

Stochastic Systems — Discrete Time Systems

Simone Santini

Class notes, Academic Year 2020/2021

1 Random variables

A **probability space** is a triple (Ω, σ, μ) where

probability space

- i) Ω is a set of *outcomes* (we shall use the notation $\omega \in \Omega$ for its elements).
- ii) σ is a σ -algebra on Ω , that is a non-empty set of subsets of Ω ($\sigma \subseteq 2^\Omega$) such that
 - a) $\Omega \in \sigma$
 - b) $A \in \sigma \implies \Omega \setminus A \in \sigma$ (we shall use the notation A^c for $\Omega \setminus A$)
 - c) If $A_n \in \sigma$ for all $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} A_n \in \sigma$.
- iii) $\mu : \sigma \rightarrow [0, 1]$ is such that
 - a) $\mu(\Omega) = 1$
 - d) Given $A_n, n \in \mathbb{N}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=0}^{\infty} \mu(A_n) \quad (1)$$

Condition ii) implies that $\bigcap_{n \in \mathbb{N}} A_n \in \sigma$, since

$$\bigcap_{n \in \mathbb{N}} A_n = \left(\bigcap_{n \in \mathbb{N}} A_n^c\right)^c \quad (2)$$

A subset $A \subset \Omega$ with $A \in \sigma$ is called an **event**. It is possible to show (we shall not do it here) that the sum in (1) is well defined, that it, it is independent on the way we index the sets A_n . With these properties, it is possible to take limit:

event

Theorem 1.1. *Is A_n , $n \in \mathbb{N}$ is a sequence of subsets of Ω and $\lim_{n \rightarrow \infty} A_n = A$, then $\lim_{n \rightarrow \infty} \mu(A_n) = \mu(A)$.*

Let M be a countable set. A (discrete) *random variable* is a function $X : \Omega \rightarrow M$ such that, for all $m \in M$,

$$\{X = m\} \triangleq \{\omega \in \Omega \mid X(\omega) = m\} \in \sigma \quad (3)$$

That is, a random variable is a mathematical onjet that associates values of M to outcomes in such a way that the back-image of any value of M is an event. The probability of the event $\{X = m\}$ is

$$P_X(m) \triangleq \mu(\{X = m\}) \quad (4)$$

The properties of μ induce corresponding properties in P_X :

i) $m \in M \implies P_X(m) \geq 0$

ii) $\sum_{m \in M} P_X(m) = 1$

If Ω and M are uncountable, the general principles are the same, but the definitions and the conditions are technically more complex, and we shall not go into the details here. In this case, P_X is a **probability density function**, events are subsets of Ω of non-zero measures, and probabilities are integrals of P_X over finite sets determined by images of events. In this case, the normalization condition is

probability density

$$\int_{\Omega} P_X(x) dx = 1 \quad (5)$$

If $\Omega = \mathbb{R}$ (as we shall often assume) we have

$$\int_{-\infty}^{\infty} P_X(x) dx = 1 \quad (6)$$

For continuous variables on \mathbb{R} one can define the **cumulative probability function**, that is, the probability that X be at most x :

cumulative probability

$$\mathcal{P}(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^x P_X(u) du \quad (7)$$

Note that

$$P_X(x) = \frac{\partial}{\partial x} \mathcal{P}(x). \quad (8)$$

From this and the positivity condition one can show that \mathcal{P} is monotonically non-decreasing and that

$$\lim_{x \rightarrow -\infty} \mathcal{P}(x) = 0 \quad \lim_{x \rightarrow \infty} \mathcal{P}(x) = 1 \quad (9)$$

A whole function might be difficult to work with; it is easier to work with an enumerable set of numbers that characterizes the function completely. **Statistical moments** are such quantities. The moment of order n of the variable X is defined as

statistical moments

$$\langle X^n \rangle = \int_M x^n P_X(x) dx \quad (10)$$

In general, given a function f defined on M , we define

$$\langle f(X) \rangle = \int_M f(x) P_X(x) dx \quad (11)$$

The n th moment is obtained for $f(x) = x^n$.

The first order moment $\langle X \rangle$ is called the **average**, or the **expected value** of X , while

expected value

$$\sigma^2 = \langle X^2 \rangle - \langle X \rangle^2 \quad (12)$$

is its **variance**; the square root of the variance, σ is the **standard deviation** of X .

variance
standard deviation

Not all distributions have finite moments, that is, the integral (10) may fail to converge. If the moments are finite, then they completely characterize the PDF. To show this, we introduce the **characteristic function** $\tilde{P}_X(\omega)$ ¹ of a PDF P_X :

characteristic function

$$\tilde{P}_X(\omega) = \langle e^{i\omega x} \rangle = \int_M e^{i\omega x} P_X(x) dx \quad (13)$$

This is simply the Fourier transform of P_X , so the PDF can be recovered from its characteristic function as

$$P_X(x) = \frac{1}{2\pi} \int e^{-i\omega x} \tilde{P}_X(\omega) d\omega \quad (14)$$

The relation with the moments becomes evident by taking the Taylor expansion of the exponential:

$$e^{i\omega x} = \sum_n \frac{(i\omega x)^n}{n!} \quad (15)$$

¹There is a possible confusion here: we have already used the symbol ω to indicate an occurrence ($\omega \in \Omega$), and now we are using it to indicate the pulsation variable in the Fourier Transform. Unfortunately, both usages are standard (there are many more mathematical concepts that letters available in the Latin and Greek alphabets, and other symbols are hard to come by in L^AT_EX). Fortunately, we shall not need the set Ω anymore, so from now on ω will indicate the Fourier transform variable.

Introducing this into (13) we get

$$\tilde{P}_X(\omega) = \sum_n \frac{(i\omega)^n}{n!} \int x^n P_X(x) dx = \sum_n \frac{(i\omega)^n}{n!} \langle X^n \rangle \quad (16)$$

A useful consequence of this expansion is that the moments of P_X can be obtained by differentiating \tilde{P}_X :

$$\langle X^n \rangle = \lim_{\omega \rightarrow 0} (-i)^n \frac{\partial^n}{\partial \omega^n} \tilde{P}_X(\omega) \quad (17)$$

* * *

The **joint probability** of two random variables X_1 and X_2 , indicated as $P_{X_1 \cap X_2}(x_1, x_2)$ measures the simultaneous probability that X_1 and X_2 take the values x_1 and x_2 , respectively. The **conditional probability** $P_{X_1|X_2}(x_1|x_2)$ denotes the probability that X_1 take value x_1 conditioned to the fact that X_2 takes value x_2 . Two variables are **independent** if for all x_1, x_2 $P_{X_1|X_2}(x_1|x_2) = P_{X_1}(x_1)$, that is, if knowing the value of X_2 does not change the distribution of X_1 . Joint and conditional probabilities are related through Bayes's theorem:

$$P_{X_1 \cap X_2}(x_1, x_2) = P_{X_1|X_2}(x_1|x_2)P_{X_2}(x_2) = P_{X_2|X_1}(x_2|x_1)P_{X_1}(x_1) \quad (18)$$

joint probability

conditional probability
independence

1.1 Useful Probability Distributions

A variable X follows a **Gaussian** (or *normal*) distribution if

$$P_X(x) = N_X(x) = N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) \quad (19)$$

Gaussian distribution

(Figure 1) or, equivalently, it has characteristic function

$$\tilde{P}_X(\omega) = \tilde{N}(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} P_X(x) dx = \exp\left(i\omega\mu - \frac{\omega^2\sigma^2}{2}\right) \quad (20)$$

The mean of the distribution is $\langle X \rangle = \mu$. Note that, for $\mu = 0$, the characteristic function has also the functional form of a Gaussian, a fact that will come handy in the following. In this special case ($\mu = 0$) the moments are given by

$$\langle X^n \rangle = \int_{-\infty}^{\infty} x^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx = \begin{cases} \frac{2^{\frac{n}{2}}\sigma^n}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2}\right) & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \quad (21)$$

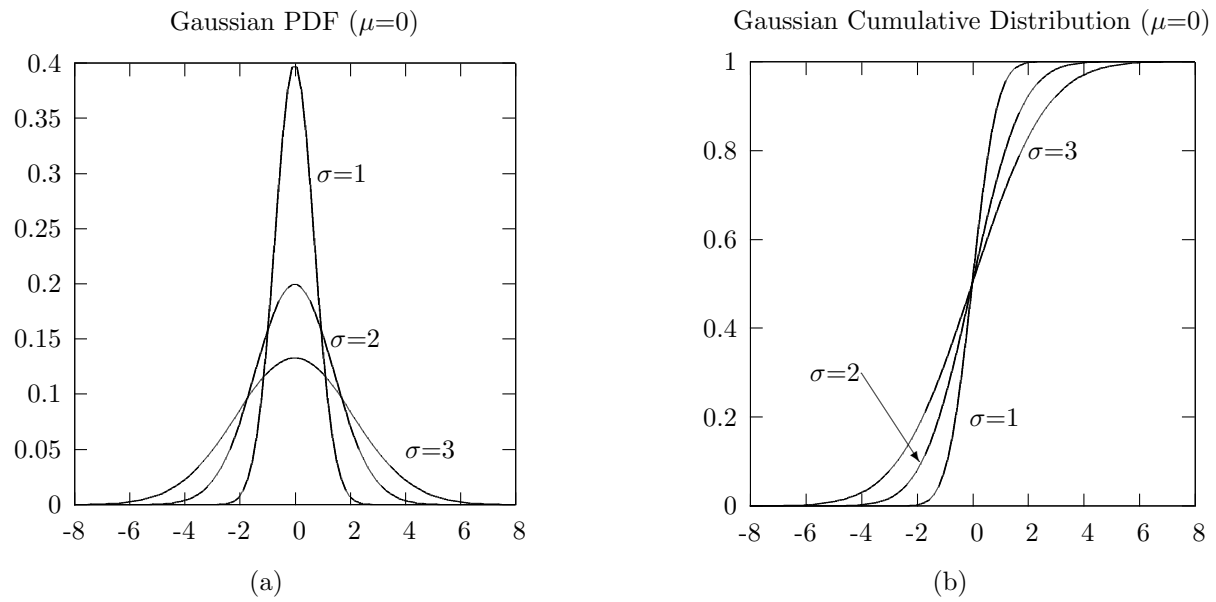


Figure 1: The Gaussian PDF (a) and the corresponding cumulative distribution (b) for various values of σ ; in all cases it is $\mu = 0$.

where Γ is Euler's Gamma function. An important moment is

$$\langle X^2 \rangle = \sigma^2 + \langle X \rangle. \quad (22)$$

One important property of the Gaussian distribution, vis-à-vis the Central Limit Theorem (which we shall consider in the following), is that it is *stable*: if X, Y are Gaussians, and $a, b \in \mathbb{R}$, then $aX + bY$ is also Gaussian.

Let X and Y be two Gaussian-distributed variables with zero mean and variance σ_1^2 and σ_2^2 , respectively. Then

$$\begin{aligned} \tilde{N}_X(\omega) &= \exp\left(-\frac{\omega^2 \sigma_1^2}{2}\right) \\ \tilde{N}_Y(\omega) &= \exp\left(-\frac{\omega^2 \sigma_2^2}{2}\right) \end{aligned} \quad (23)$$

and consequently

$$\tilde{N}_X(\omega) \tilde{P}_Y(\omega) = \exp\left(-\frac{\omega^2 (\sigma_1^2 + \sigma_2^2)}{2}\right) \quad (24)$$

that is, the product of the characteristic function of two Gaussian distributions is still the characteristic function of a Gaussian distribution.

* * *

The Gaussian distribution is defined for all $x \in \mathbb{R}$, but in the model of many interesting phenomena the variable assumes only positive values in such a way that the probability that $x = 0$ is 0 and, after reaching a maximum, decreases rapidly for high values of x . Examples of this kind of phenomena abound and are of the most diverse nature, from the length of messages in internet fori, to the price of hotels or the size of the fragments resulting from a collision. In these cases, negative values are out of the question, so we can't model them using a Gaussian distribution for which $N_X(x) > 0$ for all $x \in \mathbb{R}$.

All these phenomena can be modeled by means of a **logonormal distribution**. A variable X has logonormal distribution if $\log X$ has normal (viz. Gaussian) distribution. Let Φ and ϕ be the cumulative distribution and the density of a normally distributed variable with 0 mean and unit variance ($\mathcal{N}(0, 1)$), and assume $\log X \sim \mathcal{N}(\mu, \sigma)$, i.e. $\log X$ has a normal distribution with mean μ

Logonormal distribution

and variance σ^2 . Then

$$\begin{aligned}
 P_X(x) &= \frac{d}{dx} \mathcal{P}_X(x) = \frac{d}{dx} \mathbb{P}[X \leq x] \\
 &= \frac{d}{dx} \mathbb{P}[\log X \leq \log x] \\
 &= \frac{d}{dx} \Phi\left[\frac{\log x - \mu}{\sigma}\right] \\
 &= \phi\left[\frac{\log x - \mu}{\sigma}\right] \frac{d}{dx} \left[\frac{\log x - \mu}{\sigma}\right] \\
 &= \frac{1}{\sigma x} \phi\left[\frac{\log x - \mu}{\sigma}\right] \\
 &= \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{(\log X - \mu)^2}{2\sigma^2}\right]
 \end{aligned} \tag{25}$$

Figure 2 shows the behavior of the logonormal PDF for various values of σ and $\mu = 0$. Note that μ

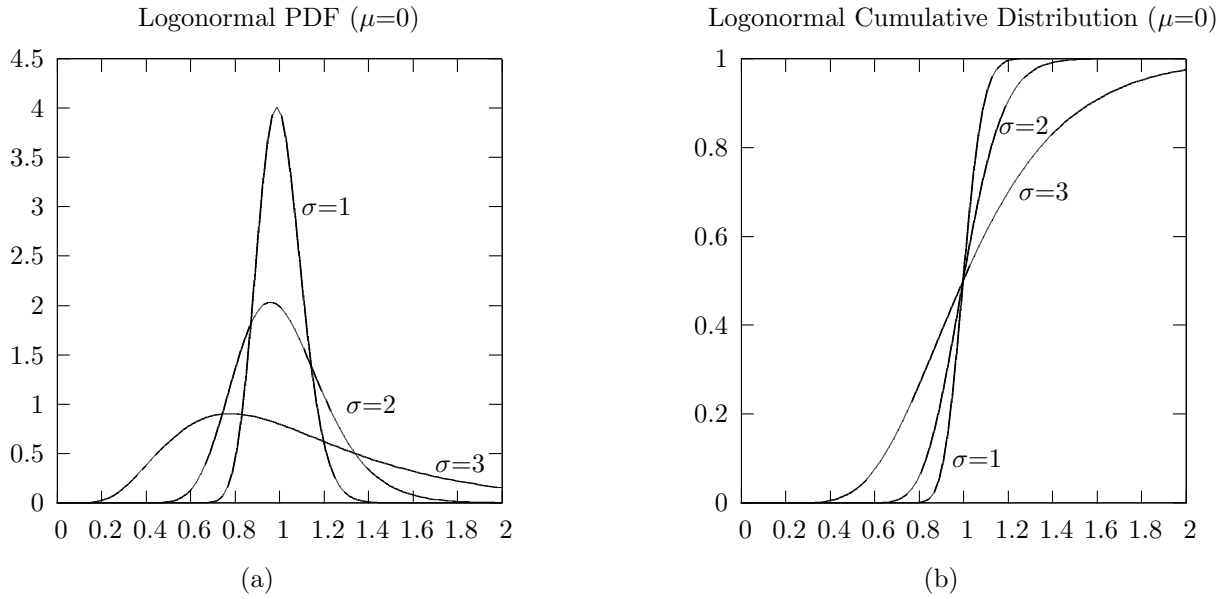


Figure 2: The Logonormal PDF (a) and the corresponding cumulative distribution (b) for various values of σ ; in all cases it is $\mu = 0$.

and σ are the mean and variance of $\log X$, *not* of X . To distinguish them, I shall indicate the mean and the variance of X as m and v , respectively.

The moments of X are given by

$$\langle X^n \rangle = \int_0^\infty x^n P_X(x) dx = \exp\left(n\mu + \frac{n^2\sigma^2}{2}\right) \tag{26}$$

as can be verified by replacing $z = \frac{1}{\sigma} [\log X - (\mu + n\sigma^2)]$ in the integral. From this we have

$$\begin{aligned} m &= \langle X \rangle = \exp\left(\mu + \frac{\sigma^2}{2}\right) \\ \langle X^2 \rangle &= \exp(2\mu + 2\sigma^2) \\ v &= \langle X^2 \rangle - \langle X \rangle^2 = \exp(2\mu + \sigma^2)(e^{\sigma^2} - 1) \end{aligned} \quad (27)$$

From these equality, one can derive the values of μ and σ^2 for desired m and v :

$$\mu = \log \frac{m}{\sqrt{1 + \frac{v}{m^2}}} \quad \sigma^2 = \log \left(1 + \frac{v}{m^2}\right) \quad (28)$$

The characteristic function $\langle \exp(i\omega x) \rangle$ is defined, but if we try to extend it to complex variables, $\langle \exp(sx) \rangle$, $s \in \mathbb{C}$ is not defined for any s with a negative imaginary part. This entails that the characteristic function is not analytical in the origin and, consequently, it can't be represented as an infinite convergent series. In particular, the formal Taylor series

$$\sum_n \frac{(i\omega x)^n}{n!} \langle x^n \rangle = \sum_n \frac{(i\omega x)^n}{n!} \exp\left(n\mu + \frac{n^2\sigma^2}{2}\right) \quad (29)$$

diverges

* * *

Other positive variables follow a different distribution, one in which the value 0 is the most probable, and the probability decreases sharply as x increases, In these cases, the variable x can be modeled using an **exponential** distribution (Figure 3).

Exponential distribution

$$P_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (30)$$

If the variable can take negative values, then the distribution takes the name of **Laplace** distribution

Laplace distribution

$$P_X(x) = \frac{\lambda}{2} e^{-\lambda|x|} \quad (31)$$

Its characteristic function is

$$\tilde{P}_X(\omega) = \frac{\lambda^2}{\lambda^2 + \omega^2} \quad (32)$$

and its moments

$$\langle X^n \rangle = \frac{1}{\lambda^n} \Gamma(n+1) \quad (33)$$

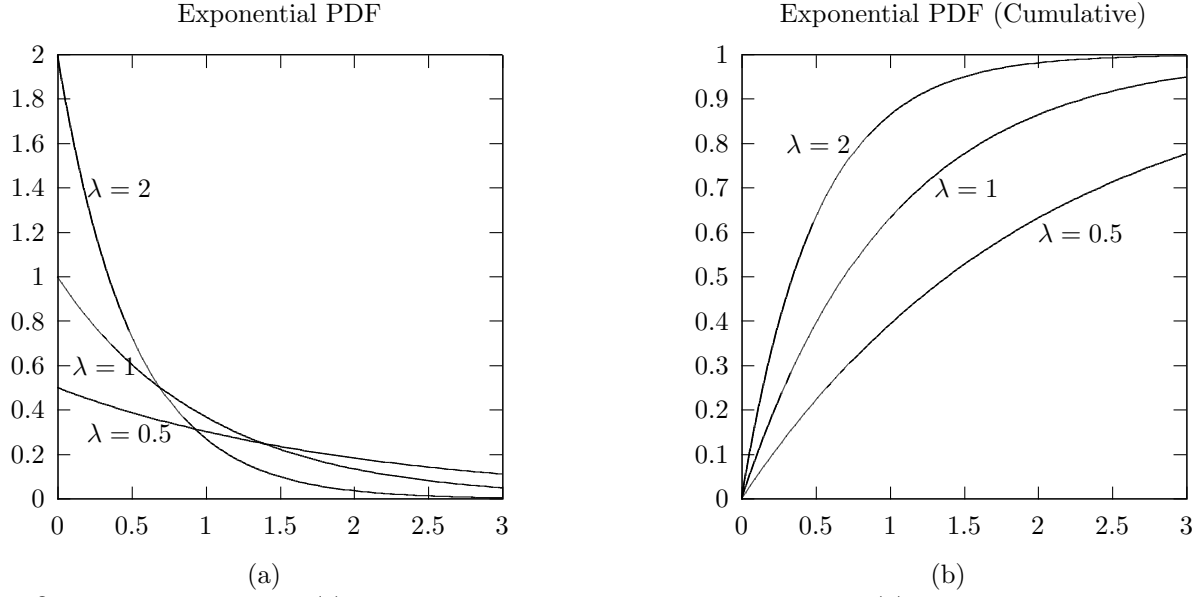


Figure 3: The exponential PDF (a), and the corresponding cumulative distribution in (b) for various values of λ .

* * *

A **uniform** or *flat* distribution assigns the same probability density to each point in Ω . So, if $\Omega = [a, b]$,

Uniform distribution

$$P_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (34)$$

The characteristic function of the uniform distribution is

$$\tilde{P}_X(\omega) = \frac{e^{i\omega b} - e^{i\omega a}}{i\omega(b-a)} \quad (35)$$

and its moments

$$\langle X^n \rangle = \frac{1}{n+1} \frac{b^{n+1} - a^{n+1}}{b-a} \quad (36)$$

* * *

A **Cauchy**, or **Lorentz** distribution has PDF

Cauchy distribution

$$P_X(x) = \frac{1}{\pi} \frac{\gamma}{x^2 + \gamma^2} \quad (37)$$

where γ is a positive parameter, and characteristic function

$$\tilde{P}_X(\omega) = e^{-\gamma|\omega|} \quad (38)$$

If one tries to compute the moments using the definition

$$\langle X^n \rangle = \frac{\gamma}{\pi} \int \frac{x^n}{x^2 + \gamma^2} dx \quad (39)$$

then, since the integrand behaves as x^{n-2} for $x \rightarrow \infty$, one observes that they diverge for $n \geq 1$. This limits the usefulness of this distribution as a model of real phenomena (which typically have finite moments), and in practice one "truncates" the distribution to a finite interval $[a, b]$.

* * *

We have mentioned that one important property of the Gaussian distribution is the preservation of the functional form of their characteristic function under multiplication, as in (24). The Gaussian distribution is not the most general distribution with this property (although it is the only one with this property *and* finite moments): it is shared by the family of **Lévy distributions**. Lévy distributions depend on four parameters: α (Lévy index), β (skewness), μ (shift), and σ (scale), and they are defined through their characteristic function:

Lévy distribution

$$\tilde{P}_{\alpha,\beta}(\omega; \mu, \sigma) = \int_{-\infty}^{\infty} e^{i\omega x} P_{\alpha,\beta}(x; \mu, \sigma) dx \triangleq \exp \left[i\mu\omega - \sigma^\alpha |\omega|^\alpha \left(1 - i\beta \frac{\omega}{|\omega|} \Phi \right) \right] \quad (40)$$

where

$$\Phi = \begin{cases} \tan \frac{\alpha\pi}{2} & \alpha \neq 1, 0 < \alpha < 2 \\ -\frac{2}{\pi} \ln |x| & \alpha = 1 \end{cases} \quad (41)$$

the four parameters determine the shape of the distribution. Of these, α and β play a major rôle in this note, while μ and σ can be eliminated through proper scale and shift transformations (much like mean and variance for the Gaussian distribution):

$$P_{\alpha,\beta}(x; \mu, \sigma) = \frac{1}{\sigma} P_{\alpha,\beta}\left(\frac{x - \mu}{\sigma}; 0, 1\right) \quad (42)$$

From now on, I shall therefore ignore μ and σ and refer to the distribution as $P_{\alpha,\beta}(x)$. Note the symmetry relation

$$P_{\alpha,-\beta}(x) = P_{\alpha,\beta}(x) \quad (43)$$

The distributions with $\beta = 0$ are symmetric, and these are the ones that are the most relevant in this context. The closed form of $P_{\alpha,\beta}$ is known only for a few cases. If $\alpha = 2$ one obtains the Gaussian distribution (β is irrelevant, since $\Phi = 0$); if $\alpha = 1, \beta = 0$ one obtains the Cauchy distribution, and for $\alpha = 1/2, \beta = 1$, the Lévy-Smirnov distribution

$$P_{1/2,1}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{3}{2}} \exp(-\frac{1}{2x}) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (44)$$

The most important property in this context is the asymptotic behavior of $P_{\alpha,\beta}$ which is given by the power law

$$P_{\alpha,0}(x) \sim \frac{C(\alpha)}{|x|^{1+\alpha}} \quad (45)$$

with

$$C(\alpha) = \frac{1}{\pi} \sin\left(\frac{\pi\alpha}{2}\right) \Gamma(1+\alpha) \quad (46)$$

This power law behavior entails that arbitrarily large values are relatively probable (compared with the exponential decay of the Gaussian). Consequently, as can be expected, $\langle X^2 \rangle$ diverges for $\alpha < 2$.

* * *

The **Dirac delta distribution** is a pathological distribution useful in many contexts; for example, when dealing with certainty in a probabilistic framework, or when analyzing discrete random variables in a context created for continuous ones. The distribution is:

Dirac δ distribution

$$P_X(x) = \delta(x - x_0) \quad (47)$$

where $\delta(\cdot)$ is the Dirac distribution. The characteristic function of the distribution is

$$\tilde{P}_X(\omega) = \exp(i\omega x_0). \quad (48)$$

The function $\delta(x)$ is zero everywhere except for $x = 0$, and

$$\int_{-\infty}^{\infty} \delta(x) dx = 1 \quad (49)$$

This property entails $\delta(ax) = \delta(x)/a$. Also

$$\int_{-\infty}^{\infty} f(x) \delta(x - x_0) dx = f(x_0) \quad (50)$$

from which we derive

$$\langle x^n \rangle = x_0^n \quad (51)$$

* * *

Unlike the previous distribution, the **binomial distribution** is defined for discrete variables, in particular for a variable X that can take two values, the first one with probability p , and the second one with probability $1 - p$. Suppose, for example, that we play a game in which, at each turn, I have a probability p of winning and $1 - p$ of losing (think of head-and-tails game with a tricked coin). If we play N rounds of the game, what is the probability that I win exactly n times? This turns out to be

binomial distribution

$$P(X = n) = \binom{N}{n} p^n (1 - p)^{N-n} = \frac{N!}{n!(N-n)!} p^n (1 - p)^{N-n} \quad (52)$$

which is precisely the binomial distribution. Its characteristic function is

$$\tilde{P}(\omega) = (1 - p + pe^{i\omega})^N \quad (53)$$

from which the moments can be derived. For example

$$\langle X \rangle = \lim_{\omega \rightarrow 0} \frac{d\tilde{P}}{d\omega} = \lim_{\omega \rightarrow 0} pN e^{i\omega} (1 - p + pe^{i\omega})^{N-1} = pN \quad (54)$$

* * *

An important and common distribution, one that appears as a limiting case of many finite processes, is the **Poisson Distribution**. Its importance will probably be more evident if we derive it as a limiting case in some examples.

Poisson distribution

Example I:

Consider events that may happen at any moment in time (the events are punctual: they have no duration). Divide the time-line in small intervals of duration Δt , so short that the probability that two or more events will take place in the same interval is negligible. Assume that the probability that one event take place in $[t, t + \Delta t)$ is constant, and proportional to the length of the interval:

$$P(1; \Delta t) = \lambda \Delta t \quad (55)$$

and, because no two events happen in the same interval,

$$P(0; \Delta t) = 1 - \lambda \Delta t \quad (56)$$

Let $P(0; t)$ be the probability that no event has taken place up to time t . Then

$$P(0; t + \Delta t) = P(0; t)(1 - \lambda \Delta t) \quad (57)$$

Rearranging the terms we get

$$\frac{P(0; t + \Delta t) - P(0; t)}{\Delta t} = -\lambda P(0; t) \quad (58)$$

and, taking the limit for $\Delta t \rightarrow 0$

$$\frac{\partial}{\partial t} P(0; t) = -\lambda P(0; t) \quad (59)$$

that is, $P(0; t) = C \exp(-\lambda t)$ or, considering the boundary condition $P(0, 0) = 1$,

$$P(0; t) = e^{-\lambda t} \quad (60)$$

This takes care of the case in which no event takes place before time t . On to the general case. There were n events by time $t + \Delta t$ if either (1) we had n events up to time t and no event occurred in $[t, t + \Delta t]$, or (2) there were $n - 1$ events at t and one event occurred in $[t, t + \Delta t]$. This leads to

$$P(n; t + \Delta t) = (1 - \lambda \Delta t)P(n; t) + \lambda \Delta t P(n - 1; t) \quad (61)$$

rearranging and taking the limit $\Delta t \rightarrow 0$, we have

$$\frac{\partial}{\partial t} P(n; t) + \lambda P(n; t) = \lambda P(n - 1; t) \quad (62)$$

In order to transform this equation into a more manageable form, we look for a function that, multiplied by the left-hand side, transforms it into the derivative of a product. That is, we look for a function $\mu(t)$ such that

$$\mu(t) \left[\frac{\partial P}{\partial t} + \lambda P \right] = \frac{\partial}{\partial t} [\mu(t) P] \quad (63)$$

It is easy to verify that $\mu(t) = \exp(\lambda t)$ fits the bill. Equation (62) therefore becomes

$$\frac{\partial}{\partial t} [e^{\lambda t} P(n; t)] = e^{\lambda t} \lambda P(n - 1; t) \quad (64)$$

For $n = 1$ we have

$$\frac{\partial}{\partial t} [e^{\lambda t} P(1; t)] = e^{\lambda t} \lambda e^{-\lambda t} = \lambda \quad (65)$$

That is, integrating both sides and multiplying by $e^{-\lambda t}$

$$P(1; t) = \lambda t e^{-\lambda t} \quad (66)$$

For arbitrary n , I'll show by induction that

$$P(n; t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (67)$$

Figure 4: The Poisson PDF for various values of n .

We have already derived the result for $n = 0$ and for $n = 1$. For arbitrary n , we have

$$\begin{aligned} \frac{\partial}{\partial t} [e^{\lambda t} P(n+1; t)] &= e^{\lambda t} \lambda P(n; t) \\ &= e^{\lambda t} \lambda \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (\text{induction hypothesis}) \\ &= \lambda \frac{(\lambda t)^n}{n!} \end{aligned} \quad (68)$$

So, integrating

$$e^{\lambda t} P(n+1; t) = \frac{\lambda}{n!} \int (\lambda t^n) dt = \frac{(\lambda t)^{n+1}}{(n+1)!} + C \quad (69)$$

where $C = 0$ because of the initial conditions, so

$$P(n+1; t) = e^{-\lambda t} \frac{(\lambda t)^{n+1}}{(n+1)!} \quad (70)$$

(end of example)

The distribution that results from this example:

$$P_X(x) = e^{-x} \frac{x^n}{n!} \quad (71)$$

is the Poisson distribution that, in the example, gives us the probability that n events take place in a time x . Figure 4 shows the shape of this distribution as a function of x for various values of n .

Example II:

The Poisson distribution can also be seen as a limiting case of the binomial distribution. If p is the probability of success, then $\nu = Np$ is the expected number of successful trials, as per (54). This approximation is valid for large N . In this case, we have

$$P(n; N) = \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N} \right)^n \left(1 - \frac{\nu}{N} \right)^{N-n} \quad (72)$$

Taking $N \rightarrow \infty$, we have

$$\begin{aligned}
 P_\nu(n) &= \lim_{N \rightarrow \infty} P(n; N) \\
 &= \lim_{N \rightarrow \infty} \frac{N \cdot (N-1) \cdots (N-n+1)}{n} \frac{\nu^n}{N^n} \left(1 - \frac{\nu}{N}\right)^N \left(1 - \frac{\nu}{N}\right)^{-n} \\
 &= \lim_{N \rightarrow \infty} \frac{N \cdot (N-1) \cdots (N-n+1)}{N^n} \frac{\nu^n}{n!} \left(1 - \frac{\nu}{N}\right)^N \left(1 - \frac{\nu}{N}\right)^{-n} \\
 &= 1 \cdot \frac{\nu^n}{n!} e^{-\nu} \cdot 1 \\
 &= \frac{\nu^n}{n!} e^{-\nu}
 \end{aligned} \tag{73}$$

So, once again, we find that the number of successes has a Poisson distribution.

(end of example)

The characteristic function of the distribution (71) is

$$\tilde{P}(\omega) = e^{\lambda(e^{i\omega} - 1)} \tag{74}$$

from which we obtain

$$\langle X \rangle = \lambda \tag{75}$$

1.2 Functions of Random Variables

If X is a random variable on Ω , and $f : \Omega \rightarrow \Omega'$, then $Y = f(X)$ is a random variable on Ω' . Here I'll consider, for the sake of simplicity, the case $\Omega = \Omega' = \mathbb{R}$ (all our considerations can be generalized to arbitrary continua Ω under fairly general conditions, essentially that Ω be a metric space). In order to determine the distribution of y , I begin with a preliminary observation. For a random variable X , let $\mathbb{P}_X[x, x + \Delta x]$ the probability that the value of X falls in $[x, x + \Delta x]$. Then, for small Δx ,

$$\begin{aligned}
 \mathbb{P}_X[x, x + \Delta x] &= P(X \leq x + \Delta x) - P(X \leq x) \\
 &= \frac{\partial}{\partial x} P(X \leq x) \Delta x + O(\Delta x^2) \\
 &= P_X(x) \Delta x + O(\Delta x^2)
 \end{aligned} \tag{76}$$

Let now f be invertible, and $g = f^{-1}$. Then

$$\begin{aligned}
 P_Y \Delta y &= \mathbb{P}_Y[y, y + \Delta y] \\
 &= \mathbb{P}_X[g(y), g(y + \Delta y)] \\
 &\approx \mathbb{P}_X \left[g(y), g(y) + \left| \frac{dg}{dy} \right| \Delta y \right] \\
 &= P_X(g(y)) \left| \frac{dg}{dy} \right| \Delta y
 \end{aligned} \tag{77}$$

from which we get

$$P_Y(y) = P_X(g(y)) \left| \frac{dg}{dy} \right| \tag{78}$$

Note that equivalently one could have defined

$$P_Y(y) = \int \delta(y - f(x)) P_X dx = \langle \delta(y - f(x)) \rangle_X \tag{79}$$

where the subscript on the average reminds us that we are taking the average with respect to the distribution of X . From this, we can determine the characteristic function of Y :

$$\begin{aligned}
 \tilde{P}_Y(\omega) &= \int e^{i\omega y} P_Y(y) dy \\
 &= \int P_X(x) \left[\int e^{i\omega y} \delta(y - f(x)) dy \right] dx \\
 &= \int e^{i\omega f(x)} P_X(x) dx \\
 &= \langle \exp[i\omega f(x)] \rangle_X
 \end{aligned} \tag{80}$$

If $Y = aX$, then

$$\tilde{P}_Y(\omega) = \langle \exp[i\omega aX] \rangle_X = \tilde{P}_X(a\omega) \tag{81}$$

* * *

Consider now the sum of two random variables: $Z = X + Y$. Each value of Z can be obtained through an infinity of events: each time X takes an arbitrary value x , and y takes a value $z - x$, Z takes the same value, namely z . Summing up all these possible events we obtain

$$P_Z(z) = \int_{-\infty}^{\infty} P_X(x) P_Y(z - x) dx \tag{82}$$

This is known as the *convolution* of P_X and P_Y , often indicated as $P_Z = P_X * P_Y$. The properties of the Fourier transform entail that the corresponding relation between characteristic functions is

$$\tilde{P}_Z(\omega) = \tilde{P}_X(\omega)\tilde{P}_Y(\omega) \quad (83)$$

* * *

Let $Y = \{y_1, \dots, y_n\}$ be a set of independent and identically distributed (i.i.d.) variables with cumulative distribution \mathcal{P}_Y and density P_Y . Consider the function $\min(Y)$: we are interested in finding its density P_{\min} and cumulative distribution \mathcal{P}_{\min} . We have:

$$\mathcal{P}_Y(x) = \mathbb{P}[\min(Y) \leq x] = 1 - \mathbb{P}[\min(Y) \geq x] \quad (84)$$

We have $\min(Y) \geq x$ iff we have $y_i \geq x$ for all i , that is

$$\begin{aligned} \mathcal{P}_{\min}(x) &= 1 - \mathbb{P}[\forall y \in Y. y \geq x] \\ &= 1 - \mathbb{P}[y \geq x]^n \\ &= 1 - \left(1 - \mathbb{P}[y \leq x]\right)^n \\ &= 1 - \left(1 - \mathcal{P}_Y(x)\right)^n \end{aligned} \quad (85)$$

The density is

$$\begin{aligned} P_{\min}(x) &= \frac{d}{dx} \mathcal{P}_{\min}(x) \\ &= n \left(1 - \mathcal{P}_Y(x)\right)^{n-1} \frac{d}{dx} \mathcal{P}_Y(x) \\ &= n \left(1 - \mathcal{P}_Y(x)\right)^{n-1} P_Y(x) \end{aligned} \quad (86)$$

For the function $\max(Y)$, working in a similar way, we have

$$\mathcal{P}_{\max}(x) = (\mathcal{P}_Y(x))^n P_{\max}(x) = n(\mathcal{P}_Y(x))^{n-1} P_Y(x) \quad (87)$$

1.3 The Central Limit Theorem

The Central Limit Theorem (important enough to be granted its own acronym: CLT) is one of the fundamental results in basic probability theory and the main reason why the Gaussian distribution

is so important and so common in modeling natural events. In a nutshell, the theorem tells us the following: if we take a lot of random variables, independent and identically distributed (i.i.d.), and add them up, the result will be a random variable with Gaussian distribution. So, for example, if we repeat an experiment many times and take the average of the results that we obtain (the average is, normalization apart, a sum), no matter what the characteristics of the experiment are, the resulting average will have (more or less) a Gaussian distribution.

But, ay, there's the rub! The theorem works only in the assumption that the moments of the distributions involved be finite. We shall see shortly what happens if this assumption is not satisfied.

Let X_1, \dots, X_n be a set of i.i.d. random variables with distribution P_X , zero mean, and (finite) variance σ^2 . Note that $Y = \sum_i X_i$ has zero mean and variance $n\sigma^2$, while $Z = (\sum_i X_i)/n$ has zero mean and variance σ^2/n . It is therefore convenient to work with the variable

$$Z_n = \frac{1}{\sqrt{n}} \sum_i X_i \quad (88)$$

which has zero mean and variance σ^2 independently of n .

Theorem 1.2. *For any distribution P_X with finite mean and variance, and X_1, \dots, X_n i.i.d. with distribution P_X , for $n \rightarrow \infty$, we have $Z_n \rightarrow Z_\infty$, where Z_∞ is a Gaussian random variable with zero mean and variance σ^2 equal to the variance of P_X .*

Proof. Consider the first terms of the expansion of the characteristic function of P_X :

$$\tilde{P}_X(\omega) = \int e^{i\omega x} P_X(x) dx = 1 - \frac{1}{2}\sigma^2\omega^2 + O(\omega^3) \quad (89)$$

The characteristic function of $Y = \sum_i X_i$ is given by (83):

$$\tilde{P}_Y(\omega) = \prod_i \tilde{P}_{X_i}(\omega) = [\tilde{P}_X(\omega)]^n \quad (90)$$

(the second equality holds because the X s have the same distribution) while (81) with $a = 1/\sqrt{n}$ gives

$$\tilde{P}_Z(\omega) = P_Y\left(\frac{\omega}{\sqrt{n}}\right) = \left[P_X\left(\frac{\omega}{\sqrt{n}}\right)\right]^n \approx \left(1 - \frac{\sigma^2\omega^2}{2n}\right)^n \xrightarrow{n \rightarrow \infty} \exp\left(-\frac{1}{2}\sigma^2\omega^2\right) \quad (91)$$

Finally, from (20) we have the inverse transform

$$P_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \quad (92)$$

□

This theorem is true, in the form in which we have presented it, only for distributions X with finite mean and variance². However, the key to the theorem is an invariance property of the characteristic function of the Gaussian. Consider the equality (90); we can split it up as:

$$\tilde{P}_Z(\omega; n) = [\tilde{P}_X(\omega)]^n = [\tilde{P}_X(\omega)]^{n/2} [\tilde{P}_X(\omega)]^{n/2} = \tilde{P}_Z(\omega; n/2) \tilde{P}_Z(\omega; n/2) \quad (93)$$

Taking the limit $n \rightarrow \infty$, this gives us $P_Z(\omega) = P_Z(\omega)P_Z(\omega)$. That is: the condition for a distribution to be a central limit is that the product of two characteristic functions have the same functional form as the original distributions. As we have seen in (24), the Gaussian distribution does have this property. Nay: it is the *only* distribution with finite moments that has this property, hence its appearance in the theorem in the finite moments case, and hence its great importance in application as a model of many processes resulting from the sum of identical sub-processes.

If we abandon the finite moment hypothesis, however, there is a more general distribution to which (93) applies: the stable Levy distribution. So, a more general form of the CLT can be enunciated as:

Theorem 1.3. *For any distribution P_X , and X_1, \dots, X_n i.i.d. with distribution P_X , for $n \rightarrow \infty$, we have*

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = Z_\infty \quad (94)$$

where Z_∞ is a random variable with Levy distribution. If the variance of P_X is finite and equal to σ^2 , then Z_∞ has a Gaussian distribution with variance σ^2 .

²I have assumed zero mean since, if the mean of the X is non-zero, the mean of Z goes to infinity; this doesn't represent a major hurdle for the theorem, which can easily be generalized by subtracting the mean from the variables X and then adding it back.

2 Stochastic Processes

A **Stochastic process** is a collection of random variables $X(t)$ (often written as X_t) indexed by a parameter t , usually identified with time. Normally, we assume $t \in \mathbb{N}$ (in which case we talk about a *discrete time* stochastic process) or $t \in \mathbb{R}^+$ (*continuous time* stochastic process). Here we shall deal mainly with the first type, although many of the definitions in this section apply to both. The instantiation of a random variable is a value $x \in M$ (the range of the variable); the instantiation of a (discrete time) stochastic process is a trajectory $x : \mathbb{N} \rightarrow M$.

Stochastic process

Let $P(X(t) = x)$ (or simply $P(x; t)$) the probability that X take value x at time t . Let $x \in M$: if M is discrete, then $P(x; t)$ is a probability, while if M is continuous, it is a probability density. Similarly, $P(X(t_1) = x_1, X(t_2) = x_2)$ (or $P(x_1, x_2; t_1, t_2)$) is the probability that $X(t_1)$ take value x_1 and $X(t_2)$ take value x_2 . The generalization to k values

$$P(x_1, \dots, x_k; t_1, \dots, t_k) \quad (95)$$

gives complete information about the process. Needless to say, these (in principle, infinite) values are virtually impossible to calculate. We need to find easier ways to get at least some information about the process.

For fixed t , X_t is a stochastic variable with distribution $P_t(x) = P(x; t)$. We can characterize this distribution using pretty much the same quantities that we used for stochastic variables, with the difference that, in this case, these quantities will be functions of time. We shall, for the sake of simplicity, consider stochastic processes in which the range is \mathbb{R} . The **mean function** $\mu : \mathbb{N} \rightarrow \mathbb{R}$ is defined as

$$\mu_t = \mathbb{E}[X_t] \quad (96)$$

The **variance** of the process is the function $\gamma : \mathbb{N}^2 \rightarrow \mathbb{R}$ defined as

$$\gamma(t, t) = \mathbb{E}[(X_t - \mu_t)^2] \quad (97)$$

You are probably wondering why the parameter t is repeated. This is so because the variance is a special case of the **covariance function**

$$\gamma(s, t) = \mathbb{E}[(X_s - \mu_s)(X_t - \mu_t)] \quad (98)$$

From this function we derive the **autocorrelation function** $\rho : \mathbb{N}^2 \rightarrow \mathbb{R}$ defined as

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}} \quad (99)$$

from the inequality $|\gamma(s, t)|^2 \leq \gamma(s, s)\gamma(t, t)$ we have

$$-1 \leq \rho(s, t) \leq 1 \quad (100)$$

mean function

variance

covariance

autocorrelation

* * *

One problem with the study of stochastic processes is that in general the distribution $P_t(x)$ is a function of t so, in order to understand the process, we must characterize a different distribution for each t . In general this is impossible, unless we can explicitly write down the functional form of this dependency. The simplest (and one of the most useful) of these functional forms is the constant; we say that a process is **stationary** if $P_t(x) = P(x)$, that is, if the distribution of the process is independent of time. A process that satisfies this condition is referred to as **strongly stationary**.

stationary process

Definition 2.1. A stochastic process X_t with distribution $P(x; t)$ is **strongly stationary** if, for all x , t_1, t_2 , it is $P(x; t_1) = P(x; t_2)$.

An equivalent condition is that, for all x_1, x_2, t_1, t_2, τ ,

$$P(x_1, x_2; t_1, t_2) = P(x_1, x_2; t_1 + \tau, t_2 + \tau) \quad (101)$$

This is a fairly strong condition, stronger than we need in most cases. For most purposes, we can weaken it.

Definition 2.2. A stochastic process X_t with distribution $P(x; t)$ is **weakly stationary** if, for all t, s, τ it is

$$\mathbb{E}[X_t] = \mu \quad (102)$$

$$\mathbb{E}[X_t^2] < \infty \quad (103)$$

$$\gamma(s, t) = \gamma(s + \tau, t + \tau) \quad (104)$$

Note that (101) implies (104), which is therefore true for strongly stationary processes, while (104) and (102) imply that, for all s, t ,

$$\mathbb{E}[X_t^2] = \mathbb{E}[X_s^2] \quad (105)$$

and, together with (104),

$$\mathbb{E}[(X_t - \mu)^2] = \sigma^2 \quad (106)$$

Condition (104) can be written as

$$\gamma(s, t) = \gamma(s - t) \quad (107)$$

That is, the covariance of X_t and X_s depends only on their distance (with sign) in time, and not on the absolute time at which they are considered.

Definition 2.3. A stochastic process X_t is **white** if $\gamma(s, t) = \delta_{s,t}\sigma^2$

white process

Normally, for us, a white (stationary) process will be a process w_t with $\mu_t = 0$ (the average, being a constant, is not very interesting: we can always subtract it, study the process with zero average, and then add it back), and $\mathbb{E}[w_t^2] = \sigma^2$, we shall write it as $w \sim wN(0, \sigma^2)$.

That a process is white means, essentially, that the past gives us no information about the present value of w_t . This doesn't mean that we can't make a prediction on w_t : the distribution $wN(0, \sigma^2)$ may very well allow us to make good predictions. What it means is that it doesn't matter whether we know the values w_0, \dots, w_{t-1} or not, our predictions and their accuracy will not change. We can write:

$$P(x_t; t | x_0, \dots, x_{t-1}; 0, \dots, t-1) = P(x_t; t) \quad (108)$$

In a stochastic process, we have two ways of computing averages: we can compute the *ensemble average* $\langle X(t) \rangle$, that is, the average of the random variable $X(t)$, or the mean value along a trajectory

$$\bar{X} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt \quad (109)$$

A process is *ergodic* if the two coincide

$$\langle X \rangle = \bar{X} \quad (110)$$

Ergodicity is an important property for us: many times we shall be interested in the characteristics of the motion of one individual, but many of the equations that we shall use involve ensemble probabilities based on a whole population. Ergodicity allows us to switch from one to the other with impunity.

2.1 Gaussian processes

A stochastic process $X(t)$ is *Gaussian* with zero mean if $\mathbb{E}[X(t)] = \mu_t = 0$ and

$$P(x_i, t_i) = \sqrt{\frac{A_{ii}}{2\pi}} \exp\left(-\frac{1}{2}A_{ii}x_i^2\right) \quad (111)$$

($A_{ii} > 0$). The joint probability $P(x_1, t_1; \dots; x_n, t_n)$ then follows a multivariate Gaussian distribution

$$P(x_1, t_1; \dots; x_n, t_n) = \frac{\det(\mathbf{A})^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i,j=1}^n x_i A_{ij} x_j\right] \quad (112)$$

Where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric (strictly) positive definite. The matrix \mathbf{A} is a measure of the covariance between two variables of the Gaussian process

$$\mathbb{E}[X(s)X(t)] = \gamma(s, t) = (\mathbf{A}^{-1})_{st} \quad (113)$$

(this is true since we assume zero mean). A process is uncorrelated if \mathbf{A} is diagonal, then

$$\gamma(s, t) = A_{st}^{-1} \delta_{s,t} \quad (114)$$

and the process is white.

2.2 Wiener Processes

A *Wiener process* W is a process in which the variables $W(t)$ are real and with independent increments $W(t_2) - W(t_1)$ that follow a Gaussian distribution. That is, they define a conditional probability

$$P(w_2, t_2 | w_1, t_1) = \frac{1}{\sigma \sqrt{2\pi(t_2 - t_1)}} \exp \left[-\frac{(w_2 - w_1)^2}{2\sigma^2(t_2 - t_1)} \right] \quad (115)$$

from which the covariance can be computed

$$\begin{aligned} \langle (W(t_2) - \langle W \rangle)(W(t_1) - \langle W \rangle) \rangle &= \langle (W(t_2) - W(0))(W(t_1) - W(0)) \rangle \\ &= \int_{-\infty}^{\infty} (w_2 - w_0) dw_2 \int_{-\infty}^{\infty} dw_1 (w_1 - w_0) P(w_2, t_2; w_1, t_1) \\ &= \sigma^2 \min(t_1, t_2) + w_0^2 \end{aligned} \quad (116)$$

Note that this entails that a Wiener process is neither white nor stationary. From this we get

$$\langle W(t)^2 \rangle = \sigma^2 t + w_0^2 \quad (117)$$

Wiener processes are related to Gaussian processes, in particular to uncorrelated (white) Gaussian processes. Let $X(t)$ be a Gaussian process with $\langle X(t_1)X(t_2) \rangle = \sigma^2 \delta(t_2 - t_1)$, and define a new stochastic process as the integral of $X(t)$:

$$Y(t) = \int_0^t X(u) du \quad (118)$$

then

$$\begin{aligned} \langle Y(t_2)Y(t_1) \rangle &= \int_0^{t_2} du_2 \int_0^{t_1} du_1 \langle X(u_1)X(u_2) \rangle \\ &= \int_0^{t_2} du_2 \int_0^{t_1} du_1 \delta(u_2 - u_1) \end{aligned} \quad (119)$$

By the properties of the Dirac function

$$\int_0^{t_1} du_1 \delta(u_2 - u_1) = \begin{cases} 1 & 0 < u_2 < t_1 \\ 0 & \text{otherwise} \end{cases} \quad (120)$$

Then

$$\langle Y(t_2)Y(t_1) \rangle = \sigma^2 \min(t_2, t_1) \quad (121)$$

which coincides with (117) for $w_1 = 0$. That is, the integral of a Gaussian process is a Wiener process.

3 Markov Chains

When we deal with a stochastic process, what we often do is to observe the process until time t , and use this knowledge to try and predict what the process will do at time $t + 1$. In a white process, of course, we may just as well save the time and effort to observe: the past will give no indication about the future, none of our observations will help, and prediction based on $P(x; t)$ is the best we can do. But, in general, we can rely on the kindness of processes: the past will indeed tell us something about the future, so we are often interested in calculating

$$P(m_{t+1}; t + 1 | m_0, \dots, m_t; 0, \dots, t) \quad (122)$$

(we use the symbols m instead of x because of tradition: in this section, the range M will be finite or countable, and x is usually the name of a real variable). This determination is, in the most general case, very complex, since it may depend on all the values m_0, \dots, m_t , that is, the amount of information that we have to keep track of grows linearly with time without bound. Fortunately for us, many interesting processes have a nice characteristic: knowing just the immediate past is enough for the prediction to be as good as it gets, and it is not necessary to keep track of remote values. We refer to this as the **Markov Property**.

Markov property

Let $\Lambda = [\lambda_1, \dots, \lambda_n]'$ be a distribution that is, a vector for which $\lambda_m \geq 0$, $\sum_m \lambda_m = 1$.

Definition 3.1. A stochastic process X_t , $t \geq 0$, with values in M (M finite or countable) is *Markov*(λ, P) if

- i) $P(m; 0) = \lambda_m$;
- ii) $P(m_t; t | m_0, \dots, m_{t-1}; 0, \dots, t-1) = P(m_t; t | m_{t-1}; t-1)$

That is knowing the value of the process at time $t - 1$ gives us the same information about the value as t as it would knowing the whole history of the process. We define the **transition matrix** at time t $\mathbf{P}(t)$ as

$$\mathbf{P}(t)|_{n,m} = P(m; t | n; t-1) \quad (123)$$

The values $m \in M$ are referred to as the **states** of the process, and $P(m; t | n; t-1)$ are the **transition probabilities** at t . A process is **stationary** if the transition probabilities are independent of time, that is, for all m, n, t, τ ,

$$P(m; t | n; t-1) = P(m; t + \tau | n; t + \tau - 1) \quad (124)$$

In this case, \mathbf{P} is a constant matrix

$$\mathbf{P}_{n,m} \triangleq P(m|n) = P(m; t | n; t-1) \quad (125)$$

transition matrix

states

transition probabilities

stationary process

Note that $\sum_m P(m|n) = 1$ (from state n it is certain that we move to some state m), therefore

$$\sum_{m \in M} \mathbf{P}_{n,m} = 1 \quad (126)$$

for all n .

Theorem 3.1. *The process X_t , $t \geq 0$ is Markov(λ, P) iff*

$$P(m_0, \dots, m_t; 0, \dots, t) = \lambda_{m_0} \mathbf{P}_{m_0, m_1} \mathbf{P}_{m_1, m_2} \cdots \mathbf{P}_{m_{t-1}, m_t} \quad (127)$$

Proof. Suppose X_t is Markov(λ, P), then

$$\begin{aligned} P(m_0, \dots, m_t; 0, \dots, t) &= P(m_t; t | m_0, \dots, m_{t-1}; 0, \dots, t-1) P(m_0, \dots, m_{t-1}; 0, \dots, t-1) \\ &\stackrel{(M)}{=} P(m_t; t | m_{t-1}; t-1) P(m_0, \dots, m_{t-1}; 0, \dots, t-1) \\ &\quad \vdots \\ &= P(m_0; 0) P(m_1; 1 | m_0; 0) \cdots P(m_t; t | m_{t-1}; t-1) \\ &= \lambda_{m_0} \mathbf{P}_{m_0, m_1} \mathbf{P}_{m_1, m_2} \cdots \mathbf{P}_{m_{t-1}, m_t} \end{aligned} \quad (128)$$

where the equality (M) is the Markov property.

Suppose not that (127) holds. We sum over m_1, \dots, m_t obtaining

$$\sum_{m_1, \dots, m_t} P(m_0, \dots, m_t; 0, \dots, t) = P(m_0; 0) = \sum_{m_1, \dots, m_t} \lambda_{m_0} \mathbf{P}_{m_0, m_1} \mathbf{P}_{m_1, m_2} \cdots \mathbf{P}_{m_{t-1}, m_t} = \lambda_{m_0} \quad (129)$$

(the last equality is a consequence of (126)). This proves property i) of definition 3.1. Summing (127) on m_t yields

$$P(m_0, \dots, m_{t-1}; 0, \dots, t-1) = \lambda_{m_0} \mathbf{P}_{m_0, m_1} \mathbf{P}_{m_1, m_2} \cdots \mathbf{P}_{m_{t-2}, m_{t-1}} \quad (130)$$

Therefore

$$P(m_t; t | m_0, \dots, m_{t-1}; 0, \dots, t-1) = \frac{P(m_0, \dots, m_t; 0, \dots, t)}{P(m_0, \dots, m_{t-1}; 0, \dots, t-1)} = \mathbf{P}_{m_{t-1}, m_t} \quad (131)$$

which proves ii). \square

The one step transition matrix \mathbf{P} can easily be extended to t steps. As a convenient notation, if A is an event, let $P_m(A) \triangleq P(A | X_0 = m)$ so that, for example, $P_m(X_1 = n) = \mathbf{P}_{m,n}$. Given an initial distribution Λ , we have

$$P(X_1 = n) = \sum_{m \in M} \lambda_m P_m(X_1 = n) = \sum_{m \in M} \lambda_m \mathbf{P}_{m,n} \quad (132)$$

Similarly,

$$\begin{aligned} P_m(X_2 = n) &= \sum_{k \in M} P_m(X_1 = k, X_2 = n) = \sum_{k \in M} \mathbf{P}_{m,k} \mathbf{P}_{k,n} = \mathbf{P}^2|_{m,n} \\ P(X_2 = n) &= \sum_{m,k \in M} \lambda_m P_m(X_1 = k, X_2 = n) = \sum_{m,k \in M} \lambda_m \mathbf{P}_{m,k} \mathbf{P}_{k,n} = \Lambda' \mathbf{P}^2|_{m,n} \end{aligned} \quad (133)$$

Continuing we get

$$\begin{aligned} P_m(X_t = n) &= \mathbf{P}^t|_{m,n} \triangleq \mathbf{P}_{m,n}^{(t)} \\ P(X_t = n) &= \sum_{m \in M} \lambda_m P_m(X_t = n) = \Lambda' \mathbf{P}^t|_n \end{aligned} \quad (134)$$

It is easy to show that the following **Chapman-Kolmogorov** equation holds:

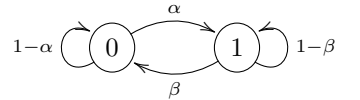
$$\mathbf{P}_{n,m}^{(t+s)} = \sum_{k \in M} \mathbf{P}_{n,k}^{(t)} \mathbf{P}_{k,m}^{(s)} \quad (135)$$

Therefore, if the chain is stationary, the t step transition matrix is simply \mathbf{P}^t , if the chain is not stationary, then

$$\mathbf{P}^{(t,t+\tau)} = \mathbf{P}_t \cdot \mathbf{P}_{t+1} \cdots \mathbf{P}_{t+\tau} \quad (136)$$

Example III:

The following diagram represents a simple two-state Markov chain



$$\mathbf{P} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \quad (137)$$

To compute $\mathbf{P}_{m,n}^{(t)}$ we use a trick that will be useful in many occasions, based on the eigenvalues of \mathbf{P} . These can easily be calculated, and they are

$$\lambda_1 = 1 \quad \lambda_2 = 1 - \alpha - \beta \quad (138)$$

The matrix \mathbf{P} can then be written as

$$\mathbf{P} = \mathbf{U} \begin{bmatrix} 1 & 0 \\ 0 & 1 - \alpha - \beta \end{bmatrix} \mathbf{U}^{-1} \quad (139)$$

and

$$\mathbf{P}^t = \mathbf{U} \begin{bmatrix} 1 & 0 \\ 0 & (1 - \alpha - \beta)^t \end{bmatrix} \mathbf{U}^{-1} \quad (140)$$

The elements of \mathbf{P}^t are linear combinations of the vectors $[1, 0]'$ and $[0, (1 - \alpha - \beta)^t]'$ so we can write, for example,

$$\mathbf{P}_{1,1}^{(t)} = A + B(1 - \alpha - \beta)^t \quad (141)$$

for some A and B . The initial conditions give us

$$\begin{aligned} \mathbf{P}_{1,1}^{(0)} &= 1 = A + B \\ \mathbf{P}_{1,1}^{(1)} &= 1 - \alpha = A + B(1 - \alpha - \beta) \end{aligned} \quad (142)$$

from which we derive

$$\begin{aligned} A &= \frac{\beta}{\alpha + \beta} \\ B &= \frac{\alpha}{\alpha + \beta} \end{aligned} \quad (143)$$

Repeating the argument for all the $\mathbf{P}_{m,n}^{(t)}$ we obtain

$$\mathbf{P}^t = \frac{1}{\alpha + \beta} \begin{bmatrix} \beta + \alpha(1 - \alpha - \beta)^t & \alpha - \alpha(1 - \alpha - \beta)^t \\ \beta - \beta(1 - \alpha - \beta)^t & \alpha + \beta(1 - \alpha - \beta)^t \end{bmatrix} \quad (144)$$

(end of example)

One useful property of Markov chains is that at any time we can "forget the past and start anew" taking as initial state the one in which we are placed.

Theorem 3.2. *Let X_t , $t \geq 0$ be Markov(λ, P). Then, conditioned on $X_t = m$, the sequence $X_{t+s}|_{s \geq 0}$ is Markov(δ_m, P), and is independent on the variables X_0, \dots, X_t .*

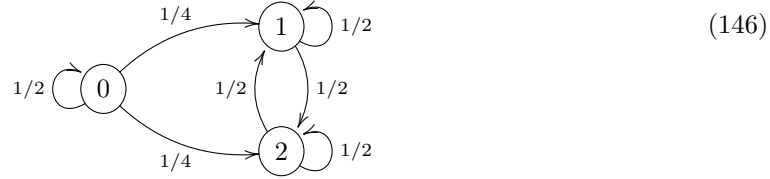
Here, $\delta_m = [0, \dots, 0, 1, 0, \dots, 0]'$, with the 1 in the m th position.

3.1 Class Structure

States have characteristics that distinguish them. Consider the following chain



It is clear that, once the chain enters the state 2, it will never leave it. In the following chain, the idea is the same but the situation is a bit more complex



In this case there is no state in which the chain remains forever, but once it has entered the set of states $\{1, 2\}$ it will never leave it, although it will keep jumping from one state to another. There is obviously something special about the set $\{1, 2\}$. In this section we shall study this and other characteristics of sets of states.

* * *

We say that n **leads to** m ($n \rightarrow m$) if $P_n(X_t = m) > 0$ for some $t > 0$. We say that n **communicates with** m ($n \leftrightarrow m$) if $n \rightarrow m$ and $m \rightarrow n$.

leads to

communicates with

Theorem 3.3. *Given two distinct states n and m , the following are equivalent:*

- i) $n \rightarrow m$
- ii) There are n_0, \dots, n_t with $n_0 = n$ and $n_t = m$ such that

$$\mathbf{P}_{n_0, n_1} \cdot \mathbf{P}_{n_1, n_2} \cdot \dots \cdot \mathbf{P}_{n_{t-1}, n_t} > 0 \quad (147)$$

- iii) $\mathbf{P}_{n, m}^{(t)} > 0$ for some $t \geq 1$.

Proof. Note that

$$\mathbf{P}_{n, m}^{(t)} \leq P[\exists t. X_t = m] \leq \sum_{s=0}^{\infty} \mathbf{P}_{n, m}^{(s)} \quad (148)$$

proving the equivalence of i) and ii). Also

$$\mathbf{P}_{n, m}^{(t)} = \sum_{n_0, \dots, n_t, n_0=n, n_t=m} \mathbf{P}_{n_0, n_1} \cdot \mathbf{P}_{n_1, n_2} \cdot \dots \cdot \mathbf{P}_{n_{t-1}, n_t} > 0 \quad (149)$$

proving the equivalence of ii) and iii). \square

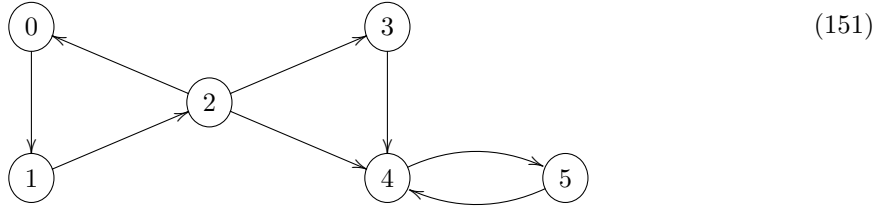
It is quite easy to see that \leftrightarrow is an equivalence relation and, as such, it partitions M into equivalence classes that we call **communicating classes**. Moreover, we say that $C \subseteq M$ is a **closed class** if

$$c \in C, c \rightarrow m \text{ Longrightarrow } m \in C \quad (150)$$

A closed class is a set of states that constitutes a trap from which there is no escape: once the chain has reached a state in C , it will forever stay in C . A state $m \in M$ is **absorbing** if $\{m\}$ is a closed class. If C is not closed, then it is **open**.

Example IV:

Given the diagram



The sets $\{0, 1, 2\}$, $\{3\}$ and $\{4, 5\}$ are communicating classes. $\{4, 5\}$ is closed, while $\{0, 1, 2\}$ and $\{3\}$ are open. (end of example)

A chain for which M is a single class is called **irreducible**.

3.2 Hitting probabilities and hitting mean time

Let $X_t|_{t \geq 0}$ a Markov chain taking values in M , with transition matrix \mathbf{P} . Given a subset $A \subseteq M$, the **first hitting time** of A is the random variable $H^A : \Omega \rightarrow \mathbb{N} \cup \{\infty\}$ given by

$$H^A(\omega) = \inf\{t \geq 0 \mid X_t(\omega) \in A\} \quad (152)$$

where, by convention, the infimum of the empty set is ∞ . The probability that starting from state m the chain ever hits A is

$$h_m^A = P(H^A < \infty) \quad (153)$$

When A is closed, h_m^A is called the **absorption probability** of A . The **mean hitting time** for reaching A is

absorption probability
mean hitting time

$$k_m^A = \mathbf{E}[H^A] = \sum_{t < \infty} t P_m(H^A = t) + \infty P_m(H^A = \infty) \quad (154)$$

Informally, we shall use sometimes the notation

$$\begin{aligned} h_m^A &\triangleq P_m(\text{hit } A) \\ k_m^A &\triangleq \mathbb{E}[t \text{ to } A] \end{aligned} \quad (155)$$

Also, we shall omit the superscript A when the set we are considering is obvious from the context.

Example V:

Consider the chain

$$\begin{array}{c} \text{1-p} \curvearrowright \textcircled{0} \xrightarrow{1-p^2} \textcircled{1} \curvearrowleft \text{1} \end{array} \quad \mathbf{P} = \begin{bmatrix} 1-p & p \\ 0 & 1 \end{bmatrix} \quad (156)$$

where 2 is an absorption state. The probability of absorption starting from 2 is clearly $h_2 = 1$, and

$$h_1 = \sum_t P(H^2 = t | X_0 = 1) = \sum_t (1-p)^{t-1} p = 1 \quad (157)$$

For the mean hitting time, clearly $k_2 = 0$, and

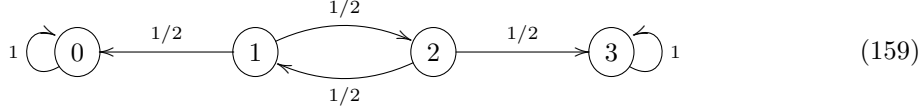
$$k_1 = \sum_t t P(H^2 = t | X_0 = 1) = \sum_t t (1-p)^{t-1} p = -p \frac{d}{dp} \sum_t (1-p)^t = \frac{1}{p} \quad (158)$$

(end of example)

The calculation done using the definition, as we have done in this example, can get very complex, but fortunately these quantities can be calculated as a solutions of a simple system of linear equations

Example VI:

Consider the chain



The chain has two absorbing states, 0 and 3. We start from 2, what is the probability of being absorbed in 3. if $h_m = P_m(\text{hit } 3)$ then, for example h_1 is equal to the probability of moving to 0 multiplied by the probability of hitting 3 starting from 0, plus the probability of moving to 2 multiplied by the probability of hitting 3 starting from 2. Reasoning in this way for all states, we have

$$\begin{aligned} h_0 &= 0 && \text{(This is so because 0 is absorbing)} \\ h_1 &= \frac{1}{2}h_0 + \frac{1}{2}h_2 \\ h_2 &= \frac{1}{2}h_1 + \frac{1}{2}h_3 \\ h_3 &= 1 \end{aligned} \tag{160}$$

which gives solutions $h_1 = 1/3$ and $h_2 = 2/3$. For the average time, the reasoning is similar, but this time each transition "costs" one time step, and the time from an absorbing state other than 3 is ∞ :

$$\begin{aligned} k_0 &= \infty \\ k_1 &= 1 + \frac{1}{2}k_0 + \frac{1}{2}k_2 \\ k_2 &= 1 + \frac{1}{2}k_1 + \frac{1}{2}k_3 \\ k_3 &= 0 \end{aligned} \tag{161}$$

which gives us $k_2 = \infty$ (if we start from 2, one third of the times we will end up in 0, from which the time to reach 3 is infinity, therefore $k_2 \leq \frac{1}{3}\infty + \frac{2}{3}A$, where A is a finite quantity.) On the other hand, if we want to know the average time to be absorbed either in 0 or 3, setting

$$k_m = \mathbb{E}[t \text{ to } \{0, 3\} | X_0 = m] \tag{162}$$

we have

$$\begin{aligned} k_0 &= 0 \\ k_1 &= 1 + \frac{1}{2}k_0 + \frac{1}{2}k_2 \\ k_2 &= 1 + \frac{1}{2}k_1 + \frac{1}{2}k_3 \\ k_3 &= 0 \end{aligned} \tag{163}$$

which gives $k_1 = k_2 = 2$.

(end of example)

The equations for the absorption probability of this example don't come out of nowhere, they are the example of a general property.

Theorem 3.4. *The vector of hitting probabilities of $A \subseteq M$, $h^A = [h_m^A | m \in M]'$ is the minimal non-negative solution of*

$$\begin{aligned} h_m^A &= 1 & m \in A \\ h_m^A &= \sum_{n \in M} \mathbf{P}_{m,n} h_n^A & m \notin A \end{aligned} \quad (164)$$

Proof. We first show that h_m^A is a solution of (164). If $X_0 = m$, $m \in A$, then $h_m^A = 1$, which satisfies the first of (164). If $X_0 \notin A$, then $H^A \geq 1$. By the Markov property,

$$\begin{aligned} h_m^A &= P_m(H^A < \infty) = \sum_{n \in M} P_m(H^A < \infty, X_1 = n) \\ &= \sum_{n \in M} P_m(H^A < \infty | X_1 = n) P_m(X_1 = n) \\ &\stackrel{(*)}{=} \sum_{n \in M} \mathbf{P}_{m,n} h_n^A \end{aligned} \quad (165)$$

where equality (*) derives from Theorem 3.2

$$P_m(H^A < \infty | X_1 = n) = P_n(H^A < \infty) = h_n^A \quad (166)$$

Assume now that $x = [x_m | m \in M]'$ is a solution of (164). If $m \in A$, then $x_m = 1 = h_m^A$. Suppose $m \notin A$, then

$$x_m = \sum_{n \in M} \mathbf{P}_{m,n} x_n = \sum_{n \in A} \mathbf{P}_{m,n} x_n + \sum_{n \notin A} \mathbf{P}_{m,n} x_n \quad (167)$$

Expanding the x_n , we have

$$\begin{aligned} x_m &= \sum_{n \in A} \mathbf{P}_{m,n} + \sum_{n \notin A} \mathbf{P}_{m,n} \left[\sum_{k \in A} \mathbf{P}_{n,k} x_k + \sum_{k \notin A} \mathbf{P}_{n,k} x_k \right] \\ &= \sum_{n \in A} \mathbf{P}_{m,n} + \sum_{n \notin A, k \in A} \mathbf{P}_{m,n} \mathbf{P}_{n,k} + \sum_{n \notin A, k \notin A} \mathbf{P}_{m,n} \mathbf{P}_{n,k} x_k \\ &= P_m(X_1 \in A) + P_m(X_1 \notin A, X_2 \in A) + \sum_{n \notin A, k \notin A} \mathbf{P}_{m,n} \mathbf{P}_{n,k} x_k \end{aligned} \quad (168)$$

Iterating we have

$$\begin{aligned} x_m &= P_m(X_1 \in A) + P_m(X_1 \notin A, X_2 \in A) + \cdots + P_m(X_1 \notin A, X_2 \in A, \dots, X_{t-1} \notin A, X_t \in A) \\ &\quad + \sum_{n_i \notin A} \mathbf{P}_{m,n_1} \mathbf{P}_{n_1,n_2} \cdots \mathbf{P}_{n_{t-1},n_t} x_{n_t} \\ &= P_m(H^A \leq t) + \sum_{n_i \notin A} \mathbf{P}_{m,n_1} \mathbf{P}_{n_1,n_2} \cdots \mathbf{P}_{n_{t-1},n_t} x_{n_t} \end{aligned} \quad (169)$$

Since $x_m \geq 0$, we have $x_m \geq P_m(H^A \leq t)$, therefore

$$x_m \geq \lim_{t \rightarrow \infty} P_m(H^A \leq t) = P_m(H^A < \infty) = h_m^A \quad (170)$$

□

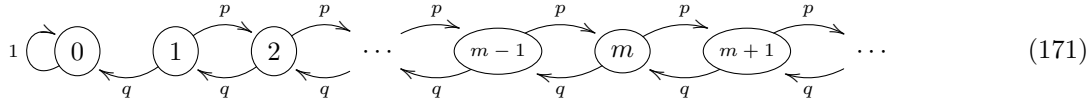
Note that, in (160), the equality $h_0 = 0$ cannot be derived from equations (164), which only yield $h_0 = h_0$. The condition $h_0 = 0$ derives from the minimality of the solution.

Example VII:

Gambler's ruin

You are a gambler playing against the house repeated rounds of a zero-sum game in which you have in each round a probability p of winning and a probability $q = 1 - p$ of losing. You gamble \$1 each round, and have an initial capital of m dollars. The house has an infinite supply of money, while you must stop playing when you go bust, that is, when you are left without money. If you play forever, what is the probability that you will end up being ruined?

The game is represented by a Markov chain, in which the state is the amount of money that you have.



Set $h_n = P_n(\text{hit } 0)$. The transition probabilities are

$$\begin{aligned} \mathbf{P}_{0,0} &= 1 \\ \mathbf{P}_{m,m-1} &= q \\ \mathbf{P}_{m,m+1} &= p \end{aligned} \quad (172)$$

therefore the probabilities that we seek are the solution of

$$\begin{aligned} h_0 &= 1 \\ h_m &= ph_{m-1} + qh_{m+1} \quad m = 1, 2, \dots \end{aligned} \quad (173)$$

If $p \neq q$, these equations have general solution

$$h_n = 1 - A + A \left(\frac{q}{p} \right)^n \quad (174)$$

and the minimal solution must be a distribution, that is, $0 \leq h_m \leq 1$. If $p < q$, the condition $h_m \leq 1$ forces $A = 0$, therefore $h_0 = 1$. If $p > q$, the minimal solution has A as large as possible consistent with $h_m \geq 0$, that is, $A = 1$, and $h_n = (q/p)^n$.

If $p = q = 1/2$, the recurrence has solution $h_n = 1 + Bn$ and the condition $h_n \leq 1$ forces $B = 0$, so $h_n = 1$: even if you find a fair casino³, the fact that you have a limited amount of money will, in the long run, bankrupt you. (end of example)

A theorem analogout to 3.4 holds for mean hitting times (we moit the proof).

Theorem 3.5. *The vector of mean hitting times of $A \subseteq M$, $k^A = [k_m^A | m \in M]'$ is the minimal non-negative solution of*

$$\begin{aligned} k_m^A &= 0 & m \in A \\ k_m^A &= 1 + \sum_{n \in M} \mathbf{P}_{m,n} k_n^A & m \notin A \end{aligned} \quad (175)$$

3.3 Recurrence and transience

Let us begin with an array of definitions, some of which we shall need only in the following (but, since they are all realted, we define them here and get it over with).

Let $X_t|_{t \geq 0}$ a Markov chain. Define

i) The **hitting time** of m is $H_m = \inf\{n \geq 0 | X_n = m\}$

hitting time

ii) The **first passage time** to m is $T_m = \inf\{n \geq 1 | X_n = m\}$

first passage time

Note that H_m and T_m differ only if $X_0 = m$.

iii) The **number of visits** to m is $V_m = \sum_{t=0}^{\infty} \chi_{X_t=m}$, where χ_A is the indicator function of a set A .

number of visits

iv) The **return probability** to m $f_m = P_m(T_m < \infty)$

return probability

³Which, of course, you won't: Las Vegas casinos are required by law to publish the odds that you have in all teh games, and in general $p/q \approx 0.98$.

v) The **mean return time** to m $\mu_m = \mathbb{E}_m[T_m]$

mean return time

vi) The **number of visits** to m **before** τ

number of visits before τ

$$V_m(\tau) = \sum_{t=0}^{\tau-1} \chi_{X_t=m} \quad (176)$$

vii) The **number of visits** to m **before the first return** to n $V_m^n = V_m(T_n)$

number of visits before first return

viii) The **mean number of visits** to m **between successive visits** to n $\gamma_m n = \mathbb{E}_m[V_m^n]$

mean number of visits

Notice that if $X_0 = m$, then $V_m^m = 1$ and therefore $\gamma + m^m = 1$.

We say that m is **recurrent** if $P_m(V_m = \infty) = 1$, otherwise we say that m is **transient**. A state is recurrent if the chain keeps returning to it, while in the case of a transient state, the chain will at some point leave it never to return.

recurrent and transient states

Lemma 3.1. For all $\tau \geq 0$, $P_m(V_m \geq \tau + 1) = (f_m)^\tau$.

Proof. The proof is by induction over τ . The lemma is true for $\tau = 0$ by definition. Assume that it is true for $\tau - 1$. Then

$$\begin{aligned} P_m(V_m \geq \tau + 1) &= P_m(V_m \geq \tau + 1 | V_m \geq \tau) P_m(V_m \geq \tau) \\ &= P_m(T_m < \infty) (f_m)^{\tau-1} \\ &= (f_m)^\tau \end{aligned} \quad (177)$$

□

From this, it is easy to show that the following is true

Theorem 3.6.

$$m \text{ recurrent} \Leftrightarrow f_m = 1 \Leftrightarrow \sum_{t=0}^{\infty} \mathbf{P}_{m,m}^{(t)} = \infty \quad (178)$$

$$m \text{ transient} \Leftrightarrow f_m < 1 \Leftrightarrow \sum_{t=0}^{\infty} \mathbf{P}_{m,m}^{(t)} < \infty \quad (179)$$

Recurrence and transience have been defined for a state, but they are really class properties

Theorem 3.7. *Take C be a communicating class, then either all the states in C are transient or they are all recurrent.*

Proof. Take $n, m \in C$ and suppose that m is transient. There are $t, s > 0$ with $\mathbf{P}_{m,m}^{(t)} > 0$ and $\mathbf{P}_{m,m}^{(s)} > 0$ by definition of C . By Chapman-Kolmogorov, for all $\tau > 0$,

$$\mathbf{P}_{m,m}^{(t+s+\tau)} \geq \mathbf{P}_{m,m}^{(t)} \mathbf{P}_{m,m}^{(\tau)} \mathbf{P}_{m,m}^{(s)} \quad (180)$$

so

$$\sum_{\tau=0}^{\infty} \mathbf{P}_{m,m}^{(\tau)} \leq \frac{1}{\mathbf{P}_{m,m}^{(t)} \mathbf{P}_{m,m}^{(s)}} \sum_{\tau=0}^{\infty} \mathbf{P}_{m,m}^{(t+s+\tau)} < \infty \quad (181)$$

□

Because of this theorem, we can speak of a transient or a recurrent class.

Theorem 3.8. *Every recurrent class is closed.*

Proof. If C is not closed, then there are $m \in C$, $n \notin C$ with $m \rightarrow n$ and $\tau \geq 1$ such that

$$P_m(X_\tau = n) > 1 \quad (182)$$

However, $n \nrightarrow m$, so

$$P_m(V_m = \infty | X_\tau = n) = 0 \quad (183)$$

Therefore

$$P_m(V_m = \infty) = \sum_k P_m(V_m = \infty | X_\tau = k) < 1 \quad (184)$$

So m is not recurrent, and neither is C

□

Theorem 3.9. *Every finite closed class is recurrent.*

Proof. Let C be such a class. Pick any initial distribution on C . Then $\sum_{m \in C} V_m = \infty$ (since C is closed). Since C is finite, some state must be visited infinitely often, so

$$1 = P\left[\bigcup_{m \in C} \{V_m = \infty\}\right] \leq \sum_{m \in C} P(V_m = \infty) \quad (185)$$

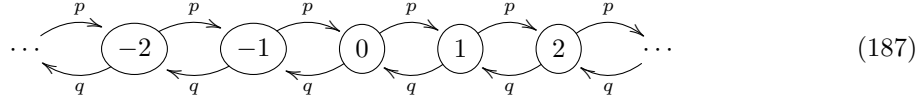
So, for some m ,

$$0 < P(V_m = \infty) = P(H_m < \infty)P_m(V_m = \infty) \quad (186)$$

But $P_m(V_m = \infty)$ can only be 0 or 1, so it must be 1. Thus m is recurrent, and so is C . \square

Example VIII:

We shall talk about random walks later on, but this is a good example to warm up. A random walk on \mathbb{Z} is one in which the walker jumps from spot m to $m + 1$ with probability p , and from m to $m - 1$ with probability q :



The walk starts from 0, and the question that we pose is: is 0 recurrent? We shall apply the first condition of theorem 3.6 to show that it is. If we start at 0, we can never be back at 0 after an odd number of steps, so $\mathbf{P}_{0,0}^{(2n+1)} = 0$ for all n . Consider now a sequence of $2n$ steps from 0 to 0. Independently of how many twists and turns we make, this will consist of n steps forward, and n steps backward and will have probability $p^n q^n$. We identify a trajectory of length $2n$ by deciding where we do the n steps forward, so there are $\binom{2n}{n}$ different trajectories, that is

$$\mathbf{P}_{0,0}^{(2n)} = \binom{2n}{n} p^n q^n \quad (188)$$

Using stirling approximation

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \quad n \rightarrow \infty \quad (189)$$

therefore, we have

$$\mathbf{P}_{0,0}^{(2n)} = \frac{(2n)!}{(n!)^2} (pq)^n \sim \frac{(4pq)^n}{\sqrt{2\pi n} \sqrt{\frac{n}{2}}} \quad (190)$$

If $p = q = 1/2$, we have $4pq = 1$ so, for N large enough and $n > N$,

$$\mathbf{P}_{0,0}^{(2n)} \geq \frac{1}{2\sqrt{2\pi n}} \quad (191)$$

and

$$\sum_{n=0}^{\infty} \mathbf{P}_{0,0}^{(2n)} \geq \frac{1}{2\sqrt{2\pi}} \sum_{n=N}^{\infty} \frac{1}{\sqrt{n}} = \infty \quad (192)$$

which shows that the walk is recurrent. If $p \neq q$ then $4pq = r < 1$ so, reasoning in a similar way,

$$\sum_n \mathbf{P}_{0,0}^{(2n)} \leq k \sum_n r^n < \infty \quad (193)$$

which shows that the walk is transient.

(end of example)

3.4 Invariant distribution

Over long periods of time, the occupation probability of the states of a Markov chain may settle to a stable distribution. As we already defined, a measure $\Lambda = [\lambda_m | m \in M]'$ is a vector with $\lambda_m \geq 0$ for all m . If $\sum \lambda_m = 1$, the vector is a **distribution**. An **invariant measure** is a measure such that

invariant measure

$$\Lambda' \mathbf{P} = \Lambda' \quad \lambda_m \geq 0 \quad (194)$$

An **invariant distribution** is defined analogously:

invariant distribution

$$\Pi' \mathbf{P} = \Pi' \quad \pi_m \geq 0 \quad \sum_{m \in M} \pi_m = 1 \quad (195)$$

If Λ is an invariant measure and $\sum_m \lambda_m < \infty$, then

$$\pi_m = \frac{\lambda_m}{\sum_{n \in M} \lambda_n} \quad (196)$$

is an invariant distribution. This is always possible if M is finite.

If M is finite, an invariant distribution always exists. We know that $\sum_n \mathbf{P}_{m,n} = 1$, which implies that for the vector $\mathbf{1} = [1, 1, \dots, 1]'$ it holds

$$\mathbf{P} \mathbf{1} = \mathbf{1} \quad (197)$$

Therefore \mathbf{P} has an eigenvalue equal to 1, and $\mathbf{1}$ is a right eigenvector of this eigenvalue. Standard linear algebra arguments show that a left eigenvalue also exists and, by the Frobenius-Perron theorem, all its components are positive. Normalizing this left eigenvector we obtain (195).

The importance of invariant distributions is related to the long-term behavior of a chain. In particular, if the Markov chain settles down to a distribution, this is an invariant one.

Theorem 3.10. *Let M be finite and, for some $m \in M$ and all $n \in M$,*

$$\lim_{t \rightarrow \infty} \mathbf{P}_{m,n}^{(t)} = \pi_n \quad (198)$$

Then $\pi = [\pi_m | m \in M]'$ is an invariant distribution

Proof. We have

$$\sum_{m \in M} \pi_m = \sum_{m \in M} \lim_{t \rightarrow \infty} \mathbf{P}_{m,n}^{(t)} = \lim_{t \rightarrow \infty} \sum_{m \in M} \mathbf{P}_{m,n}^{(t)} = 1 \quad (199)$$

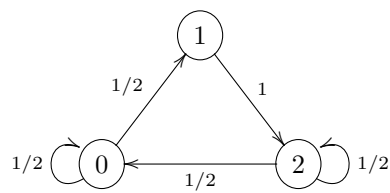
and

$$\pi_n = \lim_{t \rightarrow \infty} \mathbf{P}_{m,n}^{(t)} = \lim_{t \rightarrow \infty} \sum_{k \in M} \mathbf{P}_{m,k}^{(t-1)} \mathbf{P}_{k,n} = \sum_{k \in M} \lim_{t \rightarrow \infty} \mathbf{P}_{m,k}^{(t-1)} \mathbf{P}_{k,n} = \sum_{k \in M} \pi_k \mathbf{P}_{k,n} \quad (200)$$

that is $\pi' = \pi' \mathbf{P}$. \square

Example IX:

Consider the chain



$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \quad (201)$$

We determine the stationary distribution by solving the equilibrium equation $\pi' = \pi' \mathbf{P}$:

$$[\pi_1, \pi_2, \pi_3] = [\pi_1, \pi_2, \pi_3] \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \quad (202)$$

that is:

$$\begin{aligned} \pi_1 &= \frac{1}{2} \pi_3 \\ \pi_2 &= \pi_1 + \frac{1}{2} \pi_2 \\ \pi_3 &= \frac{1}{2} \pi_2 + \frac{1}{2} \pi_3 \end{aligned} \quad (203)$$

These equations not independent and therefore under-constrained, but we can determine a unique solution by adding the normalization condition

$$\pi_1 + \pi_2 + \pi_3 = 1 \quad (204)$$

obtaining $\pi = \frac{1}{5}[1, 2, 2]$ (end of example)

Sometimes it is more convenient to use the **balance equations**

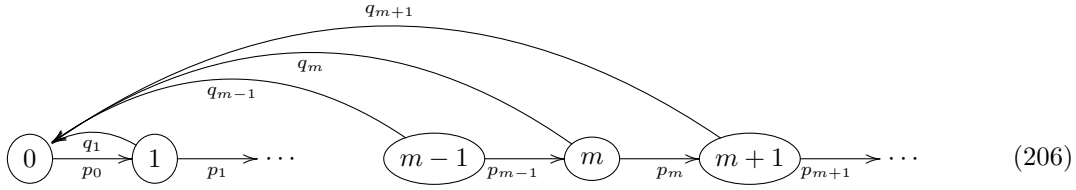
Balance equations

$$\pi_m \mathbf{P}_{m,n} = \pi_n \mathbf{P}_{n,m} \quad \text{for all } n, m \in M \quad (205)$$

If M is infinite, a stationary distribution doesn't necessarily exist.

Example X:

Consider the chain



with

$$\mathbf{P}_{m,m+1} = p_m \quad \mathbf{P}_{m,0} = q_m = 1 - p_m \quad (207)$$

Then

$$\pi_0 = \sum_{m=0}^{\infty} \pi_m q_m \implies p_0 \pi_0 = \sum_{m=1}^{\infty} \pi_m q_m \quad (208)$$

$$\pi_m = \pi_{m-1} p_{m-1} \quad m \geq 1 \quad (209)$$

Define the sequence r_m , $m \geq 0$ as

$$r_0 = 1 \quad (210)$$

$$r_m = p_0 \cdots p_{m-1} = r_{m-1} p_{m-1}$$

and choose the p_m in such a way that

$$r \triangleq \prod_{m=0}^{\infty} p_m > 0 \quad (211)$$

The definition of r_m implies that $\pi_m = r_m \pi_0$ and

$$\begin{aligned}
 \pi_0 &= \sum_m \pi_m q_m = \lim_{n \rightarrow \infty} \sum_{m=0}^n \pi_m q_m \\
 &= \lim_{n \rightarrow \infty} \sum_{m=0}^n (r_m - r_{m-1}) \pi_0 \\
 &= \lim_{n \rightarrow \infty} (r_0 - r_{n+1}) \pi_0 \\
 &= (1 - r) \pi_0
 \end{aligned} \tag{212}$$

and, since $r > 0$, the only solution is $\pi_0 = 0$, from which, via (209), we have $\pi = 0$, therefore π is not a distribution. (end of example)

Even if a stationary distribution exists, it needs not be unique.

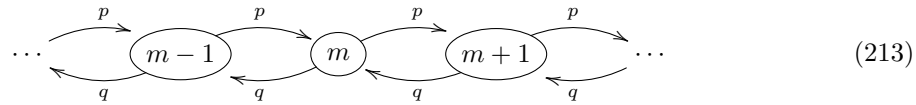
Example XI:

Let $\mathbf{P} = \mathbf{I}$ (the identity matrix), then any distribution is stationary. (end of example)

In this example, the chain is not irreducible, but the same is true for irreducible chains.

Example XII:

Consider the chain



Then $\lambda_m = 1$ and $\lambda_m = (p/q)^m$ are both invariant measures. The measures are different (viz., the solution is not unique) if $p \neq q$, that is, if the chain is transient. (end of example)

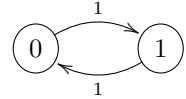
On the other hand, the following lemma guarantees unicity under certain conditions.

Lemma 3.2. *If \mathbf{P} is irreducible and recurrent, then $\gamma_k = [\gamma_m^k | m \in M]'$ satisfies $\gamma_k' = \gamma_k' \mathbf{P}$, and the solution is unique*

If the limit exists, then it must be a stationary distribution. But the limit may fail to exist.

Example XIII:

Consider the chain



$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (214)$$

Then $\mathbf{P}^{2n} = \mathbf{I}$ and $\mathbf{P}^{2n+1} = \mathbf{P}$ so the limit of \mathbf{P}^n does not exist. Intuitively, the chain jumps from one state to another without ever setting down to a stable solution. (end of example)

A state m is **aperiodic** if there is a t_0 such that $\mathbf{P}_{m,m}^{(t)} > 0$ for all $t > t_0$. The states in the previous example are not aperiodic since, for example, if the chain has $X_0 = 0$, then $\mathbf{P}_{0,0}^{(2t+1)} = 0$ for all t . A state that is not aperiodic is called **periodic** and it can be shown that there is a $\tau > 0$ such that $\mathbf{P}_{m,m}^{(t)} > 0$ if t is a multiple of τ and $\mathbf{P}_{m,m}^{(t)} = 0$ otherwise; the value τ is called the period of the state.

aperiodic state

periodic state

Lemma 3.3. *Let \mathbf{P} be irreducible, and have an aperiodic state m . Then, for all $n, k \in M$, $\mathbf{P}_{n,k}^{(t)} > 0$ for t sufficiently large. In particular, all states are aperiodic.*

We define a state m **positive recurrent** if $\mu_m < \infty$. It can be shown that if \mathbf{P} is irreducible, then if m is recurrent, all the states are.

positive recurrent state

A chain is **regular** if for some t , all the elements of $\mathbf{P}^{(t)}$ are positive.

regular chain

A chain is **ergodic** if it is aperiodic, irreducible, and positive recurrent.

ergodic chain

Theorem 3.11. *Let \mathbf{P} be the transition matrix of an ergodic Markov chain with invariant distribution π . Then, for any initial distribution,*

$$\lim_{t \rightarrow \infty} P(X_t = m) = \pi_m \quad (215)$$

In particular, for any initial distribution and any $m, n \in M$,

$$\lim_{t \rightarrow \infty} \mathbf{P}_{n,m}^{(t)} = \pi_m \quad (216)$$

In this case, we have:

$$\begin{aligned}\mu_m &= \frac{1}{\pi_m} \\ \gamma_m^n &= \frac{\pi_m}{\pi_n} \\ P\left[\lim_{t \rightarrow \infty} \frac{V_t}{t} = \pi_m\right] &= 1\end{aligned}\tag{217}$$

4 State Space models

One useful characteristic of Markov models is that the prediction of the behavior at the present time depends only on the behavior in the immediate past. In the case of a finite measurement space M , this allows us, as in the previous section, to say quite a few things about the behavior of the chain. Alas, life is not always that simple. Many natural problems and technical devices can be best modeled through a more complex **autoregressive** model (AR) with finite memory, that is, by a process of the type

$$y(t) = a_1 y(t-1) + \cdots + a_n y(t-n) + w(t) \quad (218)$$

where w is a white process and n is the **order** of the model.

autoregressive model

Before we continue, we shall do a slight change of notation. From this point on, in order to simplify some of the notation that follows, we shall adopt the common system theory convention of indicating time dependency as a subscript: $x_t \triangleq x(t)$. This will not cause confusion with subscript such as a_n in the expression above, as the meaning of the subscript (time dependency or component) is normally clear from the context. If we want to indicate the k th component of a vector x_t , we shall indicate it as $x_{t,k}$ (this will almost never be necessary, so the relatively heavy notation will not constitute a problem).

Now that this is settled, we note that the model depends on n past samples so it is not, in principle, a Markov model. However, since the output y_t depends on a finite and fixed number of past sample, we can modify it by defining a **state** that encodes all the past information that we need for predicting the value at t . If we rewrite (218) as

state

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-n+1} \end{bmatrix} = \begin{bmatrix} a_1, & \cdots & \cdots, & a_n \\ 1, & \cdots & \cdots, & 0 \\ 0, & 1, & \cdots, & 0 \\ \vdots & & & \vdots \\ 0 & \cdots, & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-n} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} w_t \quad (219)$$

then we can define the vector $x_t = [y_t, \dots, y_{t-n+1}]$ obtaining a one-step equation

$$\begin{aligned} x_t &= \mathbf{A}x_{t-1} + \mathbf{D}w_t \\ z_t &= [1, 0, \dots, 0]x_t \triangleq \mathbf{C}x_t \end{aligned} \quad (220)$$

where \mathbf{A} and \mathbf{B} are defined as in (219). Note that in this model we have a one-step evolution of the state, but we maintain the output of (218) by creating an **observation** process z_t . In many cases we assume that the state x_t is hidden, and that the only information we have about the system is the process z_t . In this case, of course, it is quite easy to reconstruct the state from the values z_t : one just has to store the last n outputs, and these will give us the value of the state x_t .

observation

We can generalize the model by transforming the matrices and the vectors, which is our example have a very specific form, into general matrix and vectors. In order to make the model more general, we assume that we have an input u_t , which we are able to control and that observation z_t is also subject to noise. We arrive in this way to the standard **state-space model**:

state-space model

$$\begin{aligned}x_{t+1} &= \mathbf{A}x_t + \mathbf{B}u_t + w_t \\ z_t &= \mathbf{C}x_t + v_t\end{aligned}\tag{221}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $z \in \mathbb{R}^q$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{C} \in \mathbb{R}^{q \times n}$; $w_t \in \mathbb{R}^n$ $v_t \in \mathbb{R}^q$ are white processes.

This is not the most general model possible; we could assume that \mathbf{A} , \mathbf{B} , and \mathbf{C} depend on time. This model, however, is enough for our purposes here. As a matter of further simplification we shall often assume that $m = q = 1$.

In this model, the state contains all the relevant information about the system, the input u_t allows us to intervene on the state, and the output z_t allows us to observe the effects of changes of state. The noises w_t and v_t are there to remind us that life is tough, that we don't always have all the information we would like to have, and that there are things outside of our control that cause interferences.

Let us for the moment ignore the noises, and assume that we have perfect knowledge of the system. We still don't have direct access to the state. All we can do is to observe z_t trying to deduce what the state might be and use u_t trying to make it do what we want. Because of these limitations, we might not be able to accomplish our goals, that is, we might not be able to know everything about the state simply by observing z_t or to make it do what we want simply by manipulating u_t . These observations lead to two important concepts in system theory: **reachability** and **observability**.

Intuitively, a state x_t is **reachable** (at time t) if we can start from an arbitrary state (say, $x_0 = 0$ —the actual value is not terribly important) and devise an input $[u_0, \dots, u_{t-1}]$ that leads us to that state. More formally, considering a state x_t , we can write

reachability

$$\begin{aligned}x_t &= \mathbf{A}x_{t-1} + \mathbf{B}u_{t-1} \\ &= \mathbf{A}(\mathbf{A}x_{t-2} + \mathbf{B}u_{t-2}) + \mathbf{B}u_{t-1} \\ &= \mathbf{A}^2x_{t-2} + \mathbf{A}\mathbf{B}u_{t-2} + \mathbf{B}u_{t-1} \\ &= \mathbf{A}^2(\mathbf{A}x_{t-3} + \mathbf{B}u_{t-3}) + \mathbf{A}\mathbf{B}u_{t-2} + \mathbf{B}u_{t-1} \\ &= \mathbf{A}^3x_{t-3} + \mathbf{A}^2\mathbf{B}u_{t-3} + \mathbf{A}\mathbf{B}u_{t-2} + \mathbf{B}u_{t-1} \\ &\vdots \\ &= \mathbf{A}^t x_0 + \mathbf{A}^{t-1}\mathbf{B}u_0 + \dots + \mathbf{A}\mathbf{B}u_{t-2} + \mathbf{B}u_{t-1}\end{aligned}\tag{222}$$

Based on this relation, we can give the following definition.

Definition 4.1. A state x_t is reachable (at time t) if, given $x_0 = 0$, there is a sequence of inputs u_0, \dots, u_{t-1} such that

$$x_t = [\mathbf{B} | \mathbf{AB} | \dots | \mathbf{A}^t \mathbf{B}] \begin{bmatrix} u_{t-1} \\ u_{t-2} \\ \vdots \\ u_0 \end{bmatrix} \quad (223)$$

The matrix

$$\mathcal{R}^+(t) = [\mathbf{B} | \mathbf{AB} | \dots | \mathbf{A}^t \mathbf{B}] \quad (224)$$

is called the **reachability matrix** at time t . Let $\chi^+(t)$ be the set of states reachable at time t , then

reachability matrix

$$\chi^+(t) = \text{Im}[\mathcal{R}^+(t)] \quad (225)$$

\mathcal{R}^t is an $n \times t$ matrix, therefore it has rank at most n , which entails

$$\{0\} = \chi^+(0) \subseteq \chi^+(1) \subseteq \dots \subseteq \chi^+(n) = \chi^+(n+1) = \dots \quad (226)$$

Set $\mathcal{R}^+ \triangleq \mathcal{R}^+(n)$. If $\text{rank}(\mathcal{R}^+) = n$, then $\chi^+(n) = \mathbb{R}^n$, that is, any state can be reached in at most n step. In this case, we say that the system itself is reachable.

reachable system

Example XIV:

To see an example of unreachable system, divide $x(t) \in \mathbb{R}^n$ into $x'(t) \in \mathbb{R}^p$ and $x''(t) \in \mathbb{R}^m$, with $n = m + p$, and assume that the system has the following structure.

$$\begin{bmatrix} x'_t \\ - \\ x''_t \end{bmatrix} = \begin{bmatrix} \mathbf{P} & | & \mathbf{Q} \\ - & - & - \\ 0 & | & \mathbf{R} \end{bmatrix} \begin{bmatrix} x'_{t-1} \\ - \\ x''_{t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{T} \\ - \\ 0 \end{bmatrix} u_{t-1} \quad (227)$$

which can be written as

$$\begin{aligned} x'_t &= \mathbf{P}x'_{t-1} + \mathbf{Q}x''_{t-1} + \mathbf{T}u_{t-1} \\ x''_t &= \mathbf{R}x''_{t-1} \end{aligned} \quad (228)$$

The input u acts only on x' , and x' has no effect on x'' , so any evolution that starts with $x_0 = 0$ will remain with $x''_t = 0$ for all t , independently of the input u . Any state in which $x'' \neq 0$ is unreachable.

To see how this works out with the reachability matrix, consider that

$$\mathbf{A}^2 = \begin{bmatrix} \mathbf{P} & | & \mathbf{Q} \\ - & - & - \\ 0 & | & \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{P} & | & \mathbf{Q} \\ - & - & - \\ 0 & | & \mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^2 & | & \mathbf{PQ} + \mathbf{QR} \\ - & - & - \\ 0 & | & \mathbf{R}^2 \end{bmatrix} \quad (229)$$

Iterating, we can see that all \mathbf{A}^t have the same structure:

$$\mathbf{A}^t = \left[\begin{array}{c|c} \mathbf{P}^t & \mathbf{K}_t \\ \hline 0 & \mathbf{R}^t \end{array} \right] \quad (230)$$

for some \mathbf{K}_t . Therefore

$$\mathbf{A}^t \mathbf{B} = \left[\begin{array}{c} \mathbf{P}^t \mathbf{T} \\ \hline 0 \end{array} \right] \quad (231)$$

and

$$\mathcal{R}^+ = \left[\begin{array}{ccc|c} \mathbf{T} & \mathbf{P}\mathbf{T} & & \mathbf{P}^{n-1}\mathbf{T} \\ \hline - & - & \cdots & - \\ 0 & 0 & & 0 \end{array} \right] \quad (232)$$

which has rank at most p

(end of example)

Similar considerations hold for observability. In this case, given a state x_0 , we let the system evolve freely (setting $u = 0$) for a time t , and by observing the outputs z_t , we want to reconstruct the state x_0 . If this is possible, x_0 is **observable**. In this case we can write

$$\begin{aligned} z_0 &= \mathbf{C}x_0 \\ z_1 &= \mathbf{C}x_1 = \mathbf{C}\mathbf{A}x_0 \\ &\vdots \\ z_{t-1} &= \mathbf{C}x_{t-1} = \cdots = \mathbf{C}\mathbf{A}^{t-1}x_0 \end{aligned} \quad (233)$$

We define, analogously to the previous case, the **observability matrix**

observability matrix

$$\mathcal{O}^+(t) = \left[\begin{array}{c} \mathbf{C} \\ - \\ \mathbf{C}\mathbf{A} \\ - \\ \vdots \\ - \\ \mathbf{C}\mathbf{A}^{t-1} \end{array} \right] \quad (234)$$

and, as before, we say that a system is **observable** if $\mathcal{O}^+ \triangleq \mathcal{O}^+(n)$, and $\text{rank} \mathcal{O}^+ = n$.

observable system

Example XV:

As a parallel to the previous case, the system

$$\begin{bmatrix} x'_t \\ - \\ x''_t \end{bmatrix} = \begin{bmatrix} \mathbf{P} & | & \mathbf{Q} \\ - & - & - \\ 0 & | & \mathbf{R} \end{bmatrix} \begin{bmatrix} x'_{t-1} \\ - \\ x''_{t-1} \end{bmatrix} \quad (235)$$

$$z_t = [\mathbf{T} \quad | \quad 0]x_t$$

has $\text{rank}(\mathcal{O}^+) = p$, and is not observable (x'' has no effect, direct or indirect, on z). (end of example)

This kind of systems is used mainly to model physical systems for the purpose of control. This can work only if the system is observable (we can observe what's going on and get the complete picture of the state) and reachable (we can use the inputs to modify the state as we wish). In this situation, lack of observability or reachability is the mark of a poor model, one that doesn't serve the purpose for which it was designed—it is generally a sign that the model has to be re-done. From our point of view, we can always assume that the systems that we consider are reachable and observable.

4.1 The Kalman Filter

Let us consider the presence of noise in the system. In one common scenario, the matrices \mathbf{A} , \mathbf{B} , \mathbf{C} are known and, by observing the output of the system, we need to know the state x_t at all times. In the absence of noise, the matrix \mathcal{O}^+ allows us this estimation, albeit with the delay of n steps (we need up to n observations in order to know x_0). If there is noise, the situation is more complicated. Here we present a classical solution to the problem: the **Kalman Filter**. In the following, the input u_t , which is controlled by us and therefore is always known without noise is not essential, so we shall assume, for the moment, that $u_t \equiv 0$. Consider therefore the system

$$\begin{aligned} x_{t+1} &= \mathbf{A}x_t + w(t) \\ z_t &= \mathbf{C}x_t + v_t \end{aligned} \quad (236)$$

where w_t and v_t are Gaussian white noises with zero mean (if the mean is $\mu \neq 0$ we can simply model it as a system with a constant input μ and a zero-mean noise) and covariance

$$\begin{aligned} \mathbf{Q} &= \mathbb{E}[w_t w'_t] \\ \mathbf{R} &= \mathbb{E}[v_t v'_t] \end{aligned} \quad (237)$$

The arguments that we present do not require the noise to be Gaussian (they do require it to be stationary) but the solution that we find is optimal only in the case of Gaussian noise. Our purpose

is to find an estimate \hat{x}_t based on the prior knowledge of the system and the observed output so as to minimize the square error

$$\mathbf{P}_t = \mathbb{E}[e_t e_t'] = \mathbb{E}[(\hat{x}_t - x_t)(\hat{x}_t - x_t)'] \quad (238)$$

Specifically, we are interested in minimizing the trace $T[\mathbf{P}_t]$, which corresponds to the component-wise mean square error

$$T[\mathbf{P}_t] = \mathbb{E}\left[\sum_k (\hat{x}_{t,k} - x_{t,k})^2\right] \quad (239)$$

Assume that we have a prior estimate of x_t , \bar{x}_t . We create \hat{x}_t by correcting for the prediction error between the output predicted through \bar{x}_t and the output that we observed:

$$\hat{x}_t = \bar{x}_t + K_t(z_t - \mathbf{C}\bar{x}_t) \quad (240)$$

K_t is called the **Kalman gain**, and it is the element that we have to determine in order to determine to minimize the error. The stochastic process

kalman gain

$$i_t = z_t - \mathbf{C}\bar{x}_t \quad (241)$$

is called the **innovation process**: it represents the part of the observable behavior of the system which is "new" respect to our prediction: the unpredictable and unexpected. Inserting the second of (236) into (240), we have

innovation process

$$\hat{x}_t = \bar{x}_t + K_t(\mathbf{C}x_t + v_t - \mathbf{C}\bar{x}_t) \quad (242)$$

and

$$\begin{aligned} e_t &= x_t - \hat{x}_t \\ &= x_t - \bar{x}_t - K_t(\mathbf{C}(x_t - \bar{x}_t) + v_t) \\ &= (\mathbf{I} - K_t\mathbf{C})(x_t - \bar{x}_t) - K_tv_t \end{aligned} \quad (243)$$

The error $(x_t - \bar{x}_t)$ does not depend on the observations at time t , therefore it is uncorrelated with v_t :

$$\mathbb{E}[(x_t - \bar{x}_t)v_t] = 0 \quad (244)$$

We have therefore

$$\begin{aligned} \mathbf{P}_t &= \mathbb{E}[e_t e_t'] \\ &= (\mathbf{I} - K_t\mathbf{C})\mathbb{E}[(x_t - \bar{x}_t)(x_t - \bar{x}_t)'](\mathbf{I} - K_t\mathbf{C})' + K_t\mathbb{E}[v_t v_t']K_t' \end{aligned} \quad (245)$$

The value

$$\bar{\mathbf{P}}_t = \mathbb{E}[(x_t - \bar{x}_t)(x_t - \bar{x}_t)'] \quad (246)$$

is the prior estimate of \mathbf{P} . From this and the second of (237) we have

$$\mathbf{P}_t = (\mathbf{I} - K_t\mathbf{C})\bar{\mathbf{P}}_t(\mathbf{I} - K_t\mathbf{C})' + K_t\mathbf{R}K_t' \quad (247)$$

Expanding \mathbf{P}_t we have

$$\begin{aligned}\mathbf{P}_t &= \bar{\mathbf{P}}_t(\mathbf{I} - K_t\mathbf{C})' - K_t\mathbf{C}\bar{\mathbf{P}}_t(\mathbf{I} - K_t\mathbf{C})' + K_t\mathbf{R}K_t' \\ &= \bar{\mathbf{P}}_t - \bar{\mathbf{P}}_t(K_t\mathbf{C})' - K_t\mathbf{C}\bar{\mathbf{P}}_t + K_t\mathbf{C}\bar{\mathbf{P}}_t(K_t\mathbf{C})' + K_t\mathbf{R}K_t' \\ &= \bar{\mathbf{P}}_t - \bar{\mathbf{P}}_t(K_t\mathbf{C})' - K_t\mathbf{C}\bar{\mathbf{P}}_t + K_t(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R})K_t'\end{aligned}\quad (248)$$

The trace is invariant to transposition ($T[\mathbf{A}] = T[\mathbf{A}']$), so we have

$$T[\mathbf{P}_t] = T[\bar{\mathbf{P}}_t] = 2T[K_t\mathbf{C}\bar{\mathbf{P}}_t] + T[K_t(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R})K_t'] \quad (249)$$

We minimize it by setting its first derivative to zero:

$$\frac{d}{dK_t}T[\mathbf{P}_t] = -2(\mathbf{C}\bar{\mathbf{P}}_t) + 2K_t(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R}) \quad (250)$$

which gives

$$K_t = \bar{\mathbf{P}}_t\mathbf{C}'(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R})^{-1} \quad (251)$$

With this Kalman gain, we can compute the error variance

$$\begin{aligned}\mathbf{P}_t &= \bar{\mathbf{P}}_t - K_t\mathbf{C}\bar{\mathbf{P}}_t - \bar{\mathbf{P}}_t(K_t\mathbf{C})' + K_t(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R})K_t' \\ &= \bar{\mathbf{P}}_t - K_t\mathbf{C}\bar{\mathbf{P}}_t - \bar{\mathbf{P}}_t(K_t\mathbf{C})' + \bar{\mathbf{P}}_t\mathbf{C}'K_t' \\ &= (\mathbf{I} - K_t\mathbf{C})\bar{\mathbf{P}}_t\end{aligned}\quad (252)$$

With these equations, we can update our estimate of \bar{x}_t and our error \mathbf{P}_t once we have computed the priors \bar{x}_t and $\bar{\mathbf{P}}_t$. We get the priors from our estimated at the previous time:

$$\begin{aligned}\bar{x}_t &= \mathbf{A}\bar{x}_{t-1} \\ \bar{\mathbf{P}}_t &= \mathbb{E}[e_t e_t'] = \mathbb{E}[(\mathbf{A}e_{t-1} + w_{t-1})(\mathbf{A}e_{t-1} + w_{t-1})'] \\ &\stackrel{(*)}{=} \mathbb{E}[\mathbf{A}e_{t-1}(\mathbf{A}e_{t-1})'] + \mathbb{E}[w_{t-1}w_{t-1}'] \\ &= \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}' + \mathbf{Q}\end{aligned}\quad (253)$$

where the equality (*) is because e_t and w_t are uncorrelated. So, suppose that we have our priors \bar{x}_t and $\bar{\mathbf{P}}_t$. One step of the Kalman filter consists of the following:

Compute the Kalman Gain	$K_t = \bar{\mathbf{P}}_t\mathbf{C}'(\mathbf{C}\bar{\mathbf{P}}_t\mathbf{C}' + \mathbf{R})^{-1}$
Update the estimate	$\hat{x}_t = \bar{x}_t + K_t(z_t - \mathbf{C}\bar{x}_t)$
Update the covariance	$\mathbf{P}_t = (\mathbf{I} - K_t\mathbf{C})\bar{\mathbf{P}}_t$
Compute the priors	$\bar{x}_{t+1} = \mathbf{A}\bar{x}_t$
	$\bar{\mathbf{P}}_{t+1} = \mathbf{A}\mathbf{P}_t\mathbf{A}' + \mathbf{Q}$