty \implies$ we can have $\pi\left((-1, 1)\right) = \infty$ as long as $x^2\pi(\mathrm{d}x)$ is integrable around 0.

For instance, $\pi(\mathrm{d}x) \approx \frac{1}{x}\,\mathrm{d}x$ around 0 is allowed.



Definition 7.3

A Lévy process on $\mathbb{R}$ is a continuous-time Cdlg process $\{X(t), t \geqslant 0\}$ such that:

- $X(0) = 0$ a.s.;
- $X(s + t) - X(s) \overset{d}{=} X(u + t) - X(u) \qquad \forall s, u, t \geqslant 0$: it has stationary increments;
- $X(s + t) - X(s) \perp\!\!\!\perp \{X(u), u \leqslant s\}$: it has independent increments.

Example 7.4

Poisson process:

- $X(0) = 0$ a.s.;
- $\underbrace{X(s+t) - X(s)}_{\perp\!\!\!\perp s, X(s)} \sim \text{Pois}(\lambda)$.

Brownian motion:

- $X(0) = 0$ a.s.;
- $X(s+t) - X(s) \sim N(0,t) \perp\!\!\!\perp s, X(s)$.

Definition 7.4

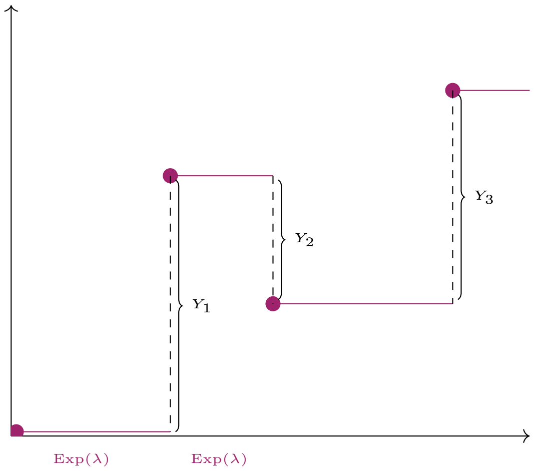
Let $N(t)$ be a rate λ Poisson process and let $Y_i \overset{i.i.d.}{\sim} F$, independent of $N(t)$. Then

$$X(t) := \sum_{i=1}^{N(t)} Y_i \qquad t \geqslant 0$$

is called **compound Poisson process**.

Exercise 7.2

Show that a CPP is also a Lévy process and when $F = \delta_1$ it is a normal Poisson process.

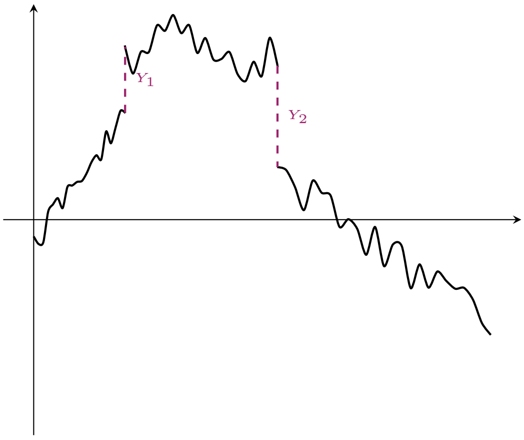


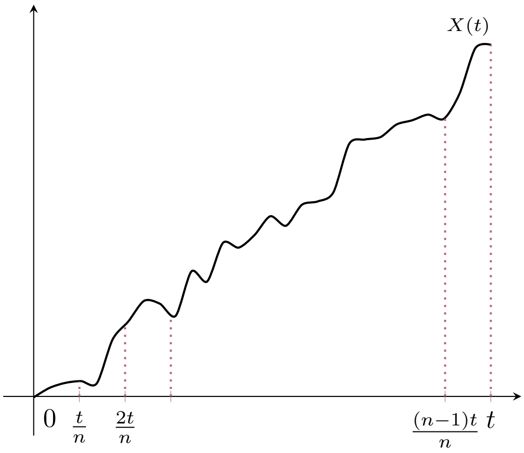
Exercise 7.3

Show that a finite sum of Lévy processes is a Lévy process.

Example 7.5

$$X(t) = \underbrace{X^{(1)}(t)}_{CP} + \underbrace{X^{(2)}(t)}_{BM}$$





Let $X(t)$ be a Lévy process and write:

$$X(t) = \underbrace{X\left(\frac{t}{n}\right) - X(0)}_{Y_1^{(n)}} + \underbrace{X\left(\frac{2t}{n}\right) - X\left(\frac{t}{n}\right)}_{Y_2^{(n)}} + \ldots + \underbrace{X\left(t\right) - X\left(\frac{(n-1)t}{n}\right)}_{Y_n^{(n)}}$$

$$\implies Y_i^{(n)} \overset{i.i.d.}{\sim} \quad \text{from indepedence and stationarity of increments}$$

$$= \sum_{i=1}^{n} Y_i^{(n)}$$

So $X(t)$ is I.D..

Define

$$\psi_t(u) := \log \mathbb{E}\left[e^{iuX(t)}\right].$$

as the c.e. of $X(t)$, from the sum:

$$\text{if } t = m \in \mathbb{N} \qquad \begin{cases} \psi_m = n\psi_{\frac{m}{n}} & n \in \mathbb{N} \\ \psi_m = m\psi_1 & n = m \end{cases} \implies \psi_{\frac{m}{n}} = \frac{m}{n}\psi_1$$

that is, $\forall t$ rational

$$\psi_t(u) = t\psi_1(u)$$

can be extended to $t \geqslant 0$.

So

$$\forall t \geqslant 0, \quad \psi_t(u) = t\psi_1(u)$$

So any Lévy process is such that

$$\mathbb{E}[e^{iuX(t)}] = e^{-t\psi_1(u)}$$

and so it is characterised by its characteristic exponent at $t = 1$; therefore, the Lévy-Khintchine formula provides a characterisation of Lévy processes.

Example: Linear Brownian Motion

$$dX(t) = mdt + sdB(t)$$

$\implies$

$$X(t) \sim N(mt, s^2t)$$

$\implies$

$$\begin{aligned}\mathbb{E}[e^{iuX(t)}] &= e^{i(mt)u-\frac{1}{2}(s^2t)u^2} \\ &= \exp\left\{\underbrace{-t\underbrace{\left[-imu + \frac{1}{2}s^2u^2\right]}_{\psi_1(u)}}_{t\psi_1(u)}\right\}\end{aligned}$$

So at time $t = 1$, we find the triplet

$$(\mu, \sigma, \pi) = (-m, s, 0)$$

This suggests:

- μ describes a constant drift with slope $-\mu$
- σ describes a Brownian component with diffusion coefficient σ^2.

Example 7.6

CP:

$$X(t) = \sum_{i=1}^{N(t)} Z_i$$

with

$$N(t) \sim Po(\lambda t)$$

and

$$Z_i \overset{i.i.d.}{\sim} F$$

$$\psi_t(u) = t\psi_1(u)$$

implies that at time $t = 1$ we get a CP R.V.

$$X(1) = \sum_{i=1}^{N(1)} Z_i$$

where $N(1) \sim Po(\lambda)$. This implies that

$$\psi_1(u) = \lambda \int_{\mathbb{R}} (1 - e^{iux})F(\mathrm{d}x)$$

We compute

$$\begin{aligned}\psi_1(u) &= \lambda \int_{\mathbb{R}} \left(1 - e^{iux} + iux\mathbb{1}_{(|x|<1)} - iux\mathbb{1}_{(|x|<1)}\right) F(\mathrm{d}x) \\ &= \int_{\mathbb{R}} \left(1 - e^{iux} + iux\mathbb{1}_{(|x|<1)}\right) \underbrace{\lambda F(\mathrm{d}x)}_{\pi(\mathrm{d}x)} - iu \underbrace{\lambda \int_{-1}^{1} xF(\mathrm{d}x)}_{-\mu}\end{aligned}$$

so the triplet is

$$\mu = -\lambda \int_{-1}^{1} xF(\mathrm{d}x), \quad \sigma = 0, \quad \pi = \lambda F$$

So:

$$\lambda = \int_{\mathbb{R}} \pi(\mathrm{d}x)$$

called **total mass of** π, is the Poisson rate for jump arrivals.

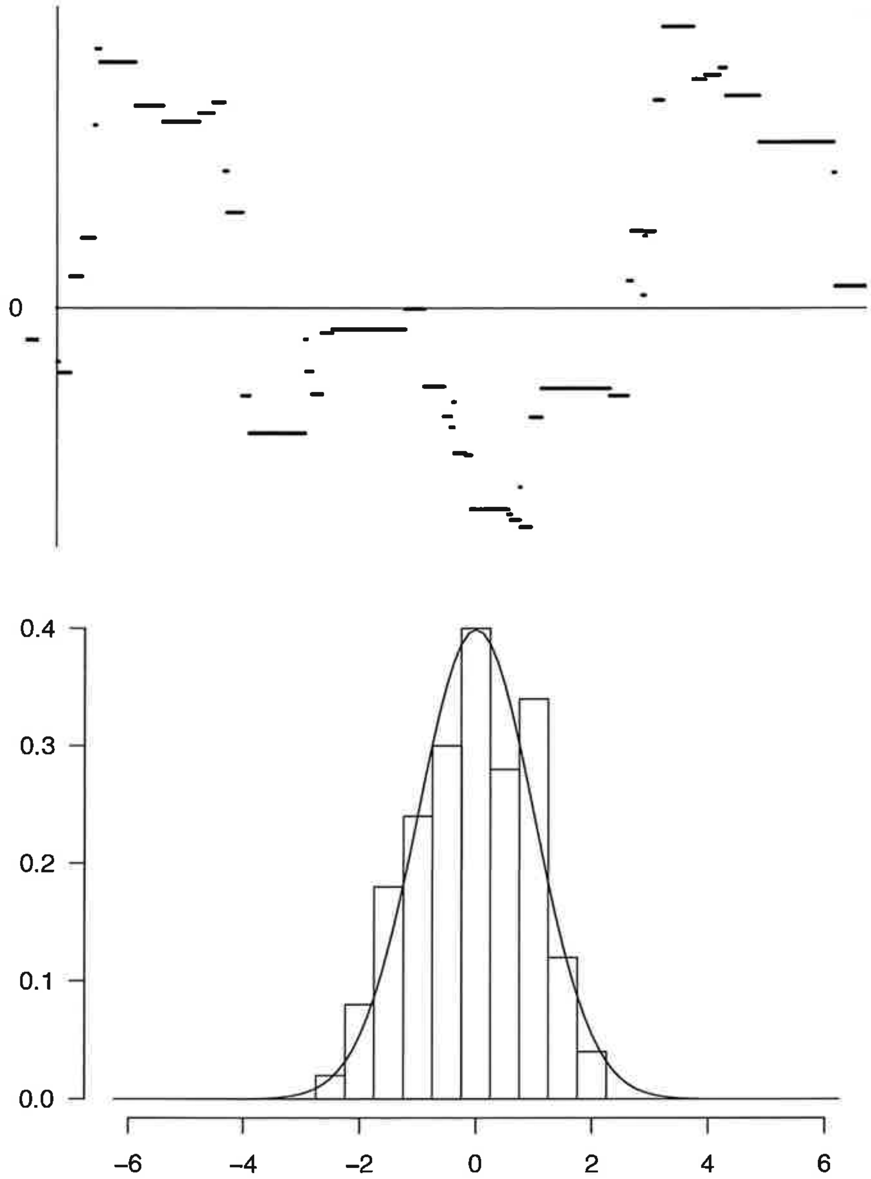
Remark

The Lévy intensity of a CP has a finite total mass by construction!

The normalized π yields

$$F = \lambda^{-1}\pi$$

which gives the distribution of the jumps.



A CP trajectory and empirical distribution of the jumps vs. a $N(0, 1)$.

We talked about Lévy processes, which are Markov processes with stationary and independent increments, characterized by the characteristic exponent at time 1 through

$$\psi_t(u) = t\psi_1(u)$$

7.1 Subclasses of Lévy Processes

We introduced a different object, a compound Poisson process:

$$X(t) = \sum_{i=1}^{N}(t)Z_i$$

with $Z_i \overset{i.i.d.}{\sim} F$ with no mass at 0. in particular

$$N(t) \sim \text{Pois}(\lambda t)$$

The characteristic exponent of a generic compound Poisson process is

$$\psi(u) = \lambda \int_{\mathbb{R}} (1 - e^{iux})F(\mathrm{d}x)$$

Remark

The F is the distribution of the jump sizes and the lambda is the Poisson rate for jumps arrival. λ is the Poisson rate for jump arrivals

$$\lambda = \int_{\mathbb{R}} \pi(\mathrm{d}x) < \infty$$

by construction.

We can add something to this. If we add a drift to the compound Poisson process, we obtain a Lévy process with characteristic exponent given by the sum of the exponents:

$$\lambda \int_{\mathbb{R}} (1 - e^{iux})F(\mathrm{d}x) - iub$$

where the drift is bt.

Exercise 7.4

Show the triplet is (μ, σ, π), where

- $\sigma = 0$
- $\pi = \lambda F$
- $\mu = -(b + \lambda \int_{-1}^{1} xF(\mathrm{d}x))$

Example 7.7

In the previous lecture we had shown that a random variable $Y \sim \Gamma(\alpha, \beta)$ is infinitely divisible with characteristic function

$$\frac{1}{(1 - \frac{iu}{\beta})^\alpha} \underbrace{=}_{\text{Frullani integral}} \exp\left\{-\int_0^\infty (1 - e^{iux})\alpha x^{-1} e^{-\beta x}\,\mathrm{d}x\right\} \tag{8}$$

> Now, clearly, the integral becomes the characteristic exponent for the random variable since we can see
>
> $$\int_0^\infty (1 - e^{iux}\alpha x^{-1}e^{-\beta x} = \psi(u)$$
>
> and
>
> $$\alpha x^{-1}e^{-\beta x}\,\mathrm{d}x = \pi(\mathrm{d}x)$$
>
> We can add and subtract $iux1_{(|x|<1)}$ and therefore, similarly to CP process, we get
>
> $$\psi(u) = \int_0^\infty (1 - e^{iux} - iux1_{(|x|<1)}\pi(\mathrm{d}x) - iu\int_0^1 x\pi(\mathrm{d}x)$$
>
> From this, we can read the Lévy triplet
>
> $$\mu = -\int_0^1 x\pi(\mathrm{d}x), \quad \sigma = 0, \quad \pi(\mathrm{d}x) = \alpha x^{-1}e^{-\beta x}\,\mathrm{d}x$$

We can define basing on this a Lévy process, then we want to understand what does is mean in terms of trajectories. We have infinitive mass around 0.
We can use $\psi_t(u) = t\psi_1(u)$ to define a Gamma process as a Lévy process with triplet as above, that is with exponent for $X(t)$ given by

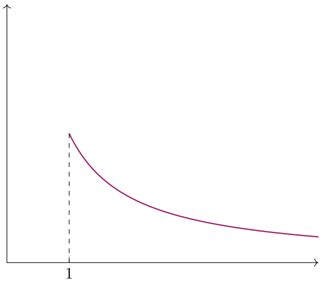
$$\psi_t(u) = \int_0^\infty (1 - e^{iux})dtx^{-1}e^{\beta x}\,\mathrm{d}x$$

Now we can work background and fine the increments of the process.
So through (8) we see that

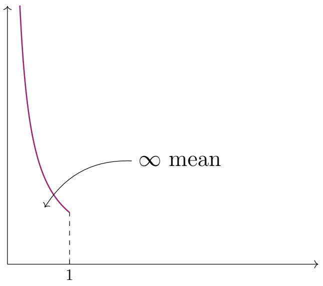
$$X(t) \overset{d}{=} X(s+t) - X(s) \sim Ga(\alpha t, \beta)$$

We would like to know the behaviour in terms of trajectories.
Note that:

- for $x \to \infty, \pi(\mathrm{d}x) \approx e^{-\beta x}$ is integrable $\implies \pi((1,\infty)) < \infty$.



- for $x \to 0^+$, $\pi(\mathrm{d}x) \approx \frac{1}{x}$ $\implies \pi\left((0,1)\right) = \infty$: there is infinite mass in every neighbourhood of 0.



  The requirement in the LévyKhintchine formula

  $$\int_{\mathbb{R}} \min(1, x^2)\pi(\mathrm{d}x) < \infty$$

  is satisfied: around 0 we have $\frac{x^2}{x}$.
  Technically we have something legitimate that we do not understand: since $\pi(\mathbb{R}_+) = \infty$ it is not a Compound Poisson.

The Gamma process belongs to the following subclass of Lévy processes.

Definition 7.5

A **subordinator** is a Lévy process with almost surely non-decreasing sample paths, hence with triplet:

- $\mu \leq 0$ (so we have a non-negative drift $-\mu$);
- $\sigma = 0$ that is, there is no Brownian component (indeed, it couldn't have this component because of its oscillations);
- $\pi((-\infty, 0]) = 0$, (se we only select positive jumps).

Remark

The Gamma process is gonna be used a lot in Bayesian Statistics where we will use it to create appropriate prior distributions.

We want to understand path properties when π has total mass. We write a generic characteristic exponent as follows

$$\psi(u) = \underbrace{i\mu u + \frac{1}{2}\sigma^2 u^2}_{\psi^{(1)}} + \underbrace{\int_{(-1,1)^c} (1 - e^{iux})\pi(\mathrm{d}x)}_{\psi^{(2)}} + \underbrace{\int_{(-1,1)} (1 - e^{iux} + iux1_{(|x|<1)})\pi(\mathrm{d}x)}_{\psi^{(3)}}$$

$\psi^{(1)}$ is the exponent of a Brownian Motion

$$dX^{(1)}(t) = -\mu dt + \sigma dB(t)$$

- $\psi^{(1)}$: is a linear BM
- $\psi^{(2)}$: outside $(-1,1)$ π has finite mass so $\psi^{(2)}$ can be written as

  $$\lambda_0 \int_{(-1,1)^C} (1 - e^{iux}) F_0(\mathrm{d}x)$$

  with $\lambda_0 := \pi((-1,1)^C)$ and $F_0 := \lambda_0^{-1}\pi|_{(-1,1)^C}$.
  We look outside a neighborhood of zero and then we can normalize measure and therefore we can have a compound Poisson.
  So $\psi^{(2)}$ corresponds to a CP process with λ_0 Poisson rate and F_0 distribution for jumps of size $|x| \geq 1$.
- $\psi^{(3)}$: we can distinguish two cases

  1. $\pi((-1,1)) < \infty$

     $$\int_{-1}^{+1} (1 - e^{iux} + iux) \underbrace{\pi(\mathrm{d}x)}_{\substack{<\infty\text{: we can split} \\ \text{the integral}}} = ux\pi(\mathrm{d}x) \underbrace{\int_{-1}^{+1} (1 - e^{iux})\pi(\mathrm{d}x)}_{C.P.} + iu \underbrace{\int_{-1}^{+1} x\pi(\mathrm{d}x)}_{\text{drift}}$$

  2. $\pi((-1,1)) = \infty$ Jumps arrive at infinite rate (the total mass is infinite) but we cannot normalize π to get the jump distribution. So, the condition

     $$\int_R \min(1, x^2)\pi(\mathrm{d}x) < \infty$$

     implies:

     · "big jumps" (size greater or equal than 1) arrive at finite rate $\pi((-1,1)^C) < \infty$ hence they occur finitely often in every bounded interval.

- If $\pi((-1,1)) = \infty$, "small jumps" arrive at infinite rate (precisely given by that mass), hence they occur infinitely often in every bounded interval.

Even if it seems that the trajectories are somewhere flat, they are not: there are infinite small jumps. The Gamma process only increases by jumps. The trajectory are nowhere continuous. Such a process is said to have **infinite activity**.

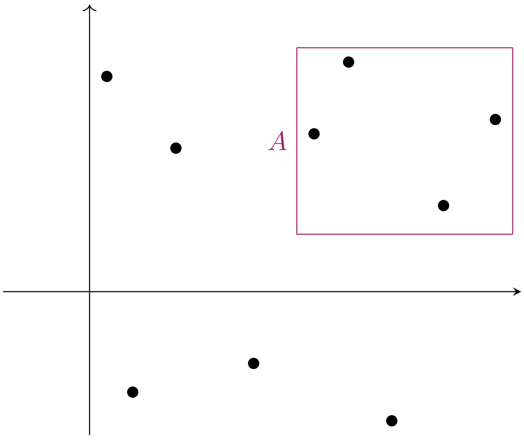
Definition 7.6

A measure N on a σ-finite measure space $(\mathbb{X}, \mathcal{X}, \mu)$ is said to be a **Poisson random measure** (PRM) with mean intensity measure μ if for mutually disjoint sets $A_1, \ldots, A_k \in \mathbb{X}$

$$\mathcal{N}(A_i) \overset{i.i.d.}{\sim} P_0(\mu(A_i))$$

Example 7.8

$\mathbb{X} = \mathbb{R}_+ \times \mathbb{R}$



$N(A) \overset{i.i.d.}{\sim} Po(\mu(A))$.
If $A = [s,t]x[a,b]$, $\mu =$ Lebesgue measure $\times\, \pi$

$$\mu(A) = \lambda(t-s) \int_a^b F(\mathrm{d}x)$$

where $\pi = \lambda F$.

Proposition 7.1

Let N be a PRM on $[0,\infty) \times \mathbb{R}$ with mean intensity $\mu = Leb \times \pi$, with π finite on $\mathbb{R}$ such that $\pi(\{0\}) = 0$.
For any Borel set $B \in \mathcal{B}(\mathbb{R})$

$$X_B(t) := \int_0^t \int_B x N(ds, \mathrm{d}x)$$

is a compound Poisson process with rate $\lambda_B := \pi(B)$ and jump distribution $F_B = \lambda_B^{-1} \pi|_B$.

Example 7.9

Assume that $B = \mathbb{R}$, π is finite. Then, we can show that

$$N = \sum_{i\geq 1} 1_{(s_i, Z_i)}$$

which gives us the random point configuration.
This yields that

$$\int_0^t \int_B x \sum_{i\geq 1} 1_{(s_i,Z_i)}(ds, \mathrm{d}x) = \bigotimes$$

What we are saying is: we take the s and accumulate the x and the x is the second coordinate z. Sum up the second coordinate, that is sum up the heights of everything.

$$\bigotimes = \sum_{i\geq 1} Z_i 1_{(s_i \in [0,t], Z_i \in B)}$$

The PRM is a random object which gives us the properties of the sample paths.
We sum points heigth of every point $(S_i, Z_i) \in [0, t] \times B$ (if $B = \mathbb{R}$, we sum all jumps sizes/points heights in $(0, t]$).

If we now let, for general π (so in every $[0, t] \times [0, \varepsilon]$ there could be infinitely many points).

$$\varepsilon_m = \frac{1}{x^m}, \quad m \geq 1$$
$$B_m = (-1, -\varepsilon_m] \cup [\varepsilon_m, 1)$$

Set

$$X^{(\varepsilon,m)}(t) := \underbrace{\int_0^t \int_{B_m} xN(ds, \mathrm{d}x)}_{\text{CP since}\pi(B_m)<\infty} - t \underbrace{\int_{B_m} x\pi(\mathrm{d}x)}_{\text{drift}}$$

which has exponent

$$\psi^{(3,m)} = \int_{B_m} (1 - e^{iux})\pi(\mathrm{d}x) + iu \int_{B_m} x\pi(\mathrm{d}x)$$

informally, as $n \to \infty$, we obtain the exponent we were looking for.

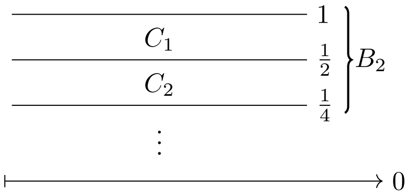
$$\psi^{(3)} = \int_{|x|<1} (1 - e^{iux} + iux)\pi(\mathrm{d}x)$$

More formally:

Proposition 7.2

Let $X^{(3,m}$ be a Lévy process with exponent $\psi^{(\varepsilon,m)}(CP + drift)$.
As $m \to \infty, X^{(3,m)} \overset{a.s.}{\to} X^{(3)}$ uniformly over $[0, T]$, where $X^{(3)}$ is a Lévy process with exponent $\psi^{(3)}$ and with at most countably-many discontinuities in every bounded interval.

If $C_n = [\frac{1}{2^n}, \frac{1}{2^{n-1}}), B_m = \bigcup_{n=1}^m C_n$:



$$\psi^{(\varepsilon,m)} = \int_{B_m} (1 - e^{iux} + iux)\pi(\mathrm{d}x) =$$

$$= \sum_{n=1}^{m} [\lambda \underbrace{\int_{C_n} (1 - e^{iux}) F_n(\mathrm{d}x)}_{C.P.} + iu\lambda_n \underbrace{\int_{C_n} x F_n(\mathrm{d}x)]}_{\text{Drift}}$$

Hence the exponent of the process that converges to the limit process is given by a sum of cp processes whose jumps sizes are in disjoint intervals.

Theorem 7.2

Lévy-Ito decomposition. Let (μ, σ, π) satisfy the conditions of the Lévy-Kinthchine formula.
Any Lévy process X is the superposition of three intependent Lévy processes, such that

$$X = X^{(1)} + X^{(2)} + X^{(3)}$$

where

1. $X^{(1)}$ is a linear Brownian Motion $dX^{(1)} = \mu dt + \sigma dB(t)$
2. $X^{(2)}$ is a CP with jumps low

$$F_0 := \lambda_0^{(-1)} \pi|_{(-1,1)^C}$$

and rate

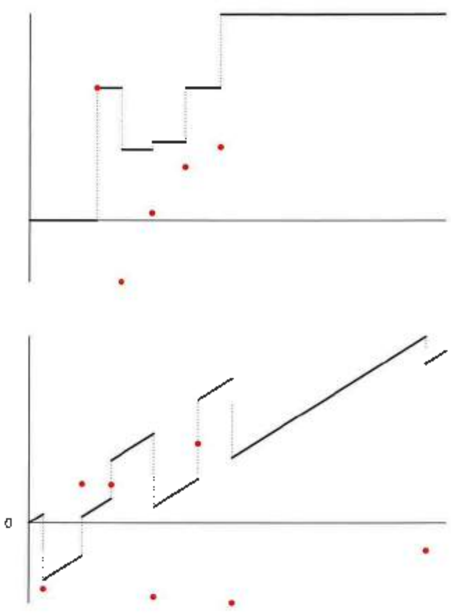
$$\lambda_0 := \pi((-1,1)^C)$$

3. $X^{(3)}$ is the sum of countably-many CP processes with drift and jumps in $(-1, 1)$.

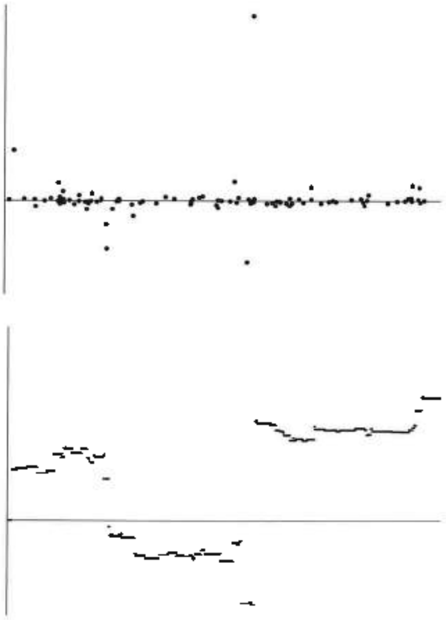
Exercise 7.5

Verify superposition of CP and parameterisation for the Gamma process.

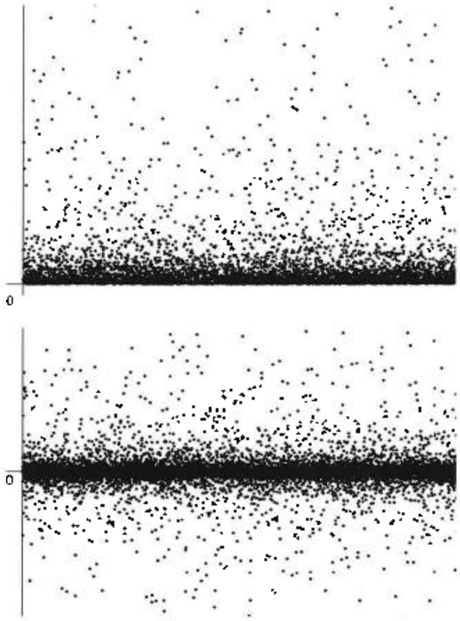
So the Gamma process has no Brownian component, π was on mR_+ and there was no drift. Hence, it is a process which there are infintely - many jumps in every small bounded interval.



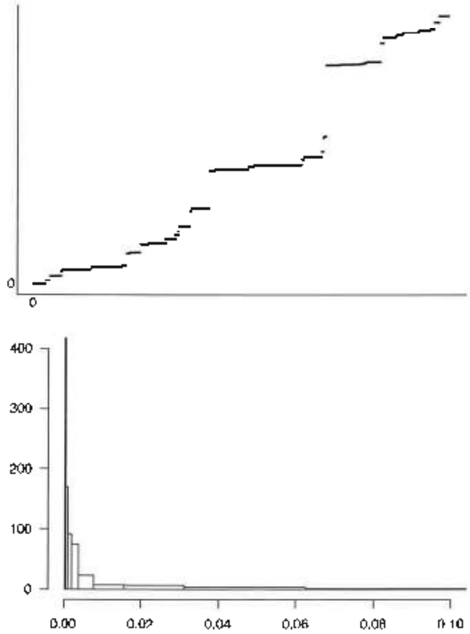
A realization of a PRM and the resulting trajectory of the CP process. Bottom: same with additional drift bt.



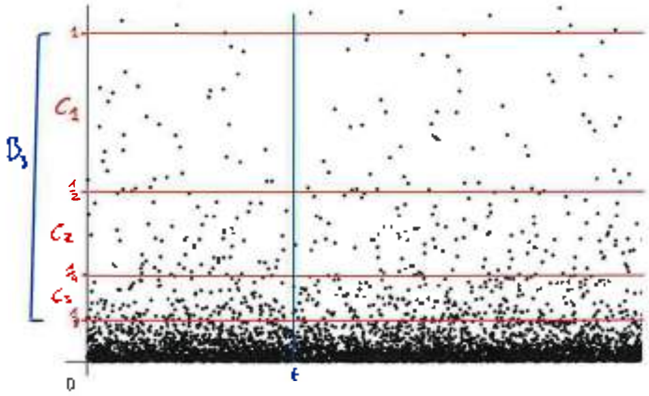
A realization of a PRM with Cauchy Lèvy intensity (finite mass) and the resulting trajectory of the Cauchy CP process.



A realization of a PRM with infinite total mass around 0, with null (top) and positive (bottom) mass on the negative half line.



Trajectory of a Gamma process, and the histogram of the jumps ("empirical" π) with size in B_n, $1 \le n \le 14$.



Slicing of the jump size space for given t into sets $C_m = [2^{-m}, 2^{-m+1})$. In each C_m every realization has almost surely finitely-many points, so

$$X_{C_m} = \int_0^t \int_{C_m} x N(\mathrm{d}s, \mathrm{d}x)$$

is a CP process. Then X_{B_m} is the superposition of finitely-many CP processes, and the a.s. limit of X_{B_m} yields a Lévy process with the desired exponent.