

Albert-Ludwigs-Universität Freiburg

nonofficial lecture notes:

Classical Complex Systems

JProf. David Gross

winter semester 2014/15

scripted by
Kai von Prillwitz

Contents

0	Motivation	3
1	Dynamical Systems	4
1.1	Discrete systems	5
1.1.1	Introduction of the tent map	6
1.1.2	Bernoulli shift and binary numbers	7
1.1.3	Stability of orbits, bifurcations, Lyapunov exponents	10
1.1.4	Invariant measure	14
1.1.5	Feigenbaum universality	18
1.1.6	Fractal dimension	23
2	Stochastic Processes	28
2.1	Probability theory	29
2.1.1	Conditional probabilities	29
2.1.2	Random variables	30
2.1.3	Cumulative distribution function, probability distribution and probability density	30
2.1.4	Two variable distributions	31
2.1.5	Expectation value	32
2.2	Discrete time and discrete space processes	34
2.2.1	Markov chains	35
2.2.2	Properties of the transition matrix	38
2.2.3	Stationary and limit distributions	39
2.2.4	Time averages	44
2.2.5	Recurrence time	47
2.2.6	Reversible Markov chains	51
2.2.7	Markov chain Monte Carlo simulation	53
2.2.8	Random walks on graphs	56
2.3	Continuous time and discrete space Markov chains	57
2.3.1	Gillespie algorithm	62

0 Motivation

Consider the following simple model for the population dynamics of some species:

$$x_{t+1} = r x_t (1 - x_t), \quad x_t \in [0, 1], \quad t \in \mathbb{N} \quad (0.1)$$

This model is known as the *logistic map*. When x_t is the (normalized) population at some time step t then the logistic map can be used to calculate the population after one additional time step. The factor $r x_t$ can be understood as the offspring, e.g. for $r = 2$ each individual will have (on average) two offsprings. With this term alone the population would grow exponentially. This is prevented by the $1 - x_t$ term which can roughly be understood as starvation due to overpopulation, e.g. if a predator population grows too large there will not be enough prey to feed all of them. The rate r is bounded between one and four. For $r < 1$ the population would shrink exponentially even without the $1 - x_t$ term. For $r > 4$ one could reach populations $x_{t+1} > 1$. The graph of the logistic map for $r = 3.35$ is shown in Fig. 1.

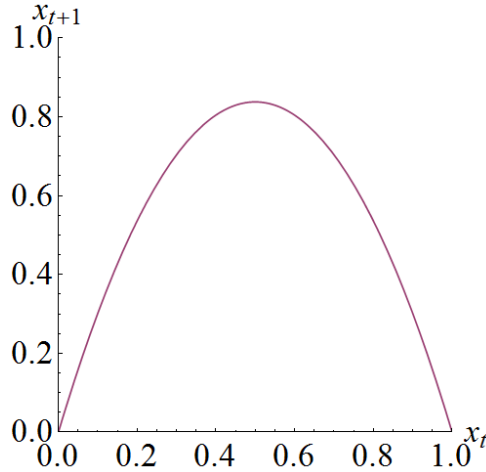


Figure 1: Logistic map for $r = 3.35$. The two competing terms $r x_t$ (offspring) and $1 - x_t$ (starvation) lead to a downward parabola. The maximum is obtained for $x_t = 0.5$. The parameter r only stretches the parabola in the up-down direction.

Phenomenology

The long-term qualitative behaviour of a population following the logistic map seems to depend only on the rate r , *not* on the initial population. We further observe:

- Fixed points that seem to attract all initial populations for $r < 3$ (Fig. 2a)
- Low period orbits for $3 < r < 3.59$ (Fig. 2b)
- “Chaotic” behaviour for $r = 4$ (Fig. 2c)
- In the chaotic regime tiny variations of the initial population get quickly amplified. A given interval of populations seems “to be visited evenly”.

- Bifurcations: For example at $r = 3$ the stable fixed point becomes unstable while a new stable period-2 orbit emerges. In general, the period of the stable orbit is doubled (see Fig. 2d).

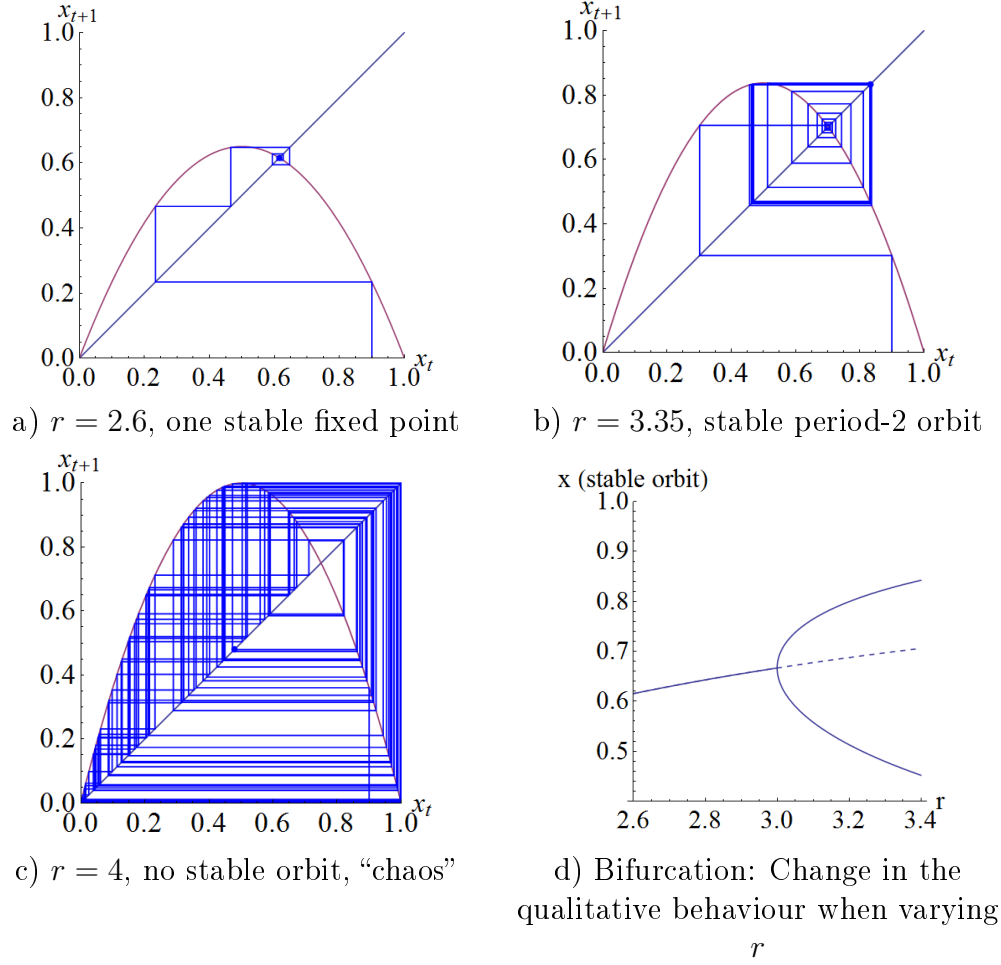


Figure 2: a) to c) show the long-term behaviour of the logistic map for different values of r . The blue path shows the stepwise progress of the population. In a) the population is driven to the stable fixed point marked by the intersection between the parabola and the identity line. In b) the population first approaches the fixed point but is then attracted by a period-2 orbit. In c) no regularity can be observed. The population seems to visit all possible values. The transition from the behaviour of a) to b) can be understood with the bifurcation depicted in d). (Dotted) solid lines represent (un)stable orbits.

1 Dynamical Systems

Definition A *dynamical system* is described by:

- a *state space* \mathcal{M} (sometimes: *phase space*)

- a family Φ^t of transformations:

$$\begin{aligned} \Phi^t : \mathcal{M} &\rightarrow \mathcal{M} && \text{where } t \in \mathbb{N} \text{ (discrete case)} \\ &&& \text{or } t \in \mathbb{R}_+ \text{ (continuous case)} \\ \text{s.t.} \quad &1) \Phi^0(x) = x \\ &2) \Phi^t \circ \Phi^s(x) = \Phi^{t+s}(x) \end{aligned}$$

Some common cases are:

1. Discrete systems:

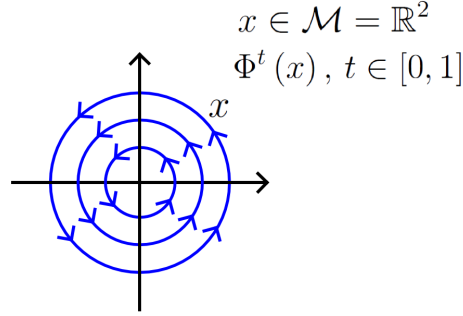
$t \in \mathbb{N}$, also called *maps*, completely specified by $\Phi^1 =: f$, example: logistic map from section 0.

2. Flows:

Continuous-time systems defined by first order differential equation

$$\partial_t \Phi^t(x) = F(x),$$

also completely defined by their vector field. Example: one dimensional harmonic oscillator:



3. Hamiltonian flows:

Flows that originate from classical mechanics. The harmonic oscillator is a representative of this special class of flows.

1.1 Discrete systems

A discrete system is completely specified by its map $\Phi^1 =: f$. The time evolution is generated by repeated application of this map. The sequence of x values obtained this way is called *orbit*:

- Let $x \in \mathcal{M}$. The orbit of x is denoted by:

$$\begin{aligned} &x, f(x), f \circ f(x), f \circ f \circ f(x), \dots \\ = &x_0, x_1, x_2, x_3, \dots \end{aligned}$$

- An orbit has *period* p if $x_0 = x_p$.
- The period is called a *proper period* if x_0, \dots, x_{p-1} are all different.

The goal of this section is to understand the behaviour of the logistic map that we observed in section 0 and to realize that this behaviour is generic for a large group of similar maps. To understand the general features of the logistic map, we first investigate the *tent map*.

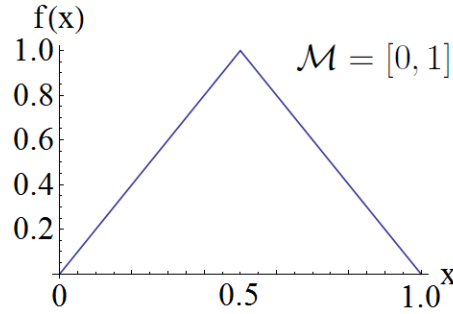
1.1.1 Introduction of the tent map

There are three equivalent ways to define the tent map.

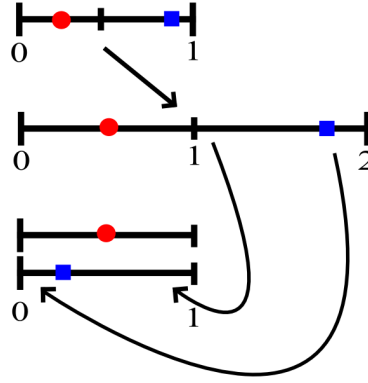
Function The tent map can be defined by its function:

$$\begin{aligned} f(x) &= 1 - 2|x - 1/2| \\ &= \begin{cases} 2x & \text{for } 0 \leq x \leq \frac{1}{2} \\ 2 - 2x & \text{for } \frac{1}{2} \leq x \leq 1 \end{cases} \end{aligned}$$

Graph The graph of the tent map also explains its name:



Stretch and fold A graphical instruction to obtain $f(x)$ for arbitrary x . First, the interval $[0, 1]$ is stretched homogeneously by a factor of 2. The $[1, 2]$ part of the new interval is then mapped back to $[0, 1]$:



We now want to find periodic orbits of the tent map. To this end, recall the definition of a periodic orbit: An orbit has *period* p if $x_0 = x_p = f^p(x_0)$, i.e. we return to the initial point after p applications of the map. Hence, a period-1 orbit is simply a fixed point of the map f . In general, a point lying on a period- n orbit is a fixed point of f^n (the n -fold application of f). It is therefore a good idea to take a look at the graph of several iterates of the map, see Fig. 3.

We see that the graph of the n -fold application of the tent map consists of 2^{n-1} tents each of width 2^{-n+1} . The minima (maxima) lie exactly at even (odd) multiples of 2^{-n} . This observations lead to the following

side remark: f^n will map any multiple of 2^{-n} to either 0 or 1
 $\Rightarrow f^{n+1}(k \cdot 2^{-n}) = 0$

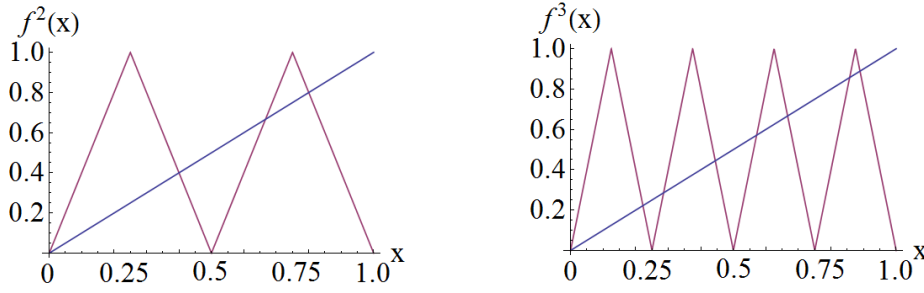


Figure 3: 2- and 3-fold application of the tent map. The number of tents doubles at each step. Each tent is intersected twice by the identity line leading to 2^n fixed points for the n -fold application of the map.

\Rightarrow This causes a problem for digital computers since they can only work with binary numbers (and not with symbolic expressions like fractions). Applying the tent map several times to any initial value will eventually lead to constant 0. Without this issue the tent map would be a good and simple random number generator. Slight changes to the tent map can correct the problem.

Coming back to the original problem of finding periodic orbits of the tent map, one can make the following observation and draw the corresponding conclusions:

- f^n has 2^n fixed points (two for each of the 2^{n-1} tents; including $x = 0$).
- There are 2^n points that lie on a period- n orbit.
- Only countably many points lie on a periodic orbit.
- The set of points with non periodic orbit is open and dense (i.e. it is much “larger” than the set of points with periodic orbit, similar to the relation between irrational and rational numbers).

The last conclusion means that if we draw a (true) random number in $[0, 1]$ as initial condition, we end up on a non periodic orbit. However, this does not necessarily imply chaotic behavior. Points could also lie in the *basin of attraction* of a periodic orbit (as we have seen for the logistic map for small values of r in Fig. 2):

$$x_0, x_1, x_2, x_3, x_4, x_5 = x_3, x_6 = x_4, x_7 = x_3, \dots$$

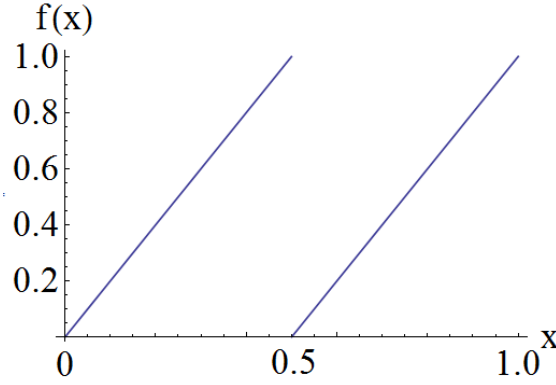
To get a proper understanding of the non periodic orbits we switch to an even simpler but strongly related system, the Bernoulli shift.

1.1.2 Bernoulli shift and binary numbers

As for the tent map there are several representations of the Bernoulli shift. The plain definition reads:

$$f(x) = (2x) \bmod 1, \quad x \in [0, 1]$$

A value x is first multiplied by two, then everything in front of the decimal point is dropped, i.e. set to zero. For $x \in [0, 1/2]$ the Bernoulli shift is the same as the tent map ($f(x) = 2x$). For $x \in [1/2, 1]$ one simply gets a shifted copy of this line, which leads to the following graph:



As for the tent map, the number of line segments doubles at each iteration of the map. Also, each of those lines will be intersected once by the identity line, leading to the same number of fixed points (2^n) for the n_{th} iteration. This behaviour suggests that results for the Bernoulli shift, concerning the behaviour of the non periodic orbits, should qualitatively also be true for the tent map.

To get another, maybe less straightforward, representation one can map the interval $[0, 1]$ to the unit circle in the complex plane: $c(x) = e^{i2\pi x}$. Due to the 2π -periodicity of the imaginary exponential function, the modulo operation of the Bernoulli shift is already naturally included. One then finds the property:

$$c(f(x)) = c((2x) \bmod 1) = e^{i2\pi[(2x) \bmod 1]} = e^{i4\pi x} = c(x)^2$$

We will not pay more attention to this representation, but go on to the very useful representation via binary numbers which will prove helpful to understand the properties of the non periodic orbits. For a better understanding we shall first make a recap of number representation in the decimal and binary system.

Recap: Decimal and binary number representation

First, consider $x \in \mathbb{N}$. The general decimal representation reads:

$$x = \sum_{k=0}^{\lfloor \log_{10}(x) \rfloor} d_k 10^k, \text{ e.g. } 17 = 1 \cdot 10^1 + 7 \cdot 10^0$$

The binary representation is completely analog:

$$x = \sum_{k=0}^{\lfloor \log_2(x) \rfloor} b_k 2^k, \text{ e.g. } 17 = 2^4 + 2^0 = 10001$$

For decimals, i.e. $x \in [0, 1)$, it is formally the same, we just need negative exponents:

$$\begin{aligned} x &= \sum_{k=1}^{\infty} d_{-k} 10^{-k} \\ &= \sum_{k=1}^{\infty} b_{-k} 2^{-k} \end{aligned}$$

The sums go to infinity since the termination of the sum is not guaranteed. Here are some facts about the binary expansion of decimals:

- With n binary digits one can exactly represent all multiples of 2^{-n} .
- As $n \rightarrow \infty$ this set becomes dense in the interval $[0, 1]$.
 \Rightarrow every $x \in [0, 1]$ can be arbitrary well approximated
 \Rightarrow get all numbers as $n \rightarrow \infty$
- The binary expansion of x terminates iff $\exists n$ s.t. $2^n \cdot x$ is an integer.
- The binary expansion of x eventually becomes periodic iff $x \in \mathbb{Q}$ (termination is a special case of periodicity as it can be considered an infinite repetition of zeros)
- The binary expansion never becomes periodic iff x is irrational

We can finally come back to the Bernoulli shift.

Binary representation of the Bernoulli shift

Recall the action of the map $f(x) = (2x) \bmod 1$: First, there is the multiplication by two. In the binary representation this means to shift the whole binary expansion one to the left. Second, everything in front of the decimal point is set to zero:

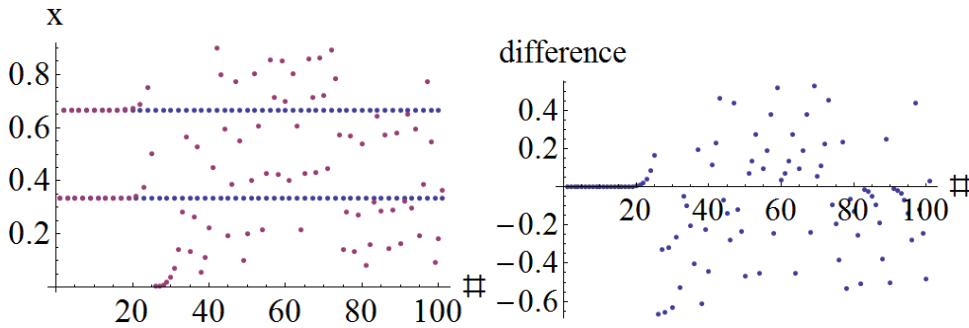
$$\begin{aligned} x &= 0.b_{-1}b_{-2}b_{-3}\dots \\ \rightarrow 2x &= b_{-1}.b_{-2}b_{-3}b_{-4}\dots \\ \rightarrow f(x) &= 0.b_{-2}b_{-3}b_{-4}\dots \end{aligned}$$

If we want x to be on a period- n orbit x must have a period- n binary expansion, e.g. for $n = 2$: $x = 0.b_{-1}b_{-2}b_{-1}b_{-2}\dots$. Hence:

- There are $2^n - 1$ numbers in $[0, 1)$ that lie on an orbit of period n (we identify $0 \cong 1$ here, without this identification there would be 2^n numbers).
- x lies on an orbit that will eventually become periodic iff $x \in \mathbb{Q}$ (i.e. if the (binary) expansion terminates or becomes periodic).
 \rightarrow these are only countably many numbers
 \rightarrow "most" x generate no such orbit at all.

There is one other behaviour of orbits that we did not investigate yet. Even if we do not start on a periodic orbit and the orbit does not eventually become periodic, we might still get closer and closer to a periodic orbit, or in other words, get attracted by it. To state the question in a different way: How stable are periodic orbits (subject to small perturbations)? Is the perturbed orbit driven back to the periodic orbit or does the disturbance grow more and more? First, we shall give a graphical example for the Bernoulli shift, then, in the next section, this question will be treated in a general way with application to the Bernoulli shift, tent map and finally also the logistic map.

The period-2 orbit of the Bernoulli shift consist of the two points $x_0 = 1/3$ and $x_1 = 2/3$. If we start with x_0 and introduce a small perturbation $x'_0 = 1/3 + 10^{-8}$ then it is plausible that the binary expansions of x_0 and x'_0 are the same at the first few positions after the decimal point, but eventually start to differ unpredictably. So we expect the orbit of x'_0 to stay very close to the orbit of x_0 for the first few iterations but then to become completely independent of it. This behaviour can indeed be observed in the following two plots.



The first plot shows the orbits of $x_0 = 1/3$ (blue dots) and $x'_0 = 1/3 + 10^{-8}$ (red dots). The second plot shows their difference. This strong (exponential) sensitivity to the initial condition is one characterization of chaos.

1.1.3 Stability of orbits, bifurcations, Lyapunov exponents

Let f be a differentiable map at x_0 and its orbit x_0, x_1, x_2, \dots . Assume further that we have a small perturbation t of the initial condition x_0 . Then the first order perturbation after applying the map once is given by:

$$f(x_0 + t) = f(x_0) + f'(x_0)t + \mathcal{O}(t^2)$$

If we apply the map a second time, we find

$$f(f(x_0 + t)) = f(f(x_0)) + \underbrace{f'(f(x_0))}_{f'(x_1)} f'(x_0)t + \mathcal{O}(t^2)$$

by using the chain rule and $f(x_0) = x_1$. Continuing this procedure, we find for the perturbation after k iterations:

$$f^k(x_0 + t) = f^k(x_0) + f'(x_{k-1})f'(x_{k-2})\dots f'(x_0)t + \mathcal{O}(t^2)$$

In the particular case that x_0 lies on a period- n orbit (i.e. $f^n(x_0) = x_0$) one gets:

$$f^n(x_0 + t) = x_0 + \prod_{i=0}^{n-1} f'(x_i)t + \mathcal{O}(t^2)$$

The expression $\partial_t f^n(x_0 + t) = \prod_{i=0}^{n-1} f'(x_i)$ is simply the product of the derivatives at all values x_i on the periodic orbit. It is therefore independent of the starting point, i.e. instead of x_0 we could also start at any other x_i (on the orbit). The expression can thus be considered a property of the orbit rather than of the particular initial condition. The expression is also important enough to get its own definition:

$$\mu_n := \prod_{i=0}^{n-1} f'(x_i)$$

If we now apply the map not only n times but $k \cdot n$ times, i.e. if we go around the total orbit k times, we get exactly in the same way as before:

$$f^{k \cdot n}(x_0 + t) = x_0 + \mu_n^k t + \mathcal{O}(t^2)$$

where we used once again the properties of the periodic orbit, namely $f^{k \cdot n}(x_0) = x_0$ and that we gather one factor μ_n for each of the k turns around the orbit. We can now make the following statement:

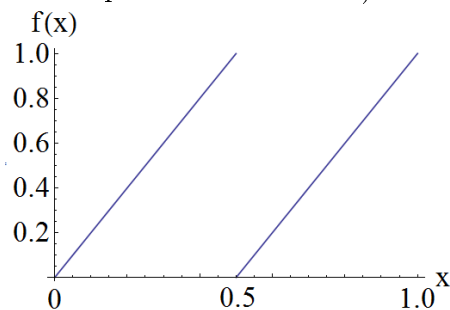
$$\text{If } |\mu_n| \begin{cases} > 1 \\ < 1 \end{cases} \rightarrow \begin{cases} \text{perturbation is amplified exponentially} \\ \text{perturbation is damped exponentially} \end{cases} \rightarrow \begin{cases} \text{orbit is unstable} \\ \text{orbit is stable} \end{cases}$$

One can also rewrite the above expression as

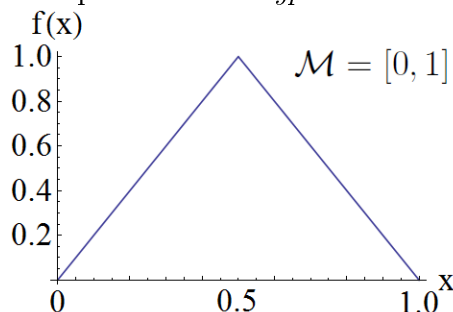
$$f^{k \cdot n}(x_0 + t) = x_0 + e^{(\ln \mu_n) \cdot k} t + \mathcal{O}(t^2)$$

where $\ln \mu_n$ is called the *Lyapunov exponent* of f^n . We can finally answer questions about the stability of orbits of the Bernoulli shift, the tent map and the logistic map.

Bernoulli shift $f(x) = (2x) \bmod 1 \rightarrow \mu_n = 2^n > 1$ for *all* orbits of period n (since the derivative is 2 for any of the n points on the orbit) \rightarrow orbits are not stable!



Tent map $f(x) = 1 - 2|x - 1/2| \rightarrow |\mu_n| = 2^n > 1$ (the derivative is either 2 or -2) \rightarrow *all* orbits are unstable. Such maps are called *hyperbolic*.



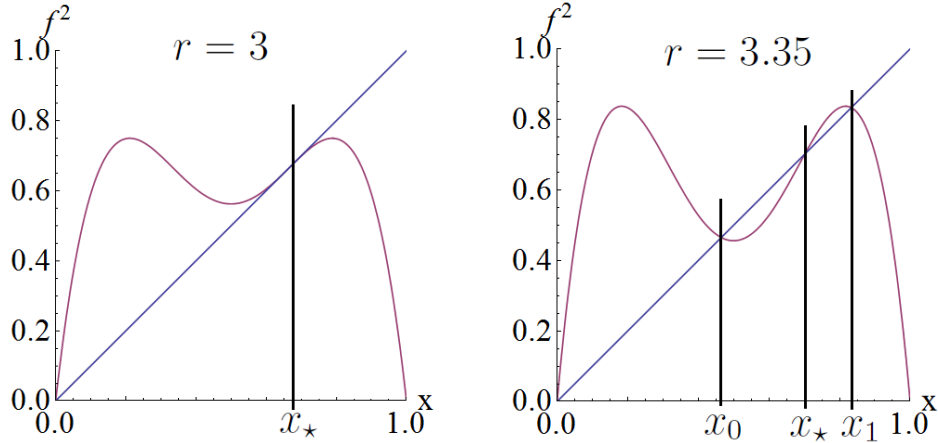
Logistic map $f(x) = rx(1-x)$. For the logistic map matters are more involved. First, we have the free parameter r on which the behaviour will depend. Second, to calculate μ_n we have to explicitly find the periodic orbit since the derivative is not just constant as for the previous two maps. We shall do this for the period-1 orbit (i.e. the fixed point of f). For the period-2 orbit we shall make a graphical investigation of the slopes at the fixed points of f^2 to determine the stability.

Period-1 orbit The fixed point equation for f is $rx(1-x) = x$ with the two solutions $x_{\star'} = 0$ and $x_{\star} = 1 - 1/r$. The derivative of the logistic map is $f'(x) = r(1-2x)$. Hence,

- for $x_{\star'} = 0$: $\mu_1 = f'(0) = r \rightarrow$ only stable for $|r| < 1$. In the interesting regime $r \in (1, 4]$ this fixed point is unstable.
- for $x_{\star} = 1 - \frac{1}{r}$: $\mu_1 = f'(x_{\star}) = 2 - r$. For $r \in (1, 4]$ we find:

$$x_{\star} = 1 - \frac{1}{r} \text{ is } \begin{cases} \text{stable for} & r \in (1, 3) \\ \text{unstable for} & r > 3 \end{cases}$$

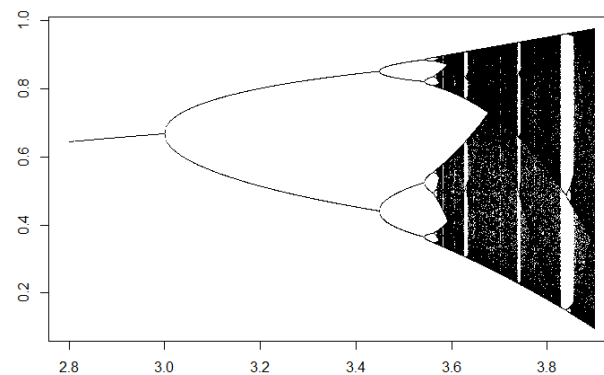
Period-2 orbit Look at f^2 for $r \gtrsim 3$. For the known fixed point x_{\star} we find $(f^2)'(x_{\star}) = f'(f(x_{\star})) \cdot f'(x_{\star}) = f'(x_{\star})^2 = (2-r)^2$. As seen before for only one iteration of f , this fixed point becomes unstable for $r > 3$. Taking a look at the graph of f^2 we see that two new fixed points arise.



- At $r = 3$ the slope of f^2 at x_{\star} becomes parallel to the identity line: $(f^2)'(x_{\star}) = 1$.
- For $r \gtrsim 3$, since the graph of f^2 intersects the identity line at x_{\star} from below, it has necessarily a slope larger than one $\rightarrow x_{\star}$ becomes unstable (as known from the previous calculation).
- At the same time two new fixed points appear, x_0 and x_1 , one to the left, one to the right of x_{\star} .
- At $r = 3$ all three points coincide: $x_{\star} = x_0 = x_1$. Therefore $(f^2)'(x_0) = (f^2)'(x_1) = 1$
- As r increases the points separate and the slopes at x_0 and x_1 decrease.
 \rightarrow two *stable* fixed points of f^2
 \rightarrow *stable* period-2 orbit of f

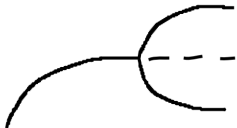
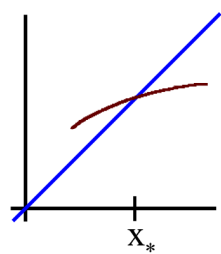
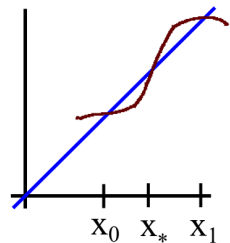
- When r is increased further the slopes at x_0 and x_1 become negative (as can already be seen in the plot for $r = 3.35$) and finally also smaller than -1 . At this value of r the two points become unstable and a stable period-4 orbit emerges. This procedure goes on and on.

The observed behaviour is a fairly general mechanism for turning a stable period-1 orbit into an unstable period-1 and a stable period-2 orbit (or in general period- 2^k into period- 2^{k+1}). This is also called a *bifurcation*, or to be more precise a *period doubling bifurcation*. A picture of a single bifurcation was already shown in section 0 and will reappear in the next subsection. The following plot shows a large segment of the total period doubling cascade of the logistic map. Only the stable orbits are shown.



Bifurcation: general classes

In addition to the period doubling bifurcation, there are two other frequently appearing types of bifurcations. Table 1 gives a quick overview over the three types. All three have in common that there is some scaling parameter that is adjusted over a critical value at which the bifurcation occurs. The period doubling bifurcation was investigated in detail in the previous section. The inverse period doubling bifurcation looks similar but the roles of stable and unstable orbits are reversed, i.e. an unstable orbit turns into a stable orbit and an unstable orbit with doubled period. For the tangent bifurcation no fixed point exists before the bifurcation. As the scaling parameter is increased two fixed points emerge, one stable, one unstable.

Type	bifurcation scheme	graph before bifurcation	graph after bifurcation
period doubling		 $f'(x_*) < 1$	 $f'(x_*) > 1$ $ f'(x_{0,1}) < 1$


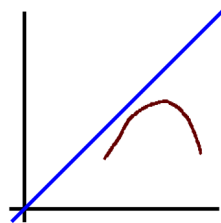
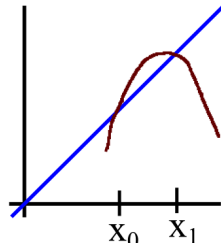

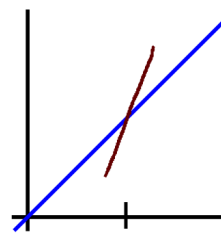
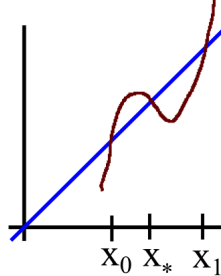
tangent		 no fixed point	 $x_0 \quad x_1$ $f'(x_0) > 1$ $ f'(x_1) < 1$
inverse period doubling		 x_* $f'(x_*) > 1$	 $x_0 \quad x_* \quad x_1$ $ f'(x_*) < 1$ $f'(x_{0,1}) > 1$

Table 1: Overview over three frequent types of bifurcations. The bifurcation schemes (second column) assume that the scaling parameter is increased from left to right. A more detailed plot of the period doubling bifurcation can be found in Fig. 2d. (Dotted) solid lines represent (un)stable orbits. The bifurcations can also occur in the backward direction. The graphs in column two can then simply be read from right to left. Columns three and four show the interesting section of the graph for parameter values just before and just after the bifurcation occurred, and also tell which of the fixed points are (un)stable. Precise details of the map are not important. Only the rather general qualitative behaviour depicted in the graphs is relevant.

1.1.4 Invariant measure

Note: This motivational paragraph does not occur in my written notes, so I am writing this freely out of my mind.

The motivation to this section is roughly the following: Why does the physical theory of thermodynamics work? That is, why does an unpredictable (be it by lack of computational power or fundamental physical limits encountered in quantum mechanics) behaviour on a microscopic scale lead to predictable behaviour on a macroscopic level. For example: Why can we measure temperature and get a precise and reliable result even though the microscopic dynamics of the single particles seem to be chaotic? Well, the answer is roughly the following: We measure over a relatively (on a microscopic scale) large period of time and average over all microscopic states that occur during this time. This will always lead approximately to the same result. To connect this average to our dynamical systems: Each time step corresponds to one iteration of the dynamical map. That is, in our case the time average (of some function) corresponds to the average of the function evaluated at all points on the orbit. Unfortunately we do not explicitly come back to this

motivation in the remaining section. It was primary meant to give a smooth start.

Now, let f be a dynamical map and g a function. Then we are interested in the limit:

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=0}^{k-1} g(x_i) \quad \text{with } x_{i+1} = f(x_i)$$

To analyze this limit use the notion of *measure*. For given f , x_0 , $k \in \mathbb{N}$ and $A \subset \text{dom}(f)$ define the measure:

- $\mu_{x_0}^{(k)}(A) = \frac{1}{k} (\# \text{ of points } x_0, x_1, \dots, x_{k-1} \in A)$
- $\mu_{x_0}(A) = \lim_{k \rightarrow \infty} \mu_{x_0}^{(k)}(A)$

In the cases considered in the previous sections the set A could be any subset of the interval $[0, 1]$.

Two properties of such measures:

1) $\mu_{f(x_0)}(A) = \mu_{x_0}(A)$

This statement says that it does not matter at which point of the orbit we start. This can be seen in the following way: As long as we average over a finite orbit, the orbit starting at x_0 and the orbit starting at $f(x_0)$ will only differ on the endpoints. Their contribution to the average becomes smaller and smaller as we increase the length of the orbit. This can be formalized in the following proof:

$$|\mu_{f(x_0)}^{(k)}(A) - \mu_{x_0}^{(k)}(A)| \leq \frac{2}{k} \xrightarrow{k \rightarrow \infty} 0 \quad \square$$

Since it does not matter if we move our initial point one step forward ($x_0 \rightarrow f(x_0)$) or the set A one step backward ($A \rightarrow f^{-1}(A)$, where $f^{-1}(A)$ is the set of all points that are mapped inside A ; this set inverse always exists), the above statement is equivalent to:

$$\mu_{x_0}(A) = \mu_{x_0}(f^{-1}(A)) (= \mu_{f(x_0)}(A))$$

This gives rise to the definition of an *invariant measure*.

Definition Let f be a map and μ a measure. Then μ is invariant under f if $\forall A : \mu(A) = \mu(f^{-1}(A))$.

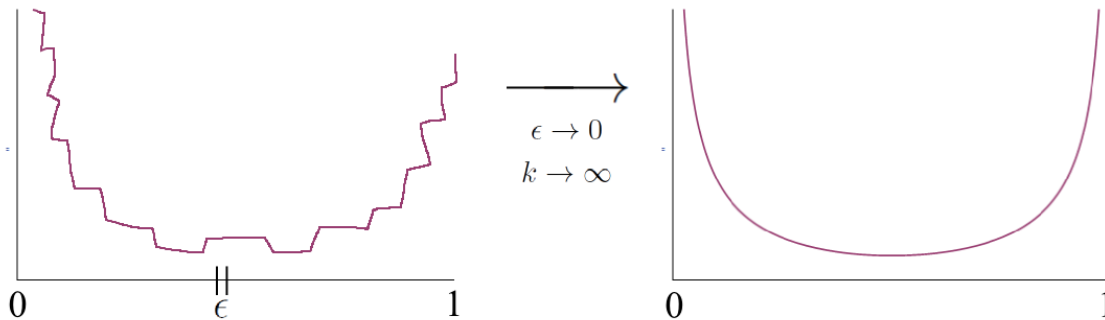
This means:

- The measures μ_x are invariant measures with respect to f
- Dynamical averages are described by integrals with respect to invariant measures:

$$\frac{1}{k} \sum_{i=0}^{k-1} g(x_i) = \int g d\mu_{x_0}^{(k)} \quad (\text{Lebesgue integral})$$

2) If x_0 is “totally aperiodic” (i.e. the orbit never repeats itself) then it is plausible that μ_x is described by a *density*.

$$\begin{aligned} \rho(y) &:= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \mu_x \left(\left[y - \frac{\epsilon}{2}, y + \frac{\epsilon}{2} \right] \right) \\ \text{s.t. } \mu_x(A) &= \int_A \rho(y) dy \end{aligned}$$



Remark: “totally aperiodic” is required. This requirement tells us roughly that the orbit “smoothly” fills some area, and is not concentrated at a few points. Then the measure is in some sense smooth as well. What goes wrong without the “total aperiodicity” can be seen by an example: Assume x_0 is a fixed point, then

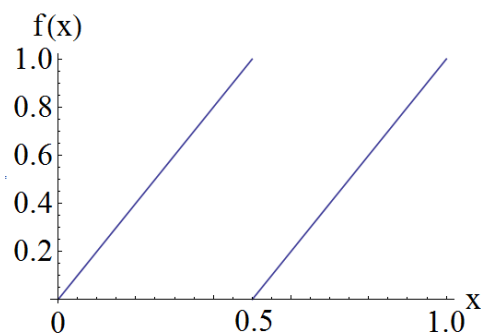
$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=0}^{k-1} g(x_i) = g(x_0)$$

i.e. we would require a density $\rho(y)$ s.t:

$$\int g(y) \rho(y) dy = g(x_0)$$

This means $\rho(y) = \delta_{x_0}(y)$ i.e. the density would be a “delta distribution” which is not a density!

Example: invariant measure of Bernoulli shift



Claim: Lebesgue measure is invariant under the Bernoulli shift. (Note: If the set A is an interval, then its Lebesgue measure is just its typical length.)

Example: $A = [0, 1/4]$, $\mu_L(A) = 1/4$. To calculate $\mu_L(f^{-1}(A))$ we have to know the preimages of the interval $[0, 1/4]$. From the graph one can read off the intervals $[0, 1/8]$ and $[1/2, 1/2 + 1/8]$. One can then calculate:

$$\begin{aligned}\mu_L(f^{-1}(A)) &= \mu_L([0, 1/8] \cup [1/2, 1/2 + 1/8]) \\ &= 2 \cdot 1/8 \\ &= 1/4 \\ &= \mu_L(A)\end{aligned}$$

To state the important steps of the calculation in words:

- “Every set has two preimages of half the size each.”
- “ $\frac{\# \text{ preimages}}{\text{stretching factor}} = \frac{2}{2} = 1$ ”

In fact, this concept can be generalized and formalized:

Frobenius-Perron-Theorem ρ is an invariant density with respect to f iff

$$\rho(y) = \sum_{z \in f^{-1}(\{y\})} \frac{1}{|f'(z)|} \rho(z) \quad \forall y$$

The sum goes over all preimages z of y . The absolute value of the derivative ($|f'(z)|$) is the stretching factor. Let's apply this to the tent map (which is completely analog to the Bernoulli shift).

- Each point has two preimages. The slope is always $|f'(z)| = 2$. For the constant density $\rho(x) = 1$ we see immediately: $1 = 2 \cdot \frac{1}{2} \cdot 1 = 1 \checkmark$.
- Not shown but true: this is the unique invariant measure.
- So: If μ_x generates a measure with a density then that density has to be the constant one.

One can make the following connection to the logistic map (see also exercise sheet 5):

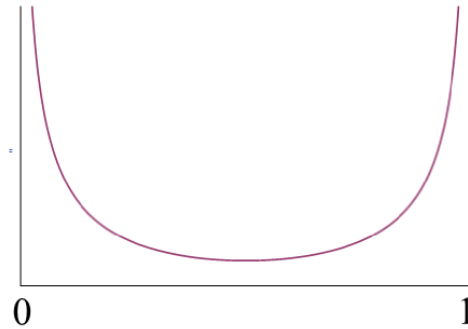
- The logistic map at $r = 4$ is the same as the tent map, up to a change of coordinates:

$$f_{\log,4} = h^{-1} f_{tent} h \quad \text{with } h(x) = \frac{2}{\pi} \arcsin \sqrt{x}$$

- The proof is not difficult, but we have only done some computer simulations.
- One can understand chaotic behaviour of the logistic map at $r = 4$ (section 0, Fig. 2c) in a basic “number theoretic way” (as we have done for the Bernoulli shift with the binary number representation; recall that the Bernoulli shift is very similar to the tent map).

- In particular, one can transform the invariant density from the constant one of the tent map to the invariant density of the logistic map at $r = 4$:

$$\rho_{f_{\log,4}}(x) = \underbrace{\rho_{f_{tent}}(h(x))}_{=1} \left| \frac{dh}{dx} \right| = \frac{1}{\pi \sqrt{x(1-x)}}$$



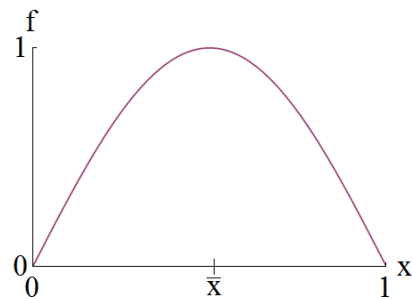
1.1.5 Feigenbaum universality

The recommended reference for this section is Feigenbaum's original paper: *Quantitative Universality for a Class of Nonlinear Transformations*.

One phenomenon that we observed for the logistic map is a cascade of period doubling bifurcations when increasing the scaling parameter r (see section 1.1.3). In this section we shall show that this phenomenon also occurs for a rather general class of maps and that the different bifurcation cascades even share quantitative features.

We consider functions f of the form:

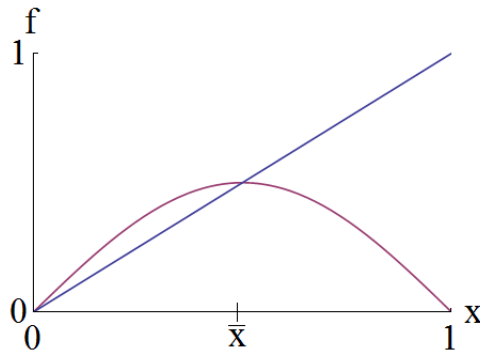
- $f : [0, 1] \rightarrow [0, 1]$
- f has only one extremal point, a maximum at, say, \bar{x} .
- f is parametrized by a scaling parameter:
 $f_r(x) = r f_1(x)$



Two examples of such maps are the logistic map and an appropriately scaled sine function. But note that the function does not have to be symmetric, in particular \bar{x} does not have to be $1/2$.

Definition An orbit is *superstable* if its Lyapunov exponent is $-\infty$ i.e. $\mu = \prod_i f'(x_i) = 0$.

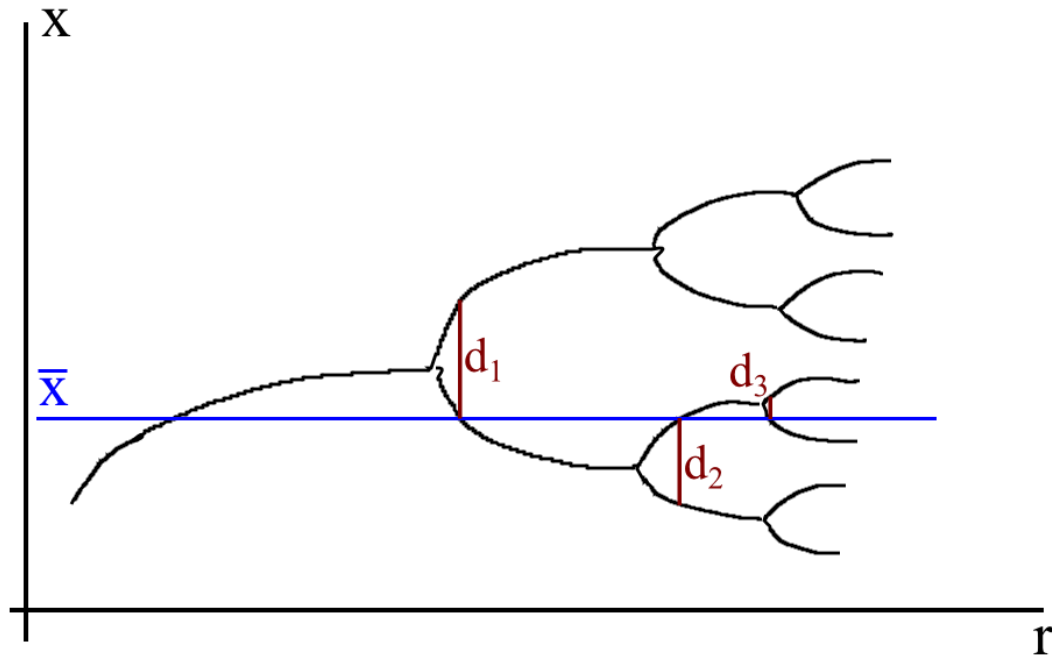
Example Assume that the identity line intersects the graph of f exactly at the maximum \bar{x} . Then $f'(\bar{x}) = 0$ and the fixed point is superstable.



In general an orbit $x_0, x_1, x_2, \dots, x_{n-1}$ is superstable iff $\exists i$ s.t. $x_i = \bar{x}$, because then $\prod_i f'(x_i) = 0$.

Statement:

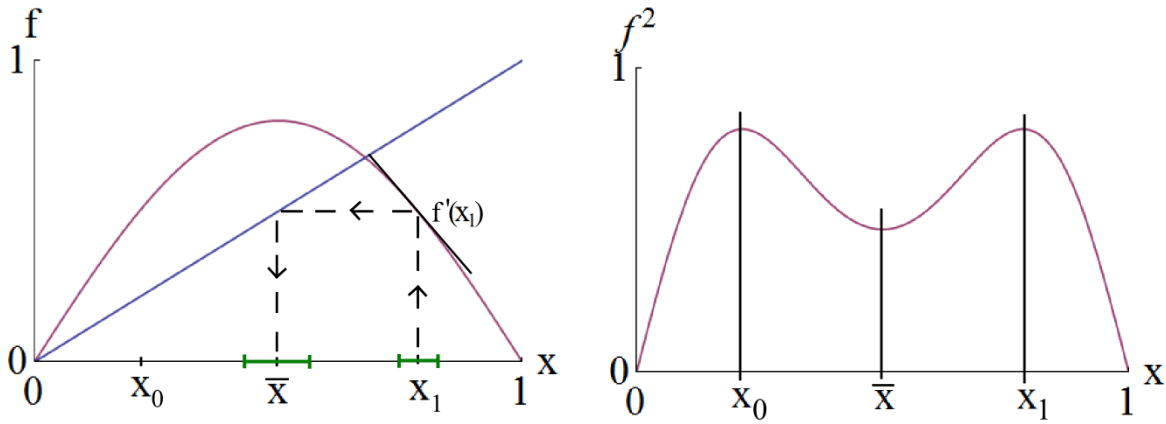
1. Every f as above goes through a period doubling cascade.
2. For every n there is a parameter r_n where the 2^n -periodic orbit becomes superstable.
3. Let d_n be the gap between \bar{x} and the closest point on the orbit. Then $\frac{d_n}{d_{n+1}} \rightarrow \alpha = 2.50\dots$, the second Feigenbaum constant.



To explain this universal behaviour, we revisit the period doubling mechanism in terms of these general functions f .

Insight 1

The behaviour of f^2 around any extremal point depends only on f 's behaviour around its maximum (and on a finite set of slopes $f'(x)$).



As an example, consider the maximum x_1 of f^2 . We have $f(x_1) = \bar{x}$ (because only then $f^2(x_1) = f(\bar{x})$ is maximal). Hence, a small interval around x_1 is mapped under f to a small interval around \bar{x} , stretched by a factor given by the slope $|f'(x_1)|$. The behaviour of x_1 under f^2 is therefore strongly connected to the behaviour of \bar{x} under f . This can be made more precise by Taylor expanding the inner application of f in the expression $f^2(x_1)$ to first order:

$$\begin{aligned} f(x_1 + \delta) &\approx \bar{x} + c \cdot \delta \quad \text{where } c = f'(x_1) \\ \rightarrow f(f(x_1 + \delta)) &\approx f(\bar{x} + c \cdot \delta) \end{aligned}$$

The factor c stretches the x -axis.

For x_0 it works completely analogous. For the minimum of f^2 at \bar{x} one should Taylor expand the outer application of f (the first order approximation of the inner f would just be constant since f is maximal at \bar{x}):

$$f(f(\bar{x} + \delta)) \approx f^2(\bar{x}) + c[f(\bar{x} + \delta) - f(\bar{x})] \quad \text{where } c = f'(f(\bar{x}))$$

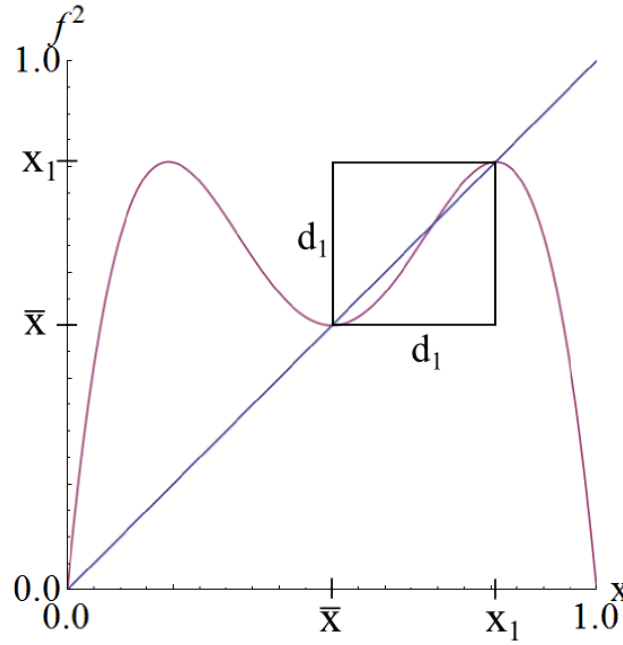
Now c stretches the y -axis and the terms $f^2(\bar{x})$ and $cf(\bar{x})$ also induce a shift along the y -axis.

So we can say in general: Around its extremal values, f^2 is approximately obtained from f by a combination of “shifts and rescalings”.

\Rightarrow The same should also be true for f^4 (just substitute $f \rightarrow f^2$) and hence for $f^{2^n} \forall n$.

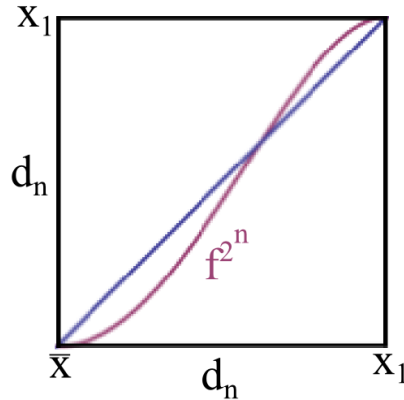
Insight 2

The mechanism for forming period doublings and superstable orbits depends only on the behaviour of f around its maximum \bar{x} . Consider the graph of f^2 for $r = r_2$ (i.e. when the period-2 orbit becomes superstable; for the logistic map $r_2 = 1 + \sqrt{5}$). First, realize that \bar{x} has to be part of the orbit in order to make it superstable. The other point on the superstable period-2 orbit is x_1 .



Thus, as we see, \bar{x} and x_1 are fixed points of f^2 for $r = r_2$. Therefore the black rectangle is indeed a square with side length $d_1 = x_1 - \bar{x}$. Also note that the slope of f^2 at \bar{x} and x_1 is zero since they are both extrema of f^2 .

In general, consider $r = r_n$ i.e. where the orbit of period 2^n becomes superstable. Then f^{2^n} around \bar{x} should look like a S-shape of length d_n in both directions.



This means: We expect f^{2^n} around \bar{x} to look roughly the same for any n on an appropriate scale. It is important to note that the scaling factor is the same in x - and y -direction.

Now combine Insight 1 and 2:

From insight 1 we get the rescaling behaviour:

$$f^{2^n}(\bar{x} + \delta) \approx c_2 f^{2^{n-1}}(\bar{x} + c_1 \delta)$$

That is, doubling the numbers of iterations we get a scaling factor of c_1 along the x -axis and a scaling factor c_2 along the y -axis. From insight 2 we know that these scaling factors should be the same, $c_1 = c_2 = c$, and hence all together:

$$f^{2^n}(\bar{x} + \delta) \approx c f^{2^{n-1}}(\bar{x} + c\delta)$$

Now one might expect that f^{2^n} around \bar{x} converges to a function that is *invariant* under simultaneous rescaling of the argument and the range (the same way that the measures $\mu_x^{(n)}$ tended to an invariant measure). This invariant function should be a solution of the **Feigenbaum Cvitanovic equation**:

$$\alpha g \circ g \left(\frac{\delta}{\alpha} \right) = g(\delta)$$

Note that we got rid of the \bar{x} in the argument due to a reparametrization. Even though this equation is very complicated (due to its nonlinearity), there is the following

Result: There is a *unique* solution (for g and α). For g one can calculate an expansion of the form:

$$g(x) = 1 - 1.52763...x^2 + 0.104815...x^4 + 0.0267057...x^6 + \dots$$

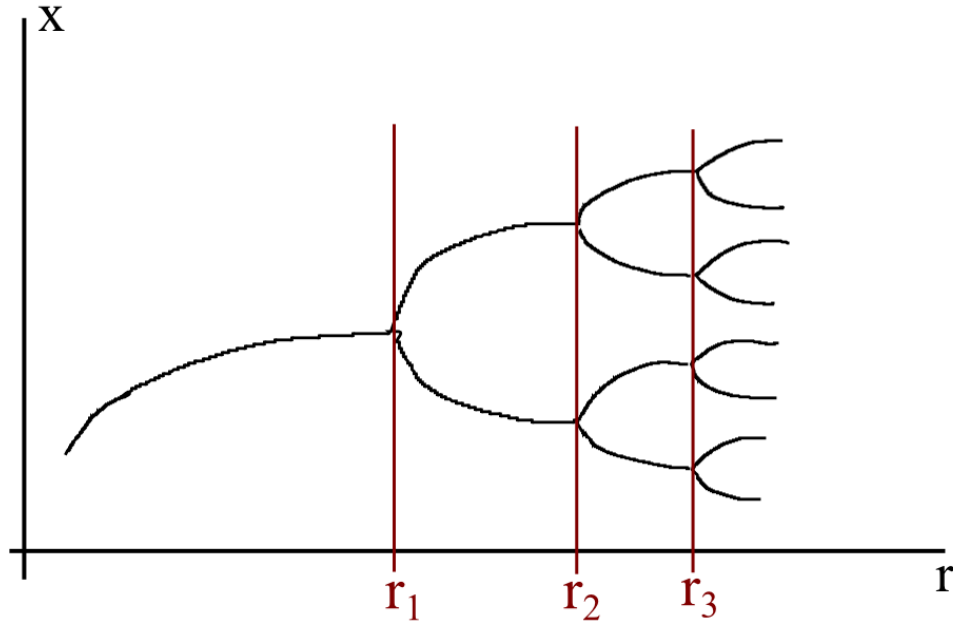
For the scaling constant α (scaling between box sizes in the above figures) one finds:

$$\alpha = 2.50290...$$

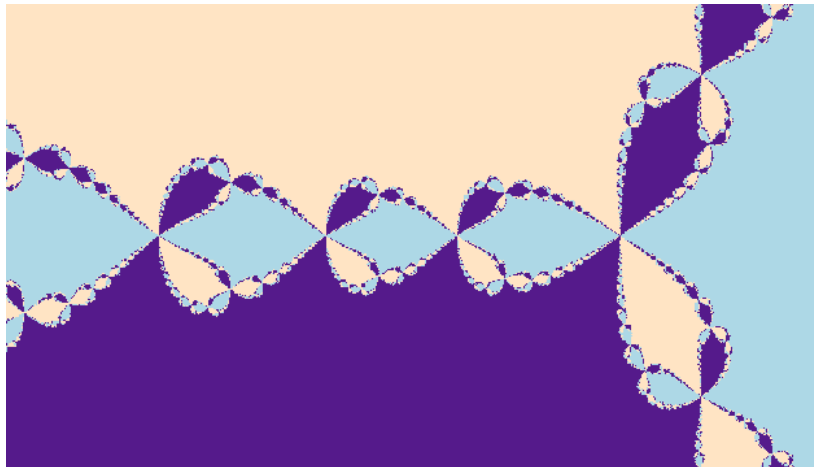
Note: The derivation of the Feigenbaum Cvitanovic equation was only heuristic and rather hand waving, but the results are precise!

Also note that α is called the second Feigenbaum constant. The first Feigenbaum constant appears in a similar universal behaviour for the distances between neighbouring bifurcations.

$$\delta_k = \frac{r_{k-1} - r_{k-2}}{r_k - r_{k-1}} \rightarrow \delta = 4.66920...$$



1.1.6 Fractal dimension



Observation:

Some structures arising in simple-to-specify dynamical systems seem to have “features on every scale”. There are “notions” of dimensionality that capture this effect quantitatively.

Construction:

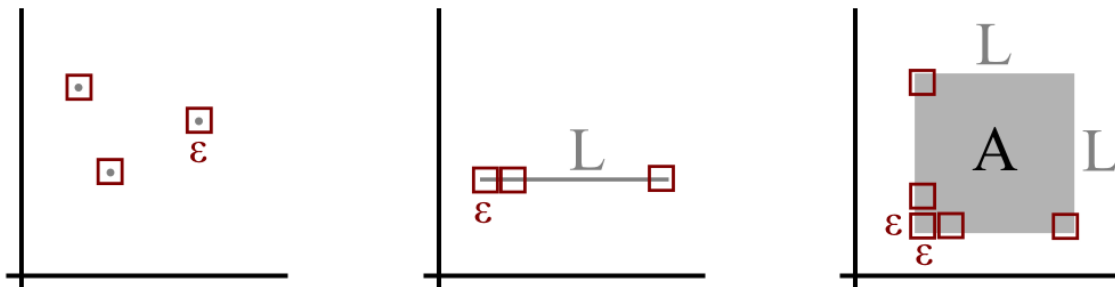
Consider a bounded shape $S \subset \mathbb{R}^d$. With every S and every scale ϵ , associate:

$$N_S(\epsilon) = \# \text{ of } \epsilon\text{-cubes required to cover } S$$

$$\Rightarrow \log N_S(\epsilon) \text{ is the } \textit{entropy number} \text{ of } S \text{ at scale } \epsilon.$$

Note: We typically use the base-2-logarithm whose unit is called *bit*. However, using another base simply corresponds to a change of units (i.e. a scaling factor).

Examples in \mathbb{R}^2 :



a) n isolated points, mutual distance $> \epsilon$:

$$\log N_S(\epsilon) = \log n$$

b) Line of length L (parallel to axis):

$$\begin{aligned} \log N_S(\epsilon) &= \log \frac{L}{\epsilon} \\ &= \log \frac{1}{\epsilon} + \log L \end{aligned}$$

c) Square of side length L :

$$\begin{aligned}\log N_S(\epsilon) &= \log \left[\left(\frac{L}{\epsilon} \right)^2 \right] \\ &= 2 \log \frac{1}{\epsilon} + 2 \log L\end{aligned}$$

The slope in front of the $\log \frac{1}{\epsilon}$ term seems to match our intuition of dimension. General statement: If S is a D -dimensional cube of side length L , contained in \mathbb{R}^d , $d \geq D$, then:

$$\begin{aligned}\log N_S(\epsilon) &= \log \left[\left(\frac{L}{\epsilon} \right)^D \right] \\ &= D \log \frac{1}{\epsilon} + D \log L\end{aligned}$$

Interpretation of entropy number:

$\log N_S(\epsilon)$ = "# of bits one has to use in order to specify a point on S up to an error of ϵ ". (In general $\log N$ is the number of bits needed to encode N numbers. For example, if $N = 100$ one needs $\log_2 100 \approx 6.64 \rightarrow 7$ bits (in the binary system) or $\log_{10} 100 = 2$ digits (in the decimal system).

Definition *Box counting dimension:*

$$D_0(S) = \lim_{\epsilon \rightarrow 0} \frac{\log N_S(\epsilon)}{\log \frac{1}{\epsilon}}$$

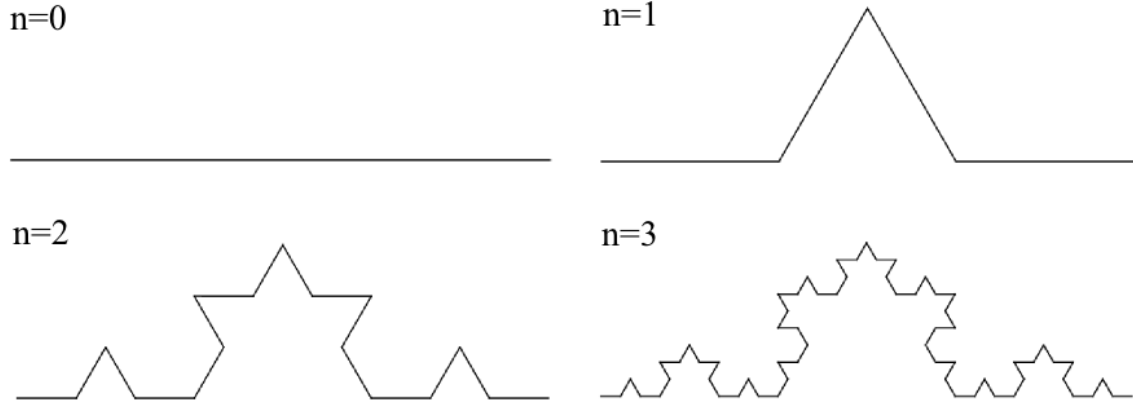
From the previous examples it follows that:

$$D_0(D\text{-dimensional cube of side length } L) = \lim_{\epsilon \rightarrow 0} \frac{D \log \frac{1}{\epsilon} + D \log L}{\log \frac{1}{\epsilon}} = D$$

This is the minimal requirement for a definition of dimension. We have to obtain the correct value for objects that we already associate a dimension with. The box counting dimension is one way (of many) to formalize and generalize our intuitive notion of dimension. We shall now calculate the box counting dimension for more complicated examples that show structure on arbitrary fine scales.

Example 1 $S = \text{Koch curve} \subset \mathbb{R}^2$:

The Koch curve is constructed in an iterative way. We start with one straight line (from 0 to 1). In the first step, we replace the middle third interval by a "tent" i.e. an upward directed equilateral triangle without its base. The figure now consists of four lines each of length $1/3$. In each following step we apply the same procedure to each line of the current figure.

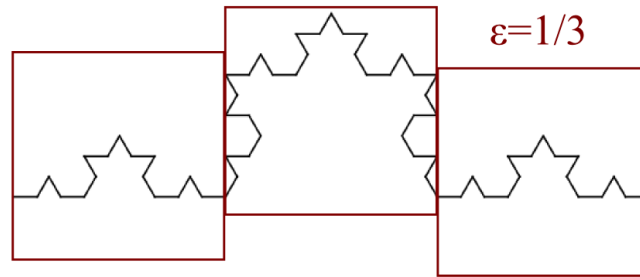


In each iteration each line of length l_n becomes four lines each of length $l_n/3$. The length of the total curve does therefore increase by a factor of $4/3$ at each iteration. So for the length of the curve we find:

$$1, \frac{4}{3}, \left(\frac{4}{3}\right)^2, \dots, \lim_{n \rightarrow \infty} \left(\frac{4}{3}\right)^n = \infty$$

Note: If we terminate the procedure after a finite number of iterations we would of course get a figure of finite length (composed of a lot of straight lines) that has the intuitive dimension $D = 1$. But we are interested in the limiting curve i.e. in the case that the iterations never stop.

To calculate the box counting dimension we consider the sequence $\epsilon_k = \left(\frac{1}{3}\right)^k$ for $k \rightarrow \infty$ (strictly speaking convergence for this subsequence does not imply convergence for the general $\epsilon \rightarrow 0$, but we ignore this difficulty here). For $k = 1$ we need $N_K\left(\frac{1}{3}\right) = 3$ boxes to cover the whole figure:



Essentially, these boxes cover the four lines that we had after one iteration. The left box covers the left line, the right box the right line and the middle box the two middle lines from the upward triangle. Since this standard shape reproduces itself four times in each iteration the required number of boxes scales with a factor of four at each iteration:

k	1	2	...	K
$N_K\left(\left(\frac{1}{3}\right)^k\right)$	3	12	...	$3 \cdot 4^{K-1}$

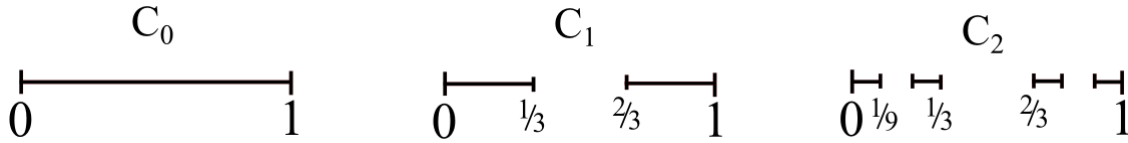
From there it follows:

$$\begin{aligned}
 D_0(\text{Koch}) &= \lim_{k \rightarrow \infty} \frac{\log \left[N_K \left(\left(\frac{1}{3} \right)^k \right) \right]}{\log [3^k]} \\
 &= \lim_{k \rightarrow \infty} \frac{\log [3 \cdot 4^{k-1}]}{\log [3^k]} \\
 &= \lim_{k \rightarrow \infty} \frac{(k-1) \log 4 + \log 3}{k \log 3} \\
 &= \frac{\log 4}{\log 3} \\
 &\approx 1.26...
 \end{aligned}$$

The Koch curve has a fractal dimension! In some sense it is more than a line but less than a two dimensional surface.

Example 2 Middle-third Cantor set $C \subset \mathbb{R}$:

As the Koch curve, the Cantor set follows an iterative construction. We start with the interval $[0, 1]$. In the first step we remove the middle third of this interval, i.e. the interval $(1/3, 2/3)$. We have now two intervals of length $1/3$. In each following step we remove the middle third of each interval of the current figure.



By construction it follows that the measure (here the length of all combined intervals) is reduced by a factor of $2/3$ at each iteration:

$$\mu(C_n) = \left(\frac{2}{3} \right)^n \xrightarrow{n \rightarrow \infty} 0$$

So the measure of the Cantor set is $\mu(C) = 0$. Intuitively this fact stands in favour of the dimension $D = 0$. On the other hand, the Cantor set is uncountable (i.e. there is no bijection to \mathbb{N}). Sloppy speaking, this means that the Cantor set is not just a collection of single points. This stands in favour of the dimension $D = 1$. Hence, with these two indicators we cannot ascribe the Cantor set an intuitive dimension. In fact, it turns out that the box counting dimension is between 0 and 1. As for the Koch curve, we consider the sequence $\epsilon_k = \left(\frac{1}{3} \right)^k$. For $k = 1$ we need two boxes to cover the whole set (essentially one box per interval of the set C_1). In each iteration the number of intervals is doubled while their length is divided by a factor of three. Hence, (for $k = 2$) we need four boxes of length $1/9$. In general:

k	1	2	...	K
$N_C \left(\left(\frac{1}{3} \right)^k \right)$	2	4	...	2^K

Thus, the box counting dimension evaluates to:

$$D_0(C) = \lim_{k \rightarrow \infty} \frac{\log(2^k)}{\log(3^k)} = \lim_{k \rightarrow \infty} \frac{k \log 2}{k \log 3} = \frac{\log 2}{\log 3} \approx 0.63...$$

As a side remark, here is a proof that the Cantor set is uncountable (proof by contradiction):

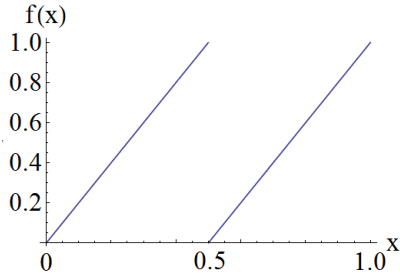
First, construct a representation of the elements of the Cantor set: At each iteration one can either choose the left or the right interval. We encode the left path with a 0 and the right path with a 1. Each element of the Cantor set is obtained by a unique path and can therefore be represented by a unique infinite sequence of 0's and 1's. Now assume that we have a correspondence $\mathbb{N} \rightarrow C$:

1	001010111001...
2	011011100010...
3	110100001100...
⋮	⋮

Then one can construct an element that is not in this list in the following way: Take the negation ($\bar{0} = 1$, $\bar{1} = 0$) of the first element of the first list entry, the second element of the second entry and in general the n_{th} element of the n_{th} entry. By construction the new element differs from any other list entry at at least one position. \square

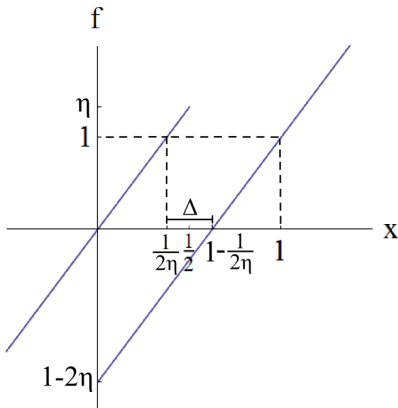
Example 3 Fractal arising in a dynamical system:

Recall the Bernoulli shift:



$$f(x) = 2x \bmod 1 = \begin{cases} 2x & , x \leq \frac{1}{2} \\ 2x - 1 & , x > \frac{1}{2} \end{cases}$$

We slightly modify this map by introducing an additional parameter that increases the slope, and by extending the map to $\pm\infty$ (we are still mainly interested in the interval $[0, 1]$):



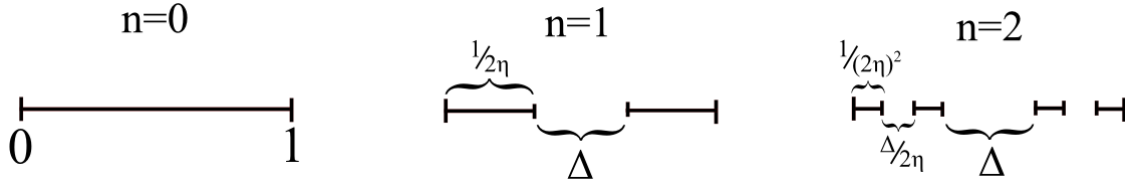
$$f(x) = \begin{cases} 2\eta x & , x \leq \frac{1}{2} \\ 2\eta(x - 1) + 1 & , x > \frac{1}{2} \end{cases}$$

Initial values:

- $x_0 < 0$ diverge to $-\infty$
- $x_0 > 1$ diverge to $+\infty$

But also values $x_0 \in [0, 1]$ that are eventually mapped to either of the two regimes will diverge to $\pm\infty$. Then, what is the structure of the set S of points that will remain inside the interval $[0, 1]$ indefinitely?

For one iteration one can see that the intervals $[0, 1/2\eta]$ and $[1 - 1/2\eta, 1]$ are mapped to $[0, 1]$, while points in the middle interval $[1/2\eta, 1 - 1/2\eta]$ are mapped to values < 0 or > 1 . Similar as for the construction of the middle-third Cantor set, this procedure repeats itself. So for two iterations we have to find the subsets of $[0, 1/2\eta]$ and $[1 - 1/2\eta, 1]$ that are mapped to $[1/2\eta, 1 - 1/2\eta]$ at the first iteration. For $[0, 1/2\eta]$ one finds the intervals $\frac{1}{2\eta} [0, 1/2\eta]$ and $\frac{1}{2\eta} [1 - 1/2\eta, 1]$. So at each iteration the number of remaining intervals is doubled while their length is divided by a factor of 2η . The gaps between the intervals scale as $\Delta = 1 - 1/\eta$, $\Delta/2\eta$, $\Delta/(2\eta)^2 \dots$



To calculate the box counting dimension we consider the sequence $\epsilon_k = \left(\frac{1}{2\eta}\right)^k$ and find:

$$\frac{k}{N_S \left(\left(\frac{1}{2\eta} \right)^k \right)} \left| \begin{array}{cccc} 1 & 2 & \dots & K \\ 2 & 4 & \dots & 2^K \end{array} \right.$$

$$D_0(S) = \lim_{k \rightarrow \infty} \frac{\log(2^k)}{\log[(2\eta)^k]} = \frac{\log 2}{\log 2\eta} \in [0, 1]$$

For $\eta = 1$ we regain the typical Bernoulli shift where the whole interval $[0, 1]$ is mapped to $[0, 1]$ and thus we find $D_0 = 1$. For $\eta \rightarrow \infty$ the whole interval $[0, 1]$ diverges to $\pm\infty$ and we find $D_0 = 0$. For any other $\eta \in (1, \infty)$ we find that the set remaining indefinitely in $[0, 1]$ is a fractal of dimension $\frac{\log 2}{\log 2\eta}$. So we actually managed to find a fractal arising from a simple dynamical map.

2 Stochastic Processes

At the heart of stochastic processes stands probability theory. We shall therefore start with a short introduction of this large subject.

2.1 Probability theory

The abstract definition of a probabilistic model starts with a probability space (Ω, \mathcal{F}, P) build up by the three parts:

- Ω : The *sample space* which consists of all possible outcomes.
- \mathcal{F} : The set of *events*. An event is any subset of Ω .
- P : The *probability measure* that assigns a probability to every event.

As an example consider a coin flipping experiment:

- $\Omega = \{\text{heads}, \text{tails}\}$
- $\mathcal{F} = \{\{\text{heads}, \text{tails}\}, \{\text{heads}\}, \{\text{tails}\}, \emptyset\}$: The event $\{\text{heads}\}$ means the coin landing heads. The event $\{\text{heads}, \text{tails}\}$ is the event that the coin lands either heads or tails, so the event that anything happens. \emptyset means that none of the possible outcomes occurs.
- $P(\{\text{heads}, \text{tails}\}) = 1$ because the probability that anything happens has to be one. Accordingly, the probability that we observe none of the possible outcomes is $P(\emptyset) = 0$. If the coin is fair then $P(\{\text{heads}\}) = P(\{\text{tails}\}) = 1/2$.

In general, the following rules apply:

1. $0 \leq P(A) \leq 1 \ \forall A \in \mathcal{F}$
2. $P(\Omega) = 1$
3. For A_1, A_2, A_3, \dots ($A_i \in \mathcal{F}$) with $A_j \cap A_k = \emptyset$ (for $j \neq k$) one has $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. That is, probabilities of disjoint events are simply added.

2.1.1 Conditional probabilities

One is often interested in the probability of an event A given that an event B occurred. This probability can be obtained as the probability that A and B occur divided by the probability of B (e.g. B : it is cloudy today, A : it was sunny yesterday):

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ \rightarrow P(A \cap B) &= P(A|B) P(B) \end{aligned}$$

Two events are called independent if the joint probability factorizes:

$$P(A \cap B) = P(A) P(B)$$

In this case we find for the conditional probabilities:

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B)$$

2.1.2 Random variables

In the abstract definition a random variable

$$X : \Omega \rightarrow \mathbb{R}$$

is a map from the sample space to the real numbers. This establishes the connection from the abstract definition to the intuitive understanding of probability theory.

Example 1: The outcomes heads and tails of a coin are often mapped to 0 and 1 (but any other numbers would be possible as well):

$$\begin{aligned} X(\text{heads}) &= 0 \\ X(\text{tails}) &= 1 \end{aligned}$$

Example 2: For a six sided die the outcome 'side with n eyes' can be mapped to the number n . In this case the step from the abstract definition to the assigned number is very intuitive (in a careless sense they coincide).

2.1.3 Cumulative distribution function, probability distribution and probability density

One is often interested in the probability that the outcome of the random variable is smaller than a certain value:

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

$F_X(x)$ is called the cumulative distribution function of the random variable X .

For discrete random variables one can also assign a probability to a single outcome value:

$$\text{probability distribution: } P_X(x) = P(X = x)$$

For continuous random variables the probability that the variable takes on one specific value x is zero. But one can often define a:

$$\text{probability density } f_X(x) \text{ s.t. } F_X(x) = \int_{-\infty}^x f_X(t) dt$$

If the distribution is “well behaved”, then:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Note: The probability density does not necessarily exist. In a sense, the distribution has to be “smooth” enough. This is the case for *absolutely continuous* random variables which always have a density. The cumulative distribution function, on the other hand, always exists.

As an example consider the Gaussian distribution with mean μ and width σ : $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

2.1.4 Two variable distributions

Consider two random variables $X, Y : \Omega \rightarrow \mathbb{R}$. Completely analogous to the case of one single variable one can define the joint cumulative distribution:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \in \Omega : X(\omega) \leq x\} \cap \{\omega' \in \Omega : Y(\omega') \leq y\})$$

Furthermore, in the discrete case, define the joint probability distribution

$$P_{X,Y}(x, y) = P(X = x, Y = y)$$

and in the continuous case the joint density

$$f_{X,Y}(x, y) = \frac{d^2}{dx dy} F_{X,Y}(x, y).$$

One can obtain the marginal distribution of a single variable by summing or integrating over the other variable:

$$\begin{aligned} \text{discrete: } P_X(x) &= \sum_y P_{X,Y}(x, y) \\ \text{continuu: } f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \end{aligned}$$

Analogous to what we did in the abstract definition, we can define conditional distributions.

In the discrete case:

$$\begin{aligned} P_{X|Y=y}(x) &= P(X = x | Y = y) \\ &= P(\{\omega \in \Omega : X(\omega) = x\} | \{\omega' \in \Omega : Y(\omega) = y\}) \\ &= \frac{P(\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega' \in \Omega : Y(\omega) = y\})}{P(\{\omega' \in \Omega : Y(\omega) = y\})} \\ &= \frac{P_{X,Y}(x, y)}{P_Y(y)} \\ \rightarrow P_{X,Y}(x, y) &= P(X = x | Y = y) P(Y = y) \\ &= P(Y = y | X = x) P(X = x) \end{aligned}$$

In the continuous case:

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} \\ \rightarrow f_{X,Y}(x, y) &= f_{X|Y=y}(x) f_Y(y) \\ &= f_{Y|X=x}(y) f_X(x) \end{aligned}$$

For independent random variables all kinds of distribution functions factorize:

$$\begin{aligned} \text{cumulative distr. fct.: } F_{X,Y}(x, y) &= F_X(x) F_Y(y) \\ \text{discrete case: prob. distr.: } P_{X,Y}(x, y) &= P_X(x) P_Y(y) \\ \text{continuous case: density: } f_{X,Y}(x, y) &= f_X(x) f_Y(y) \end{aligned}$$

2.1.5 Expectation value

The expectation values of a random variable is defined as the average over all possible outcomes weighted with the corresponding probabilities:

$$\begin{aligned} \text{discrete case: } E[X] &= \sum_x x P(X = x) \\ \text{continuous case: } E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \end{aligned}$$

The expectation does not necessarily exist. As an example consider the discrete distribution $P(X = n) = \frac{1}{C} \frac{1}{n^2}$ (with $C = \frac{\pi^2}{6}$), $n = 1, 2, 3, \dots$. One finds:

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{C} \frac{1}{n^2} &= 1 \\ \text{but } E[X] &= \sum_{n=1}^{\infty} \frac{1}{C} \frac{1}{n} = \infty \end{aligned}$$

It is even possible that one cannot assign any value (including $\pm\infty$) at all. The following subsection is dedicated to illustrate this strange behaviour.

Side remark: About ill defined expectation values

The crucial point that can cause trouble is when the expectation value has a positive and a negative part that converge individually to $\pm\infty$ (then “ $E[x] = \infty - \infty = ?$ ”). We will first consider the discrete case in a general way and then give a detailed example for the continuous case.

Consider a series that is convergent but not absolutely convergent:

$$\sum_{n=1}^{\infty} a_n = C \quad \text{but} \quad \sum_{n=1}^{\infty} |a_n| = \infty$$

Such a series is called conditionally convergent.

Riemann’s rearrangement theorem: Let $\sum_{n=1}^N a_n$ be a conditionally convergent series, then for any $r \in \mathbb{R}$ there exists a permutation σ of the \mathbb{N} s.t.:

$$\lim_{N \rightarrow \infty} \sum_{n=1}^N a_{\sigma(n)} = r$$

That is, one can rearrange the order of summation in order to converge to any desired value. One can also let the series diverge to $\pm\infty$ or to not be well defined at all. (Note: For an absolutely convergent series the order of summation is irrelevant.)

For expectation values no natural order of summation exists. Hence, if the expectation value is only conditionally convergent one obtain any (or no) value. Thus, in this case the expectation value is not well defined. An example for a conditionally convergent series is $\sum_{n=1}^{\infty} (-1)^n \frac{1}{n}$. We will not prove the above theorem but instead consider a detailed example for the continuous case.

Consider the general integral (not necessarily an expectation value) $\int_{-\infty}^{\infty} f(x) dx$. This expression contains two limits at once (upper integral limit to $+\infty$ and lower limit to $-\infty$). If both limits exist separately

$$I_+ := \int_0^{\infty} f(x) dx = U \quad \text{and} \quad I_- := \int_{-\infty}^0 f(x) dx = V$$

then we get for the total integral:

$$\int_{-\infty}^{\infty} f(x) dx = I_+ + I_- = U + V$$

However, if $I_+ = \infty$ and $I_- = -\infty$ we run into trouble. The value of the total integral might depend on “how fast we approach $+\infty$ compared to $-\infty$ ”. To formalize this, we first combine the two limits by defining two functions $G(t)$ and $H(t)$ that satisfy:

$$\lim_{t \rightarrow \infty} G(t) = \infty \quad \text{and} \quad \lim_{t \rightarrow \infty} H(t) = -\infty$$

Then we rewrite $\int_{-\infty}^{\infty} f(x) dx$ as:

$$\lim_{t \rightarrow \infty} \int_{H(t)}^{G(t)} f(x) dx$$

In the case that $I_+ = U$ and $I_- = V$ exist separately this expression is well defined:

$$\begin{aligned} \left| \int_{H(t)}^{G(t)} f(x) dx - U - V \right| &= \left| \int_0^{G(t)} f(x) dx - U + \int_{H(t)}^0 f(x) dx - V \right| \\ &\leq \underbrace{\left| \int_0^{G(t)} f(x) dx - U \right|}_{\rightarrow 0} + \underbrace{\left| \int_{H(t)}^0 f(x) dx - V \right|}_{\rightarrow 0} \\ &\rightarrow 0 \end{aligned}$$

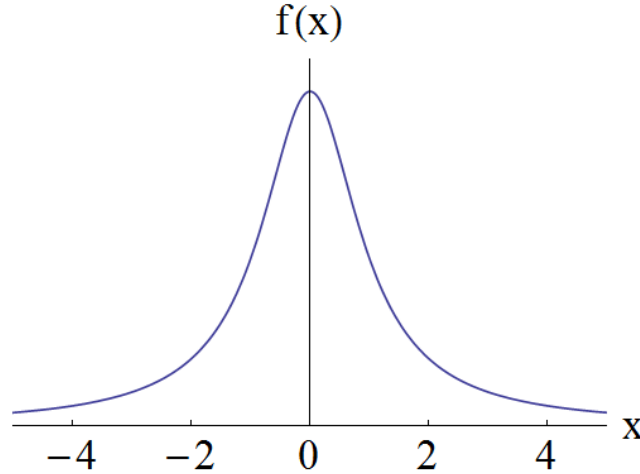
For the case that I_+ and I_- are $\pm\infty$ consider as an example the Cauchy distribution with density $f_X(x) = \frac{1}{\pi} \frac{1}{x^2+1}$ and define:

$$Q(t) = \int_{H(t)}^{G(t)} \frac{1}{\pi} \frac{x}{x^2+1} = \frac{1}{2\pi} \ln \left(\frac{G(t)^2+1}{H(t)^2+1} \right)$$

$\lim_{t \rightarrow \infty} Q(t)$ would be the expectation value if it existed. To show that it is not well defined we consider several realizations of $G(t)$ and $H(t)$.

- $G(t) = t, H(t) = -t \rightarrow Q(t) = \frac{1}{2\pi} \ln \left(\frac{t^2+1}{t^2+1} \right) \rightarrow \frac{1}{2\pi} \ln 1 = 0$
- $G(t) = te^{\pi a}, H(t) = -t \rightarrow Q(t) = \frac{1}{2\pi} \ln \left(\frac{t^2 e^{2\pi a} + 1}{t^2 + 1} \right) \rightarrow \frac{1}{2\pi} 2\pi a = a$
We can obtain any value $a \in \mathbb{R}$!
- $G(t) = t^2, H(t) = -t \rightarrow Q(t) \rightarrow \infty$
- $G(t) = t, H(t) = -t^2 \rightarrow Q(t) \rightarrow -\infty$
- $G(t) = t(1 + |\sin t|), H(t) = -t$
 $\rightarrow Q(t) = \frac{1}{2\pi} \ln \left(\frac{(1+|\sin t|)^2 + \frac{1}{t^2}}{1 + \frac{1}{t^2}} \right) \rightarrow \frac{1}{\pi} \ln(1 + |\sin t|)$
 \rightarrow no limit exists!

As one can see, the expectation value of the Cauchy distribution is not well defined. Also higher moments, like variance, are not well defined. Considering the graph of the Cauchy distribution, one might wonder why this is the case.



The graph looks qualitatively similar to that of a Gaussian distribution. Intuitively one would probably expect zero as the expectation value (as obtained for the same rate of convergence towards both $\pm\infty$). However, compared to the Gaussian distribution, the tails of the Cauchy distribution are much heavier such that the individual integrals in both directions diverge.

As a final remark, the empirical mean value (sample mean) that would converge to the expectation value if it existed, does not converge either. The fluctuations become larger with increasing sample size. The law of large numbers that typically guarantees the convergence of the sample mean to the expectation value cannot be applied here as it requires the existence of the expectation value.

2.2 Discrete time and discrete space processes

In general, A stochastic process can be characterized by a sequence of random variables $\{X_t\}_{t \in S}$ where $S = \mathbb{N}(\mathbb{R})$ corresponds to discrete (continuous) time processes. The values

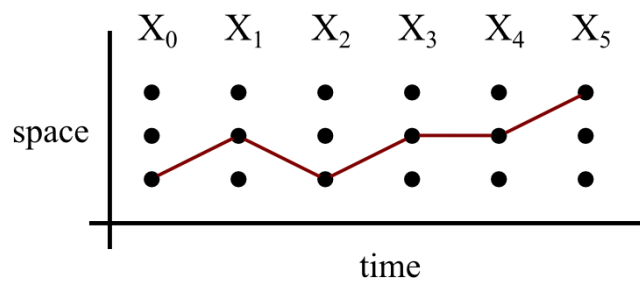
that X_t can take are elements of the *state space*. The state space can be discrete or continuous as well. All four combinations of discrete and continuous time and state space are possible.

	discrete space	continuous space
discrete time	our starting point	...
continuous time	...	e.g. Brownian motion

We start with discrete time and discrete space processes. Consider the sequence

$$X_0, X_1, X_2, X_3, \dots, X_{n+1}$$

where X_i are discrete random variables. Label the possible values of X_i by $j_i = 1, 2, 3, \dots$. The probability that the process follows one particular *path* (i.e. a particular sequence of values $X_i = j_i$)



is:

$$P(X_{n+1} = j_{n+1}, X_n = j_n, \dots, X_0 = j_0)$$

One of our goals is to calculate this probability (i.e. to simplify the expression) and other quantities arising from it (like marginal distributions). We do this for special class of processes, so called *Markov chains*.

2.2.1 Markov chains

In general the value of X_{n+1} can depend on all previous values X_n, X_{n-1}, \dots, X_0 . We want to consider the special case that the probability for the next step depends only on the current step. One then speaks of a Markov chain and also calls such systems *memoryless*. Formally the Markov condition reads:

$$P(X_{n+1} = j_{n+1} \mid X_n = j_n, X_{n-1} = j_{n-1}, \dots, X_0 = j_0) = P(X_{n+1} = j_{n+1} \mid X_n = j_n) \quad \forall n$$

A Markov chain is called *homogeneous* if the *transition probabilities*

$$P_{jk} = P(X_{n+1} = j \mid X_n = k)$$

are independent of n , i.e. if they do not change from step to step.

Example 1: Consider a person with a life that consists only of 'study', 'party' and 'sleep'. He also does any of these activities for a whole day. The probability of tomorrow's activity shall only depend on today's activity. For example the probability that he sleeps tomorrow if he parties today ($P(\text{sleep}|\text{party})$) should be rather large. Let us specify all these probabilities:

$$\begin{array}{lll} P(\text{study}|\text{study}) = 0.3 & P(\text{study}|\text{party}) = 0.01 & P(\text{study}|\text{sleep}) = 0.4 \\ P(\text{party}|\text{study}) = 0.4 & P(\text{party}|\text{party}) = 0.09 & P(\text{party}|\text{sleep}) = 0.4 \\ P(\text{sleep}|\text{study}) = 0.3 & P(\text{sleep}|\text{party}) = 0.9 & P(\text{sleep}|\text{sleep}) = 0.2 \end{array}$$

It suggests itself to write these probabilities in a matrix:

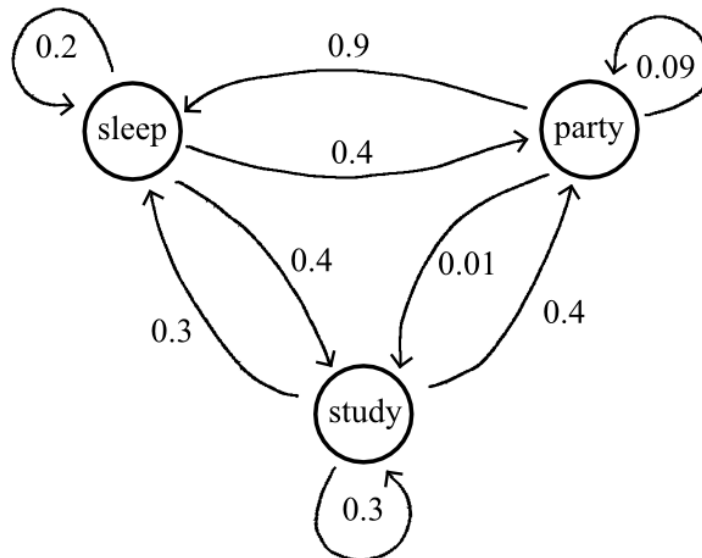
$$P = \begin{pmatrix} 0.3 & 0.01 & 0.4 \\ 0.4 & 0.09 & 0.4 \\ 0.3 & 0.9 & 0.2 \end{pmatrix}$$

The elements of this matrix are the aforementioned transition probabilities. Hence, the matrix P is called *transition matrix*. As the matrix is written down here it has to be multiplied from the left to a column vector

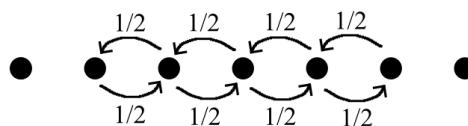
$$p = \begin{pmatrix} p_{\text{study}} \\ p_{\text{party}} \\ p_{\text{sleep}} \end{pmatrix}$$

which contains the current probabilities to study, party and sleep.

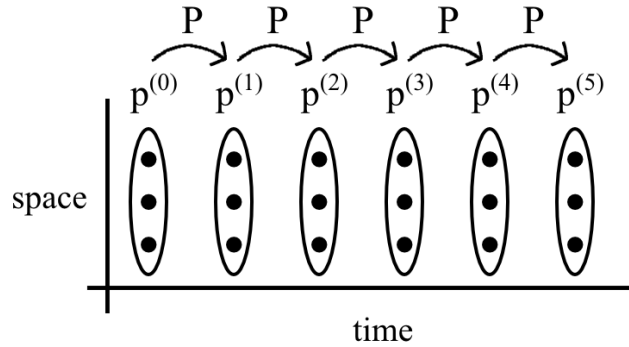
Instead of the matrix representation one can equivalently represent the transition probabilities in a *transition graph*:



Example 2 One dimensional *random walk*



That is, the initial distribution $p^{(0)}$ propagates forward in time by multiplication with the transition matrix P , one multiplication for each time step.



To identify the matrix multiplications in line 3 of the above calculation, recall the general expression for the product of two matrices A and B

$$[AB]_{ij} = \sum_k A_{ik} B_{kj}$$

and for the special case of the multiplication of a matrix A with a vector v :

$$[Av]_i = \sum_k A_{ik} v_k$$

In this way one can for example identify

$$\sum_{j_0} P_{j_1, j_0} p_{j_0}^{(0)} = [P p^{(0)}]_{j_1} = p_{j_1}^{(1)}$$

in the above calculation. Further iterations yield the final result

$$p_{j_{n+1}}^{(n+1)} = [P^{n+1} p^{(0)}]_{j_{n+1}}$$

2.2.2 Properties of the transition matrix

We consider a transition matrix $P_{jk} = P(X_{n+1} = j \mid X_n = k) \quad \forall n$. This matrix has to satisfy:

- $P_{jk} \geq 0$
- $\sum_j P_{jk} = \sum_j P(X_{n+1} = j \mid X_n = k) = 1$

That is, elements have to be non-negative as they represent probabilities and elements in one column have to sum to one since the probability to go anywhere (including to stay in the current state) is one. Such a matrix is called a *left stochastic* matrix and it has to be applied from the left to a column vector:

$$p^{(n+1)} = P p^{(n)}$$

Note that the standard in probability theory is to use *right stochastic* matrices whose rows sum to one and that are applied from the right to a row vector:

$$p^{t, (n+1)} = p^{t, (n)} \tilde{P}$$

Lemma Let P, Q be $N \times N$ left stochastic matrices. Then QP is also a left stochastic matrix. Hence, also P^n is left stochastic.

Proof:

$$\begin{aligned} \bullet [QP]_{jk} &= \sum_{l=1}^N \underbrace{Q_{jl}}_{\geq 0} \underbrace{P_{lk}}_{\geq 0} \geq 0 \\ \bullet \sum_{j=1}^N [QP]_{jk} &= \sum_{j=1}^N \sum_{l=1}^N Q_{jl} P_{lk} = \sum_{l=1}^N \underbrace{\left(\sum_{j=1}^N Q_{jl} \right)}_{=1} \underbrace{P_{lk}}_{=1} = 1 \end{aligned}$$

where we used that columns of Q and P sum to one.

2.2.3 Stationary and limit distributions

- Given P we say a distribution q is *stationary* with respect to P if

$$Pq = q \rightarrow P^n q = q$$

- A probability distribution π is called a *limit distribution* with respect to P if

$$\lim_{n \rightarrow \infty} P^n v = \pi$$

for all initial distributions v . A limit distribution does not always exist, even if a stationary distribution exists.

Examples:

- $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ does not have a limit distribution. Each application of P simply swaps the two elements of a probability vector p :

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} p_2 \\ p_1 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \rightarrow \dots$$

But the distribution $\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$ is stationary.

- The block diagonal transition matrix

$$P = \begin{pmatrix} 1/2 & 2/3 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 0 & 1/4 & 4/5 \\ 0 & 0 & 3/4 & 1/5 \end{pmatrix}$$

does not have a limit distribution. Distributions of the form $(p_1 \ p_2 \ 0 \ 0)^t$ will always be mapped to distributions of the same form and similar for $(0 \ 0 \ p_1 \ p_2)^t$. One might be able to define limit distributions for the corresponding subspaces. But no global limit exists that attracts *all* initial distribution.

While the existence of a stationary distribution does not guarantee the existence of a limit distribution, the implication in the other direction is true:

Proposition Let P be a finite left stochastic matrix. If P has a limit distribution π , then π is also the unique stationary distribution.

Proof of existence:

By assumption:

$$\pi = \lim_{n \rightarrow \infty} P^n v$$

Hence:

$$P\pi = P \lim_{n \rightarrow \infty} P^n v = \lim_{n \rightarrow \infty} P^{n+1} v = \pi$$

Proof of uniqueness:

Assume the existence of another stationary distribution $q \neq \pi$ then by assumption $Pq = q$ and hence:

$$\lim_{n \rightarrow \infty} P^n q = q$$

The definition of the limit distribution π , however, implies

$$\lim_{n \rightarrow \infty} P^n q = \pi$$

since the limit distribution attracts all other distributions. This is a contradiction and thus no such q can exist.

We are now interested in finding conditions for the existence of a limit distribution. The following theorem will be helpful:

Theorem (*Perron-Frobenius*)

Let Q be a $K \times K$ left stochastic matrix with $Q_{ij} > 0 \ \forall i, j$. Then:

- 1 is a simple eigenvalue to Q (i.e. it has multiplicity 1).
- The absolute values of all other eigenvalues of Q are strictly smaller than 1.

We will not prove this theorem here but continue with a proposition concerning the question of existence of a limit distribution.

Proposition Let Q be a $K \times K$ left stochastic matrix with $Q_{ij} > 0 \ \forall i, j$. Then Q has a limit distribution q with the property $q_i > 0 \ \forall i$.

We prove this theorem for the special case that Q has a complete set of eigenvectors. We can then take advantage of the eigenvalue decomposition of Q . In the more general case one has to use a Jordan decomposition instead. Similar arguments will lead to the same result.

Proof of existence:

According to our additional assumption there exists an invertible matrix S s.t.

$$Q = S \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{pmatrix} S^{-1}$$

where $\lambda_1, \dots, \lambda_K$ are the eigenvalues of Q . Furthermore, according to the Perron-Frobenius Theorem we know that one of these eigenvalues is 1 while the others have absolute value strictly smaller than 1. Without loss of generality we choose to order them in the following way:

$$\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_K|$$

For the limit distribution we are interested in high powers of the transition matrix Q . For high powers one obtains:

$$\lim_{n \rightarrow \infty} \lambda_1^n = 1 \text{ and } \lim_{n \rightarrow \infty} \lambda_k^n = 0 \quad \forall k \neq 1$$

We then find for high powers of the transition matrix:

$$\begin{aligned} \tilde{Q} &= \lim_{n \rightarrow \infty} Q^n \\ &= \lim_{n \rightarrow \infty} S \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{pmatrix} \underbrace{S^{-1}S}_1 \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{pmatrix} S^{-1} \dots S \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{pmatrix} S^{-1} \\ &= \lim_{n \rightarrow \infty} S \begin{pmatrix} \lambda_1^n & & \\ & \ddots & \\ & & \lambda_K^n \end{pmatrix} S^{-1} \\ &= S \begin{pmatrix} 1 & & \\ & 0 & \\ & & \ddots & \\ & & & 0 \end{pmatrix} S^{-1} \end{aligned}$$

Writing out this matrix multiplication element wise one gets:

$$\tilde{Q}_{jk} = S_{j1} [S^{-1}]_{1k}$$

Furthermore we know that powers of a stochastic matrix are stochastic matrices as well. Hence \tilde{Q} is a stochastic matrix which means $\tilde{Q}_{jk} \geq 0 \quad \forall j, k$ and columns of \tilde{Q} sum to 1:

$$\begin{aligned} 1 &= \sum_{j=1}^K \tilde{Q}_{jk} = \sum_j S_{j1} [S^{-1}]_{1k} \\ \rightarrow [S^{-1}]_{1k} &= \frac{1}{\sum_j S_{j1}} \\ \rightarrow \tilde{Q}_{jk} &= \frac{S_{j1}}{\sum_j S_{j1}} =: q_j \end{aligned}$$

Realize that this expression does not depend on the column k , i.e. all columns of \tilde{Q} are the same:

$$\tilde{Q} = \begin{pmatrix} | & | & & | \\ q & q & \cdots & q \\ | & | & & | \end{pmatrix}$$

Also note that since \tilde{Q} is a stochastic matrix we have $q_j \geq 0$ and $\sum_j q_j = 1$. Thus, q is a valid probability vector. Furthermore, q is indeed the desired limit distribution, i.e.

$$\lim_{n \rightarrow \infty} Q^n v = \tilde{Q}v = q \quad \forall \text{ probability vectors } v$$

This can be seen in the following way:

$$\begin{aligned} [\tilde{Q}v]_k &= \sum_{j=1}^K \tilde{Q}_{kj} v_j \\ &= \sum_{j=1}^K q_k v_j \\ &= q_k \underbrace{\sum_{j=1}^K v_j}_1 \\ &= q_k \\ \rightarrow \tilde{Q}v &= q \end{aligned}$$

At this point, we only know $q_j \geq 0 \quad \forall j$ since q is a probability vector. In the next part of the proof we shall show that indeed $q_j > 0$.

Proof of strict positivity:

From $q_j \geq 0 \quad \forall j$ and $\sum_j q_j = 1$ we can follow that there must exist one index l with $q_l > 0$ (otherwise the sum over all q_j could not be strictly larger than zero). From a previous proposition we also know that the limit distribution q is also the stationary distribution of Q . Hence:

$$\begin{aligned} Qq &= q \\ \rightarrow q_j &= \sum_k Q_{jk} q_k \\ &= \underbrace{Q_{jl}}_{>0} \underbrace{q_l}_{>0} + \sum_{k \neq l} \underbrace{Q_{jk}}_{>0} \underbrace{q_k}_{\geq 0} \\ &\geq Q_{jl} q_l \\ &> 0 \end{aligned}$$

Thus, from the strict positivity of one single element of q we can conclude the strict positivity of all elements of q . \square

One huge disadvantage of the proposition in its current form is the requirement $Q_{jk} > 0$ which is generally, or even typically, not satisfied. Consequently, we next strive for a proposition of greater generality. We start with some remarks that will turn out to be useful.

- $\|v\|_1 := \sum_k |v_k|$: l_1 - norm (also called “Manhattan norm” or “bounded variation distance”)
- $\|A\|_1 := \sup_{\|v\|_1=1} \|Av\|_1$
It follows: $\|Av\|_1 \leq \|A\|_1 \|v\|_1$

Lemma Given a $K \times K$ matrix A , then

$$\|A\|_1 = \max_{k=1,\dots,K} \underbrace{\sum_{j=1}^K |A_{jk}|}_{\text{sum of column } k}$$

Corollary If P is a left stochastic matrix, then $\|P\|_1 = 1$ (since each column is a probability vector).

Lemma Let P be a $K \times K$ left stochastic matrix and q be a stationary distribution to P . Then:

$$\|P^{n+1}v - q\|_1 \leq \|P^n v - q\|_1$$

This means, at each iteration we get either closer to the stationary distribution or stay at the same distance but never go farther away.

Proof:

$$\begin{aligned} \|P^{n+1}v - q\|_1 &= \|P^{n+1}v - Pq\|_1 \\ (\text{since } Pq = q) &\rightarrow \|P(P^n v - q)\|_1 \\ (\text{by the previous lemma}) &\rightarrow \underbrace{\|P\|_1}_1 \|P^n v - q\|_1 \\ &= \|P^n v - q\|_1 \end{aligned}$$

The following definition introduces the property that we require in our more general proposition concerning the existence of a limit distribution.

Definition Let P be a transition matrix. The corresponding Markov chain is called *regular* if $\exists n$ s.t.:

$$[P^n]_{jk} > 0 \quad \forall j, k$$

This means, that after sufficiently many iterations any element in the transition graph is connected with any other element.

Finally, we can state and prove the desired proposition.

Proposition Let P be a transition matrix for a finite, regular Markov chain. Then P has a limit distribution q s.t. $q_j > 0 \quad \forall j$

The only difference to the previous proposition is the less restrictive condition. Instead of requiring that each element is connected with any other element already by one application of the transition matrix, we only require this for a sufficiently large number of iterations.

Proof:

By assumption of regularity we know that $\exists n$ s.t. $[P^n]_{jk} > 0$. We define $Q := P^n$ and can use the previous proposition for Q . That is, we know $\exists q$ s.t.

$$\begin{aligned} \lim_{m \rightarrow \infty} Q^m v &= q \quad \forall v \\ \rightarrow \lim_{m \rightarrow \infty} P^{n \cdot m} v &= q \quad \forall v \end{aligned}$$

However, we are actually interested in the limit $\lim_{l \rightarrow \infty} P^l v = q$ of which $\lim_{m \rightarrow \infty} P^{n \cdot m} v$ is only a subsequence. The convergence of the subsequence does not guarantee the convergence of the full sequence. Fortunately, we can use the previous lemma (stating that by application of P we never get farther away from the limit distribution) to conclude the convergence of the full sequence. To this end, write

$$l = n \cdot c_n(l) + r_n(l)$$

where $c_n(l)$ is the (integer) number of times that n fits into l and $r_n(l)$ is the remainder (e.g. if $l = 11$ and $n = 2 \rightarrow 11 = 2 \cdot 5 + 1$). Note, that as $l \rightarrow \infty$ also $c_n(l) \rightarrow \infty$. We can now write:

$$\begin{aligned} 0 \leq \|P^l v - q\|_1 &= \|P^{n \cdot c_n(l) + r_n(l)} v - q\|_1 \\ \text{(by using the aforementioned lemma)} \rightarrow &\leq \|P^{n \cdot c_n(l)} v - q\|_1 \\ &= \|Q^{c_n(l)} v - q\|_1 \\ &\rightarrow 0 \\ &\quad l \rightarrow \infty \end{aligned}$$

So we found $\|P^l v - q\|_1 \xrightarrow{l \rightarrow \infty} 0$ and can conclude the desired convergence $P^l v \xrightarrow{l \rightarrow \infty} q$. The property $q_j > 0 \quad \forall j$ is also inherited from the previous proposition. \square

2.2.4 Time averages

Similar to the motivational part of section 1.1.4 we want to consider the loose statement “*time average is equal to ensemble average*”. One way to use this statement in practice is to approximate an inaccessible ensemble average with an average over a finite time. We will make statements concerning the (time) average for *independent and identically distributed (i.i.d.)* variables as well as for sequences generated by a Markov chain.

First, consider the i.i.d. variables $X_0, X_1, X_2, X_3, \dots$. I.i.d. means that all variables are independently drawn from the same distribution. If we consider this as a time sequence, this case is even “more memoryless” than a Markov chain. While in a Markov chain, the current state depends only on the previous state, even this dependence is lost for i.i.d. variables, i.e. a new state is drawn completely at random without considering any previous state. We also assume that the expectation of X exists: $E[X] < \infty$ (more precisely the

expectation of any X_i which is the same for all i). Then the time average, or in other words simply the mean of $X_0, X_1, X_2, X_3, \dots$ converges to $E[X]$:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^K X_k = E[X] \text{ a.s.}$$

The *a.s.* stands for *almost surely*, stating that there are sequences $X_0, X_1, X_2, X_3, \dots$ for which the mean does not converge to $E[X]$, but these sequences are so unlikely that if we would pick a sequence at random the probability to pick one of those sequences would be zero. Formally, this can be stated as (recall the abstract definition of a random variable: $X : \Omega \rightarrow \mathbb{R}$)

$$P \left(\left\{ \omega \in \Omega : \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^K X_k(\omega) = E[X] \right\} \right) = 1$$

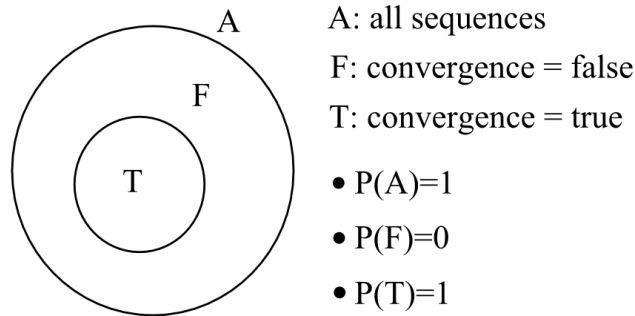
This statement is also known as the *strong law of large numbers*.

Example:

Consider a coin flipping experiment. To generate the sequence $X_0, X_1, X_2, X_3, \dots$ we flip the coin again and again (infinitely often). We assume that the probabilities for 'heads' and 'tails' do not change with time and that all flips are independent of each other. This means, that the variables are indeed i.i.d. Let us say that 'heads' corresponds to 0 and 'tails' to 1 and that the coin is fair:

$$\begin{aligned} X(\text{heads}) &= 0, \quad X(\text{tails}) = 1 \\ \text{and } P(\text{heads}) &= \frac{1}{2}, \quad P(\text{tails}) = \frac{1}{2} \\ \rightarrow E[X] &= \frac{1}{2} \end{aligned}$$

For a sequence like $(1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, \dots)$ it is reasonable that the mean $\frac{1}{K} \sum_{k=1}^K X_k$ converges to $E[X] = \frac{1}{2}$. But of course we could also have the sequence $(1, 1, 1, 1, 1, 1, 1, 1, \dots)$ for which the mean is 1. However, the latter sequences are very unlikely to occur (probability zero). The following picture serves as an illustration.



For Markov chains, a similar law exists.

Proposition Let P be an $N \times N$ transition matrix of a finite, regular Markov chain with limit distribution q . Then:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} X_k = \sum_{n=1}^N x_n q_n \text{ a.s.}$$

So the time average over an infinite chain is equal to the average over the limit distribution (the latter meaning the average of the values x_n corresponding to state n weighted with the probability q_n to be in that state in the limit distribution). The statement can be generalized to functions f of the variables X_k :

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(X_k) = \sum_{n=1}^N f(x_n) q_n \text{ a.s.}$$

Example 1:

Let f be an indicator function

$$I^{(m)}(x) = \begin{cases} 1 & , x = m \\ 0 & , x \neq m \end{cases}$$

(you can also think of the Kronecker-Delta). Then:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} I^{(m)}(X_k) = \sum_{n=1}^N \underbrace{I^{(m)}(x_n)}_{\delta_{nm}} q_n = q_m$$

With the indicator function we simply count the number of times we visit the state labeled by m . With the prefactor $1/K$ we get the relative frequency of occurrence of state m , which is given by the probability of the state in the limit distribution.

Example 2:

Consider again the person with the simple lifestyle that consists only of the activities 'study', 'party' and 'sleep'. One could be interested in the average money that this person spends per day. The function f then associates the spent money, say in €, on a single day with this day's activity, e.g. $f(\text{study}) = 10$, $f(\text{party}) = 50$ and $f(\text{sleep}) = 0$. For the time average one finds:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(X_k) = q_{\text{study}} \cdot 10 + q_{\text{party}} \cdot 50 + q_{\text{sleep}} \cdot 0$$

Note, that the formula only makes a statement for the long (infinite) time limit. For finite times, the time average will approach the limit value as time goes on but we made no statement about the magnitude of fluctuations, i.e. the speed of convergence.

Also consider the analogy to invariant measures in deterministic (chaotic) processes. The measure $\mu(A)$ counted the relative number of visits of the orbit x_0, x_1, x_2, \dots in an arbitrary set A . For the average of a function $f(x_k)$ we found:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(x_k) = \int_0^1 f(x) \rho(x) dx \text{ a.e.}$$

if the invariant measure has a density ρ . In this analogy, the invariant density plays the role of the limit (or stationary) distribution in stochastic processes. The *a.e.* stands for *almost everywhere* which has essentially the same meaning as *a.s* (*almost surely*) but for measures instead of probabilities.

Example:

For the logistic map $f(x) = rx(1-x)$ at $r = 4$ we found the invariant density

$$\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

2.2.5 Recurrence time

Consider a Markov chain with sequence $(X_k)_{k=0,1,2,\dots}$. Assuming that we start in $X_0 = j$, what is the time required to revisit state j ? This time is called the *recurrence time* (of state j):

$$T_j = \min_k \{k \geq 1, X_k = j\}$$

Since in stochastic processes transitions are governed by probabilities the recurrence time will in general be different for different realizations of the chain. One is therefore often interested in the *mean recurrence time*:

$$\tau_j = E[T_j \mid X_0 = j]$$

If $P(T_j < \infty \mid X_0 = j) = 1$ (meaning that we will eventually return to state j), then the state is called *recurrent*. If all states are recurrent, then the whole chain is called recurrent.

Lemma If a finite Markov chain is regular, then:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} I^{(m)}(X_k) = q_m > 0$$

and thus the chain is recurrent.

The implication of the lemma can be seen in the following way: If the chain would not be recurrent, then the number of occurrences $\sum_{k=0}^{K-1} I^{(m)}(X_k)$ of state m would be either 0 or 1 (depending on the initial state X_0). But this would lead to $q_m = 0$ which contradicts the assumption of regularity. One can even find a very simple expression for the mean recurrence time τ_m :

Proposition Consider a finite and regular Markov chain with $N \times N$ transition matrix P and limit distribution q . Then:

$$\tau_j = \frac{1}{q_j} \quad \forall j = 1, \dots, N$$

Proof:

- First, observe that by definition of the mean recurrence time, we can write:

$$\begin{aligned}
\tau_j &= E[\tau_j \mid X_0 = j] \\
&= \sum_{k=1}^{\infty} k P(T_j = k \mid X_0 = j) \\
&= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(T_j = k \mid X_0 = j) \\
&= \sum_{n=1}^{\infty} P(T_j \geq n \mid X_0 = j)
\end{aligned}$$

- To see the step from line 2 to line 3, introduce the shorthand notation $b_k := P(T_j = k \mid X_0 = j)$ and write:

$$\begin{aligned}
\sum_{k=1}^{\infty} k b_k &= b_1 + 2b_2 + 3b_3 + \dots \\
&= b_1 + b_2 + b_3 + \dots \left(\leftarrow \sum_{k=1}^{\infty} b_k \right) \\
&\quad + b_2 + b_3 + \dots \left(\leftarrow \sum_{k=2}^{\infty} b_k \right) \\
&\quad + b_3 + \dots \left(\leftarrow \sum_{k=3}^{\infty} b_k \right) \\
&\quad \vdots \\
&= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} b_k
\end{aligned}$$

- To see the step from line 3 to line 4, note that the sum of the probabilities that $T_j = k$ for $k = n, n+1, \dots$ is simply the probability that $T_j \geq n$.
- Now consider the product $\tau_j q_j$. We want to show that this product is 1, implying $\tau_j = \frac{1}{q_j}$. Also assume that our initial value is distributed according to the limit distribution: $P(X_0 = j) = q_j$.

$$\begin{aligned}
\tau_j q_j &= E[T_j \mid X_0 = j] q_j \\
&= \sum_{n=1}^{\infty} P(T_j \geq n \mid X_0 = j) P(X_0 = j) \\
&= \sum_{n=1}^{\infty} P(T_j \geq n, X_0 = j)
\end{aligned}$$

From line 1 to line 2 we used the above result and the assumption $P(X_0 = j) = q_j$. From line 2 to line 3 we used the general factorization of total probability into conditional probability and marginal probability.

- In the current expression $\sum_{n=1}^{\infty} P(T_j \geq n, X_0 = j)$, distinguish the cases $n = 1$ and $n \geq 2$:

$$\text{for } n = 1 : P \left(\underbrace{T_j \geq 1}_{\text{always true}}, X_0 = j \right) = P(X_0 = j)$$

$$\text{for } n \geq 2 : P(T_j \geq n, X_0 = j) = P(X_{n-1} \neq j, X_{n-2} \neq j, \dots, X_1 \neq j, X_0 = j)$$

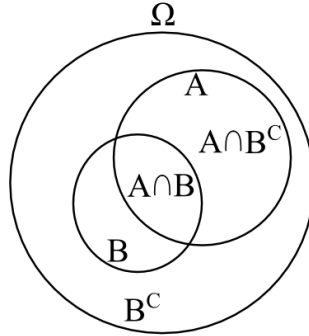
- General remark to rewrite joint probabilities:

$$P(A \cap B) = P(A) - P(A \cap B^C)$$

where B^C is the complement of B . To prove this formula, denote the total space by Ω and note that $B \cup B^C = \Omega$ and $B \cap B^C = \emptyset$. Then:

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P(A \cap (B \cup B^C)) \\ &= P((A \cap B) \cup (A \cap B^C)) \\ &= P(A \cap B) + P(A \cap B^C) \end{aligned}$$

The last equality holds since $B \cap B^C = \emptyset$ implies $(A \cap B) \cap (A \cap B^C) = \emptyset$ and probabilities of disjoint sets can simply be added. The following figure illustrates the set relations between Ω , A and B .



- To apply this general remark to $P(X_{n-1} \neq j, X_{n-2} \neq j, \dots, X_1 \neq j, X_0 = j)$ for $n \geq 2$, choose:

$$\begin{aligned} A_n &= \{X_{n-1} \neq j, \dots, X_1 \neq j\} \\ B &= \{X_0 = j\} \\ \rightarrow B^C &= \{X_0 \neq j\} \end{aligned}$$

One then obtains:

$$\begin{aligned} P \left(\underbrace{\overbrace{X_{n-1} \neq j, \dots, X_1 \neq j}^{A_n} \overbrace{X_0 = j}^B}_{A_n \cap B} \right) &= P(X_{n-1} \neq j, \dots, X_1 \neq j) \\ &\quad - P(X_{n-1} \neq j, \dots, X_1 \neq j, X_0 \neq j) \end{aligned}$$

- With all this remarks we can come back to our current expression for $\tau_j q_j$:

$$\begin{aligned}
\tau_j q_j &= \sum_{n=1}^{\infty} P(T_j \geq n, X_0 = j) \\
&= P(X_0 = j) + \sum_{n=2}^{\infty} P(X_{n-1} \neq j, \dots, X_1 \neq j, X_0 = j) \\
&= P(X_0 = j) + \sum_{n=2}^{\infty} [P(X_{n-1} \neq j, \dots, X_1 \neq j) - P(X_{n-1} \neq j, \dots, X_1 \neq j, X_0 \neq j)]
\end{aligned}$$

- Now, recall that we chose as initial condition the limit distribution. Since the limit distribution is also stationary we remain in this distribution indefinitely. This allows us to use a kind of translational invariance:

$$P(X_{n-1} \neq j, \dots, X_1 \neq j) = P(X_{n-2} \neq j, \dots, X_0 \neq j)$$

- Also introduce the notation

$$a_n := P(X_n \neq j, \dots, X_0 \neq j)$$

- Using these insights, we can write:

$$\begin{aligned}
\tau_j q_j &= P(X_0 = j) + \sum_{n=2}^{\infty} [P(X_{n-1} \neq j, \dots, X_1 \neq j) - P(X_{n-1} \neq j, \dots, X_1 \neq j, X_0 \neq j)] \\
&= P(X_0 = j) + \sum_{n=2}^{\infty} [P(X_{n-2} \neq j, \dots, X_0 \neq j) - P(X_{n-1} \neq j, \dots, X_1 \neq j, X_0 \neq j)] \\
&= P(X_0 = j) + \sum_{n=2}^{\infty} (a_{n-2} - a_{n-1})
\end{aligned}$$

- This is a telescoping sum of which only the first and “last” summand survive. For a finite sum:

$$\sum_{n=2}^m (a_{n-2} - a_{n-1}) = a_0 - a_1 + a_1 - a_2 + a_2 - a_3 + \dots + a_{m-2} - a_{m-1} = a_0 - a_{m-1}$$

- Also note that $a_0 = P(X_0 \neq j)$ and

$$\lim_{m \rightarrow \infty} a_{m-1} = \lim_{m \rightarrow \infty} P(X_{m-1} \neq j, \dots, X_0 \neq j) = 0$$

since we already know from the previous lemma (stating that regular Markov chains are recurrent) that we will eventually return to the state j .

- We can finally state the desired result:

$$\begin{aligned}
 \tau_j q_j &= P(X_0 = j) + \sum_{n=2}^{\infty} (a_{n-2} - a_{n-1}) \\
 &= P(X_0 = j) + \lim_{m \rightarrow \infty} \sum_{n=2}^m (a_{n-2} - a_{n-1}) \\
 &= P(X_0 = j) + a_0 - \lim_{m \rightarrow \infty} a_{m-1} \\
 &= P(X_0 = j) + P(X_0 \neq j) - 0 \\
 &= 1
 \end{aligned}$$

Hence, $\tau_j = 1/q_j$

□

Next, we move on to another useful property of (some) Markov chains that is related to stationary (and limit) distributions. It finds application in physical systems and in the approximation of hardly accessible expectation values.

2.2.6 Reversible Markov chains

Consider a chain X_0, X_1, X_2, \dots with limit distribution q . Assume that we start in the limit distribution $P(X_0 = j) = q_j$. Then, since q is stationary, we also find $P(X_n = j) = q_j$ and even

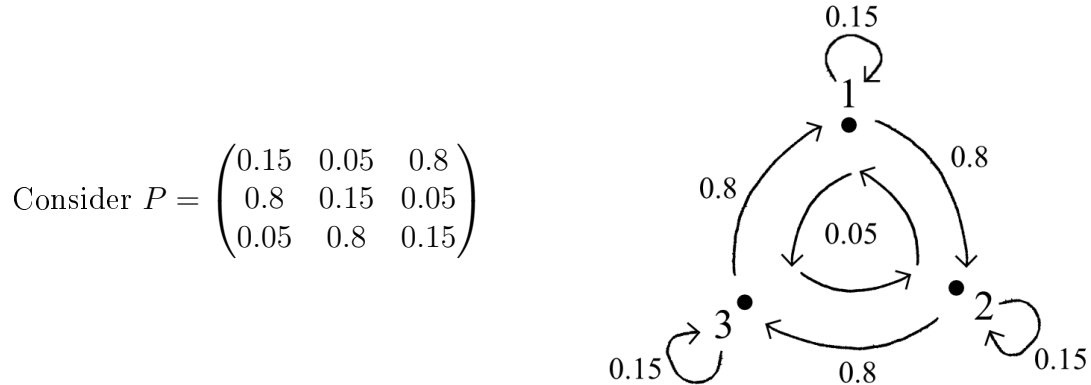
$$P(X_k = j_k, X_{k-1} = j_{k-1}, \dots, X_0 = j_0) = P(X_{k+n} = j_k, \dots, X_n = j_0)$$

i.e. we have a translational invariance. But what happens if we do not shift but inverse the order of the variables?

$$X_0, X_1, \dots, X_K \rightarrow Y_0 = X_K, Y_1 = X_{K-1}, \dots, Y_K = X_0$$

If $P(Y_k = j_K, \dots, Y_0 = j_0) = P(X_K = j_K, \dots, X_0 = j_0)$, i.e. if we see no difference when reversing the direction of time (while being in the limit distribution), then the chain is called *reversible*.

Counter example:



This chain is not reversible, as can be seen by the clockwise net flow in the plot. If no such net flow existed, the chain would be reversible. This idea is formalized in the following proposition.

Proposition A finite Markov chain with transition matrix P and limit distribution q is reversible if and only if:

$$P_{jk}q_k = P_{kj}q_j \quad \forall j, k$$

Proof: “only if” part:

We assume that the chain is reversible and want to show that $P_{jk}q_k = P_{kj}q_j \quad \forall j, k$ is satisfied. If we consider the joint distribution of two succeeding variables then the reversibility states:

$$P(Y_1 = j_1, Y_0 = j_0) = P(X_1 = j_1, X_0 = j_0)$$

By using the correspondences $Y_1 = X_0$ and $Y_0 = X_1$ we can write this as:

$$P(X_1 = j_0, X_0 = j_1) = P(X_1 = j_1, X_0 = j_0)$$

(note that one can reorder the arguments of a joint probability as one likes). Then:

$$\begin{aligned} P_{j_0j_1}q_{j_1} &= P(X_1 = j_0 \mid X_0 = j_1) P(X_0 = j_1) \\ &= P(X_1 = j_0, X_0 = j_1) \\ &= P(X_1 = j_1, X_0 = j_0) \\ &= P(X_1 = j_1 \mid X_0 = j_0) P(X_0 = j_0) \\ &= P_{j_1j_0}q_{j_0} \end{aligned}$$

The first line is simply the definition of the transition matrix via the conditional probability $P_{jk} = P(X_1 = j \mid X_0 = k)$ and the assumption that we start in the limit distribution. Line 1 to line 2 uses the decomposition of a joint probability in terms of conditional and marginal probabilities. Next, the reversibility assumption is used. The last two steps are the same as the first two steps but in backward direction.

“if” part:

Now, we assume $P_{jk}q_k = P_{kj}q_j \quad \forall j, k$ and want to show the reversibility of the chain:

$$\begin{aligned} &P(Y_K = j_K, \dots, Y_0 = j_0) \\ \text{rename } Y \leftrightarrow X &= P(X_0 = j_K, \dots, X_K = j_0) \\ \text{reorder variables} &= P(X_K = j_0, \dots, X_0 = j_K) \\ \text{definition of trans. matrix} &= P_{j_0j_1}P_{j_1j_2}\dots P_{j_{K-2}j_{K-1}}P_{j_{K-1}j_K} \underbrace{P(X_0 = j_K)}_{q_{j_K}} \\ &\quad \underbrace{P_{j_Kj_{K-1}}q_{j_{K-1}}} \\ \text{assumption } P_{jk}q_k = P_{kj}q_j &= P_{j_0j_1}P_{j_1j_2}\dots P_{j_{K-2}j_{K-1}}P_{j_Kj_{K-1}}q_{j_{K-1}} \\ \text{again} &= P_{j_0j_1}P_{j_1j_2}\dots \underbrace{P_{j_{K-2}j_{K-1}}q_{j_{K-1}}}_{P_{j_{K-1}j_{K-2}}q_{j_{K-2}}}P_{j_Kj_{K-1}} \\ &\vdots \end{aligned}$$

This procedure can be iterated further. In this way, the indices of all matrix elements are

swapped and the initial q_{j_K} becomes q_{j_0} :

$$\begin{aligned}
 & P(Y_K = j_K, \dots, Y_0 = j_0) \\
 & \vdots \\
 & = P_{j_0 j_1} P_{j_1 j_2} \dots P_{j_{K-1} j_K} q_{j_K} P_{j_K j_{K-1}} \\
 & \vdots \\
 & = P_{j_1 j_0} q_{j_0} P_{j_2 j_1} \dots P_{j_{K-1} j_K} P_{j_K j_{K-1}} \\
 & = P_{j_K j_{K-1}} P_{j_{K-1} j_{K-2}} \dots P_{j_2 j_1} P_{j_1 j_0} q_{j_0} \\
 & = P(X_K = j_K, \dots, X_0 = j_0)
 \end{aligned}$$

But this is just the defining criterion of reversibility. \square

The condition $P_{jk}q_k = P_{kj}q_j \quad \forall j, k$ is often referred to as *detailed balance*. One can easily show that if a transition matrix P and a distribution p are in detailed balance, then p is stationary (but not necessarily the limit distribution):

$$\sum_k P_{jk} p_k = \sum_k \underbrace{P_{kj} p_j}_1 = p_j \rightarrow Pp = p \quad \square$$

We can also make a short excursion to physical systems.

Consider a physical system with states $1, \dots, N$ and associated energies $E(n)$, then

$$\begin{aligned}
 G_n &= \frac{1}{Z} e^{-\beta E(n)} \\
 \text{with } Z &= \sum_n e^{-\beta E(n)} \quad (\text{partition function}) \\
 \text{and } \beta &= \frac{1}{k_B T} \quad (\text{inverse temperature})
 \end{aligned}$$

is the *equilibrium distribution*. The term 'equilibrium' is often used interchangeably with 'limit' or 'stationary'. A Markov chain with transition matrix P describing the dynamics of the system is in detailed balance with G : $P_{jk}G_k = P_{kj}G_j$.

We now come to an approximation method of expectation values that are difficult to calculate analytically (or numerically with brute force).

2.2.7 Markov chain Monte Carlo simulation

Consider some function $f(n)$ and a (limit) distribution q . We are interested in the expectation value

$$\langle f \rangle_q = \sum_n f(n) q_n$$

and assume that this expression is very difficult, if not impossible, to evaluate. This might happen if the state space is infinite and the function f as well as the distribution q do not

behave well enough. If there exists a Markov chain with q as its limit distribution then we know (“time average equals ensemble average”):

$$\sum_n f(n) q_n = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(X_k)$$

The right hand side for finite K can then be used as an approximation to the desired expectation value.

Metropolis algorithm

We want to construct a transition matrix P s.t. q is in detailed balance with P ($P_{jk}q_k = P_{kj}q_j$). Then q is also stationary under P and with some further conditions also the limit distribution.

- First, choose a symmetric transition matrix Q ($Q_{nm} = Q_{mn}$).
- Given that we are in state n_k , then:
 1. Pick a state m with probability Q_{m,n_k} at random.
 2. With probability $\min \left\{ 1, \frac{q_m}{q_{n_k}} \right\}$ accept the move $n_{k+1} = m$, otherwise stay in the current state: $n_{k+1} = n_k$.

The resulting process is a Markov chain with a transition matrix that is in detailed balance with q . One finds:

$$P_{n',n} = \begin{cases} \min \left\{ 1, \frac{q_{n'}}{q_n} \right\} Q_{n',n} & \text{if } n' \neq n \\ Q_{n,n} + \sum_{m \neq n} \left(1 - \min \left\{ 1, \frac{q_m}{q_n} \right\} \right) Q_{m,n} & \text{if } n' = n \end{cases}$$

For $n' \neq n$, $P_{n',n}$ is simply the product of the probability to pick state n' (in step 1) and the probability to accept this state (in step 2). For $n' = n$, $Q_{n,n}$ gives the probability to pick state n in step 1, which will always be accepted. But we can also stay in state n if we pick another state m (with probability $Q_{m,n}$) but reject the move which happens with probability $1 - \min \left\{ 1, \frac{q_m}{q_n} \right\}$. We still have to check that P is indeed a transition matrix that is in detailed balance with q .

- $P_{n',n} \geq 0$ is easy to see as we only have sums and products of non-negative numbers.

- Columns sum to one:

$$\begin{aligned}
\sum_{n'} P_{n',n} &= P_{n,n} + \sum_{n' \neq n} P_{n',n} \\
&= Q_{n,n} + \sum_{m \neq n} \left(1 - \min \left\{1, \frac{q_m}{q_n}\right\}\right) Q_{m,n} + \sum_{n' \neq n} \min \left\{1, \frac{q_{n'}}{q_n}\right\} Q_{n',n} \\
&= Q_{n,n} + \sum_{m \neq n} Q_{m,n} - \underbrace{\sum_{m \neq n} \min \left\{1, \frac{q_m}{q_n}\right\} Q_{m,n} + \sum_{n' \neq n} \min \left\{1, \frac{q_{n'}}{q_n}\right\} Q_{n',n}}_0 \\
&= \sum_m Q_{m,n} \\
&= 1
\end{aligned}$$

In the last step we used that Q is by assumption a transition matrix and thus its columns sum to one.

- Confirm detailed balance $P_{n',n}q_n = P_{n,n'}q_{n'}$: If $n' = n$, the equality is trivial. So consider $n' \neq n$:

$$\begin{aligned}
P_{n',n}q_n &= \min \left\{1, \frac{q_{n'}}{q_n}\right\} Q_{n',n}q_n \\
&= \min \{q_n, q_{n'}\} Q_{n',n} \\
&= \min \{q_n, q_{n'}\} Q_{n,n'} \\
&= \min \left\{1, \frac{q_n}{q_{n'}}\right\} Q_{n,n'}q_{n'} \\
&= P_{n,n'}q_{n'}
\end{aligned}$$

From line 1 to line 2 we simply shifted the non-negative q_n inside the minimum. Next, we used the assumption that Q is symmetric: $Q_{n',n} = Q_{n,n'}$. Finally, one can pull $q_{n'}$ outside the minimum which yields the desired expression. \square

Let us consider again a physical system in equilibrium:

$$G_n = \frac{1}{Z} e^{-\beta E(n)}$$

We might be interested in the expectation value:

$$\langle f \rangle_G = \sum_n f(n) G_n$$

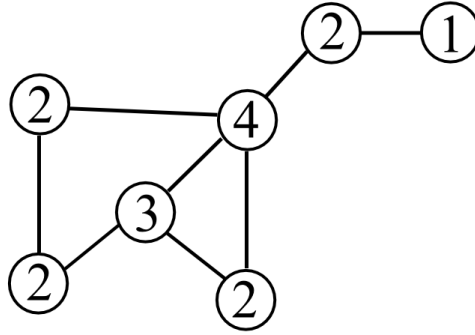
Apply the Metropolis algorithm (assume we are in state n_k and we have chosen an appropriate transition matrix Q):

1. Pick a state m with probability Q_{m,n_k} .
2. With probability $\min \{1, e^{-\beta[E(m) - E(n_k)]}\}$ accept the move $n_{k+1} = m$.

- We always accept moves to smaller energies (because the argument of the exponential function is positive in this case)!
- Moves to larger energies are allowed, but the probability to accept the move decreases with increasing energy gap.
- Also note that the partition function Z does not appear in the acceptance probability since it dropped out when taking the quotient of q_m and q_{n_k} . This property can be very helpful since Z is often extremely hard to evaluate.

2.2.8 Random walks on graphs

A graph consists of *nodes* (also called vertices) and *edges* connecting the nodes. The number of outgoing edges of a node (that is the number of connected neighbours) is called its *degree*. In the following example the degree is written inside the nodes.



Now, assume that the probability to move from one node to any of its neighbours is given by the reciprocal of its degree. The transition matrix of this process reads:

$$P_{jk} = \begin{cases} \frac{1}{\deg(k)} & \text{if } j \text{ is connected to } k \\ 0 & \text{otherwise} \end{cases}$$

The stationary distribution q is given by:

$$q_j = \frac{\deg(j)}{\sum_k \deg(k)}$$

The sum in the denominator goes over all vertices of the graph (it has to do so in order for q to be normalized to 1). Introduce the short hand notation $D := \sum_k \deg(k)$. To prove that q is stationary one can equivalently prove that q is in detailed balance with P :

$$P_{jk}q_k = \frac{\deg(k)}{D} \cdot \begin{cases} \frac{1}{\deg(k)} & \text{if } j \text{ is connected to } k \\ 0 & \text{otherwise} \end{cases} = \frac{1}{D} \cdot \begin{cases} 1 & \text{if } j \text{ is connected to } k \\ 0 & \text{otherwise} \end{cases}$$

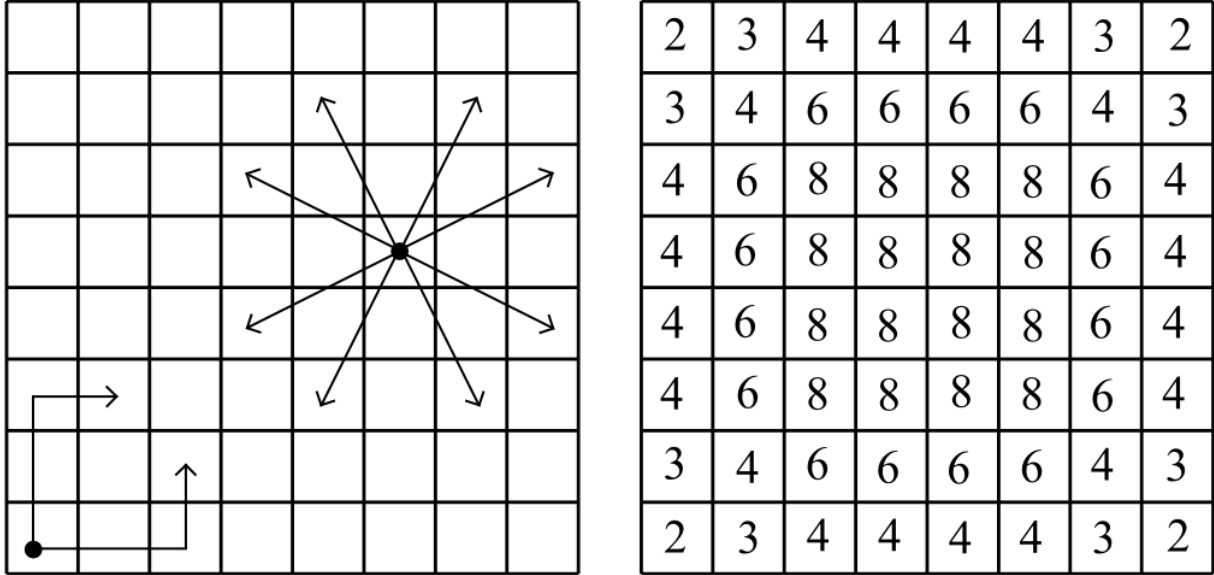
The final expression is symmetric in j and k (since j is connected to k if and only if k is connected to j). Hence, $P_{jk}q_k = P_{kj}q_j$. \square

As an example of a problem on a graph consider a knight on a chessboard. Starting at one position, what is the mean time to return to that position? The question might seem

challenging for a person inexperienced in Markov chains. But in fact, with our current knowledge it is very easy to answer. What we are looking for is the mean recurrence time τ_j of a state j . We can use the proposition that the mean recurrence time is the reciprocal of the probability to be in the state in the limit (or stationary) distribution:

$$\tau_j = \frac{1}{q_j} = \frac{\sum_k \deg(k)}{\deg(j)}$$

So all we have to do is to count the degree of all positions on the chessboard assuming the allowed movements of a knight. A knight is only allowed to make moves that consist of two steps in one direction and one step in a perpendicular direction. This rule and the degrees of all nodes are shown in the following figures.



One can calculate the total degree of all nodes to:

$$D = \sum_k \deg(k) = 4 \cdot 2 + 8 \cdot 3 + 20 \cdot 4 + 16 \cdot 6 + 16 \cdot 8 = 336$$

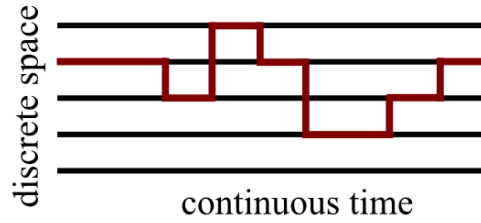
For a corner position that has degree 2 one finds:

$$\begin{aligned} q_{\text{corner}} &= \frac{2}{336} = \frac{1}{168} \\ \rightarrow \tau_{\text{corner}} &= 168 \end{aligned}$$

It requires on average 168 steps to return to a corner position!

2.3 Continuous time and discrete space Markov chains

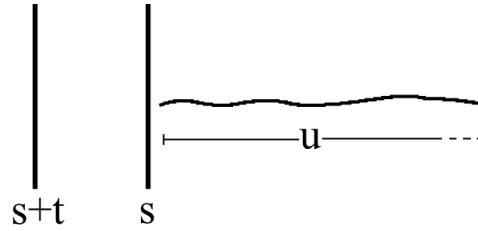
We have finished our studies of Markov chains with discrete time and discrete space. In this section we allow time to be continuous while the state space is still assumed to be discrete.



$\{X(t)\}_{t \geq 0}$ with discrete state space is a continuous time Markov chain if:

$$\begin{aligned} & P(X(s+t) = m \mid X(s) = n, X(u) = x(u), 0 \leq u < s) \\ &= P(X(s+t) = m \mid X(s) = n) \end{aligned}$$

$\forall s, t \geq 0, \forall m, n, x(u)$. That is, given that at time s we are in state n , the state at a later time $s+t$ does not depend on the history of the system before s .



A continuous time Markov chain is homogeneous if:

$$P_{mn}(t) := P(X(s+t) = m \mid X(s) = n) \quad \forall s \geq 0$$

This means that the evolution of the system shall depend only on the time difference t but not on the absolute value in time s .

Suppose at $t = 0$ the chain starts at state n (i.e. $X(0) = n$), then let σ_n denote the time it takes until the process leaves n . σ_n could be called the 'survival time' or 'waiting time'. Note that σ_n is a random variable since it will typically require a different amount of time to leave the state at each run of the chain.

Proposition Consider a homogeneous, continuous time Markov chain, then:

$$P(\sigma_n > s+t \mid \sigma_n > s) = P(\sigma_n > t)$$

The probability to leave the state in the future does not depend on the time already spent in the state.

Proof:

$$\begin{aligned} P(\sigma_n > s+t \mid \sigma_n > s) &= P(X(r) = n, r \in [0, s+t] \mid X(r') = n, r' \in [0, s]) \\ &= P(X(r) = n, r \in [s, s+t] \mid X(r') = n, r' \in [0, s]) \\ \text{used Markov property} &= P(X(r) = n, r \in [s, s+t] \mid X(s) = n) \\ \text{used homogeneity} &= P(X(r) = n, r \in [0, t] \mid X(0) = n) \\ &= P(\sigma_n > t) \end{aligned} \quad \square$$

This property implies that σ is distributed exponentially. The exponential distribution satisfies ($t \geq 0$):

$$\begin{aligned} P(\sigma > t) &= e^{-\lambda t} \\ \text{cumulative distr.: } P(\sigma \leq t) &= 1 - e^{-\lambda t} \\ \text{density function: } f_{\sigma}(t) &= \lambda e^{-\lambda t} \\ E[\sigma] &= \frac{1}{\lambda} \end{aligned}$$

This implication can be seen with the following non-rigorous argument:

- We will use without proof that in general if $P(\sigma > s + t \mid \sigma > s) = P(\sigma > t)$ is exponential, then:

$$\lim_{\Delta t \rightarrow 0} \frac{P(\sigma \leq \Delta t)}{\Delta t} = \lambda$$

- Then we can derive a differential equation for $P(\sigma > t)$ in the following way:

$$\begin{aligned} P(\sigma > t + \Delta t) &= P(\sigma > t + \Delta t \mid \sigma > t) P(\sigma > t) \\ &= P(\sigma > \Delta t) P(\sigma > t) \end{aligned}$$

where we simply used the previous proposition. Now, subtract $P(\sigma > t)$ from both sides of the equation, divide by Δt and consider the limit $\Delta t \rightarrow 0$:

$$\begin{aligned} P(\sigma > t + \Delta t) - P(\sigma > t) &= - \underbrace{[1 - P(\sigma > \Delta t)]}_{P(\sigma \leq \Delta t)} P(\sigma > t) \\ \Rightarrow \lim_{\Delta t \rightarrow 0} \frac{P(\sigma > t + \Delta t) - P(\sigma > t)}{\Delta t} &= - \underbrace{\lim_{\Delta t \rightarrow 0} \frac{P(\sigma \leq \Delta t)}{\Delta t}}_{\lambda} P(\sigma > t) \\ \Rightarrow \frac{d}{dt} P(\sigma > t) &= -\lambda P(\sigma > t) \end{aligned}$$

- The solution of this equation is just the exponential function:

$$\begin{aligned} P(\sigma > t) &= \underbrace{P(\sigma > 0)}_1 e^{-\lambda t} \\ \rightarrow P(\sigma \leq t) &= 1 - e^{-\lambda t} \end{aligned}$$

The latter is the cumulative distribution function of the exponential distribution.

The parameter λ is often called *rate* (in the lecture sometimes also *intensity*). Each state n has its own rate λ_n leading to an exponential distribution with density

$$f_{\sigma_n} = \lambda_n e^{-\lambda_n t}, \quad t \geq 0$$

for the survival time σ_n of state n . Be careful not to confuse σ_n and λ_n . σ_n is a random variable. Its value gives the time the system stays in state n until the next transition occurs. Since σ_n is a random variable this time is different for any visit of state n . The

expected value of this time is $E[\sigma_n] = \frac{1}{\lambda_n}$. In the context of “time” λ_n is not a time but a rate!

After the time σ_n has elapsed the system jumps (i.e. performs a transition) to another state m with probability:

$$p_{mn} := P(X(\sigma_n) = m \mid X(0) = n)$$

Note that by construction $p_{nn} = 0$ since σ_n is defined as the time after which we leave state n . We can choose the initial time as 0 since we are considering homogeneous Markov chains where only the time difference matters. Further remarks:

- $X(\sigma_n)$ is independent of σ_n in the sense that the probability for the next state does not depend on the waiting time. This follows by the Markov assumption that only the knowledge of the presents matters. We do not prove this statement here.
- As a consequence, in the sequence of transitions and waitings, all transitions and waitings are independent.
- Note the difference between $P_{mn}(t)$ and p_{mn} :
 - $P_{mn}(t) = P(X(t) = m \mid X(0) = n)$ is the probability to be in state m after time t given that we are currently in state n . An arbitrary number of transitions in between is allowed.
 - $p_{mn} = P(X(\sigma_n) = m \mid X(0) = n)$ is the probability that the **next** transition (which occurs after the waiting time σ_n) is to state m , given that we are currently in state n . There are no other transitions in between.
- For small times Δt one could expect that there is a strong relation between these quantities since for small times the probability for additional transitions should be small.

To make the last point precise, let us consider the “escape rate of state n ”, $\frac{1-P_{nn}(\Delta t)}{\Delta t}$, and the “escape rate from n to a specific state m ”, $\frac{P_{mn}(\Delta t)}{\Delta t}$:

Lemma

- $\lim_{\Delta t \rightarrow 0} \frac{1-P_{nn}(\Delta t)}{\Delta t} = \lambda_n$
- $\lim_{\Delta t \rightarrow 0} \frac{P_{mn}(\Delta t)}{\Delta t} = p_{mn}\lambda_n$

First, note that the second statement implies the first one (by summing over $m \neq n$ and realizing $\sum_{m \neq n} P_{mn}(\Delta t) = 1 - P_{nn}(\Delta t)$):

$$\begin{aligned}
 \lim_{\Delta t \rightarrow 0} \frac{1 - P_{nn}(\Delta t)}{\Delta t} &= \lim_{\Delta t \rightarrow 0} \sum_{m \neq n} \frac{P_{mn}(\Delta t)}{\Delta t} \\
 &= \sum_{m \neq n} \lim_{\Delta t \rightarrow 0} \frac{P_{mn}(\Delta t)}{\Delta t} \\
 &= \underbrace{\sum_{m \neq n} p_{mn}}_1 \lambda_n \\
 &= \lambda_n
 \end{aligned}$$

To see the last step, note that p_{mn} forms a probability vector for fixed n (since the probability to jump anywhere is 1) and $p_{nn} = 0$ by construction. For finite state spaces (and thus a finite sum) the swap of $\lim_{\Delta t \rightarrow 0}$ and $\sum_{m \neq n}$ is unproblematic. For infinite state spaces some technicalities may have to be taken into account. (This implication and hence the necessary arguments were not mentioned in the lecture!)

Furthermore, the statements of the lemma can be rewritten in terms of a Taylor expansion:

- $P_{nn}(\Delta t) = 1 - \lambda_n \Delta t + o(\Delta t)$
- $P_{mn}(\Delta t) = p_{mn} \lambda_n \Delta t + o(\Delta t)$

The *small-o* notation shall mean $\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$. Now, let us sketch the proof of the second statement of the above lemma. It will only be a sketch, since at one important step we assume without proof that for small times only the “one transition path” has to be taken into account while the probability for more transitions in between is negligible ($o(\Delta t)$).

$$\begin{aligned}
 P_{mn}(\Delta t) &= P(X(\Delta t) = m \mid X(0) = n) \\
 \text{unproven step} &= P(X(\sigma_n) = m, \sigma_n < \Delta t \mid X(0) = n) + o(\Delta t) \\
 \text{independence of } \sigma_n \text{ and } X(\sigma_n) &= \underbrace{P(X(\sigma_n) = m \mid X(0) = n)}_{p_{mn}} \underbrace{P(\sigma_n < \Delta t)}_{1 - e^{-\lambda_n \Delta t}} + o(\Delta t) \\
 \Rightarrow \frac{P_{mn}(\Delta t)}{\Delta t} &= \frac{p_{mn} (1 - e^{-\lambda_n \Delta t}) + o(\Delta t)}{\Delta t} \\
 &= \frac{p_{mn} (1 - 1 + \lambda_n \Delta t) + o(\Delta t)}{\Delta t} \\
 &\xrightarrow{\Delta t \rightarrow 0} p_{mn} \lambda_n
 \end{aligned}$$

In the second last step we used $e^x = 1 + x + o(x)$. □

The result of this lemma suggests the definition of the transition rate from n to m as:

$$\lambda_{mn} := p_{mn} \lambda_n$$

Remarks:

- Note again that since $p_{nn} = 0$ also $\lambda_{nn} = 0$.
- $\sum_m \lambda_{mn} = \underbrace{\sum_m p_{mn}}_1 \lambda_n = \lambda_n$
- $p_{mn} = \frac{\lambda_{mn}}{\lambda_n} = \frac{\lambda_{mn}}{\sum_{m'} \lambda_{m'n}}$

Finally, let us give an overview of the appearing (and related) quantities:

- $P_{mn}(t)$: probability to be in state m after time t given that we are currently in state n .

- p_{mn} : probability that the next state is m given that the current state is n .
- λ_{mn} : transition rate from state n to m .
- λ_n : total transition rate (“escape rate”) away from state n .

We now seek for a numerical implementation of the decomposition of a continuous time Markov chain in terms of waitings and transitions.

2.3.1 Gillespie algorithm

Consider that we want to simulate the continuous time, homogeneous Markov chain $\{X(t)\}_{t \geq 0}$. Assume that we know the rates λ_{mn} .

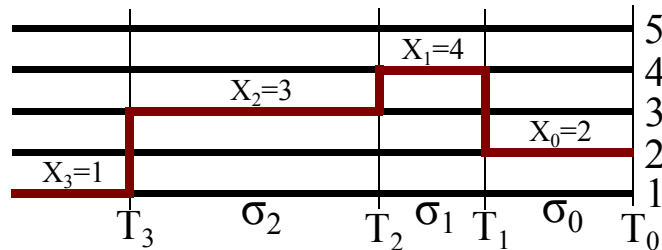
- Construct the discrete, homogeneous Markov chain $\{X_k\}_{k \in \mathbb{N}}$ with transition probabilities p_{mn} (obtained from the λ_{mn}) to generate a sequence of states.
- Construct a sequence of independent, exponentially distributed random variables $\{E_k\}_{k \in \mathbb{N}}$ with rate $\lambda = 1$.
- To obtain the final sequence of waiting times (or rather times at which the transitions occur), construct a sequence $\{T_k\}_{k \in \mathbb{N}}$. Start with $T_0 = 0$ and then calculate iteratively:

$$T_{k+1} = T_k + \frac{E_k}{\lambda_{X_k}}$$

The expression $\frac{E_k}{\lambda_{X_k}}$ is the waiting time σ_{X_k} . Since E_k was drawn from an exponential distribution with rate $\lambda = 1$, σ_{X_k} will (as desired) follow an exponential distribution with rate λ_{X_k} .

(In detail, this means that from $f_{E_k}(t) = e^{-t}$ it follows $f_{\frac{E_k}{\lambda_{X_k}}} = \lambda_{X_k} e^{-\lambda_{X_k} t}$. This transformation behaviour can for example be understood by considering expectation values: In general, the expectation value of the exponential distribution with rate λ is $\frac{1}{\lambda}$. Then $E[E_k] = 1$. From the linearity of the expectation value it follows: $E\left[\frac{E_k}{\lambda_{X_k}}\right] = \frac{1}{\lambda_{X_k}}$ (note that once the sequence $\{X_k\}$ is determined and we fix k , λ_{X_k} is just a constant). It is also intuitive that the resulting distribution should again be an exponential distribution. The expectation value then determines the rate $\lambda = \lambda_{X_k}$.)

- Finally define $X(t) = X_k$ for $t \in (T_k, T_{k+1}]$ to obtain the continuous time Markov chain. A graphical representation could be:



Application

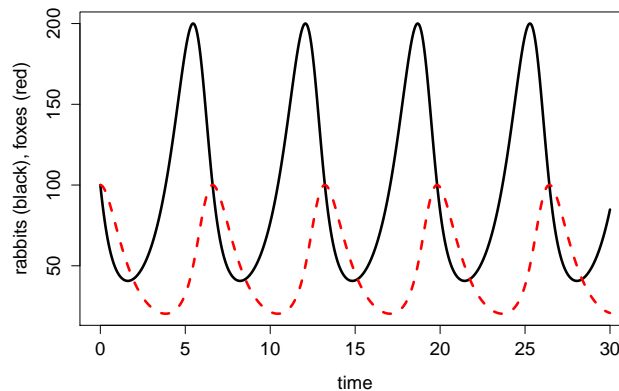
The Gillespie algorithm can be applied to the stochastic Lotka Volterra model. In general, the Lotka Volterra model describes populations of predator and prey. In the deterministic model the two populations are described by two coupled differential equations:

$$\begin{aligned}\frac{dR}{dt} &= aR - bRF \\ \frac{dF}{dt} &= cRF - gF\end{aligned}$$

R describes the population of the prey, for example rabbits, while F stands for the predator population (e.g. foxes). The single terms correspond to:

$$\begin{aligned}\text{rabbit reproduction rate:} & \quad aR \\ \text{fox reproduction rate:} & \quad cRF \\ \text{rabbit death rate:} & \quad bRF \\ \text{fox death rate:} & \quad gF\end{aligned}$$

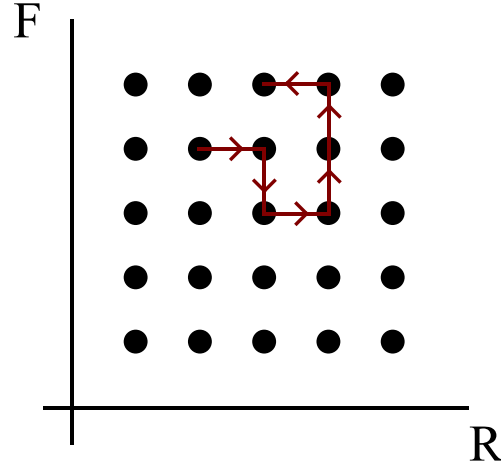
Starting with $R = 100$ rabbits and $F = 100$ foxes and choosing the parameters $a = 1$, $b = 0.02$, $c = 0.01$ and $g = 1$ one finds the following (numerical) solution of the system of equations:



The solution is perfectly periodic. In general, there are more rabbits than foxes (this is because $b = 2c$: “the birth of one fox requires the death of two rabbits”). A maximum of the rabbit population is followed by a maximum of the fox population. This makes sense, since even when there are enough rabbits to feed all the foxes it should take some time for the foxes to reproduce. But there are also several flaws in the deterministic version of the model. First, the populations can take non-natural numbers and second there is absolutely no randomness in the model. This is in contrast to nature, where only natural numbers for populations are sensible and randomness should be included as well. Both aspects are respected by the stochastic Lotka Volterra model:

- We have a discrete state space $(R, F) \in \mathbb{N}^2$.

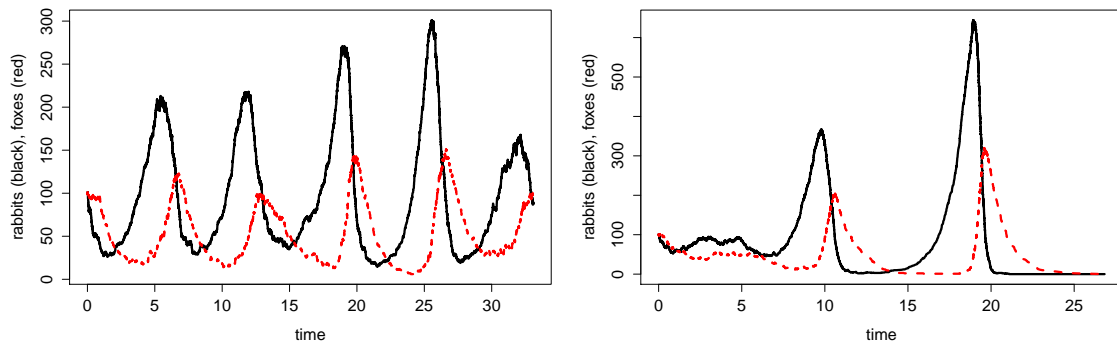
- We allow only nearest neighbour transitions. This means, only one of the actions 'rabbit is born', 'fox is born', 'rabbit dies', 'fox dies' occurs at the same time (this makes sense since on sufficiently short time scales no events should occur simultaneously).



- Similar to the meaning of the single terms in the differential equations of the deterministic model, we define the rates of the stochastic model in the following way:

$$\begin{aligned}
 \text{rabbit reproduction rate: } \lambda_{(R+1,F),(R,F)} &= aR \\
 \text{fox reproduction rate: } \lambda_{(R,F+1),(R,F)} &= cRF \\
 \text{rabbit death rate: } \lambda_{(R-1,F),(R,F)} &= bRF \\
 \text{fox death rate: } \lambda_{(R,F-1),(R,F)} &= gF
 \end{aligned}$$

One can now implement the Gillespie algorithm. As before, start with $R = 100$ rabbits and $F = 100$ foxes and choose the parameters $a = 1$, $b = 0.02$, $c = 0.01$ and $g = 1$. Since the process is stochastic we get a different solution every time we run the process. Consider the following two example solutions:



The first plot looks similar to the deterministic solution. However, the periodicity is not perfect anymore. Different peaks are of slightly different shape and in particular height.

Small fluctuations are visible. The discreteness might be difficult to see due to the choice to plot lines connecting the single points. In the second solution both predator and prey become extinct. In this case, first the fox population dropped very low (near extinction). Without enough predators the prey population could grow very large. As a consequence, the predator population recovers and grows very large as well, while at the same time the rabbit population shrinks dramatically (due to the large rabbit population the fox birth rate grows large, then, due to the large fox population the rabbit death rate hugely exceeds the rabbit birth rate). Even single fluctuations (a single dying fox or a single born rabbit) can barely change this behaviour anymore. After all rabbits died, the foxes have nothing left to eat. Hence, the foxes become extinct as well. For the chosen parameters and initial populations this scenario occurs quite often (even after this rather short time).

A third option is that the foxes die out first. (In the deterministic plot the fox population drops very low at its minima. Fluctuations can then lead to extinction.) The rabbit population would then grow exponentially.

The frequent occurrence of the extinction scenarios can be explained by the small initial populations and by the simplicity of the Lotka Volterra model. Increasing the initial populations (and the parameters b and c as well as the step size in a way that conserves the relative behaviour and the time scale) lets the evolution resemble more and more that of the deterministic model. The susceptibility to fluctuations decreases. Extinction occurs less likely. Apart from that, the model takes only two species into account. In nature there might be more predators that hunt rabbits and other preys that can feed the foxes. Also other environmental influences might play a role.