1690 Lecture notes Fall 2015

Instructor: Nicolás García Trillos

This lecture notes are based on our APMA 1690 lectures, but they are also based on the discussion of two fictional characters Solo and Pacifica . They could be anyone or anything at any point in time. Hopefully you will see how they evolve during the semester.

1. Random number generation.

1.1. **Multiplicative Congruential Generator (Optional) .** -S :The goal is to generate an i.i.d. (independent identically distributed) sample $U_1, \dots, U_n$ from the uniform distribution on the interval $[0,1]$. How do I do that?

-P: Patience Solo. Maybe a little background first?

Let us first recall some basic facts about the uniform distribution. First, the uniform distribution on the interval $[0,1]$ (denoted by $Uniform([0,1])$), is the continuous probability distribution with density $\mathbb{1}_{[0,1]}$, that is, the density is equal to 1 on the interval $[0,1]$ and it is equal to zero outside the interval $[0,1]$.

Now let us consider U a random variable with distribution $Uniform([0,1])$ (we write $U \sim Uniform([0,1])$). The cumulative distribution function (c.d.f.) of U is given by

$$F_U(u) := \mathbb{P}(U \leq u) = \int_{-\infty}^{u} \mathbb{1}_{[0,1]}(s)ds, \quad u \in \mathbb{R}.$$

Thus,

$$F_U(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ u & \text{if } u \in [0,1] \\ 1 & \text{if } 1 \leq u. \end{cases} \tag{1}$$

Remark 1.1. Let $c \in \mathbb{R}$ and let us consider U uniformly distributed on $[0,1]$. Since the distribution of U has a density, we must have,

$$\mathbb{P}(U = c) = 0,$$

indeed,

$$\mathbb{P}(U = c) = \int_c^c \mathbb{1}_{[0,1]}ds = 0.$$

Note that in the above we have used the fact that $\int_c^c f(x)dx = 0$ independently of what c is. Let us rephrase the previous mathematical statement as follows: for a fixed value $c \in \mathbb{R}$ (that we have fixed **before** seeing the outcome of U), the probability that the outcome of U is equal to the chosen number c is equal to zero. In particular, if we fix the values $c = 0$ and $c = 1$, and we consider $U \sim Uniform([0,1])$, the probability that $U = 0$ or $U = 1$ is zero. Thus, sampling from the uniform distribution on the interval $[0,1]$ or from the interval $(0,1)$ is exactly the same thing.

For a function $h : \mathbb{R} \to \mathbb{R}$, we can compute things like $\mathbb{E}(h(U))$ by using U's density. Indeed,

$$\mathbb{E}(h(U)) = \int_{-\infty}^{\infty} h(u)\mathbb{1}_{[0,1]}(u)du = \int_0^1 h(u)du.$$

S: Ok, I got it. Now what we are here for... How to generate samples from the uniform distribution?.

P:Well we may take two approaches.

The first one: We may consider *true* samples from the uniform distribution. But true randomness can not be generated in your computer! Think about it: if your computer could

generate true randomness, then your computer wouldn't be very reliable, right? Would you store your more precious data in it?. True randomness comes from physical phenomena like particles decaying, particles hitting earth's atmosphere, etc. Take a look at the webpage *random.org*, in particular go to section "Learn about Randomness". There are some interesting things there.

The second one: Since we are interested in simulating randomness with the tools we have at hand, the other possibility for generating randomness is by using your computer (or to hire people who are willing to flip coins every time you need an experiment!). But as we said before, computers can not create randomness. In what follows, we will explore a method called the *Multiplicative Congruential Generator*. This method provides us with "random" samples from the uniform distribution. We say "random" because the numbers we generate are actually deterministic.

S: Wait a second... but didn't you say we wanted to generate *random* samples from the uniform distribution? And now you are saying that your samples are going to be deterministic?

P: Well, yes. Suppose that someone gives you a list of numbers $u_1, u_2, u_3, \ldots, u_n$. Maybe these numbers were chosen randomly or "randomly". But at the end of the day, what do we care if they were randomly chosen or deterministically chosen?... as long as they *behave as if they had been chosen randomly*, they should be good enough. Right?

S: I guess if the numbers are good for our purposes...

P : That's about right dear Solo: tell me how you behave and I will tell you who you are.

Modular Arithmetic.

In order to define the *Multiplicative Congruential Generator*, we need some definitions first.

Definition 1.2. *Let m be a natural number. Let $x \in \mathbb{Z}$ be an integer (positive or negative). The number x mod$_m$ is the number r belonging to $\{0, 1, \ldots, m-1\}$ for which there exists $k \in \mathbb{Z}$ such that*

$$x = km + r.$$

In other words, x mod$_m$ is the remainder after dividing x by m.

Examples:

(1) $8 \bmod_5 = 3$
(2) $101 \bmod_5 = 1$
(3) $49 \bmod_6 = 1$
(4) $8 \bmod_9 = 8$

We may also define a type of equality called *congruence*.

Definition 1.3. *Let m be a natural number. Let $a, b \in \mathbb{Z}$. We say that*

$$a \equiv_m b$$

(to be read a is congruent to b modulo m), if there exists $k \in \mathbb{Z}$ such that

$$a = b + km.$$

Equivalently, $a \equiv_m b$ if a mod$_m = b$ mod$_m$.

Examples

(1) $8 \equiv_3 2$
(2) $6 \equiv_2 0$

(3) $1 \equiv_7 15$
(4) $16 \equiv_9 7$
(5) $m \equiv_m 0$

Multiplicative Congruential Generator: There are three parameters that we need:

(1) m : a (large) prime number.
(2) R_0 : a seed; a number that belongs to $\{1, \ldots, m-1\}$.
(3) a : a multiplier; a number that belongs to $\{2, \ldots, m-1\}$.

Given the parameters m, R_0, a, we define the sequence of numbers $R_1, R_2, \ldots$ inductively:

$$\begin{aligned} R_1 &:= (aR_0) \bmod_m \\ R_2 &:= (aR_1) \bmod_m \\ &\vdots \\ R_k &:= (aR_{k-1}) \bmod_m. \end{aligned}$$

Note that by definition of the $\bmod_m$ operation, each of the R_k is a number in the set $\{0, \ldots, m-1\}$. Below we show that R_k never attains the value 0.

Proposition 1.4. *Let m, R_0, a be parameters as defined previously. Then, for every $k \in \mathbb{N}$, R_k is **not** equal to zero.*

Proof. The proof is by induction on k.

Base Case: By assumption, R_0 is a number that belongs to $\{1, \ldots, m-1\}$ and so $R_0 \neq 0$.

Inductive Step: Suppose that $R_k \neq 0$ we want to show that $R_{k+1} \neq 0$. To see this, observe that by definition $R_{k+1} = (aR_k) \bmod_m$, so that in particular R_{k+1} is a number in $\{0, \ldots, m-1\}$. Because of this, to show that $R_{k+1} \neq 0$, it is enough to show that $R_{k+1} \not\equiv_m 0$ (go back to the definition of $\bmod_m$ and convince yourself that this is the case). Now, using the definition of $\equiv_m$, it follows that

$$R_{k+1} \equiv_m aR_k.$$

If it was true that $R_{k+1} \equiv_m 0$, then the previous identity would imply that

$$0 \equiv_m aR_k.$$

By definition, we would conclude that there exists $s \in \mathbb{Z}$ such that

$$aR_k = sm,$$

that is, aR_k is a multiple of m. However, note that the right hand side has m as a prime factor (recall that m was chosen to be a prime number), and the left hand side does not have m as prime factor because both a and R_k are less than m. Since the factorization of a positive integer number into prime factors is unique, this is impossible. Thus, it can not be true that $R_{k+1} \equiv_m 0$, or in other words it has to be true that $R_{k+1} \not\equiv_m 0$. □

Examples:

(1) Consider $m = 7$, $R_0 = 5$ and $a = 3$. You can check that

$$\begin{aligned} R_0 &= 5 \\ R_1 &= 1 \\ R_2 &= 3 \\ R_3 &= 2 \\ R_4 &= 6 \end{aligned}$$

$$R_5 = 4$$
$$R_6 = 5$$

From then on, the sequence becomes boring: it starts repeating itself ... Note however that all numbers $\{1, \ldots, m-1\}$ appear in the above sequence of numbers. Because of this, we say that the MCG is *full cycle*.

(2) Consider $m = 7$, $R_0 = 5$ and $a = 2$. You can check that

$$R_0 = 5$$
$$R_1 = 3$$
$$R_2 = 6$$
$$R_3 = 5$$

From then on, the sequence repeats itself... In contrast with the previous example, not all numbers $\{1, \ldots, m-1\}$ appear in the sequence. In this case, we say that the MCG is *not full cycle*.

Remark 1.5. Given a MCG, with parameters m, R_0, a, the property of being full cycle is independent of the seed considered. That is, if the MCG with parameters m, R_0, a is full cycle for *some* seed R_0, then it has to be full cycle for *every* seed R_0. Why?

From the above remark, we conclude that a full cycle MCG depends exclusively on properties of a and m. It turns out that a MCG is full cycle if and only if a is a *primitive root modulo* m.

Definition 1.6. *Let m be a prime number. We say that $a \in \{2, \ldots, m-1\}$ is a primitive root modulo m if*

$$a^2 \not\equiv_m 1, a^3 \not\equiv_m 1, \ldots, a^{m-2} \not\equiv_m 1$$

and

$$a^{m-1} \equiv_m 1.$$

In other words, if the first integer k for which $a^k \equiv_m 1$ is $k = m-1$.

Notice that 3 is a primitive root modulo 7: no surprise that the MCG with parameters $m = 7$,$R_0 = 5$, $a = 3$ was full cycle. On the other hand note that 2 is not a primitive root modulo 7 : no surprise that the MCG with parameters $m = 7$, $R_0 = 5$, $a = 2$ is not full cycle.

So far we have introduced the notion of *Multiplicative Congruential Generator*. Having chosen the parameters m, R_0, a, we can obtain a sequence of numbers $R_1, R_2, \ldots$. We define our "random" samples (from now on we stop using the term "random" and instead we use pseudo-random) from the uniform distribution on $[0, 1]$ by taking:

$$U_0 = \frac{R_0}{m}$$
$$U_1 = \frac{R_1}{m}$$
$$\vdots$$
$$U_k = \frac{R_k}{m}$$
$$\vdots$$

S: The numbers $U_0, U_1, U_2, \ldots$ are the numbers we say are "random" samples from $Uniform([0, 1])$? I mean pseudo-random?

P: Well yes, but provided we make a good choice for the parameters m and a. First, m has to be a large prime number (7 is not large). Second, if the MCG considered is not full cycle, we should not be using such MCG. In that case the sequence $U_0, U_1, \dots$ obtained will not quite behave like true random samples from the uniform distribution. Why? Think about it!

The bottom line is that we should consider MCGs that are full cycle. But even if we know that we should only consider MCGs that are full cycle, we still have work to do. We still need to learn how to evaluate if a given sequence of numbers behaves like a random sample, or not...

S: ... MORE work to do...

1.2. **Evaluating random number generators.** -P: Did you know that some years ago you could buy *true* random numbers? For example, a single random integer between 1 and 4000 costed $\$28.95$.

S: With that money I could buy a train ticket + a slice of pizza for 5 days in New York City.

P: Fortunately, nowadays you can find webpages that give you true random numbers for free.

S: ...meanwhile, train tickets become more expensive...

P: Death and taxes... and inflation too. When it comes to random numbers, free or not, we should be able to check their quality. They should behave like true random samples would.

The law of large numbers.

Let us recall briefly some properties about expectation and variance.

Proposition 1.7. *Given two random variables X, Y, and given numbers $a, b \in \mathbb{R}$, we have*

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y);$$

in other words, expectation is a linear operator. If in addition X, Y are uncorrelated, then

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y).$$

Remark 1.8. Recall that if X, Y are independent, in particular they are uncorrelated, but the converse statement is false! Make sure you know what each notion means.

Of course the previous statement extends to finite sums:

$$\mathbb{E}(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \mathbb{E}(X_i);$$

and

$$\operatorname{Var}(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \operatorname{Var}(X_i);$$

for the relation involving variance, we need to impose that the variables are pairwise uncorrelated.

In what follows, we consider $X_1, \dots, X_n, \dots$ to be a sequence of independent identically distributed random variables (they do not have to be uniformly distributed). The law of large numbers is one of the most basic, and at the same time most fundamental laws obeyed by i.i.d. samples.

Theorem 1.9 (Weak law of large numbers)**.** *Suppose that $X_1, \dots, X_n, \dots$ is a sequence of i.i.d samples from some distribution on $\mathbb{R}$. Suppose that this distribution is such that its mean is μ and its (finite) variance is σ^2, that is,*

$$\mathbb{E}(X_1) = \mu$$

and

$$\mathrm{Var}(X_1) := \mathbb{E}((X_1 - \mu)^2) = \sigma^2.$$

Then, for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \varepsilon \right) = 0.$$

Proof. First observe that

$$Y_n := \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu);$$

(whenever I write $:=$, it means that I am defining something; in this case, I am calling Y_n the expression $\frac{1}{n}\sum_{i=1}^{n} X_i - \mu$). Let us compute the expectation of Y_n. By linearity of $\mathbb{E}$:

$$\mathbb{E}(Y_n) = \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}(X_i) - \mu) = 0$$

Now let us compute the variance of Y_n; we use the independence of the X_i (which in particular implies they are uncorrelated) to obtain:

$$\mathrm{Var}(Y_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i - \mu) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Now we make use of Chebyshev's inequality, which states that for any given $\varepsilon > 0$ we have:

$$\mathbb{P}\left(|Y_n - \mathbb{E}(Y_n)| > \varepsilon\right) \leq \frac{\mathrm{Var}(Y_n)}{\varepsilon^2}$$

In this case we deduce that

$$\mathbb{P}\left(|Y_n| > \varepsilon\right) \leq \frac{\mathrm{Var}(Y_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon}.$$

The right hand side of the above expression goes to zero as n goes to infinity, whereas the left hand side is bounded below by zero. The squeeze theorem implies the desired result. □

Remark 1.10. If you are interested in learning why the above theorem is called *weak*, I recommend that you search in Wikipedia "Convergence of random variables". Look for "convergence in probability" and "almost sure convergence". It turns out that in the above theorem, we are stating that $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ. The strong law of large numbers on the other hand, asserts that the convergence occurs in the almost sure sense (which is stronger than convergence in probability). Moreover, in the strong law of large numbers, we do not have to assume that the variance of the random variables are finite (the variance could be infinite); the only assumption is that the expectation is finite. For our purposes, the weak law of large numbers is good enough.

We may now apply our results to the following setting. Suppose that $U_1, U_2, \ldots, U_n, \ldots$, are i.i.d. random variables with distribution $Uniform([0,1])$. Then the sum $\frac{1}{n}\sum_{i=1}^{n} U_i$ should concentrate (as $n \to \infty$) around the mean of U_1, which is

$$\mathbb{E}(U_1) = \int_0^1 u du = \frac{1}{2}.$$

Actually, something stronger holds. For arbitrary function $h : [0,1] \to \mathbb{R}$, it should also be true that $\frac{1}{n}\sum_{i=1}^{n} h(U_i)$ concentrates around

$$\mathbb{E}(h(U_1)) = \int_0^1 h(u) du.$$

This again follows from the law of large numbers applied to the sequence of random variables $h(U_1), h(U_2), h(U_3), \ldots$. Note that the independence of the U_is implies the independence of the $h(U_i)$s.

Observe that the law of large numbers is an asymptotic type result: it says that *in the limit* $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} h(U_i)$ converges to $\int_0^1 h(u)du$, but it does not tell you how large n has to be to see that the quantities are actually close to each other. Using the law of large numbers to test whether a certain sequence $U_1, U_2, \ldots$ behaves like i.i.d. samples or not, is then a qualitative test rather than a quantitative test. By this I mean the following: Suppose you are given numbers $u_1, \ldots, u_n$ that are said to be samples from $Uniform([0,1])$. Now, let us imagine that you compute the quantity

$$\frac{1}{n}\sum_{i=1}^{n} u_i$$

and it turns out to be 0.55. Of course it would be very ingenuous to expect the above sum to be exactly equal to 0.5, but in this case 0.55 "seems" to be close to 0.5. I say "seems" because I can imagine someone thinking that 0.55 is far away from 0.5 ; we have not quantified the error of the approximation and so we can only say subjectively if the quantity $\frac{1}{n}\sum_{i=1}^{n} U_i$ is close or not to what we expected.

S : But even if the numbers $u_1, u_2, \ldots, u_{100}$ were chosen randomly (truly), it is possible that

$$\frac{1}{100}\sum_{i=1}^{100} u_i \leq 0.00001.$$

So in this case the subjective test for evaluating the behavior of the sequence, would suggest that the numbers $u_1, u_2, \ldots, u_{100}$ were not randomly chosen when in fact they were.

P: It is *possible* that such thing happens, but not *probable*. We will have to quantify how unlikely that event is, but we will do that later on. Now, suppose that you are given a sequence of numbers $u_1, u_2, \ldots$ such that for every single continuous function $h : [0,1] \to \mathbb{R}$ it is true that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(u_i) = \int_0^1 h(u) du.$$

Is this enough to say that the sequence $u_1, u_2, \ldots$ does behave like an i.i.d sample from the uniform distribution?

The answer to the above question is no. Let us consider $X_1, X_2, \ldots$ a sequence of i.i.d. random variables with the uniform distribution. Now we consider the following sequence of random variables:

$$X_1, X_1, X_2, X_2, X_3, X_3, \ldots$$

Let us consider $U_1, U_2, \dots$ to be the above sequence. Then, by the law of large numbers (applied to the $h(X_i)$), you can show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(U_i) = \int_0^1 h(u)du.$$

Nevertheless, it should be clear that the sequence $U_1, U_2, \dots$ is not behaving like an i.i.d. sequence (because of the repetitions). The point is that the one dimensional functions h are not capturing the correlation among the variables! The above observation leads us to the following.

Let $U_1, U_2, \dots$ be an i.i.d sequence of uniformly distributed random variables. We consider a positive integer $d \geq 2$ and a function $h : [0,1]^d \to \mathbb{R}$. We can split the sequence $U_1, U_2, \dots$ into blocks of size d and define a sequence $\vec{U}_1, \vec{U}_2, \dots$ of d-dimensional random vectors uniformly distributed on $[0,1]^d$ as follows:

$$\begin{aligned} \vec{U}_1 &:= (U_1, \dots, U_d) \\ \vec{U}_2 &:= (U_{d+1}, \dots, U_{2d}) \\ &\vdots \end{aligned}$$

Then, the law of large numbers guarantees that as n goes to infinity, $\frac{1}{n}\sum_{i=1}^{n} h(\vec{U}_i)$ converges towards

$$\int_{[0,1]^d} h(u)du.$$

Let us see that we can use this extension to say that the "duplicated" sequence $X_1, X_1, X_2, X_2, \dots$ (that we list as $U_1, U_2, \dots$) does not behave like i.i.d. samples from the uniform distribution. Indeed, we could consider $d = 2$ and the function $h(u_1, u_2) := \mathbb{1}_{u_1 > u_2}$. Then, it is clear that for every n

$$\frac{1}{n} \sum_{i=1}^{n} h(\vec{U}_i) = 0.$$

Hence, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(\vec{U}_i) = 0$. Nevertheless, $\int_0^1 \int_0^1 h(u_1, u_2)du_1 du_2 = \frac{1}{2}$. Thus, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(\vec{U}_i) \neq \int_0^1 \int_0^1 h(u_1, u_2)du_1 du_2$.

- S: Say that one of those expensive random number generators gives me the numbers $u_1, \dots, u_n$. How many functions h should I use to be sure that they are really behaving like i.i.d samples ?

- P: I guess you know the answer...

- S: Infinitely many... And I suppose I should try with different dimensions, and different functions to try to capture pairwise correlation, correlations between three , four or six hundred of them. Rather hopeless.

-P: If you wanted to test people the way you intend to test number generators to be sure they behave the way you expect, I suppose you would be very lonely, no, Solo?. Perhaps, it is better to focus on natural things like expectation, pairwise correlation and in general things you care about the most.

1.3. **Hypothesis testing.** -P: There are more structured ways to test whether a sequence of numbers $u_1, \dots, u_n$ behaves like a true sample from the uniform distribution. Let me describe in words what hypothesis testing is about:

We assume that we have a stochastic model for some experiment. We assume that this model is correct, and we consider certain events that serve as tests for the model. Then the

experiment is performed. We look at the outcome of the experiment; we are surprised by the outcome. We quantify how big the surprise is. If this surprise is big enough we reject the hypothesis that the model was correct.

-S: Let us be more concrete. How do we specialize what you just said to our setting: test the behavior of pseudo random number generators?

-P: Suppose M is a machine that claims to provide n numbers that are chosen independently from the uniform distribution. Let H_0 be the null hypothesis: "machine M provides n numbers that are chosen independently from the uniform distribution". We may consider an alternative hypothesis H_A which in this context may simple be : "machine M does not give you numbers that are chosen independently from the uniform distribution."

Let us consider $T : \mathbb{R}^n \to \mathbb{R}$ a function that depends on n variables. For example, we may consider

$$T(x_1, \ldots, x_n) := \sum_{k=1}^{n} \mathbb{1}_{x_k > 0.9}. \tag{2}$$

Now suppose that $U_1, \ldots, U_n$ are i.i.d. random variables uniformly distributed on the interval $[0, 1]$. Then, $T(U_1, \ldots, U_n)$ is a random variable itself and hence it has its own distribution. For example, for the T defined in (2), the variable $T(U_1, \ldots, T_n)$ has distribution $Binomial(n, 0.1)$ (Why?).

Now you ask the machine M to give you n numbers. The machine gives you a list of numbers $u_1, \ldots, u_n$. Then you compute

$$t_{obs} := T(u_1, \ldots, u_n).$$

This is the *observed* value for T.

Now you quantify how big the surprise is. You compute the $p - value$:

$$p := \mathbb{P}(T(U_1, \ldots, U_n) \geq t_{obs}).$$

Finally, suppose you had pre-established a maximum level of tolerance α (a number between 0 and 1 which typically is chosen to be 0.05). If $p \leq \alpha$ then you reject H_0 in favor of H_A.

Example: Suppose $n = 100$ and suppose that T is defined as in (2). Assume that the machine M gives you numbers $u_1, \ldots, u_{100}$ such that $t_{obs} = 18$. The p-value in this case is

$$p = \mathbb{P}(T(U_1, \ldots, U_{100}) \geq 18) = 1 - \mathbb{P}(T(U_1, \ldots, U_{100}) \leq 17).$$

Note that the last term on the right hand side of the above expression is the c.d.f. of a $binomial(100, 0.1)$ random variable, evaluated at 17. With Matlab's help (use the function 'binocdf') we see that the above quantity is approximately equal to 0.01. If your maximum level of tolerance was $\alpha = 0.05$ then you would reject H_0 in this case.

S: How did you choose the function T? Could I have chosen something different?

P: The choice of T is usually influenced by what H_A is. For example, suppose that H_A was something like: "Machine M does not give independent samples from the uniform distribution ; it seems that it gives numbers close to one more frequently than what one would expect". That little piece of information suggests considering a function T as in (2). If in fact the machine has the problem suggested by H_A, then t_{obs} is going to be high in comparison to what one would expect if the numbers were chosen independently from the uniform distribution; that is, the p-value would be very small.

S: Ok. So if H_A was something like: "Machine M does not give independent samples from the uniform distribution ; it seems that it gives numbers close to *zero* more frequently

than what one should expect", then it may be a good idea to consider a function T like:

$$T(x_1, \dots, x_n) := \sum_{k=1}^{n} \mathbb{1}_{x_k < 0.2}?$$

P : *Thumbs up*.

S: If the machine M indeed had that problem, then the p-value $p = \mathbb{P}(T(U_1, \dots, U_n) \geq t_{jobs})$ would be very small...

P : *Thumbs up*.

1.4. **Transforming randomness: Using the inverse of the cumulative distribution function.** -S I am a little bit concerned. We have learned a lot about the uniform distribution, which somehow models fair behavior, but I do not think that the world is fair... In other words, we have not learned anything about this world!

-P No need to be tragic Solo. In what respects to random numbers, as we will learn now, if we know how to simulate samples from the uniform distribution then in principle we can simulate samples from any distribution we may be interested in. On the other hand, fairness or unfairness in the world does not have to do much with random actions we make, but rather with deterministic ones. It may be true that we have not learned much about the world.

The discrete case. Suppose that X is a discrete random variable with probability mass function:

$$\mathbb{P}(X = a_k) = p_k, \quad k \in \mathbb{N},$$

where the numbers a_k are the possible outcomes of X. The set $\{a_1, a_2, \dots\}$ may be finite (like in the case of the binomial distribution) or may be infinite (like in the case of the Poisson distribution). Suppose that $U \sim Uniform([0, 1])$. The idea is to use U to define a random variable X with the discrete distribution described above. Here is the idea. Recall that one of the properties of a p.m.f is that $p_k \geq 0$ for every k and that

$$\sum_{k=1}^{\infty} p_k = 1.$$

Let us break the interval $[0, 1]$ into subintervals

$$\begin{aligned} I_1 &:= [0, p_1) \\ I_2 &:= [p_1, p_1 + p_2) \\ I_3 &:= [p_1 + p_2, p_1 + p_2 + p_3) \\ &\vdots \\ I_k &:= [\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i) \end{aligned}$$

Observe that the subintervals $I_1, I_2, \dots$ are disjoint, consecutive, and they exhaust the interval $[0, 1]$. Given $U \sim Uniform([0, 1])$, we define X to be

$$X = a_k,$$

if $U \in I_k = [\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i)$. In other words, if U happens to fall in the interval I_k, then we give X the value a_k. We claim that X has the discrete distribution described before. To see this, we only have to check that for every k, the probability of $X = a_k$ is actually equal to p_k. Now, the event $\{X = a_k\}$ is equal to the event $\{U \in I_k\}$. Hence,

$$\mathbb{P}(X = a_k) = \mathbb{P}(U \in I_k) = p_k;$$

the last equality follows from the fact that $U \sim Uniform([0,1])$ and that the length of the interval I_k is p_k.

We conclude then that in order to get $X_1, \ldots, X_n$ an i.i.d. sample from the distribution of X, it is enough to consider an i.i.d. sample $U_1, \ldots, U_n$ from the $Uniform([0,1])$ distribution and then define

$$X_k = a_j$$

if $U_k \in I_j$.

Continuous case. Now let us consider a random variable X with c.d.f. F_X. For convenience we assume that F_X is continuous and strictly increasing. Note that the normal distribution and the exponential distribution satisfy these conditions. In order to obtain $X_1, \ldots, X_n$ i.i.d. samples from F_X, we consider i.i.d. samples $U_1, \ldots, U_n$ samples from the uniform distribution and define:

$$X_k := F_X^{-1}(U_k).$$

In the above formula, F_X^{-1} is the inverse of F_X; observe that F_X^{-1} exists because F_X was assumed to be strictly increasing. On the other hand, the domain of F_X^{-1} contains the interval $(0,1)$, because F_X was assumed to be continuous. This implies that the quantity $F_X^{-1}(U_k)$ is well defined. Now let us check that X_k indeed has c.d.f. F_X. To see this, considered a fixed value $t \in \mathbb{R}$; note that $F_X^{-1}(U) \leq t$ if and only if $U \leq F_X(t)$ (verify this!; this follows from the fact that F_X is strictly increasing). Hence:

$$\mathbb{P}(X_k \leq t) = \mathbb{P}(F_X^{-1}(U_k) \leq t) = \mathbb{P}(U_k \leq F_X(t)) = F_X(t);$$

the last equality follows from the fact that U_k is uniformly distributed in $[0,1]$.

General case: For a general $c.d.f$ F_X, not necessarily continuous and not necessarily strictly increasing, one may not be able to define the inverse of F_X. But even in those cases, one can still define a "pseudo inverse" F_X^{-1})(in the Statistics literature known as quantile function). For $\alpha \in (0,1)$, we define

$$F_X^{-1}(\alpha) := \inf\{t \in \mathbb{R} \,:\, F_X(t) \geq \alpha\};$$

where $\inf$ stands for the infimum of a set. Let us not go into details of what this exactly means and let us simply say that intuitively the above formula says that $F_X^{-1}(\alpha)$ is defined to be equal to the smallest number t such that $F_X(t) \geq \alpha$.

Example: Take for example the c.d.f

$$F(t) := \begin{cases} 0 & \text{if } t < 0 \\ 1/2 & \text{if } t \in [0,1) \\ 1 & \text{if } 1 \leq t. \end{cases} \tag{3}$$

This is the c.d.f of the $Bernoulli(1/2)$ distribution. Clearly F is not invertible and clearly F is not continuous. Nevertheless, we can talk about its pseudo inverse, which in this case is equal to:

$$F^{-}1(\alpha) := \begin{cases} 0 & \text{if } \alpha \in (0,1/2] \\ 1 & \text{if } \alpha \in (1/2,1) \end{cases} \tag{4}$$

(verify this!). Observe that if $U \sim Uniform([0,1])$, then $F^{-1}(U)$ is a $Bernoulli(1/2)$ r.v.

Example: In case F is continuous and strictly increasing, then F's pseudo inverse coincides with F's inverse. We have seen that if $U \sim Uniform([0,1])$, then $F^{-1}(U)$ has c.d.f. F.

The above examples illustrate that regardless of the shape a c.d.f. F has, if $U \sim Uniform([0,1])$, then $F^{-1}(U)$ has c.d.f. F; here F^{-1} is the pseudo inverse of F. This is true in general! no restriction on the c.d.f F. The bottom line is: if you know how to sample from the uniform distribution, then in principle you have a recipe to sample from any imaginable random variable.

- S: Ha! You said in principle...which means there are hidden difficulties.

- P: Of course. I can not give away all the surprises at once!

Sampling from distributions like the exponential distribution using the inverse of its c.d.f. is easy :

Example: The exponential distribution with parameter $\lambda > 0$ has the density:

$$f(x) = e^{-\lambda x} \mathbb{1}_{[0,\infty)}.$$

The corresponding c.d.f is obtained upon integration of this density:

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

$$F(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ 1 - e^{-\lambda t}, & \text{if } t \geq 0. \end{cases}$$

Observe that this c.d.f. is invertible (if we restrict the domain of definition to the region $x \geq 0$; the region $x < 0$ is irrelevant for our purposes). But it is not only invertible, more importantly, we can compute it explicitly!: for $\alpha \in (0,1)$ we have

$$F^{-1}(\alpha) = \frac{-1}{\lambda} \log(1-\alpha).$$

Thus, to obtain n independent samples from the exponential distribution with parameter λ, we sample $U_1, \ldots, U_n$ from $Uniform([0,1])$ and then define:

$$X_k := \frac{-1}{\lambda} \log(1-U_k).$$

The sequence $X_1, \ldots, X_n$ is our desired sample.

What worked for the exponential distribution does not work for the normal distribution:

Example:

The c.d.f. for a normal distribution with mean 0 and variance 1 is obtained upon integration of its density:

$$F(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx.$$

The problem is that we can not compute the above explicitly! So how can we generate samples from the normal distribution?

1.5. **Transforming randomness: Box-Muller method to sample from the normal distribution.** - P: To have a good method to sample from the normal distribution we need to learn a trick.

-S: A trick... meaning you will try to impress me and I won't learn much out of it?

-P: I know how you feel about "tricks." But in this case it is worth it, because the normal distribution appears everywhere! Better to know how to sample from it. The method we will learn is called the Box-Muller method. The basic idea is to use polar coordinates.

Recall that we can parametrize $\mathbb{R}^2 \setminus \{0\}$ using polar coordinates:

$$x = r\cos(\theta), \quad y = r\sin(\theta), \quad r \in (0,\infty), \quad \theta \in [0, 2\pi).$$

Now, suppose that X, Y are two independent random variables, both of them $N(0,1)$ (Gaussian with mean 0 and variance 1). Using polar coordinates, we can define random variables R (taking values in $(0,\infty)$) and Θ (taking values in $[0,2\pi)$) such that

$$X = R\cos(\Theta), \quad Y = R\sin(\Theta). \tag{5}$$

The plan is the following: First we compute the distribution of R, Θ. Then, we show that it is actually easy to sample from R, Θ. Then, we can simply use (5) to obtain two independent samples from the $N(0,1)$ distribution.

Since the variables X, Y are independent, their *joint density* $f_{X,Y}(x,y)$ is the product of the *marginal* densities:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}.$$

On the other hand, for arbitrary $t \in [0,\infty)$ and $\gamma \in [0,2\pi)$, we have the following:

$$\begin{aligned}
\mathbb{P}\left(R \leq t \text{ and } \Theta \leq \gamma\right) &= \int_{\{(x,y)\in\mathbb{R}^2 \,:\, r\in[0,t], \quad \theta\in[0,\gamma)\}} f_{X,Y}(x,y)dxdy \\
&= \frac{1}{2\pi}\int_0^\gamma \int_0^t e^{-r^2/2} r dr d\theta. \\
&= \left(\int_0^t e^{-r^2/2} r dr\right)\left(\frac{1}{2\pi}\int_0^\gamma d\theta\right) \\
&= \mathbb{P}(R \leq r)\cdot \mathbb{P}(\Theta \leq \gamma),
\end{aligned} \tag{6}$$

in the second equality we have changed to polar coordinates; observe that the factor r is nothing but the Jacobian associated to this change of variables. Now the above change of equalities says the following. First, the variables R and Θ are independent (Why?). Second, Θ is uniformly distributed on $[0,2\pi)$ and R is distributed according to the density

$$f_R(t) = e^{-t^2/2}t, \quad t \in [0,\infty).$$

We have determined the distributions of Θ and R. Furthermore, we have determined their joint distribution given that the variables are independent. We now show that it is easy to sample from (R,Θ). The variable R has density f_R, let us compute its c.d.f:

$$F_R(t) = \int_0^t e^{-r^2/2} r dr = 1 - e^{-t^2/2};$$

the second equality can be obtained by first considering the change of variables $u := -r^2/2$. Now we consider the inverse of F_R:

$$F_R^{-1}(\alpha) = \sqrt{-2\log(1-\alpha)}, \quad \alpha \in (0,1).$$

Thus in order to obtain two independent samples X, Y from the normal distribution we consider two independent samples $U_1, U_2 \sim Uniform([0,1])$. Then we set:

$$\Theta := 2\pi U_1$$

$$R := \sqrt{-2\log(1-U_2)}$$

Finally, we set

$$X = R\cos(\Theta); \quad Y = R\sin(\Theta).$$

We have shown that X, Y are independent $N(0,1)$ random variables. This is the Box-Muller method.

P: I think it is a beautiful "trick." It is essentially the same idea used to compute the value of the integral $\int_{-\infty}^{\infty} e^{-x^2/2}dx$. Actually, it is so beautiful and simple that I don't think it deserves the label "trick"; we should say instead that this method deserves to be part of our mathematics/statistics *folklore*.

1.6. **Transforming randomness: Rejection sampling.** -S: What is the probability that for an arbitrary distribution, we can either a) compute explicitly its c.d.f and its quantile function or b) find a "trick" that we can use to reduce the problem to a)?

P: You are not concerned about your odds for the exam, are you? If the inverse of the c.d.f. method works, then by all means use it! If not...

S: Bad luck?

P: Nah. We can always do something about it. We will study a method called rejection sampling.

Given an arbitrary density $f : \mathbb{R} \to \mathbb{R}$, the idea is that we learn how to sample from it using a method called rejection sampling. We need the following definition.

Definition 1.11. *Consider a region $A \subseteq \mathbb{R}^2$. We say that a pair of random variables (X, Y) are uniformly distributed on A if for every subregion $B \subseteq A$ we have*

$$\mathbb{P}((X, Y) \in B) = \frac{area(B)}{area(A)}.$$

Now, to the density f we associate a region in $\mathbb{R}^2$:

$$B := \{(x, y) \ : \ 0 \leq y < f(x)\}. \tag{7}$$

Suppose that we can sample (X, Y) from the uniform distribution on B (we will later have to learn how to do this). Then, we claim that X is distributed like a random variable with density f. Indeed, for $t \in \mathbb{R}$ we have

$$\mathbb{P}(X \leq t) = \mathbb{P}((X, Y) \in B_t) = \frac{area(B_t)}{area(B)} = \frac{\int_{-\infty}^{t} f(x)dx}{\int_{-\infty}^{\infty} f(x)dx} = \int_{-\infty}^{t} f(x)dx;$$

where in the above, the region B_t is defined by

$$B_t := \{(x, y) \in B \ : \ x \leq t\}.$$

Draw a graph to convince yourself that the above equalities do hold! The above chain of equalities show that indeed X has density f. We conclude then, that in order to sample from the distribution f, the only thing we need to do is to learn how to sample uniformly from the region B. We will consider different cases.

Case 1: Suppose that the density f is such that $f(x) = 0$ unless $x \in [a, b]$ and that $f(x) \leq M$ for every x. In that case we could consider the rectangle $A := [a, b] \times [0, M]$. Observe that B is contained in A. Moreover, observe that it is very easy to sample (X', Y') uniformly from A: pick $X' \sim Uniform([a, b])$ and $Y' \sim Uniform([0, M])$. The rejection sampling algorithm is:

(1) Sample (X', Y') uniformly from A.
(2) If $(X', Y') \in B$ set $X = X'$ and $Y = Y'$; return X. Else, go back to 1.

Let us show that indeed X has the desired density f; observe that it is enough to show that (X, Y) is uniformly distributed on B. Indeed, consider an arbitrary subregion $C \subseteq B$.

Then,

$$\mathbb{P}((X,Y)\in C)=\mathbb{P}((X',Y')\in C|(X',Y')\in B)=\frac{\mathbb{P}((X',Y')\in C)}{\mathbb{P}((X',Y')\in B)}$$
$$=\frac{area(C)/area(A)}{area(B)/area(A)}=\frac{area(C)}{area(B)}.$$

Case 2: Suppose that we have a density:

$$f(x)=\frac{1}{Z}\sqrt{1-x^2},\quad x\in[-1,1].$$

In the above, Z is the "normalization" constant for f. That is, Z is chosen so that f integrates to one. In this case we could compute it explicitly ($Z=\pi/2$), but in some cases this may be hard to compute. What we show now, is that it does not matter that we can not compute Z explicitly. We can do rejection sampling considering the graph of the function $g(x)=\sqrt{1-x^2}$. Indeed, consider $B:=\{(x,y)\ :\ x\in[-1,1],\quad 0\leq y\leq g(x)\}$. Also, let A be some rectangle containing B; for example take $A=[-1,1]\times[0,1]$. Then we use the rejection sampling algorithm:

(1) Sample (X',Y') uniformly from A.
(2) If $(X',Y')\in B$ set $X=X'$ and $Y=Y'$; return X. Else, go back to 1.

It is clear that X thus obtained is distributed according to f (Make sure you understand why!). Note that we never had to compute Z explicitly to be able to sample from f.

Case 3: So far we considered the case in which f has bounded support, that is, we assumed that the set in which $f(x)>0$, is bounded. In the unbounded support case, things become a little bit more complicated because we can not enclose the graph of f in a nice bounded rectangle. This is definitely a problem!

We would like to extend the analysis we have carried up to this point to the following setting: Suppose we can find a density h with the following properties

(1) We know how to draw samples from h.
(2) For some M, $f(x)\leq Mh(x)$ for every x.

In this setting we can obtain samples from f using rejection sampling:

(1) Sample X' from the density h. Then, sample Y' uniformly on the interval $[0,Mh(X')]$.
(2) If $(X',Y')\in B$ set $X=X'$; return X. Else, go back to 1.

The set B is as in (7). You can check that X thus obtained is indeed distributed according to f.

-P : By no means rejection sampling is the ultimate answer to sampling. Take a look at Case 3. It assumes that you can reduce the problem to sampling from a distribution h you know how to sample from. But in general, for arbitrary f, how can we find h satisfying the required properties? The answer is given on a case by case basis.

- S : I see a potential problem as well. In the rejection sampling algorithm, how long do I have to wait to obtain **one** single sample from the required distribution?

- P: You are right about being concerned! That actually depends on the ratio between the area of B and the area of A. If this ratio is very small, then you will probably have to throw away a lot of samples before you get to see your desired samples. This observation is a simple application of the law of large numbers. Moreover, I would like to anticipate a little bit to what we will learn when we study stochastic approximation. Suppose that we wanted to obtain samples from the uniform distribution on the unit ball $B(0,1)$ in $\mathbb{R}^d$. Suppose we do this by using samples from the hypercube $[-1,1]^d$ and using rejection sampling. I assure you that if d was something like 100, then in order to obtain one single

sample using that rejection sampling scheme, you would have to wait longer than what the universe has been waiting from its beginning to get to see us discuss these things...

-S : Is the second that has just gone by included in that waiting time?

1.7. **Summary.** The goal of this section was to simulate randomness. More specifically, to simulate samples from a given distribution. We started defining pseudo-random number generators to simulate independent random samples from the uniform distribution on the interval $[0, 1]$. We stated that a 'good' pseudo-random number generator should provide us with numbers that behave as true random numbers. In particular, the law of large numbers is a fundamental property that true random samples satisfy. We also introduced the notion of hypothesis testing as an organized method that in some cases gives us enough statistical confidence to reject the hypothesis that a given generator provides i.i.d. samples from the uniform distribution.

Initially we restricted our attention to the uniform distribution on $[0, 1]$ because we later showed that in principle we can sample from any desired distribution if we understand how to sample from the uniform distribution. We saw two methods for doing this. The first one uses the inverse of a cumulative distribution function (quantile function). This approach works well for discrete distributions and also for distributions whose c.d.f. can be written in an exact form. Rejection sampling, the other method we learned in class, is an alternative when the inverse of the c.d.f. method does not work. No method is perfect.

2. Stochastic approximation

The idea of this unit is to learn how to estimate integrals using randomness.

2.1. **Stochastic approximation: using samples from the uniform distribution.** -S: What is the relation between estimating an integral and what we did in the first section?.

-P: At first sight, nothing. Estimating an integral is a deterministic problem: in principle there are deterministic numerical methods that allow you to estimate an integral. On the other hand, in the section we learned how to sample random (or better yet pseudo-random) samples from different probability distributions. Stop for a second and think about it, in principle there is no relation between the two tasks: estimating integrals and sampling! But the really nice thing is that there is in fact a connection between the two problems. This is not something obvious, and if it is obvious to you it is probably because you have seen it before...

-S: Ok ok ... no need to go on a rant! Not very *pacifica* today.

-P: Hmmm... Estimating integrals is the first application of sampling + the law of large numbers.

The setting is very simple. Given a function $h : [0, 1]^d \to \mathbb{R}$, we want to estimate the integral:

$$I := \int_{[0,1]^d} h(x)dx$$

In the above formula, it is understood that x is a vector $x = (x_1, \dots, x_d)$.

Let $U_1, \dots, U_n, \dots$ be i.i.d. random **vectors** uniformly distributed in $[0, 1]^d$. Then the law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n} h(U_i) \approx I.$$

Observe that in particular if we wanted to estimate the volume of a region A in $[0,1]^d$, then we could consider h to be the indicator function of A. That is, we could estimate A's volume with:

$$\frac{\#\{U_i \in A\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(U_i) \approx \int_{[0,1]^d} \mathbb{1}_A(x)dx = \text{Vol}(A).$$

- S: I see that the approximation works because of the law of large numbers. But why would we use a random method to attack a deterministic problem?

Suppose d is large. Say $d = 100$. A typical naive deterministic approximation of an integral of the form $\int_{[0,1]^d} h(x)dx$ involves creating a mesh on the hypercube $[0,1]^d$. One can obtain such mesh by subdividing the interval $[0,1]$ in every coordinate. Say that each coordinate is then partitioned into $n = 10$ subintervals. Notice that 10 is actually a very small number. The points in the mesh would be of the form $(k_1/10, k_2, 10, \ldots, k_d/10)$ for integers k_i between 0 and 9. In total, there are 10^d points in this mesh, if $d = 100$, 10^{100} (more points than particles in the visible universe!) The integral I would be approximated by

$$I \approx \frac{1}{n^d}\sum_{k_1,\ldots,k_d} h(\frac{k_1}{n}, \ldots, \frac{k_d}{n}).$$

- P: There are 10^{100} summands on the right hand side of the above expression! Who or what is going to compute that sum? The usefulness of the stochastic approximation (also known as Monte Carlo method) is the fact that the error of the approximation (as we will see later on) is independent of the dimension.

-S: There was some mention to "naive" deterministic methods... Are there more sophisticated deterministic methods to approximate integrals?

-P: Absolutely. If you are interested, you can search for 'Quasi-Monte Carlo'. Let me just say that in general, in any numerical method for integral approximation (based on randomness or not), the idea is always to choose a 'good' set of points in the domain of integration (in this case $[0,1]^d$). At those points you evaluate the function h and then you take the average of those values. The Monte Carlo method is nothing but a way to select such set of points, you don't have to work on constructing sophisticated meshes or anything like that: plain random (or pseudo-random) points. And the best part is that for many purposes it works just fine!

In theory the law of large numbers guarantees that the Monte Carlo method works asymptotically. But for practical purposes the relevant question is, how big does 'n' have to be so that the approximation is actually a good approximation? This has to do with the variance of the estimator. Let I_n be

$$I_n := \frac{1}{n}\sum_{i=1}^{n} h(U_i).$$

Let us compute the variance of I_n. First, observe that $\mathbb{E}(I_n) = I$. On the other hand, using properties of the variance we obtain that

$$\text{Var}(I_n) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(h(U_i)) = \frac{\sigma^2}{n},$$

where

$$\sigma^2 := \text{Var}(h(U_i)) = \int_{[0,1]^d} (h(x))^2 dx - I^2.$$

Observe that if $\sigma^2 = 0$ then we would have perfect approximation: $I_n = I$. On the other extreme, the larger σ^2, the larger n would have to be to get a good approximation.

Example: Let us consider the set B_d^+ to be the region in $\mathbb{R}^d$ given by

$$B_d^+ := \left\{(x_1, \ldots, x_d) \in [0,1]^d \ : \ x_1^2 + \cdots + x_d^2 \leq 1\right\}.$$

If d is an even number, we can express the volume of B_d^+ by

$$\mathrm{Vol}(B_d^+) = \frac{\pi^{d/2}}{2^d \cdot (\frac{d}{2})!}.$$

If d is not even, one still has a similar formula except that now one has to use the Γ function (which generalizes the factorial). Let us try to estimate the volume of B_d^+ for $d = 2, 4, 8, 16, 32$. We do this by using Monte Carlo with $n = 10^5$ samples from the uniform distribution. We repeat this three times. These are the results I obtained:

	$d=2$	$d=4$	$d=8$	$d=16$	$d=32$
$\mathrm{Vol}(B_d^+)$	0.7854	0.3084	0.0159	3.5909×10^{-6}	1.0019×10^{-15}
MC 1	0.7861	0.3081	0.0159	5.000×10^{-6}	0
MC 2	0.7858	0.3091	0.0159	5.000×10^{-6}	0
MC 3	0.7850	0.3084	0.0156	4.000×10^{-6}	0

-P: For $d = 32$ there was no single sample in the set B_d^+. We would have had to sample $n = 10^{-15}$ points to actually see something: most of the samples were far away from the region of interest. This issue, dear Solo, leads us to the next topic: importance sampling.

2.2. **Importance sampling.** We would like to generalize the Monte Carlo method to a method that uses samples from more general density functions. As before the goal is to estimate an integral of the form $\int_{\mathbb{R}^d} h(x)dx$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a density function. Let $X_1, \ldots, X_n$ be i.i.d. samples from the density f. Then the following holds by the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)}{f(X_i)} \approx \int_{R^d} \left(\frac{h(x)}{f(x)}\right) f(x)dx = \int_{\mathbb{R}^d} h(x)dx.$$

Remark 2.1. Now that the samples are not distributed uniformly, note that we are considering the function h/f in the sum. If we had computed

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i),$$

we would have obtained an approximation to the integral $\int_{\mathbb{R}^d} h(x)f(x)dx$ ... an integral we were not interested in. We needed to modify our function so that we could cancel the density f appearing in the integral.

Let us denote by I_n the quantity

$$I_n := \frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)}{f(X_i)}.$$

It is clear that I_n is an unbiased estimator, that is, $\mathbb{E}(I_n) = I$. On the other hand, using properties of the variance, it is straightforward to check that

$$\mathrm{Var}(I_n) = \frac{\sigma^2}{n},$$

where

$$\sigma^2 = \mathrm{Var}\left(\frac{h(X_1)}{f(X_1)}\right) = \int_{\mathbb{R}^d} \frac{(h(x))^2}{f(x)} dx - I^2.$$

-P :The above expression is very revealing. We are interested in an estimator I_n with small variance: the smaller the variance, the better the approximation. But the variance of I_n depends ultimately on the quantity $\int_{\mathbb{R}^d} \frac{(h(x))^2}{f(x)} dx$. The concept of importance sampling is the following: we have a free parameter, namely we can choose the density f we want to sample from, the goal is to choose f in such a way that the expression

$$\int_{\mathbb{R}^d} \frac{(h(x))^2}{f(x)} dx,$$

is as small as possible.

-S : But in practice we will not actually solve a minimization problem, right?

- P : Well, only in a qualitative way. You can actually write down a formula for the best function f (try it!). Now, there is one hidden constraint when choosing f: you need to be able to sample effectively from f, otherwise you can not compute the estimator I_n. Hence, one looks for density functions f such that: 1) you can obtain samples from f and 2) f resembles h as much as possible. Another way of saying this is that you want f to put points on the region where "the action is happening".

Example (The Volume of B_d^+ revisited.) In order to compute the volume of B_d^+ for $d = 32$, we need to use a density f different from that of the uniform distribution. Here is what we do. First observe that B_d^+ is a $1/2^d$ fraction of the unit ball

$$B_d := \left\{(x_1, \ldots, x_d) \ : \ x_1^2 + \cdots + x_d^2 \leq 1\right\},$$

and in particular

$$\mathrm{Vol}(B_d^+) = \frac{\mathrm{Vol}(B_d)}{2^d}.$$

Hence we can focus on obtaining $\mathrm{Vol}(B_d)$. The function h we want to integrate is $h(x) = \mathbb{1}_{x \in B_d}$. To understand where the action is happening, we consider a ball concentric to B_d, with radius 0.9. Observe that in dimension 2, most of the volume of the bigger ball is due to the volume of the smaller ball. In higher dimension however, this is not the case. In fact, the ratio of the volumes is equal to $(0.9)^d$; when $d = 32$ this is equal to 0.0343, which means that the smaller ball only contributes to 3% of the unit ball! Hence, almost all of the volume concentrates towards the boundary of the unit ball. This is where the action is happening.

To choose the density f, we keep in mind the previous observation. We also keep in mind that a ball is spherically symmetric, and so its volume is isotropically distibuted. Hence, it would be a good idea to consider a density function f which is spherically symmetric as well, that is, a function that only depends on the quantity $\sum_{i=1}^d x_i^2$.

One such f is obtained as follows. Let $f : \mathbb{R}^d \to \mathbb{R}$ be given by

$$f(x_1, \ldots, x_d) = \frac{1}{(\sqrt{2\pi b^2})^d} \exp\left(-\frac{\sum_{i=1}^d x_i^2}{2b^2}\right);$$

observe that this is the joint distribution of $(Z_1, \ldots, Z_d)$, where the Z_i are i.i.d. $N(0, b^2)$. Can we sample from f? Yes, because we know how to sample from the normal distribution $N(0, 1)$ (how do you deal with the variance b^2?) Now, does this density put most of its mass where the action is happening?, that is, on the region $\left\{(x_1, \ldots, x_d) \ : \ x_1^2 + \cdots + x_d^2 = 1\right\}$? That actually depends on the choice of b.

To understand this, we consider the random variable

$$R := \sqrt{\sum_{i=1}^{d} Z_i^2}.$$

This random variable represents the distance of the point $(Z_1, \dots, Z_d)$ to the origin; we denote by f_R its density (this distribution has been well studied, look for χ-distribution). Since we want our samples to concentrate on the boundary of B_d, we analyze where the maximum of f_R occurs (better to be at $r = 1$). But in fact, for the (rescaled) χ-distribution, the maximum is achieved at $r = b\sqrt{d-1}$. Thus, if we choose $b = 1/\sqrt{d-1}$, the maximum is achieved at $r = 1$.

Implementation: Consider $n = 10^6$ samples from the density f (with $b = 1/\sqrt{31}$). Then compute

$$\frac{1}{n} \sum_{k=1}^{n} \frac{\mathbb{1}_{B_d}(X_i)}{f(X_i)}.$$

When I did this, I obtained $2^d * 9.9965 \times 10^{-16}$: this is the approximation for $\mathrm{Vol}(B_d)$. On the other hand the approximation for $\mathrm{Vol}(B_d^+)$ is 9.9965×10^{-16}. The relative error of this approximation is only about 0.2%! This is an illustration of the power of importance sampling.

2.3. **The Central Limit Theorem.** -P : You may have probably noticed that so far we have always written $\approx$ to say that as $n \to \infty$ two quantities become equal. But the $\approx$ symbol, which can be rigorously defined mathematically in terms of limits, is in many cases useless in practice.

-S : I had noticed, yes.

-P: The central limit theorem is the first result that will give us information about how big n has to be so that the $\approx$ symbol really means "approximately equal". It gives us information about the rate of convergence of the law of large numbers; the CLT will help us to understand the error of Monte Carlo approximation.

There is literature about CLT everywhere; here let us be very concrete.

Theorem 2.2 (Central Limit Theorem.)**.** *Let $Y_1, \dots, Y_n$ be **independent**, identically distributed random variables with **finite** variance σ^2 and mean μ. Then, for every interval $[a, b] \subseteq \mathbb{R}$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma} \in [a, b] \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

In simple terms, what the CLT says is that as $n \to \infty$ the random variable $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma}$ starts behaving more and more like a standard Gaussian.

Remark 2.3. Why subtracting μ? This is so that the variable $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma}$ stays centered and doesn't escape to infinity or negative infinity. Indeed,

$$\mathbb{E}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma} \right) = 0.$$

Why the scaling $\frac{1}{\sqrt{n}}$?. This is so that the variable $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{Y_i-\mu}{\sigma}$ does not spread out or concentrate at one point. Indeed,

$$\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{Y_i-\mu}{\sigma}\right) = 1.$$

In the context of stochastic approximation, for given samples $X_1, \dots, X_n$ from a density $f : \mathbb{R}^d \to \mathbb{R}$, we let

$$Y_i := \frac{h(X_i)}{f(X_i)}.$$

Then, one can check that

$$\frac{\sqrt{n}(I_n - I)}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{Y_i-\mu}{\sigma}, \tag{8}$$

where we recall σ^2 is

$$\sigma^2 = \int_{\mathbb{R}^d} \frac{(h(x))^2}{f(x)}dx - I^2.$$

Looking at (8), we deduce that for n large enough, the random variable $\frac{\sqrt{n}(I_n-I)}{\sigma}$ behaves like a standard Gaussian.

- S: Let me interrupt for a second. We are still using the phrase "n large enough"...

-P: True. But in a different way. Here the phrase "large enough" is used to guarantee that the *distribution* of $\frac{\sqrt{n}(I_n-I)}{\sigma}$ looks like the standard Gaussian distribution. You see, distributions are actually deterministic objects. If I say: "standard Gaussian distribution", you don't have to sample anything to compute it, you just write down the formula $\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ for its density. A random variable is a random object because you don't know what its outcome is going to be; on the other hand, if you know that a random variable has certain distribution (normal, exponential, binomial, etc) then you can quantify how likely the outcomes are (even if you don't know what the actual outcome is).

As we will learn when we study convolutions, in theory one can compute the exact distribution of $\frac{\sqrt{n}(I_n-I)}{\sigma}$. Thus, in principle there is no need for approximations and we can quantify how likely the outcomes for I_n are. The problem is that as we will see, computing convolutions is time consuming, specially when n is large. The nice thing about the CLT is that it says that in those cases we may think of $\frac{\sqrt{n}(I_n-I)}{\sigma}$ as standard Gaussian and we may compute probabilistic estimates under that assumption. The bottom line is: if you can compute the exact distribution of I_n, by all means use it!, if not, then cross your fingers, hope the CLT has kicked in, and use the approximation.

Computing confidence intervals. From the CLT (assuming n is large enough) we see that

$$0.95 \approx \mathbb{P}\left(\frac{\sqrt{n}(I_n - I)}{\sigma} \in [-2, 2]\right) = \mathbb{P}\left(I \in \left[I_n - \frac{2\sigma}{\sqrt{n}}, I_n + \frac{2\sigma}{\sqrt{n}}\right]\right).$$

We have just created a 95% confidence interval for the approximation of I. Roughly speaking, this says that with probability ≈ 0.95, the true value of the integral I (what we were trying to estimate) is within distance $\frac{2\sigma}{\sqrt{n}}$ from the estimator I_n. The only reason why it is ≈ 0.95 and not $= 0.95$ is because we have used the CLT approximation; if we had used the exact distribution of $\frac{\sqrt{n}(I_n-I)}{\sigma}$ we wouldn't have had to approximate.

Remark 2.4. Let us make the following remarks.

(1) Note that the size of the interval is proportional to the standard deviation σ.
(2) Note that the size of the interval depends on $\frac{1}{\sqrt{n}}$. So if one wanted to reduce in half the size of the interval, one would need to consider four times the number of samples.
(3) Note that there is no explicit dependence on the dimension d.

-S: I know you think I worry too much...
-P: I do.
-S: ...but I see a problem when computing confidence intervals. How do I compute σ? The expression for σ^2 looks more intimidating than the expression for I, which is what we are trying to compute.

-P :True, computing σ may be harder than computing I. But as in any aspect in life you do the best thing you can with what you have. If there is a theorem supporting that "the best thing you can do" is actually good, even better.

Let us consider the sample variance σ_n^2, which is defined by

$$\sigma_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h(X_i)}{f(X_i)} - I_n\right)^2.$$

The law of large numbers guarantees that σ_n^2 converges to σ^2 as $n \to \infty$. This is enough to state the following: in everything we have done (stating CLT, defining confidence intervals, etc), we can replace σ by σ_n. If you are interested in the theorem that implies this statement, look for Slutsky's theorem.

2.4. **Convolutions.** The idea of this section is to understand the true distribution of the random variable $\sum_{i=1}^{n} X_i$ where $X_1, \dots, X_n$ are independent random variables. Iterating, it is enough to understand what is the distribution of $X + Y$ for two independent random variables X, Y.

Let us denote by f_X, f_Y the densities of X and Y respectively. Since the random variables are independent, the joint density of (X, Y), $f(x, y)$, is the product of the marginals

$$f(x, y) = f_X(x) f_Y(y).$$

In order to compute the distribution of $X + Y$, we compute its c.d.f. (as always). For fixed value $t \in \mathbb{R}$, we define the region

$$A_t := \left\{(x, y) \in \mathbb{R}^2 \ : \ x + y \leq t\right\}.$$

Then,

$$\mathbb{P}(X + Y \leq t) = \mathbb{P}((X, Y) \in A_t) = \iint_{A_t} f(x, y)dxdy = \iint_{A_t} f_X(x) f_Y(y)dydx$$
$$= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{t-x} f_X(x) f_Y(y)dy\right)dx = \int_{-\infty}^{\infty} f_X(x)\left(\int_{-\infty}^{t-x} f_Y(y)dy\right)dx.$$

In order to compute the density of $X + Y$, we differentiate the above expression with respect to t to recover the density f_{X+Y} of $X + Y$:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t - x)dx.$$

Definition 2.5. *Given two functions f and g, we define the convolution of f, g (we denote it by $f * g$) to be the function:*

$$f * g(t) := \int_{-\infty}^{\infty} f(x)g(t - x)dx.$$

P : With the above definition, our computations show that the density of $X+Y$ is equal to the convolution of the densities of X and Y.

S: Why giving the name "convolution"?

P: To avoid saying " that function that you obtain by considering an integral of one function multiplied by another function which is evaluated at a shifted value..."

S: Got it. Can we use the same idea to understand the distribution of $X-Y$?

P: Sure. Just compute the distribution of $-Y$ (how?) and then note that $X-Y = X+(-Y)$.

Remark:

(1) Observe that $f*g(t) = g*f(t)$ for all values of $t \in \mathbb{R}$.

(2) Note that the formula we just deduced should not be a surprise because something similar holds for discrete random variables. Indeed, if X, Y are discrete random variable, we can use the independence of X, Y to show that

$$\mathbb{P}(X+Y=k) = \sum_x \mathbb{P}(X=x)\cdot\mathbb{P}(Y=k-x).$$

CLT revisited. Let us now revisit the CLT. Suppose that $X_1, \dots, X_n$ are i.i.d. random variables with density f. Let us denote by μ the mean and by σ^2 the variance of the variables. Let us compute the distribution of the variable $\frac{X_k-\mu}{\sqrt{n}\sigma}$. For arbitrary t,

$$\mathbb{P}\left(\frac{X_k-\mu}{\sqrt{n}\sigma} \le t\right) = \mathbb{P}\left(X_k \le \sigma t\sqrt{n}+\mu\right)$$

Upon differentiation, we deduce that the density of $\frac{X_k-\mu}{\sqrt{n}\sigma}$, denoted by f_n, is given by

$$f_n(t) = \sigma\sqrt{n} f\left(\sigma t\sqrt{n}+\mu\right).$$

By independence, the density of

$$\sum_{k=1}^{n}\left(\frac{X_k-\mu}{\sqrt{n}\sigma}\right),$$

is given by

$$f_n * \cdots * f_n;$$

that is, taking the convolution of f_n with itself n times. The central limit theorem states that the density $f_n * \cdots * f_n$, approaches that of the standard Gaussian. To be mathematically precise, we would have to specify the *topology* associated to this approximation... this goes beyond the scope of the class.

P: The bottom line is: we know how to compute the density of a sum of independent random variables.

S: But the expressions we obtain for the exact densities are quite horrible...

P: But if the variables are identically distributed we may use the CLT approximation...

S: Provided n is large (clarifying).

2.5. **Summary.** In this section we considered the deterministic problem of approximating an integral of a function or the volume of a region in space. Our approach was to consider randomness to provide us with points in space that we can use to compute an estimator. Different densities can be used to sample the points. In the end, the goal is to choose the density so that the variance of the associated estimator is as small as possible, but there is a constraint: we should be able to sample from the given density in an efficient way.

The constructed estimators are always unbiased and consistent as the number of samples grows due to the law of large numbers. Moreover, the i.i.d. structure allows us to use the

central limit theorem to understand the error of the approximation and create confidence intervals for the true value of the integral.

3. Markov processes

The goal of this section is to talk about random walks. We will start moving away from the i.i.d. structure that we have considered up to this point. In this section we consider Markovian structures.

3.1. **Definitions.** *Markovian structure*:

Definition 3.1. *A sequence of random objects $S_0, S_1, S_2, \dots$ taking values in the (discrete) state space $\mathcal{X}$ is said to be a Markov process (in discrete time and discrete state space) if for every n,*

$$\mathbb{P}(S_{n+1} = s_{n+1} | S_0 = s_0, S_1 = s_1, \dots, S_n = s_n) = \mathbb{P}(S_{n+1} = s_{n+1} | S_n = s_n),$$

for arbitrary $s_0, \dots, s_{n+1} \in \mathcal{X}$. In words, the conditional distribution of S_{n+1} given the whole trajectory up to time n ($S_0, S_1, \dots, S_n$) is equal to the conditional distribution of S_{n+1} given the state at time n (i.e. S_n).

-S: Could you give me a an example of a Markovian structure?
-P : The common saying: "Who sins and then prays, starts from zero".
-S : Huh?... a mathematical example?

To have something conccrete in mind think of $\mathcal{X} = \mathbb{Z}^d$, i.e. the d-dimensional integer lattice or a finite state space like $\mathcal{X} = \{blue, \quad red, \quad yellow\}$. A Markov chain (or Markov process) is in particular a *rule* describing how a state evolves in time. We will focus on two examples of Markov processes:

(1) Random walks on the d-dimensional integer lattice $\mathbb{Z}^d$.
(2) Markov chains designed in the context of the Metropolis-Hastings algorithm (here we will consider Markov processes in *continuous* state space)

For us, a **random walk** will be a sequence of random variables $S_0, S_1, S_2, \dots$ in $\mathbb{Z}^d$, representing the position of a particle wandering around the space at times $t = 0, 1, 2, \dots$ We will assume that S_0 is actually deterministic (it represents the starting point of the particle). The rule that describes a **random walk** is that for every $n \in \mathbb{N}$

$$S_{n+1} = S_n + X_{n+1},$$

where $X_1, X_2, X_3, \dots$ are **independent**, identically distributed random variables in $\mathbb{Z}^d$.

In particular, given the position at time n (S_n) the position at time $n+1$ is obtained by taking a step X_{n+1} from S_n. Observe that the step (X_{n+1}) is independent of the earlier steps (by assumption). Thus, the distribution of S_{n+1} **given** that we know the trajectory of the particle up to time n ($S_0 = s_0, S_1 = s_1, \dots, S_{n-1} = s_{n-1}, S_n = s_n$) is the same as the distribution of S_{n+1} **given** that we just know the position at time n ($S_n = s_n$); in other words random walks are examples of Markov processes. Note that we can alternatively write:

$$S_{n+1} = S_0 + \sum_{k=1}^{n+1} X_k.$$

Example: (d=1) We assume that the possible outcomes for the variables X_k are -1 (left) or 1 (right). Now, we let $p := \mathbb{P}(X_k = 1)$, in particular $1 - p = \mathbb{P}(X_k = -1)$. Intuitively, if $p > 1/2$, the particle will tend to drift towards the right.

Example: (d=2) Here the particle moves in the two dimensional integer lattice and we assume that X_k may be $(-1,0),(1,0),(0,1),(0,-1)$ (left, right, up ,down) with probabilities p_1, p_2, p_3, p_4 respectively.

3.2. **Recurrence.** In this section we will talk about **recurrence** of **time homogeneous** Markov chains and in particular of random walks on $\mathbb{Z}^d$. What is a time homogeneous Markov chain?

Definition 3.2. *A Markov process $S_0, S_1, \dots$ is said to be time homogeneous if*

$$\mathbb{P}(S_n = x | S_{n-1} = y) = \mathbb{P}(S_1 = x | S_0 = y),$$

for all n and all possible states x, y.

In words it means that the probability of transitioning from state y into state x in one step is independent of the time at which the transition takes place.

With this definition at hand we now focus in the following question. Suppose that a Markov process starts at state S_0. What is the probability that the particle ever returns to state S_0? In mathematical terms, we want to compute the probability of the event "there exists some $n \geq 1$ such that $S_n = S_0$". Let us denote by R the probability we are interested in:

$$R := \mathbb{P}(\exists n \geq 1 \text{s.t. } S_n = 0).$$

We will say that the state S_0 is **recurrent** if $R = 1$ and otherwise we say the state is **transient**. From now on we concentrate in the setting of the random walks in the examples form the previous Section, but the analysis below holds for general time homogeneous Markov chains. We will study the recurrence or transience of the state $0 \in \mathbb{Z}^d$.

For $n \in \mathbb{N}$ let us consider

$$Z_n = \mathbb{P}(S_n = 0),$$

and let us set $\beta := \sum_{n=0}^{\infty} Z_n$. Our first goal is to show that:

$$R = 1 - \frac{1}{\beta}. \tag{9}$$

Before proving the above formula, let us point out its implications. First observe that computing R is not a straightforward task. Indeed, our formula will strongly rely on the structure of the variables S_n. The difficulty is that when we say "the particle ever returns" we do not specify when, how, how many times , etc. There are many ways the particle can ever return to its staring point; that generality for the event makes it hard to understand and to quantify its likelihood. But the power of the above formula, is that it says that we can compute this likelihood by computing some quantities that we can understand much better. Indeed, the problem reduces to understanding $\mathbb{P}(S_n = 0)$ for any fixed n. That is, we just need to compute what is the probability that a certain random variable (in general vector) is equal to the origin; this depends only on the distribution of the random variable. In addition, we may not even have to compute explicitly the value of β. In many cases we are just interested in the qualitative behavior of the series β: if the series diverges then $R = 1$ (the particle will return to its starting point), if the series converges then $R < 1$ (the particle may or may not return).

Let us now show (9). We will do this in two ways.

First Proof: Let us denote by τ the random variable that represents the first time ≥ 1 for which $S_\tau = 0$. If $n \geq 1$ and $S_n = 0$, then we conclude that $\tau \leq n$. Then we can express the event $\{S_n = 0\}$ as $\{\tau \leq n, S_n = 0\}$. Thus, for $n \geq 1$

$$Z_n = \mathbb{P}(S_n = 0) = \mathbb{P}\left(\tau \leq n, S_n = 0\right) = \sum_{k=1}^{n} \mathbb{P}\left(\tau = k, S_n = 0\right).$$

Using conditional probabilities:

$$\sum_{k=1}^{n} \mathbb{P}\left(\tau = k, S_n = 0\right) = \sum_{k=1}^{n} \mathbb{P}\left(S_n = 0 | \tau = k\right) \cdot \mathbb{P}(\tau = k).$$

Let us now try to understand what $\mathbb{P}\left(S_n = 0 | \tau = k\right)$ is equal to. Observe that at time $\tau = k$, the particle is at the origin. Since the particle is memoryless (Markovian structure), we conclude that

$$\mathbb{P}\left(S_n = 0 | \tau = k\right) = \mathbb{P}(S_{n-k} = 0).$$

Hence, we conclude that

$$Z_n = \sum_{k=1}^{n} \mathbb{P}(\tau = k) Z_{n-k}.$$

Summing over all n, we deduce that

$$\beta = \sum_{n=0}^{\infty} Z_n = 1 + \sum_{n=1}^{\infty} Z_n = 1 + \sum_{n=1}^{\infty} \sum_{k=1}^{n} \mathbb{P}(\tau = k) Z_{n-k};$$

changing the order of summation, we deduce that

$$\beta = 1 + \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \mathbb{P}(\tau = k) Z_{n-k} = 1 + \sum_{k=1}^{\infty} \mathbb{P}(\tau = k) \cdot \left(\sum_{n=k}^{\infty} Z_{n-k} \right) = 1 + \beta \sum_{k=1}^{\infty} \mathbb{P}(\tau = k).$$

Finally, observe that R can be expressed as:

$$R = \mathbb{P}(\tau < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(\tau = k);$$

in words, if the particle ever returns to the origin is because the first time it returns to the origin is before $t = \infty$. Hence, we conclude that $\beta = 1 + \beta R$.

Second Proof: Now we define the random variable:

$$N := \#\left\{n \geq 1 \, : \, S_n = 0\right\}.$$

That is, N is the number of times the particle visits the origin after time 0. Note that,

$$N = \sum_{n=1}^{\infty} \mathbb{1}_{S_n = 0}.$$

Hence,

$$\mathbb{E}(N) = \mathbb{E}(\sum_{n=1}^{\infty} \mathbb{1}_{S_n = 0}) = \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}_{S_n = 0}) = \sum_{n=1}^{\infty} \mathbb{P}(S_n = 0) = \beta - 1.$$

Let us now compute $\mathbb{E}(N)$ in a different way. First, $\mathbb{P}(N \geq 1) = R$: the particle visits the origin at least once after time 0 if and only if the particle ever returns to the origin. On the other hand,

$$\mathbb{P}(N \geq 2 | N \geq 1) = R$$

(why?). That is, $\mathbb{P}(N \geq 2) = \mathbb{P}(N \geq 2, N \geq 1) = \mathbb{P}(N \geq 2 | N \geq 1) \cdot \mathbb{P}(N \geq 1) = R^2$. Reasoning in the same way, we can conclude that $\mathbb{P}(N \geq k) = R^k$ for all $k \in \mathbb{N}$. This shows that N has a geometric distribution with parameter R and hence

$$\mathbb{E}(N) = \frac{R}{1-R}.$$

Thus, $\frac{R}{1-R} = \beta - 1$.

Example: Recurrence for the 1d random walk. Recall $X_k \in \{-1, 1\}$ and that $\mathbb{P}(X_k = 1) = p$. The idea is to compute $\mathbb{P}(S_n = 0)$ for all n. First of all, observe that if n is odd, there is no way that $S_n = 0$, and so for n odd $\mathbb{P}(S_n = 0) = 0$. Now let us assume $n = 2k$. Recall that $S_{2k} = \sum_{i=1}^{2k} X_i$. The only way S_{2k} can be zero is if the number of X_is that are equal to -1, is equal to the number of X_is that are equal to 1. The probability that this happens is

$$\mathbb{P}(S_{2k} = 0) = \frac{(2k)!}{(k!)^2} p^k (1-p)^k.$$

We deduce that

$$\beta = \sum_{n=0}^{\infty} \mathbb{P}(S_n = 0) = \sum_{k=0}^{\infty} \mathbb{P}(S_{2k} = 0) = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)^2} p^k (1-p)^k.$$

To understand the qualitative behaviour of the series, we use *Stirling's* approximation of the factorial:

Proposition 3.3. *(Stirling's approximation). For every $n \in \mathbb{N}$,*

$$e^{1/(12n+1)} \leq \frac{n!}{n^n e^{-n} \sqrt{2\pi n}} \leq e^{1/(12n)}$$

In particular,

$$n! \approx n^n e^{-n} \sqrt{2\pi n}.$$

The approximation is really good even for small values of n, try it yourself!

Using Stirling's approximation, we conclude that

$$\frac{(2k)!}{(k!)^2} \approx \frac{4^k}{\sqrt{\pi k}},$$

and hence

$$\beta - 1 \approx \sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}}.$$

By $\approx$ we mean that we can find constants c, C such that

$$c \sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}} \leq \beta - 1 \leq C \sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}}.$$

This allows us to conclude that β diverges if and only if $\sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}}$ diverges.

Symmetric case: $p = 1/2$. When $p = 1/2$, the quantity $4p(1-p)$ is equal to 1. Hence, in this case

$$\sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}} = \sum_{k=1}^{\infty} \frac{1}{\sqrt{\pi k}} = \infty.$$

Therefore, in this case $\beta = \infty$ and so $R = 1$. That is, the particle returns to the origin. It also follows that $N = \infty$.

Assymetric case: $p \neq 1/2$. Under this case, the quantity $4p(1-p)$ is strictly less than one. This immediately implies that the series $\sum_{k=1}^{\infty} \frac{(4p(1-p))^k}{\sqrt{\pi k}}$ converges. Thus, $\beta < \infty$ and so $R < 1$, and $N < \infty$. In this case, the particle may leave forever and never return back to its starting point... :(

Example: Recurrence for the 2d random walk. Let us consider the symmetric 2d random walk on the integer lattice, that is, we let $S_0 = 0$ (here 0 denotes the origin in $\mathbb{R}^2$) and the steps X_K are such that

$$X_k = \begin{cases} (-1,0) & \text{with prob. } 1/4 \\ (1,0) & \text{with prob. } 1/4 \\ (0,1) & \text{with prob. } 1/4 \\ (0,-1) & \text{with prob. } 1/4 \end{cases} \tag{10}$$

In words, the particle moves left, right, up, or down with probability $1/4$. To understand the recurrence of this random walk, we compute $\beta = \sum_{n=0}^{\infty} \mathbb{P}(S_n = 0)$. We show that $\beta = \infty$, which in particular implies that $R = 1$ and $N = \infty$.

As in the 1d symmetric random walk, $\mathbb{P}(S_n = 0) = 0$ for n an odd integer. Hence we just need to compute $\mathbb{P}(S_{2k} = 0)$ for $k \in \mathbb{N}$.

Let us try to understand the event $S_{2k} = 0$. Indeed, if $S_{2k} = 0$, then the number of times the particle goes to the left (call this number l) has to be the same number of times the particle goes to the right. Likewise, the number of times the particle goes up (call this number u) has to be the same number of times the particle goes down. Thus, we must have $2k = 2l + 2u$ or what is the same $u = k - l$. There are $\frac{(2k)!}{(l!)(l!)(k-l)!(k-l)!}$ different trajectories in which the particle can go l times left l times right , $k - l$ times up and $k - l$ times down. Each trajectory can happen with probability $\left(\frac{1}{4}\right)^{2k}$. Therefore,

$$\begin{aligned} \mathbb{P}(S_{2k} = 0) &= \sum_{l=0}^{k} \frac{(2k)!}{(l!)(l!)(k-l)!(k-l)!} \left(\frac{1}{4}\right)^{2k} \\ &= \left(\frac{1}{4}\right)^{2k} \sum_{l=0}^{k} \frac{(2k)!}{(l!)(l!)(k-l)!(k-l)!} \\ &= \left(\frac{1}{4}\right)^{2k} \sum_{l=0}^{k} \frac{(2k)!}{(k!)(k!)} \times \frac{k!}{l!(k-l)!} \times \frac{k!}{l!(k-l)!} \\ &= \left(\frac{1}{4}\right)^{2k} \binom{2k}{k} \sum_{l=0}^{k} \binom{k}{l} \binom{k}{k-l}. \end{aligned} \tag{11}$$

In the above, we are using $\binom{n}{m}$ to represent n 'choose' m. Observe that $\sum_{l=0}^{k} \binom{k}{l}\binom{k}{k-l}$ is simply a different way to express $2k$ choose k. Hence,

$$\mathbb{P}(S_{2k} = 0) = \left(\frac{1}{4}\right)^{2k} \binom{2k}{k}^2.$$

Using Stirling's approximation for the factorial, we see that

$$\mathbb{P}(S_{2k} = 0) = \left(\frac{1}{4}\right)^{2k} \binom{2k}{k}^2 \approx \left(\frac{1}{4}\right)^{2k} \frac{4^{2k}}{\pi k} = \frac{1}{\pi k}.$$

From here we see that $\sum_{k=1}^{\infty} \mathbb{P}(S_{2k} = 0)$ behaves like $\sum_{k=1}^{\infty} \frac{1}{\pi k}$; the later series is divergent and thus $\beta = \infty$.

- S: I understand the idea of studying the qualitative bahaviour of β: If $\beta = \infty$ then $R = 1$ and in this case we also have a quantitative understanding of R. But if $\beta < \infty$ the only thing we can say about R is that $R < 1$...not something quantitative.

- P : True. Although there is at least one case where we can compute β explicitly...

Let us go back to the setting of the asymmetric $1d$ random walk. The steps are $X_k = 1$ with probability p or $X_k = -1$ with probability $q := 1 - p$. The standing assumption is that $p \neq 1/2$, or what is the same $pq < 1/4$. We claim that in this case

$$\beta = \frac{1}{\sqrt{1 - 4pq}}.$$

The important observation is the following. For $|t| < 1/4$, we can write the function $f(t) = \frac{1}{\sqrt{1-4t}}$ as a Taylor series around the origin:

$$f(t) = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)(k!)} t^k, \quad |t| < 1/4.$$

The above expression combined with the expression we had obtained for β (in the Example) gives:

$$\beta = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)(k!)} p^k q^k = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)(k!)} (pq)^k = f(pq) = \frac{1}{\sqrt{1 - 4pq}}.$$

The second to last equality holds because $pq < 1/4$.

- S: Nice. But what about other types of random walks?

- P: I would ask the computer to simulate a bunch of random walks and use Monte Carlo.

- S: Monte Carlo?

- P: Yes. You simulate many random walks. In this case Monte Carlo amounts to just computing the fraction of walks that return to the origin. One thing you would need to do is to decide when to abandon a walk and assume that it is never going to return to the origin. For example consider an asymmetric 1d random walk with $\mathbb{P}(X_k = 1) = p > 1/2$. If the walk drifts to the left, then it will tend to return shortly, but if it wanders too far to the right, then it has a very slim chance of ever returning to zero. You can try with a few different criteria for abandoning a walk and make sure that they each give similar answers.

- S : I guess I will have to try it.

- P : You may use for example *exit probabilities*.

-S: I don't know what that is .

-P: Because that is the next topic we will study.

- S: Ok. But before we move on, are symmetric random walks in $\mathbb{Z}^d$ always recurrent regardless of what d is?

-P: No. But I should say no more because in some university in some universe, there are some students working on that problem for their homework.

3.3. **Exit probabilities.** -P: Suppose I give you this \$200 to play in the casino. You decide to play the roulette , your favorite game. Suppose that you are cautious in life and prefer to take one step at a time: you always bet \$1 to even numbers (by far your favorite set of numbers). Since I assume you are not greedy, you decide to stop playing if your budget grows from \$200 to \$300. Since you are *Solo*, you have no one to borrow from (I won't give you more) and you stop playing if your budget reaches \$0. What is the probability that

you ever return to the casino hoping to earn money? Alternatively, what is the probability that you lose all your money and learn the lesson: if you go to the casino you go there to spend and not to earn?.

S: ...Let me try and let's see what happens.

Mathematically speaking we can model the situation described above as follows. Let $S_0, S_1, \dots$ be a random walk on $\mathbb{Z}$, with steps X_k such that $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = -1) = q := 1 - p$. Let us assume that $S_0 \in \{0, \dots, A\}$ and denote by τ the random variable,

$$\tau := \min \{n \in \mathbb{N} \ S_n \in \{0, A\}\}.$$

That is, τ is the exit time from the region $D = \{1, \dots, A-1\}$; it is the first time that the walk hits the boundary of D.

We are interested in S_τ, the position of the random walk when it leaves the domain D. Observe that S_τ is either 0 or A and we are interested in the probability that $S_\tau = A$. That is, we are interested in the probability that the walk hits A before it hits the origin. Let us give a name to the quantity of interest:

$$U_k := \mathbb{P}(S_\tau = A | S_0 = k).$$

The reason for introducing explicitly the intial position of the particle is the following: Although we may only be interested in some specific value of S_0, it proves useful to relate the exit probabilities for different initial conditions as we will see below. This is directly connected to the idea of *first step* analysis.

First, observe that $U_0 = 0$: If the particle starts at 0 there is no way to get to A before getting to 0. Likewise $U_A = 1$. What happens for $k \in D$? If the particle starts at k, then its next move is either left or right. Once it has reached its new position, we may think that the particle starts its motion from time zero and we may now compute what is the probability that the particle reaches A before 0 given the new initial condition. This idea relies on the Markovian structure:

$$\begin{aligned} U_k =& \mathbb{P}(S_\tau = A | S_0 = k) = \mathbb{P}(S_\tau = A, \quad S_1 = k+1 | S_0 = k) + \mathbb{P}(S_\tau = A, \quad S_1 = k-1 | S_0 = k) \\ &= \mathbb{P}(S_\tau = A, | S_0 = k, \quad S_1 = k+1) \cdot \mathbb{P}(S_1 = k+1 | S_0 = k) + \mathbb{P}(S_\tau = A, | S_0 = k, \quad S_1 = k-1) \cdot \mathbb{P}(S_1 = k - \\ &= \mathbb{P}(S_\tau = A, | S_1 = k+1) \cdot p + \mathbb{P}(S_\tau = A, | S_1 = k-1) \cdot q \\ &= pU_{k+1} + qU_{k-1}. \end{aligned}$$

Thus we have the linear system of equations:

$$\begin{gathered} U_0 = 0 \\ -qU_{k-1} + U_k - pU_{k+1} = 0, \quad k \in \{1, \dots, A-1\} \\ U_A = 1 \end{gathered}$$

We can write this system of equations using matrices:

$$C\vec{U} = \vec{b},$$

where

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -q & 1 & -p & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -q & 1 & -p \\ 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\vec{U} = \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{A-1} \\ U_A \end{bmatrix}$$

and

$$\vec{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Thus, in order to compute the exit probabilities, we simply need to solve the linear system $C\vec{U} = \vec{b}$. Does this system have a unique solution? The answer is yes. This is because the matrix C is invertible (why?).

Example: In the roulette game, $p = 9/19$ and $q = 10/19$, $A = 300$ and $S_0 = 200$. Construct the matrix C and solve the corresponding system of equations using Matlab. It turns out that U_{200} (the 201st entry of $\vec{U}$) is something like $3/10000$! What is the expected amount of money you will earn in the casino? That is, what is the expected value of S_τ?

$$\mathbb{E}(S_\tau) = 0 \cdot \mathbb{P}(S_\tau = 0) + 300 \cdot \mathbb{P}(S_\tau = 300) = \frac{9}{1000}.$$

As expected, Solo lost the \$200 Pacifica gave to him.

The previous analysis can be extended to higher dimensions. We will focus on the case $d = 2$. We need the following ingredients.

- A random walk $S_0, S_1, S_2, \ldots$ of the form

  $$S_{n+1} = S_n + X_{n+1},$$

  where the X_k are i.i.d. and

  $$X_k = \begin{cases} (1,0) & \text{with prob.} p_1 \\ (-1,0) & \text{with prob.} p_2 \\ (0,1) & \text{with prob.} p_3 \\ (0,-1) & \text{with prob.} p_4. \end{cases} \tag{12}$$

  Observe that the possible steps the walk can take induce a grid on $\mathbb{R}^2$.
- A connected region $D \subseteq \mathbb{Z}^2$. Connectedness means that one can go from any point in D to any other point in D by following a path on the grid whose vertices are all contained in D.
- The boundary B of D. B is the set of all points not in D that have a neighbor in D. By neighbor we mean the following: for $(x, y) \in \mathbb{Z}^2$, the neighbors of (x, y) are the points $(x + 1, y), (x - 1, y), (x, y + 1), (x, y - 1)$.
- We will assume that the walk starts at a point S_0 in $D \cup B$.
- τ, the exit time from the region D. Assuming the walk is such that $S_0 \in D \cup B$, we define τ by

  $$\tau := \min \{n \geq 0 \ : \ S_n \in B\}.$$

We are interested in answering questions like:

(1) Given $S_0 = (x_0, y_0)$, and given a subregion $\tilde{B} \subseteq B$, what is the probability that $S_\tau \in \tilde{B}$? That is, what is the probability that the random walk leaves the domain D through one of the points in $\tilde{B}$?
(2) Given $S_0 = (x_0, y_0)$, and given a function $f : B \to \mathbb{R}$, what is the expected value of $f(S_\tau)$? You may think of f as a 'reward' function: if the random walk leaves the domain D through point $(x, y) \in B$, then I give you $f((x, y))$ dollars.

Remark: Question 1. is a particular case of question 2. Indeed, we can define $f : B \to \mathbb{R}$ by $f(x, y) := \mathbb{1}_{\tilde{B}}((x, y))$ and recall that the expected value of the indicator function of an event is equal to the probability of the event. That is,

$$\mathbb{P}(S_\tau \in \tilde{B} | S_0 = (x_0, y_0)) = \mathbb{E}\left(\mathbb{1}_{\tilde{B}}(S_\tau) | S_0 = (x_0, y_0)\right).$$

We answer these questions the same way we did it in the 1d case. We first focus on answering question 1.

For $(x_0, y_0) \in D \cup B$, we define

$$U_{(x_0,y_0)} := \mathbb{P}\left(S_\tau \in \tilde{B} | S_0 = (x_0, y_0)\right).$$

The idea is to write down a system of equations for the variables $U_{(x_0,y_0)}$, $(x_0, y_0) \in D \cup B$.

Boundary conditions $(x_0, y_0) \in B$. Observe that if $(x_0, y_0) \in \tilde{B}$, then $U_{(x_0,y_0)} = 1$. On the other hand if $(x_0, y_0) \in B \setminus \tilde{B}$ then $U_{(x_0,y_0)} = 0$.

Interior conditions $(x_0, y_0) \in D$. Here we use a first step analysis as in the 1d case to obtain:

$$U_{(x_0,y_0)} = p_1 U_{(x_0+1,y_0)} + p_2 U_{(x_0-1,y_0)} + p_3 U_{(x_0,y_0+1)} + p_4 U_{(x_0,y_0-1)}.$$

Make sure you can reproduce the above equation.

Reorganizing, we can write:

$$U_{(x_0,y_0)} - p_1 U_{(x_0+1,y_0)} - p_2 U_{(x_0-1,y_0)} - p_3 U_{(x_0,y_0+1)} - p_4 U_{(x_0,y_0-1)} = 0.$$

Denoting by $|D|$ the number of points in D and by $|B|$ the number of points in B, we deduce that the above system of equations, is a system of $|D| + |B|$ linear equations, with $|D| + |B|$ variables. It turns out that regardless of what p_1, p_2, p_3, p_4 are (they should be positive and sum to one), the above system of equations always has a unique solution.

- S: Suppose that I am interested in $\mathbb{P}(S_\tau \in \tilde{B} | S_0 = (x_0, y_0))$ for **only one** initial condition (x_0, y_0). Do I still have to consider **all** possible initial conditions? I am only asking for one initial condition!

- P: The method above says yes! It looks like overkill, but think about it, how would you compute $\mathbb{P}(S_\tau \in \tilde{B} | S_0 = (x_0, y_0))$ otherwise? The power of the above method is that it reduces the problem you wanted to solve (in principle hard) to solving a linear system of equations (easier problem). Moreover, by solving the system of equations, you not only get an answer for the original choice of initial conditions, but for all possible initial conditions. This method gives you more than what you initially asked for.

Now let us turn to question 2. We have all the ingredients to compute $\mathbb{E}(f(S_\tau) | S_0 = (x_0, y_0))$. One way to do this is as follows. For any given $(x, y) \in B$, consider $\tilde{B}_{(x,y)}$ to be the set whose only element is (x, y). We can thus compute

$$\mathbb{P}(S_\tau = (x, y) | S_0 = (x_0, y_0)) = \mathbb{P}(S_\tau \in \tilde{B}_{(x,y)} | S_0 = (x_0, y_0)),$$

for every $(x, y) \in B$. This basically says that we know how to compute the distribution of the random variable S_τ given that $S_0 = (x_0, y_0)$. Then,

$$\mathbb{E}(f(S_\tau)|S_0 = (x_0, y_0)) = \sum_{(x,y)\in B} f((x, y)) \cdot \mathbb{P}(S_\tau = (x, y)|S_0 = (x_0, y_0)).$$

Alternatively, we can replicate the first step analysis we used when answering question 1. Introduce the variables

$$U_{(x_0,y_0)} := \mathbb{E}\left(f(S_\tau)|S_0 = (x_0, y_0)\right).$$

Then,

Boundary conditions $(x_0, y_0) \in B$. Observe that if $(x_0, y_0) \in B$, then $U_{(x_0,y_0)} = f((x_0, y_0))$.

Interior conditions $(x_0, y_0) \in D$. For $(x_0, y_0) \in D$,

$$U_{(x_0,y_0)} - p_1 U_{(x_0+1,y_0)} - p_2 U_{(x_0-1,y_0)} - p_3 U_{(x_0,y_0+1)} - p_4 U_{(x_0,y_0-1)} = 0. \tag{13}$$

- S: So, the above system of equations is similar to the one we deduced when computing $\mathbb{P}(S_\tau \in \tilde{B})$ ...

-P: Yes, the equations associated to interior points (x_0, y_0) are the same. What changes is the boundary conditions.

3.4. **(Optional) Connections to Laplace equation and Brownian motion.** In this section we consider a **symmetric random** walk: In 1d , the probabilities of moving right and left are the same, in 2d the probabilities of going up, down, right or left are all the same, etc.

Given $f : B \to \mathbb{R}$, we obtained a system of equations that the variables

$$U_{(x_0,y_0)} = \mathbb{E}(f(S_\tau)|S_0 = (x_0, y_0)), \quad (x_0, y_0) \in D \cup B$$

satisfy. Recall that we split the equations into boundary conditions and interior conditions. Let us take a closer look at the interior conditions. Indeed, the equation (13) can be rewritten as

$$0 = \big((U_{(x_0+1,y_0)} - U_{(x_0,y_0)}) - (U_{(x_0,y_0)} - U_{(x_0-1,y_0)})\big) + \big((U_{(x_0,y_0+1)} - U_{(x_0,y_0)}) - (U_{(x_0,y_0)} - U_{(x_0,y_0-1)})\big).$$

Now , take for example the term $U_{(x_0+1,y_0)} - U_{(x_0,y_0)}$. This should resemble something like $\frac{\partial U}{\partial x}(x_0, y_0)$. Note that this is only an analogy, as U is a function only defined on a discrete set and hence it does not make sense to talk about its derivative! In that order of ideas, the term

$$\big((U_{(x_0+1,y_0)} - U_{(x_0,y_0)}) - (U_{(x_0,y_0)} - U_{(x_0-1,y_0)})\big)$$

should resemble, $\frac{\partial^2 U}{\partial x^2}(x_0, y_0)$. Likewise, the term

$$\big((U_{(x_0,y_0+1)} - U_{(x_0,y_0)}) - (U_{(x_0,y_0)} - U_{(x_0,y_0-1)})\big)$$

should resemble, $\frac{\partial^2 U}{\partial y^2}(x_0, y_0)$. Thus, the interior conditions, should resemble the equation

$$\frac{\partial^2 U}{\partial x^2}(x_0, y_0) + \frac{\partial^2 U}{\partial y^2}(x_0, y_0) = 0.$$

The above heuristics lead us to consider the following setting. Suppose that D is a connected domain in $\mathbb{R}^2$ like:

This domain is not a discrete set as the sets D we have considered so far: this is a continuous domain. We denote by B its boundary and consider the partial differential equation:

$$\begin{cases} \Delta u(x,y) = 0, & \text{in } D \\ u(x,y) = f(x,y), & \text{on } B, \end{cases} \tag{14}$$

for some function f defined on B. In the above, $\Delta u(x,y) = \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y)$. The equation is interpreted as follows: find a function u whose *Laplacian* (Δu) is equal to zero inside of D and is such that u is equal to f on B. Do you see the similarity with the equations for $U_{(x_0,y_0)}$ (together with boundary conditions) ?

Now, in the previous section we saw that U, the solution to (13) together with the corresponding boundary conditions, could be written in terms of an expectation that involved a discrete random walk. The question is, can one define a continuous random walk such that the solution of (14) can be written in terms of an expectation involving such random walk? The answer is yes! This continuous random walk is called *Brownian motion*. Here, we just give an idea of how one can obtain Brownian motion. We do this in 1d and then extend to higher d.

Start with a symmetric random walk in 1d like the one defined by $S_0 = 0$ and $S_{k+1} = S_k + X_{k+1}$, where the X_k are independent and distributed according to

$$X_k = \begin{cases} 1 & \text{with prob.} 1/2 \\ -1 & \text{with prob.} 1/2 \end{cases} \tag{15}$$

Denote by σ the standard deviation of X_k. Notice that the random walk $S_0, S_1, S_2, \ldots$ is only defined at discrete times. We extend to continuous time by linear interpolation: for any $t \in [k, k+1]$ we let

$$S_t := S_k + (t-k)X_{k+1}.$$

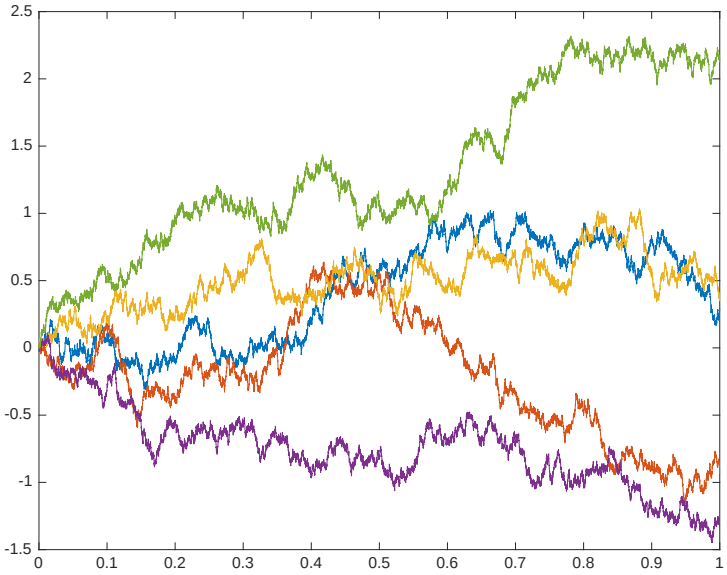
Now, for a given number $n \in \mathbb{N}$ we define a random walk (in continuous time) $Y^{(n)}$ to be:

$$Y_t^{(n)} := \frac{S_{tn}}{\sigma\sqrt{n}}, \quad t \in [0, \infty).$$

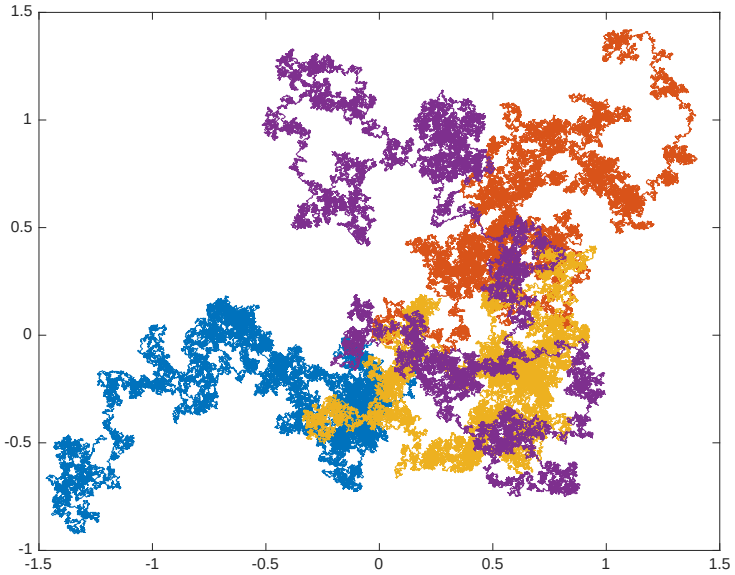
To construct $Y^{(n)}$ we have simply started from S , then rescaled time by a factor of n and then rescaled space by a factor of $\frac{1}{\sigma\sqrt{n}}$. Try to compute the expectation and the variance of $Y_t^{(n)}$. Moreover, for fixed $t \in (0, \infty)$ try to compute what is the limiting distribution of $Y_t^{(n)}$ as n goes to infinity.

Brownian motion is the continuous time, continuous space random walk obtained after taking a (non-trivial) limit of the random walks $Y^{(n)}$ as $n \to \infty$. Going into the details of what is the precise meaning of this limiting procedure is beyond the scope of the class. But for us it will be enough to say that for large enough n, the random walk $Y^{(n)}$ behaves like a Brownian motion. A Brownian motion in higher d, is simply a walk whose coordinates are independent Brownian motions in 1d.

Remark: Note that random walks (in discrete or continuous setting), are a collection of random variables. We may also see them as "random variables" on their own; under this interpretation, we may think of a trajectory as one sample from the distribution associated to such random variable. Using the function 'BrownianMotion(N , n)' with parameters $N = 5$ and $n = 10000$, I obtained five samples from Brownian motion on the time interval $[0, 1]$.



In the above, the x-axis represents time and the y-axis represents the position of the particle. Similarly, I obtained four samples from Brownian motion in 2d in the time interval $[0, 1]$,



The x, y axis are the coordinates of the position of the particle.

-S When "defining" Brownian motion, we used the symmetric random walk with steps as in (15). We called σ the variance of X_k. But isn't that variance equal to one?

-P Certainly. But, Brownian motion can be obtained using different symmetric random walks. Suppose that the steps $X_1, X_2, \ldots$ are not necessarily as in (15) but they are still

i.i.d. with mean zero (this is important) and finite variance σ^2.Then we can define $Y_t^{(n)}$ as above and as $n \to \infty$ we recover Brownian motion again.

-S I see.

- P I haven't forgotten that you owe me \$200.

3.5. **Invariant Measures of Markov chains and stochastic approximation using Markov Chains.** The set-up we consider is the following:

- A finite state space $X = \{a_1, ..., a_L\}$
- A time homogeneous Markov chain $S_0, S_1, S-2, \dots$ on X.

Such a Markov chain is characterized by:

(1) Initial state S_0 (could be chosen randomly according to some distribution on X)
(2) Transition matrix $P \in \mathbb{R}^{L \times L}$,

$$P_{ij} := \mathbb{P}(S_1 = a_j | S_0 = a_i) \quad i, j \in \{1, ..., L\}$$

Example 1: $X = \{a_1, a_2, a_3, a_4, a_5\}$

$$P = \begin{pmatrix} 0.2 & 0.4 & 0 & 0 & 0.4 \\ 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 \\ 0.4 & 0 & 0 & 0.4 & 0.2 \end{pmatrix}$$

For example : $P_{34} = 0.4$ = probability of going from state a_3 to state a_4 in one step. Notice that the sum along any row must equal 1. This is because each row represents the conditional distribution of next state given the current one.

Q: Suppose you started the Markov chain at $S_0 = a_3$. What is the probability that $S_{100} = a_i$? In other words, what is the distribution of S_{100}? Try it yourself! Simulate many chains and obtain a histogram of the values attained by S_{100}. What do you see?

- Does your answer change if $S_0 = a_1$?
- Does your answer change if S_0 is initialized randomly (always with the same distribution)?

You shouldn't see any difference! In all cases you should see that the distribution of S_{100} is roughly speaking the uniform distribution on X. We will say that the uniform distribution $\mu(a_i) = \frac{1}{5}\ \forall i$ is an **invariant measure** for the Markov chain from **Example 1**. Moreover, suppose we consider a "payoff" function $f : X \to \mathbb{R}$ (i.e. we specify $f(a_1), ..., f(a_5)$). What do you think

$$\frac{1}{n} \sum_{i=1}^{n} f(S_i),$$

converges to as $n \to \infty$? Careful!: No law of large numbers in the usual sense, because the S_i ARE NOT independent. Still it turns out that we get:

$$\frac{1}{n} \sum_{i=1}^{n} f(S_i) \underset{n \to \infty}{\longrightarrow} \sum_{l=1}^{L} f(a_l)\mu(a_l) \qquad (\mu \text{ is the invariant measure})$$

In other words, we can do stochastic approximation using Markov chains! Or at least this is what Example 1 suggests!

For general time homogeneous Markov chain (not just the one from Example 1), how do we know a distribution μ is invariant? Is μ unique? Do all Markov chains have invariant measures? Here is a criterion we will rely on for the rest of this class.

Theorem 3.4 (Detailed balance theorem:)**.** *Suppose P is a transition probability matrix satisfying:*

- *There is some $k \in \mathbb{N}$ such that P^k (multiply P k times with itself) has all strictly positive entries.*
- *There is some distribution μ ($\mu(a_1), ..., \mu(a_L)$ positive numbers adding to one) satisfying detail balance:*

$$P_{ij}\mu(a_i) = P_{ji}\mu(a_j) \quad \forall i, j$$

Then μ is the unique invariant measure of P and in particular

(1) *Regardless of where the chain starts:*

$$\mathbb{P}(S_n = a_i) \approx \mu(a_i) \qquad \text{for all } n \text{ large enough (once the chain "has mixed")}$$

(2) *We have the following stochastic approximation result:*

$$\frac{1}{n}\sum_{i=1}^{n} f(S_i) \underset{n\to\infty}{\longrightarrow} \sum_{l=1}^{L} f(a_l)\mu(a_l) \qquad \text{for arbitrary payoff function } f$$

Back to Example 1: Does $\mu(a_l) = \frac{1}{5}\ l = 1, ..., 5$ satisfy the detailed balance equations? Is it time that there is $k \in \mathbb{N}$ such that P^k has all positive entries?

What is coming next? We have seen that when a Markov chain has a unique invariant measure μ we can use approximate integrals with respect to μ using a realization of the Markov chain. How about the other way around? Namely, you give me a μ and I design a Markov chain with invariant measure μ (we say μ is a *target* distribution. This is what MCMC (Monte Carlo Markov chain) is about and in particular what the *Metropolis-Hastings algorithm* achieves.

3.6. **Monte Carlo Markov Chain (MCMC) and the Metropolis-Hastings algorithm.** The goal for this section is to design a Markov chain whose invariant measure is a target measure μ, so that we can approximate "integrals" with respect to μ using randomness.

Set-up: We will consider the following objects:

- A discrete, finite state space $X = \{a_1, \ldots, a_L\}$.
- A *target* distribution μ with $\mu(a_1), \ldots, \mu(a_L)$ such that:
  (1) $\mu(a_l) \geq 0,\ \forall\, l = 1, \ldots, L$,
  (2) $\sum_{l=1}^{L} \mu(a_l) = 1$.
- A pay-off function $f : X \to \mathbb{R}$.

We want to estimate the following expression:

$$\sum_{l=1}^{L} f(a_l)\mu(a_l)$$

Remark 3.5. Since this expression is just a sum, it should be easy to estimate it (or compute it) and so it seems to be a silly task. However, we will later consider continuous distributions in high dimensional spaces where estimating integrals may be difficult. Similar ideas to the ones we explore now will apply in continuous settings as well. In other words, the discrete case will give us some intuition that we will later use for the continuous case.

Remark 3.6. There may be many Markov chains with invariant measure μ. How do we choose the Markov chain? That's something we will discuss later on.

We will now consider a particular family of Markov chains constructed using the *Metropolis-Hastings* algorithm. The algorithm has the following inputs:

- A target measure μ on a finite state space X, defined as before.

- A "proposal distribution" Q, which is simply a transition probability matrix on X. We will adopt the notation:

$$Q(a_i, a_j) = Q_{ij} = \text{ According to } Q, \text{ how likely it is to go from } a_i \text{ to } a_j$$

To generate the Markov chain, we will first set $S_0 \in X$ (determined randomly or deterministically). In order to produce S_{k+1} from previously obtained $S_0, S_1, \ldots, S_k$, we will follow these steps:

(1) Propose $Y \sim Q(S_k, \cdot)$, i.e. Y is a sample from X as specified by the distribution given by the row of Q that corresponds to state S_k.
(2) Compute the acceptance probability:

$$\alpha(S_k, Y) := \min\left\{1, \frac{Q(Y, S_k)\mu(Y)}{Q(S_k, Y)\mu(S_k)}\right\}$$

(3) Decide whether to accept the proposal by "flipping a coin" with bias $\alpha(S_k, Y)$:

$$S_{k+1} := \begin{cases} Y, & \text{with prob. } \alpha(S_k, Y) \\ S_k, & \text{otherwise} \end{cases}$$

Remark 3.7. Take a look at the acceptance probability. Notice that when $\mu(Y)$ gets larger than $\mu(S_k)$, the larger the acceptance probability gets; this makes sense because we would like the value of Y to be chosen more often than S_k (to be able to exploit the regions where μ is large).

Remark 3.8. Notice that when $Q(Y, S_k)$ gets larger than $Q(S_k, Y)$),the larger the acceptance probability gets. This makes sense because we would like to accept proposals that are difficult to reach (to be able to explore the whole state space).

Now we wonder whether μ is invariant for the Markov chain described above (the one that is generated by the algorithm). First, let's ask what is the transition probability matrix P associated to the Metropolis-Hastings algorithm . Our claim is that P is defined as follows:

$$P(a, b) = \alpha(a, b)Q(a, b) + \left(\sum_{l=1}^{L}(1 - \alpha(a, a_l))Q(a, a_l)\right)\mathbb{1}_{a=b}$$

Indeed if $a \neq b$, we have that $P(a, b) = \alpha(a, b)Q(a, b)$, since to go from a to b we need to propose b (given that we are at a), which occurs with probability $Q(a, b)$, and later to accept b, which occurs with probability $\alpha(a, b)$. If $a = b$:

$$P(a, a) = \alpha(a, a)Q(a, a) + \sum_{l=1}^{L}(1 - \alpha(a, a_l))Q(a, a_l)$$

Note that $\alpha(a, a) = 1$, since if a is proposed, then both accepting and rejecting the proposition will lead to the new state being a. We have that a goes to a if:

- we propose a (which happens with probability $Q(a, a)$), or
- if we propose any state a_l (probability: $Q(a, a_l)$) and subsequently reject it (probability: $(1 - \alpha(a, a_l))$)

Let use demonstrate that our claim is true, i.e. that P is ergodic with unique invariant measure μ, by applying the detailed balance theorem. If we assume that Q has all positive entries then so does P. We must show that the detailed balance equations hold:

$$P(a, b)\mu(a) = P(b, a)\mu(b), \quad \forall a, b \in X$$

Note that the equality is trivial when $a = b$, so it is sufficient to consider the case $a \neq b$. Now,

$$\mu(a)P(a,b) = \alpha(a,b)Q(a,b)\mu(a)$$
$$\mu(b)P(b,a) = \alpha(b,a)Q(b,a)\mu(b)$$

Suppose that $\alpha(a,b) < 1$. Then:

$$\alpha(a,b) = \frac{Q(b,a)\mu(b)}{Q(a,b)\mu(a)} < 1 \implies \frac{Q(a,b)\mu(a)}{Q(b,a)\mu(b)} > 1$$

Recall that:

$$\alpha(b,a) = \min\left\{1, \frac{Q(a,b)\mu(a)}{Q(b,a)\mu(b)}\right\}$$

Since the second expression is greater than 1, we have that $\alpha(b,a) = 1$. Thus:

$$\mu(a)P(a,b) = \frac{Q(b,a)\mu(b)}{Q(a,b)\mu(a)} \cdot Q(a,b)\mu(a) = Q(b,a)\mu(b)$$
$$\mu(b)P(b,a) = \alpha(b,a)Q(b,a)\mu(b) = Q(b,a)\mu(b),$$

so we have demonstrated detailed balance for the case $\alpha(a,b) < 1$. If on the other hand $\alpha(a,b) = 1$, then:

$$\frac{Q(b,a)\mu(b)}{Q(a,b)\mu(a)} \geq 1$$

from which it follows that:

$$\frac{Q(a,b)\mu(a)}{Q(b,a)\mu(b)} \leq 1,$$

so:

$$\alpha(b,a) = \frac{Q(a,b)\mu(a)}{Q(b,a)\mu(b)}$$

Thus:

$$\mu(a)P(a,b) = \alpha(a,b)Q(a,b)\mu(a) = Q(a,b)\mu(a)$$
$$\mu(b)P(b,a) = \frac{Q(a,b)\mu(a)}{Q(b,a)\mu(b)}Q(b,a)\mu(b) = Q(a,b)\mu(a)$$

From the above we conclude that $P(a,b)\mu(a) = P(b,a)\mu(b)$ in all cases. Therefore, we conclude that the Metropolis-Hastings algorithm has μ as a unique invariant measure.

3.7. **Metropolis-Hastings for continuous distributions and the random walk sampler.** We now change our set-up and consider a distribution μ with a density function $\rho : \mathbb{R}^d \to [0,\infty)$. Our goal is to approximate

$$\int_{\mathbb{R}^d} f(x)\rho(x)\,dx$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is some arbitrary function, using Markov chains (now in continuous state space $\mathbb{R}^d$!).

For Metropolis-Hastings in this new setting we still need a proposal distribution Q. Q can be thought as a joint distribution with conditional **densities** $q(y|x)$ (See examples below). The algorithm looks the same as in thediscrete case (except that now we work with p.d.fs instead of p.m.fs). In other words: We construct $S_0, S_1, S_2, \cdots \in \mathbb{R}^d$. To define S_{k+1} from S_k we do:

(1) Propose $Y \sim q(\cdot|S_k)$ (sample from the conditional distribution $q(\cdot|S_k)$).

(2) Compute $\alpha(S_k, Y)$ where

$$\alpha(x, y) := \min\left\{1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right\}.$$

(3) Set

$$S_{k+1} = \begin{cases} Y & \text{with prob. } \alpha(S_k, Y) \\ S_k & \text{with prob. } 1 - \alpha(S_k, Y). \end{cases}$$

Example 1: Random walk proposal (in 1d).

$$q(y|x) = \frac{1}{\sqrt{2\pi b^2}} \exp\left(-\frac{(x-y)^2}{2b^2}\right)$$

To sample from $q(\cdot|x)$ it is enough to take

$$Y = x + \xi,$$

where $\xi \sim N(0, b^2)$. Notice that for this choice of proposal the acceptance probability in the Metropolis-Hastings algorithm simplifies to:

$$\alpha(x, y) = \min\left\{1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right\} = \min\left\{1, \frac{\rho(y)}{\rho(x)}\right\}.$$

Q: What is the role of b?

Example 2: Independent sampler proposal. We consider:

$$q(y|x) = g(y),$$

where g is another density we know how to get samples from. The acceptance probability in this case becomes:

$$\alpha(x, y) = \min\left\{1, \frac{g(x)\rho(y)}{g(y)\rho(x)}\right\}$$

Q: How to pick g?

The idea of producing **trace plots** (plot of k against $f(S_k)$) is useful because it provides qualitative evidence that a Markov chain has *mixed* and it also allows us to remove the *initialization bias*. To get rid of the initialization bias it makes sense to use:

$$\frac{1}{n-M} \sum_{i=M+1}^{n} f(S_i)$$

as an approximation of $\int_{\mathbb{R}^d} f(x)\rho(x)\,dx$, where M can be tuned by looking at the trace plot.

Back to the question of how to choose the proposal. We will focus on the random walk sampler : How do we pick b? There is a trade-off between exploration and exploitation. In particular:

- Very small b
  - No exploration
  - A lot of acceptance
- Very large b
  - Too much exploration and no exploitation.
  - A lot of rejections.

The idea is to choose b in between. One possible way to **tune** b is to select it so that the **average** acceptance rate (in the Metropolis-Hastings algorithm) is something like 23.4%. We will discuss more about this magical number in Homework 6.

3.8. **Summary.** In this section we discussed Markov chains and some of their properties. Random walks are particular examples of time homogeneous Markov chains; we studied the notion of recurrence, as well as the computation of exit probabilities. We then studied the notions of invariant measure and limiting distubution of a Markoc chain; our ultimate goal was to introduce the concept of MCMC computing and the Metropolis-Hastings algorithm. This algorithm can be used for stochastic approximation when it is difficult to generate samples from a target distribution μ.

4. Graphical Models

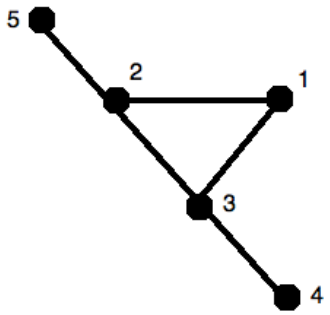
So far we have considered collections of random variables $(Y_1, \ldots, Y_n)$ that have i.i.d. structure or that have Markovian structure. In this section we consider more general structures. The idea is to associate a graph to the collection of variables $(Y_1, \ldots, Y_n)$; the structure of the variables is demonstrated with a graph. The ultimate goal of doing this, is to: make random simulations, compute expected values, compute marginal/conditional probabilities, and determine most likely configurations. We will consider two different types of graphical models : Gibbs random fields (GRFs) and Bayes nets (BN).

4.1. **Gibbs random fields (GRFs).** We first need to introduce the concept of undirected graphs.

Definition 4.1. *An undirected graph is a pair of objects (V, E) where:*

- *V is a set (the vertices, or nodes of the graph).*
- *E is the set of edges of the graph. An edge is simply an unordered pair of vertices.*

For example, we may consider $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{3, 2\}, \{1, 3\}, \{3, 4\}, \{5, 2\}\}$



We also need to introduce the concept of *clique*.

Definition 4.2. *Given a graph $G = (V, E)$, a nonempty subset C of V, is said to be a clique of G if every two vertices in C are connected by an edge.*

In the graph considered above, the set $C = \{1, 2, 3\}$ is a clique of G, as there is an edge between any two of the vertices in C. On the other hand, the set $\{2, 3, 5\}$ is not a clique of G because there is no edge between the vertices 3 and 5.

Given a graph G, we will denote by $\mathcal{C}(G)$ the set of all cliques of G. Notice that $\mathcal{C}(G)$ is a set of subsets of V.

In the example above, some of the elements of $\mathcal{C}(G)$ are : $\{1\}, \{2,3\}, \{1,2,3\}$. Observe that the only clique in G containing three or more elements is the set $\{1,2,3\}$.

Let V be a set of vertices and consider a collection of random variables $\{X_v\}_{v\in V}$; that is, for every vertex in V we associate a random variable. These random variables are not assumed to be independent or identically distributed. We will assume for simplicity that the variables X_v are all discrete. Let us now introduce some notation:

- We denote by p_V, the joint p.m.f. of the variables $\{X_v\}_{v\in V}$.
- For any sequence of vertices $A = (v_1, \ldots, v_k)$, we denote by X_A the vector $(X_{v_1}, \ldots, X_{v_k})$.
- x_v will denote a possible outcome or configuration of the variable X_v. Likewise, x_A will denote a possible outcome or configuration for X_A.
- We denote by p_v the marginal distribution of X_v and by p_A the joint p.m.f. of the variables in X_A.

With the above definitions and notation we can now define the notion of Gibbs random fields.

Definition 4.3. *Given a graph $G = (V, E)$, we say that the random variables X_V factor as a Gibbs random field **with respect to** G, if the probability mass function p_V can be written as*

$$p_V(x_V) = \frac{1}{Z} \prod_{C\in\mathcal{C}(G)} \phi_C(x_C),$$

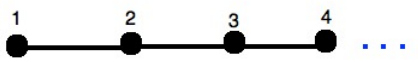
where Z (normalization constant) is a positive constant, and where the functions ϕ_C take positive values.

We will also say that $\{X_v\}_{v\in V}$ respects the graph G. The functions ϕ_C are called clique potentials.

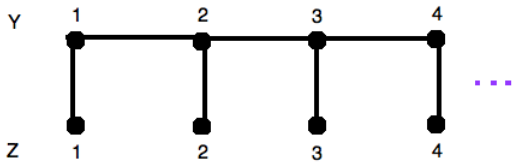
Remark: Although we introduced the notion of Gibbs random fields in the context of discrete random variables, this notion extends in the obvious way to random variables with a joint density. The idea is to replace p.m.f.s by pdfs in all of the above definitions.

Example:(Independent variables) Suppose that $X_1, X_2, \ldots, X_n$ are pairwise independent. Then, the variables respect the the trivial graph on $\{1, \ldots, n\}$: no edges between the nodes.

Example: (Markov chains) Suppose that the variables $X_1, \ldots, X_n$ have a Markovian structure. Then the variables $X_1, \ldots, X_n$ respect the graph:



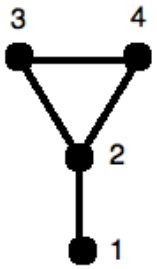
Example:(Hidden Markov chain) Here we assume we have two type of variables 'latent' variables and observed variables. Let us denote by $Y_1, \ldots, Y_n$ the latent variables and by $Z_1, \ldots, Z_n$ the observed variables. We assume that the variables $Y_1, \ldots, Y_n$ have a Markovian structure, and that given Y, the Zs are all independent. Then the variables $X_1, \ldots, X_n$ respect the graph :



Example: Consider the joint distribution on X_1, X_2, X_3, X_4 given by

$$p_V(x_V) = \frac{1}{Z} x_1^{x_2} \cos(x_2 x_3 x_4).$$

In the above, we are considering $V = \{1, 2, 3, 4\}$. Then, X_V respects the graph G:



To see this, first observe that the cliques of G are:

$$\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 2\}, \{2, 3, 4\}.$$

We may take $\phi_{(1,2)}(x_1, x_2) = x_1^{x_2}$ and $\phi_{(2,3,4)}(x_2, x_3, x_4) = \cos(x_2 x_3 x_4)$. And we may take $\phi_C(x_C) = 1$ for the rest of the cliques. Then, it is clear that

$$p_V(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C).$$

We conclude that X_V indeed respects G.

Some relevant observations about GRFs are the following.

(1) The individual clique potentials may **not** be probability distributions.
(2) If X_V respects the graph $G = (V, E)$ then X_V also respects the graph $G' = (V, E')$ where $E \subseteq E'$. In particular, a set of random variables X_V always respects the complete graph on the set of vertices V.

The previous observation shows that a set of variables X_V may respect different graphs, but we can always talk about minimal graphs respected by X_V. This is a natural notion to consider.

Definition 4.4. *We say that a graph $G = (V, E)$ is minimal for X_V if*

(1) *X_V respects G.*
(2) *X_V does not respect (V, E') for any $E' \subsetneq E$.*

In general minimal graphs may be non-unique. Nevertheless, under some additional conditions minimal graphs are indeed unique. One such condition is the following: Suppose that Ω_v denotes the set of possible outcomes for the variable X_v. Denote by Ω_V

the collection of tuples $x_V = (x_v)_{v \in V}$, where $x_v \in \Omega_v$ for every $v \in V$. If it is true that for every x_V, $p_V(x_V) > 0$, then X_V has a unique minimal graph. If you are interested in this result, perhaps you can look up the notion of Markov Random Fields and the Hammersley-Clifford theorem. We will say something about Markov Random Fields latter on.

4.2. **Independence, conditioning, and marginalizing rules.** -S : Say that X_V respects some graph G. What does that tell us about the variables?

-P : Let us first start with what it does *not* tell us!

If there is no edge between nodes v and w then the variables X_v and X_w may be dependent.

To illustrate that such situation can happen we consider the following example. Suppose that I flip a fair coin. If the outcome of the coin is H then I roll two 6-sided (fair) dice. If the outcome of the coin is T, then I roll two 1000-sided (fair) dice. We set the variable $X_1$to be equal to one if the coin's outcome is H and to be equal to zero if the coin's outcome is T. We set X_2 to be the outcome of the first die and set X_3 to be the outcome of the second die. Notice that given the outcome of X_1, the variables X_2, X_3 are independent, but it is not true that the variables X_2, X_3 are independent. Intuitively, if X_2 is greater than 6, then automatically this gives a hint to what values to expect for X_3: large numbers, because $X_2 > 6$ means that X_3 was obtained by rolling a 1000-sided dice!

We claim that the minimal graph for (X_1, X_2, X_3) does not have an edge between 2 and 3. Indeed, one can check that in this case, the joint distribution is given by:

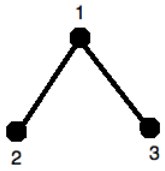
$$P_{1,2,3}(x_1, x_2, x_3) = \frac{1}{2}\left(\frac{\mathbb{1}_{x_2 \leq 6}}{6}\right)^{x_1}\left(\frac{\mathbb{1}_{x_3 \leq 6}}{6}\right)^{x_1}\left(\frac{\mathbb{1}_{x_2 \leq 1000}}{1000}\right)^{1-x_1}\left(\frac{\mathbb{1}_{x_3 \leq 1000}}{1000}\right)^{1-x_1}$$

which can be written as:

$$P_{1,2,3}(x_1, x_2, x_3) = \frac{1}{2}\alpha(x_1, x_3) \cdot \beta(x_1, x_2),$$

where $\alpha(x_1, x_3) = \left(\frac{\mathbb{1}_{x_3 \leq 6}}{6}\right)^{x_1} \cdot \left(\frac{\mathbb{1}_{x_3 \leq 1000}}{1000}\right)^{1-x_1}$ and β is defined analogously.

From the above formula we can conclude that any graph respected by X_1, X_2, X_3 must contain the edges $\{1, 2\}$ and $\{1, 3\}$. This is because the functions α and β can not be written as a product of individual functions of x_1, x_2 and x_3. We conclude then that the minimal graph for X_1, X_2, X_3 is:



P : Comments?

S: My only concern is whether the above example can be carried out in reality... I am particularly concerned about how to construct a 1000-sided fair dice.

P Look up 'platonic polyhedrons'. Proof that there is no regular polyhedron with 1000 sides: By contradiction. If there existed such a thing, players of 'Dungeons and Dragons' would definitely have one. They don't have one. Q.E.P.D.

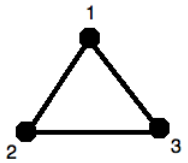
If there is an edge between nodes v and w, then the variables X_v and X_w may be independent

To illustrate that such situation can happen we consider the following example. Suppose that I flip two fair coins. If both outcomes turn out to be H then I flip another fair coin. If not, then I flip a biased coin with probability of H equal to $2/3$. We define X_i to be 1 if the i-th coin flip is H and zero otherwise. Note that the variables X_1 and X_2 are independent. Nevertheless, we show that the minimal graph for X_1, X_2, X_3 is the complete graph on $\{1, 2, 3\}$; in particular there is an edge between 1 and 2.

Indeed, the joint distribution in this case can be written as:

$$p_{1,2,3}(x_1, x_2, x_3) = \frac{1}{4}\left((2/3)^{x_3} \cdot (1/3)^{1-x_3}\right)^{1-x_1x_2} \cdot (1/2)^{x_1x_2} .$$

Note that no matter what we do, in the previous expression we can not separate the term $(2/3)^{x_1x_2x_3}$ into individual functions of x_1, x_2 and x_3. This shows that the minimal graph for X_1, X_2, X_3 is the complete graph on $1, 2, 3$.



- S : Cool, but knowing what a graph can not tell us, does not tell me much...

Proposition 4.5. *(Independence rule.) Suppose that X_V respects the graph $G = (V, E)$. Suppose that the nonempty set $A \subsetneq V$ is such that there are no edges between nodes in the set A with nodes in the set A^c. Then, X_A and X_{A^c} are independent.*

-P What do you think?

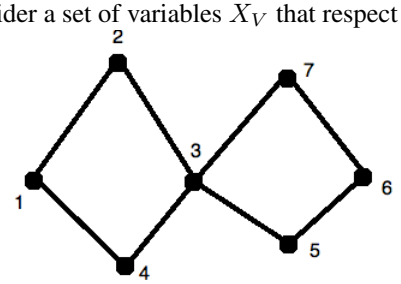
-S That helps.

- P That helps, but it is not a very useful property on its own. After all, we are trying to understand collections of variables that may have a more intricate dependence structure. The independence rule is particularly useful when combined with the next rule.

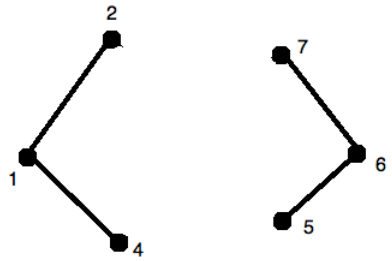
Proposition 4.6. *(Conditioning rule.) Suppose that X_V respects the graph $G = (V, E)$. We consider the conditional distribution of X_A **given** that $X_{A^c} = x^*_{A^c}$; we denote it by $P_{A|X_{A^c}=x^*_{A^c}}$. Then, this joint (conditional) distribution respects the graph $G' = (A, E')$, where E' is obtained from E by removing all edges involving nodes in A^c, i.e.,*

$$E' := \{\{u, v\} \in E \ : \ u, v \in A\} .$$

Example: Consider a set of variables X_V that respect the graph depicted below:



If we condition on X_3 we conclude that the variables $X_{V\setminus\{3\}}$ respect the graph:

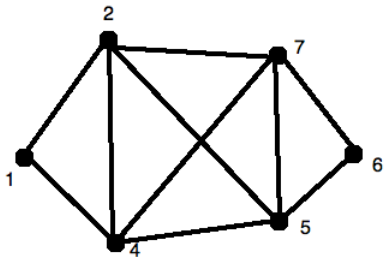


Combining with the independence rule, we conclude that (X_1, X_2, X_4) and (X_5, X_6, X_7) are **conditionally independent** given X_3. Careful! we can not conclude that (X_1, X_2, X_4) and (X_5, X_6, X_7) are independent, only conditionally independent given X_3. Think about this: two variables may be independent but not conditionally independent given a third one. Conversely, two variables may be conditionally independent given a third one, but not independent.

Now we would like to move to the marginalizing rule. This rule will say something about the marginal distribution of X_A.

Proposition 4.7. *(Marginalizing rule.) Suppose that X_V respects the graph $G = (V, E)$. Then, the variables X_A respect the graph $G' = (A, E')$, where E' is defined as follows: if $u, v \in A$ and $\{u, v\} \in E$ then $\{u, v\} \in E'$, or if $u, v \in A$ and there is a path between u, v using edges in E and only visiting nodes in A^c, then $\{u, v\} \in E'$. In words, we keep edges between nodes in A that were already there, and add edges between nodes in A when the nodes can be connected by a path with edges in E and nodes in A^c.*

Example: In the example above, the marginal of $X_{V\setminus\{3\}}$ respects the graph:



-S: I have a question regarding minimal graphs and the various rules we have learned about. Suppose that G is a minimal graph for X_V. Consider $A \subseteq V$ and consider X_A's marginal. We use the marginalization rule to obtain the graph G' as defined in Proposition 4.7. Is this graph minimal for X_A?

-P : Not necessarily. For example, in the situation described when showing that having an edge does not imply dependence, we had variables X_1, X_2, X_3 such that X_1 and X_2 were independent and such that minimal graph was the complete graph on $1, 2, 3$. If we applied the marginalization rule we would obtain a graph with one edge between 1 and 2. It is clear that this can not be the minimal graph for X_1 and X_2 because they are actually independent. In general, the marginalization rule is a worst case scenario: the graph given by the marginalization rule is guaranteed to be respected by the variables , but in some special cases (like in the previous example) some edges in the graph may be unnecessary. The same is true for the conditioning rule.

4.3. **Computing with GRF.** Detailed examples for this section, can be found in the document 'DP in a nutshell' on Canvas. Here we only review some definitions and add some remarks.

When the joint distribution of a set of random variables has a factorization into clique potentials, we may use that factorization to compute things like normalization constants, marginals, expectations, and most likely configurations, in an efficient way (at least more efficiently than using a brute force approach). Suppose that the joint distribution p_V respects the graph $G = (V, E)$. That is, we assume that

$$p_V(x_V) := \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C).$$

An ordering for the vertices V, determines an algorithm to find normalization constants, marginals, expectations, and most likely configurations. Take for example the problem of finding Z and suppose we have ordered the vertices in V as $v_1, \ldots, v_n$. Then,

$$Z = \sum_{x_V} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C) = \sum_{x_{v_n}} \sum_{x_{v_{n-1}}} \cdots \sum_{x_{v_1}} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C).$$

From now on let us just write x_i for x_{v_i}.

When evaluating the sum over x_1, some of the terms $\phi_C(x_C)$ don't depend on x_1 and hence can be extracted from the sum. Those terms that depend on x_1 will be added over x_1; the resulting expression will be a function of the variables different from x_1 that also appeared in those terms. We have thus computed the sum over x_1. We will then extract from the sum over x_2, those terms that don't depend on x_2, then compute the sum of the remaining terms just as described for the first sum. We continue in this fashion...

To describe abstractly the previous intuitive approach, we introduce the following notions.

Definition 4.8. *For label k (in particular we have fixed an ordering/labeling for the variables), we define the boundary of k to be the set:*

$$b_k = \{l > k \ : \ \exists C \in \mathcal{C}(G) \text{ s.t. } l \in C \text{ and } C \cap \{1, \ldots, k\} \neq \emptyset\}.$$

That is, b_k is the set of labels after k, who share a clique with k or with a label before k.

Definition 4.9. *For label k (in particular we have fixed an ordering/labeling for the variables), we define the set of innovation cliques at step k, to be the set of cliques:*

$$\mathcal{C}_k := \{C \in \mathcal{C}(G) \ : \ k \in C \text{ and } C \cap \{1, \ldots, k-1\} = \emptyset\}.$$

That is, $\mathcal{C}_k$ is the set of cliques, that contain k and do not contain any of the labels before k.

Now we can describe how to obtain Z. Define

$$T_1(x_{b_1}) := \sum_{x_1} \prod_{C \in \mathcal{C}_1} \phi_C(x_C).$$

Notice that this is the first sum described earlier. Having defined $T_{k-1}(x_{b_{k-1}})$, we define

$$T_k(x_{b_k}) := \sum_{x_k} \prod_{C \in \mathcal{C}_k} \phi_C(x_C) T_{k-1}(x_{b_{k-1}}).$$

Note that implicitly, we are saying that the right hand side of the above expression only depends on x_{b_k}; a fact that would need a proof. In any case, we can obtain iteratively $T_1, \ldots, T_{n-1}, T_n$. What is T_n? Well, T_n is actually a number; this number is Z.

S: So the method just described is essentially an 'use the factorization and then extract what is constant' approach?

P: Yes, that is correct.

S: Ok. But I don't see how this helps.

P: Let us understand why the factorization is important. Suppose that the variable x can take the values $1, \ldots, 10$ and that the variable y can take the values $1, \ldots, 10$ as well. In general, to specify a function $f : \{1, \ldots, 10\} \times \{1, \ldots, 10\} \to \mathbb{R}$, you would need to specify the value of $f(x, y)$ for every possible combination of (x, y). That is you would need to specify 100 numbers and then you can store those values in a look up table. Every time I give you an x and a y you can check your table and return the value $f(x, y)$.

Now I claim that if $f(x, y)$ was very special, in particular, if $f(x, y) = g(x) \cdot h(y)$, then you wouldn't need to store 100 values but only 20 in order to specify the function f completely. Indeed, in this case, you would only need to specify g (store 10 values) and h (store 10 values). With look up tables for g and h, we can recover completely f: I give you x and y, you look for $g(x)$ and $h(y)$, you multiply the two numbers and voila!

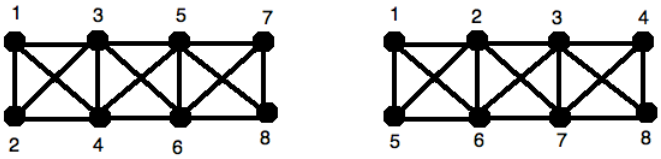
S: Touché.

The functions $T_1, \ldots, T_{n-1}, T_n$ can then be represented as look up tables (that explains why we use T to denote them). Once we compute T_1 (and we store its values), we may use T_1 to obtain T_2 (we store its values). Then we can use T_2 to obtain T_3 ...etc. The point is that we base our computations on previous computations, we don't have to compute everything again: we use the look up tables.

S: So if I understand correctly, the 'memory cost' associated to the computation of Z is determined by how big the look up tables T_i are?

P: Yes.

Example: Suppose that a set of variables respects the graph depicted below. We assume that the variables can take values in $\{1, \ldots, 10\}$. We have chosen two different orderings for the vertices.



To understand the memory cost of computing Z for each of the orderings, we simply need to understand the sets b_k in each case. For the first ordering,

$$b_1 = \{2, 3, 4\}, b_2 = \{3, 4\}, b_3 = \{4, 5, 6\}, b_4 = \{5, 6\},$$
$$b_5 = \{6, 7, 8\}, b_6 = \{7, 8\}, b_7 = \{8\}, b_8 = \emptyset.$$

For the second ordering,

$$b_1 = \{2, 5, 6\}, b_2 = \{3, 5, 6, 7\}, b_3 = \{4, 5, 6, 7, 8\}, b_4 = \{5, 6, 7, 8\},$$
$$b_5 = \{6, 7, 8\}, b_6 = \{7, 8\}, b_7 = \{8\}, b_8 = \emptyset.$$

What about the computational cost associated to an ordering? Let us for simplicity assume that we only take into account the operation of addition (for simplicity we assume that multiplications are done instantaneously). For every k, how many operations do we use to obtain the look up table T_k? Well, for every x_{b_k} we need to sum 10 numbers to obtain $T_k(x_{b_k})$ and there are $10^{|b_k|}$ possible configurations x_{b_k} ; here $|b_k|$ stands for the number of elements in b_k. Therefore, the number of operations (remember we are not counting products) needed to compute T_k is $10 \cdot 10^{|b_k|}$. In order to obtain Z we need to compute $T_1, \ldots, T_{n-1}$, so that in the end we can compute $Z = T_n$. The total number of operations is then:

$$\sum_{k=1}^{n} 10 \cdot 10^{|b_k|}.$$

Remark: From the above analysis we deduce that it would be desirable to find an ordering that minimizes the size of its corresponding boundaries. Unfortunately, the problem of finding such ordering is NP. Nevertheless, there are some ways to 'relax' that optimization problem. The result is an ordering that in many cases is good enough.

Remark: Note that in order to get an idea of the computational complexity of computing a normalization constant or the most likely configuration for a given joint distribution p, we only need to know the dependence structure of the variables, which is represented in the graph G. That, is we do not really use the exact form for the clique potentials and we only use the factorization (summarized in the graph). This is one good reason for why it is useful to consider the graph: it reveals many qualitative (and even in this case quantitative) properties of the joint distribution p_V.

When computing expectations and most likely configurations, we follow a very similar approach to the ' use factorization and extract what is constant' approach, see 'DP in a nutshell' for further details. For computing marginals, it is important to clarify certain aspects that now we describe.

Given an ordering for the variables, we may use similar ideas to compute the marginals $p_{k:n}$ for every k. Here we denote by $p_{k:n}$ the marginal of the variables $X_k, \ldots, X_n$. It is not hard to show that:

$$p_{k+1:n}(x_k, \ldots, x_n) = \frac{1}{Z} T_k(x_{b_k}) \prod_{C \in \mathcal{C}(G) \,:\, C \cap 1, \ldots, k = \emptyset} \phi_C(x_C). \tag{16}$$

Let us illustrate how we can obtain the above formula for $p_{2:n}$. Indeed, the marginal $p_{2:n}(x_2, \ldots, x_n)$ is obtained from $p(x_1, \ldots, x_n)$ after summing over x_1. That is,

$$p_{2:n}(x_2, \ldots, x_n) = \sum_{x_1} p(x_1, \ldots, x_n) = \frac{1}{Z} \sum_{x_1} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C).$$

Now from the sum, we can extract everything that does not depend on x_1, in particular we can extract the terms $\phi_C(x_C)$ for which $C \cap \{1\} = \emptyset$. Thus,

$$p_{2:n}(x_2, \ldots, x_n) = \prod_{C \cap \{1\} = \emptyset} \phi_C(x_C) \sum_{x_1} \prod_{1 \in C} \phi_C(x_C).$$

But the set of cliques for which $1 \in C$ is precisely what we called $\mathcal{C}_1$. Hence the sum in the above expression is $T_1(x_{b_1})$. Thus,

$$p_{2:n}(x_2,\dots,x_n) = \prod_{C\cap\{1\}=\emptyset} \phi_C(x_C) T_1(x_{b_1}).$$

This coincides with (16) when $k=1$.

- S Once again the 'extract what is constant' approach.

- P Indeed.

Remark: Note that if we computed Z using a certain ordering for the variables, and we stored the look up tables $T_1,\dots,T_n$, then we would have everything we need to be able to compute $p_{k:n}$ for any possible configuration: we just need to multiply a bunch of given numbers. In other words, if we were interested in obtaining Z and then obtaining one of the marginals $p_{k:n}$, it is perhaps a good idea to store the look up tables. The bottom line is, if you already have the look up tables, then you also have the marginals.

- S The above procedure is useful if we wanted to compute $p_{k:n}$ when an ordering has been fixed. But how can we compute things like $p_{8,3}$? That is the marginal of X_8 and X_3?

-P With the tools we have learned, there is no other way than to relabel the variables and start from scratch, leaving the variables X_8, X_3 for last. But there are some more advanced techniques that may help! For example, look up 'tree junctions'.

Let us move now to sampling. We will see now that if we have computed the marginals $p_{k:n}$ (which in turn means we have computed the look up tables), then in principle we can sample from p. Note that this means that in principle any ordering of the variables will give us a way to sample from p. Only in principle, because as we explored in Section 1, some distributions are easier to sample from than others.

First note that using (16) for $k=n-1$, we can obtain the marginal for the last variable X_n:

$$p_n(x_n) = \frac{1}{Z}\phi_{\{n\}}(x_n) T_{n-1}(x_n).$$

On the other hand, for every $k=1,\dots,n-1$, we can write the conditional distribution of X_k given that $X_{k+1}=x_{k+1},\dots,X_n=x_n$ as:

$$p_{k|k+1:n}(x_k) = \frac{p_{k:n}(x_k,x_{k+1},\dots,x_n)}{p_{k+1:n}(x_{k+1},\dots,x_n)}.$$

To sample from $X_1,\dots,X_n$ we do the following:

Sample from p_n to obtain $\hat{x}_n$. Use this $\hat{x}_n$ to sample from $p_{n-1|n}$: we obtain $\hat{x}_{n-1}$. Use $\hat{x}_{n-1},\hat{x}_n$ to sample from $p_{n-2|n-1:n}$ and obtain $\hat{x}_{n-2}$. Continue in this fashion. At the end we will have a list of values $\hat{x}_1,\dots,\hat{x}_n$. This is one sample from the joint distribution p.

Final remarks When we get to study Bayes nets, we will see that in some cases there are other ways to sample from p, that do not require so many computations: the approach we have considered, using DP and the GRF structure, relies on computing the look up tables $T_1,\dots,T_n$.

Also, there may be some numerical instabilities when computing the most likely configuration or the normalization constant using the approach described in this section. This is because computers have their limitations when working with very small or very large numbers. For the 'finding the most likely configuration' task, working with logarithms usually helps (see 'DP in a nutshell') . On the other hand, for the ' finding the normalization constant' task, an iterative normalization approach usually helps (see Section d in 'DP in nutshell').

4.4. **Markov Random fields.** The last graphical model we will consider is Markov random fields. This model is very similar to GRFs and in fact under some conditions the two concepts are equivalent.

Definition 4.10. *Let $G = (V, E)$ be an **undirected** graph. For $v \in V$, we denote by N_v the set of neighbors of v. That is,*

$$N_v := \{u \in V \ : \ \{u, v\} \in E\}$$

Definition 4.11. *We say that a set of random variables $\{X_v\}_{v\in V}$ is a Markov random field with respect to an **undirected** graph $G = (V, E)$ if for every $v \in V$*

$$p_{v|V\setminus\{v\}} = p_{v|N_v}.$$

In words, if the conditional distribution of X_v given the rest of the variables is equal to the conditional distribution of X_v given its neighbors.

Theorem 4.12. *(Hammersley-Clifford Theorem part 1). If $\{X_v\}_{v\in V}$ are a GRF with respect to $G = (V, E)$, then they are also a Markov random field with respect to the same graph. That is, GRFs with respect to G are MRFs with respect to G.*

The converse is true under some additional conditions. Let us denote by Ω_v the possible outcomes for the variable X_v.

Theorem 4.13. *(Hammersley-Clifford Theorem part 2). Suppose that $\{X_v\}_{v\in V}$ are a MRF with respect to $G = (V, E)$. In addition, assume that for every configuration $x_V = \{x_v\}_{v\in V}$ with $x_v \in \Omega_v$ for every $v \in V$, we have $p_V(x_v) > 0$. Then, the variables $\{X_v\}_{v\in V}$ are a GRF with respect to the same graph.*

4.5. **Bayes Nets.** -S : You talked about Bayes 'something', as a way to sample from some complicated joint distribution without computing too much...

P : Er... We need to be more precise than that. However, loosely speaking, if we can write a joint distribution as a product of known conditional distributions, then it is "easy" to sample from such joint distribution.

S: What you are talking about contrasts with the factorization into clique potentials from last section, where individual clique potentials were not necessarily marginals or conditional distributions. Correct?

P: Yes. Let us be more precise now.

Definition 4.14. *A directed graph G, consists of a set of vertices V and a set of directed edges E. The set E is a subset of ordered pairs of elements in V. For this reason, we may write an element in E as : $u \to v$.*

The vertices of a graph will represent random variables (as in the GRF setting), and the arrows some notion of causality. Indeed, we want to be able to say the following: if $u \to v$ is an edge of the graph, then in order to sample from the variable represented by v, we will need to sample first from the variable represented by u. With this in mind, we should restrict our attention to graphs that have no loops (so as to rule out an egg-chicken phenomenon).

Definition 4.15. *A directed graph $G = (V, E)$ is acyclic, if it does not have loops. By loop we mean a path $v_1 \to v_2 \to \cdots \to v_n$ with $v_1 = v_n$.*

P : We now introduce a bunch of "familiar" terms.

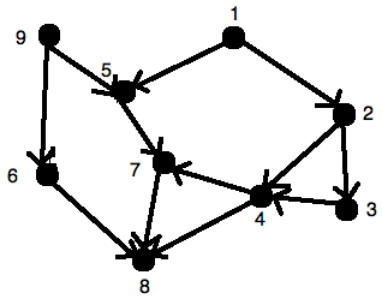


FIGURE 1. A directed acyclic graph

Definition 4.16. *Let $G = (V, E)$ be a directed acyclic graph. For a given $v \in V$, we define the set of 'parents' of v (and denote it by $pa(v)$) to be the set of all vertices w with $w \to v$. Likewise, we denote by $ch(v)$ the set of 'children' of v. This set consists of all the vertices w with $v \to w$. Finally, we say that w is an 'ancestor' of v if there exists a path $w \to v_1 \to v_2 \to \cdots \to v_n \to v$.*

S: We need the graph to be acyclic so that a vertex can not be a parent or grandparent or great-grandparent of itself?

P: Now that you mention it, that makes sense.

Definition 4.17. *Let $G = (V, E)$ be a directed acyclic graph. We denote by $\mathcal{R}$ the set of 'roots' of the graph G, that is, the set of vertices $v \in V$ for which $pa(v) = \emptyset$. Likewise, we denote by $\mathcal{L}$, the set of 'leaves' of the graph G, that is, the set of vertices v for which $ch(v) = \emptyset$.*

Remark: Note that if V is finite (as we have always assumed), then the graph has at least one root and at least one leave. We are certainly using the fact that G is acyclic to deduce this.

P: We are ready to define the so called 'something' nets:

Definition 4.18. *Let $\{X_v\}_{v \in V}$ be a set of random variables and denote by p_V their joint distribution. We say that the variables are a Bayes net with respect to a directed acyclic graph $G = (V, E)$ if*

$$p_V(x_V) = \prod_{v \in V} p_{v|pa(v)}(x_v | x_{pa(v)}).$$

In the above formula, $p_{v|pa(v)}$ stands for the conditional distribution of X_v given $X_{pa(v)}$. In other words, the conditional distribution of X_v given its parents. Finally, if $v \in \mathcal{R}$, $p_{v|pa(v)}$ must be interpreted as p_v (the marginal distribution of X_v).

P : GRF is to undirected graphs as BN is to directed graphs.

S: But what is their difference, which one is better?

P: What does 'better' mean? Let us first try to understand Bayes nets a bit *better*. In fact, let us start by stating the analogues of the independence, conditioning and marginalizing rules that we explored in the GRF setting.

Proposition 4.19. *(Independence rule) Suppose that $\{X_v\}_{v\in V}$ is a BN with respect to $G = (V, E)$. If v, w have no common ancestors, then X_v and X_w are independent. In particular, the roots of G are all independent.*

Proof. Let us show that the roots of G are independent. Observe that

$$p_V(x_V) = \prod_{v\in\mathcal{R}} p_v(x_v) \cdot \prod_{v\in V\setminus\mathcal{R}} p_{v|pa(v)}(x_v|x_{pa(v)}).$$

To compute the marginal $p_{\mathcal{R}}$, we add the above expression over all $x_{V\setminus\mathcal{R}}$. On the left hand side, we obtain precisely $p_{\mathcal{R}}$. On the right hand side, we obtain the product $\prod_{v\in\mathcal{R}} p_v(x_v)$. This shows the independence of the variables. □

Example: Suppose that the variables $\{X_v\}_{v\in V}$ are a Bayes net with respect to the graph in Figure 1. We can conclude that the variables X_1 and X_9 are independent. We can also concluded that the vector (X_6, X_9) is independent of (X_1, X_2, X_3) or that X_6 is independent of X_4. However, we can **not** conclude that X_5 and X_3 are independent (nor dependent). Indeed, in general, having a common ancestor does not imply dependence.

When we combine with the next rule, we can infer conditional independence.

Proposition 4.20. *(Conditioning rule). Suppose that $\{X_v\}_{v\in V}$ is a BN with respect to $G = (V, E)$. Let A be a subset of $\mathcal{R}$, that is, A is a collection of roots of G. Then, the conditional distribution of $X_{V\setminus A}$ given X_A, is a Bayes net with respect to $G' = (V \setminus A, E')$, where E' is the set of edges $u \to v$ in E, with $u \in V \setminus A$.*

In simple words, the resulting graph is obtained by erasing the roots in A and the arrows coming out from them.

Proof. Observe that

$$\begin{aligned} p_{V\setminus A|A}(x_{V\setminus A}|x_A) &= \frac{\prod_{v\in A} p_v(x_v) \cdot \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)})}{p_A(x_A)} \\ &= \frac{p_A(x_A) \cdot \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)})}{p_A(x_A)} \\ &= \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}). \end{aligned}$$

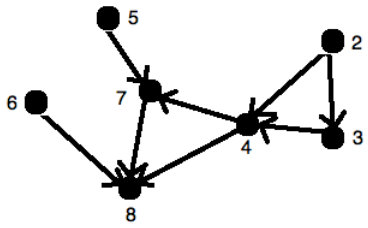
□

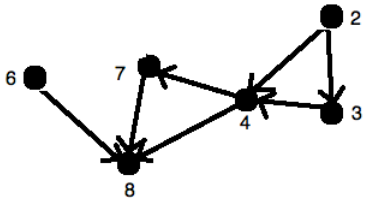
Example: Suppose that the variables $\{X_v\}_{v\in V}$ are a BN with respect to the graph in Figure 1. Conditioned on X_1 and X_9, the variables $X_{2:8}$ form a BN with respect to the graph:

In particular, using now the independence rule, we can conclude for example that the variables X_5, X_6, X_2 are conditionally independent given X_1 and X_9. Let us further condition on X_5. Then, conditioned on X_1, X_9, X_5, the variables $(X_{2:4}, X_{6:8})$ are a BN with respect to:

In particular, we can conclude that the vector (X_2, X_3) is conditionally independent of X_6 given X_1, X_5, X_9.

Remark: Note that the conditioning rule only applies when we condition on roots!





Proposition 4.21. *(Marginalizing rule). Suppose that $\{X_v\}_{v\in V}$ is a BN with respect to $G = (V, E)$. Let A be a subset of $\mathcal{L}$, that is, A is a collection of leaves of G. Then, the marginal of $X_{V\setminus A}$, is a Bayes net with respect to $G' = (V \setminus A, E')$, where E' is the set of edges $u \to v$ in E, with $v \in V \setminus A$.*

In simple words, the resulting graph is obtained by erasing the leaves in A and the arrows entering them.

Proof. Observe that

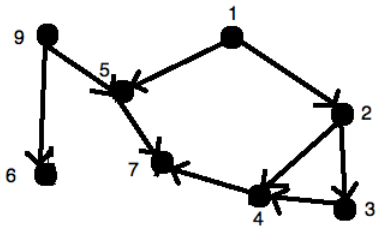
$$\begin{aligned} p_{V\setminus A}(x_{V\setminus A}) &= \sum_{x_A} \left(\prod_{v\in A} p_{v|pa(v)}(x_v|x_{pa(v)}) \cdot \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}) \right) \\ &= \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}) \cdot \sum_{x_A} \prod_{v\in A} p_{v|pa(v)}(x_v|x_{pa(v)}) \\ &= \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}) \end{aligned}$$

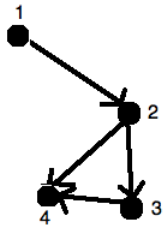
We can extract the product from the sum because it does not depend on x_A: we use the fact that $A \subseteq \mathcal{L}$. □

Example: Suppose that the variables $\{X_v\}_{v\in V}$ are a Bayes net with respect to the graph in Figure 1. The variables $X_{V\setminus\{8\}}$ form a BN with respect to the graph:

We can continue summing over leaves and for example deduce that the variables X_1, X_2, X_3, X_4 form a BN with respect to the graph:

Remark: Note that the marginalizing rule only applies when we sum over leaves!





The previous rules allow us to give the best possible interpretation of what a BN is about: a directed graph tells us a story of how to sample from some joint distribution p_V. For example, suppose that the variables $\{X_v\}_{v\in V}$ form a BN with respect to the graph from figure 1. In order to sample from p_V we first obtain a sample $\tilde{x}_1$ of X_1, then, independently, we obtain a sample $\tilde{x}_9$ from X_9. With $\tilde{x}_1, \tilde{x}_9$ we can sample $\tilde{x}_5$ from X_5 (we only need the parents of 5). Using $\tilde{x}_9$, we can also sample $\tilde{x}_6$ from X_6 (we only need the parents of 6). And so on. The graph tells us what to sample next and what needs to be sampled before sampling from a given variable. In this way we obtain one sample from X_V, namely $\tilde{x}_V$.

P So, if you know that some variables are a BN with respect to an undirected acyclic graph G, and moreover, you know the relevant conditional distributions (the conditional distributions $p_{v|pa(v)}$), sampling becomes very "easy". You just follow the arrows!

S "Easy"?

P Well, as always, you have to be able to sample efficiently from a given one dimensional distribution. This was what we discussed in Section 1. Now, notice that we have been considering discrete random variables. Do you remember how to sample a discrete random variable?

S...

P What's that face?

S Err...I was just thinking about GRFs and BNs. I think that the concept of BN is easier to understand than that of GRF. The independence and conditioning rules for BN are very simple and the marginalizing rule for BN is much easier to get than the one for the GRF... Why using GRF instead of BN? Is it fair to say that BNs are better than GRFs?

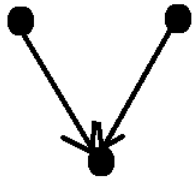
S Not really, it depends. Consider for example $p(x_1, x_2, x_3, x_4)$ given by:

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} x_1^{x_2} x_2^{x_3} x_3^{x_4} x_4^{x_1}.$$

In this case you can associate an undirected graph G in a straightforward way. What about a directed graph ?

P: I guess I would have to compute something... The BN structure is not so explicit in this case.

S: Yes, exactly. Now suppose I tell you a story of how to carry out a stochastic experiment. For example, I tell you to flip two unbiased coins. If the outcome of both coins is H then, flip an unbiased coin again. Otherwise, flip a biased coin. The joint distribution of the three coin flips has a very explicit BN structure!



S: Right, so suppose that we know that some variables form a BN with respect to some directed graph, and suppose that we know the relevant conditional distributions; as in the story you just told me. In that case, having a directed graph is better than having an undirected graph associated to the variables?

P: Let me explain with the following example. Suppose that we have a set of variables that form a BN with respect to the following graph:

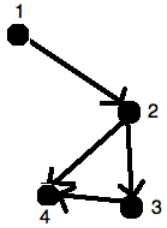


FIGURE 2. A directed graph.

From what we have discussed up to this point, can you conclude that the variable X_1 is independent of X_3 given X_2? Remember that the conditioning rule for BN can only be used when conditioning on roots.

S: I don't know.

P : What if I told you that the variables are a Gibbs random field with respect to the undirected graph:

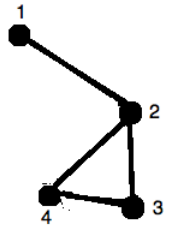


FIGURE 3. Moral graph of the graph in Figure 2.

S: Yes of course, we just use the independence and conditioning rules for GRFs.

P: Correct. In this case, the GRF structure allows us to conclude the conditional independence. Now actually, we will see that, if the variables are a BN with respect to the graph in Figure (2), then they are also a GRF with respect to the graph in Figure (3). In fact, we will see that if a set of variables is a BN with respect to some graph G, then there is an undirected graph G' (we will call it moral graph) that is respected by the variables (i.e. the variables are a GRF with respect to G'). Perhaps, in that sense, BNs are better than GRF: from a directed graph you can obtain an associated undirected graph without computing, but not the other way around (unless you compute conditional distributions). Remember that we want to compute as little as possible. The problem is that sometimes moral graphs don't help that much.

Definition 4.22. *Let $G = (V, E)$ be a directed acyclic graph. The moral graph associated to G, is the undirected graph $G' = (V, E')$ with set of vertices V and with edges E' that are determined as follows:*

- $\{u, v\} \in E'$, *if* $u \to v$.
- $\{u, v\} \in E'$, *if there exists* $w \in V$ *with* $v \to w$ *and* $u \to w$. *In simple words, we marry the parents.*

S : Is there an easy way to remember what a moral graph is?

P: "Parents should stay together and be with their children" or if you prefer names of songs : 'Stay together for the kids'.

S : Or the 'kids aren't alright'?

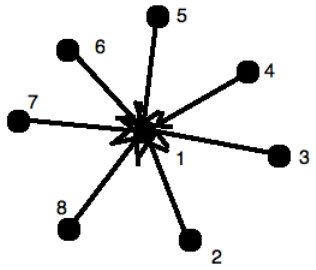
P: Actually, that one does not work.

Proposition 4.23. *If $\{X_v\}_{v \in V}$ is a Bayes net with respect to a directed acyclic graph G, then the variables are a Gibbs random field with respect to G's moral graph.*

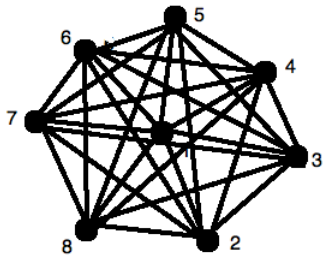
Example: The graph G from before has G' as moral graph. It is clear now that the variables X_1 and X_3 are independent given X_2.

In some cases, moral graphs are quite useful, as we can infer (conditional) independence of some of the variables from it; this is specially true when the moral graph does not have too many edges. In other cases, moral graphs are very uninformative.

Example: Suppose that a set of variables is a BN with respect to the graph:



From the independence rule for BN, we can say that the variables $X_{2:8}$ are independent: quite informative. Now let us consider the associated moral graph.



The above is the complete graph on eight vertices. Any collection of eight random variables would respect such graph. The moral graph is completely uninformative.

Pill against loss of memory: Remember, the ultimate goal of introducing graphical models was to help us compute quantities of interest associated to some joint distribution. In the GRF model, we used DP and exploited the factorization into clique potentials to compute efficiently. We will see that in the BN setting, we can obtain certain marginals and conditional distributions with no computations or with few computations. The bottom line is that graphical models have to be combined to take the best of all worlds.

4.6. **Computing using Bayes Nets.** Suppose that the variables $\{X_v\}_{v\in V}$ are a Bayes net with respect to the directed acyclic graph $G = (V, E)$. That is, we know that the joint distribution p_V can be written as:

$$p_V(x_V) = \prod_{v\in V} p_{v|pa(v)}(x_v|x_{pa(v)}).$$

We will assume that we know the conditional distributions $p_{v|pa(v)}$. In particular, for any given x_v and $x_{pa(v)}$, we know what $p_{v|pa(v)}(x_v|x_{pa(v)})$ is equal to. For simplicity, we may assume that there is no computation cost associated to obtaining $p_{v|pa(v)}(x_v|x_{pa(v)})$

(think of these functions as compressed in look up tables). Any other distribution (not one of the $p_{v|pa(v)}$) would have to be computed.

Computing marginals.

- P: 'Computing marginals using Bayes nets' should actually be : 'Use the marginalizing rule on leaves as much as you can, then start computing'.

- S: Very descriptive.

Suppose that A is a set of leaves of G. It follows from the proof of the conditioning rule, that the marginal of $p_{V\setminus A}$ is:

$$p_{V\setminus A}(x_{V\setminus A}) = \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}).$$

That is, the marginal of $p_{V\setminus A}$ (for A a set of leaves!) is obtained without computing anything.

Example: Suppose that the variables $\{X_v\}_{v\in V}$ are a BN with respect to the graph in Figure 1. We can compute (at zero cost) the marginal of the variables $\{X_1, X_2, X_3, X_4\}$ by pruning leaves (using the marginalizing rule repeatedly). Namely,

$$p_{1234}(x_{1:4}) = p_1(x_1)\cdot p_{2|1}(x_2|x_1)\cdot p_{3|2}(x_3|x_2)\cdot p_{4|2,3}(x_4|x_2,x_3).$$

On the other hand, notice that we can not obtain the marginal of $\{X_1, X_2, X_4\}$ by pruning any other leaves. At this point there is no other option but to compute.

$$\begin{aligned} p_{124}(x_{1:4}) &= \sum_{x_3}\big(p_1(x_1)\cdot p_{2|1}(x_2|x_1)\cdot p_{3|2}(x_3|x_2)\cdot p_{4|2,3}(x_4|x_2,x_3)\big) \\ &= p_1(x_1)\cdot p_{2|1}(x_2|x_1)\sum_{x_3}\big(p_{3|2}(x_3|x_2)\cdot p_{4|2,3}(x_4|x_2,x_3)\big) \end{aligned} \tag{17}$$

How many computations do we need to obtain p_{124}? Suppose that the range for all the variables is $\{1,\dots,10\}$ and suppose that we want to store the function p_{124} in a look up table. The memory cost of doing this is 10^3. To compute p at a fixed configuration x_{124}, we would need to compute 10 sums. To be able to store p as a look up table we would need to compute something like 10^4 sums.

Now , notice that we can not exploit the BN structure to compute all possible marginals. Take for example the variables X_1, X_8. The BN structure is not helpful in this case because X_8 is itself a leave (we can not add over x_8 because we need the marginal of X_8!). What to do in this case? Perhaps the best idea is to work with the moral graph and use DP to compute the desired marginal. Remember, in general, graphical models have to be combined to take the best of all worlds.

Computing conditional distributions.

- P: 'Computing conditional distributions using Bayes nets' should actually be : 'Use the conditioning rule on roots as much as you can, then start computing'.

Suppose that A is a set of roots of G. It follows from the proof of the conditioning rule, that the marginal of $p_{V\setminus A}$ is:

$$p_{V\setminus A|A}(x_{V\setminus A}|x_A) = \prod_{v\in V\setminus A} p_{v|pa(v)}(x_v|x_{pa(v)}).$$

That is, the conditional distribution of $p_{V\setminus A|A}$ (for A a set of leaves!) is obtained without computing anything.

Example: Suppose that the variables $\{X_v\}_{v\in V}$ are a BN with respect to the graph in Figure 1. We can compute (at zero cost) the conditional distribution of the variables $\{X_3, X_4, X_7, X_8\}$ given the rest of the variables, by removing roots (using the conditioning rule repeatedly). Namely,

$$p_{3478|12569}(x_{3478}|x_{12569}) = p_{3|2}(x_3|x_2)\cdot p_{4|23}(x_4|x_2,x_3)\cdot p_{7|45}(x_7|x_4,x_5)\cdot p_{8|467}(x_8|x_4,x_6,x_7).$$

We can also obtain the conditional distribution of X_2, X_3, X_4 given X_1 (at zero cost). First we use the marginalizing rule to obtain the distribution p_{1234}. Then we use the conditioning rule. We deduce that

$$p_{234|1}(x_{2,3,4}|x_1) = p_{2|1}(x_2|x_1)\cdot p_{3|2}(x_3|x_2)\cdot p_{4|23}(x_4|x_2,x_3).$$

What about something like the conditional distribution of $p_{4|2}$? We can not extract any other roots and so we need to compute. Let us simplify things first. Remember the conditional Bayes rule:

$$\mathbb{P}(A\cap B|C) = \mathbb{P}(A|B,C)\cdot\mathbb{P}(B|C)$$

Using Bayes rule, we deduce that:

$$p_{34|2}(x_3,x_4|x_2) = p_{4|23}(x_4|x_2,x_3)\cdot p_{3|2}(x_3|x_2)$$

Note that in the above expression, both terms on the right hand side are known. Adding over x_3, we obtain the desired conditional distribution:

$$p_{4|2}(x_4|x_2) = \sum_{x_3} p_{4|23}(x_4|x_2,x_3)\cdot p_{3|2}(x_3|x_2).$$