Basics

To recap, any probability space is a tuple of a sample space, a collection of subsets of that sample space, and a function from the event space to $[0,1]$, where the event space and probability function each satisfy three natural conditions ($\mathcal{F}$ is a σ-algebra, plus the basic propeties of $\mathbb{P}$). We take random variables as functions from Ω, representing an observable. Formally they must also have that each $\{X(\omega) \le x\} \in \mathcal{F}$.

In prelims probability there was a distinction between discrete and continuous random variables. These do not cover every possible notion of a random variable, and so we ideally want to unify these definitions to a more abstract notion. Beginning in this way we define expectation axiomatically:

Definition 1 (Expectation)
- *$\mathbb{E}I_A = \mathbb{P}(A)$ for any event A.*
- *If $\mathbb{P}(X \ge 0) = 1$ then $\mathbb{E}X \ge 0$.*
- *$\mathbb{E}(X + aY) = \mathbb{E}X + a\mathbb{E}Y$ for any $a \in \mathbb{R}$.*

We immediately get consequences of these axioms for notions of variance and covariance, so we need not add additional baggage to each of these for the moment.

Definition 2 (Independence) *A collection of events $\{A_i \mid i \in I\}$ are independent if*
$$\mathbb{P}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbb{P}(A_i)$$
The application of this to random variables is exactly as one would expect.

Convergence of random variables

Take X, Y random variables. In certain cases we would like a concept of distance between X and Y.

Definition 3 (Convergence) *Take a sequence (X_n) of random variables, and random variable X.*
- *$X_n \to X$ (almost surely) if $\mathbb{P}(\{X_n \to X \text{ as } n \to \infty\}) = 1$.*
- *$X_n \xrightarrow{P} X$ (in probability) as $n \to \infty$ if for every $\varepsilon > 0$, $\mathbb{P}(|X_n - X| < \varepsilon) \to 1$ as $n \to \infty$.*
- *$X_n \xrightarrow{d} X$ (in distribution) as $n \to \infty$ if for every $x \in \mathbb{R}$ such that F is continuous at x, $F_n(x) \to F(x)$ as $n \to \infty$.*

We should find that the above notions are decreasing in strength. By its nature we can often write distribution convergence not with a random variable X, but just with its distribution.

To show that almost sure convergence implies probabilistic convergence, we first state the following lemma:
Lemma 1 *Let A_n be an increasing sequence of events (For all $k \in \mathbb{N}$, $A_k \subseteq A_{k+1}$). Then*
$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup_{k=0}^{\infty} A_k\right)$$
.
As proof, write
$$\begin{aligned}\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=0}^{n} A_k\right) &= \mathbb{P}\left(A_0 \cup \bigcup_{k=1}^{n} A_k \setminus A_{k-1}\right)\\ &= \mathbb{P}(A_0) + \sum_{k=1}^{n} \mathbb{P}(A_k \setminus A_{k-1})\\ &\to \mathbb{P}(A_0) + \sum_{k=1}^{\infty} \mathbb{P}(A_k \setminus A_{k-1})\\ &= \mathbb{P}\left(\bigcup_{k=0}^{\infty} A_k\right).\end{aligned}$$
We can then consider the event defined in almost sure convergence:
$$\begin{aligned}\{X_n \to X \text{ as } n \to \infty\} &= \{\forall \varepsilon > 0 \,.\, \exists N \ge 0 \,.\, \forall n \ge N \,.\, |X_n - X| < \varepsilon\}\\ &= \bigcap_{\varepsilon > 0} \bigcup_{N=0}^{\infty} \{\forall n \ge N \,.\, |X_n - X| < \varepsilon\}\\ &\subseteq \bigcup_{N=0}^{\infty} \{\forall n \ge N \,.\, |X_n - X| < \varepsilon\} \qquad \text{for any } \varepsilon > 0\end{aligned}$$
Thus we turn the event of convergence into an infinite union of increasing sets, which is itself an event of probability 1, so we have $\mathbb{P}(\{\forall n \ge N \,.\, |X_n - X| < \varepsilon\}) \to 1$ as $N \to \infty$. Further,
$$\begin{aligned}\{\forall n \ge N \,.\, |X_n - X| < \varepsilon\} &= \bigcap_{n=N}^{\infty} \{|X_n - X| < \varepsilon\}\\ &\subseteq \{|X_n - X| < \varepsilon\} \qquad \text{for any } n \ge N\end{aligned}$$
so we get $1 \ge \mathbb{P}(|X_n - X| < \varepsilon) \ge \mathbb{P}(\{\forall n \ge N \,.\, |X_n - X|\}) \to 1$ and by sandwiching probabilistic convergence is achieved.

To show that the inverse doesn't hold, just take a sequence of random variables wherein the probability clearly converges, but not so quickly as to have the probability of an infinite tail being within a small range being likely. See $X_n \sim \mathrm{Ber}(1/n)$.

To show that probabilistic convergence implies distributive convergence, note that in the limit we can get $F_n(x)$ in terms of an arbitrary $\varepsilon > 0$ and X. Then we may bound $F_n(x)$ and use continuity of F to show convergence.

Theorem 2 *For (X_n) all defined on the same probability space, $X_n \xrightarrow{d} c$ for some constant c implies that $X_n \xrightarrow{p} c$.*

This follows fairly immediately from algebra.

Theorem 3 (Weak law of large numbers) *Suppose (X_n) are i.i.d. with mean $\mu < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$. Then*
$$\frac{S_n}{n} \xrightarrow{p} \mu \text{ as } n \to \infty$$

We can prove this statement using characteristic functions:
$$\begin{aligned}\phi_{S_n/n}(t) &= \phi_X(t/n)^n\\ &= \left(1 + i\mathbb{E}[X]\frac{t}{n} + o(t/n)\right)^n\\ &\to e^{it\mathbb{E}[X]} \qquad \text{by continuity of exp and log}\end{aligned}$$
and by the characteristic function continuity result $S_n/n \xrightarrow{d} \mu$, which then means $S_n/n \xrightarrow{p} \mu$ as μ is constant.

Theorem 4 (Strong law of large numbers) *Suppose (X_n) are iid with mean $\mu < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$. Then*
$$\frac{S_n}{n} \to \mu \text{ almost surely as } n \to \infty$$

The proof of this is not examinable, and a full proof is given in Probability, Measure and Martingales.

Theorem 5 (Central limit theorem) *Suppose (X_n) are i.i.d., $\mathbb{E}[X_k] = \mu$, $\mathrm{Var}\, X_k = \sigma^2 < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$, then*
$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

We define $Y_n = \frac{X_n - \mu}{\sigma}$, so $S_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Y_n$, and thus
$$\begin{aligned}\phi_{S_n/\sqrt{n}}(t) &= \phi_Y\left(\frac{t}{\sqrt{n}}\right)^n\\ &= \left(1 - \frac{t^2}{2n} + o(t^2/n)\right)^n\\ &\to e^{-t^2/2}\end{aligned}$$
so by continuity $S_n/\sqrt{n} \xrightarrow{d} N(0,1)$.

Conditional Densities

Definition 4 *For two events A and B with $\mathbb{P}(A) > 0$,*
$$\mathbb{P}(B \mid A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},$$
and in application to random variables, we get
$$\mathbb{P}(X \le x \mid A) = \frac{\mathbb{P}(\{X \le x\} \cap A)}{\mathbb{P}(A)}$$

The second function gives a conditional cdf for X, implying the existence of a pdf $f_{X|A}$ for which
$$\mathbb{P}(X \in C \mid A) = \int_C f_{X|A}(x)\,\mathrm{d}x$$
A problem which we come to is trying to observe the conditional density of Y for $X = x$, as for continuous random variables $\mathbb{P}(X = x) = 0$. To resolve this, we take the distribution of Y conditioned on $\{x \le X \le x + \varepsilon\}$, and for nice enough $f_{X,Y}(x,y)$, $f_X(x)$ we get
$$\begin{aligned}\mathbb{P}(Y \le y \mid x \le X \le x + \varepsilon) &= \frac{\int_{-\infty}^{y}\int_{x}^{x+\varepsilon} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v}{\int_x^{x+\varepsilon} f_X(u)\,\mathrm{d}u}\\ &\sim \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)}\,\mathrm{d}v \qquad \text{as } \varepsilon \to 0\end{aligned}$$

Markov Chains

Definition 5 *Let $X = (X_0, X_1, X_2, \dots)$ be a sequence of random variables taking values in I. The process X is called a Markov chain if for any $n \ge 0$ and $i_0, i_1, \dots, i_{n+1} \in I$,*
$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n)$$
In addition, the Markov chain is homogeneous if $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ is constant in $n \ge 0$.

Intuitively, a Markov chain is a sequence wherein one need not keep track of previous states in order to determine the distribution over future states, but rather one only needs to know where they are (and potentially the time at which they are there). In the case of a homogeneous Markov chain, we can write $P = (p_{ij})$ as the matrix with the ith row the distribution of X_{n+1} given $X_n = i$.

We almost always talk about homogeneous Markov chains in this course.

From the Markov property, we very quickly get a formula for n-step probabilities.

$$\begin{aligned}p_{ij}^{(n+m)} &= \mathbb{P}(X_{n+m+r} = j \mid X_r = i)\\ &= \sum_{k\in I} \mathbb{P}(X_{m+r} = k \mid X_r = i)\mathbb{P}(X_{n+m+r} = j \mid X_{m+r} = k)\\ &= \sum_{k\in I} p_{ik}^{(m)} p_{kj}^{(n)}\\ &= (P^{(m)}P^{(n)})_{ij}\end{aligned}$$
so $P^{(n)} = P^{(n-1)}P$ so by induction $P^{(n)} = P^n$.

It is not quite correct to say that in a Markov chain X_n depends *only* on X_{n-1} - there is certainly still randomness involved, and this would imply a functional relationship which doesn't quite exist. We *can* however say that for each n we can have a random variable $Y_n = f(Y_{n-1}, X_n)$ where X_n is independent of $(Y_0, \dots, Y_{n+1})$. Then (Y_n) is a markov chain.

We say that i leads to j, or $i \to j$ where for some $n \ge 0$, $p_{ij}^{(n)} > 0$, and we say that i communicates with j, or $i \leftrightarrow j$ where $i \to j$ and $j \to i$. This is an equivalence relation, thus partitioning I into communicating classes. We say that a chain for which I is a single equivalence class is irreducible. Further we say that a class is closed if the probability for ever exiting is 0. If the singleton of a state is closed then that state is absorbing.

Definition 6 (Period) *The periodicity of state i is defined as $\gcd\{n \mid p_{ii}^{(n)} > 0\}$. If this is 1 then we say the state is aperiodic.*

All states within the same communicating class have the same period. To see this note that if i and j communicate, then we can get a, b such that $p_{ij}^{(a)} > 0$ and $p_{ji}^{(b)} > 0$, so $p_{ii}^{(a+b)} > 0$. Further, if $p_{jj}^{(m)} > 0$, then $p_{ii}^{(a+b+m)} > 0$. Thus if i has period d, then $d \mid a + b + m$ and $d \mid a + b$, so $d \mid m$. Thus the period of i divides the period of j, and by symmetry thus the reverse holds, so the period of i is equal to the period of j.

Definition 7 *Let (X_n) be a Markov chain, and $A \subseteq I$. Define*
$$h_i^A = \mathbb{P}\left(\bigcup_{n\ge 0}\{X_n \in A\} \mid X_0 = i\right)$$
as the hitting probability of A from i.

Theorem 6 *The vector of hitting probabilities $(h_i^A \mid i \in I)$ is the minimal non-negative solution to the recurrence equations*
$$h_i^A = \begin{cases} 1 & \text{if } i \in A\\ \sum_{j\in I} p_{ij}h_j^A & \text{if } i \notin A\end{cases}$$

The base case is obvious. For the recurrence we partition and use the Markov property. To show that the minimal non-negative solution is correct, take an arbitrary non-negative solution $\boldsymbol{x}$, and show that for all $M \in \mathbb{N}$, $i \in I$ that
$$x_i \ge \mathbb{P}\left(\bigcup_{n\le M}\{X_n \in A\} \mid X_0 = i\right).$$
For $M = 0$ we get if $i \in A$ that $x_i = 1$, and if $i \notin A$ that the right hand side is 0. Further, if the statement is true for $M - 1$, then if $i \in A$ then again $x_i = 1$ so the equation holds, and otherwise we can partition to maintain the inequality.

Recurrence and Transience

For $\mathbb{P}(X_n = i \text{ for some } n \ge 1) < 1$, we have that the total number of visits to i has geometric distribution with parameter $1 - p$, and so the probability that i is hit infinitely often is 0, so we call the state *transient.* If however we have $\mathbb{P}(X_n = i \text{ for some } n \ge 1) = 1$, then clearly the probability of hitting i infinitely often is 1, so we call the state *recurrent.*

Theorem 7 *A state i is recurrent iff $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$.*

The total number of visits to i is $\sum \mathbb{1}(X_n = i)$, which has expectation equal to $\sum p_{ii}^{(n)}$. If i is transient this expectation is finite, whereas if it is recurrent then the expectation is infinite.

Theorem 8 *Let C be a communicating class. Either all states in C are recurrent, or all are transient. Further, every recurrent class is closed, and every finite closed class is recurrent.*

Take a C with a recurrent state, so $\sum_{n=0}^{\infty} p_{ii}^{(n)}$ is infinite. For some a, b, $p_{ji}^{(a)}, p_{ij}^{(b)}$ are positive, so $p_{ji}^{(a)}p_{ii}^{(n)}p_{ij}^{(b)} \le p_{jj}^{(a+b+n)}$, so $\frac{1}{p_{ji}^{(a)}p_{ij}^{(b)}}\sum_{n=0}^{\infty} p_{jj}^{(n)}$ is infinite.

If a drunk person was wandering with uniform random distribution around town, they would return to their original position eventually with probability 1. If, however, they have access to a spaceship, then there is positive probability that they never come home.

Definition 8 *$H^A = \min\{n \ge 0 \mid X_n \in A\}$ is the hitting time of A.*

Theorem 9 *The vector of mean hitting times k^A is the minimal non-negative solution to*
$$k_i^A = \begin{cases} 0 & \text{if } i \in A\\ 1 + \sum_j p_{ij}k_j^A & \text{otherwise}\end{cases}$$

The proof here follows straightforwardly from conditional expectations, and minimality using the same idea as for hitting probabilities.

From this we get the notion of a mean return time, $m_i = 1 + \sum_j p_{ij}k_j^{\{i\}}$. If i is recurrent but m_i is infinite, we say that i is null recurrent. If however $m_i < \infty$ then i is positive recurrent.

Generating Functions

We have an existing notion of generating functions for discrete random variables from prelims probability. That is, $G_X(s) = \mathbb{E}[s^X]$, defined on the radius of convergence of the corresponding power series. We have various results about these functions, such as that the exact distribution of X may be extracted via differentiation, demonstrating uniqueness, and that with (X_n), N independent, each X_n identically distributed, $G_{\sum_{i=1}^{N} X_i}(s) = G_N(G_X(s))$.

Theorem 10 *If each X_n for $n \ge 1$ and X have generating functions G_{X_n} and G_X, then $G_{X_n} \to G_X$ pointwise if and only if $X_n \xrightarrow{d} X$.*

This is hopefully clear from definitions.

Definition 9 *The moment generating function of a random variable X is defined as*
$$M_X(t) = \mathbb{E}[e^{tX}]$$
For example, for $\mathrm{Exp}(\lambda)$:
$$\begin{aligned}M_X(t) &= \mathbb{E}[e^{tX}]\\ &= \int_0^{\infty} e^{tx} f(x)\,\mathrm{d}x\\ &= \int_0^{\infty} \lambda e^{(t-\lambda)x}\,\mathrm{d}x\\ &= \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } t < \lambda\\ \infty & \text{otherwise}\end{cases}\end{aligned}$$
We get fairly quickly a few similar results as for generating functions. For X with a generating function M_X defined for t,
$$\begin{aligned}M_{aX+b}(t) &= \mathbb{E}[e^{t(aX+b)}]\\ &= e^{bt}\mathbb{E}[e^{atX}]\\ &= e^{bt}M_X(at)\end{aligned}$$
and for $\{X_1, \dots, X_n\}$ independent with generating functions defined for each on t,
$$\begin{aligned}M_{\sum_{k=1}^{n} X_k}(t) &= \mathbb{E}[e^{t\sum_{k=1}^{n} X_k}]\\ &= \mathbb{E}\left[\prod_{k=1}^{n} e^{tX_k}\right]\\ &= \prod_{k=1}^{n} \mathbb{E}[e^{tX_k}]\\ &= \prod_{k=1}^{n} M_{X_k}(t).\end{aligned}$$
Furthermore, we have a convergence result, that if $M_{|X|}(t_0)$ exists for some $t_0 > 0$, then for $t \in [-t_0, t_0]$:
$$\begin{aligned}M_{|X|}(t_0) &= \int_0^{\infty} e^{t_0 x}(f(x) + f(-x))\,\mathrm{d}x\\ &\ge \int_0^{\infty} e^{tx}(f(x) + f(-x))\,\mathrm{d}x \qquad \text{for } |t| \le t_0\\ &\ge \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x\\ &= M_X(t)\end{aligned}$$
so $M_X(t)$ is defined on this interval.

Theorem 11 *Suppose $\mathbb{E}[e^{t_0|X|}]$ is finite for some $t_0 > 0$. Then we both have that*
$$M_X(t) = \sum_{k=0}^{\infty} \mathbb{E}[X^k]\frac{t^k}{k!} \quad \text{for } |t| \le t_0$$
and
$$M_X^{(k)}(0) = \mathbb{E}[X^k]$$

One needs a bit of work not included in this course (Fubini's theorem) to show that the expectation operator and infinite sums can commute in this case, but assuming that the result follows.

Although not proven in this course, it should be hopefully clear that if there is a non-trivial interval containing 0 on which the moment generating function exists, then all positive moments must be finite.

An equivalent statement to the existence of the MGF on some neighbourhood of 0 is that for some $t_0 > 0$, $\mathbb{P}(|X| > x) = O(e^{-t_0 x})$. If $M_X(t)$ is finite on $[-t_0, t_0]$, then $\mathbb{P}(|X| > x) \le e^{-t_0 x}M_X(t_0)$ for all $x \ge 0$ by Markov's inequality. In the reverse direction, we can use $\mathbb{E}[e^{tX}] \le \mathbb{E}[e^{t|X|}]$, from which we get
$$\begin{aligned}\mathbb{E}[e^{t|X|}] &= \int_0^{\infty} \mathbb{P}(e^{t|X|} > x)\,\mathrm{d}x\\ &\le 1 + \int_1^{\infty} \mathbb{P}\left(|X| > \frac{\log x}{t}\right)\mathrm{d}x\\ &\le 1 + \int_1^{\infty} Cx^{-t_0/t}\,\mathrm{d}x\end{aligned}$$
which is a finite integral for $0 < t < t_0$.

Theorem 12 *If X and Y have the same moment generating function, which is finite on $[-t_0, t_0]$ for some $t_0 > 0$, then X and Y have the same distribution.*

More generally, if we have a sequence of random variables (X_n) and X with finite moment generating functions on $[-t_0, t_0]$, and $M_{X_n}(t) \to M_X(t)$ as $n \to \infty$ for all $t \in [-t_0, t_0]$, then $X_n \xrightarrow{d} X$ as $n \to \infty$.

The proofs of both the above are beyond the scope of this course.

Definition 10 *The characteristic function of X is $\phi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$.*

Not only can we extend all of the basic results for MGFs to characteristic functions, but our convergence result becomes that the characteristic function always exists. This follows as $\cos(tX)$ and $\sin(tX)$ have image $[-1,1]$, so the function is just the sum of two finite integrals.

Thanks to this convergence result we get the following power series result:
$$\phi_X(t) = \sum_{n=0}^{\infty} \frac{i^n t^n \mathbb{E}[X^n]}{n!}$$
Both the uniqueness and continuity statements hold in a similar way as for MGFs, but as before their proofs are beyond the scope of this course.

TODO: Write down the generatinng and characteristic functions of all basic distributions

Joint distributions

Definition 11 *The joint cumulative distribution function of two random variables X, Y is defined by*
$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y).$$
X and Y are said to be jointly continuous with joint pdf $f_{X,Y}$ if their cdf can be written as an integral
$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}u\,\mathrm{d}v$$
While we *can* change $f_{X,Y}$ at finitely many points without changing the integral, thus violating continuity, in general where $F_{X,Y}$ is differentiable it is natural to write
$$f_{X,Y}(x,y) = \frac{\partial F_{X,Y}}{\partial x \partial y}(x,y)$$
For suitably nice (Borel measurable) sets $A \subseteq \mathbb{R}^2$,
$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
We also get the obvious results of $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}y$, $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}x$.

Theorem 13 *Suppose $T : (x,y) \mapsto (u,v)$ is a bijection from some $D \subseteq \mathbb{R}^2$ to some $R \subseteq \mathbb{R}^2$. We define the jacobian as*
$$\begin{aligned}J &= \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u}\\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v}\end{vmatrix}\\ &= \left|\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}\right|\end{aligned}$$
If X, Y have joint pdf $f_{X,Y}$ which is 0 outside D, then the random variables $(U,V) = T(X,Y)$ are jointly continuous with joint pdf
$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}(x(u,v), y(u,v))J(u,v) & \text{if } (u,v) \in R\\ 0 & \text{otherwise}\end{cases}$$
With $T(A) = B$:
$$\begin{aligned}\mathbb{P}((U,V) \in B) &= \mathbb{P}((X,Y) \in A)\\ &= \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \iint_B f_{X,Y}(x(u,v), y(u,v))J(u,v)\,\mathrm{d}u\,\mathrm{d}v.\end{aligned}$$
So the result is immediate via substitution.

I'm keeping the notation from lectures here, although in all honesty some weird choices were made here. For instance, the function $(u,v) \mapsto (x(u,v), y(u,v))$ is just T^{-1}. The Jacobian used is that of T^{-1} rather than that of T, so in fact the entire statement might be better expressed using T^{-1} than T.

The above can then be generalised to the case of joint distributions of $n > 2$ random variables, for which the Jacobian becomes the determinant of an $n \times n$ matrix. With $Z_1, Z_2, \dots, Z_n$ standard normal variables, their joint density function can be written as
$$\begin{aligned}f_{\boldsymbol{Z}}(\boldsymbol{z}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z_i^2}{2}\right)\\ &= \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\boldsymbol{z}^{\top}\boldsymbol{z}\right)\end{aligned}$$
and we can then define $W_1, W_2, \dots, W_n$ by
$$\begin{pmatrix} W_1\\ W_2\\ \vdots\\ W_n\end{pmatrix} = A\begin{pmatrix} Z_1\\ Z_2\\ \vdots\\ Z_n\end{pmatrix} + \begin{pmatrix} \mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}.$$
For A invertible then we can apply change of variables to get a joint distribution $f_{\boldsymbol{W}}$, giving
$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \frac{1}{(2\pi)^{n/2}|\det A|}\exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^{\top}(AA^{\top})^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right)$$

Stationary distributions

Let X be a markov chain with transition matrix P. A distribution over X_0, π, is a stationary distribution if
$$\pi P = \pi$$
so we have that if X_0 is distributed by π, then so will X_n be for all $n \ge 0$.

Theorem 14 (Ergodic theorem) *Let P be irreducible. Let $V_i(n)$ be the number of visits to state i before time n, that is*
$$V_i(n) = \sum_{r=0}^{n-1} \mathbb{1}(X_r = i).$$
Then for any initial distribution, and for all $i \in I$,
$$\frac{V_i(n)}{n} \to \frac{1}{m_i} \quad \text{almost surely as } n \to \infty$$

The result is immediate for transient chains, as with probability 1, $V_i(n)$ is bounded. If instead the chain is recurrent, take R_k as the time between the kth and $(k+1)$th visits to i, which are i.i.d. with mean m_i, and by the strong law of large numbers their sample mean tends to m_i almost surely, indicating that where T_k is the time of the kth visit to i, as T_1 is finite thus $T_k/k \to m_i$ almost surely. We get that $V_i(T_k) = k$, so $V_i(T_k)/T_k = k/T_k \to 1/m_i$ almost surely as $k \to \infty$, and $T_k \to \infty$ as $k \to \infty$ almost surely. $V_i(n)/n$ is a bounded increasing sequence, so it is known to converge and by the previous statement it must converge to m_i.

Theorem 15 *Let P be an irreducible transition matrix. Then P has a stationary distribution if and only if P is positive recurrent, and the stationary distribution π is unique, given by $\pi_i = 1/m_i$.*

If P is positive recurrent we get that $\pi_i = 1/m_i$ is an eigenvector immediately for finite state spaces, and for infinite we get an upper bound on π_j in terms of π_i which gives the same result. The converse is determined by taking the expected rate of visits $\mathbb{E}V_n(i)/n$ for X_0 distributed by π, noting that this is π_i, and that by probabilistic convergence we can get $\mathbb{E}V_n(i)/n \to 1/m_i$.

Theorem 16 *If P is irreducible and aperiodic with stationary distribution π, then for any initial distribution, for all $i \in I$, $\mathbb{P}(X_n = i) \to \pi_i$ as $n \to \infty$, and in particular for all $i, j \in I$, $p_{ij}^{(n)} \to \pi_j$.*

Let (X_n) be Markov distributed with initial distribution λ, transition matrix P, and (Y_n) initially distributed by π with transition matrix also P. With $T = \inf\{n \ge 0 \mid X_n = Y_n\}$, we can consider $W_n = (X_n, Y_n)$ as a markov chain, which is irreducible with a stationary distribution, so is positive recurrent and $\mathbb{P}(T < \infty) = 1$. Thus we can define the chain Z_n as X_n for $n < T$ and Y_n for $n \ge T$, and it turns out that this is Markov. Thus the result follows from here (**check**).

Time reversal

Theorem 17 *For P an irreducible transition matrix with stationary distribution π, and $(X_0, \dots, X_N) \sim \mathrm{Markov}(\pi, P)$. Then for $0 \le n \le N$, with $Y_n = X_{N-n}$, $(Y_0, \dots, Y_N) \sim \mathrm{Markov}(\pi, Q)$ with $Q = (q_{ij})$ for*
$$q_{ij} = \frac{\pi_j}{\pi_i}p_{ji}$$
and Q also has stationary distribution π.

First we take the matrix Q, and observe that it is stochastic by taking the sum of each row.
$$\begin{aligned}\mathbb{P}(Y_0 = i_0, \dots, Y_N = i_N) &= \mathbb{P}(X_0 = i_N, \dots, X_N = i_0)\\ &= \mathbb{P}(X_N = i_0 \mid X_{N-1} = i_1, \dots, X_0 = i_N)\\ &\quad \mathbb{P}(X_0 = i_N, \dots, X_{N-1} = i_1)\\ &= p_{i_1 i_0}\mathbb{P}(X_0 = i_N, \dots, X_{N-1} = i_1)\\ &= \pi_{i_N}\prod_{k=1}^{N} p_{i_k i_{k-1}}\\ &= \pi_{i_N}\prod_{k=1}^{N} \frac{\pi_{i_{k-1}}}{\pi_{i_k}} q_{i_{k-1} i_k}\\ &= \pi_{i_0}\prod_{k=1}^{N} q_{i_{k-1} i_k}\end{aligned}$$
and consequently we immediately get that $\mathbb{P}(Y_0 = i) = \pi_i$, as well as that $\mathbb{P}(Y_n = j \mid Y_{n-1} = i_{n-1}, \dots, Y_0 = i_0) = q_{i_{n-1}j}$, so independent of $i_0, \dots, i_{n-2}$ and thus $Y \sim \mathrm{Markov}(\pi, Q)$.

We say that a transition matrix P is reversible if $P = Q$.

Theorem 18 *Let P be an irreducible transition matrix with stationary distribution π. P is reversible iff for all $i, j \in I$*
$$\pi_i p_{ij} = \pi_j p_{ji}$$

This follows immediately from the definitions. These equations are sometimes referred to as the detailed balance equations.

Theorem 19 *If the matrix P and the distribution π are in detailed balance, then π is stationary for P.*

This follows as $\pi_j = \sum_i \pi_j p_{ji} = \sum_i \pi_i p_{ij}$ for any j. It is this characterisation of stationary distributions which makes time reversal so useful.