Elementary Stochastic Processes

These are lecture notes for the course "Elementary Stochastic Processes" (STAT GR 5207) taught at Columbia during Spring 2025, which is a Masters-level introduction to theory and applications of stochastic processes. The course is example-oriented, in the sense students are expected to be able to perform detailed analyses and calculations for concrete models of interest, but they are not expected to give proofs of the main theorems. Additionally, this version of the course gives less emphasis to some classical topics (e.g., branching processes, renewal theory) in exchange for further emphasis on topics of statistical interest (e.g., Hawkes processes, autoregressive processes, Karhunen-Loève representation).

Contents

1 Introduction

1.1 About this class

In previous probability classes, we were typically interested in a single random variable, say X, and various of its properties. For example, we learned how to compute things like its expectation $\mathbb{E}[X]$, variance $\mathrm{Var}(X)$, cumulative distribution function $F(x) := \mathbb{P}(X \leq x)$, and more.

Somtimes we were interested in a sequence of random variables $X_1, X_2, \ldots$ which are independent samples of X, and in this case we studied properties of the sum

$$S_n := \sum_{i=1}^{n} X_i,$$

usually as the number n grows. For example, the law of large numbers (LLN) tells us the sense in which S_n/n converges to a limit, the central limit theorem (CLT) tells us the sense in which $S_n/\sqrt{n}$ converges to a limit (after suitable centering), and more.

These calculations usually had statistical motivations. For example, suppose that $X_1, \ldots, X_n$ were independent samples from a Gaussian distribution with mean θ and variance 1, where θ is an unknown parameter; this is often denoted $\mathcal{N}(\theta, 1)$. Then, we can use the random variable S_n/n to try to make inferences about the unknown θ, since the LLN guarantees that we have $S_n/n \to \theta$ as $n \to \infty$. Additionally, we can exactly calculate the distribution of the estimation error $S_n/n - \theta$, and this allows us to construct confidence intervals for the unknown θ.

In this class, the focus is slightly different. We are also interested in collections of random variables, but they will no longer have the same form of independence; rather, they will be related in some interesting ways. These considerations make the study quite different from previous courses.

Sometimes we will have a sequence $X_1, \ldots, X_n$ of random variables, and we will denote this $\{X_n\}_{n \geq 0}$. Other times we will have an uncountable collection of random variables, which we can think of as a random function $t \mapsto X_t$, and we will denote this $X := \{X_t\}_{t \geq 0}$. These are called the *discrete-time* and *continuous-time* cases, respectively. For this section, we will not try to distinguish between these cases, but we will later see that there are many important differences.

Like in previous classes, we are interested in computing various aspects of $\{X_t\}_t$. For example, the expectation $\mathbb{E}[X_t]$, variance $\mathrm{Var}(X_t)$, cumulative distribution function $F_t(x) := \mathbb{P}(X_t \leq x)$, and more.

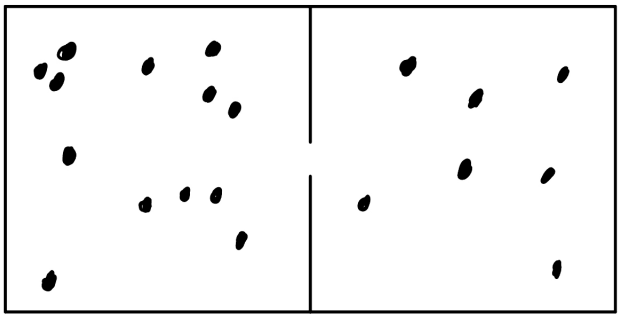
However, these computations can become very complicated, so sometimes we are satisfied by computing what happens in the limit as t grows.

1.2 Motivating examples

Now we give some examples of models of interest that we will study throughout the class.

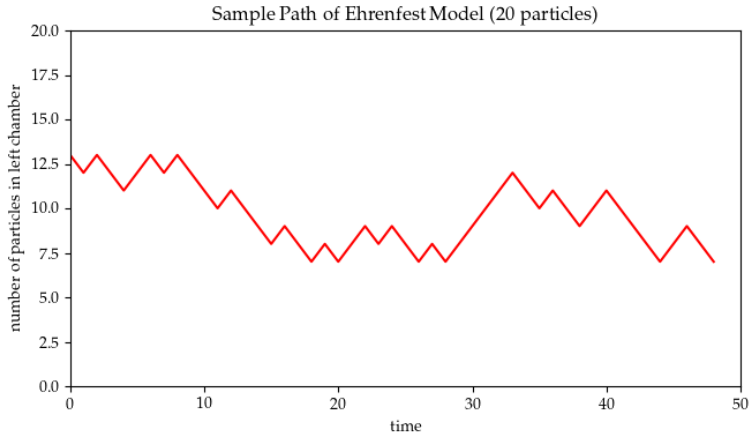
EX (Ehrenfest gas model)**.** Around the year 1900, physicists were looking for mathematical models to explain real-world thermodynamic phenomena. One of the most important models is the following model for movement of particles in a gas.

We suppose that there are M particles of gas that are divided into two chambers, with a small opening between them. For example, for $M = 20$, we should picture something like this:



We write X_t for the number of particles in the left chamber at time t, so $0 \leq X_t \leq M$. To evolve the system from time t to time $t+1$, we select one particle uniformly at random, and we move it from its current chamber to the opposite chamber. So, for example, we have $X_t = 13$ above, and, conditional on $\{X_t = 13\}$, we have $X_{t+1} = 12$ with probability $13/20 = 0.65$ and X_{t+1} with probability $7/20 = 0.35$.

Now suppose $M = 20$ and $X_0 = 13$, so that the system begins in the state shown above. Then, the state X_1 is a random variable which depends on X_0. Also, X_2 is a random variable which depends on X_1. And so on. If we plot the random trajectory $X_0, \ldots, X_{50}$, we might see something like this:

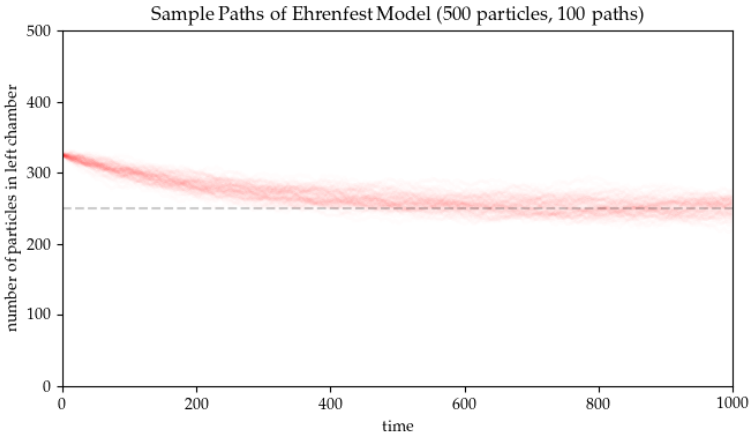


Of course, we will get a different plot if we re-run this system from the same initial state. But the resulting plots will always share some qualitative similarities.

Now, it is natural to ask what happens to this system when M is large and t tends to infinity. Does the system reach some sort of equilibrium? What does this equilibrium look like? Is it the same equilibrium,

regardless of the initial state? Intuition from physics suggests that X_t should fluctuate around $M/2$ when M is large and t goes to infinity. Roughly speaking, this is the claim that the number of particles in either chamber should balance itself after a long time.

To see this in an example, suppose we have $M = 500$ particles and $X_0 = 325$. (This is proportional to the system shown above, but 25 times larger.) Also, let's plot 100 different trajectories instead of just 1 trajectory, so that we can get a sense for the average properties of this system. We get the following:

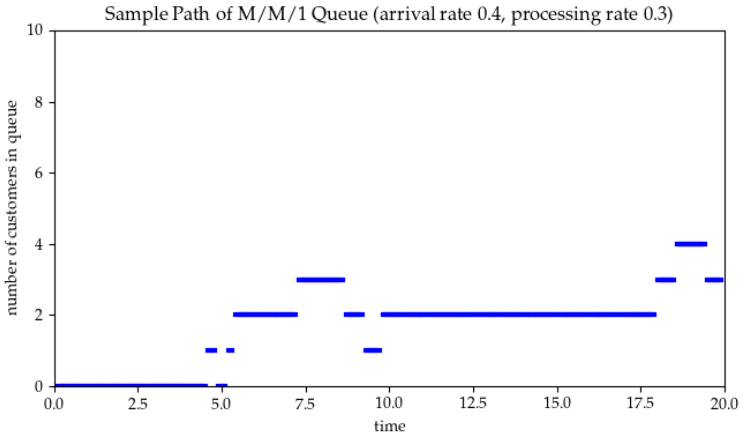


Here, we added a dashed line at $M/2 = 250$ just for the sake of visualization. As expected, it seems like this system reaches an equilibrium which is fluctuating around $M/2$, and it reaches this equilibrium from any possible starting state.

EX (M/M/1 queue)**.** An important field of operations research is *queueing theory*, which studies models of processes which must be completed in sequence. Originally, this study was motivated by problems involving customers arriving at a service facility. But there are many modern applications, as we will see. For instance, we will consider a model of packets arriving at an internet server.

Suppose that packets arrive at the server randomly but at a statistically-regular rate. (We will precisely define what this means later.) Also, each packet requires an exponentially-distributed amount of time to be processed. However, there is only one processor, so packets are simply processed in the order that they arrive. Let's write X_t for the number of packets in the queue at time t; this model is called the *M/M/1 queue* and it depends on two parameters, the arrival rate and the processing rate.

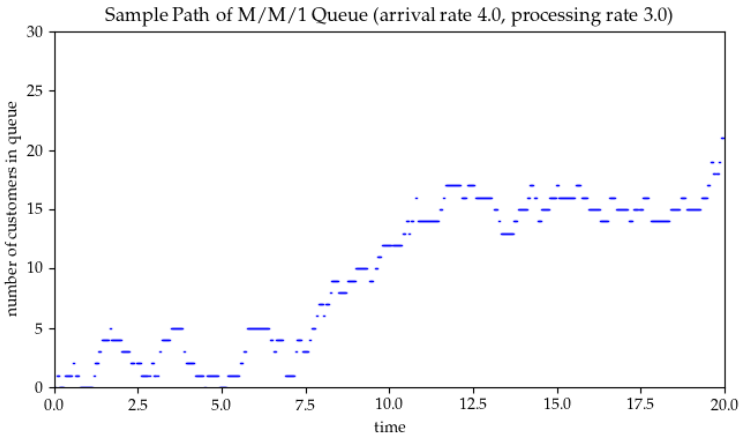
In order to visualize a sample path of this process, suppose that $X_0 = 0$, meaning there are no packets initially. Also, suppose that we have an arrival rate of 0.4 packets per minute, and a processing rate of 0.3 processings per minute. Then we get something like the following:



Of course, if we run this process again, we will get a different random path.

Now consider that you are in charge of designing and managing the server for large website. A quantity of interest is the waiting time for each packet, that is the time elapsed between its arrival and the completion of its processing. How long should is a typical waiting time? If the arrival rate is larger than the processing rate, then it is natural to expect that waiting times may be arbitrarily large.

To see this in a simulation, suppose now that we have an arrival rate of 4 packets per minute, and a processing rate of 3 processings per minute. (This is the same as the system above, except 10 times larger.) Then a sample path of our process may look like the following:

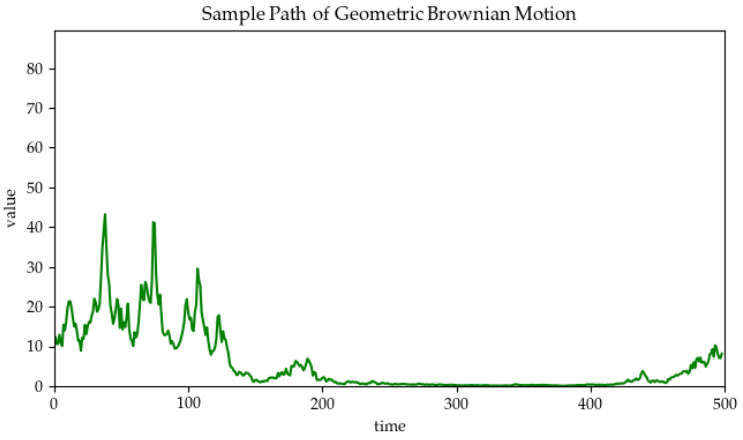


Notice that this process now has very large fluctuations. Also notice that this looks almost like the graph of a continuous function if we zoom out enough.

EX (Geometric Brownain motion)**.** In mathematical finance there are many different models for the evolution of a stock price. Presently, we will explore just one of the simplest models.

Suppose that $\{X_t\}_{t\geq 0}$ is a collection of random variables, where $X_0 > 0$ is some fixed (non-random) initial stock price, and the conditional distribution of $\log(X_t/X_s)$ given $\{X_s = x\}$ follows a Gaussian distribution $\mathcal{N}(\log(x), (t-s)\sigma)$, for every possible $0 \leq s \leq t$. In other words, the process $\{X_t\}_{t\geq 0}$ has "multiplicative increments" with scale $\sigma > 0$. The process $\{X_t\}_{t\geq 0}$ is called a *geometric Brownian motion*.

To see a sample path of a geometric Brownian motion, let's fix $X_0 = 10$ and $\sigma = 0.2$. Then we get the following:



As we can see, the process has large fluctuations; sometimes it has rapidly oscillating peaks and valleys, and other times it has a small and somewhat constant value.

As we will see later, geometric Brownian motion is somewhat "universal". This can be interpreted in various senses. On the one hand, we will show that it is the natural limit of many different models of stock prices. On the other hand, we can show that it is the unique stochastic process satisfying some natural assumptions. Although real-world stock prices do not really follows this sort of pattern, it shows why this is an important process for the theory of mathematical finance.

1.3 Probability background

Lastly, let's do some examples of calculations that illustrate the probability background needed in the course. If you are somewhat comfortable calculations like these (and if you are comfortable with the problems in HW1), then you are sufficiently prepared for the rest of the course.

EX (Binomial with Poisson Number of Trials)**.** Suppose that X follows the Poisson distribution with rate λ, and that the conditional distribution of Y given $\{X = k\}$ follows the Binomial(X, p) distribution. What is the marginal distribution of Y?

To do this, we compute:

$$\begin{aligned}
\mathbb{P}(Y = m) &= \sum_{k=0}^{\infty} \mathbb{P}(Y = m \,|\, X = k)\mathbb{P}(X = k) && \text{(law ot total probability)} \\
&= \sum_{k=m}^{\infty} \mathbb{P}(Y = m \,|\, X = k)\mathbb{P}(X = k) && \text{(terms with } k < m \text{ vanish)} \\
&= \sum_{k=m}^{\infty} \frac{k!}{m!(k-m)!} p^m (1-p)^{k-m} \cdot e^{-\lambda} \frac{\lambda^k}{k!}.
\end{aligned}$$

Now we need to simplify this sum. First, we make the change of variables $n = k - m$ so that the sum ranges over all $n = 0, 1, \ldots,$, and second we can take out all the terms that do not depend on n:

$$\begin{aligned}
\mathbb{P}(Y = m) &= \sum_{k=m}^{\infty} \frac{k!}{m!(k-m)!} p^m (1-p)^{k-m} \cdot e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \frac{(\lambda p)^m}{m!} \sum_{n=0}^{\infty} \frac{1}{n!} (\lambda(1-p))^n.
\end{aligned}$$

Notice that this sum is just the power series for e^z, evaluate at $z = \lambda(1-p)$, so we get

$$\begin{aligned}
\mathbb{P}(Y = m) &= e^{-\lambda} \frac{(\lambda p)^m}{m!} \sum_{n=0}^{\infty} \frac{1}{n!} (\lambda(1-p))^n \\
&= e^{-\lambda} \frac{(\lambda p)^m}{m!} e^{\lambda(1-p)} \\
&= e^{-\lambda p} \frac{(\lambda p)^m}{m!}.
\end{aligned}$$

Therefore, the marginal distribution of Y is Poisson with rate λp.

EX (Independent Poissons Conditional on Sum)**.** Suppose that X follows the Poisson distribution with rate λ, that Y follows the Poisson distribution with rate μ, and that X and Y are independent. What is the conditional distribution of X given $\{X + Y = n\}$?

To do this, we simply compiute:

$$\begin{aligned}
&\mathbb{P}(X = k \,|\, X + Y = n) \\
&= \frac{\mathbb{P}(X = k, X + Y = n)}{\mathbb{P}(X + Y = n)} && \text{(definition of conditioning)}
\end{aligned}$$

$$= \frac{\mathbb{P}(X = k, Y = n - k)}{\mathbb{P}(X + Y = n)} \qquad (\{X = k, X + Y = n\} = \{X = k, Y = n - k\})$$
$$= \frac{\mathbb{P}(X = k)\mathbb{P}(Y = n - k)}{\mathbb{P}(X + Y = n)}. \qquad \text{(independence)}$$

Notice that we can compute both terms in the numerator, but that we do not know what the denominator is. However, we do not really need to know the denominator; if we understand the distribution as a function of k, then the denominator is simply the required normalizing constant! So, we continue the calculation as:

$$\begin{aligned}
\mathbb{P}(X = k \,|\, X + Y = n) &\propto \mathbb{P}(X = k)\mathbb{P}(Y = n - k) \\
&= e^{-\lambda}\frac{\lambda^k}{k!} \cdot e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} \\
&\propto \frac{\lambda^k}{k!} \cdot \frac{\mu^{n-k}}{(n-k)!} \\
&= \frac{1}{k!(n-k)!}\lambda^k\mu^{n-k}.
\end{aligned}$$

We recognize that this looks similar to the probability mass function of a binomial random variable. Indeed, if we multiply by $n!(\lambda + \mu)^{-n}$, then we have shown:

$$\mathbb{P}(X = k \,|\, X + Y = n) \propto \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k}.$$

In conclusion, the conditional distribution of X given $\{X + Y = n\}$ is Binomial with n trials and success probability $\lambda/(\lambda + \mu)$.

EX (Gaussian with Gaussian Centering)**.** Suppose that X follows the Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and that the conditional distribution of Y given $\{X = x\}$ follows the Gaussian distribution $\mathcal{N}(x, \tau^2)$. What is the condtional distribution of X given $\{Y = y\}$?

By Bayes rule, we have:

$$f_{X\,|\,Y}(x \,|\, y) = f_{Y\,|\,X}(y \,|\, x)\frac{f_X(x)}{f_Y(y)},$$

and we can compite some of these terms. As before, recall that we don't need to compute this density exactly; we only need to identify it as a function of x, and then the normalizing constant will be automatically determined. So we can compute the following, by completing the square:

$$\begin{aligned}
f_{X\,|\,Y}(x \,|\, y) &\propto f_{Y\,|\,X}(y \,|\, x)f_X(x) \\
&\propto \exp\left(-\frac{1}{2\tau^2}(y - x)^2\right)\exp\left(-\frac{1}{2\sigma^2}x^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)x^2 + \frac{xy}{\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)\left(x - \frac{\sigma^2}{\sigma^2 + \tau^2}y\right)\right).
\end{aligned}$$

Notice that this is just a Gaussian density, and we can see its mean and variance. Therefore, the conditional distribution of X given $\{Y = y\}$ is

$$\mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + \tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right).$$

2 Discrete-time Markov chains

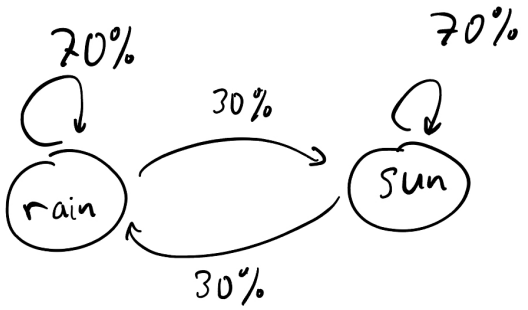
2.1 Basic definitions and examples

DEF. A *Markov chain (MC)* is a stochastic process $\{X_n\}_{n\geq 0}$ such that the conditional distribution of X_{n+1} given $\{X_0 = i_0, \ldots, X_n = i_n\}$ equals the conditional distribution of X_{n+1} given $\{X_n = i_n\}$, for all states $i_0, \ldots, i_n$

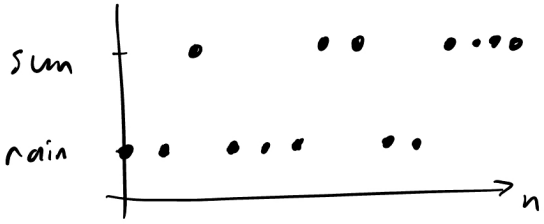
In other words: "The distribution of the future depends only on the present, not on the past". That is, if we know $\{X_n = i\}$, then we do not need to know how we arrived at state i in order to predict the value of X_{n+1}. This is the opposite of what some other fields call *path-dependence*.

Typically, a MC may be represented in many different ways. It is useful to be able to "translate between" these representations since sometimes a problem will be hard in one representation but easy in another representation. To illustrate this, we give a simpe example about the weather, represented 4 different ways:

- In plain language: Tomorrow there is 50% of rain and 50% of sun. Every subsequent day, there is 70% chance of keeping the same weather, and 30% chance of changing weather.
- Directed weighted graph:



- Plot of sample path:



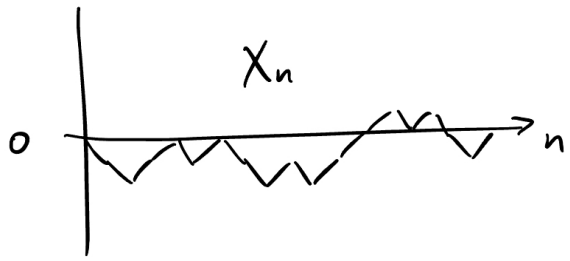
- Linear algebra:

$$\pi = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \qquad \text{and} \qquad P = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$$

These all represent the same process, but there are some advantages and disadvantages to each.

We give some further examples of MCs:

EX (simple symmetric random walk)**.** This is a MC with state space $\mathbb{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ defined by the property that, at each time step, we move up or down with equal probability. Its sample paths look like



and its transition matrix is given by

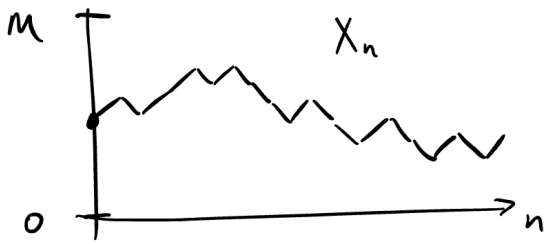
$$P_{i,j} = \begin{cases} 1/2 & \text{if } j = i-1 \\ 1/2 & \text{if } j = i+1 \\ 0 & \text{else.} \end{cases}$$

for all $i \in \mathbb{Z}$.

EX (Ehrenfest gas model)**.** For fixed positive integer M, this a MC on the state space $\{0, 1, \ldots, M\}$ which has the interpretation of counting the number of particles in the left of two chambers, where, at each time step, we select one particle uniformly at random and move it to the other chamber. So, its transition matrix is

$$P_{i,j} = \begin{cases} i/M & \text{if } j = i-1 \\ 1 - i/M & \text{if } j = i+1 \\ 0 & \text{else.} \end{cases}$$

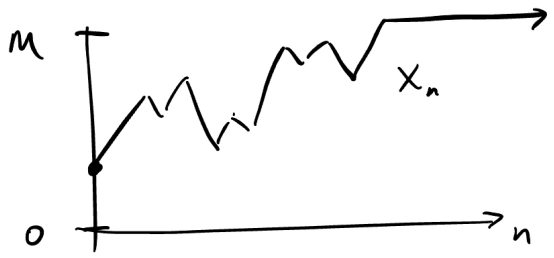
for all state i, and its sample paths look like



EX (Wright-Fisher model)**.** For fixed positive integer M, this is a MC on the state space $\{0, 1, \ldots, M\}$ which represents the number of individuals in a population posessing a gene of type $\times$ out of two possible genes $\{\times, \circ\}$. The dynamics are defined via

$$(X_{n+1} \,|\, X_n = i) \sim \text{Binom}(M, i/M)$$

which has the following interpretation: Every individual born at time $n+1$ inherits a gene from an individual born at time n, uniformly among all possible parents. Notice in this process that, if $X_n = 0$ for some n, then $X_k = 0$ for all $k \geq n$; similarly for M replacing 0. So, sample paths of this process look something like the following:



Additionally, there are some "pathological" examples of MCs that are important to consider. That is, these examples may seem trivial or irrelevat, but they are indeed MCs and it's useful to keep them in mind.

EX (independent, identically-distibuted sequence)**.** Suppose that $X_0, X_1, \ldots$ are independent, identically-distibuted (IID) samples from some distribution. Then we of course have

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

since both side are just equal to $\mathbb{P}(X_{n+1} = i_{n+1})$ by independence.

EX (deterministic sequence)**.** Suppose that f is a deterministic function such that $X_{n+1} = f(X_n)$ for all n. Then $X_0, X_1, \ldots$ is a MC with transition matrix given by

$$P_{ij} = \begin{cases} 1 & \text{if } j = f(i) \\ 0 & \text{else.} \end{cases}$$

This illustrates something important in general: If the ith row of P has a single entry equal to 1 and all other entries equal to 0, then X_{n+1} is a non-random function of X_n when $\{X_n = i\}$.

2.2 Transition matrices

Although it is useful to keep in mind many possible representations for a given MC, it is usually most convenient to work with the linear algebra perspective. As we will see, this representation is usually the most effective for calculations.

DEF. A *transition matrix* is a (possibly infinite) matrix satisfying

- $P_{ij} \geq 0$ for all states i, j, and
- $\sum_j P_{ij} = 1$ for all states i.

The second condition requires that the sum of the entries of P along each row must equal 1. It is not true in general that the sum of the entries of P along each column must equal 1; you will think about this a bit on HW3.

If $\{X_n\}_{n\geq 0}$ is a MC, its transition matrix P is defined via

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

for all states i, j. That is, P_{ij} represents the probability of going to state j conditional on being in state i. Strictly speaking, it is possible for MC $\{X_n\}_{n\geq 0}$ may have a different transition matrix at each time n; in this class we will not consider that possibility, and instead we will focus on MCs with a single transition matrix over all time.

One of the reasons the linear algbera is useful, is the following:

LEM. *For all $n \geq 0$ and all states i, j,*

$$(P^n)_{ij} = \mathbb{P}(X_n = j \,|\, X_0 = i)$$

PF: For $n = 1$, the claim is equivalent to the definition of the transition matrix. For $n = 2$, we calculate as follows:

$$\begin{aligned}
&\mathbb{P}(X_2 = j \,|\, X_0 = i) \\
&= \sum_k \mathbb{P}(X_2 = j, X_1 = k \,|\, X_0 = i) && \text{(law of total probability)} \\
&= \sum_k \mathbb{P}(X_2 = j \,|\, X_1 = k, X_0 = i)\mathbb{P}(X_1 = k \,|\, X_0 = i) && \text{(Bayes' rule)} \\
&= \sum_k \mathbb{P}(X_2 = j \,|\, X_1 = k)\mathbb{P}(X_1 = k \,|\, X_0 = i) && \text{(Markov property)} \\
&= \sum_k P_{kj} P_{ik} && \text{(definition of } P\text{)} \\
&= (P^2)_{ij} && \text{(definition of matrix-matrix multiplication)}
\end{aligned}$$

For $n = 3$, we execute the same steps except that we insert the $n = 2$ result in the end:

$$\begin{aligned}
&\mathbb{P}(X_3 = j \,|\, X_0 = i) \\
&= \sum_k \mathbb{P}(X_3 = j \,|\, X_2 = k)\mathbb{P}(X_2 = k \,|\, X_0 = i) && \text{(same steps as above)} \\
&= \sum_k P_{kj}(P^2)_{ik} && \text{(definition of } P \text{ and } n = 2 \text{ case)} \\
&= (P^3)_{ij} && \text{(definition of matrix-matrix multiplication)}
\end{aligned}$$

Continuing in this way (or, doing proof by induction) shows the desired result for all n.

We used the notation $(P^n)_{ij}$ to emphasize that this is the i, j entry of the matrix power P^n, since P^n_{ij} might be mistaken as the nth power of the i, j entry of P. In the rest of the course, we will just use the notation P^n_{ij}.

Another useful calculation is the following. Let α denote a row vector which is a probability vector; this means it has non-negative entries, and its entries sum to 1.

LEM. *If X_0 is distributed according to a probability vector α, then for any $n \geq 0$, the random variable X_n is distributed according to the probability vector αP^n.*

PF: For $n = 1$, we compute:

$$\begin{aligned}
\mathbb{P}(X_1 = i) &= \sum_j \mathbb{P}(X_1 = i \,|\, X_0 = j)\mathbb{P}(X_0 = j) && \text{(law of total probability)} \\
&= \sum_j P_{ji}\alpha_j && \text{(definitions of } P, \alpha\text{)} \\
&= (\alpha P)_i
\end{aligned}$$

This shows that the distribution of X_1 is αP, and the general case follows from the previous result.

When you first learn linear algebra, the definition of matrix-matrix multiplicaiton and vector-matrix multiplication may seem a bit weird. But the MC perspective is a nice way to justify it: The sum appearing in the definition is just an example of the law of total probability!

2.3 Limiting and stationary distributions

One of the most important questions we will try to answer this semester is: What happens the distribution of a MC in the long-run? Here, we make this notion precise, although it will take a while before we have a comprehensive answer. Throughout this subsection, $\{X_n\}_{n\geq 0}$ is a MC with transition matrix P.

DEF. A probability distribution π is called a *limiting distribution* for $\{X_n\}_{n\geq 0}$ (or for P) if we have

$$\lim_{n\to\infty} \mathbb{P}(X_n = j \,|\, X_0 = i) = \pi_j \text{ for all states } i, j.$$

Because of the results of the previous subsection, this condition is equivalent to all of the following

- $\lim_{n\to\infty} P_{ij}^n = \pi_j$ for all states i, j
- $\lim_{n\to\infty} \alpha P^n = \pi$ for all probability vectors α, and
- $\lim_{n\to\infty} P^n = \begin{pmatrix} \text{———} & \pi & \text{———} \\ & \vdots & \\ \text{———} & \pi & \text{———} \end{pmatrix}$.

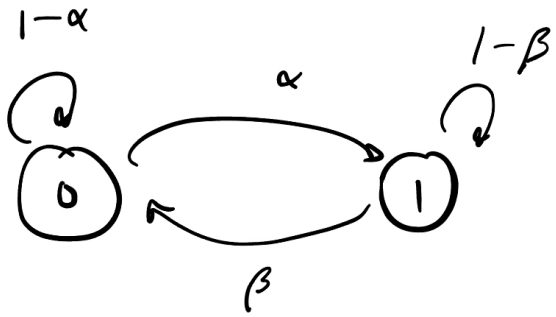
If a limiting distribution exists, then it is unique.

As we will see next, a limiting distribution may or may not exist. One of our main goals is to understand when this is the case, and how to find it if so.

EX (two-state chain)**.** For $\alpha, \beta \in [0, 1]$, consider

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

Or, in graph form:



If you go through and do the calculations directly, it is possible to show

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

for all n. So, if $\alpha + \beta < 1$, then taking $n \to \infty$ gives

$$\lim_{n\to\infty} P^n = \frac{1}{\alpha+\beta} \begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix}.$$

In other words, the limiting distribution is

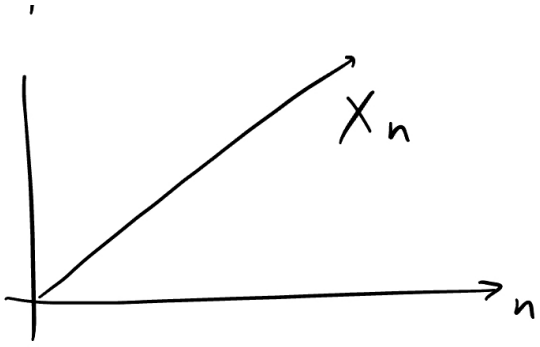
$$\pi = \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}$$

However, note that this calculation is already quite complicated!

EX (increasing chain)**.** Suppose that P is the transition matrix defined via

$$P_{ij} = \begin{cases} 1 & \text{if } j = i+1 \\ 0 & \text{else.} \end{cases}$$

In other words, this is the transition matrix for the Markov chain $\{X_n\}_{n\geq 0}$ such that $X_{n+1} = X_n + 1$ (equivalently, $X_n = X_0 + n$) for all $n \geq 0$:



We can see that P does not have any limiting distribution:

$$\lim_{n\to\infty} \mathbb{P}(X_n = j \,|\, X_0 = i) = 0$$

This is because, as long as i, j are fixed, it is possible to take n large enough that the probability is zero. This shows that the only possible limiting distribution is π satisfying $\pi_i = 0$ for all i, but this is not a probability vector.

As we can see, limiting distributions can be quite complicated: they may not exist, they may be hard to find even if they do exist, etc. So, we will need to develop some tools for studying them. A key ingredient is another specialize type of distribution:

DEF. A probability distribution π is called a *stationary distribution* for $\{X_n\}_{n\geq 0}$ (or for P) if we have $\pi = \pi P$. By the results of the previous subsection, this is equivalent to the statement that $X_0 \sim \pi$ implies $X_1 \sim \pi$.

As we will see later, finding a stationary distribution is not usually too difficult since $\pi = \pi P$ is just a linear system of equations! However, stationary distributions need not be unique. (You will see an example in HW2.)

EX (two-state chain)**.** As above, consider the transition matrix

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

and let us assume $\alpha + \beta > 0$. and let us try to solve the system of equations $\pi = \pi P$. Writing $\pi = (\pi_0, \pi_1)$, this system of equations is just

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

However, we must remember that we also have the constraint $\pi_0 + \pi_1 = 1$. So, the first equation implies $\alpha\pi_0 = \beta\pi_1$, so combining this with $\pi_0 + \pi_1 = 1$ yields $\pi_0 = \beta/(\alpha+\beta)$. Then we also get $\pi_1 = \beta/(\alpha+\beta)$. Thus, the unique stationary distribution is

$$\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right).$$

Notice above that the limiting distribution and the stationary distribution for the two-state chain are the same. As we see, these notions are related:

LEM. *If π is the limiting distribution for P, then π is a stationary distribution for P.*

PF: Suppose π is the limiting distribution for P. Then for any probability vector α we have

$$\pi = \lim_{n\to\infty} \alpha P^n = \lim_{n\to\infty} \alpha P^{n+1} = \left(\lim_{n\to\infty} \alpha P^n\right) P = \pi P.$$

This shows that π is a stationary distribution for P.

Unfortunately, it is not always true that a stationary distribution must be a limiting distribution. This would have been nice, since it would allow us to find a limiting distribution via solving a linear system of equations. But we have a counter-example:

EX (alternating chain)**.** Consider the transition matrix

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

which represents a MC that alternates (deterministically) between two states. We can directly calculate

$$P^2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and more generally

$$P^n := \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{if } n \text{ is even} \\ \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \text{if } n \text{ is odd.} \end{cases}$$

Thus, $\lim_{n\to\infty} P^n$ does not exist, hence P does not have any limiting distribuion. But P does have a stationary distribution—We showed in the previous example that

$$\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$$

is stationary for P. Thus, stationary distributions need not be limiting distributions.

Interestingly, there are some conditions under which a stationary distribution is guaranteed to be a limiting distribution. In such cases, we can find a limiting distribution easily! Our next goal is to understand when this is the case.

2.4 Regularity

DEF. A transition matrix P is called *regular* if there exists some $n \geq 0$ such that all entries of P^n are strictly positive. A MC $\{X_n\}_{n\geq 0}$ is called regular when its transition matrix is regular.

EX (two-state chain)**.** In the example of the two-state chain above, P is regular if $0 < \alpha < 1$ and $0 < \beta < 1$. As we already saw, it is not regular if $\alpha = \beta = 0$.

EX. For $0 < p < 1$, considr the transition matrix:

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

We can directly compute:

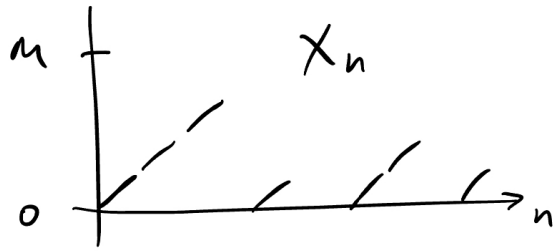
$$P^2 = \begin{pmatrix} 2p(1-p) & p^2 & (1-p)^2 \\ (1-p)^2 & 2p(1-p) & p^2 \\ p^2 & (1-p)^2 & 2p(1-p) \end{pmatrix},$$

hence P is regular.

EX (finite success runs)**.** For some $0 < p < 1$, consider the MC with transition matrix

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

on the state space $\{0, 1, \ldots, 4\}$; more generally, we can think of a similar transition matric on the state space $\{0, 1, \ldots, M\}$ for some $M \geq 0$. This has a concrete interpretation: At each step, the MC either increments by 1, or it gets reset back to state 0. However, in the last state M, we always go back to state 0. This is usually called a *success runs* chain since it counts the number of consecutive successes before a failure, in some independent trials. You can visualize the sample paths like this:



To see that this is regular, we introduce a small trick: Notice thay we only care about whether entries are 0 or positive, but that we do not care about their precise value. So, it makes sense to just replace all of the positive entries with $+$, and compute the matrix powers with respect to the usual arithmetic rules. We can see

$$P = \begin{pmatrix} + & + & 0 & 0 & 0 \\ + & 0 & + & 0 & 0 \\ + & 0 & 0 & + & 0 \\ + & 0 & 0 & 0 & + \\ + & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad P^2 = \begin{pmatrix} + & + & + & 0 & 0 \\ + & + & 0 & + & 0 \\ + & + & 0 & 0 & + \\ + & + & 0 & 0 & 0 \\ + & + & 0 & 0 & 0 \end{pmatrix},$$

$$P^3 = \begin{pmatrix} + & + & + & + & 0 \\ + & + & + & 0 & + \\ + & + & + & 0 & 0 \\ + & + & + & 0 & 0 \\ + & + & + & 0 & 0 \end{pmatrix}, \qquad \cdots$$

and so on. Thus, P^5 has all positive entries, thus P is regular. (P is regular for general $M > 0$ as well, since we can see that P^M has all positive entries.)

Regular transition matrices are important because they satisfy the following:

THM (fundamental theorem of regular MCs)**.** *If $\{X_n\}_{n\geq 0}$ is a regular MC, then it has a unique stationary distribution π, and π is also a limiting distribution.*

EX (two-state chain)**.** As we showed directly, the two-state chain has a unique stationary distribution, and this stationary distribution is also a limiting distribution. Now we now that this is actually a consequence of the fundamental theorem, since we showed that the transition matrix is regular.

EX (finite success runs)**.** Since the success runs chain is regular, we can find its limiting distribution simply by solving the stationarity equation $\pi = \pi P$. For general M, this is just the linear system:

$$\begin{cases} \pi_0 = (1-p)(\pi_0 + \cdots + \pi_{M-1}) + \pi_M \\ \pi_i = p\pi_{i-1} \text{ for all } 1 \leq i \leq M \end{cases}$$

Notice that the second equation implies $\pi_i = p^i \pi_0$ for all $1 \leq i \leq M$. So, the condition $\pi_0 + \cdots + \pi_M = 1$ implies

$$\pi_0 + p\pi_0 + p^2\pi_0 + \cdots + p^M \pi_0 = 1,$$

which means

$$\pi_0 = \frac{1}{1 + p + p^2 + \cdots + p^M} = \frac{1-p}{1-p^{M+1}}.$$

Therefore,

$$\pi_i = \frac{1-p}{1-p^{M+1}} p^i$$

for all $0 \leq i \leq M$. (Notice that this looks slightly like a geometric distribution, but not quite. In fact it is a geometric distribution conditioned on being less than or equal to M. That is why the density is proportional to that of a geometric, but the normalizing constant is different.)

Our next goal is to study the Ehrenfest gas model, but—as we will see later in the course—it does not have a limiting distribution. So, we will instead study a variant of the model which is more well-behaved.

DEF. The *lazy version of* a transition matrix P is $\tilde{P} := \frac{1}{2}(I + P)$.

As we saw in HW2, this has two interpretations (1) each step of $\tilde{P}$ is either a step of P or a "hold" in the current state, or (2) $\tilde{P}$ holds in the current state for a geometric amount of time, and then takes a step from P. Lazy chains are important because of the following:

LEM. *If P is a transition matrix and $\tilde{P}$ is the lazy version of P, then P and $\tilde{P}$ have the same stationary distributions.*

EX (lazy Ehrenfest gas model)**.** Fix $M > 0$, and let $\tilde{P}$ denote the lazy version of the Ehrenfest gas model on M particles, which we saw in the previous lectures.

First let us show that $\tilde{P}$ is regular. Indeed, we claim that P^M has all positive entries. To see this, note that for $0 \leq i \leq j \leq M$ we have

$$P^M_{ij} \geq 2^{M-(j-i)} \cdot \frac{1-i}{M} \cdot \frac{1-(i+1)}{M} \cdots \frac{1-(j-1)}{M} > 0$$

and similar for $0 \leq j \leq i$. In other words, it is always possible to travel between any two states in M steps.

By the fundamental theorem, $\tilde{P}$ has a unique stationary distribution. But $\tilde{P}$ and P have the same stationary distributions, so we can instead look for the stationary distribution of P. What should this stationary distribution be? We can of course set up a linear system of equations to solve for π, but we can just as well do guess-and-check: by physical intuition, we might expect that the stationary distribution is

$\pi = \text{Binom}(M, \frac{1}{2})$. All we need to do is verify this, since we know from the fundamental theorem that the stationary distribution is unique. So we check:

$$\begin{aligned}
(\pi P)_i &= \sum_{j=0}^{M} \pi_j P_{ji} && \text{(definition of vector-matrix product)} \\
&= \pi_{i-1} P_{i-1,i} + \pi_{i+1} P_{i+1,i} && \text{(support of } P\text{)} \\
&= \binom{M}{i-1} 2^{-M} \left(1 - \frac{i-1}{M}\right) + \binom{M}{i+1} 2^{-M} \cdot \frac{i+1}{M} && \text{(definition of } P\text{)} \\
&= \binom{M}{i} 2^{-M}. && \text{(algebra)}
\end{aligned}$$

Therefore, the limiting distribution of $\tilde{P}$ is just $\text{Binom}(M, \frac{1}{2})$.

At this point, we have develop a rather nice theory for how to compute limiting distributions for some MCs of interest. However, there are two drawbacks of what we have learned so far:

- Regularity can be hard to verify, since we have to compute matrix powers. So, it would be nice to have some sufficient conditions for regularity, which are easier to check.
- Many MCs on infinite state spaces are not regular. So, it would be nice to have a more general fundamental theorem which also applies to some MCs of interest on infinite state spaces.

The next part of the course will address these problems.

2.5 State space decomposition

As above, let $\{X_n\}_{n \geq 0}$ be a MC.

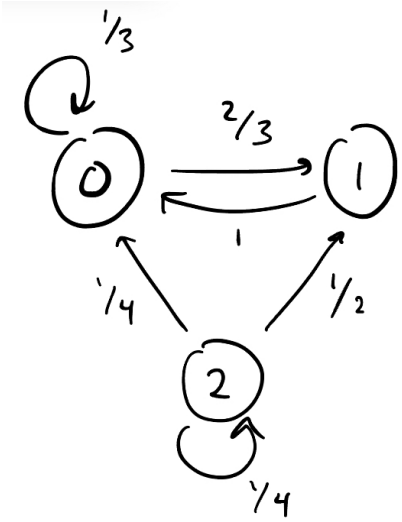
DEF. We say that state i *reaches* state j if there exists some $n \geq 0$ such that $P_{ij}^n > 0$. We say that states i, j *communicate* if i reaches j and j reaches i.

It can be shown that the state space S of a MC $\{X_n\}_{n \geq 0}$ can be partitioned into a maximal collection $S_1, S_2, \ldots$ where all states with S_k communicate with each other, for each k. This is called the *state space decomposition*. (This is possible because "communication" is an *equivalence relation*, but we do not go into the details here.)

EX. Consider the transition matrix

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

on the state space $\{0, 1, 2\}$. We can visualize this as follows:



To find its communicating classes, note that no state reaches state 2, so 2 must be in its own communicating class. Also, 0 and 1 reach each other, so they are in the same communicating class. So, the communicating classes are $\{0, 1\}$ and $\{2\}$.

EX (simple symmetric random walk)**.** Consider the transition matrix P of a simple symmetric random walk. It seems intuitive that every state can reach every other state, hence that there is one communicating class. To show this, consider arbitrary state i, j with $i < j$; it is always possible to from state i to state j in $j - i$ steps, by taking an upward step at each time, and this small but positive probability:

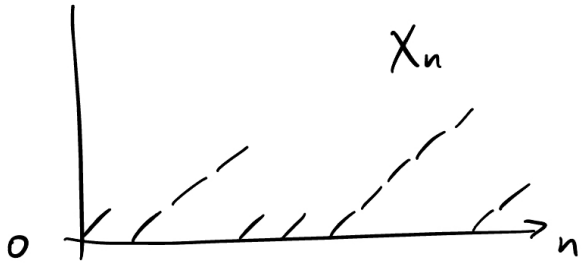
$$P_{ij}^{j-i} \geq 2^{j-i} > 0.$$

Thus, i reaches j, and the same argument (but with downward steps) shows that j reaches i. So P has one communicating class, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

EX (infinite success runs)**.** For $0 < p < 1$, consider the transition matrix

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

on the state space $\{0, 1, 2, \ldots\}$. Note that this is just the infinite version of the finite success runs MC that we studied last week, so its sample paths are like the following:



In this MC, we can show that 0 communicates with every state, and this will imply that there is only one communicating class. To see that 0 reaches state $i \geq 1$, note that we can always go from 0 to i in i steps by increasing at each step, i.e.

$$P_{0i}^i \geq p^i > 0.$$

So see that state i reaches state 0, simply note that we can always go from i to 0 in one step, since $P_{i0} = 1 - p > 0$. Thus, there is only one communicating class $\{0, 1, 2, \ldots\}$.

DEF. A Markov chain is *irreducible* if it has exactly one communicating class.

2.6 Recurrence and Transience

Throughout this subsection, we write $\{X_n\}_{n\geq 0}$ for a MC and P for its transition matrix. For a state i, we write $T_i := \min\{n \geq 1 : X_n = i\}$ for the *first return time* to i.

DEF. A state i is called *recurrent* if we have

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

and it is called *transient* if we have

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

In other words, recurrent states are ones which we certainly return to after long enough time, and transient states are ones that we might never return to.

Note that the definitions of recurrence and transience only directly say whether we will return to state i eventually. But, by the Markov property, the process "resets" each time we return to state i. So we actually learn something more: If i is recurrent, then it will return to state i infinitely many times. If i is transient, then it will return to state i finitely many times, and the total numbre of visits to i is a geometric random variable.

We also have another useful characterization of recurrence and transience:

LEM. *A state i is recurrent if and only if*

$$\sum_{n=0}^{\infty} P_{ii}^n = \infty$$

and it is transient if and only if

$$\sum_{n=0}^{\infty} P_{ii}^n < \infty.$$

We also have the following, which is useful in applications; if we want to show that all states are recurrent, it suffices to show that the MC is irreducible and that there exists at least one recurrent state.

LEM. *Within a communicating class, either all states are recurrent or all states are transient.*

EX (success runs)**.** Consider the (infinite) success runs example from before. It is easy to show that state 0 is recurrent: If we start at 0 but we have not returned to 0 in n steps, this means we must have made n upward steps, so

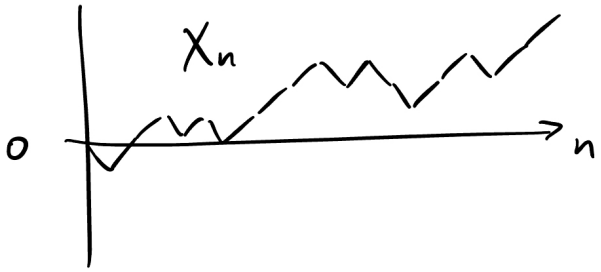
$$\mathbb{P}(T_0 > n \,|\, X_0 = 0) = p^n.$$

But taking $n \to \infty$ shows $\mathbb{P}(T_0 = \infty \,|\, X_0 = 0) = 0$, so state 0 is recurrent. Since we showed before that this MC is irreducible, this implies that all states are recurrent.

EX (biased simple random walk)**.** For $0 < p < 1$ consider the transition matrix

$$P_{i,j} = \begin{cases} p & \text{if } j = i-1 \\ 1-p & \text{if } j = i+1 \\ 0 & \text{else} \end{cases}$$

for all $i \in \mathbb{Z}$. Note that this is just the simple symmetric random walk if $p = 1/2$, but it allows for some biased motion if $p \neq 1/2$. For example, if $p < 1/2$, then it has an upward bias, so its sample paths look something like:



How should we check recurrence/transience? Recall that this MC is irreducible, so it suffices to check whether 0 is recurrent or transient. To do this, we use the alternative characterization; note that $P_{00}^n = 0$ if n is odd, and if n is even then we can use Stirling's approximation $n! \sim \sqrt{2\pi n}(n/e)^n$ to get:

$$\begin{aligned} P_{00}^{2n} = \mathbb{P}(X_{2n} = 0 \,|\, X_0 = 0) &= \binom{2n}{n} p^n (1-p)^n && \text{(Binomial distribution)} \\ &\sim \frac{4^n}{\sqrt{\pi n}} p^n (1-p)^n && \text{(Stirling's formula)} \\ &\sim \frac{(4p(1-p))^n}{\sqrt{\pi n}} \end{aligned}$$

If $4p(1-p) \geq 1$ then this decays like $n^{-1/2}$ hence its sum diverges. If $4p(1-p) < 1$ then this decays exponentially hence its sum converges. Now note that $4p(1-p) = 1$ if and only if $p = 1/2$. In other words, this MC is recurrent if $p = 1/2$ and it is transient if $p \neq 1/2$. This makse a bit of sense, since if $p < 1/2$, then it looks like $\{X_n\}_{n \geq 0}$ runs away to ∞.

In this setting we have one of our important theorems.

THM (fundamental theorem of irreducible MCs)**.** *If $\{X_n\}_{n \geq 0}$ is an irreducible MC, then it has a unique stationary distribution π, and π satisfies*

$$\pi_i = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mathbb{E}[T_j \,|\, X_0 = j]}$$

for all states i, j.

An important remark about the theorem is that we allow $\mathbb{E}[T_i \,|\, X_0 = j] = \infty$, in which case the right side should be interpreted as $1/\infty = 0$. This of course occurs if $\{X_n\}_{n \geq 0}$ is transient, since then we have

$$\mathbb{E}[T_j \,|\, X_0 = j] \geq \infty \cdot \mathbb{P}(T_j = \infty \,|\, X_0 = j) = \infty.$$

Interestingly, it may also occur if $\{X_n\}_{n\geq 0}$ is recurrent, since recurrence only guarantees that T_j is finite but not that its expectation $\mathbb{E}[T_j \mid X_0 = j]$ is finite. We will discuss more about this later.

A second remark about the theorem is that the conclusion is different than, but somewhat related to, the conclusion of the fundamental theorem of regular Markov chains. Indeed, for regular $\{X_n\}_{n\geq 0}$ we concluded that

$$\lim_{n\to\infty} P_{ij}^n$$

exists, while for irreducible $\{X_n\}_{n\geq 0}$ we concluded that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k$$

exists. It turns out that the second conclusion is weaker than the first; this follows from the general real analysis fact that, for any sequence $\{a_n\}_{n\geq 0}$ of real numbers, $\lim_{n\to\infty} a_n = a$ implies $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k = a$.

A last remark is to give some interpretation to the quantity

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k.$$

Indeed, let's compute:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(X_k = j \mid X_0 = i) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\mathbf{1}\{X_k = j\} \mid X_0 = i\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{X_k = j\} \,\middle|\, X_0 = i\right].
\end{aligned}$$

Thus, the limit is exactly the long-term expected proportion of time spent in state j. In some sense, it makes sense that this should be the reciprocal of the value $\mathbb{E}[T_j \mid X_0 = j]$, which is time in between successive visits to state j.

2.7 Periodicity

Some of the MCs we have studied have the weird behavior that there are some states that may only be visited at even times and some states that may only be visited at odd times. This causes problems for the existence of limiting distributions (like we saw in the example of the Ehrenfest model).

In order to exclude such behaviors, we need to introduce another definition. Throughout, let $\{X_n\}_{n\geq 0}$ denote a MC with transition matrix P.

DEF. The *period* of a state i is

$$d(i) := \gcd\{n \geq 1 : P_{ii}^n > 0\}$$

where gcd is the *greatest common divisor* of a set of integers. A state i is called *aperiodic* if $d(i) = 1$.

Whereas irreducibility means the MC cannot be "broken up in space", aperiodicity means the MC cannot be "broken up in time".

EX (simple symmetric random walk)**.** Consider the MC $\{X_n\}_{n\geq 0}$ with transition matrix given by

$$P_{i,j} = \begin{cases} 1/2 & \text{if } j = i-1 \\ 1/2 & \text{if } j = i+1 \\ 0 & \text{else} \end{cases}$$

for all $i \in \mathbb{Z}$; recall this this plainly means that the MC moves up or down at each time step, with equal probability. Note for every state i it is only possible to return to state i after an even number of steps; that is, $P_{ii}^n = 0$ if n is odd. In fact, we also have $P_{ii}^n \geq \binom{n}{n/2} 2^{-n} > 0$ if n is even, since there is positive probability of achieveing exactly $n/2$ up steps and $n/2$ down steps. Thus,

$$d(i) = \gcd\{n \geq 1 : P_{ii}^n = 0\} = \gcd\{2, 4, 6, \ldots\} = 2$$

for all $i \in \mathbb{Z}$. In other words, every state has period 2.

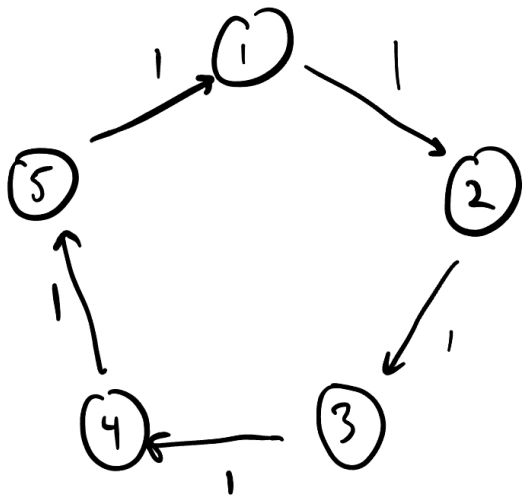
EX (cycle)**.** For a positive integer M, let $\{X_n\}_{n \geq 0}$ be the MC on $\{1, 2, \ldots, M\}$ defined as follows: If $i \neq M$, then

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

and, also

$$\mathbb{P}(X_{n+1} = 1 \,|\, X_n = M) = 1.$$

In other words, this MC is non-random; for $M = 5$ it can be represented by the graph



Note that $P_{ii}^n = 1$ if n is a multiple of M and that $P_{ii}^n = 0$ otherwise. Thus,

$$d(i) = \gcd\{M, 2M, 3M, \ldots\} = M$$

for all states i. Thus, all states have period M.

EX (Ehrenfest models)**.** Let P denote the transition matrix for the Ehrenfest gas model, and let $\tilde{P} = \frac{1}{2}(I+P)$ denote the transition matrix for the lazy version of the Ehrenfest gas model. Along the same lines as the simple symmetric random walk, one can show $P_{ii}^n > 0$ if and only if n is even, hence all states have period 2 in the standard Ehrenfest model. However, one can see $\tilde{P}_{ii}^n > 0$ for all $n \geq 1$, hence all states have period 1 in the lazy Ehrenfest model.

This last example illustrates something elementary but important: If a state i has $P_{ii} > 0$, then we must have $d(i) = 1$. This partially explains why lazy versions are convenient: They are always aperiodic!

Another important result is:

LEM. *Within a communicting class, all states have the same period.*

Thus, we usually do not need to compute the period of every state. Instead, we use a similar method that we used in the study of recurrence/transience: We should find one state for which the period is easy to compute, and it follows that all other states in this communicating class must have the same period.

The most important thing about periodicity is the following, which gives us an easy way to apply the fundamental theorem of regular MCs:

THM. *A MC on a finite state space is regular if and only if it is irreducible and aperiodic.*

2.8 Ergodicity

Our next goal is to understand which MCs on infinite state spaces satisfy the same conclusion as the fundamental theorem of regular MCs. Let $\{X_n\}_{n\geq 0}$ be a MC with transition matrix P, and let us assume that its state space is infinite.

This requires some more notions. Recall that a state j is called *recurrent* if the conditional distribution of T_j given $\{X_0 = j\}$ never takes the value ∞, and that it is called *transient* otherwise. However, note that the definition of recurrence does not say whether the condition expectation $\mathbb{E}[T_j \,|\, X_0 = j]$ is finite or infinite. This leads to the following.

DEF. A recurrent state j is called *positive-recurrent* if $\mathbb{E}[T_j \,|\, X_0 = j] < \infty$ and is called *null-recurrent* if $\mathbb{E}[T_j \,|\, X_0 = j] = \infty$.

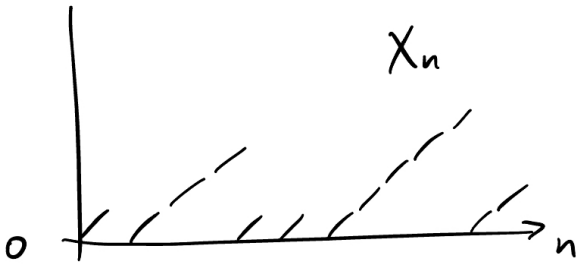
DEF. A MC is *ergodic* if it is irreducible, aperiodic, and positive-recurrent.

THM (fundamental theorem of ergodic MCs)**.** *If $\{X_n\}_{n\geq 0}$ is an ergodic MC, then it has a unique stationary distribution π, and π is a limiting distribution.*

EX (success runs)**.** For $0 < p < 1$, consider the transition matrix

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

on the state space $\{0, 1, 2, \ldots\}$. Of course, its sample paths look like



as we saw before

We claim that this MC is ergodic. As we already saw, it is irreducible. By irreducibility, all states are aperiodic as soon as at least one state is aperiodic, and we see that state 0 is aperiodic since we have $P_{00} = 1 - p > 0$. Also by irreducibility, all states are positive-recurrent as soon as at least one state is positive-recurrent, and we claim that state 0 is positive-recurrent. As we already calulated, we have

$$\mathbb{P}(T_0 > n \,|\, X_0 = 0) = p^n$$

for all $n \geq 0$, since not returning to state 0 in n steps occurs if and only if all of the first n steps are upward jumps. This implies $\mathbb{E}[T_0 \,|\, X_0 = 0] = 1/p$ by the tail sum formula or by noting that the conditional distribution is in fact Geo(p). Thus, this MC is ergodic.

Now the fundamental theorem of ergodic MCs implies that we can find a limiting distribution by finding the unique stationary distribution. To do this, let $\pi = (\pi_0, \pi_1, \ldots)$ be a probability vector satisfying $\pi = \pi P$, and note that this is just

$$\begin{aligned} \pi_0 &= (1-p)(\pi_0 + \pi_1 + \pi_2 + \cdots) \\ \pi_i &= p\pi_{i-1} \text{ for } i \geq 1. \end{aligned}$$

By iterating the second equation, we get $\pi_i = p^i \pi_0$ for all $i \geq 1$. But then the condition

$$\pi_0 + \pi_1 + \cdots = 1$$

implies

$$\pi_0(1 + p + p^2 + \cdots) = \frac{\pi_0}{1-p} = 1,$$

hence we have $\pi_0 = 1 - p$ hence $\pi_i = (1-p)p^i$ for all $i \geq 0$. In other words, π is a Geo(p) random variable.

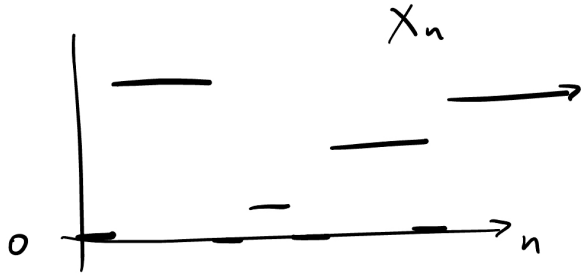
In summary, this MC is ergodic, and its limiting distribution is Geo(p).

EX (jump-hold chain)**.** Let $\alpha = (\alpha_0, \alpha_1, \alpha_2, \ldots)$ be a probability vector with $\alpha_i > 0$ for all $i \geq 0$, and consider the MC with transition matrix

$$P := \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ 1/2 & 1/2 & 0 & 0 & 0 & \cdots \\ 1/3 & 0 & 2/3 & 0 & 0 & \cdots \\ 1/4 & 0 & 0 & 3/4 & 0 & \cdots \\ 1/5 & 0 & 0 & 0 & 4/5 & \cdots \end{pmatrix}$$

For which α is this MC ergodic?

First let's give some interpretation of this MC. From state 0, it jumps to state i with probability α_i. From state $i \geq 1$, it holds in the current state for some length of time, and then jumps back to state 0; the holding time is a geometric random variable on $\{0, 1, 2, \ldots\}$ with parameter $1/(i+1)$ (hence it has expectation i). In other words, its sample paths look like this:



Note that this MC, on average, holds for longer when it is in larger states. So, we should expect that it is ergodic only if α does not put too much mass on large states.

Now we consider ergodicity. Since $\alpha_i > 0$ for all $i \geq 0$, we know that it is irreducible and aperiodic. So, we only need to check positive-recurrence; by irreducibility, it suffices to find a single state that is positive-recurrent. As usual, we consider state 0.

$$\mathbb{E}[T_0 \,|\, X_0 = 0] = \sum_{i=0}^{\infty} \mathbb{E}[T_0 \,|\, X_0 = 0, X_1 = i]\alpha_i \qquad \text{(law of total probability)}$$

$$
\begin{aligned}
&= \sum_{i=0}^{\infty} \mathbb{E}[T_0 \,|\, X_1 = i]\alpha_i && \text{(Markov property)} \\
&= \sum_{i=0}^{\infty} \mathbb{E}\left[1 + \mathrm{Geo}\left(\frac{1}{i+1}\right)\right]\alpha_i && \text{(holding time is geometric)} \\
&= \sum_{i=0}^{\infty} (i+1)\alpha_i \\
&= 1 + \sum_{i=0}^{\infty} i\alpha_i.
\end{aligned}
$$

Therefore, this MC is ergodic if and only if $\sum_{i=0}^{\infty} i\alpha_i < \infty$, meaning α has finite expectation.

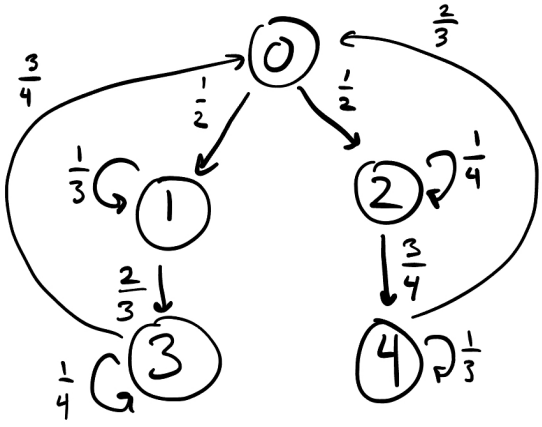
2.9 First transition analysis

Our next focus is a general method for computing some expectations and probabilities of some aspects of MCs. This is *first transition analysis*, which is more like a method rather than a general theory. So, we will illustrate this concept by examples.

EX. In HW2 Q4, we considered the question of computing the expected return time to state 0, denoted $\mathbb{E}[T_0 \,|\, X_0 = 0]$, in the MC on the state space $\{0, 1, 2, 3, 4\}$ whose transition matrix is

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

Previously, you answered this by noting that the MC can be represented by the graph



and then by using the Markov property to show that that the length of any sample path going from 0 to 0 must be the sum of some (independent) geometric random variables.

Now we'll see that we can also solve it by brute calculation. To do this, set

$$
v_i := \mathbb{E}[T_0 \,|\, X_0 = i]
$$

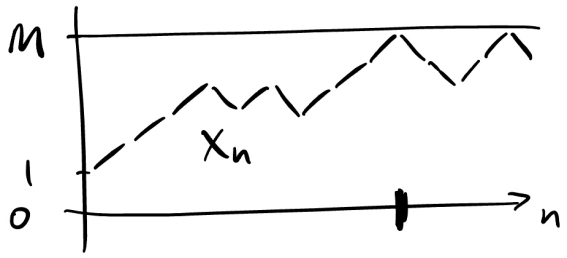
for all states i. Conceptually, this may be somewhat surprising: Even though we are interested only in the value v_0, we can make our life easier by trying to compute all the values $v_0, \ldots, v_4$ simultaneously. The

reason that this is easier is that the law of total probability will imply that these values satisfy a certain linear system which we can directly solve. More specifically, we have

$$\begin{cases} v_0 &= \frac{1}{2}v_1 + \frac{1}{2}v_2 \\ v_1 &= \frac{1}{3}v_1 + \frac{2}{3}v_3 + 1 \\ v_2 &= \frac{1}{4}v_2 + \frac{3}{4}v_4 + 1 \\ v_3 &= \frac{3}{4}0 + \frac{1}{4}v_3 + 1 \\ v_4 &= \frac{2}{3}0 + \frac{1}{3}v_4 + 1. \end{cases}$$

To solve this, note that the last equation implies $v_4 = 3/2$ and that the second to late implies $v_3 = 4/3$. Plugging these into the second and third equations yields $v_2 = 17/6$ and $v_1 = 17/6$. So, we get to the desired answer $v_0 = 17/6$.

EX (SSRW exit location)**.** If $X = \{X_n\}_{n\geq 0}$ is a simple symmetric random walk, what is $\mathbb{P}(X \text{ hits } M \text{ before } 0 \,|\, X_0 = 1)$ as a function of M? Of course we expect this to go to 0 as $M \to \infty$, since for large M it will be very unlikely. A picture to keep in mind for this is the following:



As before, we solve this by defining

$$v_i := \mathbb{P}(X \text{ hits } M \text{ before } 0 \,|\, X_0 = 1)$$

for all states i, and we aim to find a linear system of equations relating these values. As before, this may be unintuitive since we are really just interested in v_1. At any rate, notice that $v_0 = 0$ and $v_M = 1$, as well as

$$v_i = \frac{1}{2}(v_{i-1} + v_{i+1}) \text{ for } 1 \leq i \leq M-1.$$

To solve for v, we note that this is equivalent to

$$v_{i+1} - v_i = v_i - v_{i-1} \text{ for } 1 \leq i \leq M-1.$$

This just says that the increments of v are constant, and that is an equivalent way of stating that v is linear. Thus, there exists some constant $c > 0$ such that $v_i = ci$ for all states i. But the condition $v_M = 1$ then implies $c = 1/M$, hence we have shown

$$v_i = \frac{i}{M} \text{ for } 0 \leq i \leq M.$$

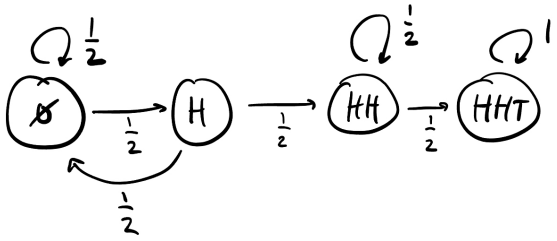
Notice that this gives the correct value for the boundary terms v_0, v_M, as it must. Then we can go back to our original goal and conclude

$$\mathbb{P}(X \text{ hits } M \text{ before } 0 \,|\, X_0 = 1) = \frac{1}{M}.$$

On HW5 you will see an example of a similar calculation, but for the time it takes to hit either 0 or M.

EX (patterns in coin-tossing)**.** Let $Y_1, Y_2, \ldots$ denote successive flips of a fair coin, meaning these are i.i.d. and for each i the value Y_i is equal to either H or T with equal probability (representing "heads" and "tails"). What is the expected number of flips until the most recent three coins show the pattern HHT?

This problem can be solved by introducing a MC which records the "progress towards" seeing HHT. That is, consider a MC $\{X_n\}_{n\geq 0}$ represented by the graph:



Then, the value we are interested in is just $v_\varnothing$, since we start with no progress.

In order to find $v_\varnothing$, we note that the values $v_\varnothing, v_H, v_{HH}, v_{HHT}$ must satisfy the system of equations:

$$\begin{cases} v_\varnothing &= 1 + \frac{1}{2}v_\varnothing + \frac{1}{2}v_H \\ v_H &= 1 + \frac{1}{2}v_\varnothing + \frac{1}{2}v_{HH} \\ v_{HH} &= 1 + \frac{1}{2}v_{HHT} + \frac{1}{2}v_{HH} \\ v_{HHT} &= 0 \end{cases} \tag{2.1}$$

Solving this system of linear equations yields

$$\begin{pmatrix} v_\varnothing \\ v_H \\ v_{HH} \\ v_{HHT} \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \\ 2 \\ 0 \end{pmatrix}.$$

In particular, we get $v_\varnothing = 8$. In HW5, you will do a similar calculation for the expected time it takes to see the pattern HTH.

2.10 Reversibility

In addition to limiting distributions and stationary distributions, there is another type of probability distribution that we should consider in our study of MCs. This gives some simplified calculations in some models of interest.

DEF. A probability distribution π is called a *reversible distribution* for $\{X_n\}_{n\geq 0}$ if we have $\pi_i P_{ij} = \pi_j P_{ji}$ for all states i, j. Equivalently, this is means $X_0 \sim \pi$ implies that (X_0, X_1) and (X_1, X_0) have the same distribution.

Roughly speaking, a reversible distribution is an initial distribution such that the resulting MC looks the same whether we view time going forwards or backwards. For example, the Ehrenfest gas model admits a reversible distribution, but the success runs does not. More generally:

MC	Reversible?
simple symmetric random walk	yes
Ehrenfest gas model	yes
Wright-Fisher model	no
IID sequence	yes
alternating chain	yes
success runs	no
cycle chain	no
jump-hold chain	yes

Reversibility is important because of the following relationship with stationary:

LEM. *If π is reversible for P, then it is distribution for P.*

PF: We directly calculate:

$$\begin{aligned}
(\pi P)_i &= \sum_j \pi_j P_{ji} && \text{(definition of vector-matrix product)} \\
&= \sum_j \pi_i P_{ij} && \text{(definition of reversibility)} \\
&= \pi_i \sum_j P_{ij} && \text{(algebra)} \\
&= \pi_i. && (P \text{ has row sums equal to 1})
\end{aligned}$$

This shows that we can sometimes find a stationary distribution by instead finding a reversible distribution. This is useful since the reversibility condition—sometimes called the *detailed balance condition*—is usually easier to compute than the stationary condition. However, there exist regular MCs which do not admit reversible distributions, so this strategy does not always work.

EX. Consider the MC with transition matrix:

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

This MC is regular (P^2 has all positive entries) and doubly-stochastic hence the uniform distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is the unique stationary distribution. But this distribution is not reversible, since

$$\pi_1 P_{12} = \frac{1}{3} \cdot \frac{1}{3} \neq \frac{1}{3} \cdot \frac{2}{3} = \pi_2 P_{21}.$$

Therefore, P has no reversible distributions.

We saw previously that MCs can be represented (at least visually) as graphs. It turns out that reversibility is closely connected to this, so to explain it precisely let $G = (V, E)$ denote a graph. This means V is an arbitrary vertex set, and $E \subseteq V \times V$ is a collection of edges. Importantly, we think of G as unweighted and undirected. The *degree* of a vertex i is the numer of vertices it is connected to, and this is denoted $\deg(i)$.

The *random walk on* G is the MC on the state space V and with transition matrix

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

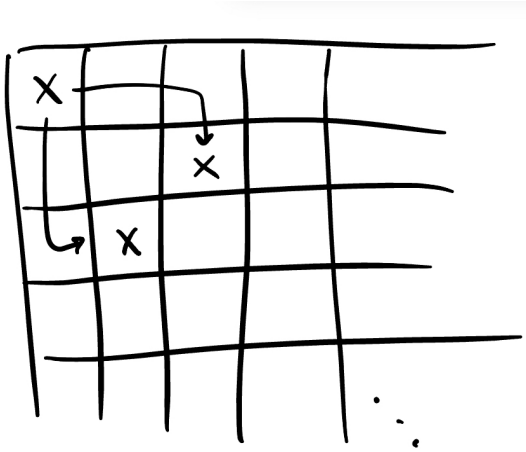
In other words, from the state i, this MC transitions uniformly at random to one of the states that i is connected to. It can easily be shown that every random walk on a graph admits a reversible distribution of the form

$$\pi_i \propto \deg(i).$$

The normalization constant is, of course, $\sum_i \deg(i)$.

This perspective allows us to answer some questions quite easily, for example the following famous problem:

EX (knight's random walk)**.** Consider a knight situated in the corner of a 8×8 chess board, and let $\{X_n\}_{n \geq 0}$ is the MC consisting of the position of the knight, where at each time step it moves uniformly at random to any of the states which constitute a legal chess move. For example, from a corner it has two legal moves:



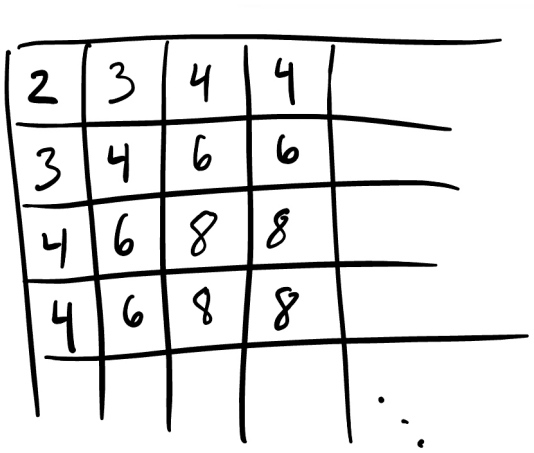
Our question is: How many steps does it take, in expectation, before the knight returns to its starting corner?

To solve this, we observe that this MC is equal to a random walk on a certain graph $G = (V, E)$: We have the vertex set V as the 64 states of the chess board, and we have E consisting of all pairs of states that correspond to a legal move. This MC is irreducible (since the graph G is connected) but it is not aperiodic (all states have period 2). Thus, the fundamental theorem of irreducible MCs applies, and it tells us that we have

$$\mathbb{E}[T_c \,|\, X_0 = c] = \frac{1}{\pi_c}$$

where $c \in V$ represents a corner state.

To compute this, we use the known form of the stationary distribution for random walks on graphs. To get this, we need to compute the degree of each vertex, which is equal to the number of legal moves from each states. We can fill this in as folllows:



Continuing in this way, we find $\sum_{i \in V} \deg(i) = 336$ and $\deg(c) = 2$, hence

$$\mathbb{E}[T_c \mid X_0 = c] = \frac{336}{2} = 168.$$

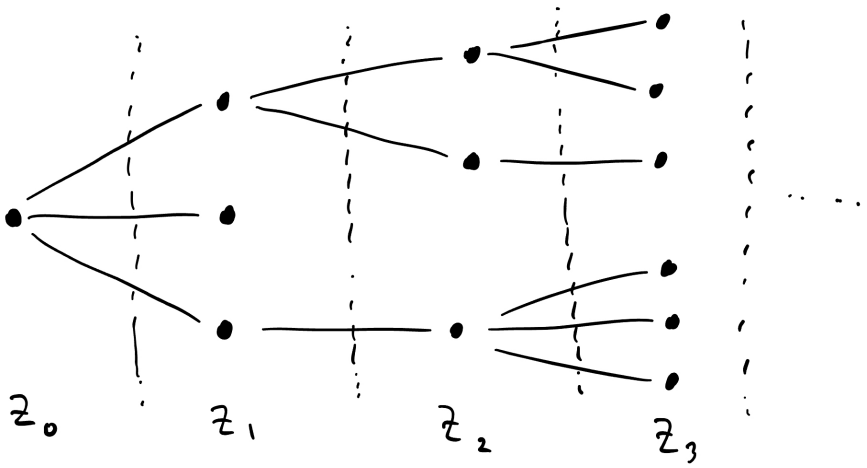
In other words, it takes an expected 168 steps for the knight to return to it starting corner!

2.11 Branching processes

This last part of the class focuses on an important model of population growth from mathematical biology, but which also has applications in many areas. It is a cool subject, since lots of calculations can be done exactly, although with slightly different methods than we used in the previous part of the course.

DEF. Let p be a probability vector on $\{0, 1, 2, \ldots\}$ and define $\{Z_n\}_{n \geq 0}$ as follows. First, $Z_0 = 1$. Then, the conditional distribution of Z_{n+1} given $Z_n = i$ equals $\sum_{j=1}^{i} X_j$, where $X_1, X_2, \ldots$ are IID random variables with distribution p. We say that $\{Z_n\}_{n \geq 0}$ is a *branching process with offspring distribution p.*

A picture to keep in mind for a branching process $\{Z_n\}_{n \geq 0}$ is the following:



Each dot represents an individual, and the vertical collections of dots represent the number of individuals in a given generation. As we can see, each individual gives rise to a random number of individuals in the next generation, and some lineages may die out in finite time.

One of the basic aspects of branching processes is the growth of the expected generation size over time, $\mathbb{E}[Z_n]$. It turns out we can compute this easily. Of course, we have $\mathbb{E}[Z_0] = 1$ and $\mathbb{E}[Z_1] = \sum_{i=0}^{\infty} ip_i$, so let us write $\mu := \sum_{i=0}^{\infty} ip_i$ for this value, which is the mean number of offspring per individual. Then we can compute:

$$\begin{aligned}
\mathbb{E}[Z_2] &= \sum_{i=0}^{\infty} \mathbb{E}[Z_2 \mid Z_1 = i]\mathbb{P}(Z_1 = i) && \text{(law of total probability)} \\
&= \sum_{i=0}^{\infty} \mathbb{E}\left[\sum_{j=1}^{i} X_j\right] p_i && \text{(definition of branching process)} \\
&= \sum_{i=0}^{\infty} i\mu p_i \\
&= \mu^2.
\end{aligned}$$

More generally, one can iterate this process to find

$$\mathbb{E}[Z_n] = \mu^n$$

for all $n \geq 0$. In particular, we have

$$\lim_{n\to\infty} \mathbb{E}[Z_n] = \begin{cases} 0 & \text{if } \mu < 1 \\ 1 & \text{if } \mu = 1 \\ \infty & \text{if } \mu > 1 \end{cases} \tag{2.2}$$

This shows that the theory of branching process can be divided into three cases, based on the mean μ of the offspring distribution p: A branching process $\{Z_n\}_{n\geq 0}$ is called *sub-critical* if $\mu < 1$, *critical* if $\mu = 1$, and *super-critical* if $\mu > 1$.

Another interesting aspect of branching process is to determine the probability of "extinction", which means there exists some $n \geq 0$ with $Z_n = 0$. Of course, if $p_0 > 0$, then there is always positive probability of extinction, since we have $\mathbb{P}(Z_1 = 0) = p_0$. So, a better question is whether we have extinction with probablity 1, or with some probability strictly between 0 and 1. It turns out we can answer this very precisely, but we need to have a small digression.

DEF. For a probability vector p on $\{0, 1, 2, \ldots\}$, its *probability generating function (PGF)* is the function $G_p : [0, 1] \to \mathbb{R}$ defined via $G_p(s) := \sum_{i=0}^{\infty} s^i p_i$. Equivalently, if X is a non-negative-integer-valued random variable, its PGF is the funciton $G_X : [0, 1] \to \infty$ defined via $G_X(s) = \mathbb{E}[s^X]$.

For instance, if X is Poisson random variable with rate λ, then we can compute:

$$\begin{aligned}
G_X(s) &= \sum_{i=0}^{\infty} s^i \cdot e^{-\lambda} \frac{\lambda^i}{i!} && \text{(definition of PGF)} \\
&= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda s)^i}{i!} && \text{(algebra)} \\
&= e^{-\lambda} e^{\lambda s} && \text{(power series for } e^z\text{)} \\
&= e^{\lambda(s-1)}.
\end{aligned}$$

It's similarly easy to explicit compute the PGF for other random variables of interest.

There are many reasons why the PGF is important, and they are all related to the fact that the PGF encodes lots of information about a random variable or its distribution. For instance, if X has distribution p, then note

$$G_p(0) = p_0 = \mathbb{P}(X = 0)$$

and

$$G_p'(0) = p_1 = \mathbb{P}(X = 1)$$

and

$$G_p'(1) = \sum_{i=0}^{\infty} i p_i = \mathbb{E}[X].$$

Similarly, we can compute the variance and higher moments of X through higher derivatives of G_X.

In fact, the PGF contains "all of the information" in a non-negative-integer-valued random variable, in the following sense:

LEM. *Random variables X, Y with values in $\{0, 1, 2, \ldots\}$ have the same distribution if and only if $G_X = G_Y$.*

For our purposes, something useful about the PGF is that it is easy to analyze the PGF of a random sum of IID random variables. This of course is related to branching processes, since each generation in a branching process is an IID sum of a random number of random variables.

LEM. *If N is a non-negative-integer-valued random variable, and $X_1, X_2, \ldots$ are IID non-negative-integer-valued random variables which are also independent of N, then the random variable*

$$Z := \sum_{i=1}^{N} X_i$$

has PGF equal to $G_Z(s) = G_N(G_X(s))$.

PF: Just compute

$$\begin{aligned}
G_Z(s) &= \mathbb{E}[s^Z] && \text{(definition of PGF)} \\
&= \sum_{n=0}^{\infty} \mathbb{E}[s^Z \mid N = n]\mathbb{P}(N = n) && \text{(law of total probability)} \\
&= \sum_{n=0}^{\infty} \mathbb{E}[s^{\sum_{i=1}^{n} X_i}]\mathbb{P}(N = n) && \text{(definition of } Z\text{)} \\
&= \sum_{n=0}^{\infty} \left(\mathbb{E}[s^{X_1}]\right)^n \mathbb{P}(N = n) && \text{(independence)} \\
&= \sum_{n=0}^{\infty} \left(G_X(s)\right)^n \mathbb{P}(N = n) && \text{(definition of PGF)} \\
&= G_N(G_X(s)) && \text{(definition of PGF)}
\end{aligned}$$

for all $0 \leq s \leq 1$.

An application is the following:

LEM (Wald's identity)**.** *If N is a non-negative-integer-valued random variable and $X_1, X_2, \ldots$ are IID non-negative-integer-valued random variables which are also independent of N, then*

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\mathbb{E}[X_i].$$

PF: Writing $Z := \sum_{i=1}^{N} X_i$ and using $G_Z(s) = G_N(G_X(s))$ from above, we have

$$\mathbb{E}[Z] = G_Z'(1)$$

$$
\begin{aligned}
&= G_N'(G_X(1))G_X'(1) && \text{(chain rule)}\\
&= G_N'(1)G_X'(1) && (G_X(1) = 1)\\
&= \mathbb{E}[N]\mathbb{E}[X_1].
\end{aligned}
$$

One can also derive this result by applying the law of total probability directly, but it's nice to see that it can be deduced from results which we already know.

Now let us return to branching processes. We can use the result above to, in principle, find the distribution of Z_n for each $n \geq 0$, where $\{Z_n\}_{n\geq 0}$ is a branching process with an arbitrary offspring distribution p. Indeed, we have $G_{Z_1} = G_p$ by definition. Then applying the result above gives $G_{Z_2}(s) = G_p(G_p(s))$. Similarly,

$$G_{Z_3}(s) = G_p(G_p(G_p(s))).$$

Continuing in this way, we get

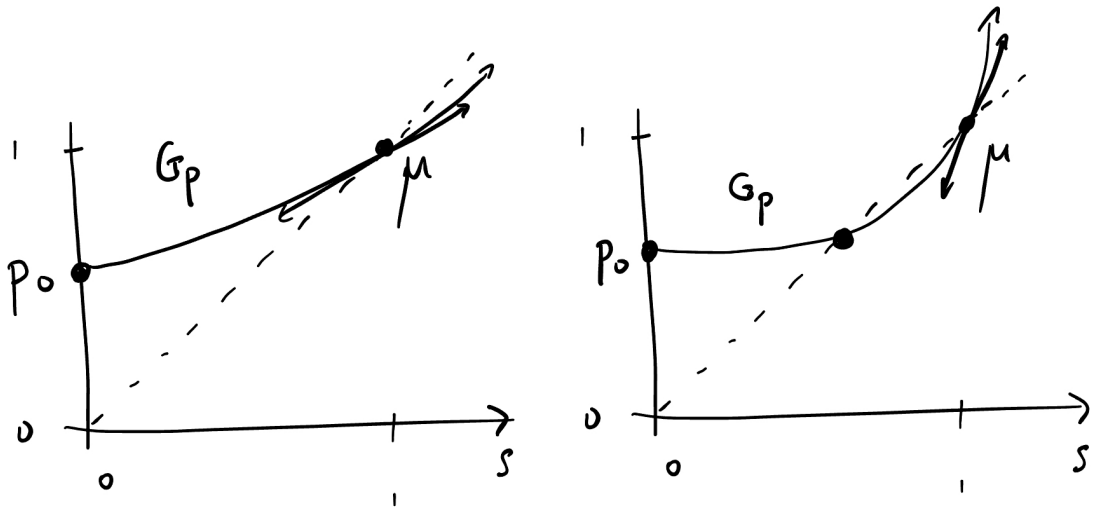
$$G_{Z_n}(s) = (\underbrace{G_p \circ \cdots \circ G_p}_{n \text{ times}})(s)$$

for all $n \geq 0$. Of course, this may be hard to compute exactly.

Nonetheless, iterations of the PGF arise in the following result which describes how to find the extinction probability for a general branching process:

THM. *In a branching process with offspring distribution p, the probability of extinction equals the smallest solution $0 \leq s \leq 1$ to the equation $s = G_p(s)$.*

Lastly, we give some remarks about this result. First, note that there is a nice visualization of the theorem as follows, where we assume $p_0 > 0$:



The left plot represents the sub-critical case where $\mu < 1$ and the right plot represents the super-critical case where $\mu > 1$; recall that $\mu = \mathbb{E}[X]$ is exactly $G_p'(1)$, the slope of G_p at $s = 1$. Since G_p is always convex (it is easy to show $G_p''(s) \geq 0$) we see that $\mu < 1$ implies that there are no solutions of $s = G_p(s)$ on $0 \leq s < 1$, and that $\mu > 1$ implies that there exists a solution of $s = G_p(s)$ on $0 \leq s < 1$. In particular, in a sub-critical branching process, the probability of extinction is 1, and in the super-critical branching process, the probability of extinction is strictly in between 0 and 1.

The critical case $\mu = 1$ deserves further attention. If the slope of G_p is exactly 1, then we see that—like the sub-critical case—there are no solutions to $s = G_p(s)$ on $0 \leq s < 1$. Therefore, critical branching

processes also have extinction with probability 1. This may seem to contradict our earlier calculation that the mean generating size is constant in a critical branching process. In fact, this is not a contradiction; a critical branching process has extinction occurring at a very large random time, so this does not effect very much the expected generation size!

3 Poisson processes

3.1 Memorylessness

In the first part of the class (on discrete-time MCs), we saw that geometric random variables appeared very often. In this next part of the class (on Poisson processes and continuous-time MCs), we will see that exponential random variables appear very often. This is because they are closely related to the Markov property.

DEF. A non-negative random variable T is *memoryless* if

$$\mathbb{P}(T > t \,|\, T > s) = \mathbb{P}(T > t - s)$$

for all $0 < s < t$.

It is usually useful to think of T as the waiting until an event A occurs. Then T being memoryless, roughly speaking, means the following: If we know that A has not yet occurred, the remaining time it takes for A to occur has the same distribution as the original time it takes for A to occur.

Now consider a MC $\{X_n\}_{n\geq 0}$ and let A be the event that X_n first differs from X_0. It makes sense that the waiting time for A to occur must be memoryless; if it was not, then $\{X_n\}_{n\geq 0}$ would not be a MC!

In HW6 you will show that the only memoryless random variable taking values in $\{0, 1, 2, \ldots\}$ is geometric random variable, and that the only memoryless random variable taking values in $[0, \infty)$ is an exponential random variable. This partially explains why we saw so many geometric random variables, and why we will soon see so many exponential random variables.

3.2 Basic definitions

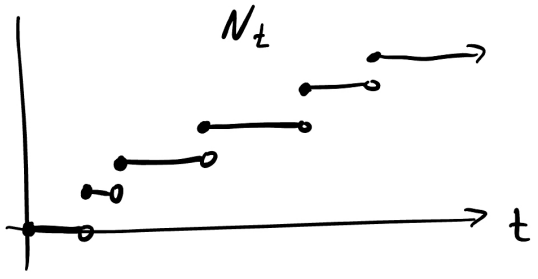
DEF. A *counting process* is a continuous-time stochastic process $\{N_t\}_{t\geq 0}$ satisfying

(1) $N_0 = 0$,

(2) N_t is a non-negative integer for each $t \geq 0$, and

(3) We have $N_s \leq N_t$ for all $0 \leq s \leq t$.

Note that the definition of counting process does not involve any probability; it is just a description of the sample paths. They look something like this:



(Observe that there are discontinuities, and we have to decide whether to fill in the top of bottom of the jump. This corresponds to whether we want $\{N_t\}_{t\geq 0}$ to be "left-continuous" or "right-continuous". While we will not care about that in this class, we emphasize that they are usually assumed to be right-continuous, as shown above.)

The probabilistic properties we impose are the following:

DEF. A *Poisson process of rate* $\lambda > 0$ is a counting process $\{N_t\}_{t\geq 0}$ satisfying

(1) $N_t - N_s$ has a $\text{Poi}(\lambda(t-s))$ distribution for all $0 \leq s \leq t$, and

(2) $N_t - N_s$ is independent of N_s for all $0 \leq s \leq t$.

In any counting process $\{N_t\}_{t\geq 0}$, terms like $N_t - N_s$ are called *increments*. Thus, the defining properties of Poisson processes are sometimes called (1) *stationary increments* and (2) *independent increments*. As we will see, Poisson processes are convenient to study because many calculations involving them can be done exactly. Some simple examples of this are as follows.

EX. For non-negative integers $0 \leq a \leq b$ and times $0 \leq s \leq t$ we have

$$\begin{aligned}
&\mathbb{P}(N_s = a, N_t = b) \\
&\qquad = \mathbb{P}(N_s = a, N_t - N_s = b - a) \\
&\qquad = \mathbb{P}(N_s = a)\mathbb{P}(N_t - N_s = b - a) && \text{(independent increments)} \\
&\qquad = \mathbb{P}(\text{Poi}(\lambda s))\mathbb{P}(\text{Poi}(\lambda(t-s)) = b - a) \\
&\qquad = e^{-\lambda s}\frac{(\lambda s)^a}{a!} \cdot e^{-\lambda(t-s)}\frac{(\lambda(t-s))^{b-a}}{(b-a)!} && \text{(Poisson distribution)}
\end{aligned}$$

EX. For any $t \geq 0$ we have

$$\mathbb{P}(N_{2t} = N_t) = \mathbb{P}(N_{2t} - N_t = 0) = \mathbb{P}(\text{Poi}(\lambda t) = 0) = e^{-\lambda t}$$

As we will see in the next subsection, it is not actually necessary to specify that N_t has a Poisson distribution for each $t \geq 0$; rather, we will show that this follows from other assumptions.

3.3 Law of rare events

You are probably easily able to "spot Binomial distributions in the wild" since we very often count the number of successes in independent trials of events which have the same probability of success. At the same time, you are probably not so familiar with "spotting Poisson distributions in the wild"; when you first learn about Poisson random variables, it's not really clear why they should be an important example of a count-type random variable.

It turns out that Poisson random variables are canonical in a similar sense that Gaussian random variables are canonical: They arise as limits of Binomials under certain conditions. This is captured in the following result.

THM (law of rare events)**.** *For fixed* $\lambda > 0$ *and* $k \geq 0$*, we have*

$$\mathbb{P}\left(\text{Binom}\left(n, \frac{\lambda}{n}\right) = k\right) \to \mathbb{P}(\text{Poi}(\lambda) = k)$$

as $n \to \infty$*.*

PF: The proof is actually quite simple; we have

$$\mathbb{P}\left(\text{Binom}\left(n, \frac{\lambda}{n}\right) = k\right)$$

$$
\begin{aligned}
&= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} && \text{(definition of Binomial)}\\
&= \frac{\left(1-\frac{0}{n}\right)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{n-k-1}{n}\right)}{k!}\lambda^k\left(1-\frac{\lambda}{n}\right)^{-k}\left(1-\frac{\lambda}{n}\right)^{n} && \text{(algebra)}\\
&= \frac{e^{-\lambda}\lambda^k}{k!} && \text{(evaluating limits)}\\
&= \mathbb{P}(\text{Poi}(\lambda)=k) && \text{(definition of Poisson)}
\end{aligned}
$$

for all k.

Using this result, we can get the following:

THM. *If a counting process $\{N_t\}_{t\geq 0}$ satisfies*

(1) $N_t - N_s$ has the same distribution as N_{t-s} for all $0 \leq s \leq t$, and

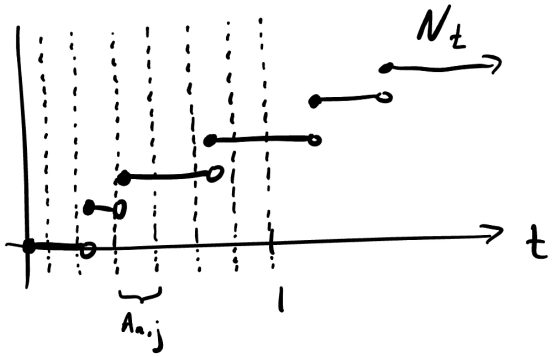
(2) $N_t - N_s$ is independent of N_s for all $0 \leq s \leq t$,

then it is a Poisson process.

PF: We only need to show that there exists some $\lambda > 0$ such that N_t has a $\text{Poi}(\lambda t)$ distribution for all $t \geq 0$. To do this, we focus on time $t = 1$ and we will use the law of rare events to show that N_t has distribution $\text{Poi}(\lambda)$ for some λ. To set it up, consider dividing time $[0, 1]$ into n sub-intervals,

$$\left[0, \frac{1}{n}\right], \left[\frac{1}{n}, \frac{2}{n}\right], \ldots, \left[\frac{n-1}{n}, 1\right].$$

Generally, let's write $A_{n,j}$ for the jth interval above, for $1 \leq j \leq n$. We can picture this set-up as follows:



Now define

$$Y_n := \sum_{j=1}^{n} \mathbf{1}\Big\{\{N_t\}_{t\geq 0} \text{ has a jump during the time interval } A_{n,j}\Big\}.$$

Observe that (2) implies that Y_n is a count of the number of success in n independent trials, and that (1) implies that each trial has the same success probability,

$$p_n := \mathbb{P}\Big(\{N_t\}_{t\geq 0} \text{ has a jump during the time interval } A_{n,1}\Big).$$

In particular, Y_n has distribution Binom(n, p_n). In fact, we can show that there is some $\lambda > 0$ such that $p_n \approx \lambda/n$. To do this, we calculate:

$$\begin{aligned}
\mathbb{E}[N_1] &= \mathbb{E}\left[\sum_{j=1}^{n}\left(N_{\frac{j}{n}} - N_{\frac{j-1}{n}}\right)\right] && \text{(telescoping sum)} \\
&= \sum_{j=1}^{n}\mathbb{E}\left[N_{\frac{jt}{n}} - N_{\frac{(-1}{n}}\right] && \text{(linearity of expectation)} \\
&= n\mathbb{E}\left[N_{\frac{1}{n}}\right] && \text{(stationary increments)} \\
&\geq n\mathbb{P}\left(N_{\frac{1}{n}} \geq 1\right) && \text{(Markov's inequality)} \\
&\approx n\mathbb{P}\left(\{N_t\}_{t\geq 0} \text{ has a jump during } A_{n,1}\right) \\
&= np_n.
\end{aligned}$$

Here, we used Markov's inequality to get an upper bound on p_n, and it turns out that one can use similar methods to get a lower bound on p_n. (But we skip the details for now.) This shows

$$np_n \approx \mathbb{E}[N_1].$$

So, if we set $\lambda := \mathbb{E}[N_1]$, then the law of rare events finishes the proof.

The precise details of the preceding proof are not so important. (In fact, we skipped a lot of details, anyway.) What is important is that this result highlights the role of the Poisson distribution: Any counting process with stationary and independent increments must have Poisson increments!

3.4 Markov property

It turns out that a Poisson process posesss a form of the Markov property, similar to what we saw for discrete-time MCs. We will discuss this in more detail in the next part of the course on continuous-time Markov chains, but for now we state the following:

THM. *A Poisson process $\{N_t\}_{t\geq 0}$ has the* Markov property *that the conditional distribution of N_t given $\{N_r = a, N_s = b\}$ equals the conditional distribution of N_t given $\{N_s = b\}$ for all $0 \leq r \leq s \leq t$ and all $0 \leq a \leq b$.*

As before, this means the conditional distribution of the future of a Poisson process given its entire past is the same as its conditional distribution given the current state.

In fact, it is true (roughly speaking) that any counting process with the Markov property is a Poisson process. But this takes a lot of notions to state precisely, so we do not pursue it in this class.

The Markov property is useful for a few reasons. First, it allows us to derive a representation of Poisson processes in terms of some exponential and Gamma random variables, as we will do in the next part of the class. Second, it provides some conceptual insight into some calculations we have already done in the class and in the homework.

EX. In HW6 Q4, you directly calculated the distribution N_T of a Poisson process $\{N_t\}_{t\geq 0}$ of rate $\lambda > 0$ evaluated at an independent exponential time T with rate μ, and the result is a certain geometric random variable. This raises the question: What was the independent trial that was being repeated until observing a success? (You should ask yourself this question any time you find a geometric random variable in the wild!)

To get at this, first note that the event $\{N_T \geq 1\}$ is the same as $\{T \leq S_1\}$. But we know that T has distribution Exp(μ) and S_1 has distribution Exp(λ) and that they are independent, hence from HW Q1 we

have

$$\mathbb{P}(N_T \geq 1) = \mathbb{P}(T \leq S_1) = \frac{\mu}{\lambda + \mu}.$$

We interpret this as follows: There is a race between T and S_1, and we want to S_1 to finish first. So, we think of $\{T \leq S_1\}$ as a failure and $\{T > S_1\}$ as a success. The calculation above shows that the probability of success is just $\mu/(\lambda + \mu)$.

Next consider the event $\{N_T \geq 2\}$. Because $\{N_T \geq 2\} \subseteq \{N_T \geq 1\}$, we of course have:

$$\mathbb{P}(N_T \geq 2) = \mathbb{P}(N_T \geq 2 \,|N_T \geq 1)\mathbb{P}(N_T \geq 1)$$

We already computed $\mathbb{P}(N_T \geq 1)$ above, so it only remans to compute $\mathbb{P}(N_T \geq 2 \,|N_T \geq 1)$. We interpret this as follows: There is a race between T and S_2, and we want to S_2 to finish first. So, we think of $\{T \leq S_2\}$ as a failure and $\{T > S_2\}$ as a success. However, we now want to compute the conditional probability of success in the second race given success in the first race. Here is the conceptual key: By the Markov property, this has the same probability of success as the first race! So we have

$$\mathbb{P}(N_T \geq 2) = \left(\frac{\mu}{\lambda + \mu}\right)^2.$$

More generally, we have

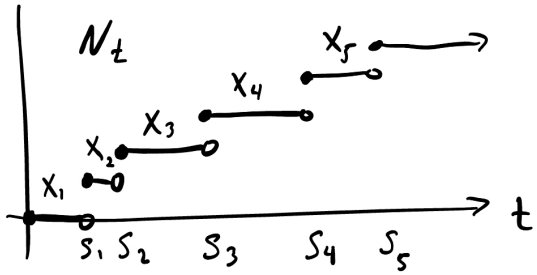
$$\mathbb{P}(N_T \geq k) = \left(\frac{\mu}{\lambda + \mu}\right)^k$$

for $k \geq 0$. In other words, N_T has a geometric distribution, as you already computed.

3.5 Arrival and interarrival times

Our earlier definition of Poisson process (a counting process with stationary and independent increments) is useful for some calculations, but not for others. Now we give an alternative (but equivalent) definition which is also sometimes useful.

THM. *Let $X_1, X_2, \ldots$ be IID exponential random variables with rate $\lambda > 0$, and set $S_n := X_1 + \cdots + X_n$ and $N_t := \max\{n \geq 0 : S_n \leq t\}$. Then, $\{N_t\}_{t \geq 0}$ is a Poisson process of rate $\lambda > 0$. The random variables $S_1, S_2, \ldots$ are called the arrival times and the random variables $X_1, X_2, \ldots$ are called the interarrival times.*



Observe that S_k has a Gamma distribution with shape k and rate λ, simply because it is a sum of k IID exponential random variables.

EX. Let $\{N_t\}_{t \geq 0}$ be a Poisson process of rate λ, let $\{Y_n\}_{n \geq 0}$ be a Markov chain on the state space $\{0, 1, 2, \ldots\}$ satsifying $P_{00} = 0$, and suppose that $\{N_t\}_{t \geq 0}$ and $\{Y_n\}_{n \geq 0}$ are independent. For $\{\tilde{Y}_t\}_{t \geq 0}$ defined via $\tilde{Y}_t = Y_{N_t}$ for $t \geq 0$, we want to compute

$$\mathbb{E}\left[\tilde{T}_0 \,|\, \tilde{Y}_0 = 0\right]$$

where $\tilde{T}_0$ is the time of the first return of $\{\tilde{Y}_t\}_{t\geq 0}$ to state 0.

There is an analogous quantity T_0 for the discrete-time MC, and we know that T_0 and $\tilde{T}_0$ are related via

$$\tilde{T}_0 = \sum_{i=1}^{T_0} T_i.$$

In other words, $\tilde{T}_0$ is a random sum of exponentials, where the number of terms in the sum is the return time for a discrete-time MC.

By Wald's identity, we therefore have

$$\mathbb{E}\left[\tilde{T}_0 \,|\, \tilde{Y}_0 = 0\right] = \mathbb{E}[X_1]\mathbb{E}[T_0 \,|\, Y_0 = 0] = \frac{\mathbb{E}[T_0 \,|\, Y_0 = 0]}{\lambda}.$$

We can interpret this as a sort of time-dilation: the return time for the continuous-time MC is just proportional to the return time for the discrete-time MC, where the proportionality constant is exactly the speed of the Poisson process.

(Note that we needed $P_{00} = 0$ in order for this to work; if $P_{00} > 0$, then the jumps of $\{Y_n\}_{n\geq 0}$ will be impossible to detect in $\{\tilde{Y}_t\}_{t\geq 0}$.)

3.6 Conditioning and invariances

Our existing discussions of Poisson processes have been mostly conceptual. In order to do interesting calculations, we will exploit some fundamental properties that we now explain.

The first result is that the arrival times of a Poisson process become easy to analyze if we condition on the number of arrivals in a fixed interval:

THM. *If $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$, then the conditional distribution of $S_1, \ldots, S_{N_t}$ given $\{N_t = k\}$ is equal to a sorted list of IID uniformly-distributed random variables on $[0, t]$.*

Applications of this theorem usually require some notation and results on sorting IID random variables. That is, if $U_1, \ldots, U_k$ are any random variables, we write $U_{(1)}, \ldots, U_{(k)}$ for these values sorted into increasing order. In fact, if $U_1, \ldots, U_k$ are IID uniform random variables on $[0, 1]$, then $U_{(j)}$ has distribution $\text{Beta}(j, n + 1 - j)$ for all $1 \leq j \leq k$.

We also note that the rate λ does not appear in the result above; this is not a mistake! It means that the rate of a Poisson processes enters only through the number of arrivals, but not in the positions of the arrivals. In some sense, this means λ represents the "speed" of the Poisson process.

Now we can see some applications:

LEM (Campbell's formula)**.** *If $\{N_t\}_{t\geq 0}$ is a Poisson process of rate $\lambda > 0$ with arrival times $S_1, S_2, \ldots$ then we have*

$$\mathbb{E}\left[\sum_{k=1}^{N_t} f(S_k)\right] = \lambda \int_0^t f(u)\,\mathrm{d}u$$

for any function $f : [0, \infty) \to \mathbb{R}$.

PF: Let us compute the following, where $U_1, \ldots, U_k$ denote an IID sequence of uniform random variables on $[0, t]$, and $U_{(1)} \ldots, U_{(k)}$ denote the order statistics:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{N_t} f(S_k)\right] \\
&\quad = \sum_{i=0}^{\infty} \mathbb{E}\left[\sum_{k=1}^{N_t} f(S_k) \,\middle|\, N_t = i\right] \mathbb{P}(N_t = i) && \text{(law of total probability)}
\end{aligned}$$

$$= \sum_{i=0}^{\infty} \mathbb{E}\left[\sum_{k=1}^{i} f(U_{(k)})\right] \mathbb{P}(N_t = i) \qquad \text{(conditioning theorem)}$$

$$= \sum_{i=0}^{\infty} \mathbb{E}\left[\sum_{k=1}^{i} f(U_k)\right] \mathbb{P}(N_t = i) \qquad \text{(rearranging the sum)}$$

$$= \sum_{i=0}^{\infty} \frac{i}{t} \int_0^t f(u)\,\mathrm{d}u\, \mathbb{P}(N_t = i) \qquad \text{(uniform distribution)}$$

$$= \frac{1}{t} \int_0^t f(u)\,\mathrm{d}u\, \mathbb{E}[N_t] \qquad \text{(algebra)}$$

$$= \lambda \int_0^t f(u)\,\mathrm{d}u. \qquad \text{(Poisson distribution)}$$

An important special case is the following application in finance:

EX (expected present value of future income)**.** Suppose that $\{N_t\}_{t\geq 0}$ is a Poisson process or rate $\lambda > 0$, and that at each arrival time we obtain \$1. However, the value today of \$1 received at time $t > 0$ in the future is considered to be $e^{-\beta t}$, for some factor $\beta > 0$ called the *discount factor*. Then, the expected present value of al future income is

$$\mathbb{E}\left[\sum_{k=1}^{\infty} e^{-\beta S_k}\right].$$

We can use Campbell's formula to compute this. Indeed, fix $t > 0$ and consider the function $f(u) = e^{-\beta u}$. Then we have

$$\mathbb{E}\left[\sum_{k=1}^{N_t} e^{-\beta S_k}\right] = \lambda \int_0^t e^{-\beta u}\,\mathrm{d}u = \lambda \cdot \frac{1 - e^{-\beta t}}{\beta},$$

hence

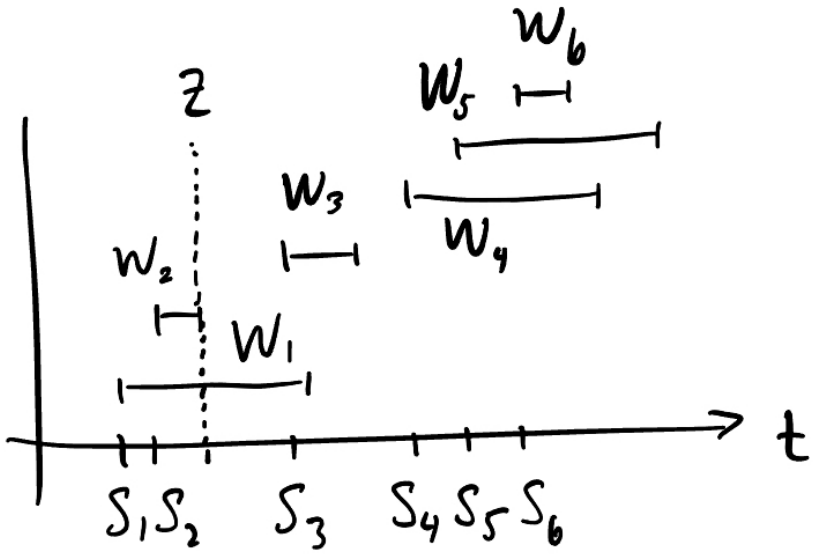
$$\mathbb{E}\left[\sum_{k=1}^{\infty} e^{-\beta S_k}\right] = \lim_{t\to\infty} \mathbb{E}\left[\sum_{k=1}^{N_t} e^{-\beta S_k}\right] = \lambda \cdot \lim_{t\to\infty} \frac{1 - e^{-\beta t}}{\beta} = \frac{\lambda}{\beta}.$$

This simple conclusion (expected present value of future income is inversely proportional to discount factor) is very important in mathematical finance.

EX (first job completion in $M/G/\infty$ queue)**.** Let $\{N_t\}_{t\geq 0}$ denote a Poisson process of rate $\lambda > 0$ with arrival times $S_1, S_2, \ldots$ and let $W_1, W_2, \ldots$ be IID samples from some non-negative continuous distribution, such that all W are independent of all S. Set

$$Z := \min\{S_1 + W_1, S_2 + W_2, \ldots\},$$

and let us find the distribution of Z. We can visualize the problem as follows:



Here, the vertical axis does not mean anything in particular; we just use the vertical spacing to separate the horizontal bars representing the processing times.

This model has an interpretation in terms of queueing theory. Suppose that S_k represents the arrival time of customer k, that W_k represents the processing time of customer k, and that there are infinitely many servers to process customers. Then, Z is the time of the first completed job in this queue. This model is usually referred to as a $M/G/\infty$ queue (which corresponds to Markov interarrival times, General processing times, and ∞-ly many servers.)

Our goal is to find $\mathbb{P}(Z > z)$, in terms of the rate λ of $\{N_t\}_{t\geq 0}$ and the CDF F of W. To get there, we will first study

$$Z_t := \min\{S_1 + W_1, S_2 + W_2, \ldots, S_{N_t} + W_{N_t}\}$$

and use the fact that $Z_t \to Z$ hence $\mathbb{P}(Z_t > z) \to \mathbb{P}(Z > z)$ as $t \to \infty$. What does Z_t represent? It is the time of the first completed job, among the jobs which have started by time t. Note that this is different than

$$Z_k := \min\{S_1 + W_1, S_2 + W_2, \ldots, S_k + W_k\},$$

which represents the of the first completed job, among the first k jobs. As we will see, the randomization actually helps us since we can apply the conditioning formula.

To compute $\mathbb{P}(Z_t > z)$, let $U_1, \ldots, U_i$ denote IID random variables that are uniform on $[0, t]$, and note:

$$\begin{aligned}
&\mathbb{P}(Z_t > z) \\
&= \mathbb{P}(S_k + W_k > z \text{ for all } 1 \leq k \leq N_t) && \text{(definition of } Z_t\text{)} \\
&= \sum_{i=0}^{\infty} \mathbb{P}(S_k + W_k > z \text{ for all } 1 \leq k \leq N_t \,|\, N_t = i)\mathbb{P}(N_t = i) && \text{(law of total probability)} \\
&= \sum_{i=0}^{\infty} \mathbb{P}(U_{(k)} + W_k > z \text{ for all } 1 \leq k \leq i)\mathbb{P}(N_t = i) && \text{(conditioning theorem)} \\
&= \sum_{i=0}^{\infty} (\mathbb{P}(U_1 + W_1 > z))^i \mathbb{P}(N_t = i). && \text{(independence)}
\end{aligned}$$

To simplify this, write $p(t,z) := \mathbb{P}(U_1 + W_1 > z)$, so that

$$\begin{aligned}\mathbb{P}(Z_t > z) &= \sum_{i=0}^{\infty} (p(t,z))^i \mathbb{P}(N_t = i) && \text{(above)}\\ &= \exp(-\lambda t(1-p(t,z))) && \text{(Poisson PGF)}\end{aligned}$$

for all $t, z \geq 0$. Now we compute $p(t,z)$. As long as $t \geq z$, we have:

$$\begin{aligned}1 - p(t,z) &= \mathbb{P}(U_1 + W_1 \leq z)\\ &= \frac{1}{t}\int_0^t \mathbb{P}(u + W_1 \leq z)\,\mathrm{d}u && \text{(law of total probability)}\\ &= \frac{1}{t}\int_0^z \mathbb{P}(W_1 \leq z-u)\,\mathrm{d}u\\ &= \frac{1}{t}\int_0^z \mathbb{P}(W_1 \leq v)\,\mathrm{d}v && \text{(change variable } v = z-u)\end{aligned}$$

So, by fixing $z \geq 0$ and taking $t \to \infty$, we conclude:

$$\mathbb{P}(Z > z) = \lim_{t\to\infty} \mathbb{P}(Z_t > z) = \exp\left(-\lambda \int_0^z F(x)\,\mathrm{d}x\right).$$

When F is known then usually the right side above can be exactly computed.

It is also interesting to think of what happens when $z \to \infty$. That, what is the probability that the first job completion takes a very long time? Since $F(x) = 1 - \mathbb{P}(W_1 > x)$ for all $x \geq 0$, we have the following for large z:

$$\begin{aligned}\mathbb{P}(Z > z) &= \exp\left(-\lambda \int_0^z F(x)\,\mathrm{d}x\right) && \text{(above)}\\ &= \exp\left(-\lambda z + \lambda \int_0^z \mathbb{P}(W_1 > x)\,\mathrm{d}x\right) && \text{(algebra)}\\ &\approx \exp\left(-\lambda z + \lambda \mathbb{E}[W]\right) && \text{(tail sum formula)}\\ &\propto e^{-\lambda z}.\end{aligned}$$

In other words, the probability that the first job completion takes a very long time is approximately the same as the probability that an $\mathrm{Exp}(\lambda)$ random variable is large.

In HW7, you will study some further aspects of the $M/G/\infty$ queue using similar techniques.

Next we discuss a pair of properties that are useful for some further conceptual problems.

THM (superposition)**.** *If $\{N_t^{(1)}\}_{t\geq 0}, \ldots, \{N_t^{(k)}\}_{t\geq 0}$ are independent Poisson processes with rates $\lambda_1, \ldots, \lambda_k > 0$ respectively, then the counting process $\{N_t\}_{t\geq 0}$ defined via*

$$N_t := N_t^{(1)} + \cdots + N_t^{(k)}$$

is a Poisson process with rate $\lambda := \lambda_1 + \cdots + \lambda_k$.

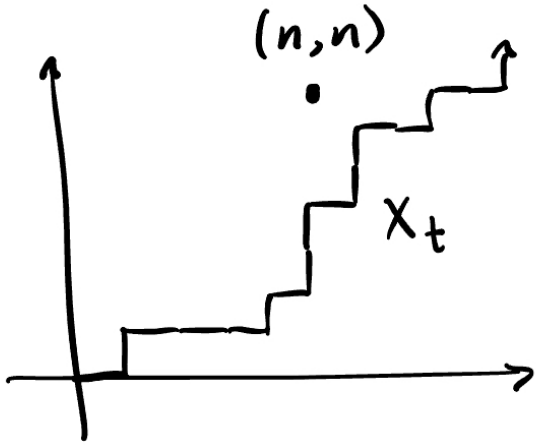
THM (thinning)**.** *Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate $\lambda > 0$, and let $p = (p_1, \ldots, p_k)$ be a probability vector. Suppose that each arrival of $\{N_t\}_{t\geq 0}$ is assigned a* type *in the set $\{1, \ldots, k\}$ according to the probability vector p, and that all types are independent of each other and independent of $\{N_t\}_{t\geq 0}$. Now let $\{N^{(j)}\}_{t\geq 0}$ denote the counting process of all type j arrivals for each $1 \leq j \leq k$. Then, $\{N_t^{(1)}\}_{t\geq 0}, \ldots, \{N_t^{(k)}\}_{t\geq 0}$ are independent Poisson processes with rates $\lambda p_1, \ldots, \lambda p_k > 0$.*

It is good to think of thinning and superposition as inverses of each other, although they are often used together.

EX. Suppose $\{N_t\}_{t\geq 0}$ and $\{M_t\}_{t\geq 0}$ are independent Poisson processes with the same rate $\lambda > 0$, and define the stochastic process $\{X_t\}_{t\geq 0}$ in $\mathbb{Z}^2$ defined as $X_t = (N_t, M_t)$. Note that the sample paths of $\{X_t\}_{t\geq 0}$ are "up-right paths" in $\mathbb{R}^2$. For $n \geq 0$, let us compute the probability

$$\mathbb{P}\left(\{X_t\}_{t\geq 0} \text{ ever visits the state } (n,n)\right)$$

which we can visualize as follows:



Of course, it is possible for $\{X_t\}_{t\geq 0}$ to pass (n,n) without ever hitting it, so the probability should be less than 1. In fact, we should expect that the probability goes to 0 as $n \to \infty$, since the distribution of X_t becomes more diffuse as $t \to \infty$.

To compute this, note that the process $\{N_t + M_t\}_{t\geq 0}$ is a Poisson process (of rate 2λ). Moreover,

$$\begin{aligned}&\{\{X_t\}_{t\geq 0} \text{ ever visits the state } (n,n)\}\\ &= \{\text{When } N_t + M_t = 2n, \text{ we have } N_t = n\}.\end{aligned}$$

Thus, by thinning:

$$\begin{aligned}&\mathbb{P}\left(\{X_t\}_{t\geq 0} \text{ ever visits the state } (n,n)\right)\\ &= \mathbb{P}\left(\text{When } N_t + M_t = 2n, \text{ we have } N_t = n\right)\\ &= \mathbb{P}\left(N_t = n \,|\, N_t + M_t = 2n\right)\\ &= \mathbb{P}(\text{Binom}(2n, 1/2) = n)\\ &= \binom{2n}{n} 2^{-n}.\end{aligned}$$

By Stirling's formula $k! \approx (k/e)^k \sqrt{2\pi k}$, we have

$$\mathbb{P}\left(\{X_t\}_{t\geq 0} \text{ ever visits the state } (n,n)\right) \approx \frac{1}{\sqrt{\pi n}}$$

for large n.

Note that this does not depend on the rate λ. This makes sense, since the probability does not depend on the speed of time, but rather on the particular jumps.

3.7 Variations on Poisson processes

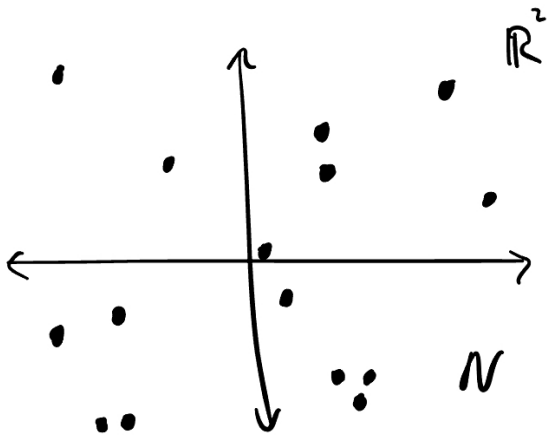
There are many ways to generalize Poisson processes, and we will discuss a few of them now.

DEF. A *spatial Poisson point process* in $\mathbb{R}^2$ with rate $\lambda > 0$ is a collection of points $N := \{S_1, S_2, \ldots\} \subseteq \mathbb{R}^2$ satisfying the following, where $N(A)$ denotes the number of points of N contained in a region $A \subseteq \mathbb{R}^2$:

(1) $N(A)$ has distribution $\text{Poi}(\lambda \cdot \text{area}(A))$ for all $A \subseteq \mathbb{R}^2$, and

(2) $N(A)$ and $N(B)$ are independent if $A, B \subseteq \mathbb{R}^2$ are disjoint.

We now make some remarks about this definition. First, notice that the points $S_1, S_2, \ldots$ correspond, in some sense, to the arrivals of N, although not exactly: In the spatial setting, these points do not have an intrinisic order. Instead, we think of them as a random unordered set of points, as follows:



Second, notice that there is a natural generalization to arbitrary dimension: A spatial Poisson point process in $\mathbb{R}^1$ involves length instead of area (hence it corresponds to our original definition of Poisson process), and a spatial Poisson point process in $\mathbb{R}^3$ involves volume instead of area.

Many aspects of Poisson processes are also true for spatial Poisson point processes, for instance:

THM. *If N is a spatial Poisson process in $\mathbb{R}^2$ with rate $\lambda > 0$, then the conditional distribution of $\{S_1, \ldots, S_{N_t}\}$ given $\{N(A) = k\}$ is equal to a set of k IID uniformly-distributed random variables in $A \subseteq \mathbb{R}^2$.*

Next, we return to the temporal setting.

DEF. An *inhomogeneous Poisson process with intensity function* $\lambda : [0, \infty)] \to [0, \infty)$ is a counting process $\{N_t\}_{t \geq 0}$ satisfying the following:

(1) $N_t - N_s$ has distribution $\text{Poi}(\int_s^t \lambda(u)\, du)$ for all $0 \leq s \leq t$, and

(2) $N_t - s$ is independent of N_s for all $0 \leq s \leq t$.

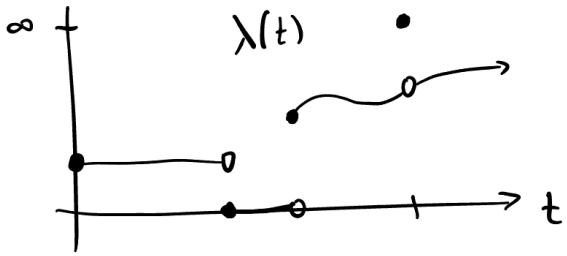
Note that an inhomogeneous Poisson process has increments that are independent but not stationary. Also, if λ is a constant function, then we recover our original definition of (homogeneous) Poisson process.

Inhomogeneous Poisson processes have some interesting behavior that is not seen in the case of (homogeneous) Poisson processes. For example, if the intensity function satisfies $\int_0^\infty \lambda(u)\, du < \infty$, then the total number of arrivals

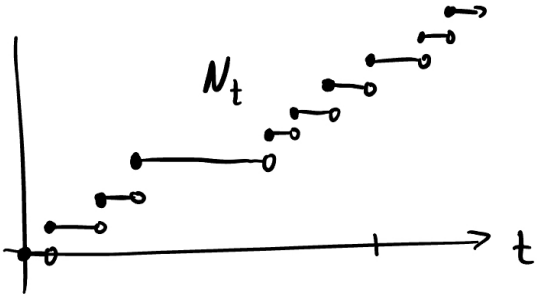
$$\lim_{t \to \infty} N_t$$

exists and is finite, and and it has distribution $\text{Poi}(\int_0^\infty \lambda(u)\, du)$. Moreover, one can make an inhomogeneous Poisson process exhibit deterministic behaviors: if $\lambda(t) = 0$ for some $a \leq t \leq b$, then $\{N_t\}_{t \geq 0}$ has no arrivals

during $[a, b]$ with probability 1, and if $\lambda(t) = \infty$ then $\{N_t\}_{t \geq 0}$ has an arrival exactly at time t with probability 1. For example, an inhomogeneous Poisson process with intensity function



has sample paths looking like



We can also do some exact calculations with inhomogeneous Poisson processes:

EX. Suppose that $\{N_t\}_{t \geq 0}$ is an inhomogeneous Poisson process with intensity function $\lambda : [0, \infty) \to [0, \infty)$. We may easily compute the distribution of the first arrival time S_1, since

$$\begin{aligned} \mathbb{P}(S_1 > t) &= \mathbb{P}(N_t = 0) \\ &= \mathbb{P}\left(\text{Poi}\left(\int_0^t \lambda(u)\, \mathrm{d}u \right) = 0 \right) \\ &= \exp\left(-\int_0^t \lambda(u)\, \mathrm{d}u \right) \end{aligned}$$

for all $t \geq 0$.

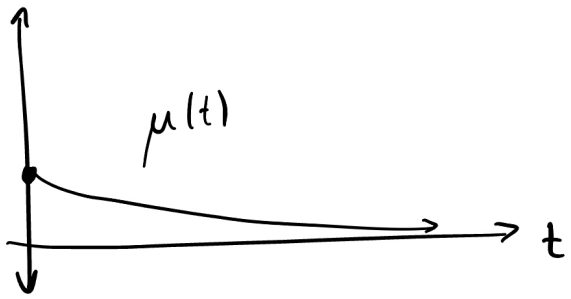
Lastly, we give an important model of "self-exciting" counting processes which are important for applications in neuroscience, finance, seismology, etc. The starting observation is that we can in fact define inhomogeneous Poisson processes where the intensity function is itself a stochastic process, which we denote $\{\Lambda_t\}_{t \geq 0}$ and refer to as the *intensity process*.

DEF. A *Hawkes process with base rate* $\lambda \geq 0$ *and excitation function* $\mu : [0, \infty) \to [0, \infty)$ is an inhomogeneous Poisson process $\{N_t\}_{t \geq 0}$ with arrivals $S_1, S_2, \ldots$ whose intensity process $\{\Lambda_t\}_{t \geq 0}$ is given by

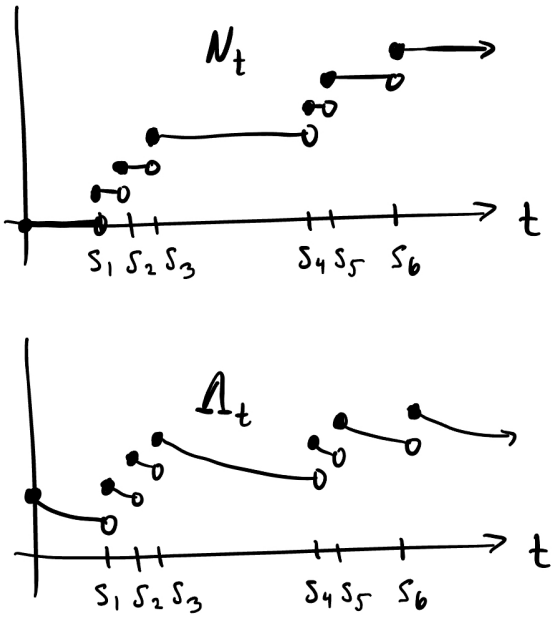
$$\Lambda_t = \begin{cases} \mu(t) + \sum_{k=1}^{N_t} \mu(t - S_k) & \text{if } \lambda = 0 \\ \lambda \quad + \sum_{k=1}^{N_t} \mu(t - S_k) & \text{if } \lambda > 0 \end{cases}$$

for all $t \geq 0$. Note that we need to distinguish the $\lambda = 0$ case, since otherwise there are no arrivals at all.

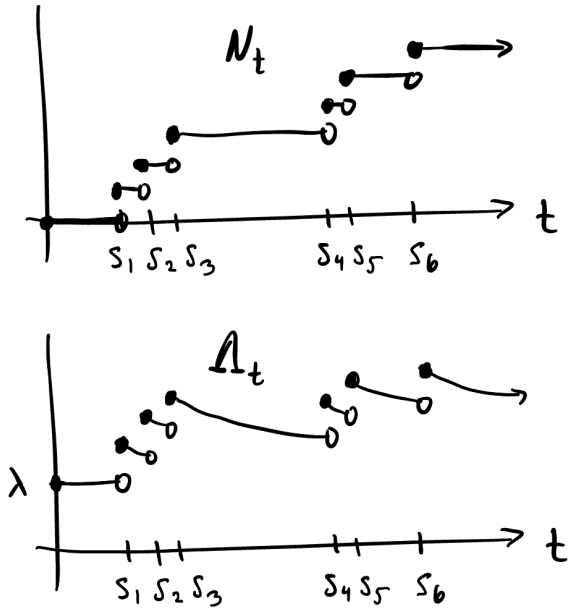
There is a classic picture to illustrate this. First, we consider an excitation function which, say, looks like this:



Then, we consider simultaneously plotting the counting process and the intensity process. For $\lambda = 0$ we get something like this:



For $\lambda > 0$, we get something like this:



If you think the definition of a Hawkes process seems somewhat circular (the definition of Λ depends on N, and the definition of N depends on Λ), you are absolutely right! This is exactly what leads to the "clustering" structure of the arrivals.

One fundamental problem is to calculate the total number of arrivals in a Hawkes with no base rate, or to check when the total number of arrivals is infinite. Of course, we should assume $\int_0^\infty \mu(u)\,\mathrm{d}u < \infty$ since otherwise there will obviously be infinitely many arrivals. While Hawkes processes seem quite complicated, it turns out that the following structure provides a way to analyze them:

- First, the initial copy of μ "gives rise to" some number of arrivals; the total number of such arrivals has a $\mathrm{Poi}(\int_0^\infty \mu(u)\,\mathrm{d}u)$ distribution.
- Second, each of these arrivals increases the intensity by a copy of μ and "gives rise to" some further arrivals; each arrival yields a number of further arrivals which has a $\mathrm{Poi}(\int_0^\infty \mu(u)\,\mathrm{d}u)$ distribution.
- Continue in this way.

This insight shows that Hawkes processes have a branching process embedded within them!

It is useful to give some notation to this. First, let $\{N_t^{(0)}\}_{t\geq 0}$ denote the Poisson process of arrivals due to the initial copy of μ. Then, let $\{N_t^{(1)}\}_{t\geq 0}$ denote the Poisson process of arrivals due to the induced copies of μ. Then, continue in this way. By construction, we have

$$N_t = \sum_{n=0}^{\infty} N_t^{(n)}$$

which can be seen roughly as a form of superposition, although not literally so since these (inhomogeneous) Poisson process are not independent. Also, as we noted above, the stochastic process

$$\left\{\lim_{t\to\infty} N_t^{(n)}\right\}_{n\geq 0}$$

is exactly a branching process whose offspring distribution is Poisson with parameter $\int_0^\infty \mu(u)\,\mathrm{d}u$.

This structure allows us to answer some questions about Hawkes processes in terms of related results for branching processes, like the following:

EX. Suppose that a Hawkes process $\{N_t\}_{t\geq 0}$ has no base rate and with excitation function $\mu : [0,\infty) \to [0,\infty)$. When do we have

$$\mathbb{P}\left(\lim_{t\to\infty} N_t < \infty\right) = 1$$

in terms of μ?

Note that $\{\lim_{t\to\infty} N_t < \infty\}$ is exactly the same event as the extinction of the embedded branching process. So, by our early theory of branching processes, we know that extinction occurs with probability 1 only in the sub-critical and critical cases. In other words, we have

$$\mathbb{P}\left(\lim_{t\to\infty} N_t < \infty\right) = 1$$

if and only if $\int_0^\infty \mu(u)\,\mathrm{d}u \leq 1$.

On HW8 you will see that some other events related to Hawkes processes may be analyzed through their embedded branching process.

3.8 Renewal theory

There is a more general class of counting processes, which includes Poisson processes, for which many results are known at least asymptotically.

DEF. A *renewal process* is a counting process $\{N_t\}_{t\geq 0}$ whose interarrival times $X_1, X_2, \ldots$ are IID non-negative continuous random variables.

If X_i are Exp(λ) for some $\lambda > 0$, then the resulting renewal process is a Poisson process with rate $\lambda > 0$. However, for general X_i, a renewal process is not a Markov process. For example, is X_i are uniformly-distributed on $[0,1]$, then the arrival times are certainly not memoryless.

There are a few results about renewal processes that we will discuss, but our treatment is very brief and there is a lot we will not study. Both results will be related to the function $m : [0,\infty) \to [0,\infty)$ defined via $m(t) := \mathbb{E}[N_t]$ called the *renewal function* which counts the expected number of arrivals up to time $t \geq 0$.

THM (renewal equation)**.** *For a renewal process whose interarrival times have density f and CDF F, the renewal function satisfies*

$$m(t) = F(t) + \int_0^t m(t-x)f(x)\,\mathrm{d}x$$

for all $t \geq 0$.

The renewal equation looks a bit like a recursion for the function m, and indeed it is proven via a form of first-transition analysis like we saw during our study of discrete-time MCs; you will do this in detail in HW8.

Another result (which we will not prove) is the following:

THM (renewal theorem)**.** *For a renewal process whose interarrival times have mean $\mu > 0$, we have*

$$\lim_{t\to\infty} \frac{m(t)}{t} = \frac{1}{\mu}.$$

It is somewhat intuitive that the renewal theorem states that the expected number of arrivals per time is inversely proportional to the expected length of time between arrivals. Also note that if $\{N_t\}_{t\geq 0}$ is in fact a Poisson process of rate $\lambda > 0$, then we have

$$\frac{m(t)}{t} = \frac{1}{\lambda}.$$

for all $t \geq 0$. In other words, the Poisson process has exact number of expected arrivals per time, and a general renewal process has an asymptotically exact number of expected arrivals per time.

Lastly, here is another application of renewal theory:

EX. Let $X_1, X_2, \ldots$ be IID random variables which are uniformly distributed on $[0, 1]$, and set

$$T := \min\{n \geq 0 : X_1 + \cdots + X_n > 1\}$$

which is the number of uniforms that we sample before their running sum exceeds 1. What is $\mathbb{E}[T]$?

To do this, write $\{N_t\}_{t\geq 0}$ for the renewal process corresponding to the interarrival times $X_1, X_2, \ldots$ and note that $T = N_1 + 1$. Thus, $\mathbb{E}[T] = \mathbb{E}[N_1] + 1 = m(1) + 1$, so it suffices to compute the renewal function m. In order to find the renewal function m, note

$$\begin{aligned} m(t) &= t + \int_0^t m(t-x)\,\mathrm{d}x && \text{(renewal equation)} \\ &= t + \int_0^t m(u)\,\mathrm{d}u && \text{(change of variables)} \end{aligned}$$

for $0 \leq t \leq 1$. So, by differentiating the above, we see that m satisfies the ODE

$$\begin{cases} m'(t) = 1 + m(t) \text{ for } 0 \leq t \leq 1 \\ m(0) = 0 \end{cases}$$

We won't get into the details about differential equations, but it turns out that this ODE has a unique solution given by $m(t) = e^t - 1$. Therefore, we have $\mathbb{E}[T] = e$.

4 Continuous-time Markov chains

4.1 Basic definitions and examples

DEF. A *continuous-time Markov chain* is a stochastic process $\{X_t\}_{t\geq 0}$ such that the conditional distribution of X_t given $\{X_r = i, X_s = j\}$ equals the conditional distribution of X_t given $\{X_s = j\}$ for all times $0 \leq r \leq s \leq t$ and all states i, j.

In the remainder of the course, we will usually say *discrete-time Markov chain (DTMC)* and *continuous-time Markov chain (CTMC)* to distinguish between these notions.

We need to figure out the correct analog, continuous-time setting, of transition matrices from the discrete-time setting. Previously we had a single transition matrix P representing the one-step transition probabilties, and it followed that P^n represented the n-step transition probabilities. Unfortunately, this is not possible in continuous time, since there is no smallest increment of time. Instead, we have to keep track of the transition matrices over all time:

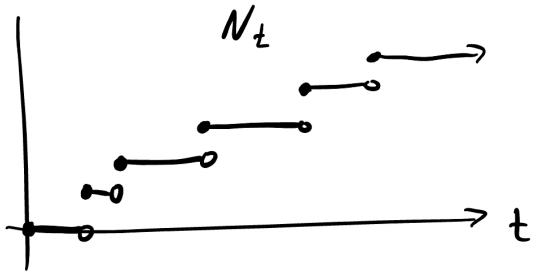
DEF. The *transition function* $\{P(t)\}_{t\geq 0}$ of a CTMC $\{X_t\}_{t\geq 0}$ is the family of transition matrices defined via

$$P_{ij}(t) = \mathbb{P}(X_t = j \,|\, X_0 = i)$$

for all $t \geq 0$ and all states i, j.

While we will later see many examples of CTMCs, it is instructive to consider just two examples for now:

EX (Poisson process)**.** If $\{N_t\}_{t\geq 0}$ is a Poisson process of rate $\lambda > 0$, then we saw in the last part of the class that has the Markov property hence it is a CTMC. Remember that the sample paths look like this:



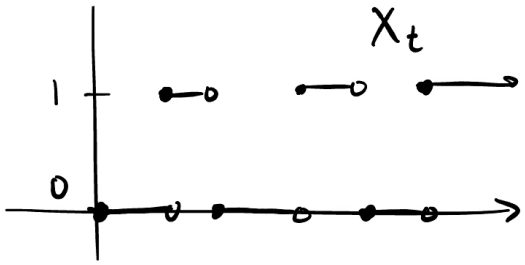
In fact, we can directly compute that its transition function has entries equal to the values of a Poisson PMF with parameter λt. That is, for all $t \geq 0$ we have:

$$P(t) = e^{-\lambda t} \begin{pmatrix} 1 & \lambda t & \frac{(\lambda t)^2}{2} & \frac{(\lambda t)^3}{6} & \cdots \\ 0 & 1 & \lambda t & \frac{(\lambda t)^2}{2} & \cdots \\ 0 & 0 & 1 & \lambda t & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

EX (two-state chain)**.** For $\lambda, \mu > 0$, consider a stochastic process $\{X_t\}_{t \geq 0}$ on the state space $\{0, 1\}$ given as follows:

- When in state 0, hold for Exp(λ) time and then move to state 1.
- When in state 1, hold for Exp(μ) time and then move to state 0.

Here, we assume that all of the holding times are independent of each other. We can visualize the sample paths of $\{X_t\}_{t \geq 0}$ as follows:



By the memorylessness of the exponential distribution, this process is a CTMC. We will later see that the transition function for this CTMC is

$$P(t) = \frac{1}{\lambda + \mu} \begin{pmatrix} \mu + \lambda e^{-(\lambda+\mu)t} & \lambda - \lambda e^{-(\lambda+\mu)t} \\ \mu - \mu e^{-(\lambda+\mu)t} & \lambda + \mu e^{-(\lambda+\mu)t} \end{pmatrix}$$

for all $t \geq 0$, but the derivation is a bit complicated. (Recall that it was also complicated the case for the two-state DTMC).

Any transition function $\{P(t)\}_{t \geq 0}$ must satisfy the property

$$P(s + t) = P(s)P(t)$$

for all $0 \leq s \leq t$, which is called the *Chapman-Kolmogorov equation*. Note that this has an simple meaning if we understand $P(t)$ to be the operation of "advancing time by t units": Any CTMC has the property that advancing time by t units and then advancing time by s units is equivalent to advancing time by $t+s$ units. Although this seems tautological, it does not need to be true for processes without the Markov property.

DEF. If $\{X_t\}_{t\geq 0}$ is a CTMC, its *embedded chain* is the DTMC consisting of the sequence of states it visits.

Notice that the transition matrix P of an embedded chain must always satisfy $P_{ii} = 0$ for all states i. This is because self-transitions in the embedded chain would be invisible from the CTMC.

EX (Poisson process)**.** For a Poisson process of any rate, the embedded chain has transition matrix

$$\tilde{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

since it only moves by jumping up.

EX (two-state chain)**.** For the two-state chain described above, its embedded chain has transition matrix

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

In other words, the embedded chain is exactly the alternating chain.

In both of the preceding examples, the embedded chain was deterministic. In general this need not be true. Next we will describe a three-state CTMC whose embedded chain is more interesting.

EX (three-state chain)**.** Take six arbitrary positive numbers Q_{12},Q_{13},Q_{21}, Q_{23},Q_{31}, and Q_{32}. Then we set

$$\begin{aligned} Q_1 &= Q_{12} + Q_{13} \\ Q_2 &= Q_{21} + Q_{23} \\ Q_3 &= Q_{31} + Q_{32}. \end{aligned}$$

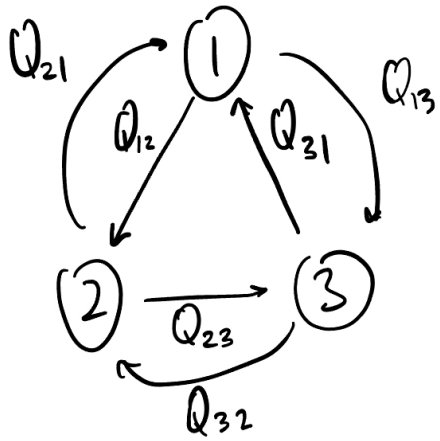
Then, we define a CTMC on the state space $\{1, 2, 3\}$ as follows:

- When in state i, hold for $\text{Exp}(Q_i)$ time.
- Then, move to state $j \neq i$ with probability proportional to Q_{ij}.

As we can see, this CTMC has an embedded chain with transition matrix

$$\begin{pmatrix} 0 & Q_{12}/Q_1 & Q_{13}/Q_1 \\ Q_{21}/Q_2 & 0 & Q_{23}/Q_2 \\ Q_{31}/Q_3 & Q_{32}/Q_3 & 0 \end{pmatrix}$$

Then repeat this process. Visually speaking, we can think of representing this process with the following graph:



Here, the arrows have transition rates rather than transition probabilities like we had in the case of DTMCs.

By properties of exponential random variables, the three-state chain above can also be described in a different way. When in state i, suppose that we independently sample T_{ij} from $\text{Exp}(Q_{ij})$ for $j \neq i$, and then we transition to state

$$j := \arg\min\{T_{ij} : j \neq i\}$$

after holding for time

$$T_i := \min\{T_{ij} : j \neq i\}.$$

In other words, there are many independent "exponential clocks" in a race, and the winner of the race determines the location of the next state.

As we will later see, it turns out that every CTMC can be (in some sense) understood in this form.

4.2 Infinitesimal generator

In the case of DTMCs, the transition matrix was a convenient object to do calculations with. In the case of CTMCs, we can do similar calculations with the transition function, but it already seems much more complicated. In this part we will see that even a CTMC admits a single matrix that we can do calculations with.

DEF. A *generator matrix* is a (possibly infinite) matrix satisfying

- $Q_{ij} \geq 0$ for all states $i \neq j$, and
- $\sum_j Q_{ij} = 0$ for all states i.

The second condition requires that the sum of the entries of Q along each row must equal 0. In particular, these conditions together imply that the diagonal entries of Q satisfy $Q_{ii} \leq 0$.

Let us show that from each generator matrix we can define a CTMC. Indeed, given a generator matrix Q, we let a stochastic process $\{X_t\}_{t \geq 0}$ evolve as follows:

- When in state i, hold for time $\text{Exp}(-Q_{ii})$ time.
- Then, move to state $j \neq i$ according to the transition matrix $\tilde{P}$ given by

$$\tilde{P}_{ij} = \frac{Q_{ij}}{-Q_{ii}}$$

for all states i, j.

As we already saw in the example of the three-state chain, this perspective allows us to represent each CTMC as a combination of a (inhomogeneous) PP and a DTMC.

Next we point out that a transition function $\{P(t)\}_{t\geq 0}$ and a generator matrix Q encode exactly the same information. To show this, we note that there are several equivalent ways to describe the relationship between them.

(1) Derivative: If $\{P(t)\}_{t\geq 0}$ is given, then Q may be computed via

$$Q_{ij} = \lim_{t\to 0} \frac{P_{ij}(t)}{t} \text{ for } i \neq j,$$

and $Q_{ii} = -\sum_{j\neq i} Q_{ij}$.

(2) Differential equations: $\{P(t)\}_{t\geq 0}$ and Q together solve the ODEs

$$\begin{cases} P'(t) &= QP(t) \text{ for all } t \geq 0 \\ P(0) &= I \end{cases}$$

and

$$\begin{cases} P'(t) &= P(t)Q \text{ for all } t \geq 0 \\ P(0) &= I, \end{cases}$$

called the *Kolmogorov forward equation* and the *Kolmogorov backward equation*, respectively.

(3) Matrix exponential: If Q is given, then $\{P(t)\}_{t\geq 0}$ may be computed via

$$P(t) = e^{tQ},$$

for all $t \geq 0$. Here, the matrix exponential is defined as

$$e^Z = I + Z + \frac{1}{2}Z^2 + \frac{1}{6}Z^3 + \cdots$$

which generalizes the usual formula for the power series of e^z for a real number z.

EX (Poisson process)**.** The infinitesimal generator Q of a Poisson process of rate $\lambda > 0$ is

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ 0 & 0 & 0 & -\lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

EX (two-state chain)**.** The infinitesimal generator Q of the two-state chain above is

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

4.3 Limiting and stationary distributions

DEF. A probability distribution π is called a *limiting distribution for* $\{X_t\}_{t\geq 0}$ if we have

$$\lim_{t\to\infty} P_{ij}(t) = \pi_j$$

for all states j.

DEF. A probability distribution π is called a *stationary distribution for* $\{X_t\}_{t\geq 0}$ if we have $\pi = \pi P(t)$ for all $t \geq 0$.

LEM. *A probability distribution π is a stationary distribution for $\{X_t\}_{t\geq 0}$ if and only if $\pi Q = 0$.*

PF: First, suppose that π satisfies $\pi = \pi P(t)$ for all $t \geq 0$. Then:

$$\pi Q = \pi \lim_{t\to 0} \frac{P(t) - I}{t} = \lim_{t\to 0} \frac{\pi P(t) - \pi}{t} = \lim_{t\to 0} \frac{0}{t} = 0.$$

Second, suppose that π satisfies $\pi Q = 0$. Then use the Kolmogorov backward equation to get:

$$\pi P'(t) = \pi Q P(t) = 0 P(t) = 0$$

for all $t \geq 0$. This means $\pi P(t)$ is a constant. Since for $t = 0$ we have $\pi P(0) = \pi I = \pi$, this implies $\pi P(t) = \pi$ for all $t \geq 0$.

As before, we want to know when a stationary distribution is a limiting distribution. We can do this under some assumptions, but we need to introuce the relevant notions for CTMCs.

First, irreducibility:

DEF. A CTMC with transition function $\{P(t)\}_{t\geq 0}$ is *irreducible* if for all states $i \neq j$ there exists some $t > 0$ such that $P_{ij}(t) > 0$.

LEM. *A CTMC is irreducible if and only if its embedded DTMC is irreducible.*

Second, recurrence/transience:

DEF. Let $\{X_t\}_{t\geq 0}$ denote a CTMC. For state i, write $E_i := \min\{t \geq 0 : X_t \neq i\}$ for the first exit time from i, and $T_i := \min\{t \geq E_i : X_t = i\}$ for the first time after E_i returning to i. We say that state i is *transient* if

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

and *recurrent* if

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

Furthermore, a recurrent state i is called *positive-recurrent* if

$$\mathbb{E}[T_i \,|\, X_0 = i] < \infty$$

and called *null-recurrent* if

$$\mathbb{E}[T_i \,|\, X_0 = i] = \infty.$$

LEM. *In an irreducible CTMC, a state i transient, positive-recurrent, or null-recurrent according to whether in the embedded DTMC state i transient, positive-recurrent, or null-recurrent.*

Lastly, ergodicity:

DEF. A CTMC is *ergodic* if it is irreducible and positive-recurrent.

LEM. *A CTMC is ergodic if and only if its embedded DTMC is irreducible and positive-recurrent.*

Note that aperiodicity is not required. This is because, in some sense, all states are automatically aperiodic in a CTMC. Stated precisely, the reason is the following (which you will prove carefully on HW9):

LEM. *If states i, j and a transition function $\{P(t)\}_{t\geq 0}$ satisfy $P_{ij}(t) > 0$ for some $t > 0$, then they satisfy $P_{ij}(t) > 0$ for all $t \geq 0$.*

Now the main result:

THM (fundamental theorem of ergodic CTMCs)**.** *If $\{X_t\}_{t\geq 0}$ is an ergodic CTMC, then it has a unique stationary distribution π, and π is a limiting distribution.*

Now we'll do some examples of this.

EX (Jukes-Cantor model)**.** This is a model for mutations appearing in the nucleotide sequence of DNA. That is, we assume that each nucleotide, taking values in $\{A, G, C, T\}$, mutates at constant rate λ and that each nucleotide is equally likely to mutate into any other. So, the infinitesimal generator matrix is

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

for $\lambda > 0$. We can easily see that the transition matrix of the embedded DTMC is

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

It is clear that $\tilde{P}^2$ has all positive entries, hence this transition matrix is ergodic. Therefore, the fundamental theorem applies to Q. We may guess that the uniform distribution $\pi = (1/4, 1/4, 1/4, 1/4)$ is stationary for Q, and indeed it satisfies $\pi Q = 0$. Therefore, π is the limiting distribution.

EX (reflecting biased random walk)**.** Fix $\lambda, \mu > 0$, and consider

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \cdots \\ 0 & \mu & -(\mu+\lambda) & \lambda & \cdots \\ 0 & 0 & \mu & -(\mu+\lambda) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The embedded DTMC has transition matrix

$$\tilde{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 1-p & 0 & p & 0 & \cdots \\ 0 & 1-p & 0 & p & \cdots \\ 0 & 0 & 1-p & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where $p = \lambda/(\lambda+\mu)$. Note that this is just the transition matrix for a biased random walk, except in the state 0 where we always move to state 1. We easily see that $\tilde{P}$ is irreducible. Also, we know from earlier in the class that this DTMC is positive-recurrent if and only if $p < 1$ which is equivalent to $\lambda < \mu$. Therefore, this CTMC is ergodic if and only if $\lambda < \mu$.

EX (Ehrenfest gas model)**.** For fixed $M > 0$, consider the following continuous-time version of the Ehrenfest gas model: We have M particles distributed into two chambers, and each particle moves to the opposite chamber according to a Poisson process of rate λ, where all M Poisson processes are independent of each other. This is a CTMC with infinitesimal generator given by

$$Q_{ij} = \begin{cases} i\lambda & \text{if } j = i-1 \\ -N\lambda & \text{if } j = i \\ (N-i)\lambda & \text{if } j = i+1 \end{cases}$$

Alternatively, we can visualize Q as follows.

$$Q = \begin{pmatrix} -M\lambda & M\lambda & 0 & \cdots & 0 & 0 & 0 \\ \lambda & -M\lambda & (M-1)\lambda & \cdots & 0 & 0 & 0 \\ 0 & 2\mu & -M\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2M\lambda & 2\lambda & 0 \\ 0 & 0 & 0 & \cdots & (M-1)\lambda & -M\lambda & \lambda \\ 0 & 0 & 0 & \cdots & 0 & M\lambda & -M\mu \end{pmatrix}.$$

We want to show that this CTMC has a limiting distribution, and identify what it is.

First, that the embedded DTMC is exactly the discrete-time Ehrenfest gas model on M particles. Since we already showed this DTMC is irreducible and positive-recurrent (but not aperiodic), it follows that the CTMC is ergodic. Thus, the fundamental theorem of ergodic CTMC applies, and we know that there is a unique stationary distribution and that it is also a limiting distribution.

Because of the analogy with the discrete-time model, we might guess that a stationary distribution is $\pi = \text{Binom}(M, 1/2)$. Indeed, it is straightforward to check that this is indeed stationary. So, the fundamental theorem implies that it is the unique stationary distribution and also a limiting distribution!

Interestingly, in continuous-time there is another (easier) we can identify the limiting distribution. Let $\{\tilde{X}_t^1\}_{t\geq 0}, \ldots, \{\tilde{X}_t^M\}_{t\geq 0}$ be independent copies of the CMTC on $\{0,1\}$ with generator matrix

$$\tilde{Q} = \begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix},$$

and observe that $\{X_t\}_{t\geq 0}$ is exactly

$$X_t = \sum_{i=1}^{M} \tilde{X}_t^i.$$

We know that the limiting distribution of $\{\tilde{X}_t^i\}_{t\geq 0}$ is just $\tilde{\pi} = (1/2, 1/2)$. Therefore, the limiting distribution of $\{X_t\}_{t\geq 0}$ is just a sum of M independent copies of $\tilde{\pi}$. This is exactly the Binomial distribution $\text{Binom}(M, 1/2)$.

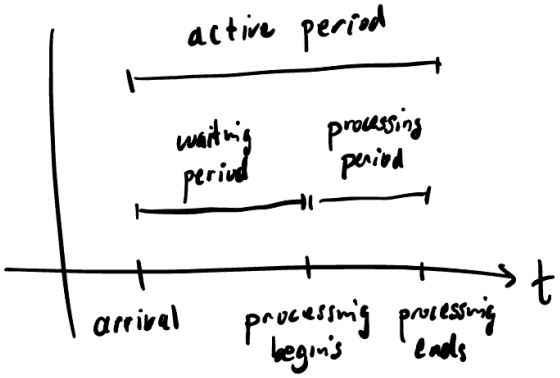
In the preceding example, we saw that systems of non-interacting particles are easy to analyze!

4.4 Queueing theory

A classical application of al of the ideas we have learned so far is to queueing theory, which the mathematical study of processed completed in sequence. When we develop a mathematical model for queuing, we need to specify three things:

- How do units arrive in the system?
- How long is the processing of each unit?
- When do arrived units begin being processed?

We think of the following figure for this:



A unit is said to be *active* if it is either waiting or being processed.

There is a standard notation for a nice class of queuing models that we can analyze:

interarrival model/processing model/number of processors

(This is called *Kendall's notation*.) The first two arguments (the times) can be either of: M for "Markov" or "memoryless" to mean that times are IID exponential, with some specific rates, or they can be G for "general" to mean that the times of are IID from some specified continuous distribution. The number of processors must be either a positive integer, of infinity.

We will always write $\{X_t\}_{t\geq 0}$ for the number of active units in a given queue. Now we can discuss a few models, and see how the results of the class allow us to determine lots of things about these queues!

We begin with the M/M models since they are CTMCs and hence we can do many explicit calculations. Note that when the interarrival model is M, then the process of arrivals is just a PP $\{N_t\}_{t\geq 0}$.

EX $(M/M/1)$**.** Suppose the arrival process has rate $\lambda > 0$ and the processing times are exponential with rate $\mu > 0$, and let's find the infinitesimal generator Q of this CTMC. If $X_0 = 0$, then the only thing that can happen is the arrival of a new unit, which occurs at rate λ. If $X_0 = i$ for $i \geq 1$, then the only things that can happen are the arrival of a new unit (which occurs at rate λ) or the completion of an existing processing (which occurs at rate μ, since there is only one processing at a time); in particular, note that this does not depend on i. Thus, the generator is:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \cdots \\ 0 & \mu & -(\mu+\lambda) & \lambda & \cdots \\ 0 & 0 & \mu & -(\mu+\lambda) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

We recognize that this is exactly the generator of the reflecting biased random walk, which we already studied.

EX $(M/M/\infty)$**.** In this queue (since there are infinitely many processors), every arrival begins processing immediately. If the arrival process has rate $\lambda > 0$ and the processing times are exponential with rate $\mu > 0$, then the generator of $\{X_t\}_{t\geq 0}$ is just:

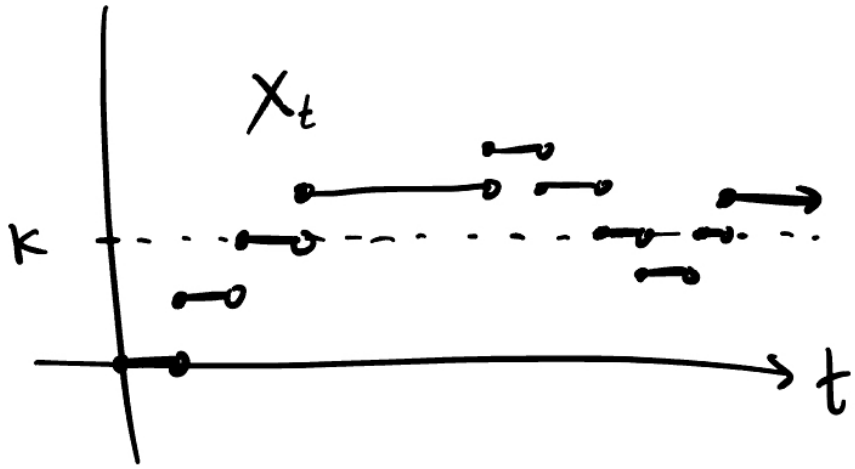
$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \cdots \\ 0 & 2\mu & -(2\mu+\lambda) & \lambda & \cdots \\ 0 & 0 & 3\mu & -(3\mu+\lambda) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

In some sense, this means the X_t has a downward bias, and that the bias gets stronger as when X_t is larger. This can be seen in the transition matrix of the embedded DTMC, in which we take $\lambda = \mu$ for the sake of simplicity:

$$\tilde{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 1/2 & 0 & 1/2 & 0 & \cdots \\ 0 & 2/3 & 0 & 1/3 & \cdots \\ 0 & 0 & 3/4 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Indeed, one can show that $\{X_t\}_{t\geq 0}$ is always positive recurrent, although we won't prove it in this class.

EX $(M/M/k)$**.** By combining the two previous examples, we can analyze many properties of $\{X_t\}_{t\geq 0}$ from a $M/M/k$ queue. First of all, notice that $\{X_t\}_{t\geq 0}$ "looks like" $M/M/\infty$ when $X_t \leq k$ and that $\{X_t\}_{t\geq 0}$ "looks like" $M/M/1$ when $X_t \geq k$ (but with μ replaced with $\tilde{\mu} = k\mu$).



Thus, one can analyze this by comparison with the previous examples; for instance, in HW9 you will show that $M/M/k$ is recurrent if $k\mu > \lambda$ and transient if $k\mu < \lambda$.

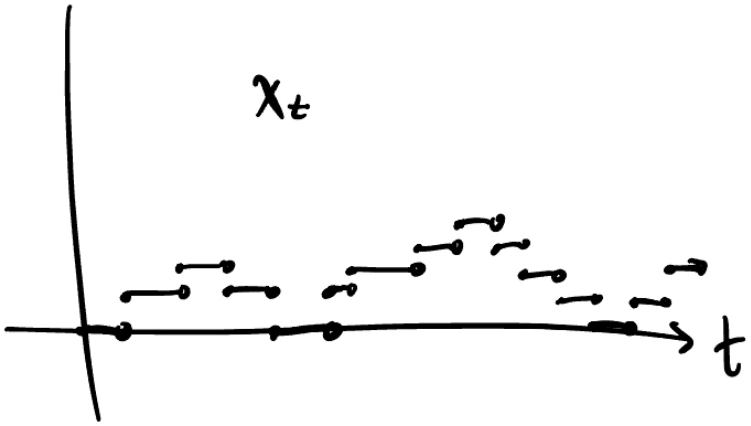
Next we consider the M/G models. Although they are not CTMCs, there are many things we can study.

EX $(M/G/\infty)$**.** Suppose arrivals occur at rate $\lambda > 0$, that processing times are IID continuous random variables $W_1, W_2, \ldots$, and that there are infinitely many processors. (So, processing beings immediately for each unit, hence there are no waiting times.) During our study of Poisson processes, we were able to analyze the limiting distribution of $\{X_t\}_{t\geq 0}$ directly; you showed in the HW that

$$\lim_{t\to\infty} \mathbb{P}(X_t = k) = \mathbb{P}\left(\text{Poi}(\lambda \mathbb{E}[W_1])\right) = k)$$

Of course, one can prove this directly from the fundamental theorem of ergodic CTMCs in the case of $M/M/\infty$. But it's interesting that we can even do this in the $M/G/\infty$ case too!

EX $(M/G/1)$**.** Suppose arrivals occur at rate $\lambda > 0$, that processing times are IID continuous random variables $W_1, W_2, \ldots$, and that there is only one processor. Although this process it not a CTMC, it is still of interest to ask whether the state 0 is recurrent: That is, we want to know when the sample paths look like this:



In engineering terms, recurrence simply means that the queue will always become empty eventually.

The difficulty is that each arriving unit causes future units to wait, and each of the subsequently waiting units causes future units to wait, and so on. Remarkably, this is exactly an embedded branching process! In fact, the "height" of each "excursion from state 0" is exactly the total number of individuals in such a branching process. In particular, the lengths time between visits to state 0 are all finite if the embedded branching process is critical or sub-critical, and they have a positive probability of being infinite if the embedded branching process is super-critical. The next goal is to make this argument precise.

For convenience, suppose W_k has density f, CDF F, and mean $1/\mu$. To get started, let us write P_k for the time that arrival unit k begins processing, and W_k for the waiting time of arrived unit k. In terms of the PP of arrivals $\{N_t\}_{t\geq 0}$ and its arrival times $S_1, S_2, \ldots$, we have $P_1 = S_1$, and

$$P_2 = \begin{cases} S_1 + W_1 & \text{if } N_{P_1+W_1} - N_{P_1} \geq 1 \\ S_1 + W_1 + \mathrm{Exp}(\lambda) & \text{if } N_{P_1+W_1} - N_{P_1} = 0. \end{cases}$$

One can write down some explicit formulas for P_k, but we don't need to do that for our current problem. The only thing we need to observe is that the intervals

$$[P_1, P_1 + W_1), [P_2, P_2 + W_2), [P_3, P_3 + W_3), \ldots$$

are disjoint (because there is only one processor).

Now we can describe the branching process. We set $Z_0 = 1$ as always, and we think of this as the first arrived unit. Then we set

$$Z_{n+1} = \sum_{k=1}^{Z_n} (N_{P_k+W_k} - N_{P_k})$$

for $n \geq 0$, and we think of this as the number of arrived units whose waiting is due to the wating periods of units in Z_n. By the disjointness of the intervals, this is exactly a branching process whose offspring distribution has mean

$$\begin{aligned} &\mathbb{E}[N_{P_k+W_k} - N_{W_k}] \\ &= \int_0^\infty \mathbb{E}[N_{P_k+W_k} - N_{P_k} \,|\, W_k = w]] f(w)\,\mathrm{d}w && \text{(law of total probability)} \\ &= \int_0^\infty \mathbb{E}[N_w - N_0] f(w)\,\mathrm{d}w && \text{(stationary increments)} \end{aligned}$$

$$= \int_0^\infty \lambda w \, f(w) \, \mathrm{d}w \qquad \text{(Poisson increments)}$$
$$= \lambda \mathbb{E}[W_k]$$
$$= \frac{\lambda}{\mu}$$

Thus, the branching process is

$$\begin{cases} \text{sub-critical if} & \lambda < \mu \\ \text{critical if} & \lambda = \mu \\ \text{super-critical if} & \lambda > \mu. \end{cases}$$

In particular, the branching process has certain extinction if and only if $\lambda \leq \mu$.

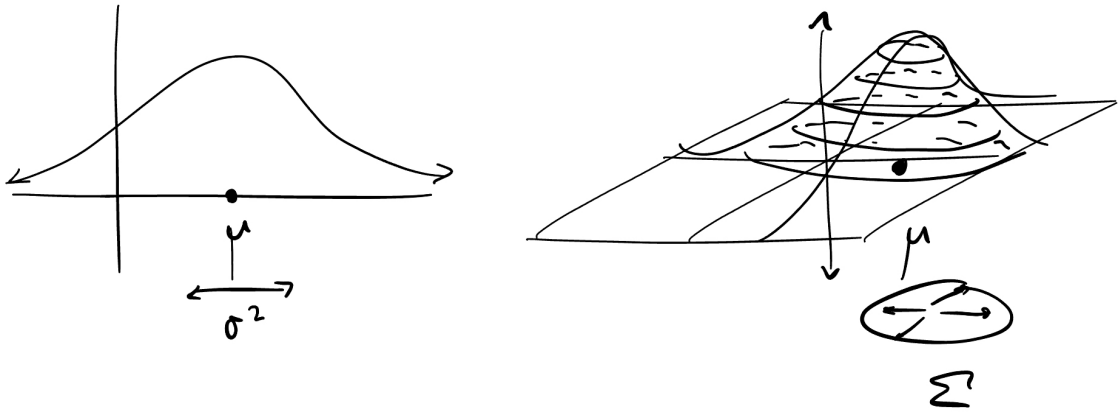
In summary, we showed that $\{X_t\}_{t \geq 0}$ is recurrent if $\lambda \leq \mu$ and that it is transient if $\lambda > \mu$. (Note that this is similar to what we observed for $M/M/1$ by comparison with biased random walk.) In fact, it is possible to show that it is positive-recurrent if $\lambda < \mu$ and null-recurrent if $\lambda = \mu$, but this is a bit harder to prove.

One can also study similar problems for G/M and G/G queues, but this requires a few more results on renewal theory in addition to existing results on PPs. So we don't pursue it in this class.

5 Gaussian processes

In most of the class before this, we studied stochastic processes with the Markov property, and this allowed us to deduce many things. Presently, we'll focus on a different dependence structure—Gaussian distribution—which also allows us to do many calculations and deduce many thins.

Before, we get into the details, let's give a picture. You have probably seen the "bell curve" depiction of a Gaussian distribution with mean μ and variance σ^2. We will try to generalize this to higher dimensions, leading to pictures like the following:



We will later see what μ and Σ are, but for now you can think of them as encoding the "center" and "spread". So, Σ will play the role of σ^2.

Recall that the *(univariate) Gaussian density* with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

For a random variable X with density f, we write $X \sim \mathcal{N}(\mu, \sigma^2)$, and it follows that $\mathbb{E}[X] = \mu$, and $\mathrm{Var}(X) = \sigma^2$. There are also many other exact calculations we can do with Gaussian random variables, and

we will see many of them throughout this next part of the course. For now let's just recall a fundamental fact that the sum of indepedent Gaussian random variables is Gaussian.

5.1 Multivarite Gaussian distributions

To begin:

DEF. A random vector $(X_1, \ldots, X_m)$ is *multivariate Gaussian* if the random variable $\lambda_1 X_1 + \cdots + \lambda_m X_m$ has univariate Gaussian distribution for all real numbers $\lambda_1, \ldots, \lambda_m$.

Geometrically speaking, the operation of "taking the linear combination" means "projecting onto a line". So, a joint distribution is multivariate Gaussian if and only if all of the one-dimensional projections are univariate Gaussian.

EX (IID Gaussians)**.** Suppose $\varepsilon_1, \ldots, \varepsilon_m$ are IID with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $\sigma^2 > 0$. We claim that $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ is multivariate Gaussian. To see this, take arbitrary $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ and note that

$$\lambda_1 \varepsilon_1 + \cdots + \lambda_m \varepsilon_m$$

is just a sum of independent (but not identically-distributed) Gaussians, hence it is Gaussian.

EX (copies)**.** Suppose that $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ and that $X_i = X_1$ for all $2 \leq i \leq m$. So, $X_1, \ldots, X_m$ are all copies of the same random variable. We claim that $X = (X_1, \ldots, X_m)$ is a multivariate Gaussian. To show this, take arbitrary $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ and note that

$$\lambda_1 X_1 + \cdots + \lambda_m \lambda_1 X_m = (\lambda_1 + \cdots + \lambda_m) X_1$$

is just a constant times a Gaussian hence it is Gaussian.

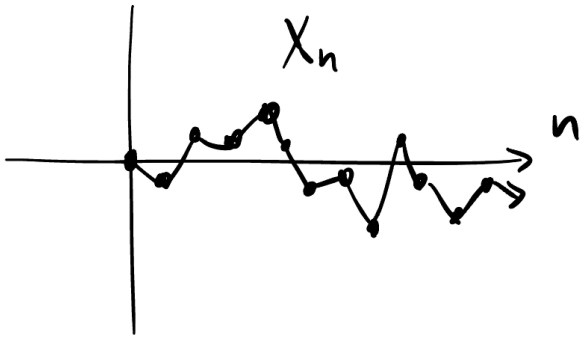
It is also convenient to be able to discuss multivariate Gaussians with infinite index sets. We will later focus on this a lot more, and with extensions to continuous time.

DEF. A *discrete-time Gaussian process (DTGP)* is a stochastic process $\{X_n\}_{n \geq 0}$ such that $(X_1, \ldots, X_n)$ is multivariate Gaussian for all $n \geq 0$.

EX (Gaussian RW)**.** Suppose $X_0 = 0$ and $\varepsilon_1, \varepsilon_2, \ldots$ are IID from $\mathcal{N}(0, \sigma^2)$, and let the *Gaussian random walk* be the process $\{X_n\}_{n \geq 0}$ defined via

$$X_n = X_{n-1} + \varepsilon_n$$

for $n \geq 1$. We can picture the sample paths of this process as follows:



Note that this is qualitatively very similar to the SSRW that we studied as a DTMC, but now we have a continuous state space intead of a discrete state space. Nonetheless, we will later see that these two models are closely related!

Anyway, we claim that $\{X_n\}_{n\geq 0}$ is a Gaussian process. To do this, we observe that we can write $X_n = \sum_{i=1}^n \varepsilon_i$, which means X is just the process of cumulative sums of the Gaussian errors. Then for any $\lambda_1, \dots, \lambda_n$ we have

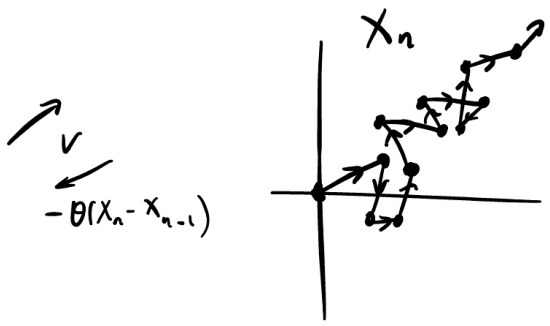
$$\begin{aligned}&\lambda_1 X_1 + \cdots + \lambda_n X_n \\ &= (\lambda_1 + \cdots + \lambda_n)\varepsilon_1 + (\lambda_2 + \cdots + \lambda_n)\varepsilon_2 + \cdots + (\lambda_{n-1} + \lambda_n)\varepsilon_{n-1} + \lambda_n \varepsilon_n.\end{aligned}$$

As before, this is a sum of independet (but not identically distributed) Gaussians, hence Gaussian. This shows that the Gaussian RW is a Gaussian process.

EX (particle in viscous fluid)**.** We consider the following physical model for a propelling particle passing through a viscous fluid. That is, for $v \in \mathbb{R}^2$ and $\theta > 0$ and $X_0 = 0 \in \mathbb{R}^2$, set

$$X_n - X_{n-1} = v - \theta(X_n - X_{n-1}) + \varepsilon_n$$

for $n \geq 1$. For a picture:



Here, we think of v as "velocity" and θ as the "drag coefficient". This means that the motion is slowed by applying a force which acts opposite to the direction of motion.

We claim that $\{X_n\}_{n\geq 0}$ is a Gaussian process. To see this, rewrite the above:

$$X_n = \frac{v}{1+\theta} + X_{n-1} + \frac{1}{1+\theta}\varepsilon_n,$$

and note that this implies

$$X_n = \frac{n}{1+\theta} v + \frac{1}{1+\theta} \sum_{i=1}^n \varepsilon_n.$$

In order to check the desired condition, we need to compute an arbitrary linear combination of the preceding values. But notice that we are now in $\mathbb{R}^2$ rather than $\mathbb{R}$. So, we can write $\lambda^\top X_n$ for an arbitrary linear combination, and this will simplify the notation a bit. Then, for $\lambda_1, \dots, \lambda_n \in \mathbb{R}^2$ we compute:

$$\begin{aligned}&\lambda_1^\top X_1 + \cdots + \lambda_n^\top X_n \\ &= \frac{(\lambda_1 + 2\lambda_2 + \cdots + n\lambda_n)^\top}{1+\theta} v \\ &+ \frac{(\lambda_1 + \cdots + \lambda_n)^\top}{1+\theta}\varepsilon_1 + \frac{(\lambda_2 + \cdots + \lambda_n)^\top}{1+\theta}\varepsilon_2 + \cdots + \frac{(\lambda_{n-1} + \lambda_n)^\top}{1+\theta}\varepsilon_{n-1} + \frac{\lambda_n^\top}{1+\theta}\varepsilon_n.\end{aligned}$$

As we already saw, this is a linear sum of independent Gaussians (plus a consant) hence univariate Gaussian.

EX (autoregressive process)**.** Fix $p \geq 1$ and $\theta_1, \ldots, \theta_p$. Suppose that $\varepsilon_1, \varepsilon_2, \ldots$ are IID from $\mathcal{N}(0, \sigma^2)$, that $X_{-(p-1)}, \ldots, X_0$ are fixed and known, and that

$$X_n = \theta_1 X_{n-1} + \cdots + \theta_p X_{n-p} + \varepsilon_n$$

for $n \geq 1$. We won't go through the details here, but it turns out that this is also a Gaussian process. This is one of the most important models in time series analysis, mathematical finance, etc. since it's quite simply but also very expressive.

Just like how we reduced the study of MCs to the study of their transition matrices / infinitesimal generator matrices, we want to reduce the study of Gaussian distributions to some calculations involving linear algebra.

DEF. For a random vector $X = (X_1, \ldots, X_m)$, its *covariance matrix* is the symmetric matrix $\mathrm{Cov}(X) \in \mathbb{R}^{m \times m}$ defined via

$$\mathrm{Cov}(X)_{ij} = \mathrm{Cov}(X_i, X_j).$$

Equivalently, if we think of X as a column vector, then

$$\mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top].$$

DEF. A matrix $\Sigma \in \mathbb{R}^{m \times m}$ is called *positive semi-definite (PSD)* if it is symmetric and we have $\lambda^\top \Sigma \lambda \geq 0$ for all $\lambda \in \mathbb{R}^m$. It is called *positive definite (PD)* or *strictly positive definite* if $\lambda^\top \Sigma \lambda > 0$ for all $\lambda \in \mathbb{R}^m$ with $\lambda \neq 0$.

We observe that a covariance matrix is always PSD. Indeed, for X any random vector in $\mathbb{R}^m$ and any $\lambda \in \mathbb{R}^m$, note:

$$\begin{aligned}
\lambda^\top \mathrm{Cov}(X) \lambda &= \lambda^\top \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \lambda \\
&= \mathbb{E}[\lambda^\top (X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top \lambda] && \text{(linearity of expectation)} \\
&= \mathbb{E}\left[\langle \lambda, X - \mathbb{E}[X] \rangle^2\right] && \text{(definition of inner product)} \\
&\geq 0 && \text{(square is non-negative.)}
\end{aligned}$$

This calculation also allows us to characaterize when a covariance matrix is PD. That is, note that we have

$$\mathbb{E}\left[\langle \lambda, X - \mathbb{E}[X] \rangle^2\right] = 0 \text{ if and only if } \langle \lambda, X - \mathbb{E}[X] \rangle = 0.$$

The latter condition just means that X is always contained in a subspace of $\mathbb{R}^m$. So, covariance matrices degenerate exactly when the random variable is "not full-dimensional".

For every Gaussian random vector X we can make its covariance. It turns out that we can also go the other way! We do not prove the details here, but the following result states that Gaussian distributions are exactly in correspondence with covariance matrices.

THM. *For every $\mu \in \mathbb{R}^m$ and every PSD $\Sigma \in \mathbb{R}^{m \times m}$, there exists a unique multivariate Gaussian X with $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$. The distribution of X is denoted $\mathcal{N}(\mu, \Sigma)$.*

There are a few other things we should establish about covariance matrices before we move onto our larger study of Gaussian processes.

LEM. *If $X \sim \mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{m \times m}$ is arbitrary, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^\top)$.*

PF: First we need to check that AX is indeed Gaussian. To do this, we need to take arbitrary λ and show that $\lambda^\top AX$ is a univariate Gaussian. To do this, write

$$\langle \lambda, AX \rangle = \langle A^\top \lambda, X \rangle$$

and note that the right hand side is a univariate Gaussian, because of our assumption that X is multivariate Gaussian. Now we just need to check the mean and covariance of X. We of course have $\mathbb{E}[AX] = A\mathbb{E}[X]$ by the linearity of expectation. Also:

$$\begin{aligned}
&\lambda^\top \mathrm{Cov}(AX)\lambda \\
&= \lambda^\top \mathbb{E}\left[(AX - \mathbb{E}[AX])(AX - \mathbb{E}[AX])^\top\right]\lambda && \text{(definition of covariance matrix)} \\
&= \lambda^\top \mathbb{E}\left[A(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top A^\top\right]\lambda && \text{(algebra)} \\
&= \lambda^\top A\mathbb{E}\left[A(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right]A^\top\lambda && \text{(linearity of expectation.)}
\end{aligned}$$

This completes the proof.

Here are some other important facts:

LEM. *If (X_1, X_2) are mutlivariate normal, then $\mathrm{Cov}(X_1, X_2) = 0$ if and only if X_1 and X_2 are independent.*

LEM. *If $X_n \sim \mathcal{N}(\mu_n, \Sigma_n)$ and $X \sim \mathcal{N}(\mu, \Sigma)$, then $X_n \to X$ if and only if $\mu_n \to \mu$ and $\Sigma_n \to \Sigma$.*

LEM. *If $X_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ are multivariate Gaussian $A \sim \mathrm{Ber}(\alpha)$ for $0 < \alpha < 1$ is independent of X_0, X_1, then the random vector*

$$\tilde{X} := \begin{cases} X_0 & \text{if } A = 0 \\ X_1 & \text{if } A = 1 \end{cases}$$

is multivariate Gaussian if and only if $\mu_0 = \mu_1$ and $\Sigma_0 = \Sigma_1$.

Now we revisit our earlier examples.

EX (IID Gaussians)**.** If $X_1, \ldots, X_n$ are IID Gaussians with mean 0 and variance σ^2, then $X = (X_1, \ldots, X_n)$ is a Gaussian random vector with mean vector

$$\mu = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

We often denote this covariance matrix as $\sigma^2 I$ or $\sigma^2 I_n$ for convenience.

EX (copies)**.** Suppose $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $X_i = X_1$ for all $2 \le i \le m$. Then $X = (X_1, \ldots, X_n)$ is a Gaussian random vector with mean vector

$$\mu = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2 & \cdots & \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 & \sigma^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 & \sigma^2 \end{pmatrix}.$$

We can also extend these definitions easily to the case of discrete-time Gaussian process, where our covariance matrix is now infinite. For instance:

EX (Gaussian RW)**.** Let $\varepsilon_1, \varepsilon_2, \ldots$ be IID Gaussian from $\mathcal{N}(0, \sigma^2)$, and $X_0 = 0$ and

$$X_n = X_{n-1} + \varepsilon_n$$

for $n \geq 1$. Then

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & \sigma^2 & \sigma^2 & \sigma^2 & \cdots \\ 0 & \sigma^2 & 2\sigma^2 & 2\sigma^2 & \cdots \\ 0 & \sigma^2 & 2\sigma^2 & 3\sigma^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \sigma^2 \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 2 & \cdots \\ 0 & 1 & 2 & 3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Sometimes it is useful to view the mean and covariance in "block form", meaning the entries are themselves matrices.

EX (particle in viscous fluid)**.** Fix $v \in \mathbb{R}$ and $\theta > 0$ and $X_0 = 0 \in \mathbb{R}^2$, set

$$X_n - X_{n-1} = v - \theta(X_n - X_{n-1}) + \varepsilon_n$$

for $n \geq 1$. We already saw this this can be represented as

$$X_n = \frac{n}{1+\theta} v + \frac{1}{1+\theta} \sum_{i=1}^{n} \varepsilon_i,$$

hence its mean and covariance are just:

$$\mu = \frac{1}{1+\theta} \begin{pmatrix} 0 \\ v \\ 2v \\ 3v \\ \vdots \end{pmatrix}$$

and

$$\Sigma = \frac{\sigma^2}{1+\theta} \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & I_2 & I_2 & I_2 & \cdots \\ 0 & I_2 & 2I_2 & 2I_2 & \cdots \\ 0 & I_2 & 2I_2 & 3I_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

5.2 Definitions and examples

DEF. For any set T, a *Gaussian process (GP)* is a random function $X : T \to \mathbb{R}$ such $(X_{t_1}, \ldots, X_{t_m})$ is a Gaussian random vector for all $t_1, \ldots, t_n \in T$.

Usually we think of T as "time". If $T = \{1, \ldots, m\}$, then a Gaussian process is just a Gaussian vector. If $T = \{0, 1, 2, \ldots\}$, then this is equivalent to our earlier definition of discrete-time GP. If $T = [0, \infty)$ then we refer to this as a *continuous-time GP (CTGP)*. Later we will see that one can take $T = \mathbb{R}^2$ in which case the random function is usually called a *Gaussian field*.

DEF. For any Gaussian process $X : T \to \mathbb{R}$, its *mean function* $\mu : T \to \mathbb{R}$ and *coavariance function* $\Sigma : T \times T \to \mathbb{R}$ are defined via

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

and

$$\Sigma(s,t) := \mathrm{Cov}(X_s, X_t)$$

for all $s, t \in T$.

Note that if $T = \{1, \ldots, m\}$ then the mean function and covariance function are just the same as the mean vector and covariance matrix! If $T = \{0, 1, 2, \ldots\}$ then someties we think of Σ as being an "infinite matrix" instead of as a function, but these are of course equivalent.

Like in the case of Gaussian vectors, we want to understand which functions μ and Σ correspond to the mean and covariance functions of some Gaussian process.

DEF. A function $\Sigma : T \times T \to \mathbb{R}$ is called a *positive semi-definite (PSD) kernel* if it is $\Sigma(s,t) = \Sigma(t,s)$ for all $s, t \in T$ and if the matrix

$$\begin{pmatrix} \Sigma(t_1,t_1) & \cdots & \Sigma(t_1,t_m) \\ \vdots & \ddots & \vdots \\ \Sigma(t_m,t_1) & \cdots & \Sigma(t_m,t_m), \end{pmatrix}$$

is PSD for all $t_1, \ldots, t_m \in T$. Matrices of the above form are called *minors* of Σ.

THM. *For every function $\mu : T \to \mathbb{R}$ and for every PSD kernel $\Sigma : T \times T \to \mathbb{R}$, there exists a unique Gaussian process $X : T \to \mathbb{R}$ with mean function μ and covariance function Σ.*

If X is a GP with mean function μ and covariance function Σ, then $\tilde{X}$ defined via $\tilde{X}_t = X_t - \mu(t)$ is a GP with mean zero and with covariance function Σ. Moreover, most aspects of X can be re-stated in terms of $\tilde{X}$. So we occasionally assume that the mean function is zero, in which case the GP is called *centered*.

Now we can explore many examples. For now let's stick to the case of $T = [0, \infty)$, but there are many interesting examples where $T = \mathbb{R}^d$ for $d \geq 2$. First we will see examples where we can explicitly construct the process $\{X_t\}_{t \geq 0}$ from some auxiliary random variables, and that will tell us everything we want to know about its structure, its sample paths, etc.

EX (cosine process)**.** For $\sigma^2 > 0$ and $\lambda > 0$, consider the covariance function

$$\Sigma(s,t) = \sigma^2 \cos(\lambda(t-s)).$$

To see what this process looks like, it suffices to construct a GP with this covariance function. To do this, let $\varepsilon_1, \varepsilon_2$ be IID samples from $\mathcal{N}(0, \sigma^2)$, and set

$$X_t = \varepsilon_1 \cos(\lambda t) + \varepsilon_2 \sin(\lambda t).$$

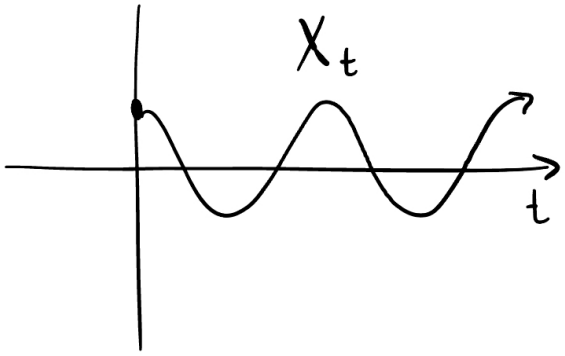
Now we just need to compute

$$\mu(t) = \mathbb{E}[X_s] = \mathbb{E}[\varepsilon_1] \cos(\lambda t) + \mathbb{E}[\varepsilon_2] \sin(\lambda t) = 0,$$

and

$$\begin{aligned} \mathrm{Cov}(X_s, X_t) &= \mathbb{E}[X_s X_t] \\ &= \mathbb{E}[\varepsilon_1^2] \cos(\lambda s) \cos(\lambda t) \\ &\quad + \mathbb{E}[\varepsilon_1 \varepsilon_2] \cos(\lambda s) \sin(\lambda t) \\ &\quad + \mathbb{E}[\varepsilon_2 \varepsilon_1] \sin(\lambda s) \cos(\lambda t) \\ &\quad + \mathbb{E}[\varepsilon_2^2] \sin(\lambda s) \sin(\lambda t) && \text{(algebra)} \\ &= \sigma^2 (\cos(\lambda s) \cos(\lambda t) + \sin(\lambda s) \sin(\lambda t)) && \text{(expectation)} \\ &= \sigma^2 (\cos(\lambda(s-t)). && \text{(cosine difference formula)} \end{aligned}$$

Thus, the sample paths of the cosine process look like this:



EX (linear GP)**.** For $a^2, b^2 > 0$, consider the covariance function

$$\Sigma(s,t) = a^2 st + b^2$$

for $s, t \geq 0$. In order to construct a GP with this covariance kernel, take any $\mu : [0,\infty) \to \mathbb{R}$, and set

$$X_t = At + B$$

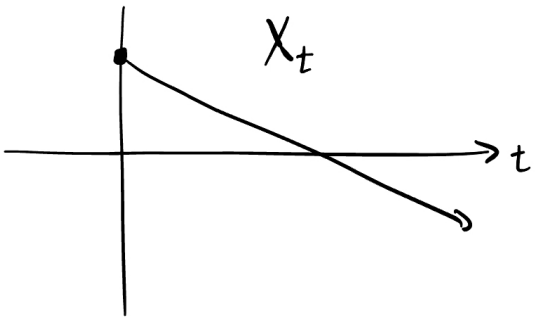
where A and B are independent Gaussian random variables with mean zero and variance a^2 and b^2, respectively. Note that we can compute:

$$\mathbb{E}[X_t] = \mathbb{E}[A]t + \mathbb{E}[B] = 0$$

and

$$\begin{aligned} \mathrm{Cov}(X_s, X_t) &= \mathrm{Cov}(At + B, As + B) \\ &= \mathrm{Var}(A)st + \mathrm{Var}(B) \\ &= a^2 st + b^2 \\ &= \Sigma(s,t). \end{aligned}$$

Thus, a GP with this covariance is just a random line, whose slope and intercept are independent Gaussian random variables.



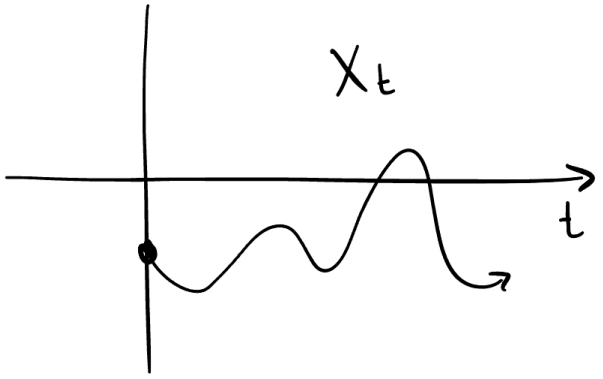
In HW11 you will consider a class of examples which generalizes the linear GP. That is, for any $\phi : \mathbb{R} \to \mathbb{R}^k$, one can set $\Sigma(s,t) = \langle \phi(s), \phi(t) \rangle$, and it turns out that Σ is always a covariance function.

In some other cases, we cannot easily construct $\{X_t\}_{t \geq 0}$, although its existence is guaranteed by the representation theorem above. This motivates our general theory; we would like to be able to deduce properties of $\{X_t\}_{t \geq 0}$ directly from its covariance function Σ, and we will indeed be able to do this soon!

EX (square-exponential GP)**.** For $\ell > 0$, consider

$$\Sigma(s,t) = \exp\left(-\left(\frac{s-t}{\ell}\right)^2\right).$$

It is not so easy to show directly, but this is indeed a PSD kernel. Its sample paths look something like this:

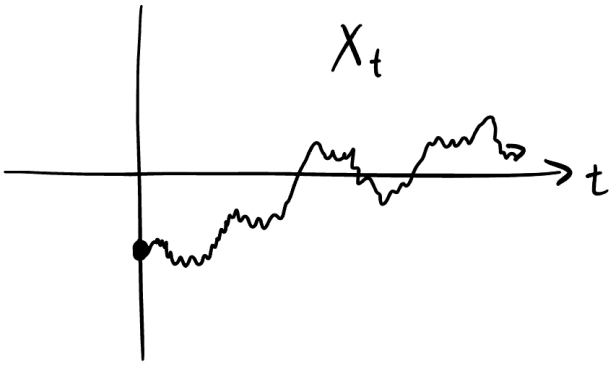


Note that $\ell > 0$ controls the strength of the covariance across time.

EX (Ornstein-Uhlenbeck process)**.** For $\ell > 0$, consider

$$\Sigma(s,t) = \exp\left(-\frac{|s-t|}{\ell}\right).$$

Again it is not so easy to think about directly, but we will later see that its sample paths look like this:



Note that they look qualitative similar to the sample paths of the square-exponential GP, but with a much "rougher" structure.

5.3 Smoothness properties

When we think of a stochastic process as a random function, it natural to wonder when this random function is smooth in some sense (meaning continuous, differentiable, etc.) In discrete-time, this question does not make any sense, but in continuous-time it is rather interesting.

Throughout this part, we let $\{X_t\}_{t\geq 0}$ denote a general CTGP with mean function μ and covariance function Σ.

DEF. We say $\{X_t\}_{t\geq 0}$ is *mean-square continuous (MSC)* if

$$\lim_{h\to 0} \mathbb{E}\left[|X_{t+h} - X_t|^2\right] = 0$$

for all $t \geq 0$.

THM. *A CTGP is MSC if and only if its mean function and covariance function are continuous.*

One can also ask for more smoothness beyond continuity. For example,

DEF. We say $\{X_t\}_{t\geq 0}$ is *mean-square differentiable (MSD)* if there exists a stochastic process $\{Y_t\}_{t\geq 0}$ such that

$$\lim_{h\to 0} \mathbb{E}\left[\left(\frac{|X_{t+h} - X_t|}{h} - Y_t\right)^2\right] = 0$$

for all $t \geq 0$. More generally, we say that $\{X_t\}_{t\geq 0}$ is *mean-square differentiable or order k (MSD-k)* if there exists a stochastic process $\{Y_t\}_{t\geq 0}$ such that

$$\lim_{h\to 0} \mathbb{E}\left[\left(\frac{\Delta_h^k X_t}{h^k} - Y_t\right)^2\right] = 0,$$

where Δ_h is the discrete-difference operator. (For example, $\Delta_h X_t = X_{t+h} - X_t$, $\Delta_h^2 X_t = X_{t+2h} + X_t - 2X_{t+h}$, etc.)

Note that MSC is just MSD-0 and MSD is just MSD-1. In general, MSD-k is roughly saying that the sample paths of $\{X_t\}_{t\geq 0}$ are k-times differentiable. While in general it can be difficult to characterize higher-order smoothness of GPs, we can do this exactly for a restricted class of GPs:

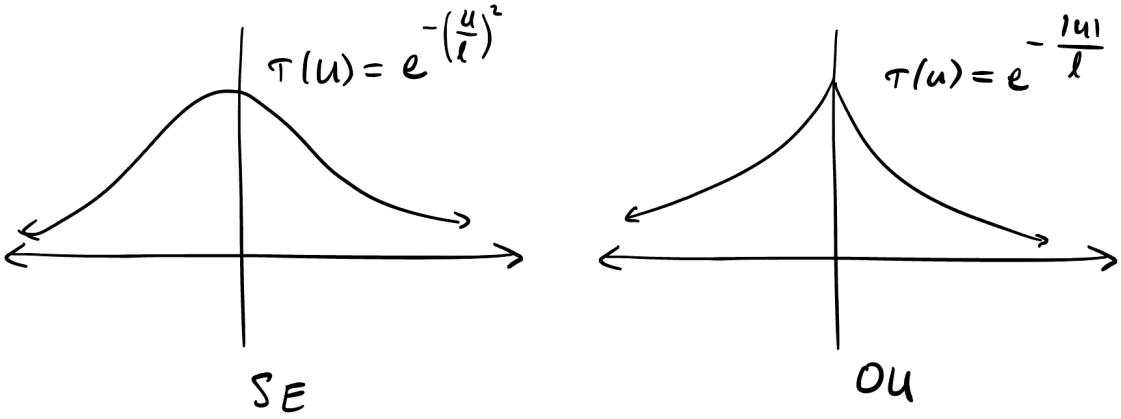
DEF. A Gaussian process $\{X_t\}_{t\geq 0}$ is called *stationary* if there exists a function $\tau : \mathbb{R} \to \mathbb{R}$ such that its covariance function Σ can be written as $\Sigma(s,t) = \tau(t-s)$.

Here, "stationary" is being used in the same way as "stationary increments" from Poisson processes but not in the same way at "stationary distributin" from Markov chains. Intuitively, a stationary GP is one for which the correlations in the window $[s,t]$ are the same as those in the window $[0, t-s]$.

LEM. *If $\{X_t\}_{t\geq 0}$ is a stationary Gaussian process with covariance function $\tau : \mathbb{R} \to \mathbb{R}$, then $\{X_t\}_{t\geq 0}$ is MSD-k if and only if the derivative $\tau^{(k)}(0)$ exists and is finite.*

EX (cosine process)**.** Because of the explicit formulation of the cosine process above, we know that its sample paths are infinitely differentiable. This is reflected by the fact that the covariance function is $\Sigma(s,t) = \sigma^2 \cos(\lambda(t-s))$, and the function $\tau(u) = \sigma^2 \cos(\lambda u)$ is infinitely differentiable.

EX (SE versus OU)**.** Consider $\Sigma(s,t) = \exp(-|t-s|/\ell)$ Since this function is continuous, we know that $\{X_t\}_{t\geq 0}$ is MSC. But $\tau(u) = \exp(-|u|/\ell)$ is not differentiable at $u = 0$, so $\{X_t\}_{t\geq 0}$ is not MSD. This explains why its sample paths are quite "rough".



5.4 Conditioning

One of the most important things about Gaussian processes is that, if we condition on observing some value at a fixed time, then the resulting process is still Gaussian, and we have an explicit formula for its covariance function. As we will see later, this allows us to do many exact calculations that are not possible for other stochastic processes!

To begin, we give a concrete statement for Gaussian random vectors.

LEM. *Suppose that $(X, Y) \in \mathbb{R}^m \times \mathbb{R}^n$ is a multivariate Gaussian with*

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} \right),$$

so $\mu_X \in \mathbb{R}^m, \mu_Y \in \mathbb{R}^n, \Sigma_X \in \mathbb{R}^{m\times m}, \Sigma_Y \in \mathbb{R}^{n\times n}, \Sigma_{XY} \in \mathbb{R}^{m\times n}$, and $\Sigma_{YX} \in \mathbb{R}^{n\times m}$. Then, the conditional distribution of

$$X \quad \text{given} \quad \{Y = y\}$$

is a multivariate Gaussian random vector with mean vector

$$\tilde{\mu} = \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_Y)$$

and covariance matrix

$$\tilde{\Sigma} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}.$$

The proof is not hard; it just requires applying Bayes rule and then completing the square, using sufficient linear algebra notation. Also, the interpretation of the result is simpler than it seems: The value $y - \mu_r$ is just the distance of y from its mean, and multiplying this by $\Sigma_{XY}\Sigma_Y^{-1}$ is like a "change of scale", where we translate from Y units to X units.

This result is very powerful, and it let us do very many explicit calculations.

EX. Let $\{X_n\}_{n\geq 0}$ be a Gaussian RW, with $X_0 = 0$ and with noise variance $\sigma^2 > 0$. Our goal is to find the conditional distribution of

$$X_{2n} \text{ given } \{X_n = x\}$$

for fixed $n \geq 0$. One can think of this as trying to predict the value of X_{2n} given the value $X_n = x$; knowing the exact distribution further allows one to construct confidence intervals for the predictions.

To begin, recall that we earlier computed $\mathrm{Cov}(X_n, X_m) = \sigma^2 \min\{n, m\}$, hence we have

$$\begin{pmatrix} X_n \\ X_{2n} \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} n & n \\ n & 2n \end{pmatrix} \right).$$

Thus, by applying the conditoning fornula, we find that the desired conditional distribution of X_{2n} given $\{X_n = x\}$ has mean

$$0 + \sigma^2 n \, \frac{x - 0}{\sigma^2 n} = x$$

and variance

$$\sigma^2 2n - \frac{(\sigma^2 n)^2}{\sigma^2 n} = \sigma^2 n.$$

In other words:

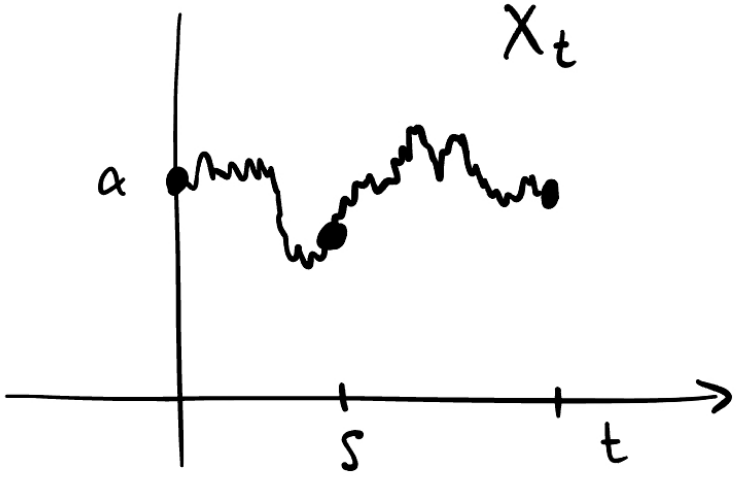
$$(X_{2n} \,|\, X_n = x) \sim \mathcal{N}\left(x, \sigma^2 n\right).$$

You may notice that this is just the marginal distribution of X_n given $X_0 = x$. Indeed, this is true and it has to do with the fact that Gaussian RW has independent increments. (We will come back to this when we study Brownian motion later.)

EX (optimal execution in OU bridge)**.** Let $\{X_t\}_{t\geq 0}$ be an Ornstein-Uhlenbeck process with covariance function $\Sigma(s,t) = \exp(-\gamma|t-s|)$. For some fixed $\alpha > 0$ and $t \geq 0$, let us try to solve the problem

$$\max_{0\leq s\leq t} \mathbb{E}[X_s \,|X_0 = X_t = \alpha].$$

We visualize this as follows:



In words, we aim to find a time to maximize the expected value of the process, conditional on knowing that it must begin and end in the same value.

To do this, note that the joint distribution of X_s, X_0, and X_t is a multivariate Gaussian parameterized as follows:

$$\begin{pmatrix} X_s \\ X_0 \\ X_t \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & e^{-\gamma s} & e^{-\gamma(t-s)} \\ e^{-\gamma s} & 1 & e^{-\gamma t} \\ e^{-\gamma(t-s)} & e^{-\gamma t} & 1 \end{pmatrix} \right),$$

Now we apply the conditioning formula, for $X := X_s$ and $Y := (X_0, X_t)^\top$. We will need invert the 2×2 matrix Σ_Y which in this case is

$$\Sigma_Y = \begin{pmatrix} 1 & e^{-\gamma t} \\ e^{-\gamma t} & 1 \end{pmatrix} \qquad \Rightarrow \qquad \Sigma_Y^{-1} = \frac{1}{1-e^{-2\gamma t}} \begin{pmatrix} 1 & -e^{-\gamma t} \\ -e^{-\gamma t} & 1 \end{pmatrix}.$$

Therefore, the conditional mean of X_s given $Y = (\alpha, \alpha)^\top$ is just

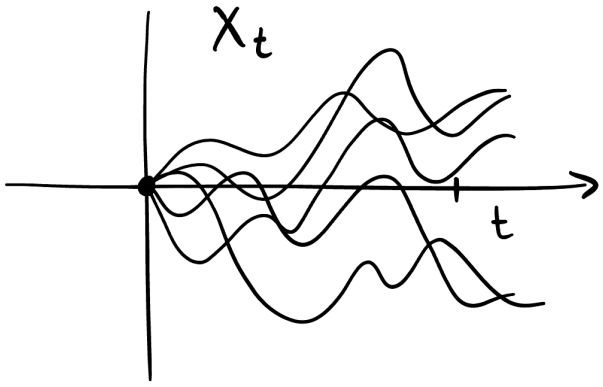
$$\begin{aligned} \tilde{\mu} &= 0 + \begin{pmatrix} e^{-\gamma s} & e^{-\gamma(t-s)} \end{pmatrix} \frac{1}{1-e^{-2\gamma t}} \begin{pmatrix} 1 & e^{-\gamma t} \\ e^{-\gamma t} & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} \\ &= \frac{\alpha}{1+e^{-\gamma t}} \left(e^{-\gamma s} + e^{-\gamma(t-s)} \right). \end{aligned}$$

Notice that this is just $e^{-\gamma s} + e^{-\gamma(t-s)}$ as a function of s, so the maximum occurs at $s = 0, t$. (Also, the minimum occurs at $s = t/2$.) In other words, the conditional expectation of the process is maximized at the endpoints of the time interval $[0, t]$.

EX. Let $\{X_t\}_{t\geq 0}$ be a GP with mean function $\mu : [0,\infty) \to \mathbb{R}$ satisfying $\mu(0) = 0$ and with square-exponential covariance function $\Sigma(s,t) = \exp(-\gamma(t-s)^2)$ for some $\gamma > 0$. Our goal to find the conditional distribution

$$\begin{pmatrix} X_t \\ X_t' \end{pmatrix} \quad \text{given} \quad \{X_0 = 0\} \tag{5.1}$$

for fixed $t \geq 0$. What do we expect the answer to be? Let's just draw some sample paths to try to get some intution:



By inspecting these, we might guess that X_t and X_t' are positively correlated under this conditioning.

To find the conditional distribution exactly, first note that we can determine the following joint distribution of X_0, X_t, and X_{t+h}, for any $t, h \geq 0$:

$$\begin{pmatrix} X_0 \\ X_t \\ X_{t+h} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(0) \\ \mu(t) \\ \mu(t+h) \end{pmatrix}, \begin{pmatrix} 1 & e^{-\gamma t^2} & e^{-\gamma(t+h)^2} \\ e^{-\gamma t^2} & 1 & e^{-\gamma h^2} \\ e^{-\gamma(t+h)^2} & e^{-\gamma h^2} & 1 \end{pmatrix} \right),$$

In particular, we can apply the fact that linear transformatins of Gaussians remain Gaussian to get:

$$\begin{pmatrix} X_0 \\ X_t \\ (X_{t+h} - X_t)/h \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(0) \\ \mu(t) \\ (\mu(t+h) - \mu(t))/h \end{pmatrix}, \begin{pmatrix} 1 & e^{-\gamma t^2} & \frac{e^{-\gamma(t+h)^2} - e^{-\gamma t^2}}{h} \\ e^{-\gamma t^2} & 1 & \frac{e^{-\gamma h^2} - 1}{h} \\ \frac{e^{-\gamma(t+h)^2} - e^{-\gamma t^2}}{h} & \frac{1 - e^{-\gamma h^2}}{h^2} & \frac{2(1 - e^{-\gamma h^2})}{h^2} \end{pmatrix} \right).$$

Now take $h \to 0$ and note that the entries reduce to just calculating the value of derivatives, hence:

$$\begin{pmatrix} X_0 \\ X_t \\ X_t' \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu(t) \\ \mu'(t) \end{pmatrix}, \begin{pmatrix} 1 & e^{-\gamma t^2} & -2\gamma t e^{-\gamma t^2} \\ e^{-\gamma t^2} & 1 & 0 \\ -2\gamma t e^{-\gamma t^2} & 0 & 2\gamma \end{pmatrix} \right).$$

This tells us that the random vector $(X_0, X_t, X_t')^\top$ is multivariate Gaussian, so now we can just apply the conditioning formula to get:

$$\left(\begin{pmatrix} X_t \\ X_t' \end{pmatrix} \,\middle|\, X_0 = 0 \right) \sim \mathcal{N}\left(\begin{pmatrix} \mu(t) \\ \mu'(t) \end{pmatrix}, \begin{pmatrix} 1 - e^{-2\gamma t^2} & 2\gamma t e^{-2\gamma t^2} \\ 2\gamma t e^{-2\gamma t^2} & 2\gamma(1 - 2\gamma t^2 e^{-2\gamma t^2}) \end{pmatrix} \right).$$

Note that this confirms our intution since, for example we can see that X_t and X_t' are postively correlated under this conditioning.

It's also instructive to compare this with the unconditional distribution of $(X_t, X_t')^\top$, which we can read off from our earlier calculation:

$$\begin{pmatrix} X_t \\ X_t' \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(t) \\ \mu'(t) \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2\gamma \end{pmatrix} \right).$$

Interestingly, observe that X_t and X_t' are independent without the conditioning!

It is also possible to apply the conditioning formula to the continuous-time GP directly. This leads to the following, which is also useful in some applications:

LEM. *If $\{X_t\}_{t\geq 0}$ is a GP with mean function μ and covariance function Σ, then the conditional distribution of*

$$\{X_t\}_{t\geq 0} \text{ given } \{X_r = x\}$$

is a GP with mean function

$$\tilde{\mu}(t) = \mu(t) + \Sigma(r,t)\,\frac{x-\mu(r)}{\Sigma(r,r)}$$

covariance function

$$\tilde{\Sigma}(s,t) = \Sigma(s,t) - \frac{\Sigma(s,r)\Sigma(r,t)}{\Sigma(r,r)}$$

The analogous statement for discrete-time is true, as well.

5.5 Karhunen-Loève Expansion

Our last topic concerns a fundamental representation theorem for Gaussian processes. To begin, we should recall the following representation of a multivariate Gaussian random vector:

THM. *Let $\Sigma \in \mathbb{R}^{m\times m}$ any covariance matrix. Then, there exists real numbers $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ and vectors $\phi_1, \ldots, \phi_m \in \mathbb{R}^m$ such that the following holds:*

(1) $\|\phi_1\| = 1$ for all i, and $\langle \phi_i, \phi_j \rangle = 0$ for all $i \neq j$.

(2) For each i, we have $\Sigma\phi_i = \lambda_i \phi_i$.

In particular, if $\varepsilon_1, \cdots, \varepsilon_n$ are IID from $N(0,1)$, then the random vector

$$X := \sum_{i=1}^{n} \sqrt{\lambda_i}\,\varepsilon_i \phi_i$$

is a centered multivariate Gaussian with covariance matrix Σ.

The proof of this result is basically an application of the spectral decomposition for symmetric matrices. But we will not focus on the details since the linear algebra is not so central to this class. Instead, we use this as a point of comparison for the following:

THM. *Let $\Sigma : T \times T \to \mathbb{R}$ be any PSD kernel. Then, there exists real numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and functions $\phi_1, \phi_2, \ldots : T \to \mathbb{R}$ such that the following holds:*

(1) $\int_T (\phi_i(t))^2 \,\mathrm{d}t = 1$ for all i, and $\int_T \phi_i(t)\phi_j(t)\,\mathrm{d}t = 0$ for all $i \neq j$.

(2) For each i, we have $\int_T \Sigma(s,t)\phi_i(t)\,\mathrm{d}t = \lambda_i \phi_i(s)$ for all $s \in T$.

In particular, if $\varepsilon_1, \varepsilon_2, \ldots$ are IID from $N(0,1)$, then the stochastic process $\{X_t\}_{t\in T}$ defined via

$$X_t := \sum_{i=1}^{\infty} \sqrt{\lambda_i}\,\varepsilon_i \phi_i(t)$$

is a centered Gaussian process with covariance function Σ.

Intuitively, conditions (1) and (2) in the preceding two results are saying the same thing. The functions ϕ are eigenvectors of the covariance Σ, in the same sense as the vectors v are eigenvectors of the covariance matrix Σ. Indeed, notice that if the integrals with sums in (1) and (2) above, then we recover exactly the definition of vector norm and matrix-vector multiplication! We also note that the existence of λ and ϕ satisfying (1) and (2) follows, again, from some sort of spectral theorem, but we will not focus on the details in this class.

As a sanity check, let's see that the second part of the theorem follows easily from the first part: If $\{X_t\}_{t\in T}$ is defined above, then we can simply compute:

$$\begin{aligned}
\mathrm{Cov}(X_s, X_t) &= \mathrm{Cov}\left(\sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \varepsilon_i \phi_i(s), \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \varepsilon_j \phi_i(t)\right) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sqrt{\lambda_i \lambda_j}\phi_i(s)\phi_j(t)\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) && \text{(linearity of covariance)} \\
&= \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t) && \text{(property (1))} \\
&= \sum_{i=1}^{\infty} \int_T \Sigma(r,s)\phi_i(r)\,\mathrm{d}r\phi_i(t) && \text{(property (2))} \\
&= \int_T \Sigma(r,s) \sum_{i=1}^{\infty} \phi_i(r)\phi_i(t)\,\mathrm{d}r && \text{(algebra)} \\
&= \Sigma(s,t)
\end{aligned}$$

In the last line, we used the fact that $\sum_{i=1}^{\infty} \phi_i(r)\phi_i(t) = \delta_t(r)$, which is similar to the case of covariance matrices.

This result has a few interpretations. First of all, it gives a way to simulate a GP using just some IID Gaussian random variables, which appear as the coefficients above. Second, it tells us (in some generality) how to directly construct a GP $\{X_t\}_{t\geq 0}$ with a desired covariance, and this can help us learn things about $\{X_t\}_{t\geq 0}$ from Σ. (Of course, this requires us to be able to find ϕ and λ which is not always easy to do. But often it reduces to solving a system of ODEs.)

EX (cosine process)**.** Recall that for $\sigma^2 > 0$ and $\lambda > 0$, a cosine process $\{X_t\}_{t\geq 0}$ is a Gaussian process of the form

$$X_t = \varepsilon_1 \cos(\lambda t) + \varepsilon_2 \sin(\lambda t).$$

where $\varepsilon_1, \varepsilon_2$ are IID samples from $\mathcal{N}(0, \sigma^2)$. If we just write $\varepsilon_1 = \sqrt{\sigma^2}\tilde{\varepsilon}_1$ and $\varepsilon_2 = \sqrt{\sigma^2}\tilde{\varepsilon}_2$ so that $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2$ are IID centered Gaussians with variance 1, then we can see that this is already in the form of the KL representation. So, the explicit representation for cosine processes that we derived earlier was in fact a KL representation!

EX (OU process)**.** Suppose $\{X_t\}_{t\geq 0}$ is an OU process with covariance $\Sigma(s,t) = \exp(-\gamma|t-s|)$ for γ. It turns out that there exists a sequence $\{w_i\}_{i\geq 1}$ of real numbers such that the λ and ϕ can be written as

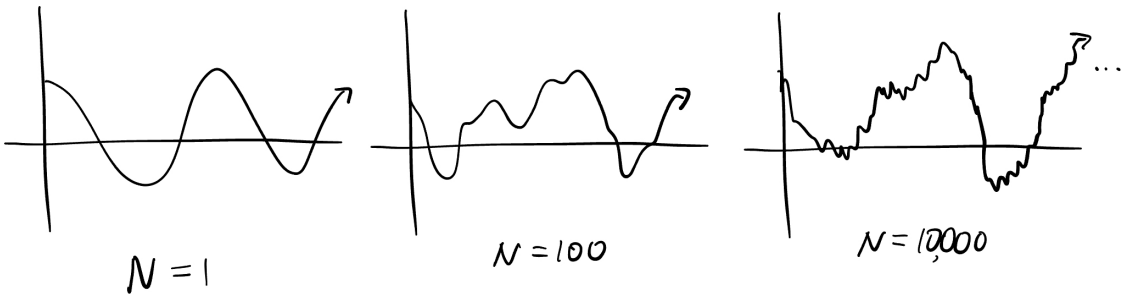
$$\begin{aligned}
\lambda_i &= \frac{2\gamma}{w_i^2 + \gamma^2} \\
\phi_i(t) &= \sqrt{\frac{2w_i^2}{2\gamma + w_i^2 + \gamma^2}}\cos(w_i t) + \sqrt{\frac{\gamma^2}{2\gamma + w_i^2 + \gamma^2}}\sin(w_i t).
\end{aligned}$$

There is no exact formula for the $\{w_i\}_{i\geq 1}$ but they approximately $w_i \approx i\pi$ for large i, hence $\lambda_i \approx 1/(i^2\pi^2)$. We will not do this in detail, but you will verify in HW 11 that these indeed satisfy the desired conditions.

In many applications, the importance of the KL representation is that it gives a way to approximate X_t as $\sum_{i=1}^{N} \sqrt{\lambda_i}\, \varepsilon_i \phi(t)$ for some large N. But we should be careful about a few things:

- How should we choose N? In fact, the "approximation error" is like $\sum_{i=N+1}^{\infty} \lambda_i$, which goes to 0 as $N \to \infty$.
- How can $\sum_{i=1}^{N} \sqrt{\lambda_i}\, \varepsilon_i \phi(t)$ be smooth whilte X_t is non-smooth? This involves some real analysis, but basically a limit of smooth functions does not need to be smooth.

We can visualize these considerations in the case of the OU process.



Note that the leftmost plot looks like a cosine process, and that the sample paths get rougher as N increases.

5.6 Convergence of Gaussian processes

Just like how multivariate Gaussian random vectors converge exactly when the mean vectors and covariance matrices converge, something similar is true for Gaussian processes. However, we will not study any precise theorems in this part; rather, we will give some heuristic arguments, and we will mention that it takes more real analysis and measure to make them precise.

EX (AR(1) $\to$ OU)**.** For each N, take $\theta_N > 0$ and suppose that $\varepsilon_1, \varepsilon_2, \ldots$ are independent identically-distributed samples from $\mathcal{N}(0,1)$. Define the first-order Gaussian autoregressive process $\{X_n^{(N)}\}_{n \geq 0}$ via

$$X_n^{(N)} = \theta_N X_{n-1}^{(N)} + \varepsilon_n$$

for $n \geq 1$. Recall that the mean function of this process is $\mu^{(N)}(n) = 0$ and that the covariance is just

$$\Sigma^{(N)}(n,m) = \theta_N^{|m-n|}.$$

Now let $\tilde{X}^{(N)}$ be defined as

$$\tilde{X}_t^{(N)} := X_{Nt},$$

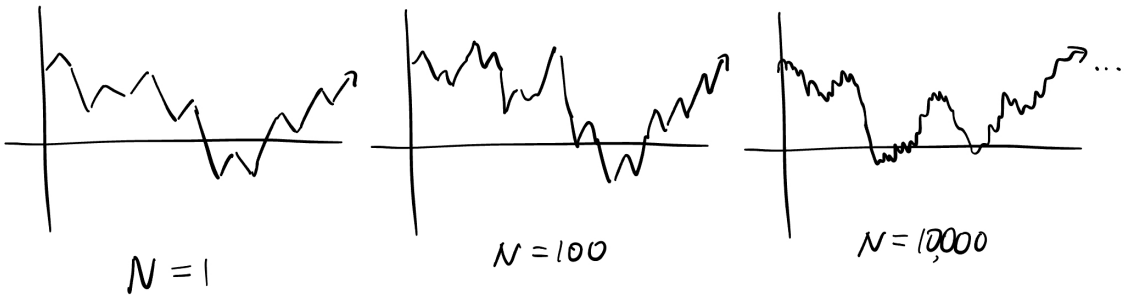
which means we "speed up time" by a factor of N. (In other words, what previously took 1 unit of time now takes $1/N$ units of time. So this is like "zooming out" in the time axis.) Now compute:

$$\begin{aligned} \tilde{\Sigma}^{(N)}(s,t) &= \operatorname{Cov}\left(\tilde{X}_s^{(N)}, \tilde{X}_t^{(N)}\right) \\ &= \operatorname{Cov}\left(X_{Ns}, X_{Nt}\right) \\ &= \theta_N^{N|t-s|}. \end{aligned}$$

In order for this to converge to a limit as $N \to \infty$, we should take $\theta_N = 1 - \frac{\gamma}{N}$, so that

$$\tilde{\Sigma}^{(N)}(s,t) = \theta_N^{N|t-s|} = \left(\left(1 - \frac{\gamma}{N}\right)^N\right)^{|t-s|} \to e^{-\gamma|t-s|}.$$

In other words, we have shown that the AR(1) process converges to the Ornstein-Uhlenbeck process when the scaling is chosen appropriately! We visualize this as follows:



This give another way to simulate an OU process, in addition to the KL representation that we saw before. It is possible to show that the AR approximation has a larger error, but that it reflects the roughness property which may be desirable in applications.

EX (Gaussian RW $\to$?)**.** Let $\varepsilon_1, \varepsilon_2, \ldots$ be IID samples from $\mathcal{N}(0, \sigma^2)$ and define the Gaussian RW via

$$X_n = X_{n-1} + \varepsilon_n$$

and $X_0 = 0$. Now consider the stochastic process

$$\left\{ \frac{1}{\sqrt{N}} S_{Nt} \right\}_{t \geq 0}.$$

The interpretation is also somewhat simple. By replacing the time index t with Nt, we compress time by a factor of N or equivalently zoom out on the time axis with a factor fo N. By multiplying the outside by $1/\sqrt{N}$, we compress space by a fator of $\sqrt{N}$ or equivalent zoom on the space axis with a factor of $\sqrt{N}$. As we will now compure, these time- and space-scaling gives a covariance function with a limit:

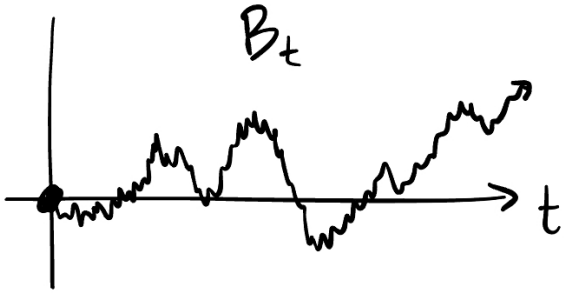
$$\begin{aligned}
\Sigma^{(N)}(s,t) &= \operatorname{Cov}\left(\frac{1}{\sqrt{N}} S_{Ns}, \frac{1}{\sqrt{N}} S_{Nt} \right) \\
&= \frac{1}{N} \operatorname{Cov}\left(S_{Ns}, S_{Nt} \right) \\
&= \frac{1}{N} \min\{Ns, Nt\} \\
&= \min\{s, t\}.
\end{aligned}$$

What is this Gaussian process that arises as the limit of Gaussian RW under this particular scaling? This is exactly Brownian motion that we will study for the rest of the class.

6 Brownian motion

In this last part of the class we will study a fundamental stochastic process representing continuous-time motion, called *Brownian motion*. As we will see, this connects many of the ideas from different parts of the class!

Before we get into the details, here is a cartoon of Brownian motion that you can keep in mind:



6.1 Basic definition

It turns out that there are many equivalent ways to define Brownian motion. Here is a good first definition:

DEF. A *Brownian motion (BM)* is a stochastic process $\{B_t\}_{t\geq 0}$ with $B_0 = 0$ and posessing continuous sample paths, satisfying

(1) $B_t - B_s$ has a distribution $\mathcal{N}(0, t-s)$ for all $0 \leq s \leq t$, and

(2) $B_t - B_s$ is independent of B_s for all $0 \leq s \leq t$.

You may notice some similarities with our first definition of Poisson process. (Indeed, Poisson Processes and Brownain motion are closely related, although we will not discuss it in this class.) That is, BM has stationary increments (condition (1) above) and independent increments, (condition (2) above), although the increments have Gaussian distributions instead of Poisson distributions. Also, there is a qualitative difference: we required a PP to be a counting process, but we require a BM to be a continuous process.

Like in our discussion of PPs, it turns out that one does not need to require the Gaussian part of condition (1); remarkably, it is implied by the other parts! In the case of PPs this was due to the law of rare events, but in the case of BM it is simply due to the central limit theorem.

THM. *If a stochastic process $\{B_t\}_{t\geq 0}$ has $B_0 = 0$, has continuous sample paths, and satisfies*

(1) $B_t - B_s$ has the same distribution as B_{t-s} for all $0 \leq s \leq t$, and

(2) $B_t - B_s$ is independent of B_s for all $0 \leq s \leq t$,

then it is a Brownian motion.

PF: It suffices to show that B_t has distribution $\mathcal{N}(0,t)$ for all $t \geq 0$. To do this, let's just consider the special case $t = 1$. For any $n \geq 0$, define

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{n} \left(B_{\frac{i}{n}} - B_{\frac{i-1}{n}} \right).$$

Since this is just a telescoping sum, we have $Y_n = B_1$ for all $n \geq 0$. But (1) and (2) tell us exactly that Y_n is a sum of IID random variables with the same variance. So

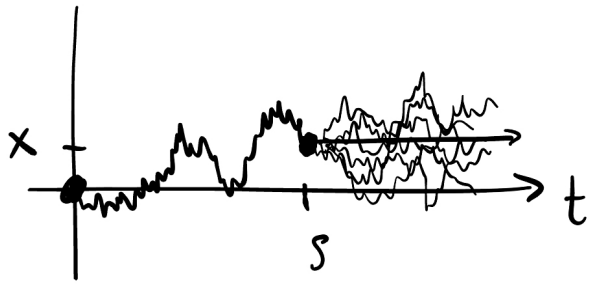
$$B_1 = Y_n \to \mathcal{N}(0,1)$$

as claimed.

EX. For $0 \leq s \leq t$ and $x \in \mathbb{R}$, we compute:

$$\begin{aligned}
\mathbb{E}[B_t \,|\, B_s = x] &= \mathbb{E}[B_s + (B_t - B_s) \,|\, B_s = x] \\
&= \mathbb{E}[B_s \,|\, B_s = x] + \mathbb{E}[B_t - B_s \,|\, B_s = x] && \text{(linearity of conditional expectation)} \\
&= x + \mathbb{E}[B_t - B_s \,|\, B_s = x] && \text{(definition of conditional expectation)} \\
&= x + \mathbb{E}[B_t - B_s] && \text{(property (2))} \\
&= x && \text{(property (1))}
\end{aligned}$$

This calculation can be interpreted as a prediction problem: If we know that the value of BM at time s is x, then our best guess for the value at all future times is still x. A picture to visualize this is as follows



where after time s we plot many different possible sample paths, and we observe that their mean is always x.

EX. For $0 \leq s \leq t$, we compute:

$$\begin{aligned}
\mathrm{Cov}(B_s, B_t) &= \mathrm{Cov}(B_s, B_s + (B_t - B_s)) \\
&= \mathrm{Var}(B_s) + \mathrm{Cov}(B_s, B_t - B_s) && \text{(linearity of covariance)} \\
&= s + \mathrm{Cov}(B_s, B_t - B_s) && \text{(property (1))} \\
&= s. && \text{(property (2))}
\end{aligned}$$

In other words, for all $s, t \geq 0$ we have $\mathrm{Cov}(B_s, B_t) = \min\{s, t\}$. Note that already saw this last week when we discussed BM in the context of limits of Gaussian RW; it's nice to see that we can show this directly from the definition above! Also recall from the midterm that we did a similar calculation for PPs.

Notice that in both problems we used the same trick of writing B_s as $B_s + (B_t - B_s)$. This is the key trick for many problems involving Brownian motion: figuring out how to write the quantities of interest in terms of increments.

In the next two parts of the class, we will situate the study of BM in the context of two types of stochastic processes we have already studied: Gaussian processes and random walks. This will allow us to understand lots of aspects of BM by applying things we learned earlier in the class.

6.2 As a Gaussian process

Once we show that BM is a Gaussian process, we will be able to apply all of the tools from the previous part of the class. So the starting point is the following:

THM. *Suppose $\{B_t\}_{t \geq 0}$ is a centered Gaussian process with continuous sample paths and with covariance function $\Sigma(s,t) = \min\{s,t\}$. Then $\{B_t\}_{t \geq 0}$ is a Brownian motion.*

PF: By the uniqueness part of the representation theorem for Gaussian processes, it suffices to show that BM is a Gaussian process. That is, we fix $0 \le t_1 \le \cdots \le t_k$ and we need to show that $(B_{t_1}, \ldots, B_{t_k})$ has a multivariate Gaussian distribution. To do this, we take arbitrary $\lambda_1, \ldots, \lambda_k$, and we need to show that $\lambda_1 B_{t_1} + \cdots + \lambda_k B_{t_k}$ has a univariate Gaussian distribution. To do this, we need to represent each B_{t_i} as a sum of increments, since we know they are Gaussian and independent, For simplicty let's just take $k = 2$ and note

$$\begin{aligned}\lambda_1 B_{t_1} + \lambda_2 B_{t_2} &= \lambda_1 B_{t_1} + \lambda_2 (B_{t_1} + (B_{t_2} - B_{t_1})) \\ &= (\lambda_1 + \lambda_2) B_{t_1} + \lambda_2 (B_{t_2} - B_{t_1}).\end{aligned}$$

Since B_{t_1} and $B_{t_2} - B_{t_1}$ are independent Gaussians by definition, we see that $\lambda_1 B_{t_1} + \lambda_2 B_{t_2}$ indeed has a univariate Gaussian distribution. The general case is similar, but the notation is a bit more involves; so, the proof is complete.

This allows us to do immediately derive some interesting things about BM:

EX (smoothness)**.** The BM covariance function is continuous, so we know from general theory of GPs that BM is MSC. (Of course, we required in the definition that BM has contininuous sample paths, so this is not new information.) Although the covariance function is not stationary (so our GP theory does not directly apply), one can show that BM is not MSD (hence also not MSD-k for any $k \ge 1$.) This explains why BM has "rough" sample paths in our cartoons. In HW 13 you will see that BM and OU are very closely related, and in fact they have the same smoothness properties.

EX (Brownian bridge)**.** If $\{B_t\}_{0 \le t \le 1}$ is a BM, let's compute its conditional distribution given $\{B_1 = 0\}$. We know that this is itself a GP and we can find its covariance as follows:

$$\tilde{\Sigma}(s,t) = \min\{s,t\} - \frac{\min\{s,1\}\min\{t,1\}}{\min\{1,1\}} = \min\{s,t\} - st,$$

for $0 \le s, t \le 1$. Note the similarity with the analogous problem on Gaussian RW from HW 11; for BM, conditioning the endpoint to be equal to 0 is the same as subtracting a certain linear GP!

EX (Karhunen-Loève for BM)**.** If $\{B_t\}_{t \ge 0}$ is a BM, then its KL representation is

$$B_t = \sum_{i=1}^{\infty} \sqrt{2}\, \varepsilon_i \, \frac{\sin\left(\left(i - \frac{1}{2}\right)\pi t\right)}{\left(i - \frac{1}{2}\right)\pi t}$$

for $0 \le t \le 1$, where $\varepsilon_1, \varepsilon_2, \ldots$ are IID Gaussians. As usual, determining these ϕ and λ requires solving some differential equations, which we will not focus on in this class.

6.3 As a a random walk

We have already seen that BM can be regarded as a limit of Gaussian RW in a suitable sense. As we explain next, BM can be regarded as a limit of many different kinds of RW, including also the SSRW.

DEF. For IID random variables $X_1, X_2, \ldots$ with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$, the *random walk with increments* $X_1, X_2, \ldots$ is the discrete-time stochastic process $\{S_n\}_{n \ge 0}$ defined via

$$S_n := \sum_{i=1}^{n} X_i.$$

(Note that this is not usually equivalent to being aRW on a graph.)

If $X_1 = \pm 1$ with equal probability, then this is just the simple symmetric RW that we studied in the beginning of the course, and if $X_1 \sim \mathcal{N}(0,1)$ then this is just the Gaussian RW that we studied during the part of the course on Gaussian processes. But this construction is much more general; we could have, for example, $X \sim \text{Exp}(1)$ with probability $1/2$ and $X = -1$ with probability $1/2$.

The following famous result states that all RWs look like Brownian motion if we rescale time and space appropriately:

THM (Donsker)**.** *Suppose that $X_1, X_2, \ldots$ are IID with $\mathbb{E}[X_1] = 0$ and $\text{Var}(X_1) = 1$, and set $S_n := \sum_{i=1}^n X_i$. Then,*

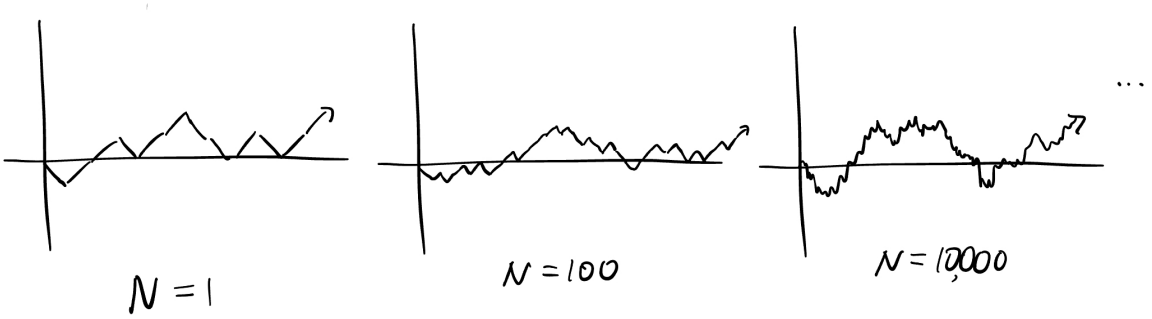
$$\left\{ \frac{1}{\sqrt{N}} S_{Nt} \right\}_{t \geq 0}$$

converges to BM as $N \to \infty$. More precisely, for (most) functions $f : \{\text{sample paths}\} \to \mathbb{R}$, we have

$$\mathbb{E}\left[f\left(\left\{ \frac{1}{\sqrt{N}} S_{Nt} \right\}_{t \geq 0} \right) \right] \to \mathbb{E}\left[f\left(\{B_t\}_{t \geq 0} \right) \right] \tag{6.1}$$

as $N \to \infty$.

The value of this is that we can "transfer" existing results for RWs onto results for BM, and this is useful since we can often do exact calculations for, say, SSRW or Gaussian RW. In some sense, it can be seen as a "definition" of BM, since it shows that BM is exactly what arises in the limit of RWs. For instance,



This picture provides some explanation for the roughness of the samples paths of BM: in some sense, a BM is "moving both up and down with equal probability" at all times!

In order to apply Donsker's theorem, we usually need to come up with a clever choice of an observable f. Usually, we write $\{\omega_t\}_{t \geq 0}$ for an arbitrary path (which may be a sample path of a random walk, a rescaling thereof, a sample path of a BM, etc.)

EX (recurrence)**.** Let $\{B_t\}_{t \geq 0}$ be a BM, and for $s > 0$ set $T(s) := \min\{t \geq s : B_t = 0\}$ as the first time after s that the BM returns to state 0. (Note that we usually need a complicated definition like this for continuous-time Markov processes, since they can come back to state 0 in some infinitesimal time.) We claim that BM is recurrent in the sense that $\mathbb{P}(T(s) < \infty) = 1$ for all $s > 0$. In other words, BM will return to state 0 at arbitarily large times in the future.

To prove this, fix $s > 0$ and take a path $\{\omega_t\}_{t \geq 0}$, and define

$$f_s(\{\omega_t\}_{t \geq 0}) := \begin{cases} 1 & \text{if } \{\omega_t\}_{t \geq 0} \text{ returns to 0 at some time after } s \\ 0 & \text{if } \{\omega_t\}_{t \geq 0} \text{ never returns to 0 after time } s. \end{cases}$$

Wwe know from the recurrence of the SSRW that

$$\mathbb{E}\left[f_s\left(\left\{ \frac{1}{\sqrt{N}} S_{Nt} \right\}_{t \geq 0} \right) \right] = 1$$

for all $N \geq 0$, but also by Donsker's theorem that

$$\mathbb{E}\left[f_s\left(\left\{\frac{1}{\sqrt{N}}S_{Nt}\right\}_{t\geq 0}\right)\right] \to \mathbb{E}\left[f_s\left(\{B_t\}_{t\geq 0}\right)\right].$$

Therefore, we have shown

$$\mathbb{P}(T(s) < \infty) = \mathbb{E}\left[f_s\left(\{B_t\}_{t\geq 0}\right)\right] = 1.$$

In other words, BM is recurrent. (It is also natural to ask whether BM is null-recurrent or positive recurrent; you will do this on HW 13.)

EX (reflection principle)**.** A fundamental problem in mathematical finance is to check whether BM exceeds some value x *at any point* during a time interval $[0, t]$. Typically, this is because such "crossing events" in options markets are conditions for executing a constract, so the crossing probabilities are important in determine the price of an option.

To understand such crossing events, we focus first on a SSRW $\{S_n\}_{n\geq 0}$, where we can do some exact calculations. That is, our goal is to compute the probability

$$\mathbb{P}\left(\max_{0\leq j\leq n} S_j \geq k\right)$$

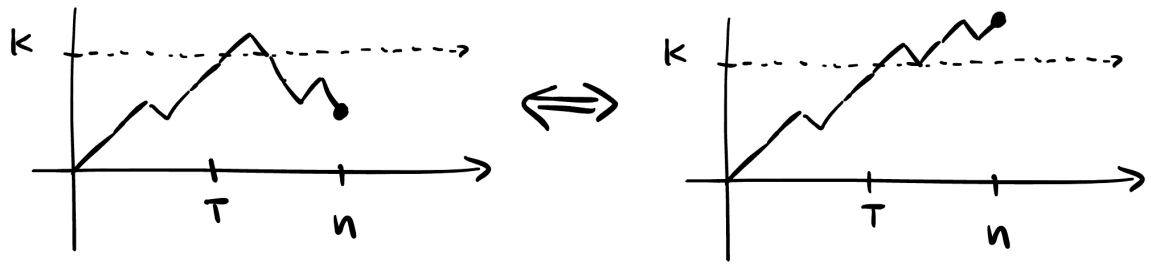
for fixed $n, k \geq 0$. Note that this is a rather complicated event, since it depends on the entire trajectory of $\{S_n\}_{n\geq 0}$. Nonetheless, we can compute its probability exactly. To do this, note that we can break up the event based on the value of the endpoint S_n, yielding:

$$\begin{aligned}\left\{\max_{0\leq j\leq n} S_j \geq k\right\} &= \left\{\max_{0\leq j\leq n} S_j \geq k, S_n \leq k-1\right\} \cup \left\{\max_{0\leq j\leq n} S_j \geq k, S_n \geq k\right\} \\ &= \left\{\max_{0\leq j\leq n} S_j \geq k, S_n \leq k-1\right\} \cup \{S_n \geq k\}\end{aligned}$$

The last equality is due to the fact that S_n exceeding k certainly implies that the running max exceeds k. Now comes the key observation: we have a bijection between samples paths satisfying the events

$$\left\{\max_{0\leq j\leq n} S_j \geq k, S_n \leq k-1\right\} \quad \text{and} \quad \{S_n \geq k+1\}.$$

This is best explained in a picture:



In words, we simply "reflect" each sample path across the horizontal line at state k, for all parts of the path after the time T of the first visit to state k. In particular, putting it all together shows

$$\mathbb{P}\left(\max_{0\leq j\leq n} S_j \geq k\right) = \mathbb{P}\left(S_n \geq k+1\right) + \mathbb{P}\left(S_n \geq k\right).$$

This is called the *reflection principle for the SSRW*. Also note that the right side above can be exactly computed, since the marginal distribution of S_n is just $\text{Binom}(n, 1/2)$.

Now we show the analogous statement for BM. To do this, fix $t > 0$ and $x \geq 0$, and define the obserable f via

$$f(\{\omega_t\}_{t\geq 0}) := \begin{cases} 1 & \text{if } \max_{0\leq s\leq t} \omega_s \geq x \\ 0 & \text{if } \max_{0\leq s\leq t} \omega_s < x. \end{cases}$$

Using the reflection principle for SSRW, we find

$$\begin{aligned} \mathbb{E}\left[f\left(\left\{\frac{1}{\sqrt{N}}S_{Nt}\right\}_{t\geq 0}\right)\right] &= \mathbb{P}\left(\max_{0\leq s\leq t} S_{Ns} \geq x\sqrt{N}\right) \\ &= \mathbb{P}\left(\max_{0\leq j\leq Nt} S_j \geq x\sqrt{N}\right) \\ &= \mathbb{P}\left(S_{Nt} \geq x\sqrt{N}+1\right) + \mathbb{P}\left(S_{Nt} \geq x\sqrt{N}\right) \\ &= \mathbb{P}\left(\frac{1}{\sqrt{N}}S_{Nt} \geq x + \frac{1}{\sqrt{N}}\right) + \mathbb{P}\left(\frac{1}{\sqrt{N}}S_{Nt} \geq x\right). \end{aligned}$$

Therefore, Donsker's theorem and the central limit theorem together show:

$$\begin{aligned} \mathbb{P}\left(\max_{0\leq s\leq t} B_s \geq x\right) &= \mathbb{E}\left[f\left(\{B_t\}_{t\geq 0}\right)\right] \\ &= \lim_{N\to\infty} \mathbb{E}\left[f\left(\left\{\frac{1}{\sqrt{N}}S_{Nt}\right\}_{t\geq 0}\right)\right] \\ &= \lim_{N\to\infty}\left(\mathbb{P}\left(\frac{1}{\sqrt{N}}S_{Nt} \geq x + \frac{1}{\sqrt{N}}\right) + \mathbb{P}\left(\frac{1}{\sqrt{N}}S_{Nt} \geq x\right)\right) \\ &= \mathbb{P}\left(\mathcal{N}(0,t) \geq x\right) + \mathbb{P}\left(\mathcal{N}(0,t) \geq x\right) \\ &= 2\mathbb{P}\left(\mathcal{N}(0,t) \geq x\right). \end{aligned}$$

Lastly, recall that $B_t \sim \mathcal{N}(0,t)$ by definition. Thus, we have shown

$$\mathbb{P}\left(\max_{0\leq s\leq t} B_s \geq x\right) = 2\mathbb{P}\left(B_t \geq x\right).$$

which is called the *reflection principle for BM.*

In some sense, the reflection principle for BM should also be understood as a statement about a bijection between certain sets of paths satisfying two different events. That is, paths which cross x at some time and are beow x at the final time are in bijection with paths that are above x at the final time. However, it's a bit hard to make this precise directly in continuous-time. But in discrete-time, the combinatorial structure of SSRW allows us to set up this bijection exactly!

A consequence of the reflection principle is the following:

EX (first passage time)**.** Let $\{B_t\}_{t\geq 0}$ be a BM and for $x > 0$, define $T_x := \min\{t \geq 0 : B_t = x\}$ as the time of the first visit to state x. Then we can compute the CDF of T_x as follows, using the reflection principle:

$$\begin{aligned} \mathbb{P}(T_x \leq t) &= \mathbb{P}(B_s \geq x \text{ for some } 0 \leq s \leq t) \\ &= \mathbb{P}\left(\max_{0\leq s\leq t} B_s \geq x\right) \\ &= 2\mathbb{P}\left(B_t \geq x\right) \\ &= 2\mathbb{P}\left(\mathcal{N}(0,t) \geq x\right) \\ &= \mathbb{P}\left(\left(\frac{x}{\mathcal{N}(0,1)}\right)^2 \leq t\right). \end{aligned}$$

In other words, T_x has the same distribution as $(x/Z)^2$ where $Z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable. (On HW 13, you will use this formula to compute the expectation $\mathbb{E}[T_1]$.)

6.4 Invariances

When we studied PPs, we showed that there were some transformations that we can apply to a PP (thinning, superposition) which yield another PP. Presently we will do the same thing for some transformations of BM. We will only discuss a few of these invariances, but there are actually very many of them.

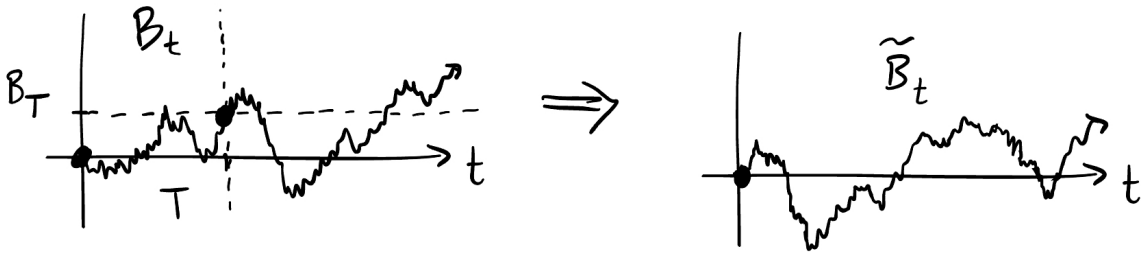
The first invariance is the following sort of translation:

LEM. *If $\{B_t\}_{t\geq 0}$ is a BM and $T > 0$ is fixed, then $\{\tilde{B}_t\}_{t\geq 0}$ defined via*

$$\tilde{B}_t := B_{t+T} - B_T$$

is also a BM.

We can visualize this transformation as follows:



In words, we take the space-time point (T, B_T), and we translate it to $(0,0)$. So the resulting sample path $\tilde{B}$ is just the sample path B after time T, suitably placed at the origin.

This result is easy to prove, and we will actually prove it two different ways; once using the definition with continuous sample paths and stationary independent increments, and once using the definition in terms of GPs.

PF: (via stationary and independent increments) It is clear that $\{\tilde{B}_t\}_{t\geq 0}$ has continuous sample paths, since each of its sample paths is contained in a sample path of B which is continuous by definition. Moreover, we easily check

$$\tilde{B}_0 = B_{0+T} - B_T = 0$$

as needed. To see the stationary increments property (1), we need to show that $\tilde{B}_t - \tilde{B}_s$ has distribution $\mathcal{N}(0, t-s)$. But we can compute

$$\tilde{B}_t - \tilde{B}_s = B_{t+T} - B_T - (B_{s+T} - B_T) = B_{t+T} - B_{s+T}$$

and the right side has distribution $\mathcal{N}(0, t+T-(s+T)) = \mathcal{N}(0, t-s)$ since B is a Brownian motion. To see the independent increments property (2), we fix $0 \leq s \leq t$ and we note that $\tilde{B}_t - \tilde{B}_s$ and $\tilde{B}_s$ are just equal to

$$\begin{aligned} \tilde{B}_t - \tilde{B}_s &= B_{t+T} - B_T - (B_{s+T} - B_T) = B_{t+T} - B_{s+T} \\ \tilde{B}_s &= B_{s+T} - B_T \end{aligned}$$

Since $0 \leq s+T \leq t+T$ and B is a Brownian motion, it follows that these random variables are independent.

PF: (via Gaussian processes) Note that $\{\tilde{B}_t\}_{t\geq 0}$ is a linear transformation of $\{B_t\}_{t\geq 0}$, so it is also a GP. Moreover, its covariance function is

$$\begin{aligned}\mathrm{Cov}(\tilde{B}_s, \tilde{B}_t) &= \mathrm{Cov}(B_{s+T} - B_T, B_{t+T} - B_T) \\ &= \mathrm{Cov}(B_{s+T}, B_{t+T}) - \mathrm{Cov}(B_{s+T}, B_T) \\ &\quad - \mathrm{Cov}(B_{s+T}, B_T) + \mathrm{Var}(B_T) \\ &= \min\{s+T, t+T\} - T - T + T \\ &= \min\{s, t\}.\end{aligned}$$

So, by the theorem above, it is also a Brownian motion.

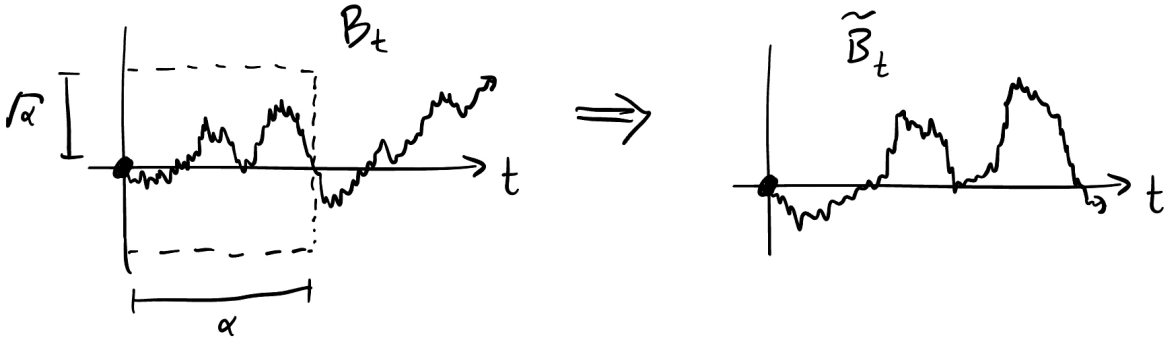
The second invariance is the following sort of scaling:

LEM. *If $\{B_t\}_{t\geq 0}$ is a BM and $\alpha > 0$ is fixed, then $\{\tilde{B}_t\}_{t\geq 0}$ defined via*

$$\tilde{B}_t := \sqrt{\alpha} B_{t/\alpha}$$

is also a BM.

We can visualize this transformation as follows:



In other words, we rescale time by a factor of α and space by a factor of $\sqrt{\alpha}$.

PF: (via Gaussian processes) It is clear that $\{\tilde{B}_t\}_{t\geq 0}$ is a Gaussian process, since it is a linear function of $\{B_t\}_{t\geq 0}$. Moreover, its covariance function is

$$\mathrm{Cov}(\sqrt{\alpha}B_{s/\alpha}, \sqrt{\alpha}B_{t/\alpha}) = \alpha \mathrm{Cov}(B_{s/\alpha}, B_{t/\alpha}) = \alpha \min\left\{\frac{s}{\alpha}, \frac{t}{\alpha}\right\} = \min\{s, t\}.$$

By the theorem above, this implies that $\{\tilde{B}_t\}_{t\geq 0}$ is a BM.

One can also prove this result via the definition involving stationary and independent increments, of course. But we omit the details here.

6.5 Markov property

One of the most important properties of BM is that it satisfies a form of continuous-time Markov property:

THM. *If $\{B_t\}_{t\geq 0}$ is a BM, then the conditional distribution of X_t given $\{B_r = x, B_s = y\}$ equals the conditional distribution of X_t given $\{B_s = y\}$, for all times $0 \leq r \leq s \leq t$ and all states $x, y \in \mathbb{R}$.*

Like in many of our earlier results, we can prove this two different ways:

PF: (via Gaussian processes) Consider the multivariate Gaussian random vector

$$\begin{pmatrix} B_t \\ B_r \\ B_s \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} t & r & s \\ r & r & r \\ s & r & s \end{pmatrix} \right)$$

and let us compute its conditional distribution given the two different events, $\{B_r = x, B_s = y\}$ and $\{B_s = y\}$.

For the first, we set $X := B_t$ and $Y := (B_r, B_s)$ and apply the conditioning formulas. Note

$$\Sigma_Y = \begin{pmatrix} r & r \\ r & s \end{pmatrix} \quad \Rightarrow \quad \Sigma_Y^{-1} = \frac{1}{r(s-r)} \begin{pmatrix} s & -r \\ -r & r \end{pmatrix},$$

so we have

$$\begin{aligned} \tilde{\mu} &= \mu_X + \Sigma_{XY}\Sigma_Y^{-1}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \mu_Y \right) \\ &= \frac{1}{r(s-r)} \begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} s & -r \\ -r & r \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \frac{1}{r(s-r)} \begin{pmatrix} rs - rs & -rs + -r^2 + rs \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= y \end{aligned}$$

and

$$\begin{aligned} \tilde{\sigma}^2 &= t - \frac{1}{s(t-s)} \begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} s & -r \\ -r & r \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} \\ &= t - \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} \\ &= t - s. \end{aligned}$$

Therefore, $(B_t \, B_r = x, B_s = y) \sim \mathcal{N}(y, t-s)$.

For the second, we set $X := B_t$ and $Y := B_s$ and compute

$$\tilde{\mu} = \mu_X + \sigma_{XY}^2 \frac{y - \mu_Y}{\sigma_Y} = s \cdot \frac{y}{s} = y$$

and

$$\tilde{\sigma}^2 = t - \frac{(\sigma_{XY})^2}{\sigma_Y^2} = t - \frac{s^2}{s} = t - s.$$

Therefore, $(B_t \, B_s = y) \sim \mathcal{N}(y, t-s)$. Since these agree, the result is complete.

PF: (via random walk) Let $\{S_n\}_{n \geq 0}$ denote a SSRW, and compute:

$$\begin{aligned} &\mathbb{P}(B_t = a \,|\, B_r = x, B_s = y) \\ &= \frac{\mathbb{P}(B_t = a, B_r = x, B_s = y)}{\mathbb{P}(B_r = x, B_s = y)} && \text{(definition of conditional probability)} \\ &= \lim_{N \to \infty} \frac{\mathbb{P}(S_{Nt} = a\sqrt{N}, S_{Nr} = x\sqrt{N}, S_{Ns} = y\sqrt{N})}{\mathbb{P}(S_{Nr} = x\sqrt{N}, S_{Ns} = y\sqrt{N})} && \text{(Donsker)} \end{aligned}$$

$$
\begin{aligned}
&= \lim_{N\to\infty} \mathbb{P}(S_{Nt} = a\sqrt{N} \,|\, S_{Nr} = x\sqrt{N}, S_{Ns} = y\sqrt{N}) && \text{(definition of conditional probability)}\\
&= \lim_{N\to\infty} \mathbb{P}(S_{Nt} = a\sqrt{N} \,|\, S_{Ns} = y\sqrt{N}) && \text{(Markov property of SSRW)}\\
&= \lim_{N\to\infty} \frac{\mathbb{P}(S_{Nt} = a\sqrt{N}, S_{Ns} = y\sqrt{N})}{\mathbb{P}(S_{Ns} = y\sqrt{N})} && \text{(definition of conditional probability)}\\
&= \frac{\mathbb{P}(B_t = a, B_s = y)}{\mathbb{P}(B_s = y)} && \text{(Donsker)}\\
&= \mathbb{P}(B_t = a \,|\, B_s = y). && \text{(definition of conditional probability)}
\end{aligned}
$$

This finishes the proof.

BM satisfies, in fact, a stronger form of the Markov property in which the times $0 \le r \le s \le t$ are allowed to be random, provided that they do not "see into the future" (e.g. first passage times). However, we will not pursue this during this class.

Since BM satisfies the Markov property, it is natural to ask all of the questions that we studied during the first part of the course. For example, transience/recurrence, existence of stationary and limiting distributions, first-transition analysis, etc.

EX (no limiting distribution)**.** Let's say that a probability density g on $\mathbb{R}$ is a *limiting distribution for BM* if for all $a < b$ we have

$$\lim_{t\to\infty} \mathbb{P}(a \le B_t \le b) = \int_a^b f(x)\,\mathrm{d}x.$$

In fact, we can show that no limiting distribution exists. To do this, just compute

$$\lim_{t\to\infty} \mathbb{P}(a \le B_t \le b) = \lim_{t\to\infty} \mathbb{P}\left(\frac{a}{\sqrt{t}} \le \mathcal{N}(0,1) \le \frac{b}{\sqrt{t}}\right) = \lim_{t\to\infty} \mathbb{P}\left(\mathcal{N}(0,1) = 0\right) = 0.$$

This implies that $g(x) = 0$ for all x, but then g cannot be a probability density.

EX (first-transition analysis)**.** While first-transition analysis for MCs reduced our probabilistic quantities of interest into systems of linear equations, we will see that first-transition analysis for BM reduces our probability quantities of interest into differential equations. One important example is as follows: Fix $a > 0$, and define $T_a := \min\{t \ge 0 : B_t \in \{-a, a\}\}$ as the first exit time from the region $(-a, a)$. Our goal is to compute $\mathbb{E}[T_a]$.

To do this, recall that in HW5 we did this for the SSRW (with a slightly different but equivalent parameterization). That is, for $K > 0$ we defined $v_i := \mathbb{E}[T_K \,|\, S_0 = i]$, and we determined the following system of equations for v:

$$\begin{cases} v_i &= 1 + \frac{1}{2}v_{i-1} + \frac{1}{2}v_{i-1} \text{ for } -K < i < K \\ v_{-K} &= 0 \\ v_K &= 0 \end{cases}$$

In order to pass to the limit, let us take $K = a\sqrt{N}$ and define the function $V : [-a, a] \to \mathbb{R}$ via $V(i/\sqrt{N}) = v_i$. Then the system above rearranges to

$$\begin{cases} V\left(\frac{i-1}{\sqrt{N}}\right) + V\left(\frac{i+1}{\sqrt{N}}\right) - 2V\left(\frac{i}{\sqrt{N}}\right) = -2 \quad \text{for } -a < i < a \\ v_{-a} = 0 \\ v_a = 0 \end{cases}$$

which, as $N \to \infty$, converges to

$$\begin{cases} V''(x) = -2 \quad \text{for } -a < x < a \\ V(-a) = 0 \\ V(a) = 0 \end{cases}$$

But this is just an ODE that we can solve!

Indeed, note that $V''(x) = 2$ implies that we must have $V(x) = -x^2 + bx + c$ for some $b, c \in \mathbb{R}$. Then $V(-a) = V(a) = 0$ implies $b = -a$ and $c = a^2$. This implies $V(x) = a^2 - x^2$, hence

$$\mathbb{E}[T_a] = a^2.$$

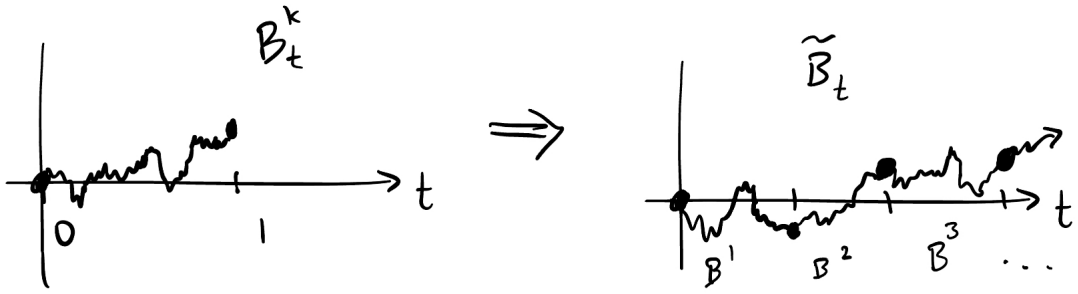
Note that this agrees with our calculation from the SSRW case.

It also natural to wonder whether BM has infinitesimal generator, since this was a key tool for computations involving CTMCs. Since BM has a continuous state space, it is not clear what an infinitesimal generator should be (but of course it cannot be a matrix). It turns out that there is a precise sense in which the infinitesimal generator of BM is the second-order differential operator $L := \mathrm{d}^2/\,\mathrm{d}x^2$, but we will not pursue it in this class.

6.6 Construction of Brownian motion

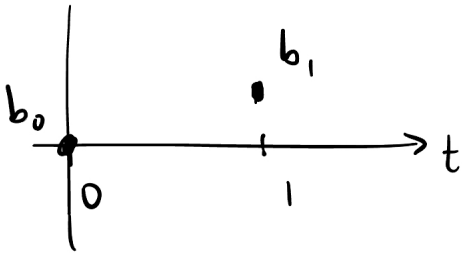
There is an important mathematical question that we have not addressed so far: How do we know that there exists a stochastic process satisfying the definition we have given for BM? We have already seen that we can use general theory of GPs to ensure existence, but it is also of interest to see how to construct BM directly. In this section we will sketch the proof of this construction (mostly in pictures) but details can be found in most textbooks on measure-theoretic probability.

Step 1. We claim that it suffices to construct BM just for the time interval $[0, 1]$, denoted $\{B_t\}_{0 \le t \le 1}$. Indeed, suppose we have done so, and let us show how to use this to construct BM for all time, denoted $\{B_t\}_{t \ge 0}$. First, let $\{B_t^1\}_{0 \le t \le 1}, \{B_t^2\}_{0 \le t \le 1}, \ldots$ denoted IID copies of $\{B_t\}_{0 \le t \le 1}$. Then define the process $\{\tilde{B}_t\}_{t \ge 0}$ by "stitching together" these processes to form a continuous sample path. This is best understood in the following picture:



By the translation-invariance property we studied before, one can show that $\{\tilde{B}_t\}_{t \ge 0}$ is a BM. Thus, it suffices to focus on constructions for time $0 \le t \le 1$.

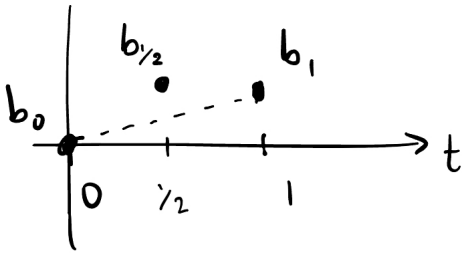
Step 2. We show how to construct B_t for times $t \in D \subseteq [0, 1]$, where $D := \{k2^{-n} : n \ge 0 \text{ and } 0 \le k \le 2^n\}$. We do this via an iterative procedure, as follows. First, it is easy to define the joint distribution of (B_0, B_1) since $B_0 = 0$ is non-random, and $B_1 \sim \mathcal{N}(0, 1)$. So we can simply sample (b_0, b_1) from this joint distribution, whch we visualize as follows:



Now, how can we update this picture to include the value $B_{1/2}$? We need to compute the conditional distribution

$$B_{1/2} \quad \text{given} \quad \{(B_0, B_1) = (b_0, b_1)\}.$$

This is just a Gaussian conditioning problem, which we know how to do from the previous part of the class; the result is just $\mathcal{N}(\frac{1}{2}b_1, \frac{1}{4})$, which is a Gaussian distribution whose mean is the midpoint of the existing samples. So, we add this point to the picture, yielding:



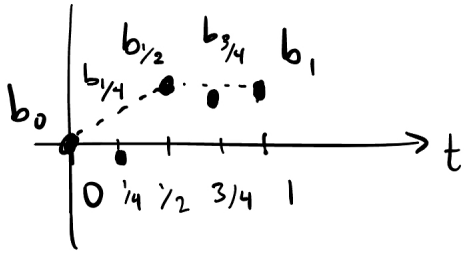
Next, we want to include the values $B_{1/4}$ and $B_{3/4}$. That is, we need to compute the conditional distributions

$$\begin{aligned} B_{1/4} \quad &\text{given} \quad \{(B_0, B_{1/2}, B_1) = (b_0, b_{1/2}, b_1)\} \\ B_{3/4} \quad &\text{given} \quad \{(B_0, B_{1/2}, B_1) = (b_0, b_{1/2}, b_1)\}. \end{aligned}$$

Here is an important observation: by the Markov property, these are, respectively, equivalent to the following conditional distributions

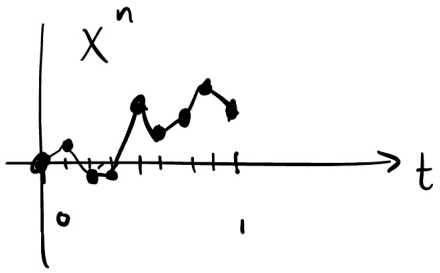
$$\begin{aligned} B_{1/4} \quad &\text{given} \quad \{(B_0, B_{1/2}) = (b_0, b_{1/2})\} \\ B_{3/4} \quad &\text{given} \quad \{(B_{1/2}, B_1) = (b_{1/2}, b_1)\}. \end{aligned}$$

In other words, only the previously-sampled points at nearby times are needed for sampling a point at a new time; the conditional means are the midpoints of these adjacent values, and the conditional variances are equal to 1/8. So we can sample them and get the next picture:



We can continue in this way for all steps n. Interestingly, note that the number of terms involved in the conditioning at the nth step is 2^n, but the Markov property allows us to reduce the number of terms in the conditioning to just 2. (So, we save exponential amounts of computation, by the Markov property.)

Now for each $n \geq 0$ we let $\{X_t^n\}_{t \geq 0}$ be the linear interpolation of the values which have been determined by step n. For example, for $n = 3$ we have



and $\{X_t^{n+1}\}_{t \geq 0}$ "refines" $\{X_t^n\}_{t \geq 0}$ for all $n \geq 0$.

Step 3. We show how to take the limit of the processes $\{X_t^n\}_{t \geq 0}$ as $n \to \infty$. First of all, note what happens if t is of the form $t = k2^{-N}$ for some N: for $n \geq N$, the value X_t^n does not change. In particular, we may define the stochastic process $\{X_t\}_{t \in D}$ via

$$X_t := \lim_{n \to \infty} X_t^n.$$

However, we still need to extend this from D to $[0, 1]$. This requires more real analysis than we have in this class, so for now let's just recall that a uniformly continuous function on a dense set can be uniquely extended to a continuous function on the closure of its domain. Thus, we just need to show that $\mathbb{P}(\{X_t\}_{t \in D}$ is uniformly continuous$) = 1$. This can be done using the reflection principle, although we won't discuss the details here. Then, we let $\{B_t\}_{0 \leq t \leq 1}$ be this (uniquely-defined) extension of $\{X_t\}_{t \in D}$.

Step 4. To finish the construction, we need to check that $\{B_t\}_{0 \leq t \leq 1}$ is indeed a Brownian motion. By design, it has $B_0 = 0$ and continuous sample paths. Also, it is not hard to show that $\{B_t\}_{0 \leq t \leq 1}$ is indeed a Gaussian processes with covariance function $\Sigma(s, t) = \min\{s, t\}$. Therefore, the proof is complete.

6.7 Variations on Brownian Motion

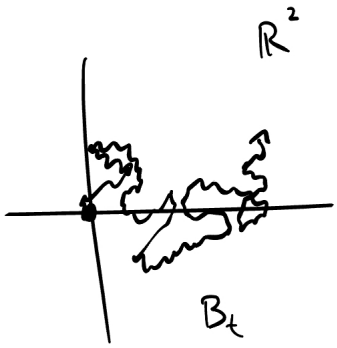
There are many ways to generalize Brownian motion, and we discuss a few of them now.

DEF. For positive integer d, a *d-dimensional Brownian motion* is an $\mathbb{R}^d$-valued stochastic process $\{B_t\}_{t \geq 0}$ with $B_0 = 0$ and posessing continuous sample paths, satisfying

(1) $B_t - B_s$ has distribution $\mathcal{N}(0, (t - s)I_d)$ for all $0 \leq s \leq t$, and

(2) $B_t - B_s$ is independent of B_s for all $0 \leq s \leq t$.

This is entirely analogous to the case of univariate BM, except now the increments are multivariate Gaussian distributions. This process also has an interpretation as the movement of a particle in $\mathbb{R}^d$, which we can visualize as follows:



You may notice some similarity with the particle in a viscous fluid example from earlier in the course. It turns out that there is indeed a Donsker-type theorem for multidimensional BM which you can use to show that a suitable scaling limit of the particle in a viscous fluid converges to a 2-dimensional BM.

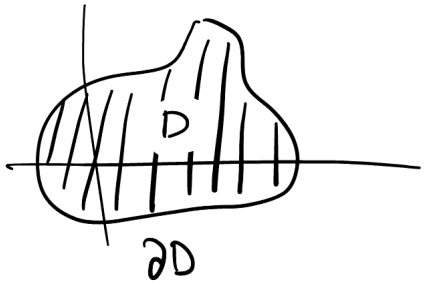
Although multidimensional BM seems a bit complicated, we can see in the following that it is actually not much more complicated than the usual univariate BM:

LEM. *A $\mathbb{R}^d$-valued stochastic process $\{B_t\}_{t\geq 0}$ is a d-dimensional Brownian motion if and only if the coordinate processes $\{B_t^1\}_{t\geq 0}, \ldots, \{B_t^d\}_{t\geq 0}$ defined via $B_t = (B_t^1, \ldots, B_t^d)$ are independent univariate Brownian motions.*

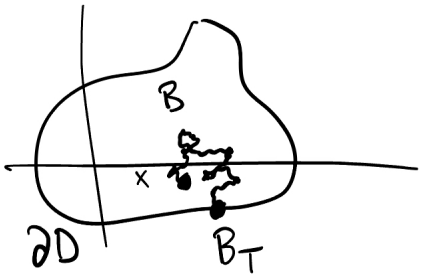
We do not give the proof here, but it is intuitive for the following reasons: The increments have covariance matrix proportional to the identity, and we know that uncorrelated jointly Gaussian processes must be independnent.

A qualitative difference between multidimensional BM and univariate BM is the geometric structure of first passage time problems. Recall that, in the univariate case, first passage problems are simple since entering or existing a region (i.e., an interval) is only possible from one or two locations. But in higher dimensions this is not true:

EX (location at first passage time for 2-dimensional BM)**.** Write $D \subseteq \mathbb{R}^2$ for some domain and $\partial D \subseteq \mathbb{R}^2$ for its boundary. We can visualize this as follows:



If $\{B_t\}_{t\geq 0}$ is a 2-dimensional BM started from $B_0 = x$ on the interior of D, then what is B_T where $T := \min\{t \geq 0 : B_t \in \partial D\}$ is the first passage time to the boundary? For instance, one sample path may look like the following:



But now note that B_T is a continuous random variable taking values in the boundary ∂D. This is different than we have seen for univariate BM, where the location at first passage time is discrete, since it can only take on 1 or 2 different values.

This idea of running a multivariate BM until it hits the boundary of a given region is very rich. As you will see on HW 14, it can be used to construct solution to some PDEs that are interesting in their own right.

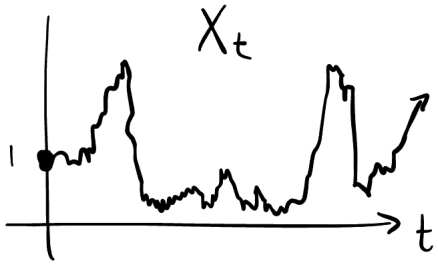
Next we give a non-linear transformation of BM which is an important model of a stock price in mathematical finance:

DEF. For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, a *geometric Brownian motion (GBM)* is a stochastic process $\{X_t\}_{t\geq 0}$ defined via

$$X_t = \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma B_t\right)$$

where $\{B_t\}_{t\geq 0}$ is some Brownian motion.

The sample paths of $\{X_t\}_{t\geq 0}$ look something like the following:



There are a few reasons why GBM is a useful model for an asset price in mathematical finance:

- It takes only non-negative values, which is desirable since real-world assets (usually) cannot have negative value
- It has a "multiplicative" structure rather than an "additive" structure, which is useful because real-world problems involve optimizing *multiplicative* returns on a principal.
- It has rough sample paths like BM, which reflects the empirical volatility observed in real-world asset prices.

We can also do some exact calculations with GBM, because it is an explicit function of a BM. For example, it is possible to compute

$$\mathbb{E}[X_t] = e^{\mu t} \quad \text{and} \quad \text{Var}(X_t) = e^{2\mu t}(e^{\sigma^2 t} - 1)$$

for all $t \geq 0$ (although we won't go through the details in this class). Note the similarity with the PP-based stock price model we studied earlier in the course!

One of the most imporant things about GBM is that, when $\mu = 0$, we have

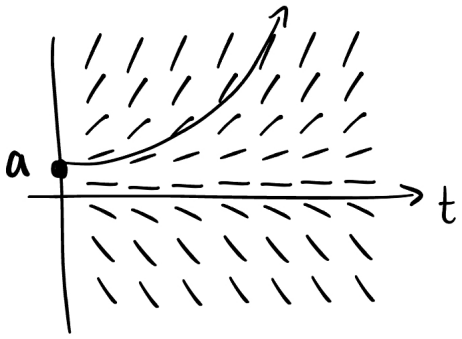
$$\mathbb{E}[X_t \mid X_s = x] = x$$

for all $0 \leq s \leq t$ and all $x \in \mathbb{R}$. (You will see this in detail on HW14). This is similar to a property we found for BM itself, and we interpreted this in a statistical sense: If I know the value of GBM at time s, then my best guess for its value at a future time $t \geq s$ is simply the current value. This is called the *martingale property* and it, roughly speaking, means that financial markets with GBM assets cannot have arbitrage opportunities. If you take further courses in stochastic calculus and mathematical finance, you will study the martingale property in a lot of detail.

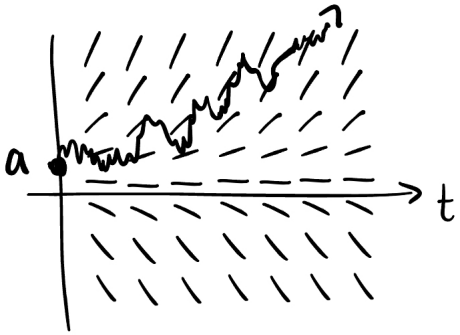
Our last goal is to study differential equations which involve Brownian motion. To do this, let's review some basic aspects of ordinary differential equations (ODEs). Recall that by an ODE we usually mean the equation

$$\begin{cases} \frac{\mathrm{d}x_t}{\mathrm{d}t} = \mu(x_t) & \text{for } t > 0 \\ x_0 = a \end{cases}$$

for some function $\mu : \mathbb{R} \to \mathbb{R}$ and some constant $a \in \mathbb{R}$. This has a physical interpretation: the function μ determines a (time-homogeneous) velocity field, and a path $\{x_t\}_{t \geq 0}$ is a solution to the ODE above if it satisfies the initial condition and always stays tangent to this velocity field. We can visualize this as follows:



Our goal is to understand what is means to have a "random version" of this ODE; this should mean, for example that a solution is a stochastic process $\{X_t\}_{t \geq 0}$ whose average behavior is like $\{x_t\}_{t \geq 0}$ but "perturbed independently" at each time step. For example, we want something like this:



However, it will take some work to define this mathematically.

To get there, let's recall that there are three equivalent ways we can define the ODE above. They are

$$\text{derivative formulation:} \quad \begin{cases} \frac{\mathrm{d}x_t}{\mathrm{d}t} = \mu(x_t) & \text{for } t > 0 \\ x_0 = a \end{cases}$$

$$\text{differential formulation:} \quad \begin{cases} \mathrm{d}x_t = \mu(x_t)\,\mathrm{d}t & \text{for } t > 0 \\ x_0 = a \end{cases}$$

$$\text{integral formulation:} \quad \begin{cases} x_t - x_0 = \int_0^t \mu(x_s)\,\mathrm{d}s & \text{for } t > 0 \\ x_0 = a \end{cases}.$$

These all have some advantages and disadvantages. For example, in the derivative formulation it is quite easy to see the picture in terms of velocity fields. But the differential formulation tells us how to make a numerical approximation scheme, by replacting $\mathrm{d}t$ with $t_i - t_{i-1}$ and $\mathrm{d}x_t$ with $x_{t_i} - x_{t_{i-1}}$ for some fine grid $0 \le t_1 \le t_2 \le \cdots$ approximating $[0, \infty)$. Lastly, the integral formulation is interesting because it does not require the path $\{x_t\}_{t \ge 0}$ to be differentiable!

Now we come to the main definition:

DEF. A *stochastic differential equation (SDE)* is an equation of the form

$$\begin{cases} \mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t \text{ if } t > 0 \\ X_0 = a \end{cases}$$

where $\mu : \mathbb{R} \to \mathbb{R}$ is called the *drift function* and $\sigma^2 > 0$ is called the *diffusion coefficient*, and $\{B_t\}_{t \ge 0}$ is a Brownian motion. We say that a stochastic process $\{X_t\}_{t \ge 0}$ *is a solution to the stochastic differential equation (SDE)* if we have if we have

$$X_t - a = \int_0^t \mu(X_s)\,\mathrm{d}s + B_t$$

for all $t \ge 0$.

Although SDEs are usually presented in the differential formulation, we must actually work with the integral formulation if we want to be mathematically precise. This is because solutions to SDEs will always have a roughness which is locally like BM, hence they will never be differentiable.

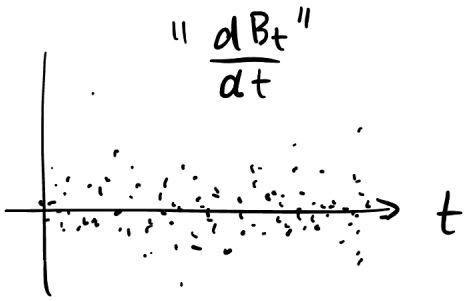
The definition above connects the differential formulation and the integral formulation, but what is the derivative formulation? Heuristically, it says $\{X_t\}_{t \ge 0}$ is a solution to the SDE above if we have

$$\begin{cases} \frac{\mathrm{d}X_t}{\mathrm{d}t} = \mu(X_t) + \frac{\mathrm{d}B_t}{\mathrm{d}t} & \text{for } t > 0 \\ X_0 = a \end{cases}.$$

But, what is the stochastic process $\{\frac{\mathrm{d}B_t}{\mathrm{d}t}\}_{t \ge 0}$? On the one hand, we know that this does not exist since the BM sample paths are not differentiable. On the other hand, we know from general theory of GPs that if this process exists (which is does not!) then it would have covariance function

$$\Sigma(s, t) = \mathbf{1}\{s = t\}.$$

In other words $\{\frac{\mathrm{d}B_t}{\mathrm{d}t}\}_{t \ge 0}$ is a "continuum of IID samples from $\mathcal{N}(0, 1)$". You should think of its sample paths like this:



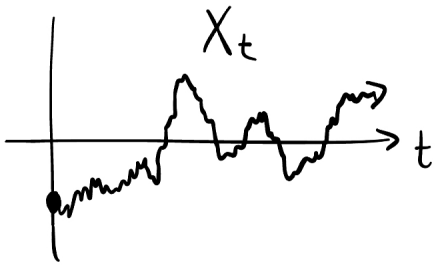
In statistical signal processing this is called a "white noise" and there is a saying that "the derivative of Brownian motion is white noise". But that is just a heuristic way of explaining the correspondence between the integral and derivative formulations of an SDE.

Now we can give some examples to make this all a bit more concrete.

EX (OU process)**.** For $\gamma > 0$, consider the SDE

$$\begin{cases} \mathrm{d}X_t = -\gamma X_t \,\mathrm{d}t + \,\mathrm{d}B_t \text{ if } t > 0 \\ X_0 = A \end{cases}$$

where $A \sim \mathcal{N}(0,1)$. It turns out that this SDE has a unique solution given by the OU process:

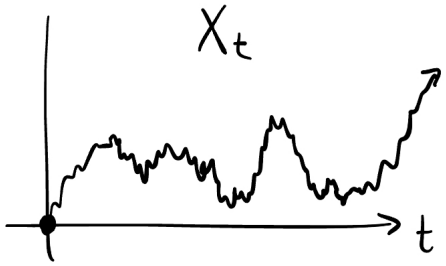


We have already seen (via scaling limits) that an OU process is mean-reverting since it is a limit of AR(1) processes that by definition are mean-reverting. But this SDE formulation gives us another reason why it mean-reverting: It has drift function $\mu(x) = -\theta x$ which means the average velocity points to state 0, and the strength of the velocity increases when X_t is far away.

EX (Bessel process)**.** For $c > 0$, consider the SDE

$$\begin{cases} \mathrm{d}X_t = \frac{c}{X_t} \,\mathrm{d}t + \,\mathrm{d}B_t \text{ if } t > 0 \\ X_0 = 0 \end{cases}$$

which has drift function $\mu(x) = c/x$; this means we have average velocity pushing away from state 0, and its strength is inversely proportional to the size of X_t. In particular, the drift is negligible for large X_t hence the process looks like a Brownian motion, and the drift is extremely strong for small X_t. So the sample paths look like this:



Note that "instantaneous" escape from state 0 at time $t = 0$.

Lastly, we discuss SDEs whose diffusion is not a coefficient, but rather a function of the current state That is, we aim to solve

$$\begin{cases} \mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t \text{ if } t > 0 \\ X_0 = a \end{cases}$$

for some drift function $\mu : \mathbb{R} \to \mathbb{R}$, some diffusion function $\mu : \mathbb{R} \to [0, \infty)$, and some $a \in \mathbb{R}$. One reason for doing this is financial; the factor $\sigma(X_t)$ represents the instantaneous volatility of $\{X_t\}_{t\geq 0}$ and it is sometimes desirable that this should depend on the current value of an asset. For instance, GBM is represented by the SDE

$$\begin{cases} \mathrm{d}X_t = \mu X_t\,\mathrm{d}t + \sigma X_t\,\mathrm{d}B_t \text{ if } t > 0 \\ X_0 = 1 \end{cases}.$$

However, making sense of this is a bit difficult. We should of course translate it into integral form as follows:

$$X_t - a = \int_0^t \mu(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}B_s.$$

But now we need to assign a meaning to the term

$$\int_0^t \sigma(X_s)\,\mathrm{d}B_s.$$

This is like "integrating a stochastic process with respect to BM" and is usually called the *Ito integral*. If you take further courses in stochastic processes then you will precisely define this, and you will see that you can do all kinds of calculus with Brownian motion, usually referred to as *Ito calculus*.