Introduction to Stochastic Processes

Dr. Naomi Feldheim, `naomi.feldheim@biu.ac.il`
Summary by Ari Feiglin

Contents

1 Introduction

This course will focus on tools which can be used to study random processes. A random process is a sequence of random variables which represent measurements of the process. Examples of random processes are random walks (these are commonly described as the path a drunk man would take while trying to get home), card shuffles (which can be viewed as choosing a card and placing it randomly in the deck), and branching (for example the population of bunnies in a specific area: the random variable being the number of bunnies in each generation).

2 Markov Chains

2.0.1 Definition

A **discrete-time Markov process** is a sequence of random variables $\{X_n\}_{n\geq 0}$. This sequence is called a **Markov chain** on a set of states S if:

(1) For every n, $X_n \in S$ almost surely (meaning $\mathbb{P}(X_n \in S) = 1$),

(2) For every $n \geq 0$ and for every $s_0, \ldots, s_{n+1} \in S$,

$$\mathbb{P}(X_{n+1} = s_{n+1} \mid X_0 = s_0, \ldots, X_n = s_n) = \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$$

ie. the probability of the next measurement being some arbitrary value is dependent only on the previous measurement. This is only necessary if $\mathbb{P}(X_0 = s_0, \ldots, X_n = s_n) > 0$.

In this course S will always be countable. We can also write the second condition using distributive equivalence:

$$X_{n+1}|X_0, \ldots, X_n \overset{d}{=} X_{n+1}|X_n$$

Notice how the Markov property can be strengthened in various ways, for example if $n > m$ then

$$\begin{aligned}
&\mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}, \ldots, X_m = s_m)\\
&\quad= \sum_{s_m,\ldots,s_0} \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) \cdot \mathbb{P}(X_{m-1} = s_{m-1}, \ldots, X_0 = s_0 \mid X_{n-1} = s_{n-1}, \ldots, X_m = s_m)\\
&\quad= \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}) \cdot \sum \mathbb{P}(X_{m-1} = s_{m-1}, \ldots, X_0 = s_0 \mid X_{n-1} = s_{n-1}, \ldots, X_m = s_m)\\
&\quad= \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1})
\end{aligned}$$

This can be viewed as the base case for

$$\mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \ldots, X_m = s_m) = \mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \ldots, X_{m'} = s_{m'})$$

where $m' < m$. This is since for $k = 1$, both of these are equal to $\mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$. The induction step follows by

$$\begin{aligned}
&\mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_n = s_n, \ldots, X_m = s_m)\\
&\quad= \sum_{s_{n+1}} \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_{n+1} = s_{n+1}, \ldots, X_m = s_m) \cdot \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, \ldots, X_m = s_m)\\
&\quad= \sum_{s_{n+1}} \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_{n+1} = s_{n+1}, \ldots, X_{m'} = s_{m'}) \cdot \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, \ldots, X_{m'} = s_{m'})\\
&\quad= \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_n = s_n, \ldots, X_{m'} = s_{m'})
\end{aligned}$$

By taking $m' = 0$ and $m = n$ we get $\mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n) = \mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \ldots, X_0 = s_0)$, or in other words for all $m < n$,

$$\mathbb{P}(X_n = s_n \mid X_m = s_m, \ldots, X_0 = s_0) = \mathbb{P}(X_n = s_n \mid X_m = s_m)$$

This can be even further strengthened: let $\varnothing \neq B \subseteq \{0, \ldots, n-1\}$ and $m = \max B$ then

$$\mathbb{P}(X_n = s_n \mid \forall i \in B\colon X_i = s_i) = \mathbb{P}(X_n = s_n \mid X_m = s_m)$$

To prove this let $C = \{0, \ldots, m\} \setminus B$ then

$$\begin{aligned}
\mathbb{P}(X_n = s_n \mid \forall i \in B\colon X_i = s_i) &= \sum_{(s_i)_{i\in C} \in S^C} \mathbb{P}(X_n = s_n \mid X_m = s_m, \ldots, X_0 = s_0) \cdot \mathbb{P}(\forall i \in C\colon X_i = s_i \mid \forall i \in B\colon X_i = s_i)\\
&= \mathbb{P}(X_n = s_n \mid X_m = s_m) \cdot \sum \mathbb{P}(\forall i \in C\colon X_i = s_i \mid \forall i \in B\colon X_i = s_i)\\
&= \mathbb{P}(X_n = s_n \mid X_m = s_m)
\end{aligned}$$

A consequence of this is that if $\{X_n\}_{n\geq 0}$ is a Markov chain and $\{a_n\}_{n\geq 0}$ is strictly monotonic then $Y_n = X_{a_n}$ is also a Markov chain. After all if we let $B = \{a_{n-1}, \ldots, a_0\}$ then $\max B = a_{n-1}$ and so

$$\begin{aligned}\mathbb{P}(Y_n = s_{a_n} \mid Y_{n-1} = s_{a_{n-1}}, \ldots, Y_0 = s_{a_0}) = \mathbb{P}(X_{a_n} = s_{a_n} \mid \forall i \in B \colon X_i = s_i) &= \mathbb{P}(X_{a_n} = s_{a_n} \mid X_{a_{n-1}} = s_{a_{n-1}}) \\ &= \mathbb{P}(Y_n = s_{a_n} \mid Y_{n-1} = s_{a_{n-1}})\end{aligned}$$

as required.

2.0.2 Definition

For a Markov chain $\{X_n\}_{n\geq 0}$ on a finite set of states S, we define the **adjacency matrix** at the nth measurement by

$$P_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_{n-1} = i)$$

for $i, j \in S$. This is also sometimes written as $P_n(i \to j)$ (the probability measuring i on the $n-1$th measurement gives j on the next). If $P^{(n)}$ is the same for all n, then we say that the chain is **homogeneus in time**, and we generally write P in place of $P^{(n)}$.

For example, suppose a frog is hopping between N leaves. The frog can hopping from every leaf to every other leaf, and it always chooses a leaf in an independent and uniform manner. This defines a Markov chain where the states are the leaves, and X_n is the leaf the frog is on after n hops. This Markov chain is even homogeneus since the frog makes its choices in a manner which does not take the current number of hops into account. The adjacency matrix is defined by

$$P_{ij} = \begin{cases} \frac{1}{N-1} & i \neq j \\ 0 & i = j \end{cases}$$

This is the simple random process on the complete graph of N vertices, K_N.

Suppose $N = 4$, and supppose that at the beginning the frog is on either the first or second leaf with equal probability. What is the probability that after one hop the frog is on the fourth leaf? The following notation will be used: $X \sim (a_0, \ldots, a_n)$ will be used to mean $\mathbb{P}(X = s_i) = a_i$, where s_i is some understood ordering of the set of states S. Then

$$\mathbb{P}\left(X_1 = j \;\middle|\; X_0 \sim \left(\frac{1}{2}, \frac{1}{2}, 0, 0\right)\right) = \mathbb{P}(X_1 = j \mid X_0 = 1) \cdot \frac{1}{2} + \mathbb{P}(X_1 = j \mid X_0 = 2) \cdot \frac{1}{2}$$

as the rest of the terms are zero. For $j = 4$ we get that this is equal to $\frac{1}{3}$. Notice that we can generalize this and get

$$\mathbb{P}(X_{n+1} = j \mid X_n \sim \vec{v}) = \sum_{i\in S} \mathbb{P}(X_{n+1} = j \mid X_n = i) \cdot \mathbb{P}(X_n = i) = \sum_{i\in S} P_{ij}^{(n+1)} \vec{v}_i = (\vec{v} \cdot P^{(n+1)})_j$$

So we have proven the following:

2.0.3 Proposition

If $X_n \sim \vec{v}$ then $X_{n+1}|X_n \sim \vec{v} \cdot P^{(n+1)}$, and so $X_n|X_0 \sim \vec{v} \cdot P^{(n)} \cdots P^{(1)}$. In particular if the Markov chain is homogeneus, $X_n|X_0 \sim \vec{v} \cdot P^n$.

This simplifies dealing with Markov chains, especially homogeneus ones.

2.0.4 Example

Suppose $\{Y_n\}_{n=1}^{\infty}$ is a sequence of random variables which have the distribution $Y_n \sim \mathrm{Ber}\left(\frac{1}{n}\right)$ (recall that $X \sim \mathrm{Ber}(p)$ means that X is 1 with probability p and zero otherwise). And we define $X_n = \chi\{(\exists m \leq n)\, Y_m = 1\}$, the indicator of the set of all values such that there is an index before n where $Y_m = 1$ (χ_S is the *indicator function* of the set S, defined by $\chi_S(x) = 1$ for $x \in S$ and zero otherwise). We will prove X_n is a Markov chain. Notice that

$$X_n = \chi\{(\exists m \leq n)\, Y_m = 1\} = \chi\{(\exists m \leq n-1)\, Y_m = 1\} \vee \chi\{Y_n = 1\} = X_{n-1} \vee \chi\{Y_n = 1\}$$

$\vee$ is bitwise or, or equivalently the maximum. And therefore we get that $X_n = \bigvee_{i=1}^{n} \chi\{Y_i = 1\}$. This means that if $X_{n-1} = 1$ then $X_n = 1$, and if $X_{n-1} = 0$ then $X_n = 1$ if and only if $Y_n = 1$. And so X_n's value depends only on X_{n-1}'s and not any previous X_i. So $\{X_n\}_{n=1}^{\infty}$ is indeed a Markov chain.

Notice that

$$\mathbb{P}(X_n = 0 \mid X_{n-1} = 0) = \mathbb{P}(Y_n = 0) = \frac{n-1}{n}, \quad \mathbb{P}(X_n = 1 \mid X_{n-1} = 0) = \mathbb{P}(Y_n = 1) = \frac{1}{n},$$

$$\mathbb{P}(X_n = 0 \mid X_{n-1} = 1) = 0, \quad \mathbb{P}(X_n = 1 \mid X_{n-1} = 1) = 1$$

And so we get that

$$P^{(n)} = \begin{pmatrix} \frac{n-1}{n} & \frac{1}{n} \\ 0 & 1 \end{pmatrix}$$

2.0.5 Definition

A real $n \times n$ matrix P such that $P_{ij} \geq 0$ for every i, j, and for every row i we have $\sum_{j=1}^{n} P_{ij} = 1$ then P is called an **stochastic matrix**.

Notice that we can draw a diagram for every stochastic matrix and it will be the transition matrix of a Markov chain. Meaning every stochastic matrix is the transition matrix of some Markov chain, and every transition matrix is stochastic. Notice that the second condition for a matrix to be stochastic can be written as $P\mathbf{1} = \mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)^\top$.

2.1 Hitting Times and Classifying States

2.1.1 Definition

Let $\{X_n\}_{n \geq 0}$ be a Markov chain over a state space S, and let $A \subseteq S$. Then we define the **hitting time** to A to be the random variable

$$T_A = \min\{t \geq 1 \mid X_t \in A\}$$

Note that if X_t is never in A then T_A can be ∞, and so T_A is a function from the probability space to the extended reals: $\Omega \longrightarrow \mathbb{R} \cup \{\infty\}$. This means that $T_A^{-1}\{\infty\}$ must also be measurable (an event).

In the case that A is a singleton $A = \{a\}$ then we write T_a in place of T_A. Notice that T_A measures starting from $t = 1$, while it is possible that the initial condition is in A, ie. $X_0 \in A$. So in the case that $X_0 \in A$, T_A measures the *return time* to A, in particular if $X_0 \sim \delta_a$ where $\delta_a = (0, \ldots, 1, \ldots, 0)$ (1 is at the index corresponding to the state a). We also use the following notation

$$\mathbb{P}_V(E) = \mathbb{P}(E \mid X_0 \sim V), \qquad \mathbb{P}_{\delta_a}(E) = \mathbb{P}_a(E) = \mathbb{P}(E \mid X_0 = a)$$

If P is the transition matrix of a homogeneus Markov chain, then $P^n(a \to b)$ means $P^n_{ba} = \mathbb{P}(X_n = b \mid X_0 = a)$.

2.1.2 Lemma

If $\{X_n\}$ is a homogeneus Markov chain, then

$$P^n(a \to b) = \sum_{m=1}^{n} \mathbb{P}_a(T_b = m) P^{n-m}(b \to b)$$

$$\begin{aligned} P^n(a \to b) = \mathbb{P}_a(X_n = b) = \mathbb{P}\left(\bigcup_{m=1}^{n} \{T_b = m\}, X_n = b \,\middle|\, X_0 = b \right) &= \sum_{m=1} \mathbb{P}(T_b = m, X_n = b \mid X_0 = b) \\ &= \sum_{m=1}^{n} \mathbb{P}(X_n = b \mid T_b = m, X_0 = a) \cdot \mathbb{P}(T_b = m \mid X_0 = a) \end{aligned}$$

Now, $\mathbb{P}(X_n = b \mid T_b = m, X_0 = a) = \mathbb{P}(X_n = b \mid X_m = b, X_{m-1} \neq b, \ldots, X_1 \neq b, X_0 = a) = \mathbb{P}(X_n = b \mid X_m = b)$ by the Markov property. Since $\{X_n\}$ is homogeneus this is just equal to $P^{n-m}(b \to b)$. Thus this formula is equal to

$$\sum_{m=1}^{b} \mathbb{P}(X_n = b \mid X_m = b) \cdot \mathbb{P}_a(T_b = m) = \sum_{m=1}^{b} P^{n-m}(b \to b) \cdot \mathbb{P}_a(T_b = m) \qquad \blacksquare$$

Let us introduce some more notation:

$$f_{a \to b} = \mathbb{P}(T_b < \infty \mid X_0 = a), \qquad f_{a \to a} = f_a = \mathbb{P}(T_a < \infty \mid X_0 = a)$$

thus $f_{a \to b}$ is the probability that if we start at a, we eventually reach b.

2.1.3 Lemma

$f_{a\to c} \geq f_{a\to b}\cdot f_{b\to c}$

Notice that $\{T_c<\infty\} = \{(\exists t>0)X_t = c\} \supseteq \dot\bigcup_{k>0}\{T_b = k, (\exists t>k)X_t = c\}$. Thus we get

$$\begin{aligned}
f_{a\to c} = \mathbb{P}(T_c<\infty \mid X_0 = a) &\geq \sum_{k=1}^{\infty} \mathbb{P}(T_b = k, (\exists t>k)X_t = c \mid X_0 = a)\\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a)\cdot\mathbb{P}((\exists t>k)X_t = c \mid T_b = k, X_0 = a)\\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a)\cdot\mathbb{P}((\exists t>k)X_t = c \mid X_k = b, X_{k-1}\neq b,\dots,X_1\neq b, X_0 = a)\\
\text{(Markov property)} &= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a)\cdot\mathbb{P}((\exists t>k)X_t = c \mid X_k = b)\\
\text{(homogeneity)} &= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a)\cdot\mathbb{P}((\exists t>0)X_t = c \mid X_0 = b)\\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a)\cdot f_{b\to c} = f_{a\to b}\cdot f_{b\to c} \qquad \blacksquare
\end{aligned}$$

In particular this means

$$f_a \geq f_{a\to b}\cdot f_{b\to a}$$

For every $a\in S$ we define the random variable $N(a) = \sum_{n=1}^{\infty}\chi\{X_n = a\}$, which is the number of times the state a is visited from time 1 and onward. When $X_0\sim V$ we write $N_V(a)$. Notice then that $f_{a\to b} = \mathbb{P}(N(b)\geq 1 \mid X_0 = a)$ and so $f_a = \mathbb{P}(N(a)\geq 1\mid X_0 = a)$.

2.1.4 Proposition

$\mathbb{P}(N(a)\geq k \mid X_0 = a) = f_a^k$

We prove this by induction, for $k=1$ this is simply what we just said. Now

$$\begin{aligned}
\mathbb{P}(N(a)\geq k+1 \mid X_0 = a) &= \sum_{m=1}^{\infty}\mathbb{P}(T_a = m, |\{j>m \mid X_j = a\}|\geq k \mid X_0 = a)\\
\text{(Markov property)} &= \sum_{m=1}^{\infty}\mathbb{P}(T_a = m \mid X_0 = a)\cdot\mathbb{P}(|\{j>m\mid X_j = a\}|\geq k \mid X_m = a)\\
\text{(homogeneity)} &= \sum_{m=1}^{\infty}\mathbb{P}(T_a = m \mid X_0 = a)\cdot\mathbb{P}_a(N(a)\geq k)\\
\text{(induction)} &= f_a^k\sum_{m=1}^{\infty}\mathbb{P}(T_a = m\mid X_0 = a) = f_a^{k+1} \qquad \blacksquare
\end{aligned}$$

Notice then that

$$\mathbb{P}(N(a) = k \mid X_0 = a) = \mathbb{P}_a(N(a)\geq k) - \mathbb{P}_a(N(a)\geq k+1) = f_a^k - f_a^{k+1} = f_a^k(1-f_a)$$

Thus $N_a(a)\sim \mathrm{Geo}(1-f_a)-1$ (the $+1$ is since $X\sim\mathrm{Geo}(p)$ means $\mathbb{P}(X=k) = p(1-p)^{k-1}$). Thus

$$\mathbb{E}[N_a(a)] = \frac{1}{1-f_a} - 1 = \frac{f_a}{1-f_a}$$

2.1.5 Definition

A state $b\in S$ is **recurrent** if $f_b = 1$, equivalently if $\mathbb{P}_b(T_b<\infty)$ (the probability of returning to b is 1). A non-recurrent state is called **transient**. b is **absorbing** if $P(b \to b) = 1$.

Notice that if b is recurrent then if $f_b = 1$, $N_b(b) \sim \text{Geo}(0) - 1$, meaning $\mathbb{P}_b(N(b) = \infty) = 1$. And if b is transient then $N_b(b)$ is a finite geometric variable and so $\mathbb{P}_b(N(b) < \infty) = 1$. And so

$$\begin{aligned} b \text{ is recurrent} &\iff \mathbb{P}(N(b) = \infty \mid X_0 = b) = 1, \\ b \text{ is transient} &\iff \mathbb{P}(N(b) < \infty \mid X_0 = b) = 1 \iff \mathbb{P}(N(b) < \infty \mid X_0 \sim v) = 1 \end{aligned}$$

2.1.6 Definition

Let $a, b \in S$ be states. Then b is **reachable** from a if $f_{a \to b} \neq 0$ or $a = b$, this is denoted $a \to b$. a and b are **connected** if both $a \to b$ and $b \to a$, this is denoted $a \leftrightarrow b$.

This means that $a \to b$ if and only if there exists some $n \geq 0$ such that $P^n(a \to b) > 0$. Furthermore, connectivity is an equivalence relation: it is obviously reflexive and symmetric and if $a \to b$ and $b \to c$, since $f_{a \to c} \geq f_{a \to b} \cdot f_{b \to c} > 0$, we get that reachability and therefore connectivity is transitive. Thus S can be partitioned into *connectivity classes*.

2.1.7 Lemma

If $a \to b$ and $a \neq b$ then $\mathbb{P}(T_b < T_a \mid X_0 = a) > 0$.

Since $a \to b$, there exists a sequence of states $a = s_0, \ldots, s_m = b$ such that $P_{s_i s_{i+1}} > 0$ for all i. We can assume that for every $i > 0$, $a \neq s_i$. So we have a sequence whose probability is positive and where the hitting time of b is before that of a, so the probability that $T_b < T_a$ must be positive. ■

2.1.8 Definition

$A \subseteq S$ is **closed** if for every $a \in A$ and every $b \notin A$, b is not reachable from a. A is also called **irreducible** if it is closed and connected.

2.1.9 Theorem

If a is recurrent and $a \to b$, then also $b \to a$ and b is recurrent.

We know

$$f_{a \to b} = \mathbb{P}_a(T_a > T_b) + \mathbb{P}_a(T_a < T_b) \cdot \mathbb{P}(T_b < \infty \mid T_a < T_b)$$

by the above lemma $p = \mathbb{P}_a(T_b < T_a) > 0$ and so by homogeneity

$$= p + (1 - p) \cdot \mathbb{P}(T_b < \infty \mid X_0 = a) = p + (1 - p) f_{a \to b}$$

Thus we get that $p \cdot f_{a \to b} = p$ and since $p \neq 0$, $f_{a \to b} = 1$. Now

$$f_{a \to b}(1 - f_{b \to a}) = \mathbb{P}(X_n \text{ hits } b \text{ and never returns to } a \mid X_0 = a) \leq \mathbb{P}_a(N(a) < \infty) = 0$$

Thus $f_{b \to a} = 1$. Now $f_b \geq f_{b \to a} \cdot f_{a \to b} = 1$ so b is also recurrent. ■

So if $a \leftrightarrow b$, then a is recurrent if and only if b is. If b is reachable from a but a is not reachable from b, then a is transient. And if a is recurrent and $a \to b$ then $\mathbb{P}_b(N(a) = \infty) = 1$.

2.1.10 Theorem

A finite closed set of states $A \subseteq S$ contains a recurrent state.

Suppose A has only transient states. This means that $\mathbb{P}_v(N(a) < \infty) = 1$ for every $a \in A$, and so we get that $\mathbb{P}_v((\forall a \in A) N(a) < \infty) = 1$ (as the intersection of a countable number of events with probability one). And this means $\mathbb{P}_v\left(\sum_{a \in A} N(a) < \infty\right) = 1$ since A is finite. But since A is closed, we can never leave A and so if v's support is in A then $\sum_{a \in A} N_v(a) = \infty$. ■

In particular, since S is closed, if S is finite it contains a recurrent state.

2.1.11 Theorem

If S is a finite state space, then it can be uniquely partitioned into

$$S = T \uplus C_1 \uplus \cdots \uplus C_k$$

where T is the set of all transient states, and C_i are all disjoint irreducible (closed and connected) sets.

So T is the set of all transient states, and for every recurrent state $a \in S \setminus T$ let $C_a = \{b \mid a \to b\}$. By a previous theorem, for every $b \in C_a$, $b \to a$ so and if $b \to b'$ then $a \to b'$ meaning $b' \in C_a$, so C_a is closed. And if $b, b' \in C_a$ then $a \to b$ and $a \to b' \implies b' \to a$ and therefore $b' \to b$, so C_a is connected and therefore irreducible. By taking representatives of each C_a, let $C_i = C_{a_i}$, we get the partition.

This partition is unique: since if $C_1 \uplus \cdots \uplus C_k = C_1' \uplus \cdots \uplus C_m'$ let $a \in C_1$ then $a \in C_i'$ for some i, without loss of generality assume $a \in C_1'$. Then for every $b \in C_1$, since C_1 is connected $a \to b$ and so $b \in C_1'$ since C_1' is closed, thus $C_1 = C_1'$. Continuing inductively we get $k = m$ and $C_i = C_i'$ as required. ∎

2.1.12 Example

Suppose Elise is in a room 0, and can either stay in the room with probability $1 - p_1 - p_2$, go to room 1 with probability p_1 or go to room 2 with probability p_2. If she goes to a new room, she stays there forever. Knowing that ends up in room 2, what is the expected amount of time she spends waiting in room 0?

So we want to find the expected value of $N_0(0)$ knowing that $T_2 < \infty$. So we will compute

$$\mathbb{P}(N_0(0) = k \mid T_2 < \infty) = \frac{\mathbb{P}(N_0(0) = k, T_2 < \infty)}{\mathbb{P}(T_2 < \infty)} = \frac{\mathbb{P}(X_1 = \cdots = X_k = 0, X_{k+1} = 2)}{\mathbb{P}(T_2 < \infty)}$$

Now, utilizing conditional probability and the Markov property (this is all done under the assumption $X_0 = 0$),

$$\mathbb{P}(X_1 = \cdots = X_k = 0, X_{k+1} = 2) = \mathbb{P}(X_{k+1} = 2 \mid X_k = 0) \cdot \mathbb{P}(X_k = 0 \mid X_{k-1} = 0) \cdots \mathbb{P}(X_1 = 0) = p_2 \cdot (1 - p_1 - p_2)^k$$

And $\mathbb{P}(T_2 < \infty) = \frac{p_2}{p_1 + p_2}$ since to get to room 2 we must visit room 0 an arbitrary number of times, and then go to room 2, so

$$\mathbb{P}(T_2 < \infty) = \sum_{n=0}^{\infty} p_2 \cdot (1 - p_1 - p_2)^n = \frac{p_2}{p_1 + p_2}$$

Thus

$$\mathbb{P}(N_0(0) = k \mid T_2 < \infty) = (p_1 + p_2) \cdot (1 - p_1 - p_2)^k$$

Which means that

$$(N_0(0) \mid T_2 < \infty) \sim \mathrm{Geo}(p_1 + p_2) - 1 \implies \mathbb{E}[N_0(0) \mid T_2 < \infty] = \frac{1 - p_1 - p_2}{p_1 + p_2}$$

Notice two things: firstly, by symmetry this means that $(N_0(0) \mid T_1 < \infty) \sim \mathrm{Geo}(p_1 + p_2) - 1$ which is the same distribution. And secondly, this is the same distribution as $N_0(0)$, so the expected time Elise waits at room 0 does not change if we know which room she ends up in.

2.1.13 Definition

Let $a \in S \setminus T$ be a recurrent state, then we define its **period** to be

$$d(a) = \gcd\{n \geq 1 \mid P^n(a \to a) > 0\}$$

An irreducible Markov chain is called **periodic** if every state is recurrent and has the same period greater than 1, which is the **period** of the Markov chain.

Notice that if $P(a \to a) > 0$ then $d(a) = 1$, and so a periodic chain can be made aperiodic by adding a self-edge whose probability is nonzero.

2.1.14 Proposition

If P is the transition matrix of some periodic chain with a period of d, then P^d is reducible.

2.1.15 Proposition

If the Markov chain is irreducible then every state has the same period.

Let P be the transition matrix of the chain. Since $x \leftrightarrow y$, there exist natural r, ℓ such that $P^r(x,y), P^\ell(y,x) > 0$. So let $m = r + \ell$ and so

$$P^m(x,x) \geq P^r(x,y) \cdot P^\ell(y,x) > 0, \qquad P^m(y,y) \geq P^\ell(y,x) \cdot P^r(x,y) > 0$$

So let $\tau(a) = \{n \geq 1 \mid P^n(a,a) > 0\}$, and by above we have shown that $m \in \tau(x) \cap \tau(y)$. Now for every $n \in \tau(x)$ we have that $P^{\ell+n+r}(y,y) \geq P^\ell(y,x)P^n(x,x)P^r(x,y) > 0$ and so $n + m \in \tau(y)$. Thus $m + \tau(x) \subseteq \tau(y)$. By definition we have $d(y) = \gcd(\tau(y))$ and since $m \in \tau(y)$ we have $d(y)|m$ and since $m + \tau(x) \subseteq \tau(y)$ we must have that $d(y)|\tau(x)$. Thus $d(y)|d(x)$, and since x, y are arbitrary we get $d(x)|d(y)$ and so $d(x) = d(y)$ as required. ∎

This means that every irreducible Markov chain has a period, and if the period is > 1, it is periodic. So in order for an irreducible Markov chain to be periodic, it is sufficient for there to exist a state a with $d(a) > 1$.

A common Markov chain is a random walk on $\mathbb{Z}$, where

$$P(i, i+1) = p, \quad P(i, i-1) = 1 - p, \quad P(i,j) = 0 \text{ for } j \notin \{i \pm 1\}$$

Another way of representing X_n is by $X_n = \sum_{k=1}^n B_k$ where $B_k = 1$ with probability p and $B_k = -1$ with probability $1 - p$. $\{B_k\}$ is independent. If $p = \frac{1}{2}$, the walk is called *fair*.

2.1.16 Theorem

If $p \neq \frac{1}{2}$, every state in $\mathbb{Z}$ is transient.

Since all the states are connected, it is sufficient to show that 0 is transient. So we set $X_0 = 0$ and notice that $\frac{B_k+1}{2} \sim \mathrm{Ber}(p)$ and thus $\frac{X_n+n}{2} \sim \mathrm{Bin}(n,p)$ thus

$$\mathbb{P}(X_{2n} = 0) = \mathbb{P}\left(\frac{X_{2n} + 2n}{2} = n\right) = \binom{2n}{n} p^n (1-p)^n$$

and $\mathbb{P}(X_{2n+1} = 0) = \mathbb{P}\left(\frac{X_{2n+1}+2n+1}{2} = n + \frac{1}{2}\right) = 0$ since binomial distributions take on only integer values. By Stirling's approximation: $k! \in \Theta(k^{k+1/2}e^{-k})$, we get that there exists some $c > 0$ such that

$$\mathbb{P}(X_{2n} = 0) = \frac{(2n)!}{n!n!} p^n (1-p)^n \leq cp^n(1-p)^n \frac{(2n)^{2n+1/2}e^{-2n}}{n^{2n+1}e^{-2n}} = cp^n(1-p)^n \frac{2^{2n+1/2}}{\sqrt{n}} = c' \frac{\big(4p(1-p)\big)^n}{\sqrt{n}}$$

This can be bound by a q^n where $q \in [0,1)$, since $4p(1-p) < 1$ for $p \neq \frac{1}{2}$. Thus we get that $\sum_{k=1}^\infty \mathbb{P}(X_k = 0 \mid X_0 = 0)$ and so by Borel-Cantelli we then get that $\mathbb{P}(X_k = 0 \text{ i.o.} \mid X_0 = 0) = 0$, meaning that the probability $X_k = 0$ an infinite number of times is zero. Thus $\mathbb{P}(N(0) = \infty \mid X_0 = 0) = 0$, and so this means 0 is transient as required. ∎

If we have a Markov chain, and $A \subseteq S$, we can ask questions about hitting times in A by removing all the states in A and adding a new state $\hat{A}$. This can only be done if for every $a, a' \in A$ and $b \notin A$, $P(a \to b) = P(a' \to b)$, and we define that the probability $P(\hat{A} \to b) = P(a \to b)$. And $P(b \to \hat{A}) = \sum_{a \in A} P(b \to a)$. In particular this can be done if A is closed.

2.1.17 Example

Suppose we have the following Markov chain:

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

where $p, q, r \geq 0$ and $p + q + r = 1$. If we know that $X_0 = 2$, what is the probability that the chain will be absorbed into 1 or 3?

Let us define

$$\ell_j = \mathbb{P}(T_1 < \infty \mid X_0 = j)$$

since 1 and 3 are absorbing states, $\ell_1 = 1$ and $\ell_3 = 0$. Now, we want to compute ℓ_2:

$$\ell_2 = \mathbb{P}(T_1 < \infty \mid X_0 = 2) = \sum_{j=1}^{3} \mathbb{P}(T_1 < \infty \mid X_1 = j, X_0 = 2) \cdot \mathbb{P}(X_1 = j \mid X_0 = 2)$$
$$= \sum_{j=1}^{3} \mathbb{P}(T_1 < \infty \mid X_1 = j) \cdot P_{2j}$$

where the last step is due to homogeneity. This is equal to $\sum_{j=1}^{3} \ell_j P_{2j} = \ell_1 p + \ell_2 q + \ell_3 r = p + \ell_2 r$. Thus we get that $\ell_2 = p + \ell_2 r$ and so $\ell_2 = \frac{p}{1-r}$. Thus the probability that starting from $X_0 = 2$ we are absorbed into 1 (meaning $T_1 = \infty$) is $1 - \ell_2 = \frac{q}{1-r}$. Since 2 is transient, we are either absorbed into 1 or 3, so the probability of being absorbed into 3 is $\frac{p}{1-r}$.

Let us now ask what the expected time until being absorbed is. By the law of total expectation: Now, $\mathbb{E}\left[T_{\{1,3\}} \mid X_1 = 2\right] = \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right] + 1$ since it takes one more step, and so

$$= \left(1 + \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right]\right) \cdot \mathbb{P}_2(X_1 = 2) + \mathbb{P}_2(X_1 = 1) + \mathbb{P}_2(X_1 = 3)$$

So let $x = \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right]$, we get

$$x = (1 + x)r + p + q = (1 + x)r + (1 - r) \implies x = \frac{1}{1 - r}$$

2.1.18 Example

Suppose we have the following Markov chain:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ a_1 & a_2 & 0 & a_4 & 0 & 0 \\ 0 & 0 & b_3 & 0 & b_5 & b_6 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

What is the probability of being absorbed into one of the absorbing states $(1, 2, 5, 6)$ if it starts on one of the non-absorbing states $(3, 4)$?

Let us define $\ell_{m,k} = \mathbb{P}_m(T_k < \infty)$. Now, let us notice that

$$\ell_{m,k} = \mathbb{P}(T_k < \infty \mid X_0 = m) = \sum_{j=1}^{6} \mathbb{P}(T_k < \infty \mid X_1 = j) \cdot \mathbb{P}(X_1 = j \mid X_0 = m) = \sum_{j=1}^{6} P_{mj} \ell_{jk}$$

So if we define $L_{ij} = \ell_{ij}$ then we get that $L = PL$ and we can solve for L.

What is the expected time until being absorbed? We can consolidate $A = \{1, 2, 5, 6\}$ to a state we will call 1, then the new transition matrix is

$$P' = \begin{pmatrix} 1 & 0 & 0 \\ a_1 + a_2 & 0 & a_4 \\ b_5 + b_6 & b_3 & 0 \end{pmatrix}$$

Now let us define $r_j = \mathbb{E}[T_1 \mid X_0 = j]$, then we get

$$r_j = \sum_{i=1}^{3} \mathbb{E}[T_1 \mid X_1 = i] P_{ji} = P_{j1} + \sum_{i=2}^{3} (r_i + 1) P_{ji} = P_{j1} + P_{j2} + P_{j3} + r_2 P_{j2} + r_3 P_{j3}$$

Which is a linear system of equations which can be solved.

2.2 Stationary Distributions and the Convergence of Markov Chains

2.2.1 Definition

Suppose $|S| = N$, then a **stationary distribution** of P is a row vector π which represents a distribution (meaning $\pi_i \geq 0$ and $\sum_{i=1}^N \pi_i = 1$) such that $\pi = \pi P$.

A stationary distribution is an eigenvector (or the transpose of one) of $P^\top$ whose eigenvalue is 1. If π is a stationary distribution, then $\pi P = \pi \implies \pi P^n = \pi$ for every $n \geq 0$. This means that if $X_0 \sim \pi$ then $X_n \sim \pi$ for every n (since $\mathbb{P}(X_n = k \mid X_0 \sim \pi) = (\pi P^n)_k = \pi_k$).

For example if $G = (V, E)$ is an undirected graph where $|V| = N$ and the transitions from each state are all uniform (meaning $\mathbb{P}(X_n = v \mid X_{n-1} = u) = \frac{1}{\deg(u)}$ if $v \leftrightarrow u$), then let

$$\tilde{\pi} = (\deg(v_1), \ldots, \deg(v_N))$$

Then (using the notation $\delta\varphi$ which is 1 if φ is true and 0 otherwise) we have that $P_{xy} = \frac{1}{\deg(x)}\delta(x \leftrightarrow y)$, so

$$(\tilde{\pi}P)_y = \sum_{x\in V} \tilde{\pi}_x P_{xy} = \sum_{x\in V} \deg(x)\frac{1}{\deg(x)}\delta(x \leftrightarrow y) = \sum_{x\in V} \delta(x \leftrightarrow y) = \deg(y) = \tilde{\pi}_y$$

So $\tilde{\pi}$ is a non-negative row vector, but it must be normalized to become a distribution, so we define

$$\pi_v = \frac{\deg(v)}{\sum_{u\in V}\deg(u)} = \frac{\deg(v)}{2|E|}$$

If the degree of each vertex is constant, suppose $\deg(v) = d$ for all $v \in V$, then $\pi_v = \frac{d}{dN} = \frac{1}{N}$ so π is a uniform distribution.

2.2.2 Theorem (Existence and Uniqueness Theorem)

Let P be the transition matrix of irreducible finite-state Markov chain, then there exists a unique stationary distribution π for P.

We know that $P\mathbf{1} = \mathbf{1}$ and so 1 is an eigenvalue for P, and since P and $P^\top$ are similar, they share eigenvalues. Thus $P^\top$ has an eigenvalue of 1 and therefore must have a stationary distribution. To show that this eigenvector is unique, we will show that the column eigenspace of P has a dimension of one, and since the eigenspaces of a matrix and its transpose are equal (think Jordan normal forms), this is sufficient. So we will show that if $h \in \mathbb{R}^N$ is an eigenvector of P with an eigenvalue of 1, it is of the form $h = (c, \ldots, c)^\top$. Because S is finite, there exists a state $a \in S$ such that $h_a = M$ is maximal. Now suppose there exists a $z \in S$ such that $h_z < M$ and $P_{az} > 0$ then

$$h_a = (Ph)_a = \sum_{y\in S} P_{ay}h_y = P_{az}h_z + \sum_{y\neq z} P_{ay}h_y < M\left(\sum_{y\in S} P_{ay}\right) = M = h_a$$

since $P_{az} > 0$ and $h_z < M$, and this is a contradiction. So for every state where $P_{az} > 0$, $h_z = M$. If we continue this proof (since $P^n h = h$), we get that if $a \to z$ then $h_z = M$. Since the Markov chain is irreducible, it is closed and therefore $h_z = M$ for every $z \in S$. ∎

Notice that the proof of existence here assumes nothing about S other than it being finite. But in the case that the chain is irreducible, we can also provide a constructive proof of the existence of a stationary distribution. But first, a lemma:

2.2.3 Lemma

For every two states $x, y \in S$ in a finite irreducible state space $\mathbb{E}_x[T_y] < \infty$.

Since S is irreducible and finite, there exists an $\varepsilon > 0$ and a $r \in \mathbb{N}$ such that for every $a, b \in S$, there exists a $j \leq r$ such that $P^j(a, b) > \varepsilon$. This is since S is connected and so between every two states there exists a path of length $\leq r$ (taking the maximum length of all paths, or just N) and so $P^j(a, b) > 0$. Take ε to be less than the minimum of all such $P^j(a, b)$, which we can do since S is finite.

Thus

$$\mathbb{P}((\exists m \in [0, \ldots, r])X_m = b \mid X_n = a) > \varepsilon$$

Now we know that $T_b > kr$ if and only if $X_0, \ldots, X_r \neq b$ and then we don't hit b for another $(k-1)r$ rounds, meaning $T_b > (k-1)r$. By homogeneity this means

$$\begin{aligned}\mathbb{P}(T_b > kr \mid X_0 = a) \leq \max_{a'} \mathbb{P}(T_b > (k-1)r \mid X_0 = a')\, \mathbb{P}((\forall m \in [0, r]) X_m \neq b \mid X_0 = a) \\ \leq \max_{a'} \mathbb{P}(T_b > (k-1)r \mid X_0 = a') \cdot (1-\varepsilon)\end{aligned}$$

and so by induction, this is $\leq (1-\varepsilon)^k$. Thus

$$\mathbb{E}[T_b \mid X_0 = a] = \sum_{n=0}^{\infty} \mathbb{P}(T_b > n \mid X_0 = a) \leq r \sum_{k=0}^{\infty} \mathbb{P}(T_b > kr \mid X_0 = a) \leq r \sum_{k=0}^{\infty} (1-\varepsilon)^k < \infty$$

The first inequality is due to the series being decreasing, and so we can take a summand and copy it r times, then take the rth next. ∎

Now we can construct a stationary distribution. Let us define

$$\tilde{\pi}_y = \mathbb{E}_{z_0}\begin{bmatrix}\text{the number of times } y \text{ is visited,} \\ \text{including at time } 0, \\ \text{before returning to } z_0\end{bmatrix} = \sum_{n=0}^{\infty} \mathbb{P}(X_n = y, T_{z_0} > n \mid X_0 = z_0)$$

The last equality is since this probability is equal to the number of visits being $\geq n$. This is well-defined as

$$\tilde{\pi}_y \leq \sum_{n=0}^{\infty} \mathbb{P}(T_{z_0} > n \mid X_0 = z_0) = \mathbb{E}_{z_0}[T_{z_0}]$$

and this is finite by the above lemma, so $\tilde{\pi}_y < \infty$. Now we will compute $(\tilde{\pi}P)_y$:

$$\begin{aligned}(\tilde{\pi}P)_y &= \sum_{x \in S} \tilde{\pi}_x P_{xy} \\ &= \sum_{x \in S} \sum_{n=0}^{\infty} \mathbb{P}_{z_0}(X_n = x, T_{z_0} > n) P_{xy} \\ &= \sum_{n=0}^{\infty} \sum_{x \in S} \mathbb{P}_{z_0}(X_n = x, T_{z_0} \geq n+1)\, \mathbb{P}(X_{n+1} = y \mid X_n = x) \\ &= \sum_{n=0}^{\infty} \sum_{x \in S} \mathbb{P}_{z_0}(X_{n+1} = y, X_n = x, T_{z_0} \geq n+1) \\ &= \sum_{n=0}^{\infty} \mathbb{P}_{z_0}(X_{n+1} = y, T_{z_0} \geq n+1) \\ &= \sum_{k=1}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} \geq k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} \geq k) + \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} = k) - \mathbb{P}_{z_0}(X_0 = y, T_{z_0} = 0) \\ &= \tilde{\pi}_y + \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} = k) - \delta(y = z_0)\end{aligned}$$

Notice that $X_k = y, T_{z_0} = k$ if and only if $T_{z_0} = k$ and $y = z_0$, and so the sum is equal to $\delta(y = z_0)$. So we get that $\tilde{\pi}P = \tilde{\pi}$ as required. So we just need to normalize it by

$$\sum_{x \in S} \tilde{\pi}_S = \mathbb{E}_{z_0}[T_{z_0}]$$

And thus the stationary distribution is

$$\pi_x = \frac{\tilde{\pi}_x}{\mathbb{E}_{z_0}[T_{z_0}]}$$

∎

2.2.4 Corollary

If P is irreducible then $\pi_a = \frac{1}{\mathbb{E}_a[T_a]}$.

Since π is unique we can choose any z_0 and get the same result. So we can choose $z_0 = a$ and so

$$\pi_a = \frac{\mathbb{E}\left[\begin{matrix}\text{The number of times we visit } a \\ \text{before returning to } a \\ \text{including } t = 0\end{matrix} \middle| X_0 = a\right]}{\mathbb{E}_a[T_a]}$$

The numerator here is obviously 1, and so $\pi_a = \frac{1}{\mathbb{E}_a[T_a]}$. ■

For example, we showed that for a connected graph where the degree of each vertex is d (a connected d-regular graph), $\pi_v = \frac{1}{N}$ where $N = |V|$. Thus since P is irreducible, we get that

$$\frac{1}{N} = \pi_v = \frac{1}{\mathbb{E}_a[T_a]} \implies \mathbb{E}_a[T_a] = N$$

This is independent of the structure of the graph. But importantly, T_a is dependent on the structure of the graph!

As another example, if P is symmetric then $\mathbf{1}^\top P = (P\mathbf{1})^\top = \mathbf{1}^\top$ and so $\frac{1}{N}\mathbf{1}$ is a stationary distribution of P. And thus $\mathbb{E}_a[T_a] = N$ where $N = |S|$.

2.2.5 Theorem

If $a \in S$ is a transient state and S is finite, then for every stationary distribution π, $\pi_a = 0$.

There are two cases we will consider: that a is connected to only transient states, and that there exists a recurrent state b such that $a \to b$. In the second case we have that $b \not\to a$ since a is transient and b is recurrent. Let $a_0 = a \to a_1 \to \cdots \to a_n \to b$ be the path from a to b, and we can assume that all a_i are transient (as otherwise we could set $b = a_i$ for the minimum i where a_i is recurrent). Let C be the connected component of b and π be a stationary distribution on all of S. Then

$$\sum_{z\in C}\pi_z = \sum_{z\in C}(\pi P)_z = \sum_{z\in C}\left(\sum_{y\in C}\pi_y P(y,z) + \sum_{y\notin C}\pi_y P(y,z)\right) = \sum_{y\in C}\pi_y\sum_{z\in C}P(y,z) + \sum_{z\in C}\sum_{y\notin C}\pi_y P(y,z)$$

Since C is closed and $y \in C$, we have that $\sum_{z\in C} P(y,z) = 1$ and thus we get that the left sum is $\sum_{y\in C}\pi_y$, and since the entire expression is equal to $\sum_{z\in C}\pi_z$, we must have that the right sum is zero. So for every $z \in C$ and $y \notin C$, $\pi_y P(y,z) = 0$.

This must be true in particular for $y = a_n$ and $z = b$, and since $P(a_n, b) > 0$ this means $\pi_{a_n} = 0$. And we claim inductively that $\pi_{a_k} = 0$, since

$$\pi_{a_k} = \sum_{y\in S}\pi_y P(y, a_k)$$

and so if $\pi_{a_k} = 0$ then $\pi_y P(y, a_k) = 0$ for all $y \in S$. Since $P(a_{k-1}, a_k) > 0$ this means $\pi_{a_{k-1}} = 0$. And so in particular we have that $\pi_a = \pi_{a_0} = 0$ as required. ■

Now suppose S is a finite state space, then it can be uniquely partitioned into

$$S = T \uplus C_1 \uplus \cdots \uplus C_n$$

where T is the set of all transient states, and C_i are irreducible components. We showed that for every stationary distribution π, for every $a \in T$ we have $\pi_a = 0$. And we also showed that for every $1 \leq i \leq n$ there exists a unique stationary distribution π_i whose support is C_i (meaning for every $a \notin C_i$, $\pi_i(a) = 0$). Thus a general stationary distribution is a normalized vector (meaning the sum of its coefficients is one) in $\operatorname{span}\{\pi_1, \ldots, \pi_n\}$. This is since the transition matrix P can be viewed as a block matrix over the partition of S.

For the next lemma, let us state a combinatorical fact: if $A \subseteq \mathbb{N}$ is closed under addition and has a greatest common divisor of 1, then $\mathbb{N} \setminus A$ is finite. This is trivial if $1 \in A$.

2.2.6 Lemma

Suppose P is the transition matrix of an irreducible, aperiodic, finite-state, homogeneus Markov chain. Then there exists an $r_0 > 0$ such that for all $r \geq r_0$ and $a, b \in S$, $P^r(a, b) > 0$.

Let us define as before $\tau(a) = \{n \geq 1 \mid P^n(a, a) > 0\}$. Since P is aperiodic, $d(a) = \gcd \tau(a) = 1$, and $\tau(a)$ is closed under addition since $P^{n+m}(a, a) \geq P^n(a, a)P^m(a, a)$. This means that $\mathbb{N} \setminus \tau(a)$ is finite. This means that $\bigcup_{a\in S}(\mathbb{N} \setminus \tau(a)) = \mathbb{N} \setminus \bigcap_{a\in S} \tau(a)$ is finite as well as the finite union of finite sets. Let t_0 be an upper bound for $\mathbb{N} \setminus \bigcap_{a\in S} \tau(a)$, so for every $t \geq t_0$ we have that $t \in \bigcap_{a\in S} \tau(a)$ meaning $P^t(a, a) > 0$ for all $a \in S$.

Since P is irreducible, for every $a, b \in S$ there exists an $n = n(a, b)$ such that $P^n(a, b) > 0$. Now n is bound by $|S|$ and therefore we can define $n_0 = \max_{a,b\in S} n(a, b)$ and so for every $r \geq t_0 + n_0$ we have that $r - n_0 \geq t_0$ and so $P^{r-n_0}(a, a) > 0$. Thus

$$P^r(a, b) \geq P^{r-n_0}(a, a)P^{n_0}(a, b) > 0$$

so $r_0 = t_0 + n_0$ satisfies the condition. ∎

2.2.7 Lemma

Again suppose P is irreducible and aperiodic, and let π be its unique stationary distribution. Then there exists an $0 < \alpha < 1$ and a constant $c > 0$ such that for every $k \in \mathbb{N}$ and every distribution vector v,

$$\left\|vP^k - \pi\right\|_1 \leq c\alpha^k$$

where $\|\cdot\|_1$ is the 1-norm on $\mathbb{R}^n$: $\|u\|_1 = \sum_{k=1}^{n}|u_i|$.

By the previous lemma, there exists an $r > 0$ such that $P^r > 0$ (meaning every coefficient of P^r is positive). Since P is finite, there exists a $0 < \delta < 1$ such that for every $a, b \in S$: $P^r(a, b) \geq \delta\pi_b$. Let Π be the matrix whose rows are all π. Then let us define the matrix Q by

$$P^r = \delta\Pi + (1 - \delta)Q$$

and since $P^r \geq \delta\Pi$ (pointwise), we have $Q \geq 0$ (pointwise). Now notice that Π is stochastic since $(\Pi\mathbf{1})_i = \pi\mathbf{1} = 1$, and so Q is also stochastic:

$$\mathbf{1} = P^r\mathbf{1} = \delta\mathbf{1} + (1 - \delta)Q\mathbf{1} \implies (1 - \delta)\mathbf{1} = (1 - \delta)Q\mathbf{1}$$

and since $\delta < 1$, $1 - \delta \neq 0$. Let us define $\theta := 1 - \delta$ and we will prove by induction that for all $k \geq 1$,

$$P^{rk} = (1 - \theta^k)\Pi + \theta^k Q^k$$

for $k = 1$ this is trivial. For the induction step,

$$P^{r(k+1)} = P^{rk}P^r = \left((1 - \theta^k)\Pi + \theta^k Q^k\right)P^r = (1 - \theta^k)\Pi P^r + \theta^k Q^k P^r$$

Since $\Pi P = \Pi$, we have that $\Pi P^r = \Pi$ and so this is equal to

$$= (1 - \theta^k)\Pi + \theta^k\left((1 - \theta)Q^k\Pi + \theta Q^{k+1}\right) = (1 - \theta^k)\Pi + \theta^k(1 - \theta)Q^k\Pi + \theta^{k+1}Q^{k+1}$$

Now since Q^k is stochastic and Π's columns are constant, $Q^k\Pi = \Pi$. And so this is equal to

$$= (1 - \theta^k + \theta^k - \theta^{k+1})\Pi + \theta^{k+1}Q^{k+1} = (1 - \theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}$$

as required.

And so now we have for all $j \geq 0$, $P^{rk+j} = (1 - \theta^k)\Pi + \theta^k Q^k P^j$ and so

$$P^{rk+j} - \Pi = \theta^k\left(Q^k P^j - \Pi\right)$$

Since $Q^k P^j$ and Π are all stochastic matrices and thus their coefficients all are bound by 1, the coefficients of $Q^k P^j - \Pi$ all have an absolute value bound by 1 as well. Now since $(v\Pi)_i$ is equal to v times the ith column of Π which is $\pi_i\mathbf{1}$, we have $(v\Pi)_i = \pi_i v\mathbf{1} = \pi_i$ (since v is a distribution, $v\mathbf{1} = 1$). And so $\pi = v\Pi$, so

$$\left\|vP^{rk+j} - \pi\right\|_1 = \left\|vP^{rk+j} - v\Pi\right\|_1 = \left\|v(P^{rk+j} - \Pi)\right\|_1 = \theta^k\left\|v(Q^k P^j - \Pi)\right\|_1$$

since $Q^k P^j - \Pi$'s coefficients are all bound by 1, the norm is bound by a constant (which is the norm of v times the matrix of all ones, since v is positive). So we have that $\left\|vP^{rk+j} - \pi\right\|_1 \leq c\theta^k$ and finding the appropriate values, we can bound this by some $c'\alpha^{rk+j}$. ∎

2.2.8 Theorem

Let P be the transition matrix of an irreducible aperiodic Markov chain, and let π be its unique stationary distribution. Then for every initial distribution v, $vP^n \xrightarrow{n\to\infty} \pi$ pointwise (meaning $(vP^n)_i \xrightarrow{n\to\infty} \pi_i$). Since $(vP^n)_i = \mathbb{P}_v(X_n = i)$, equivalently $\mathbb{P}_v(X_n = i) \xrightarrow{n\to\infty} \pi_i$ or $X_n \xrightarrow{d} \pi$.

So we must simply show that $|(vP^n)_i - \pi_i| \xrightarrow{n\to\infty} 0$. This is an immediate consequence of the previous lemma, which gave us that $\|vP^n - \pi\|_1 \leq c\alpha^n$ and so in particular $\|vP^n - \pi\|_1 \xrightarrow{n\to\infty} 0$. Since $\|vP^n - \pi\|_1 = \sum_{i=1}^N |(vP^n)_i - \pi_i|$, certainly $|(vP^n)_i - \pi_i| \xrightarrow{n\to\infty} 0$, as required. (In general convergence in the p-norms of $\mathbb{R}^N$ is equivalent to pointwise convergence.) ■

2.2.9 Corollary

If P is a stochastic matrix then all of its eigenvalues are bound by 1 (in absolute value).

Let γ be an eigenvalue of P, then there exists a vector v such that $Pv = \lambda v$. Let j be the state in S such that $|v_j| = \max_{i\in S}|v_i|$ and so

$$|\lambda||v_j| = |(Pv)_j| = \left|\sum_{i\in S} P_{ji}v_i\right| \leq \sum_{i\in S} P_{ji}|v_i| \leq |v_j| \sum_{i\in S} P_{ji} = |v_j|$$

Thus $|\lambda| \leq 1$. ■

2.3 Mixing Times

2.3.1 Definition

Let μ and ν be two be two probability measures over the same σ-algebra $\mathcal{F}$, then we define their **total variation** to be

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A\in\mathcal{F}}|\mu(A) - \nu(A)|$$

This is also denoted $d_{\mathrm{TV}}(\mu, \nu)$, and this is in fact a metric over the space of probability measures on $\mathcal{F}$.

If μ and ν are discrete probability distributions on Ω, then let $B = \{x \mid \mu(x) \geq \nu(x)\}$, and let $A \subseteq \Omega$ be any event. Then for any $x \in A \cap B^c$, $\mu(x) - \nu(x) < 0$ and so $\mu(A \cap B^c) - \nu(A \cap B^c) \leq 0$. Thus

$$\mu(A) - \nu(A) = \mu(A\cap B) - \nu(A \cap B) + \mu(A \cap B^c) - \nu(A \cap B^c) \leq \mu(A \cap B) - \nu(A \cap B)$$

and for every $x \in B \cap A^c$, $\mu(x) - \nu(x) \geq 0$ and so $\mu(B \cap A^c) - \nu(B \cap A^c) \geq 0$ so

$$\leq \mu(B) - \nu(B)$$

And similarly we have that $\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c) = \mu(B) - \nu(B)$. Thus we have that for every event A, $|\mu(A) - \nu(A)| \leq \mu(B) - \nu(B)$ and so

$$\begin{aligned}\|\mu - \nu\|_{\mathrm{TV}} = \mu(B) - \nu(B) = \frac{1}{2}\big(\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)\big) &= \frac{1}{2}\left(\sum_{x\in B}\big(\mu(x) - \nu(x)\big) + \sum_{x\notin B}\big(\nu(x) - \mu(x)\big)\right) \\ &= \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|\end{aligned}$$

the second equality is since $\mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$. So we have proven

2.3.2 Proposition

If μ and ν are two discrete probability measures over the same space, then their total variation distance is equal to half of their L^1 distance, ie.

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|$$

This holds in particular for when μ and ν are distribution vectors.

2.3.3 Definition

Let P be the transition matrix of an irreducible Markov chain whose stationary distribution is π, the we define

$$d(k) = \max_{j \in S} d_{\mathrm{TV}}(e_j P^k, \pi)$$

Since $e_j P^k$ and π are both distributions, they can be viewed as probability measures, and so we can discuss their total variation. $e_j P^k$ is the distribution of X_k if $X_0 = j$, and so $d(k)$ gives us the maximum total variation of the distribution of X_k and π over all possible initial states. Let us also define the **mixing time** to be

$$t_{\mathrm{mix}}(\varepsilon) = \min\{k \mid d(k) \leq \varepsilon\}$$

$t_{\mathrm{mix}}(\varepsilon)$ gives us the minimum k where the total variation of the distribution X_k and π is less than ε, independent of the initial state. Though generally if we talk about the "mixing time" of a Markov chain, we set $\varepsilon = \frac{1}{4}$. And finally we also define

$$\bar{d}(k) = \max_{i,j \in S} d_{\mathrm{TV}}(e_i P^k, e_j P^k)$$

By the triangle inequality, $\bar{d}(k) \leq 2d(k)$. And in fact $d(k) \leq \bar{d}(k)$ so

$$d(k) \leq \max_{i,j \in S} \left\| e_i P^k - e_j P^k \right\|_{\mathrm{TV}}$$

2.3.4 Definition

A **coupling** of two probability measures μ and ν over the same σ-algebra $\mathcal{F}$ is a pair of random variables (X, Y) such that $X \sim \mu$ and $Y \sim \nu$. Formally, a coupling is a new probability space and random variables whose codomain is $\mathcal{F}$ such that for every $A \in \mathcal{F}$, $\mathbb{P}(X \in A) = \mu(A)$ and $\mathbb{P}(Y \in A) = \nu(A)$.

2.3.5 Proposition

If μ and ν are probability measures over the same σ-algebra, then

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \inf\{\mathbb{P}(X \neq Y) \mid (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$$

Let (X, Y) be a coupling and $A \in \mathcal{F}$ then

$$\mu(A) - \nu(A) = \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \in A, Y \notin A) \leq \mathbb{P}(X \neq Y)$$

and taking the infimum over all couplings (X, Y) preserves this inequality. ∎

In fact, there is actually an equality here but the other direction is harder to prove.

2.3.6 Theorem

Suppose $\{X_n\}$ and $\{Y_n\}$ are two Markov chains with the same transition matrix P. Further suppose that if $X_s = Y_s$ then $X_t = Y_t$ for all $t \geq s$, then

$$\left\| e_x P^t - e_y P^t \right\|_{\mathrm{TV}} \leq \mathbb{P}(X_t \neq Y_t \mid X_0 = x, Y_0 = y)$$

This is as $e_x P^t$ and $e_y P^t$ are the distributions of X_t and Y_t under the assumption that $X_0 = x$ and $Y_0 = y$. And (X_t, Y_t) is certainly a coupling of these distributions in $\mathbb{P}(\cdot \mid X_0 = x, Y_0 = y)$. ∎

This means that if $\{X_n\}$ is a Markov chain, and $\{Y_n\}$ is some other Markov chain with the same transition matrix then $d(k)$ (for either $\{X_n\}$ or $\{Y_n\}$) can be bound by:

$$d(k) \leq \max_{i,j \in S} \mathbb{P}(X_k \neq Y_k \mid X_0 = i, Y_0 = j)$$

2.3.7 Example

What is the mixing time of the random walk on the circle C_N (this is the graph of N nodes, $\{v_1, \dots, v_N\}$ with the edges $\{v_i, v_i\}$ and $\{v_i, v_{i+1}\}$)? Let us define two Markov chains X_n and Y_n where at every step we choose a random chain with equal probability and that will be the chain which will make the next step. As soon as the two chains intercept, they step together. Let T be the time that the two chains intercept, then by above and Markov's inequality

$$d(t) \leq \max_{x,y} \mathbb{P}_{x,y}(T > t) \leq \max_{x,y} \frac{\mathbb{E}_{x,y}[T]}{t}$$

The expected hitting time of $k \in \{0, \dots, N\}$ for a random walk on the circle is $k(N-k)$ (this will be shown later), which takes a maximum at $k = \frac{N}{2}$, and so $\max \mathbb{E}_{x,y}[T] \leq \frac{N^2}{4}$. And since we measure mixing times with $\varepsilon = \frac{1}{4}$ we get that if $\frac{N^2}{4t} \leq \frac{1}{4}$, meaning $t \geq N^2$ (and in particular if $t = N^2$), then $d(t) \leq \frac{1}{4}$. So $t_{\text{mix}} \leq N^2$.

2.4 Famous Markov Chains

In this subsection we will discuss various useful Markov chains.

1 Gambler's Ruin

The first one we will discuss is called the *Gambler's Ruin*: suppose a gambler goes to a casino with the goal of winning n dollars. If the gambler reaches his goal of n dollars or fails and loses all his money (reaches 0 dollars), he leaves the casino. Suppose he bets a single dollar each time, and has a fair chance of winning. We can ask two questions: how much time will it take for the gambler to leave the casino, and what is the probability that the gambler goes broke (reaches 0 before n)?

So we can define a Markov chain X_n where X_n is the amount of money the gambler has after n bets. The transition matrix here is

$$P(i \to i+1) = P(i \to i-1) = \frac{1}{2} \text{ for } 0 < i < n, \qquad P(0 \to 0) = P(n \to n) = 1$$

Let us define $\tau = \min\{T_0, T_n\}$ which is the time the gambler will leave the casino. We make two claims:

$$\mathbb{P}_k(X_\tau = n) = \frac{k}{n}, \qquad \mathbb{E}_k[\tau] = 4k(n-k)$$

so if the gambler starts with k dollars, the probability he gets his goal of n dollars is $\frac{k}{n}$, and the expected time it takes him to leave the casino is $4k(n-k)$. Note that $X_\tau = n$ is equivalent to $T_n < T_0$.

To prove this let us define $p_k = \mathbb{P}_k(X_\tau = n)$. Then $p_0 = 0$ and $p_n = 1$. And using first step analysis,

$$p_k = \mathbb{P}(T_n < T_0 \mid X_0 = k) = \frac{1}{2}\mathbb{P}(T_n < T_0 \mid X_1 = k+1) + \frac{1}{2}\mathbb{P}(T_n < T_0 \mid X_1 = k-1) = \frac{1}{2}p_{k+1} + \frac{1}{2}p_{k-1}$$

Now given the initial conditions, there must be a unique solution to this. And since $p_k = \frac{k}{n}$ works as a solution, it must be the unique solution.

Let us denote $\mu_k = \mathbb{E}_k[\tau]$, and then $\mu_0 = \mu_n = 0$. And we also get

$$\begin{aligned}
\mu_k = \mathbb{E}[\tau \mid X_0 = k] = \frac{1}{2}\mathbb{E}[\tau]X_1 = k+1 + \frac{1}{2}\mathbb{E}[\tau \mid X_1 = k-1] \\
&= \frac{1}{2}\big(1 + \mathbb{E}[\tau \mid X_0 = k+1]\big) + \frac{1}{2}\big(1 + \mathbb{E}[\tau \mid X_0 = k-1]\big) \\
&= 1 + \frac{1}{2}\mu_{k-1} + \frac{1}{2}\mu_{k+1}
\end{aligned}$$

Again, this must have a unique solution due to the initial conditions, and $4k(n-k)$ satisfies this.

2 Coupon Collector

There are n types of coupons, and we would like to collect them all. When we are given a coupon, the probability it is a specific type distributes uniformly. So let us define X_k to be the number of types of coupons we have after collecting k coupons, and so

$$P(i \to i) = \frac{i}{n}, \qquad P(i \to i+1) = 1 - \frac{i}{n}$$

n is the only absorbing state and so all other states are transient and we will eventually be absorbed by n with probability 1. What is the expected time that we hit n for the first time? Now, $T_{k+1} - T_k$ is the number of times that we get one of the k types of coupons we already have, and so $T_{k+1} - T_k \sim \text{Geo}\big(1 - \frac{k}{n}\big)$. Thus

$$\mathbb{E}_0[T_n] = \sum_{k=0}^{n-1} \mathbb{E}_0[T_{k+1} - T_k] = \sum_{k=0}^{n-1} \frac{1}{1 - \frac{k}{n}} = n\sum_{k=1}^{n} \frac{1}{k} \sim n\log(n)$$

What is the probability it took "much more time" to get to n? We claim

$$\mathbb{P}(T_n > \lceil n\log n + cn\rceil) \le e^{-c}$$

Let A_i be the probability that after $\lceil n\log n + cn\rceil$ time, we have not collected the ith coupon type. Then

$$\mathbb{P}(T_n > \lceil n\log n + cn\rceil) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n}\mathbb{P}(A_i) = \sum_{i=1}^{n}\left(1-\frac{1}{n}\right)^{\lceil n\log n+cn\rceil}$$

since $1-x\le e^{-x}$ this can be further bound by

$$\le n\left(e^{-\frac{1}{n}}\right)^{n\log n+cn} = ne^{-\log n - c} = e^{-c}$$

as required.

3 Pólya Urn

In an urn there are two balls: one white and one black. At every step we choose an arbitrary ball and add a new one of the same color. Let us define (B_k, W_k) to be the number of black and white balls, respectively. The state space is then $S=\mathbb{N}^2$. The transistion probabilities are

$$P\big((i,j)\to(i+1,j)\big) = \frac{i}{i+j}, \qquad P\big((i,j)\to(i,j+1)\big) = \frac{j}{i+j}$$

Is B_k a Markov chain? Well if we know B_k then we know that $W_k=(k+2)-B_k$ since the total number of balls after k steps is $k+2$. And so we can then determine the probability for transitioning, so it is indeed a Markov chain. But it is not homogeneus: in order to determine the transition probability we must use k: $\mathbb{P}(B_{k+1}=i+1 \mid B_k=i) = \frac{i}{k+2}$, which is dependent on k.

Nevertheless we claim that $B_k\sim \mathrm{Unif}\{1,2,\dots,k+1\}$. We prove this by induction: for $k=1$ this is trivial as the probabilities that $B_k=1$ and $B_k=2$ are the same. And if $B_{k-1}\sim\mathrm{Unif}\{1,2,\dots,k\}$ then

$$\begin{aligned}\mathbb{P}(B_k=j) &= \mathbb{P}(B_k=j\mid B_{k-1}=j-1)\cdot\mathbb{P}(B_{k-1}=j-1)+\mathbb{P}(B_k=j\mid B_{k-1}=j)\cdot\mathbb{P}(B_{k-1}=j)\\ &= \frac{j-1}{k+1}\cdot\frac{1}{k}+\left(1-\frac{k-1}{k+1}\right)\cdot\frac{1}{k} = \frac{1}{k}\cdot\left(1-\frac{1}{k+1}\right) = \frac{1}{k+1}\end{aligned}$$

as required.

4 Random Walks on $\mathbb{Z}$

This is a homogeneus chain on $\mathbb{Z}$ whose transitions are

$$P(k\to k+1)=p,\quad P(k\to k)=r,\quad P(k\to k-1)=q \qquad\qquad (p+q+r=1)$$

We will first focus on the case that $r=0$ and $p=q=\frac{1}{2}$, which is a *fair* walk. In this case we claim that for all $t,j,k>0$,

$$\mathbb{P}_k(T_0<t, X_t=j)=\mathbb{P}_k(X_t=-j), \qquad \mathbb{P}_k(T_0<t,X_t>0)=\mathbb{P}_k(X_t<0)$$

This is as if $T_0=s$ then the walk starting from time s is equivalent to the walk starting from $X_0=0$. Thus

$$\mathbb{P}_k(T_0=s,X_t=j)=\mathbb{P}_k(T_0=s)\cdot\mathbb{P}_0(X_{t-s}=j)$$

by symmetry this is equal to

$$=\mathbb{P}_k(T_0=s)\cdot\mathbb{P}_0(X_{t-s}=-j)=\mathbb{P}_k(T_0=s,X_t=-j)$$

And so we get that

$$\mathbb{P}_k(T_0<t,X_t=j)=\sum_{s<t}\mathbb{P}_k(T_0=s,X_t=j)=\sum_{s<t}\mathbb{P}_k(T_0=s,X_t=-j)=\mathbb{P}_k(T_0<t,X_t=-j)$$

if we start at k and at time t, $X_t=-j<0$ then necessarily $T_0<t$ so this is equal to $\mathbb{P}_k(X_t=-j)$ as required. And now

$$\mathbb{P}_k(T_0<t,X_t>0)=\sum_{j>0}\mathbb{P}_k(T_0<t,X_t=s)=\sum_{j>0}\mathbb{P}_k(X_t=-s)=\mathbb{P}_k(X_t<0)$$

as required.

We further claim that for every $k>0$,

$$\mathbb{P}_k(T_0>t)=\mathbb{P}_0(-k<X_t\le k)$$

This is as

$$\mathbb{P}_k(X_t > 0) = \mathbb{P}_k(X_t > 0, T_0 \le t) + \mathbb{P}_k(T_0 > t)$$

now we showed that $\mathbb{P}_k(X_t > 0, T_0 \le t) = \mathbb{P}_k(X_t < 0)$ and by symmetry about k this is equal to $\mathbb{P}_k(X_t > 2k)$. Thus we get that

$$\mathbb{P}_l(T_0 > t) = \mathbb{P}_k(X_t > 0) - \mathbb{P}_k(X_t > 2k) = \mathbb{P}_k(0 < X_t \le 2k) = \mathbb{P}_0(-k < X_t \le k)$$

as required.

Notice that if we start at $X_0 = 0$ then in order to get to $X_t = k$, we must take r steps to the right and ℓ steps to the left where $r + \ell = t$ and $r - \ell = k$. This means that $r = \frac{t+k}{2}$ and $\ell = \frac{t-k}{2}$. Thus

$$\mathbb{P}_0(X_t = k) = \begin{cases} \binom{t}{\frac{t-k}{2}} 2^{-t} & t \equiv k \pmod 2 \\ 0 & \text{else} \end{cases}$$

And so by Stirling we get

$$\mathbb{P}_0(X_t = k) \le \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{t}} = \frac{c}{\sqrt{t}}$$

And this means that

$$\mathbb{P}_k(T_0 > t) = \mathbb{P}_0(-k < X_t \le k) = \sum_{j=-k+1}^{k} \mathbb{P}_0(X_t = j) \le \frac{2ck}{\sqrt{t}}$$

2.4.1 Proposition

For a fair walk on $\mathbb{Z}$, every state is transient but the expected return time to each state is infinite.

We will prove that every state is transient in two ways. For the first way, let us define $A_k = \{T_{\pm 2^k} < T_0\}$ the event that we get to 2^k or -2^k before 0. And so $\mathbb{P}_0(A_{k+1} \mid A_k) = \mathbb{P}_{2^k}(T_0 < T_{2^{k+1}}) = \frac{1}{2}$ since the distance between 0 and 2^k is the same as the distance between 2^k and 2^{k+1}. Since $\mathbb{P}_0(A_1) = \frac{1}{2}$ and $\mathbb{P}_0(A_k) = \mathbb{P}_0(A_k \mid A_{k-1}) \cdot \mathbb{P}_0(A_{k-1})$, by induction $\mathbb{P}_0(A_k) = 2^{-k}$. And

$$\mathbb{P}_0(T_0 < \infty) = \mathbb{P}_0\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} 2^{-n} = 0$$

where the second equality is since $\{A_k\}$ is a decreasing sequence and so this is due to the continuity of probability. So 0 is recurrent and since all states are connected, so is every other state.

For the second proof, by Stirling $\mathbb{P}_0(X_{2n} = 0) = 2^{-2n}\binom{2n}{n} \ge \frac{c}{\sqrt{n}}$ and so

$$\sum_{n=1}^{\infty} \mathbb{P}_0(X_n = 0) = \sum_{n=1}^{\infty} \mathbb{P}_0(X_{2n} = 0) \ge \sum_{n=1}^{\infty} \frac{c}{\sqrt{n}} = \infty$$

and since $N_0(0) = \sum_{n=1}^{\infty} \chi\{X_n = 0\}$, we get that

$$\mathbb{E}[N_0(0)] = \sum_{n=1}^{\infty} \mathbb{P}(X_n = 0) = \infty$$

which means that 0 is recurrent (since $N_0(0) \sim \text{Geo}(1 - f_0)$ if $f_0 < 0$ then its expected value would be finite, so $f_0 = 1$ meaning 0 is recurrent).

To compute the expected return time, let us denote $\alpha = \mathbb{E}_1[T_0] = \mathbb{E}_n[T_{n-1}]$. And by first step analysis,

$$\alpha = \mathbb{E}_1[T_0] = \frac{1}{2}\mathbb{E}[T_0 \mid X_1 = 0] + \frac{1}{2}\mathbb{E}[T_0 \mid X_1 = 2] = \frac{1}{2} + \frac{1}{2}\big(1 + \mathbb{E}_2[T_0]\big)$$

Now, $\mathbb{E}_2[T_0] = \mathbb{E}_2[T_1] + \mathbb{E}_1[T_0]$ since this is the expected time to go from 2 to 1 to 0 (the only path from 2 to 0), and this is equal to 2α. Thus we get that $1 + \alpha = \alpha$. But no finite number satisfies this, so $\alpha = \infty$. And so we have show that

$$\mathbb{P}_0(T_0 < \infty) = 1, \qquad \mathbb{E}_0[T_0] = \infty$$

■

2.5 Asymptotic Behavior

2.5.1 Definition

Let $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ be a sequence of events, then let us define

$$\{A_n \text{ i.o.}\} = \{\omega \in \Omega \mid (\forall k)(\exists m \geq k)\omega \in A_m\} = \bigcap_{k=1}^{\infty}\bigcup_{m=k}^{\infty} A_m = \limsup A_n$$
$$\{A_n \text{ a.e.}\} = \{\omega \in \Omega \mid (\exists k)(\forall m \geq k)\omega \in A_m\} = \bigcup_{k=1}^{\infty}\bigcap_{m=k}^{\infty} A_m = \liminf A_n$$

So $\{A_n \text{ i.o.}\}$ is the set of all elements which are in infinitely many A_ns, and $\{A_n \text{ a.e.}\}$ is the set of all elements which are in all but finitely many A_ns.

Notice that in general

$$\{A_n \text{ i.o.}\}^c = \{A_n^c \text{ a.e.}\}, \qquad \{A_n \text{ a.e.}\} \subseteq \{A_n \text{ i.o.}\}$$

So for example, let $(\{0,1\}^{\mathbb{N}}, \mathcal{F}, \mathbb{P})$ be the probability space of the flipping of a fair coin. Then let us define $A_n = \{\omega \mid \omega_n = 1\}$, the event that the nth flip resulted in 1. Then $\{A_n \text{ i.o.}\}$ is the set of all ω with infinitely many 1s, and $\{A_n \text{ a.e.}\}$ is the set of all ωs with finitely many 0s. Notice that all A_n are independent since the coin flips are independent and so

$$\mathbb{P}(A_n \text{ a.e.}) = \mathbb{P}\left(\bigcup_k \bigcap_{m\geq k} A_m\right) \leq \sum_k \mathbb{P}\left(\bigcap_{m\geq k} A_m\right) = \sum_k \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=k}^{k+n} A_m\right) = \sum_k \lim_n \prod_{m=k}^{k+n} \mathbb{P}(A_m) = \sum_k \lim_n \frac{1}{2^n} = \sum_k 0 = 0$$

2.5.2 Lemma (Borel-Cantelli Lemma)

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events, then

(1) If $\sum \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

(2) If $\sum \mathbb{P}(A_n) = \infty$ and $\{A_n\}$ is independent then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

For the first, due to the continuity of measures

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0$$

the final equality is since the series $\sum_{k=1}^{\infty} \mathbb{P}(A_k)$ converges and so its tail must converge to zero. For the second,

$$\mathbb{P}(A_n \text{ i.o.}) = 1 - \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{k=m}^{\infty} A_k^c\right) \geq 1 - \sum_{m=1}^{\infty} \mathbb{P}\left(\bigcap_{k=m}^{\infty} A_k^c\right)$$

We will show that for every m, $\mathbb{P}(\bigcap_{k=m}^{\infty} A_k^c) = 0$ and this will be sufficient.

$$\mathbb{P}\left(\bigcap_{k=m}^{\infty} A_k^c\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=m}^{m+n} A_k^c\right) = \lim_{n\to\infty} \prod_{k=m}^{m+n} \mathbb{P}(A_k^c) = \lim_{n\to\infty} \prod_{k=m}^{m+n} (1 - \mathbb{P}(A_k))$$

since $1 - x \leq e^{-x}$,

$$\leq \lim_{n\to\infty} \exp\left(-\sum_{k=m}^{m+n} \mathbb{P}(A_k)\right)$$

Since the sum goes to $-\infty$, this goes to zero, as required. ■

2.5.3 Example

We say that a number is *normal* if in its base 10 representation, every finite string occurs infinitely many times. What is the probability that a number chosen uniformly in $[0,1]$ is normal? Suppose we choose $U = 0.X_1X_2X_3\ldots \in [0,1]$ where X_i is uniformly chosen, $X_i \sim \mathrm{Unif}\{0,\ldots,9\}$. Let us set a finite string $S = S_1 \cdots S_N$ and define

$$A_i = \left\{X_{iN+1} = S_1, X_{iN+2} = S_2, \ldots, X_{(i+1)N} = S_N\right\}$$

A_i is the event that the string S occurs in U beginning at index $iN+1$. $\{A_i\}$ are all independent since A_i looks at a disjoint set of X_ks than A_j does. And $\mathbb{P}(A_i) = \frac{1}{10^N}$ for every i, so $\sum_{i=1}^{\infty}\mathbb{P}(A_i) = \infty$ and so by the Borel-Cantelli Lemma we have that $\mathbb{P}(A_n \text{ i.o}) = 1$. Meaning that the probability S occurs infinitely many times in U is 1.

Notice that this does not mean the probability of U being normal is 1, rather that the probability that U has an arbitrary finite string occurring infinitely many times is 1. But this does not necessarily mean that the probability of every finite string occurring infinitely many times is 1. Fortunately it does, since if we denote the events by A_i^S, then we want to compute the probability of $\bigcap_S(A_n^S \text{ i.o.})$. Now, the countable intersection of probability-1 events also has probability 1: if $\mathbb{P}(B_n) = 1$ then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n^c\right) \geq 1 - \sum_{n=1}^{\infty}\mathbb{P}(B_n^c) = 1$$

And so $\mathbb{P}\big(\bigcap_S(A_n^S \text{ i.o.})\big) = 1$ as required.

2.5.4 Example

Suppose we have a random process $\{X_n\}$ where each step is independent and distributes the same. Let us define $S_0 = 0$ and $S_n = \sum_{i=1}^{n} X_i$ then we claim

$$\mathbb{P}(S_n = S_m \text{ i.o.}) \in \{0,1\}$$

Suppose there exists an N such that $p = \mathbb{P}(X_1 + \cdots + X_N = 0) > 0$ then let us define $A_i = \left\{\sum_{j=iN}^{(i+1)N} X_j = 0\right\}$ then since X_i are all independent and have the same distribution, $\mathbb{P}(A_i) = p > 0$ for all i. Then since A_i are also all independent, by the Borel-Cantelli Lemma, $\mathbb{P}(A_i \text{ i.o.}) = \mathbb{P}\big(S_{iN-1} = S_{(i+1)N} \text{ i.o.}\big) = 1$ and this implies $\mathbb{P}(S_n = S_m \text{ i.o}) = 1$. Alternatively, for every N, $\mathbb{P}(X_1 + \cdots + X_N) = 0$ and this means that we can never have that $S_n = S_m$ for $n > m$, as then $X_{n+1} + \cdots + X_m = 0$. Thus $\mathbb{P}((\exists n > m)S_n = S_m) = 0$, as required.

2.5.5 Definition

Let $\{A_j\}$ be a sequence of events. Then a **tail event** is an event in the σ-algebra $\sigma\big(\{A_j\}_{j=k}^{\infty}\big)$ for every $k > 0$. Ie. tail events are elements of the σ-algebra $\bigcap_{k=1}^{\infty}\sigma(\{A_j\}_{j=k}^{\infty})$. Recall that $\sigma(\mathcal{F})$ is the σ-algebra generated by the family of sets $\mathcal{F}$.

2.5.6 Theorem (Kolmogorov's Zero-One Law)

If $\{A_j\}$ is a sequence of independent events, then every tail event is trivial.

We lack tools to fully justify each step, but an outline of the proof is as follows: it can be shown that if the generators of a σ-algebra are independent then so is the σ-algebra. Thus for every k, $\sigma(A_1,\ldots,A_k)$ and $\sigma(A_{k+1},\ldots)$ are independent. So let B be a tail event, thus $B \in \sigma(A_{k+1},\ldots)$ and so B is independent of every $\sigma(A_1,\ldots,A_k)$. And this means that B is independent of $\sigma(A_1,\ldots)$, and in particular B is independent of itself. And so $\mathbb{P}(B) = \mathbb{P}(B \cap B) = \mathbb{P}(B)^2$ so $\mathbb{P}(B) \in \{0,1\}$. ■

For random variables there exists a variation of the Zero-One law:

2.5.7 Theorem

If $\{X_j\}$ are independent random variables, and if Y is a random variable (measurable) with respect to $\sigma(\{X_j\})$ then there exists a constant c such that $\mathbb{P}(Y = c) = 1$.

2.5.8 Theorem (Hewitt-Savage)

Suppose $\{X_j\}_{j=1}^{\infty}$ is a sequence of independent and equal-distribution random variables. Let $A \in \sigma(\{X_j\}_{j=1}^{\infty})$ be an event such that for every for every finite permutation of indexes π, $\pi A = A$ (since elements of A are of the form $\omega = (\omega_1, \omega_2, \ldots)$, and so $\pi A = \{\pi\omega \mid \omega \in A\}$ where π acts on the vector ω). Then $\mathbb{P}(A) \in \{0, 1\}$.

There must exist a sequence of events $A_n \in \sigma(X_1, \ldots, X_n)$ such that $\mathbb{P}(A_n \triangle A) \to 0$. Then let us define

$$\pi(j) = \begin{cases} j+n & 1 \le j \le n \\ j-n & n+1 \le j \le 2n \\ j & j > 2n \end{cases}$$

this is a permutation of only a finite number of indexes, and notice that $\pi^2 = \mathrm{id}$. Now we have that

$$\mathbb{P}(\{\omega \mid \omega \in A_n \triangle A\}) = \mathbb{P}(\{\omega \mid \pi\omega \in A_n \triangle A\})$$

And since $\{\omega \mid \pi\omega \in A\} = \pi^{-1}A = \pi A = A$ and A_n is of the form $\{\omega \mid (\omega_1, \ldots, \omega_n) \in B_n\}$ so $\{\omega \mid \pi\omega \in A_n\} = \{\omega \mid (\omega_{n+1}, \ldots, \omega_{2n}) \in B_n\} = A'_n$, we get that $\mathbb{P}(A \triangle A_n) = \mathbb{P}(A \triangle A'_n)$. In general one has $|\mathbb{P}(B) - \mathbb{P}(C)| \le \mathbb{P}(B \triangle C)$, and so $\mathbb{P}(A_n) \to \mathbb{P}(A)$ and $\mathbb{P}(A'_n) \to \mathbb{P}(A)$.

$$\mathbb{P}(A_n \triangle A'_n) \le \mathbb{P}(A_n \triangle A) + \mathbb{P}(A'_n \triangle A) \to 0$$

(since $A \triangle B \subseteq (A \triangle C) \cup (B \triangle C)$.) And so $\mathbb{P}(A_n) - \mathbb{P}(A_n \cap A'_n) \le \mathbb{P}(A_n \triangle A'_n) \to 0$, meaning that $\mathbb{P}(A_n \cap A'_n) \to \mathbb{P}(A)$. But at the same time, since A_n and A'_n are independent (as they refer to different X_is), we get that $\mathbb{P}(A_n \cap A'_n) = \mathbb{P}(A_n) \cdot \mathbb{P}(A'_n) \to \mathbb{P}(A)^2$. Thus $\mathbb{P}(A) = \mathbb{P}(A)^2$ meaning $\mathbb{P}(A) \in \{0, 1\}$. ∎

3 Brownian Motion

A general random process is a sequence of random variables $\{S_n\}_{n=1}^{\infty}$ where $S_n = X_1 + \cdots + X_n$ where $\{X_i\}_{i=1}^{\infty}$ are independent and equal-distribution. By the law of large numbers, if $\mathbb{E}[X_n] = 0$ then $\frac{S_n}{n} \xrightarrow{a.s.} 0$, and if further $\mathbb{E}\big[X_n^2\big] = 1$ (meaning $\mathrm{Var}(X_n) = 1$) then by the central limit theorem $\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$. In general, we get the following by Hewitt-Savage (this was in homework):

3.0.1 Theorem

Let S_n be a general random process, then one of the following occurs with probability 1:

$$(1)\ (\forall n)S_n = 0, \quad (2)\ S_n \to \infty \quad (3)\ S_n \to -\infty \quad (4)\ \limsup S_n = \infty,\ \liminf S_n = -\infty$$

3.0.2 Definition

A collection of random variables $\{B(t)\}_{t\geq 0}$ (meaning that for every $0 \leq t \in \mathbb{R}$, $B(t)$ is a random variable) is called **Brownian motion** which starts at $x \in \mathbb{R}$ if the following conditions are met:

(1) $B(0) \overset{as}{=} x$,

(2) Differences are independent: for every $0 \leq t_1 \leq \cdots \leq t_n$, $B(t_2) - B(t_1), \dots, B(t_n) - B(t_{n-1})$ are independent,

(3) Differences are normal: for every $t, h \geq 0$, $B(t+h) - B(t) \sim \mathcal{N}(0,h)$,

(4) Continuity: with probability 1, $t \mapsto B(t)$ is continuous.

Since each $B(t)$ is a random variable, meaning a function $B(t)\colon \Omega \longrightarrow \mathbb{R}$, we can view Brownian motion as a function $B\colon [0,\infty) \times \Omega \longrightarrow \mathbb{R}$ where $B(t,\omega) = B(t)(\omega)$. The final condition can then be stated with more formality:

$$\mathbb{P}(\{\omega \in \Omega \mid t \mapsto B(t,\omega) \text{ is continuous}\}) = 1$$

Brownian motion describes a continuous random process, unlike Markov chains whose steps are all discrete.

Now suppose $B(t)$ is Brownian motion which starts at 0, then for every $h > 0$ let us focus on $B(0), B(h), B(2h), \dots$ and so

$$B(nh) = \sum_{k=1}^{n} \big(B(kh) - B((k-1)h)\big)$$

which is the sum of independent normal distributions $\mathcal{N}(0,h)$, and so $\{B(nh)\}_{n=0}^{\infty}$ is a general random process whose steps distribute with $\mathcal{N}(0,h)$. This has the following properties:

(1) $\limsup_{t\to\infty} B(t) = \infty$ and $\liminf_{t\to\infty} B(t) = -\infty$ with probability 1. This is since for every $h > 0$, $B(nh)$ is a general random process and so either for every n it is equal to 0, or $B(nh) \to \pm\infty$ or $\limsup B(nh) = \infty$ and $\liminf B(nh) = -\infty$. The first is not true since then $B(kh) - B((k-1)h) = 0$ with probability 1, which is not true. The second and third can be ignored by symmetry.

(2) $B(t)$ is transient: it visits every open interval an infinite number of times. This is due to the above point, as since $B(t)$ is continuous if there is an open interval (a,b) which it visits a finite number of times, then after the final time it must always be above b or below a, and so $\liminf B(t) \geq b$ or $\limsup B(t) \leq a$ in contradiction.

(3) $\limsup \frac{|B(t)|}{\sqrt{t}} \geq 1$ almost surely.

(4) $B(t) \sim \mathcal{N}(0,t)$ since $B(t) - B(0) \sim \mathcal{N}(0,t)$ and $B(0) = 0$.

(5) $\mathrm{Cov}(B(t), B(s)) = \min\{t,s\}$. Suppose $s \leq t$ then $B(t) - B(s) \sim \mathcal{N}(0, t-s)$ and $B(t) \sim \mathcal{N}(0,t)$ and $B(s) \sim \mathcal{N}(0,s)$. This means that $\mathbb{E}[B(t)] = \mathbb{E}[B(s)] = 0$ so $t = \mathrm{Var}(B(t)) = \mathbb{E}\big[B(t)^2\big]$ and similarly $s = \mathbb{E}\big[B(s)^2\big]$ and $t - s = \mathbb{E}\big[(B(t) - B(s))^2\big]$. Thus

$$\begin{aligned}
\mathrm{Cov}(B(t), B(s)) = \mathbb{E}[B(t)B(s)] &= \mathbb{E}\left[-\frac{(B(t) - B(s))^2 - B(t)^2 - B(s)^2}{2}\right] \\
&= -\frac{\mathbb{E}\big[(B(t) - B(s))^2\big] - \mathbb{E}\big[B(t)^2\big] - \mathbb{E}\big[B(s)^2\big]}{2} = -\frac{(t-s) - t - s}{2} = s
\end{aligned}$$

as required.

Now suppose $B(t)$ is Brownian motion, then we generally study it by studying the marginal distributions (the distribution functions) of every finite sampling of $B(t)$, ie. the distribution of $\big(B(t_1),\dots,B(t_n)\big)$ for every $0 \leq t_1 < \cdots < t_n$. This is essentially what is dictated in the first three conditions of Brownian motion, but the fourth condition on continuity cannot be proven by the study of these distributions. This is since if $U \sim \mathrm{Unif}([0,1])$ then defining

$$\tilde{B}(t) = \begin{cases} B(t) & t \neq U \\ 0 & t = U \end{cases}$$

gives us a collection of random variables $\tilde{B}(t)$ which have the same marginal distributions as $B(t)$ but is almost surely not continuous.

Now recall the following properties of normal distributions:

$$Z \sim \mathcal{N}(\mu,\sigma^2) \implies \mathbb{E}[Z] = \mu,\ \mathrm{Var}(Z) = \sigma^2$$
$$Z_1 \sim \mathcal{N}(\mu_1,\sigma_1^2),\ Z_2 \sim \mathcal{N}(\mu_2,\sigma_2^2) \text{ are independent} \implies Z_1 + Z_2 \sim \mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

Now for another, if $Z \sim \mathcal{N}(0,1)$ then then for every $0 < t$, we have the following series of inequalities:

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{t} - \frac{1}{t^3}\right)e^{-t^2/2} \leq \mathbb{P}(Z > t) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-t^2/2}$$

This is since

$$\begin{aligned}
\mathbb{P}(Z > t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}\,du \\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\left(t+\frac{v}{t}\right)^2/2}\cdot\frac{1}{t}\,dv \qquad (\text{substituting } u = t + \tfrac{v}{t}) \\
&= \frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t}\int_0^\infty e^{-v-\frac{v^2}{2t^2}}\,dv
\end{aligned}$$

Since $1 - x \leq e^{-x} \leq 1$ for $x > 0$ so we have $1 - \frac{v^2}{2t^2} \leq e^{-\frac{v^2}{2t^2}} \leq 1$ and so

$$\int_0^\infty e^{-v-\frac{v^2}{2t^2}}\,dv \leq \int_0^\infty e^{-v}\,dv = 1$$
$$\int_0^\infty e^{-v-\frac{v^2}{2t^2}}\,dv \geq \int_0^\infty e^{-v} - \frac{v^2}{2t^2}e^{-v}\,dv = 1 - \frac{1}{2t^2}\int_0^\infty v^2e^{-v}\,dv = 1 - \frac{1}{t^2}$$

which finishes the proof.

Now, our samplings are of the form $\big(B(t_1),\dots,B(t_n)\big) \in \mathbb{R}^n$ so we have to now understand vectors of normal distributions: multi-normal vectors (also known as Gaussian vectors).

3.0.3 Definition

If $z_1,\dots,z_n$ are all independent and have the distribution $\mathcal{N}(0,1)$ then the vector $Z = (z_1,\dots,z_n)$ has the **standard normal distribution in** $\mathbb{R}^n$. This is denoted $Z \sim \mathcal{N}_n(0,I)$. And a vector of random variables $X = (x_1,\dots,x_n)$ is called **Gaussian** (or multi-normal) if there exists a matrix $A \in M_{n\times m}(\mathbb{R})$ and a vector $\mu \in \mathbb{R}^n$ such that

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \overset{d}{=} A\begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} + \mu$$

where $Z \sim \mathcal{N}_m(0,I)$. A is called the **transition matrix** and μ the **expected value vector**. The **covariance matrix** of X is defined to be $\Sigma_{ij} = \mathrm{Cov}(x_i,x_j)$.

This means that

$$x_i = \sum_{j=1}^m A_{ij}z_j + \mu_i$$

and since $z_j \sim \mathcal{N}(0,1)$ are independent, this means $x_i \sim \mathcal{N}\Big(\mu_i, \sum_{j=1}^m A_{ij}^2\Big)$. Thus our definition of Gaussian vectors are equivalent to just having a vector of normal random variables.

Notice then that $((X-\mu)\cdot(X-\mu)^\top)_{ij} = (x_i-\mu_i)(x_j-\mu_j)$ and $\mathbb{E}[x_i] = \mu_i$ (since x_i is some linear combination of z_js and μ_i, and $\mathbb{E}[z_j] = 0$), so

$$\Sigma_{ij} = \mathbb{E}[(x_i-\mu_i)(x_j-\mu_j)] = \mathbb{E}\big[(X-\mu)(X-\mu)^\top\big]_{ij} = \mathbb{E}\big[AZZ^\top A^\top\big] = A\,\mathbb{E}\big[ZZ^\top\big]A^\top = AA^\top$$

the final equality is since $\mathbb{E}[ZZ^\top]_{ij} = \mathrm{Cov}(z_i, z_j) = \delta_{ij}$ so $\mathbb{E}[ZZ^\top] = I$. Meaning that

$$\Sigma = AA^\top$$

Now notice that $AA^\top$ is invertible if and only if $A^\top$ has full column rank, meaning A has full row rank. If it does not have full column rank then there exists an x such that $A^\top x = 0$ and so $\Sigma x = 0$ for $x \neq 0$ so Σ is not invertible. And if Σ is not invertible then there exists an $x \neq 0$ such that $AA^\top x = 0$ and so $A^\top x$ is in A's nullspace. But the nullspace of A and $A^\top$'s range are orthogonal complements and so $A^\top x = 0$ meaning $A^\top$ does not have full column rank.

Thus if Σ is invertible, then A has full row rank, and so we can assume that it is invertible.

3.0.4 Proposition

If Σ is invertible then X has a density

$$f_\Sigma(x) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det \Sigma}} e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2}$$

Since $\Sigma = AA^\top$ and $Z \sim \mathcal{N}_n(0, I)$ are independent (let the densities of its coefficients be f_i), then

$$f_I(z) = \prod_{i=1}^n f_i(z_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{\sqrt{2\pi}^n} e^{-\sum_{i=1}^n z_i^2/2} = \frac{1}{\sqrt{2\pi}^n} e^{-z^\top z}$$

Now if $x = Az + \mu$ then $z = A^{-1}(x - \mu)$ and so $dz = \det(A^{-1})\, dx$ by the Jacobian. And $\det(A^{-1}) = \frac{1}{\det A} = \frac{1}{\sqrt{\det \Sigma}}$, so for every event E,

$$\mathbb{P}(Z \in E) = \int_E f_I(z)\, dz = \int_{E' = AE+\mu} f_I(A^{-1}(x-\mu)) \frac{1}{\det A}\, dx = \int_{E'} \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det \Sigma}} e^{-(x-\mu)^\top (A^{-1})^\top A^{-1}(x-\mu)/2}\, dx$$

Since $(A^{-1})^\top A^{-1} = \Sigma^{-1}$ we get that

$$\mathbb{P}(X \in E') = \mathbb{P}(Z \in E) = \int_{E'} \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det \Sigma}} e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2}\, dx$$

as required. ■

Notice that a Gaussian vector may have multiple transition matrices (for example if $Z \sim \mathcal{N}_n(0, I)$ and P is orthogonal then $PZ \overset{d}{=} Z$). So we denote the distribution of X by its covariance matrix and expected variable vector, which are unique to X:

$$X \sim \mathcal{N}(\mu, \Sigma)$$

Here are some properties of Gaussian vectors:

(1) If $X \sim \mathcal{N}(\mu, \Sigma)$ then $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$ (shown/by definition),

(2) If $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$ then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$. This results directly from the linearity of expected values and covariance.

(3) If X is Gaussian and B is a matrix, then BX is Gaussian (since $BX = BAZ + B\mu$),

(4) Conditioning a Gaussian vector on the value of some of its coordinates, or their values on a linear combination of coordinates, is still Gaussian,

(5) If $X \sim \mathcal{N}(\mu, \Sigma)$ then there exists an upper (or lower) triangle matrix U which serves as its transition matrix.

Some more important notes: if two Gaussian vectors have the same covariance matrix and the same expected value vector, then their joint probabilities are the same (since joint probabilities are dependent only on Σ and μ). And if $(X_1, \dots, X_n)$ is a Gaussian vector such that the coefficients are pairwise uncorrelated, then Σ is a matrix which is zero except on the diagonal. Thus $X_i = \alpha_i Z_i + \mu_i$ where Z_i are independent normal distributions, meaning X_i are independent. So in order to show that a Gaussian vector's coefficients are independent, it is sufficient to show that they are pairwise uncorrelated.

3.1 Wiener Process

We will construct Brownian motion as the limit of continuous random functions on $[0, 1]$. We will then continue this construction onto the intervals $\{[n, n+1]\}_{n\in\mathbb{Z}}$, but for now we focus on $[0, 1]$. Let us define

$$D = \bigcup_{n=0}^\infty D_n, \qquad D_n = \left\{ \frac{k}{2^n} \,\middle|\, 0 \leq k \leq 2^n \right\}$$

D is a dense countable subset of $[0,1]$. Then let $\{Z_t\}_{t\in D}$ be a set of independent random variables which distribute $\mathcal{N}(0,1)$. Then let us define $B(0)=0$ and $B(1)=Z_1$ and for every $d\in D_n\setminus D_{n-1}$,

$$B(d)=\frac{B(d-2^{-n})+B(d+2^{-n})}{2}+\frac{Z_d}{\sqrt{2^{n+1}}}$$

This is since $d=\frac{k}{2^n}$ for some odd k, and so $d+2^{-n}=\frac{k+1}{2^n}$ which is of the form $\frac{k'}{2^{n-1}}$ since $k+1$ is even, so $d\pm 2^{-n}\in D_{n-1}$ so this is inductive definition is well-defined. Then $B(d)$ is continuous on D and can therefore be uniquely extended (since D is dense) to a continuous function on all of $[0,1]$. We will show inductively that

(1) For every $r<s<t$ in D_n, $B(t)-B(s)$ and $B(s)-B(r)$ are independent, and $B(t)-B(s)\sim\mathcal{N}(0,t-s)$.

(2) The set $\{B(d)\mid d\in D_n\}$ is independent of $\{Z_t\mid t\in D\setminus D_n\}$ (this is obvious from the construction).

For $n=0$ this is true trivially. Let $d\in D_n\setminus D_{n-1}$, then by the inductive assumption

$$X=\frac{1}{2}\left(B\left(d+\frac{1}{2^n}\right)-B\left(d-\frac{1}{2^n}\right)\right)\sim\mathcal{N}\left(0,\frac{1}{2^{n+1}}\right)$$

and this is independent of $\{Z_t\mid t\in D\setminus D_{n-1}\}$ and in particular Z_d. Similarly

$$Y=\frac{1}{\sqrt{2^{n+1}}}Z_d\sim\mathcal{N}\left(0,\frac{1}{2^{n+1}}\right)$$

And so $X+Y$ and $X-Y$ are independent and distribute $\mathcal{N}(0,2^{-n})$. But

$$X+Y=B(d)-B\left(d-\frac{1}{2^n}\right),\qquad X-Y=B\left(d+\frac{1}{2^n}\right)-B(d)$$

And so $\{B(d)-B(d-2^{-n})\}_{0\neq d\in D_n}$ are independent (since for Gaussian vectors, pairwise independence implies independence).

For every $d\in D_n$ let us define $G_n(d)=B(d)$, and G_n is linear between points in D_n (so it is continuous).

3.1.1 Lemma

Let $\sqrt{2\log 2}<c$ then

$$\mathbb{P}\big((\exists N)(\forall n\geq N)(\forall d\in D_n)\,|Z_d|<c\sqrt{n}\big)=1$$

Let us define

$$A_n=\big\{(\forall d\in D_n)\,|Z_d|<c\sqrt{n}\big\}\implies A_n^c=\big\{(\exists d\in D_n)\,|Z_d|\geq c\sqrt{n}\big\}$$

Then

$$\sum_{n=1}^{\infty}\mathbb{P}(A_n^c)=\sum_{n=1}^{\infty}\mathbb{P}\big((\exists d\in D_n)\,|Z_d|\geq c\sqrt{n}\big)\leq\sum_{n=1}^{\infty}\sum_{d\in D_n}\mathbb{P}\big(|Z_d|\geq c\sqrt{n}\big)$$

Since $Z_n\sim\mathcal{N}(0,1)$, $\mathbb{P}(|Z_d|\geq c\sqrt{n})\leq\frac{2}{\sqrt{2\pi}}\frac{1}{c\sqrt{n}}e^{-c^2n/2}$ (we showed this before). So

$$\leq\sum_{n=1}^{\infty}(2^n+1)e^{-c^2n/2}$$

Since $c>\sqrt{2\log 2}$, $e^{-cn^2/2}\leq 2^{-2n}$, so this series converges. By the Borel-Cantelli Lemma this means that $\mathbb{P}(A_n^c\text{ i.o.})=0$ so $\mathbb{P}(A_n\text{ a.e.})=1$, which is precisely the probability we're trying to compute. ∎

And so

$$\sup_{t\in[0,1]}|G_n(t)-G_{n-1}(t)|\leq\sup_{d\in D_n}\frac{|Z_d|}{\sqrt{2^{n+1}}}\overset{as}{<}c\sqrt{\frac{n}{2^{n+1}}}$$

the first inequality is since the difference is bound by when $t\in D_n\setminus D_{n-1}$ in which case the difference is $\frac{|Z_d|}{\sqrt{2^{n+1}}}$ (since $G_{n-1}(t)=\frac{B(d-2^{-n})+B(d+2^{-n})}{2}$). The final inequality is due to the above lemma.

Finally we define

$$G_\infty(t)=\lim_{n\to\infty}G_n(t)=\sum_{n=1}^{\infty}\big(G_n(t)-G_{n-1}(t)\big)$$

The rightmost series does converge almost surely since it is bound by $\sum c\sqrt{\frac{n}{2^{n+1}}}$. This means that by the Weierstrass M-test, $\sum G_n-G_{n-1}$ converges uniformly to G_∞, and so G_∞ is a continuous function. We claim that G_∞ is indeed our Brownian motion (in $[0,1]$), so we now denote it by $B(t)=G_\infty(t)$.

Let $t_1<\cdots<t_n$ in $[0,1]$, then let $t_{1,k}<\cdots<t_{n,k}$ in D such that $t_i=\lim_{k\to\infty}t_{i,k}$. By the continuity of B, we have $B(t_{i+1})-B(t_i)=\lim_{k\to\infty}B(t_{i+1,k})-B(t_{i,k})$.

3.1.2 Proposition

If $\{X_n\}_{n=1}^\infty$ is a sequence of Gaussian vectors such that $\lim_{n\to\infty} X_n \overset{as}{=} X$ then if the limits $\mu = \lim \mathbb{E}[X_n]$ and $\Sigma = \lim \operatorname{Cov}(X_n)$ exist, then $X \sim \mathcal{N}(\mu, \Sigma)$.

Then since $B(t_{i+1,k}) - B(t_{i,k}) \sim \mathcal{N}(0, t_{i+1,k} - t_{i,k})$ we get by this above proposition, $B(t_{i+1}) - B(t_i) \sim \mathcal{N}(0, t_{i+1} - t_i)$. Now in order to continue B to $[n, n+1]$ for all $n \in \mathbb{N}$, we continue this construction (though $B(n)$ must not be redefined).

3.1.3 Definition

We provide an equivalent definition of Brownian motion: $\{B(t)\}_{t\geq 0}$ is Brownian motion starting at $x \in \mathbb{R}$ if

(1) $B(0) \overset{as}{=} x$,

(2) It is a Gaussian process: for every $0 \leq t_1 < \cdots < t_n$, $(B(t_1), \dots, B(t_n))$ is a Gaussian vector,

(3) For every t, s, $\mathbb{E}[B(t)] = 0$ and $\mathbb{E}[B(t)B(s)] = \min\{t, s\}$.

(4) $t \mapsto B(t)$ is continuous.

We have already shown that the first definition implies this one. Now suppose $B(t)$ satisfies this definition, then let $s \leq t \leq u$ then since $(B(s), B(t), B(u))$ is a Gaussian vector, so is $(B(u) - B(t), B(s))$ (as it is equal to the product of the original Gaussian vector and a matrix). Then by the minimum property

$$\operatorname{Cov}(B(u) - B(t), B(s)) = \mathbb{E}[(B(u) - B(t))B(s)] = \mathbb{E}[B(u)B(s)] - \mathbb{E}[B(t)B(s)] = s - s = 0$$

In a Gaussian vector, if two coordinates are uncoorelated then they are independent, and so $B(t) - B(s)$ and $B(u)$ are independent, so we have shown that differences are independent.
And let $t, h > 0$ then $(B(t), B(t+h))$ is Gaussian and therefore $B(t+h) - B(t)$ must have a normal distribution as well. Now

$$\mathbb{E}[B(t+h) - B(t)] = \mathbb{E}[B(t+h)] - \mathbb{E}[B(t)] = 0$$

and

$$\operatorname{Var}(B(t+h) - B(t)) = \mathbb{E}\big[B(t+h)^2\big] - 2\,\mathbb{E}[B(t+h)B(t)] + \mathbb{E}\big[B(t)^2\big] = t + h - 2t + t = h$$

so $B(t+h) - B(t) \sim \mathcal{N}(0, h)$ as required. So we have shown the equivalence of these two definitions.

3.1.4 Proposition (Scaling Invariance)

Suppose $\{B(t)\}_{t\geq 0}$ is Brownian motion (starting at 0) and $a > 0$, then $\big\{\frac{1}{a}B(a^2 t)\big\}_{t\geq 0}$ is also Brownian motion.

Let us define $X(t) = \frac{1}{a}B(a^2 t)$, then obviously $X(0) \overset{as}{=} 0$ and $X(t)$ is almost surely continuous. And

$$X(t+h) - X(t) = \frac{1}{a}\big(B(a^2(t+h)) - B(a^2 t)\big) \sim \frac{1}{a}\mathcal{N}(0, a^2 h) = \mathcal{N}(0, h)$$

And of course if $0 \leq t_0 < \cdots < t_n$ then $X(t_i) = \frac{1}{a}B(a^2 t_i)$ and since $\big\{B(a^2 t_{i+1}) - B(a^2 t_i)\big\}$ are independent, so is $\{X(t_{i+1}) - X(t_i)\}$. ■

Let $B(t)$ be Brownian motion, then for every $a < 0 < b$, let

$$T_{a,b} = \inf\{t \geq 0 \mid B(t) \in \{a, b\}\}$$

ie $T_{a,b} = \min\{T_a, T_b\}$. Then let us define $X(t) = aB(t/a^2)$ which by the above proposition is also Brownian motion, meaning if $B(t^2/a) = 1$ then $X(t) = a$ and if $B(t^2/a) = \frac{b}{a}$ then $X(t) = b$. This means that $T^X_{a,b} = a^2 T^B_{1,b/a}$. But $X \overset{d}{=} B$ so $T_{a,b} = a^2 T_{1,b/a}$, and so $\mathbb{E}[T_{a,b}] = a^2\,\mathbb{E}\big[T_{1,b/a}\big]$ and in particular $\mathbb{E}[T_{-b,b}] = b^2\,\mathbb{E}[T_{-1,1}]$. Thus the expected time it takes to leave the interval $(-b, b)$ takes b^2 more time than it takes to leave $(-1, 1)$. Due to this association we found, we have that

$$\mathbb{P}(B(T_{a,b}) = a) = \mathbb{P}\big(B(T_{1,b/a}) = 1\big)$$

3.1.5 Proposition (Time Inversion Invariance)

If $B(t)$ is Brownian motion, then so is $X(t) = tB(1/t)$ for $t > 0$ and $X(0) = 0$.

We will prove this using the second definition of Brownian motion. Obviously $X(0) \stackrel{as}{=} 0$, and $X(t) \sim \mathcal{N}(0,t)$ so $(X(t_1), \dots, X(t_n))$ is a Gaussian vector. And if $t \leq s$ then

$$\mathbb{E}[X(t)X(s)] = \mathbb{E}[tsB(1/t)B(1/s)] = ts \cdot \frac{1}{s} = t$$

as required. All that remains is to show that $X(t)$ is almost surely continuous. Obviously for $t \neq 0$ this is true since $B(t)$ is continuous, so we must only show that $X(t) \xrightarrow{t \to 0} 0$. Now,

$$\left\{\lim_{t\to 0} X(t) = 0\right\} = \bigcap_{n=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{q \in (0,1/m)} \left\{|X(q)| < \frac{1}{n}\right\}$$

Now let us enumerate the rationals in $(0, 1/m)$ by $q_1^m, q_2^m, \dots$ and so we can write this as

$$= \bigcap_{n=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} \left\{\left|X(q_1^m), \dots, |X(q_k^m)| < \frac{1}{n}\right|\right\}$$

So by the continuity of probability

$$\mathbb{P}\Big(\lim_{t\to 0} X(t) = 0\Big) = \lim_n \lim_m \lim_k \mathbb{P}\left(|X(q_1^m)|, \dots, |X(q_k^m)| < \frac{1}{n}\right)$$

Now since $\mathrm{Cov}(X(t), X(s)) = \mathrm{Cov}(B(t), B(s))$, every finite Gaussian vector $(X(t_1), \dots, X(t_n))$ has the same covariance matrix and thus joint distribution as $(B(t_1), \dots, B(t_n))$ (this fact is true in general for all two Brownian motions). And so

$$= \lim_n \lim_m \lim_k \mathbb{P}\left(|B(q_1^m)|, \dots, |B(q_k^m)| < \frac{1}{n}\right) = \mathbb{P}\Big(\lim_{t\to 0} B(t) = 0\Big) = 1$$

as required. ■

3.1.6 Theorem (Law of Large Numbers)

Suppose $B(t)$ is Brownian motion then $\frac{B(t)}{t} \xrightarrow{n\to\infty} 0$ almost surely.

As above, let $X(t) = tB(1/t)$ for $t > 0$ and $X(0) = 0$.

$$\lim_{t\to\infty} \frac{B(t)}{t} = \lim_{t\to\infty} X(1/t) = \lim_{t\to 0} X(t) = X(0) = 0$$

since $X(t)$ is continuous. ■

3.1.7 Proposition (Initialization Invariance)

If $B(t)$ is Brownian motion, for every $A > 0$, $X(t) = B(A+t) - B(A)$ is Brownian motion.

$X(0) = 0$ and it is also trivially almost surely continuous. And

$$X(t+h) - X(t) = B(A+t+h) - B(A+t) \sim \mathcal{N}(0,h)$$

And in general $X(t_n) - X(t_{n-1}) = B(t_n) - B(t_{n-1})$ differences are independent. ■

3.1.8 Proposition (Superposition of Brownian Motion)

If $B^1, \ldots, B^k$ are independent Brownian motion and $\sum_{j=1}^k \alpha_j^2 = 1$ then $X(t) = \sum_{j=1}^k \alpha_j B^j(t)$ is Brownian motion as well.

$X(0) \overset{as}{=} 0$ and $X(t)$ is also almost surely continuous.

$$X(t+h) - X(t) = \sum_{j=1}^k \alpha_j \big(B^j(t+h) - B^j(t)\big) \sim \sum_{j=1}^k \alpha^j \mathcal{N}(0,h) = \mathcal{N}\left(0, h\sum_{j=1}^k \alpha_j^2\right) = \mathcal{N}(0,h)$$

And since

$$\begin{aligned}\operatorname{Cov}(X(a)-X(b), X(c)-X(d)) &= \sum_{j=1}^k \sum_{i=1}^k \alpha_j\alpha_i \operatorname{Cov}\big(B^j(a)-B^j(b), B^i(c)-B^i(d)\big) \\ &= \sum_{j=1}^k \alpha_j^2 \operatorname{Cov}\big(B^j(a)-B^j(b), B^j(c)-B^j(d)\big) = 0\end{aligned}$$

Since for $i \neq j$, B^j and B^i are independent. Thus $X(t_n) - X(t_{n-1}), \ldots, X(t_1) - X(t_0)$ are independent. ■

We know by the definition of Brownian motion that it is (almost surely) continuous. But we can state a stronger result:

3.1.9 Theorem

If $B(t)$ is Brownian motion then there exists a constant $c > 0$ such that

$$\mathbb{P}\left(\limsup_{h\to 0} \sup_{0\le t\le 1-h} \frac{|B(t+h)-B(t)|}{\sqrt{h\log(1/h)}} \le c\right) = 1$$

Meaning that almost surely, for every $h > 0$ small enough and every $0 \le t \le 1-h$, $|B(t+h) - B(t)| \le c\sqrt{h|\log h|}$.

Recall that

$$B(t) = \sum_{n=1}^\infty \big(G_n(t) - G_{n-1}(t)\big) = \sum_{n=1}^\infty F_n(t)$$

where F_n is linear on a finite number of intervals, and we also showed that $\|F_n\|_\infty \le c\sqrt{\frac{n}{2^{n+1}}}$ for every $c > \sqrt{2\log 2}$ and for n large enough ($n > N$). F_n is linear except for on D_n, and so F_n' exists for every point not in D_n, and so

$$\sup_{t\in[0,1]\setminus D_n} |F_n'(t)| \le \frac{2\sup|F_n|}{2^{-n}} \le 2c\sqrt{n}2^{n/2}$$

And so for $\ell > N$,

$$\begin{aligned}|B(t+h) - B(t)| &\le \sum_{n=1}^\infty |F_n(t+h) - F_n(t)| \le \sum_{n=1}^\ell h\sup|F_n'| + \sum_{n=\ell+1}^\infty 2\sup|F_n| \\ &\le h\sum_{n=1}^N |F_n'|_\infty + 2ch\sum_{n=N}^\ell \sqrt{n}2^{n/2} + 2c\sum_{n=\ell+1}^\infty \sqrt{n}2^{-n/2}\end{aligned}$$

If $h > 0$ is small enough then $h\sum_{n=1}^N |F_n|_\infty < \sqrt{h\log(1/h)}$ so let us choose an ℓ where $2^{-\ell} < h < 2^{-\ell+1}$, then the second and third terms above are also bound by $\sqrt{h\log(1/h)}$. ■

This theorem gives an upper bound for how fast the differences of $B(t)$ tend to zero, and the following gives a lower bound:

3.1.10 Theorem

There exists a $c < \sqrt{2}$ such that for every $\varepsilon > 0$ there exists an $h \in (0,\varepsilon)$ such that for every $t \in [0, 1-h]$: $|B(t+h) - B(t)| \ge c\sqrt{h\log(1/h)}$ almost surely. Meaning

$$\mathbb{P}\Big((\forall\varepsilon > 0)(\exists h \in (0,\varepsilon))(\forall t \in [0,1-h])\, |B(t+h) - B(t)| \ge c\sqrt{h\log(1/h)}\Big) = 1$$

I'm sorry but I don't really feel like typing this out right now. Pretend like I did, I'm sure the proof isn't really that important anyway. ■

3.1.11 Definition

f is α-**Hölder continuous** if there exists a constant c such that $|f(x) - f(y)| \leq c|x-y|^\alpha$.

3.1.12 Corollary

Brownian motion is almost surely α-Hölder continuous for every $0 < \alpha < \frac{1}{2}$.

Notice that for small enough h, $\sqrt{h\log(1/h)} \leq h^\alpha$ for $\alpha < \frac{1}{2}$, so this is a direct result of the above theorem (the one before last). ■

We showed that almost surely $\frac{B(t)}{t} \xrightarrow{t\to\infty} 0$, and in a similar vein we prove

3.1.13 Proposition

Almost surely,

$$\limsup_{t\to\infty} \frac{B(t)}{\sqrt{t}} = \infty, \qquad \liminf_{t\to\infty} \frac{B(t)}{\sqrt{t}} = -\infty$$

Let $c > 0$, then let us define

$$A_n = \left\{B(n) > c\sqrt{n}\right\}$$

and we know $B(n) = \sum_{i=1}^n X_i$ where $X_i = B(i) - B(i-1) \sim \mathcal{N}(0,1)$ are independent. By Hewitt-Savage, since $\{A_n \text{ i.o.}\}$ is preserved under finite permutations of the indexes of X_i, we have that $\mathbb{P}(A_n \text{ i.o.}) \in \{0,1\}$. But on the other hand

$$\mathbb{P}\big(B(n) > c\sqrt{n} \text{ i.o.}\big) \geq \limsup_{n\to\infty} \mathbb{P}\big(B(n) > c\sqrt{n}\big) = \limsup_{n\to\infty} \mathbb{P}(B(1) > c) > 0$$

the final equality is since $\frac{B(n)}{\sqrt{n}} \sim \mathcal{N}(0,1)$. So $\mathbb{P}(A_n \text{ i.o.}) = 1$, meaning almost surely we can choose a subsequence m_n such that $B(m_n) > c\sqrt{m_n}$, so $\limsup \frac{B(n)}{\sqrt{n}} > c$. Taking the countable intersection over all integers c, we get that $\limsup \frac{B(n)}{\sqrt{n}} > c$ for every integer c almost surely, meaning $\limsup \frac{B(n)}{\sqrt{n}} = \infty$ almost surely. Similar for $c < 0$ and the limit inferior. ■

3.1.14 Theorem

Let $0 \leq t_0$ then almost surely $B(t)$ is not differentiable at t_0. Furthermore almost surely

$$\limsup_{h\to 0} \frac{B(t_0+h) - B(t_0)}{h} = \infty, \qquad \liminf_{h\to 0} \frac{B(t_0+h) - B(t_0)}{h} = -\infty$$

Let $X(t) = tB(1/t)$ for $t > 0$ and $X(0) = 0$, this is Brownian motion as we showed. Then

$$\limsup_{h\to 0} \frac{X(h) - X(0)}{h} = \limsup_{h\to 0} \frac{hB(1/h)}{h} = \limsup_{t\to\infty} B(t) \geq \limsup_{t\to\infty} \frac{B(t)}{\sqrt{t}} = \infty$$

And similarly we have that

$$\liminf_{h\to 0} \frac{X(h) - X(0)}{h} = -\infty$$

So $X(t)$ is not differentiable at 0 almost surely, and so neither is $B(t)$ as they are both Brownian motion. And for $t_0 > 0$, define $Y(s) = B(t_0+s) - B(t_0)$ which is Brownian motion as well, and not differentiable at 0 which is equivalent to $B(t)$ not being differentiable at t_0. ■

Notice that this does not mean that $B(t)$ is almost surely nowhere differentiable, just that if we set any specific point $B(t)$ is almost surely not differentiable there. Though it does happen that $B(t)$ is almost surely nowhere differentiable:

3.1.15 Theorem (Paley, Wiener, Zygmund)

Almost surely, Brownian motion is nowhere differentiable.

The proof of this can be found in other literature on Brownian motion.

3.1.16 Theorem

$\{B(A+t)-B(A)\}_{t\geq 0}$ is Brownian motion independent of the history $\{B(s)\}_{0\leq s\leq A}$.

We have already shown that $B(A+t)-B(A)$ is Brownian motion, and it is independent of its history since differences in Brownian motion are independent. ■

So in a way, Brownian motion is a homogeneus Markov process. It is Markov in the way that conditioning under movement until time t is equivalent to conditioning under movement at time t. And it is homogeneus meaning that starting measuring at time A is equivalent to measuring at time 0.

3.1.17 Theorem

$\{B(A+t)-B(A)\}_{t\geq 0}$ is independent of $\mathcal{F}_A^+ = \bigcap_{\varepsilon>0}\sigma\big(\{B(s)\}_{0\leq s\leq A+\varepsilon}\big)$.

By the continuity of Brownian motion, almost surely $B(A+t)-B(A) = \lim_{n\to\infty} B(A_n+t)-B(A_n)$ where $A < A_n \to A$. And so for every finite number of measurements $t_1 < \cdots < t_m$,

$$\big(B(A+t_1)-B(A),\cdots,B(A+t_m)-B(A)\big) = \lim_{n\to\infty}\big(B(A_n+t_1)-B(A_n),\ldots,B(A_n+t_m)-B(A_n)\big)$$

the right hand side is the limit of a Gaussian vector which is dependent only on $B(t)$ for $t > A$, thus for small enough $\varepsilon > 0$, it is independent of $\sigma\big(\{B(s)\}_{0\leq s\leq A+\varepsilon}\big)$. In particular it is independent of $\mathcal{F}_A^+$. Since Gaussian vectors converge to a Gaussian vector whose covariance matrix is the limit of the covariance matrices, we can conclude that $(B(A+t_1)-B(A),\ldots,B(A+t_m)-B(A))$ is independent of $\mathcal{F}_A^+$ as required. ■

Notice that $\mathcal{F}_A^+$ is the sigma algebra of all events which are dependent only on $B(s)$ for $s\leq A$ and a measurement an infinitesimal amount of time after A.

3.1.18 Theorem (Blumenthal's Zero-One Law)

For every $G\in\mathcal{F}_0^+$ and every initial measurement $B(0)=x$,

$$\mathbb{P}(G \mid B(0)=x) = \mathbb{P}_x(G)\in\{0,1\}$$

By the above theorem, for $A=0$, $\{B(t)\}_{t\geq 0}$ is independent of $\mathcal{F}_0^+$ and this means that every event in $\mathcal{F}_0^+$ is independent of itself (since it is in $\sigma\big(\{B(t)\}_{t\geq 0}\big)$), and so must have trivial probability. ■

3.1.19 Theorem

Let us define $\tau = \inf\{t>0 \mid B(t)>0\}$ and $\sigma = \inf\{t>0\mid B(t)=0\}$ then $\mathbb{P}_0(\tau=0)=\mathbb{P}_0(\sigma=0)=1$.

We know that

$$\{\tau=0\} = \bigcap_{n=1}^{\infty}\left\{\left(\exists\delta<\frac{1}{n}\right)B(\delta)>0\right\}\in\bigcap_{\varepsilon>0}\sigma\big(\{B(s)\}_{0\leq s\leq\varepsilon}\big) = \mathcal{F}_0^+$$

so by Blumenthal's zero-one law, $\mathbb{P}(\tau=0)\in\{0,1\}$ so it is sufficient to show that $\tau=0$ has nonzero probability. Let $t>0$,

$$\mathbb{P}_0(\tau\leq t)\geq\mathbb{P}_0(B(t)>0)=\frac{1}{2}$$

and by the continuity of probability

$$\mathbb{P}_0(\tau=0)=\lim_{t\to 0^+}\mathbb{P}_{\tau\leq t}(\geq)\frac{1}{2}$$

as required. Analogously we can show that $\mathbb{P}_0(\inf\{t>0\mid B(t)<0\}=0)=1$. And so by the mean value theorem we get that $\sigma=0$ almost surely, since for every $\delta>0$ we have that there exists a $t_1<\delta$ and $t_2<\delta$ such that $B(t_1)>0$ and $B(t_2)<0$, so there must exist a $t<\delta$ where $B(t)=0$. And so for every $\delta>0$, $\mathbb{P}_0(\sigma<t)=1$, and taking this limit as $t\to 0^+$ gives $\mathbb{P}_0(\sigma=0)=1$ by the continuity of probability. ■

3.1.20 Lemma

For every $a_1 < b_1 \leq a_2 < b_2$, then let $m_i = \max_{[a_i,b_i]} B(t)$ for $i = 1, 2$. Then $\mathbb{P}(m_1 = m_2) = 0$.

By the volatility of Brownian motion, there exists a t in every neighborhood of 0 such that $B(t) > 0$, now by the Markov property of Brownian motion we can generalize this to mean that for every $\varepsilon > 0$ there exists a $a_2 < t < a_2 + \varepsilon$ such that $B(t) > B(a_2)$. And so $\mathbb{P}(B(a_2) < m_2) = 1$. So there must exist an n such that $m_2 = \max_{[a_2+1/n,b_2]} B(t)$, and so if we swap a_2 with $a_2 + 1/n$ we can assume that $b_1 < a_2$. By the Markov property at b_1, $m_1 - B(b_1)$ is independent of $B(a_2) - B(b_1)$ and these are both independent of $m_2 - B(a_2)$. And so the event that $\{m_1 = m_2\}$ can be written as

$$B(a_2) - B(b_1) = \big(m_1 - B(b_1)\big) - \big(m_2 - B(a_2)\big)$$

so let $X = B(a_2) - B(b_1)$, $Y = m_1 - B(b_1)$ and $Z = m_2 - B(a_2)$. These are independent continuous random variables, which means that $\mathbb{P}(X = Y - Z) = 0$ (since $X - Y + Z$ is a continuous random variable, and so the probability it is equal to a singelton is zero). ■

3.1.21 Theorem

For Brownian motion $\{B(t) \mid 0 \leq t \leq 1\}$, almost surely

(1) every local maximum is a strong local maximum (meaning that there exists a neighborhood of the maximum in which it is *strictly* larger than all other points)

(2) the set of points in which there is a local maximum is a countably dense set,

(3) the global maximum in $[0, 1]$ is obtained at a single point.

(1) By the above lemma, there cannot exist a weak maximum, as two adjacent intervals almost surely have distinct maximums.

(2) For every $a < b$, almost surely $\max_{t\in[a,b]} B(t) \neq B(a), B(b)$. This is true in particular for rational $a < b$. And so almost surely for every rational $a < b$ there exists a local maximum in (a, b). And this means that the set of local maximums must be dense, as $\mathbb{Q}$ is.

(3) For every rational q, almost surely $\max_{[0,q]} B(t) \neq \max_{[q,1]} B(t)$, but if there were two global maxima then there would be a rational between them and this would not be the case. ■

3.1.22 Theorem (The Reflection Principle)

$\mathbb{P}\big(\max_{[0,t]} B \geq \alpha,\, B(t) \leq \alpha\big) = \mathbb{P}(B(t) \geq \alpha)$.

Let us define $\tau_\alpha = \inf\{t > 0 \mid B(t) = \alpha\}$. And so since $B(t)$ is almost surely continuous,

$$\mathbb{P}\left(\max_{[0,t]} B \geq \alpha, B(t) \leq \alpha\right) = \mathbb{P}(\tau_\alpha \leq t,\, B(t - \tau_\alpha) \leq 0)$$

since $B(t)$ has a normal distribution, this is equal to

$$= \mathbb{P}(\tau_\alpha \leq t,\, B(t - \tau_\alpha) \geq 0) = \mathbb{P}\left(\max_{[0,t]} B \geq \alpha,\, B(t) \geq \alpha\right) = \mathbb{P}(B(t) \geq \alpha)$$ ■

Furthermore, we get that

$$\mathbb{P}\left(\max_{[0,t]} B \geq \alpha\right) = \mathbb{P}\left(\max_{[0,t]} B \geq \alpha,\, B(t) \geq \alpha\right) + \mathbb{P}\left(\max_{[0,t]} B \geq \alpha, B(t) \leq \alpha\right) = 2\,\mathbb{P}(B(t) \geq \alpha)$$

and since $B(t) \sim \mathcal{N}(0, t)$ we get that

$$\mathbb{P}\left(\max_{[0,t]} B \geq \alpha\right) = 2\int_\alpha^\infty e^{-\frac{x^2}{2t}} \frac{dx}{\sqrt{2\pi t}} = 2\int_{\frac{\alpha}{\sqrt{t}}}^\infty e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}$$

And if we let $T_a = \inf\{t > 0 \mid B(t) = a\}$ then we get that

$$\mathbb{P}(T_a \le t) = \mathbb{P}\big(\sup_{[0,t]} B \ge a\big) = 2\,\mathbb{P}(B(t) \ge a) = 2\left(1 - \Phi\left(\frac{a}{\sqrt{t}}\right)\right) = 2\int_{a/\sqrt{t}}^{\infty} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}$$

and so we get that the probability density of T_a is, by differentiating the above expression,

$$f_{T_a}(t) = \frac{a}{\sqrt{2\pi}} t^{-3/2} e^{-a^2/2t}$$

3.2 Universal Scaling Limit of Random Walks

Let us return to a general random process, $S_n = X_1 + \cdots + X_n$ where $\{X_i\}$ are independent equivalently-distributed random variables where $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1$. In Einstein's physical model, a particle impact with other particles and each impact displaces it slightly. Suppose within one unit of time there are N impacts, and each impact displaces the particle by $\varepsilon \cdot X_k$. Then the placement after t units of time is

$$B_t^N = \varepsilon X_1 + \cdots + \varepsilon X_{tN}$$

Then we have that

$$\sigma^2 = \mathrm{Var}\big(B_1^N\big) = \varepsilon^2 N\,\mathrm{Var}(X_1) = \varepsilon^2 N$$

This means that $\varepsilon = \frac{\sigma}{\sqrt{N}}$ and so in the case that $\sigma = 1$,

$$B_t^N = \frac{\sigma}{\sqrt{N}}(X_1 + \cdots + X_{tN}) = \frac{1}{\sqrt{N}}(X_1 + \cdots + X_{tN})$$

So by the central limit theorem, as $N \to \infty$, $B_t^N \xrightarrow{d} B(t) \sim \mathcal{N}(0,t)$. Now notice that if $t \le s$,

$$\mathrm{Cov}\big(B_t^N, B_s^N\big) = \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{tN} X_j \sum_{i=1}^{sN} X_i\right] = \frac{1}{N}\sum_{j=1}^{tN}\sum_{i=1}^{sN} \mathrm{Cov}(X_i, X_j)$$

for $i \ne j$ the covariance is zero, and when $i = j$ it is 1, thus we get that this is equal to t. So we have that

$$\big(B_{t_1}^N, \ldots, B_{t_k}^N\big) \xrightarrow{d} \big(B_{t_1}, \ldots, B_{t_k}\big) \sim \mathcal{N}(0, \Sigma), \qquad \Sigma_{ij} = \min\{i, j\}$$

$\{B_t\}$ is indeed Brownian motion, as by the second definition we have proven all but almost sure continuity. This is true since B_t^N and $B_{t+\delta}^N$ have a difference of

$$\big|B_t^N - B_{t+\delta}^N\big| \stackrel{d}{=} \lim_N \left|\frac{1}{\sqrt{N}}\sum_{i=0}^{\delta t} X_i\right|$$

which can be made to be arbitrarily small.

4 Galton-Watson Processes

Suppose we had a petri dish containing a bacterial colony, and at constant intervals of time the bacteria reproduce randomly, in a way which is independent of the current time and of other bacteria. We can model this mathematically as follows: let $\{\xi_{i,n}\}_{i,n\in\mathbb{N}}$ be independent random variables which all have equivalent distributions, where $\xi_{i,n}$ represents the number of bacteria the ith bacterium reproduced into on the nth measurement. We denote the distribution vector of $\xi_{i,n}$ by $\bar{B} = (B_0, B_1, \ldots)$ (meaning $\mathbb{P}(\xi_{i,n} = k) = B_k$). If we denote X_n to be the number of bacteria on the petri dish at time n, then

$$X_n \overset{d}{=} \sum_{i=1}^{X_{n-1}} \xi_{i,n-1}$$

Let us notice that $\{X_n\}$ is a Markov chain; $(X_n \,|\, X_{n-1}, \ldots, X_0) \overset{d}{=} (X_n \,|\, X_{n-1})$. And at any point if $X_N = 0$, then $X_n = 0$ for every $n \geq N$, meaning 0 is an absorbing state. 0 is also the only absorbing state, unless $\mathbb{P}(\xi = 1) = 1$.

A central question to the study of such a process is the probability that the colony survives forever. We of course may also be interested in computing the expected values and variances of X_n.

4.0.1 Lemma

Let $\{X_j\}$ be a set of independent, equivalently-distributed random variables and let Y be a random variable whose support is in $\mathbb{N}_0$ which is independent of $\{X_j\}$. Then

$$\mathbb{E}\left[\sum_{j=1}^{Y} X_j\right] = \mathbb{E}[Y]\,\mathbb{E}[X_1]$$

By the law of total expectation,

$$\mathbb{E}\left[\sum_{j=1}^{Y} X_j\right] = \sum_{n=1}^{\infty} \mathbb{P}(Y = n)\,\mathbb{E}\left[\sum_{j=1}^{n} X_j\right] = \mathbb{E}[X_1]\sum_{n=1}^{\infty} n\,\mathbb{P}(Y = n) = \mathbb{E}[Y]\,\mathbb{E}[X_1] \qquad \blacksquare$$

Thus we have that

$$\mathbb{E}[X_n] = \mathbb{E}\left[\sum_{j=1}^{X_{n-1}} \xi_{j,n-1}\right] = \mathbb{E}[X_{n-1}]\,\mathbb{E}[\xi]$$

let us denote $\mu := \mathbb{E}[\xi]$, which is equal to $\sum_{n=0}^{\infty} nB_n$, and so we have that

$$\mathbb{E}[X_n] = \mu\,\mathbb{E}[X_{n-1}] \implies \mathbb{E}[X_n] = \mu^n\,\mathbb{E}[X_0]$$

We will assume that $X_0 \overset{as}{=} 1$, and so $\mathbb{E}[X_n] = \mu^n$. Similarly let us define $\sigma^2 = \mathrm{Var}(\xi) = \sum n^2 B_n - (\sum nB_n)^2$. Then

$$\begin{aligned}\mathrm{Var}(X_n) = \mathbb{E}\big[(X_n - \mu^n)^2\big] &= \mathbb{E}\big[(X_n - \mu X_{n-1} + \mu X_{n-1} - \mu^n)^2\big] \\ &= \mathbb{E}\big[(X_n - \mu X_{n-1})^2\big] + 2\,\mathbb{E}\big[(X_n - \mu X_{n-1})\mu(X_{n-1} - \mu^{n-1})\big] + \mu^2\,\mathbb{E}\big[(X_{n-1} - \mu^{n-1})^2\big]\end{aligned}$$

Using the law of recurring expectation,

$$\begin{aligned}\mathbb{E}\big[(X_n - \mu X_{n-1})\mu(X_{n-1} - \mu^{n-1})\big] &= \mathbb{E}\big[\mathbb{E}\big[(X_n - \mu X_{n-1})\mu(X_{n-1} - \mu^{n-1}) \;\big|\; X_{n-1}\big]\big] \\ &= \mathbb{E}\big[\mu(X_{n-1} - \mu^{n-1})\big(\mathbb{E}[X_n \mid X_{n-1}] - \mu X_{n-1}\big)\big]\end{aligned}$$

Notice that

$$\mathbb{E}[X_n \mid X_{n-1}] = \mathbb{E}\left[\sum_{j=1}^{X_{n-1}} \xi_{j,n-1}\right] = \mu X_{n-1}$$

so this is equal to zero. Thus we have that

$$\mathrm{Var}(X_n) = \mathbb{E}\big[(X_n - \mu X_{n-1})^2\big] + \mu^2\,\mathrm{Var}(X_{n-1})$$

Now again using the law of total expectation,

$$\begin{aligned}
\mathbb{E}\big[(X_n - \mu X_{n-1})^2\big] &= \sum_{N=0}^{\infty} \mathbb{E}\big[(X_n - \mu X_{n-1})^2 \mid X_{n-1} = N\big]\, \mathbb{P}(X_{n-1} = N) \\
&= \sum_{N=0}^{\infty} \mathbb{E}\left[\left(\sum_{j=1}^{N} (\xi_{j,N-1} - \mu)\right)^2\right] \mathbb{P}(X_{n-1} = N) \\
&= \sum_{N=0}^{\infty} \sum_{j=1}^{N} \mathrm{Var}(\xi_{j,N-1})\, \mathbb{P}(X_{n-1} = N) \\
&= \sigma^2 \sum_{N=0}^{\infty} N\, \mathbb{P}(X_{n-1} = N) = \sigma^2\, \mathbb{E}[X_{n-1}] = \sigma^2 \mu^{n-1}
\end{aligned}$$

Thus we have the following linear recurrence

$$\mathrm{Var}(X_n) = \sigma^2 \mu^{n-1} + \mu^2\, \mathrm{Var}(X_{n-1})$$

Solving it yields

$$\mathrm{Var}(X_n) = \begin{cases} \sigma^2 \mu^{n-1} \dfrac{1-\mu^n}{1-\mu} & \mu \neq 1 \\ n\sigma^2 \mu^{n-1} & \mu = 1 \end{cases}$$

4.1 Extinction Time

Similar as with Markov chains and Brownian motion, let us define

$$T_0 = \min\{t \in \mathbb{N} \mid X_t = 0\}$$

As stated before, a big focus of the study of Galton-Watson processes is computing the probability that $T_0 = \infty$. Obviously if $B_0 \neq 0$ then $\mathbb{P}(T_0 < \infty) > 0$, as it is possible for $X_1 = 0$ for example. Furthermore, if $\mu < 1$ then $\mathbb{E}[X_n] = \mu^n \to 0$ and so $\mathbb{P}(T_0 < \infty) = 1$. This is due to the Markov inequality: $\mathbb{P}(X_n = 0) = 1 - \mathbb{P}(X_n \geq 1) \geq 1 - \mathbb{E}[X_n]$, so $\mathbb{P}(X_n = 0) \to 1$ so $X_n \xrightarrow{as} 0$. Finally it is obvious that the whether or not the colony almost surely survives or goes extinct is independent on the initial state of the colony.
Let us define the probability generating function of B:

$$\varphi_\xi(s) = \varphi_B(s) = \sum_{k=0}^{\infty} B_k s^k$$

4.1.1 Theorem (The Galton-Watson Theorem)

$\mathbb{P}(T_0 < \infty \mid X_0 = 1) = \min_{s\in[0,1]}\{\varphi_B(s) = s\}$; meaning that the minimal fixed point of φ_B is the probability that the colony goes extinct. Furthermore, $\mathbb{P}(T_0 < \infty \mid X_0 = 1) = 1$ if and only if $\mu \leq 1$ and $B_0 > 0$.

We will prove this after investigating probability generating functions more.

4.1.2 Lemma

Suppose $\{X_j\}_{j=1}^{\infty}$ are independent, equivalently-distributed random variables taking on values in $\mathbb{N}_0$. Further let Y be another random variable taking on values in $\mathbb{N}_0$ independent of $\{X_j\}$. Let us define $S = \sum_{j=1}^{Y} X_j$, then

$$\varphi_S = \varphi_Y \circ \varphi_X$$

Let us define $S_k = \mathbb{P}(S = k)$ and so

$$\begin{aligned}
\varphi_S(t) = \sum_{k=0}^{\infty} S_k t^k &= \sum_{k=0}^{\infty} \sum_{\substack{n\in\mathbb{N} \\ \ell_1+\cdots+\ell_n=k}} \mathbb{P}(Y = n)\, \mathbb{P}(X_1 = \ell_1) \cdots \mathbb{P}(X_n = \ell_n) t^k \\
&= \sum_{\substack{n,k\in\mathbb{N}_0 \\ \ell_1+\cdots+\ell_n=k}} \mathbb{P}(Y = n)\, \mathbb{P}(X_1 = \ell_1) \cdots \mathbb{P}(X_n = \ell_n) t^k \\
&= \sum_{n=0}^{\infty} \mathbb{P}(Y = n) \left(\sum_{\ell=0}^{\infty} \mathbb{P}(X_1 = \ell) t^\ell\right)^n = \varphi_Y(\varphi_X(t))
\end{aligned}$$

as required. ■

Here are some traits of probability generating functions (we consider them as functions over the domain $[0,1]$):

(1) φ is nondecreasing: since if $s_1 < s_2$ then $s_1^t \le s_2^t$ we have that $\varphi(s_1) = \sum_t X_t s_1^t \le \sum_t X_t s_2^t = \varphi(s_2)$.

(2) φ is concave: this is immediate due to the fact that $(x+y)^n \ge x^n + y^n$.

(3) φ is smooth: for every $0 < \varepsilon < 1$, we have that $|X_t s^t| \le \varepsilon^t$ on $[0,\varepsilon]$ and so by the Weierstrass M-test, φ converges uniformly and since the partial sums are smooth, so to is φ on every $[0,\varepsilon]$ for $0 < \varepsilon < 1$. Smoothness at 1 can also be shown.

(4) $\varphi(0) = B_0$ and $\varphi(1) = \sum_{k=0}^{\infty} X_k = 1$.

(5) If $\varphi'(1) > 1$ then $\varphi(s) = s$ has a unique solution on $(0,1)$, and otherwise there is no solution. Let us define $f(s) = \varphi(s) - s$ then $f'(1) > 0$, meaning f is increasing at 1, and so 1 is a local maximum of f, so it is negative in a neighborhood of 1 since $f(1) = 0$. Since $f(0) = B_0 > 0$ this means that f must be zero somewhere.

Proof of The Galton-Watson Theorem:

As we proved, $\varphi_{X_n} = \varphi_{X_{n-1}} \circ \varphi_B$, and so by induction we have that $\varphi_{X_n} = \varphi_B \circ \cdots \circ \varphi_B$ (n times), under the assumption $X_0 = 1$. Thus

$$\mathbb{P}(X_n = 0 \mid X_0 = 1) = \varphi_{X_n}(0) = \varphi_B \circ \cdots \circ \varphi_B(0)$$

And since $\{T_0 < \infty\} = \bigcup \{X_n = 0\}$ which is an increasing sequence, we have that $\mathbb{P}(T_0 < \infty \mid X_0 = 1) = \lim_n \varphi_B^n(0)$. In particular we have that $\varphi_B(\mathbb{P}(T_0 < \infty \mid X_0 = 1)) = \mathbb{P}(T_0 < \infty \mid X_0 = 1)$, meaning $\mathbb{P}(T_0 < \infty \mid X_0 = 1)$ is a fixed point of φ_B. We assume that $B_0 > 0$ (as otherwise the probability of extinction is zero). Let us notice that since φ_B converges uniformly, $\mu = \varphi_B'(1)$ and so if $\mu \le 1$ then the only fixed point is 1, meaning $\mathbb{P}(T_0 < \infty \mid X_0 = 1) = 1$ as required.

Otherwise, there exists another fixed point $r \in (0,1)$. Since φ_B is non-decreasing,

$$B_0 = \mathbb{P}(X_1 = 0 \mid X_0 = 1) = \varphi_B(0) \le \varphi_B(r) = r$$

and again since φ_B is non-decreasing, $\varphi_B^2(0) \le \varphi_B(\varphi_B(r)) = r$ so inductively $\varphi_B^n(0) \le r$. And we have shown

$$0 < \mathbb{P}(T_0 < \infty \mid X_0 = 1) = \lim_n \varphi_B^n(0) \le r < 1$$

but since r is the unique fixed point in $(0,1)$, this means that $\mathbb{P}(T_0 < \infty \mid X_0 = 1) = r$ is the minimum fixed point. ■

Notice that if we begin with k bacteria, then we can view this as k independent colonies and so

$$\mathbb{P}(T_0 < \infty \mid X_0 = k) = \mathbb{P}(T_0 < \infty \mid X_0 = 1)^k = \min_{s \in [0,1]} \{\varphi_B(s) = s\}^k$$

And what is the probability the colony goes extinct within k generations? Let us define $u_k = \mathbb{P}(T_0 \le k \mid X_0 = 1)$, then using first step analysis,

$$u_k = \mathbb{P}(T_0 \le k \mid X_0 = 1) = \sum_{j=0}^{\infty} \mathbb{P}(T_0 \le k \mid X_1 = j) \cdot \mathbb{P}(X_1 = j \mid X_0 = 1) = \sum_{j=0}^{\infty} (u_{k-1})^j B_j$$

which could be solved if the B_js are known. Not that you'd know what a BJ is though.