Stochastic Processes

Varun Varanasi

May 5, 2024

Contents

1 Markov Chains

Markov chains refer to systems that evolve over time according to probability distributions. A key detail of markov chains is that the evolution of the system only depends on the most recent state of the system.

Specifying and simulating a Markov Chain

Markov chains are defined by three details:

State Space, $\mathcal{S}$

The state space is a finite or countable set of states that represents the possible realizations of a random variable X.

Initial Distribution, π_0

The initial distribution represents the probaiblity distribution of the Markov chaing at $t = 0$. For each state in the state space $\pi_0(i) = \mathcal{P}[X = i]$ represents the probability that the markov chain begins in state i. Each entry in the initial distribution abides by $\pi_0(i) \geq 0$ and $\sum_i \pi_0(i) = 1$.

Probability Transition Rule, P

If the state space has size N, the proability transition matrix $P \in \mathcal{R}^{N \times N}$. Each entry P_{ij} of the matrix can be interpretted as the conditional probability of transitioning from state i at time n to state j at time $n + 1$. We refer to probability transition matricies that do not depend on n to be time homogenous markov chains. To abide by probability rules, we further require the each entry to be non-zero and the sum of each row to be equal to 1.

The Markov Property

Definition 1 (Markov Property) *A process $X_0, X_1, ...X_n$ is said to satisfy the Markov property if*

$$P\{X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, ...X_0 = i_0\} = P\{X_{n+1} = i_{n+1} | X_n = i_n\} \qquad \forall n, i \in \mathcal{S}$$

Intuitively, the Markov property can be understood as the transition probaiblity solely depending on the previous step. The Markov Property allows the simpliest form of dependence relations between the variables without over complicating the system. Application of the markov property allows us to decompose the path of a random variable in state space into the probabilities of initial position and each transition.

$$\begin{aligned} P\{X_0 = i_0, X_1 = i_1, ...X_n = i_n\} &= P\{X_0 = i_0, \}P\{X_1 = i_1 | X_0 = i_0\}...P\{X_n = i_n | X_0 = i_0, X_1 = i_1, ...X_{n-1} = i_{n-1}\} \\ &= P\{X_0 = i_0, \}P\{X_1 = i_1 | X_0 = i_0\}...P\{X_n = i_n | X_{n-1} = i_{n-1}\} \\ &= \pi_0(i_0)P(i_0, i_1)...P(i_{n-1}, i_n) \end{aligned}$$

Markvo chains can be generalized to capture more complex dependence relations:

Definition 2 (rth Markov Property) *A process $X_0, X_1, ...X_n$ is said to be an rth order Markov if*

$$P\{X_{n+1} = i_{n+1} | X_n = i_n, ...X_0 = i_0\} = P\{X_{n+1} = i_{n+1} | X_n = i_n, ...X_{n-r+1} = i_{n-r+1}\} \qquad \forall n, i \in \mathcal{S}$$

It's all just matrix theory

Let's consider the probability distribution in state space at time n via the probability vector π_n. Suppose that the markov chain evolves through a state space of size N with a time homogenous probability transition matrix. For each entry in our probability vector π_n, we can compute its value by considering the following summation:

$$\pi_{n+1}(j) = P\{X_{n-1} = j\} = \sum_{i=1}^{N} P\{X_n = i\}P\{X_{n+1} = j | X_n = i\} = \sum_{i=1}^{N} \pi_n(i)P(i, j)$$

We can equivalently note this value in matrix notation as $\pi_{n+1} = \pi_n P$. Extrapolating this out, we state that $\pi_n = \pi_0 P^n$

1.1 The basic limit theorem of Markov chains

Theorem 1.1 (Basic Limit Theorem) *Let $X_0, X_1, ...$ be an irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Let X_0 have an arbitrary distribution π_0. The* $\lim_{n\to\infty} = \pi_n(i) = \pi(i) \quad \forall i \in \mathcal{S}$.

Proof to follow later. First, we will define stationary distributions, irreducible, and aperiodic.

Stationary distributions

Definition 3 (Stationary Distribution) *We say that π, a distribution across state space, is stationary if*

$$\pi = \pi P \quad \textit{or equivalently} \qquad \pi(j) = \sum_{i\in\mathcal{S}} \pi_i P(i,j) \ \forall j \in \mathcal{S}$$

If the probability transition matrix is symmetric, then the uniform distriubtion, $\pi(i) = \frac{1}{N}$, is stationary. The uniform distribution is more generally a stationary distribution for any transition matrix that is doubly stochastic, both rows and columns sum to 1.

We can solve for the stationary distribution of markov chain computationally by solving for higher powers of P^n or solving the equivalent algebraic expressions $\pi P = \pi$, $\pi(P-I) = 0$ or $(P^T - I)\pi = 0$ under the constriant $\sum_i \pi(i) = 1$.

Stationary distributions do not necessarily exist nor are they unique. For example, consider the situation where $P = I$. In this case there are infinitely many stationary distributions as each initial distribution is simply stationary. We can see examples of Markov chains without stationary distributions when we begin to consider infinite state spaces.

Definition 4 (Probability Flux) *For two subsets of state space A and B we define the probability flux to be*

$$\textit{flux}(A,B) = \sum_{i\in A}\sum_{j\in B} \pi(i)P(i,j)$$

Notice that when $B = A^c$ $\text{flux}(A, A^c) = \text{flux}(A^c, A)$.

1.2 Irreducibility, Periodicity, and Recurrence

Irreducibility

Notation Note: $\mathbb{P}_i$ and $\mathbb{E}_i$ refers to the probability/expectation of an event given that the chain was in state i at time $t = 0$.

Definition 5 (Accessibility) *A state, j, is said to be accessible from state i if it is possible for the chain to visit state j starting in state i over infinite time.*

$$\mathbb{P}_i = \left(\bigcup_{n=0}^{\infty} \{X_n = j\} \right) > 0$$

Equivalently,

$$\sum_{n=0}^{\infty} P^n(i,j) = \sum_{n=0}^{\infty} \mathbb{P}_i\{X_n = j\} > 0$$

Two states are said to ***communicate*** *if they are mutually acessible.*

Definition 6 (Irreducible) *A markov chain is said to be irreducible if every pair of states communicates with eachother.*

We can decompose the state space of the markov chain into groups, or classes, that communicate. Under this definition, a state space that can be partitioned into a single communicating class is an irreducible class. Notice that to maek this determination we only rely on the state space $\mathbb{S}$ and the probability transition matrix P.

Periodicity

To account for markov chains bouncing between probability distributions we introduce the constraint of periodicity in the basic limit theorem.

Definition 7 (Periodicity) *For a given markov chain $\{X_0, X_1, ...X_n\}$ the period of a state i is given by*

$$d_i = gcd\{n : P^n(i,i) > 0\}$$

Theorem 1.2 *If states i and j communicate, then $d_i = d_j$*

Proof: By the definition of communicate we have that for states i and states j $P^{n_1}(i,j) > 0$ and $P^{n_2}(j,i) > 0$. This implies that

$$P^{n_1+n_2}(i,i) > 0$$

From our definition of preiodicity, we know that d_i must divide $n_1 + n_2$ Now we can introduce a quantity n such that $P^n(j,j) > 0$. It necessarily follows that $P^{n_1+n_2+n} > 0$ and we can apply the definition of periodicity. Intuitively, we have introduced an additional path from i to i that involves a sub loop from j to j($i- > j- > j- > i$). We know that this quanitity must also be divided by d_1. Since both $n_1 + n_2 + n$ and $n_1 + n_2$ are divided by d_i, it is clear that d_i divides n. Since d_j is the gcd of all path lengths from j to j, $d_j \geq d_i$. We can prove this inequality in the opposite direction by replicating the argument with swapped indicies. Therefore, $d_i = d_j$.

The theorem implies that all communicating classes have the same periodic number. Therefore, in the context of an irreducible chain with a single communicating class, we can meaningfully talk about the periodicity of the markov chain.

Definition 8 (Aperiodic) *An aperiodic, irreducible markov chain is an irreducible markov chain with period 1. All other periods are considered periodic.*

Recurrence

Intuitively, recurrence is a measure of whether a Markoc chain in state i will eventually return to state i. If we define the hitting time of state i as

$$T_i = \inf\{n > 0 : X_n = i\}$$

Definition 9 (Recurrent) *A state is said to be recurrent if $\mathbb{P}\{T_i < \infty\} = 1$. Otherwise the state is considered* ***transient****.*

Note that recurrent differs from accessible in that it is absolute that the state will return. In terms of hitting times, we can refer to accessibility as $\mathbb{P}\{T_i < \infty\} > 0$.

Theorem 1.3 *If i is a recurrent state and j is accessible from i, then*

(i) $\mathbb{P}_i\{T_j < \infty\} = 1$

(ii) $\mathbb{P}_j\{T_I < \infty\} = 1$

(iii) j is recurrent

Proof: It is obvious that iii) follows from i) and ii). We informally can argue i) by thinking about each cylce the chain takes from i to i. Since it is a recurrent state we know that this cycle will occur with definite probability in a finite amount of time. Furthermore, we know that j is accessible from i. We can then construct a bernoulli random variable I_n to denote whether the nth cycle from i passes through j. In the limit as $n \to \infty$ for non-zero probability,p, we can say that $I_n = 1$ with certainity. Statement ii) is also implied by this argument since if a finite cycle from i to i passing through j occurs, it is necessary that the hitting time between j and i is finite.

Theorem 1.4 *The state i is recurrent if and only if $\mathbb{E}_i(N_i) = \infty$ where $N_i = \sum_{n=0}^{\infty} I\{X_n = i\}$*

Proof: The forward direction is clear since for a recurrent state i, $\mathbb{P}_i\{N_i \infty\} = 1$. For the converse, we consider that i is transient such that $\mathbb{P}_i\{T_i = \infty\} > 0$. If we use the same repeated cycle argument from above, we see that we can construe the evolution of the markov chain as repeated cycles around i until it hits a path that takes infinite time. We can model N_i as a geometric distribution with probability $q = \mathbb{P}_i\{T_i = \infty\} > 0$. The expected value of this distribution is $1/q$. Therefore, we see that a transient state necessarily has a finite $\mathbb{E}_i(N_i)$.

Corollary 1.4.1 *If j is transient, then $\lim_{n\to\infty} P^n(i,j) = 0$ for all states i*

Proof: Suppose j is transient. From our previous analysis we know that $\mathbb{E}_j(N_j)$ is finite. We now argue that $\mathbb{E}_i(N_j) \leq \mathbb{E}_j(N_j)$ by decomposing $\mathbb{E}_i(N_j) = \mathbb{P}_i\{T_j < \infty\}\mathbb{E}_i\{N_j | T_j < \infty\}$. Intuitively, we understand this as needing to get to j in finite steps and then counting the cycles through j. So, we can conclude that $\mathbb{E}_i(N_j) \leq \mathbb{E}_j(N_j)$. Since we define the expected visits as $\mathbb{E}_i(N_j) = \sum_{n=1}^{\infty} P^n(i,j) < \infty$. This can only be true if $\lim_{n\to\infty} P^n(i,j) = 0$

1.3 Random Walks

A random walk in 1-D is a symmetric random walk where steps are taken in the $+1$ or -1 directions with probability $1/2$. The position of the random walk at time n is given by $S_n = X_1 + ... + X_n$ where X_i is the outcome of each step (± 1). For multiple dimensions, we simply concatenate additional random walks to produce a d dimensional vector.

1-D In the 1-D case we will argue that the markov chain is recurrent by showing that state 0 is recurrent and then citing that recurrence is a class property.

To show recurrence we want to show

$$\mathbb{E}_0(N_0) = \sum_{n}^{\infty} P^n(0,0) = \infty$$

Notice that if n is odd $P^n(0,0)$ is immediately 0. Intuitively, you can think of needing an equal number of steps to the right and to the left to end up at the origin. Therefore, we can restrict our summation to $n = 2k$ for some integer k. Now, we want to think about $P^{2k}(0,0)$. Using the same intuition as above, we need an equal number of steps to the right and left. Therefore, if we have a total of $2k$ steps we need k steps to the left and k to the right. Since each step occurs with probability $1/2$ we can see that this probability is given by the binomial distribution $Binom(2k, 1/2)$

$$P^{2k}(0,0) = \binom{2k}{k} 2^{-2k}$$

We can approximate this value using Stirling's formula.

$$n! \approx \sqrt{2\pi n}\left(\frac{k}{e}\right)^k$$

Applying the approximation we find

$$P^{2k}(0,0) = \frac{(2k)!}{k!k!2^{2k}} \approx \frac{\sqrt{2\pi 2k}(2k/e)^{2k}}{2\pi k(k/e)^{2k}2^{2k}} = \frac{1}{\sqrt{\pi k}}$$

Now, we can argue

$$\mathbb{E}_0(N_0) = \sum_{k}^{\infty} P^{2k}(0,0) = \sum_{n}^{\infty} \frac{1}{\sqrt{\pi k}} = \infty$$

2-D In two dimensions we can recreate this argument. Since the steps in each direction are independent we can make the following statement:

$$\mathbb{P}_{(0,0)}\{S_{2k} = (0,0)\} = \mathbb{P}_{(0)}\{S_{2k}^X = (0)\}\mathbb{P}_{(0)}\{S_{2k}^Y = (0)\} \approx \frac{1}{\sqrt{\pi k}} \cdot \frac{1}{\sqrt{\pi k}} = \frac{1}{\pi k}$$

The summation of these values are also infinite. Therefore, the 2-D random walk is also recurrent.

3-D However, replicating the same argument in 3-D we find that the markov chain is transient.

$$\mathbb{P}_{(0,0,-)}\{S_{2k} = (0,0,0)\} = \mathbb{P}_{(0)}\{S^X_{2k} = (0)\}\mathbb{P}_{(0)}\{S^Y_{2k} = (0)\}\mathbb{P}_{(0)}\{S^Z_{2k} = (0)\} \approx \left(\frac{1}{\sqrt{\pi k}}\right)^3$$

Theorem 1.5 *Suppose a Markov chainh as a stationary distribution π. If state j is transient, then $\pi(j) = 0$.*

Proof: Given that π is stationary we canstate $\pi P^n = \pi$ for all n. Deconstructing this summation for index j

$$\sum_i^n \pi(i)P^n(i,j) = \pi(j) \quad \forall n$$

However, since we know that state j is transient, $\lim_{n\to\infty} P^n(i,j) \to 0$ for all i. As n approaches ∞ we see that the summation approaches 0 and since this equality must hold for all n, we can state that $\pi(j) = 0$.

Corollary 1.5.1 *If an irreducible Markov chain has a stationary distribution, then it must be recurrent.*

Proof: Irreducibility allows us to apply class properties to the chain. Therefore, it must be either recurrent or transient. From the previous theorem we know that if a transient state exists, $\pi(j) = 0$. Since transience is a class property this implies that all states will have $\pi(j) = 0$ which cannot be a stationary distribution.

At this point it is important to introduce the concept of **null recurrence**. Null recurrence refers to the condition where $\mathbb{P}\{T_i < \infty\}$ but $\mathbb{E}_i[T_i] = \infty$. Null recurrence can be thought of as states that are barely recurrent in that despite T_i is almost assuredly finite, the expectation is infinite. States that are not null recurrent are referred to as positive recurrent. These are recurrent states with less than infinite expectations.

These statements will be explained in depth later, but positive recurrence is a class property and any irreducible Markov chain has a stationary distribution if and only if it is positive recurrent.

An aside on coupling

At a high level coupling arguments involve constructing two objects from the same set of random numbers. These objects are then compared to make a probabilistic argument.

Add more ig

1.4 Proof of the Basic Limit Theorem

The Basic Limit Theorem states that if a irreducible aperiodic Markov chain has a stationary distribution π then for any initial distribution π_0 $\pi_n \to \pi$ as $n \to \infty$. In proving this theorem it is useful to define a distance between probability distributions and show that this distance converges to 0 as $n \to \infty$.

Definition 10 *Let μ and λ be two probability distributions over the sample space $\mathcal{S}$. The **total variational distance** is defined as*

$$||\lambda - \mu|| = \sup_{A\subset\mathcal{S}} [\lambda(A) - \mu(A)]$$

Alternatively,

$$||\lambda - \mu|| = \frac{1}{2}\sum_{i\in\mathcal{S}} |\lambda(i) - \mu(i)| = 1 - \sum_{i\in\mathcal{S}} \min\{\lambda(i), \mu(i)\}$$

Add proofs later

Another interpretation for the total variational distance is the largest possible discrepancy between probailities assigned by the two distributions.

2 Markov Chains: Examples and Applications

2.1 Branching Processes

The problem of Branching Processes was first studied by Sir Francis Galton when thinking about the life time of family names. If family name is only carried by a male heir, what is the probability that the family name goes extinct?

As a formal description let's consider G_t as the number of males in the $t-th$ generation. If we begin with $G_0 = 1$ and $G_t = n$, we can write a recursive relationship for the number of male descendenants in the next generation as $G_{t+1} = \sum_{k=1}^{n} X_{tk}$ where X_{tk} represents the number of sons fathered by the k-th individual in the previous generation. Each X_{tk} is an iid random variable following a pre-defined probability mass distribution.

We can think of this process as a Markov chain. In this language, we can see that state 0 is clearly an absorbing state. Similarly, we argue that any state i is transient. Notice that for any state i, $\mathbb{P}_i\{G_{t+1} = 0\} = f(0)^i > 0$. We can see that for sufficiently large i, $P_i\{T_0 < \infty\} < 1$ and therefore state i must be transient. Therefore, each state i is visited only a finite number of times. Consequently, the chain must either absorb at 0 or approach ∞.

With this intuition in hand we can start thinking about defining a recursive definition of the probability of extinction ρ.

$$\rho = \sum_{k=0}^{\infty} \mathbb{P}\{G_1 = k | G_0 = 1\} \mathbb{P}\{\text{extinction} | G_1 = k\}$$

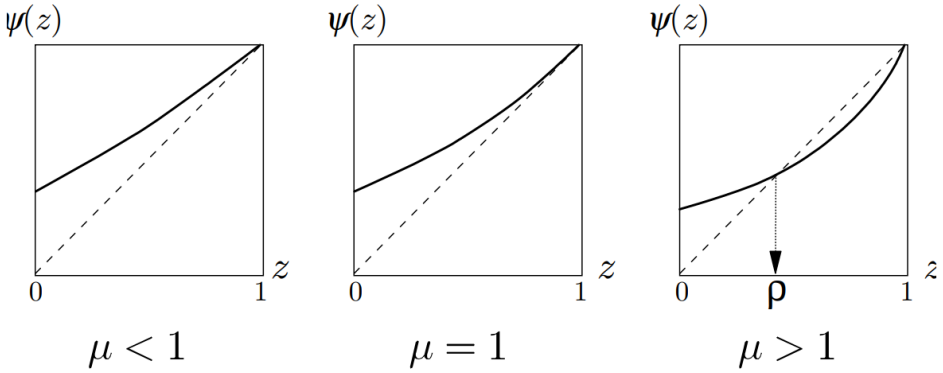
Since each man has a son with independent probability, a lineage can only die off if each of the k lineages from G_1 die off. Therefore, $\mathbb{P}\{\text{extinction}|G_1 = k\} = \rho^k$

$$\rho = \sum_{k=0}^{\infty} f(k)\rho^k = \psi(|rho)$$

We introduce the function ψ to construct a function only dependent on rho. The extinction probability is then the fixed point of this function. We refer to ψ as the probability generating function for the probability mass function f. Using derivative test we can see that the function is strictly increasing and convex. $\psi(0) = f(0)$ and $\psi(1) = 1$. An interesting quantity is $\psi'(1)$, the expected number of sons.

$$\psi'(1) = \sum_{k=0}^{\infty} kf(k) = \mu = \mathbb{E}[X]$$

Depending on the value of μ, we get three possible graphs:



For $\mu \leq 1$ the only solution is $\rho = 1$ and the probability of extinction is certain. We see that a non-trivial fixed point ρ only exists for the case $\mu > 1$.

Add in proof that the lower fixed point is the correct one

2.2 Time Reversibility

Time Reversibility captures the idea that you can watch the movement of a Markov chain and have no probabilistic intuition on which direction the chain is moving. Alternatively, it can be understood as being unable to tell whether the Markov chain movement you are watching is being played in reverse.

Definition 11 *A Markov chain is said to be time-reversible if for each n*

$$(X_0, ...X_n) = (X_n, X_{n-1}, ...X_0)$$

Equivalently, the joint distribution of the forward and backward chains are equivalent.

As a consequence of time-reversibility, we can claim that every time reversible chain must also be stationary. Consider the joint distribution $(X_0, X_1) = (X_1, X_0)$. This implies that $X_1 = X_0$ and consequently $\pi_1 = \pi_0$. Since $pi_1 = pi_0 P$ we can equivalently write $pi_0 = pi_0 P$ which is our definition of stationarity

Theorem 2.1 *The Markov chain $\{X_n\}$ is time-reversible if and only if the distribution of X_0 π satisfies $\pi = \pi P$ and $\pi(i)P(i,j) = \pi(j)P(i,j)$ for all i,j.*

The conditions imposed by this theorem, particularly that the probability flux from i to j is the same as j to i, is known as local or detailed balance. These conditions are necessary to ensure the propogation of the distribution that results in the index swapping required for time-reversibility. Stationarity is characterized by global balance: $\pi(j) = \sum_i \pi(i)P(i,j)$ or equivalently $\sum_j \pi(j)P(j,i) = \sum_i \pi(i)P(i,j)$. Intuitively, global balance enforces that the total probability flux for each state is 0 while local balance states that the probability flux betweens pairs of nodes is equal.

Theorem 2.2 *If the local balance conditions hold, then the distribution π is stationary.*

Lemma 2.3 *All birth and death chains are stationary*

For a birth death chain we know that the current chain can only go up or down by 1. Therefore, we can restrict our flux analysis to indicies above and below i. Using the fact that probability flux of a subset of the states pace must equal its complement, we can subset our space from $\{0, ..., i\}$ such that the two transitions of interest are in differing subsets. It follows that the flux from each subset and thus transition must be equal. Therefore, the chain must be stationary.

Random walks are also time-reversible Markob chains. Add example + walk through why

2.3 The Metropolis Method

The Metropolis method is a way of simulating Markov chains. For example, if we hope to simulate a random draw from a distribution π we can run an irreducible aperiodic Markov chain with stationary distribution π for a sufficiently long time. One way we can do this is by constructinga graph on which we can run a random walk. From previous analysis we know that the probability transition matrix for a random walk on a graph is

$$P_{rw}(i,j) = \begin{cases} \frac{1}{d(i)} & j \in N(i) \\ 0 \end{cases}$$

and has stationary distribution

$$\pi_{rw}(i) = \frac{d(i)}{\sum_j d(j)}$$

If we want to compute a different or more complciated distribution, we can consider another distribution π where

$$\pi(i) \propto d(i)f(i)$$

where $f(i)$ is a scaling function.

The Metropolis method proposes the following augmented transition matrix:

$$P(i,j) = \begin{cases} \frac{1}{d(i)} \min\{1, \frac{f(j)}{f(i)}\} & j \in N(i) \\ 1 - \sum_k P(i,k) & j = i \\ 0 \end{cases}$$

To verify that this method works, let's begin by showing that π is stationary under this Markov chain. For $j \in N(i)$,

$$\pi(i)P(i,j) \propto d(i)f(i) \cdot \frac{1}{d(i)} \min\{1, \frac{f(j)}{f(i)}\} = f(i) \min\{1, \frac{f(j)}{f(i)}\} = \min\{f(i), f(j)\}$$

It is clear that this quantity is symmetric across i and j so $\pi(i)P(i,j) = \pi(j)P(j,i)$ We can similarly verify this for $j = i$ and $j \notin N(i)$. Now, if we sum over the indicies i

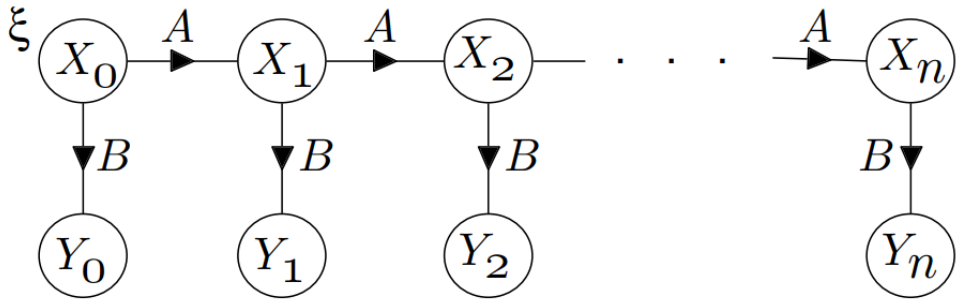
$$\sum_i \pi(i)P(i,j) = \pi(j) \sum_i P(j,i) = \pi(j)$$

Intuitively, the Metropolis method is the same as a random walk; however, now, after selecting an edge to traverse, we decide whether to traverse this candidate edge with probability $\min\{1, \frac{f(j)}{f(i)}\}$. If we dont traverse this edge, we stay in state i.

2.4 Hidden Markov Chains

Model Description

A Hidden Markov Chain is a subset of Markov Random Fields with many applications in speech recognition and bioinformatics among other fields. They are characterized by only observing some of the random variables.



In the above model we have a Markov chain $X_0, ...X_n$ with observed variables $Y_0, ...Y_n$. The model is parameterized by ξ the marginal distribution of X_0, and the two transition matricies A and B (assuming time-homogeneity). If we have u states and v observed states, it follows that $\xi \in \mathbb{R}^u$, A is a $u \times u$ matrix and B is a $u \times v$ matrix.
How well does an average of a time inhomogeneous MC model the chain?

Calculating Likelihoods

A likelihood function is the probability of the observed data given the model parameters. In the HMM model,

$$L(\theta) = p_\theta(y_0, y_1, ...y_n) = \sum_{x_0} \sum_{x_1} ... \sum_{x_n} p_\theta(x_0, ...x_n, y_0, y_n) = \sum_x p_\theta(x, y)$$

We can cleverly calculate this value by computing a recursive calculation. First, let's define

$$\alpha_t(x_t) = p_\theta(y_0, ...y_y, x_t)$$

For the base case, lets consider $t = 0$.

$$\alpha_0(x_0) = p_\theta(x_0, y_0) = \xi(x_0)B(x_0, y_0)$$

For the general case, we can decompose the sum as follows:

$$\begin{aligned}\alpha_{t+1}(x_{t+1}) &= p_\theta(y_0, ...y_{t+1}, x_{t+1}) \\ &= \sum_{x_t} p_\theta(y_0, ...y_{y+1}, x_t, x_{t+1}) \\ &= \sum_{x_t} p_\theta(y_0, ...y_t, x_t) p_\theta(x_{t+1}|x_t) p(y_{t+1}|x_{t+1}) \\ &= \sum_{x_t} \alpha_t(x_t) A(x_t, x_{t+1}) B(x_{t+1}, y_{t+1})\end{aligned}$$

Now we can see that the likelihood function is just a summation of these values:

$$L(\theta) = p_\theta(y_0, ..., y_n) = \sum_{x_n} p_\theta(y_0, ..., y_n, x_n) = \sum_{x_n} \alpha_n(x_n)$$

3 Martingales

Martingales are intended to model some notion of fairness.

Definition 12 (Martingales) *A process is said to be a martingale if* $\forall n \in \mathbb{N}$

$$\mathbb{E}[M_{n+1}|M_{0:n}] = M_n$$

Intuitively, this mean this that a process is martingale if the conditional expectation of the next step given the previous steps is the current state. In the context of gambling, we can interpret this definition as fair odds since your fortune at the next time step is expected to be equivalent to your current wealth We can also define a process to be martingale with respect to another stochastic process.

Definition 13 (Martingales) *A process is said to be a martingale with respect to another process if* $\forall n \in \mathbb{N}$

$$\mathbb{E}[M_{n+1}|W_{0:n}] = M_n$$

A process is martingale if it is martingale with respect to itself.

Branching Process Example

Consider a branching process $X_0, ...X_n$ with an offspring distribution μ.

$$X_{n+1} = \sum_{i=1}^{X_n} Z_{n,i} \qquad Z_{n,i} \sim \mu$$

We can verify the martingale property of this process by considering

$$\mathbb{E}[X_{n+1}|X_{0:n}] = \mathbb{E}[\sum_{i=1}^{X_n} Z_{n,i}|X_n] = \sum_{i=1}^{X_n} \mathbb{E}[Z_{n,i}] = X_n \cdot \mathbb{E}[\mu]$$

If $\mathbb{E}[\mu] = 1$ then the process is clearly martingale. We can also define the process U_n

$$U_n = \frac{X_n}{\mu^n}$$

Under this definition,

$$\mathbb{E}[U_{n+1}|U_{0:n}] = \mathbb{E}[\frac{X_{n+1}}{\mu^{n+1}}|X_n] = \frac{1}{\mu^{n+1}} \sum_{i=1}^{X_n} \mathbb{E}[Z_{n,i}] = \frac{1}{\mu^n} \frac{X_n}{\mu} \cdot \mathbb{E}[\mu] = U_n$$

Therefore, U_n is martingale with respect to itself.

Definition 14 (Submartingale) *A proceess is said to be Submartingale with respect to another process W_n if* $\forall n \in \mathbb{N}$

$$\mathbb{E}[X_{n+1}|W_{0:n}] \geq X_n$$

The current state is below the expected value for the future.

Definition 15 (Supermartingale) *A proceess is said to be Supermartingale with respect to another process W_n if* $\forall n \in \mathbb{N}$

$$\mathbb{E}[X_{n+1}|W_{0:n}] \leq X_n$$

The current state is above the expectation for the future.

3.1 Optional Stopping Theorem

Consider a martingale, U_n, with respect to W_n. For an arbitrary n, we can use the tower property to write

$$\mathbb{E}[U_{n+1}] = \mathbb{E}[\mathbb{E}[U_{n+1}|W_{0:n}]] = \mathbb{E}[U_n]$$

It follows that for any $n \in \mathbb{N}$

$$\mathbb{E}[U_n] = \mathbb{E}[U_0]$$

This statement can be extended further to include random stopping times T (under a few assumptions) to say $\mathbb{E}[U_T] = \mathbb{E}[U_0]$. Intuitively, we can understand thsi extensionas saying that even if a gambler has a strategy on when to stop gambling, in expectation, they will have the same fortune.

Definition 16 (Random Stopping Time) *A random variable T taking values in $\mathbb{Z} \geq 0$ is called a stopping time with respect to W_n if the indicator $\mathbb{1}(T = k)$ is a function of $W_{0:k}$.*

- Stopping times can take infinite values
- Stopping times cannot rely on information in the future

Theorem 3.1 (Optional Stopping Theorem (Strong Assumptions)) *Consider M_n a martingale with respect to W_n. Let T denote a random stopping time such that for some $B \leq \infty$, if $\mathbb{P}(T \leq B) = 1$, then*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0]$$

Proof:

First, we can decompose M_T into a sum of the differences between subsequent realizations of the process.

$$M_T = M_0 + \sum_{k=1}^{T}(M_k - M_{k-1}) = M_0 + \sum_{k=1}^{B}(M_k - M_{k-1})\mathbb{1}(T \geq k)$$

We can now take the expectation over this and write

$$\mathbb{E}[M_T] = \mathbb{E}[M_0] + \sum_{k=1}^{B}\mathbb{E}[(M_k - M_{k-1})\mathbb{1}(T \geq k)]$$

Notice that we can rewrite the indicator variable as $\mathbb{1}(T \leq k) = 1 - \mathbb{1}(T \leq k-1)$. This term is a function of $W_{0:n}$. Now we use the tower property to write

$$\begin{aligned}
\mathbb{E}[(M_k - M_{k-1})\mathbb{1}(T \geq k)] &= \mathbb{E}[\mathbb{E}[(M_k - M_{k-1})\mathbb{1}(T \geq k)|W_{0:n}]] \\
&= \mathbb{E}[\mathbb{1}(T \geq k)\mathbb{E}[(M_k - M_{k-1})|W_{0:n}]] \\
&= \mathbb{E}[\mathbb{1}(T \geq k) \cdot 0] \\
&= 0
\end{aligned}$$

Therefore,

$$\mathbb{E}[M_T] = \mathbb{E}[M_0] + 0$$

We can weakenthe bounded condition and write

Theorem 3.2 (Optional Stopping Theorem (Weak Assumptions)) *Consider M_n a martingale with respect to W_n. Let T denote a random stopping time.*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0]$$

if $T \leq B$ with probability 1 or $\mathbb{E}[T] \leq \infty$ and $\mathbb{E}[|M_{n+1} - M_n|W_{0:n}] \leq B$

Proof: Beyond the scope of the class

If M_n is a Supermartingale or Submartingale, then the OST tells us that

$$[E][M_t] \leq \mathbb{E}[M_0](\text{ Supermartingale}) \qquad [E][M_t] \leq \mathbb{E}[M_0](\text{ Submartingale})$$

There are many applications of the OST. Below, we outline a simple proposition that makes use of the OST.

Theorem 3.3 *For any X_n with $\mathbb{E}[(X_n - X_{n-1})^2|X_{0:n}] = 1$, for $T = \inf\{n \geq 0 : X_n = -a \text{ or } X_n = b\}$, then*

$$\mathbb{E}[T] = |a| \cdot b$$

The proof of this follows from defining a markovchain $U_n = X_n^2 - n$ and applying the OST for the random stopping time T. Add Proof later.

Theorem 3.4 *Let D_i be a Supermartingale on $[0, n]$ and $\mathbb{E}[(D_{i+1} - D_i)^2|D_{0:i}] \geq \sigma^2$. If $T = \inf\{i \geq 0 : D_i = 0\}$, then*

$$\mathbb{E}[T] \leq \frac{n^2}{\sigma^2}$$

Proof: Follows from showing $Y_i = X_i^2 - 2nX_i - \sigma^2 i$ is a Submartingale and then applying the OST.

3.2 Martingale Convergence

Theorem 3.5 (Martingale Convergence Theorem) *Any nonnegative supermartingale converges with probability 1*

Lemma 3.6 *Let M_n represent a supermartingale such that S and T are bounded stopping times with $S \leq T$ with probability 1. Then,*

$$\mathbb{E}[M_S] \geq \mathbb{E}[M_T]$$

When $S = 0$ this is identical to the OST for supermartingales.

Lemma 3.7 *Let M_n be a nonnegative supermartingale. For $b > 0$, if $T_b = \inf\{t \geq 0 : M_n \geq b\}$ then*

$$\mathbb{P}(T_b \leq \infty) \leq \frac{\mathbb{E}[M_0]}{b}$$

If $\mathbb{E}[M_0] < a$, we can write

$$\mathbb{P}(T_b \leq \infty) \leq \frac{a}{b}$$

Proof:

We first introduce $N \in \mathbb{N}$ and define $T_b \wedge N = \min\{T_b, N\}$. This bounds our stopping time by N. We can then apply the Optimal Stopping Theorem and state

$$\mathbb{E}[M_{T_b \wedge N}] \leq \mathbb{E}[M_0]$$

If $T_b \leq N$, we have $M_{T_b \wedge N} = M_{T_b} = b$. Therefore, we can write the expression

$$M_{T_b \wedge N} \geq \mathbb{1}(T_b \leq N)b$$

Now, taking the expectation over each side,

$$\mathbb{E}[M_{T_b \wedge N}] \geq \mathbb{P}\{T_b \leq N\} b$$

We can rearrange this and take the limit $N \to \infty$ to recover

$$\mathbb{P}(T_b \leq \infty) \leq \frac{\mathbb{E}[M_0]}{b}$$

This results statse that any supermartingale that starts below some $a \geq 0$ will stay below $b > a$ with probability at least $1 - \frac{a}{b} > 0$.

Proof (Martingale Convergence Theorem)

We can prove convergence by arguing that the probability the supermartingale tends towards infinity or the probability that the submartingale oscillates between two values is infinitely many times is 0.

$$\mathbb{P}\left(\lim U_n = \infty \bigcup \{U_n < a \text{ infinitely often } U_n > b\}\right) = 0$$

We can apply a union bound and show that the above statement is bound by

$$\mathbb{P}\left(\lim U_n\right) + \sum_{a<b} \{U_n < a \text{ infinitely often } U_n > b\}$$

We show that the first term approaches 0 by applying the earlier lemma.

$$\mathbb{P}\left(U_n \geq b \text{ for some n }\right) \leq \mathbb{E}[U_0]/b$$

First, note that we can construct the following chain of inequalities:

$$\mathbb{P}\left(\lim U_n\right) \leq \mathbb{P}(\bigcap_{b'>0} \{U_n \geq b'\}) \leq \mathbb{P}\left(U_n \geq b \text{ for some n }\right)$$

Taking $b \to \infty$ the right side approaches 0.

$$\mathbb{P}\left(\lim U_n\right) \leq \mathbb{P}\left(U_n \geq b \text{ for some n }\right) \leq \mathbb{E}[U_0]/b = 0$$

To show that the second term vanishes, we begin by defining a sequence of stopping times:

$$\begin{gathered} T_0 = 0 \\ S_1 = \inf\{t \geq T_0 : U_t \leq a\} \\ T_1 = \inf\{t \geq S_1 : U_t \geq b\} \\ \ldots \\ S_n = \inf\{t \geq T_{n-1} : U_t \leq a\} \\ T_n = \inf\{t \geq S_n : U_t \geq b\} \end{gathered}$$

These hitting time definitions formalize the idea of exiting the bounded region $[a, b]$. Our goal is to show that the number of n is finite. We bound both stopping times by N and consider the iequality:

$$\mathbb{E}(U_{T_k \wedge N}) \leq \mathbb{E}(U_{S_k \wedge N})$$

This can be rewritten as

$$\mathbb{E}(U_{T_k \wedge N} - U_{S_k \wedge N}) \leq 0$$

We can conditional decompose each term in this expression into

$$U_{T_k \wedge N} \geq b\mathbb{1}(T_k \leq N) + U_N \mathbb{1}(T_k > N) \tag{1}$$
$$U_{S_k \wedge N} \geq a\mathbb{1}(S_k \leq N) + U_N \mathbb{1}(S_k > N) \tag{2}$$

Now, considering the difference between the two,

$$U_{T_k \wedge N} - U_{S_k \wedge N} \geq b\mathbb{1}(T_k \leq N) - a\mathbb{1}(S_k \leq N) + U_N\left(\mathbb{1}(T_k > N) - \mathbb{1}(S_k > N)\right)$$

Since $T_k \geq S_k$ by definition the final term is identically 0.
Now, we can write the following:

$$0 \geq \mathbb{E}(U_{T_k \wedge N} - U_{S_k \wedge N}) = b\mathbb{P}(T_k \leq N) - a\mathbb{P}(S_k \leq N)$$

Rearranging this expression and taking $N \to \infty$

$$\mathbb{P}(T_k \leq \infty) \leq \frac{a}{b}\mathbb{P}(S_k \leq \infty)$$

Note that S_k is dependent on T_{k-1}. Therefore, for $S_k < \infty$ we require $T_{k-1} < \infty$.

$$\mathbb{P}(T_k \leq \infty) \leq \frac{a}{b}\mathbb{P}(T_{k-1} \leq \infty) = \left(\frac{a}{b}\right)^2 \mathbb{P}(T_{k-2} \leq \infty) = ... = \left(\frac{a}{b}\right)^k$$

As k approaches ∞ we see that the probability becomes 0. Therefore, we can conclude that the Supermartingale does not oscillate infinite times.

Recurrence of Random Walks

4 Gaussian Random Variables

The multivariate gaussian distribuion has mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in R^{n \times n}$.

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

The Moment Generating Function for $X \sim N(\mu, |sigma)$ is

$$\mathbb{E}[e^{\alpha^T X}] = e^{\alpha^T \mu + \alpha^T \Sigma \alpha/2}$$

This defines the guassian distribution if we know the covariance matrix. Recall that $cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Corollary 4.0.1 *$x_1, ..., x_n$ have a joint distribution which is gaussian if and only if $\sum_i \alpha_i x_i$ has a normal distribution*

Definition 17 (Gaussian Process) *A process $w(t)$ is said to be gaussian if and only if $w(t_1), ..., w(t_n)$ are jointly gaussian for all $n \in \mathbb{N}$*

If $w(t)$ is a mean-zero gaussian process, then the distribution is solely determined by the covariances of $cov(w(t), w(s))$

5 Brownian Motion

The brownian motion problem attempts to model a continous time random walk on the real line. We can think of the continuous stochastic process as a random variable over paths the process can take. Instead of simply mapping the state space to a real value, in this perspective, we map both the time and the state space to a real value. If we fix time the continous process is a random variable. If we fix the state space, we define a path.

Definition 18 (Standard Brownian Motion) *A Standard Brownian Motion, $w(t)$, is a stochastic process with 1) continuous paths, 2) stationary and independent increments, and 3) $\forall t > 0, w(t) \sim N(0,t)$*

The 3rd statement is equivalent to saying that the time marginal is the normal distribution with mean 0 and variance t. A process is said to have independent increments if $w(t_1) - w(t_0)$, $w(t_2) - w(t_1)$, ... $w(t_n) - w(t_{n-1})$ are all independently distributed. This can be interpretted as the variation in the paths over these time steps being independent. Increments are stationary if $\forall 0 \leq s < t$, $w(t) - w(s)$ is only dependent on $t - s$.

Lemma 5.1 *For a SBM process,* $w(t) - w(s) \sim N(0, t-s)$

By the stationary increments argument, $w(t) - w(s)$ must have the same distribution as $w(t-s) - w(0)$. We know that $w(0) = 0$ so $w(t-s) - w(0) = w(t-s) \sim N(0, t-s)$. SBM is a path that starts at 0 and progress as a path of indpeendent gaussian increments. It is likea gaussian random walk.

Lemma 5.2 *$w(t)$ is a Gaussian process or $w(t_1), ..., w(t_n)$ follow a gaussian distribution.*

Lemma 5.3 *For SBM,* $r(s,t) = cov(w(s), w(t)) = \min(s,t)$

Add covariance proof (just algebra with covariance definition + clever use of independent increments)

Definition 19 (Standard Brownian Motion 2) *$w(t)$ is SBM if and only if its a gaussian process with* $r(s,t) = \min\{r,t\}$ *and continuous paths*

Lemma 5.4 *If $w(t)$ is a SBM, then the process defined* $X(t) = tW(\frac{1}{t})$ *is also SBM.*

Proof:
First, we want to show that X is a gaussian process.

$$a_1X(t_1) + ... + a_nX(t_n) = a_1t_1W(1/t_1) + ... + t_na_nW(1/t_n)$$

It is clear that the right hand side is just a linear combination of a process we know to be gaussian. We know $X(t)$ is continuous since $W(t)$ is and we are not transforming the function to change continuity. The tricky part is showing this property for when $t = 0$. You can show this by breaking $w(s)$ into a summation of increments and then argue that by the strong law of large numbers these increments will attain their expectation 0. Therefore, $w(s)/s \to 0$.For the covariance function, we can apply the properties of covariance.

$$cov(X(t), X(s)) = cov(tW(\frac{1}{t}), sW(\frac{1}{s})) = stcov(W(\frac{1}{t}), W(\frac{1}{s})) = st\min\{\frac{1}{t}, \frac{1}{s}\} = \min\{s,t\}$$

Theorem 5.5 (Markov Property) *Suppose $w(t)$ is SBM.* $\forall c > 0, X(t) = W(t+c) - W(c)$ *is also an SBM. $X(t)$ is independent of* $\{w(t) : 0 \le t \le c\}$

If we know the realization of the process at a certain point, then the path is independent of previous realizations conditional on this point.

5.1 Reflection Principle

The reflection principle allows us to calculate the distribution of the hitting time τ_b or the first time the process reaches b

Theorem 5.6 (Reflection Principle)

$$\mathbb{P}\{\tau_b \le t\} = 2\mathbb{P}\{w_t > b\} = 2\left(1 - \Phi(\frac{b}{\sqrt{t}})\right)$$

Proof:

First, notice that we can write

$$\{w_t > b\} \bigcap \{\tau_b \le t\} = \{w_t > b\}$$

since the process is continuous. We can then write the following equalities

$$\mathbb{P}(\tau_b \le t) = \mathbb{P}(\tau_b \le t, w_t \le b) + \mathbb{P}(\tau_b \le t, w_t > b) = \mathbb{P}(w_t < b | \tau_b \le t)\mathbb{P}\{\tau_b \le t\} + \mathbb{P}\{w_t > b\}$$

We can use the properties of SBM to argue that $\mathbb{P}(w_t < b | \tau_b \le t) = 1/2$. In particular, we invoke the Markov property and say that conditioning on $w_s = b$ for some s, by symmetry, the probability that $w_t < b - w_t > b$ for some time $t > s$.

$$\mathbb{P}(\tau_b \leq t) = \frac{1}{2}\mathbb{P}\{\tau_b \leq t\} + \mathbb{P}\{w_t > b\}$$
$$\mathbb{P}(\tau_b \leq t) = 2\mathbb{P}\{w_t > b\}$$

Since we know that $w_t \sim N(0,t)$ we can write that

$$\mathbb{P}\{\tau_b \leq t\} = 2\mathbb{P}\{w_t > b\} = 2\left(1 - \Phi(\frac{b}{\sqrt{t}})\right)$$

5.2 Conditional Distribution

For $0 \leq t, u$ consider $W_t|W_u$. Suppose we have $u \leq t$.

$$W_t = W_t - W_u + W_u \sim N(W_u, t-u)$$

This follows from the distribution of SBM with the addition of a constant W_u. However, if we have $t \leq u$, we have a more interesting problem. Recall that if (X,Y) follow a jointly Gaussian distribution, then their conditional distributions $X|Y$ also follow a gaussian distribution. Therefore, if we can characterize $\mathbb{E}[W_t|W_u]$ and $Var[W_t|W_u]$ we can characterize the distribution of $W_t|W_n$. We show this by showing the independence of $W_t - \frac{t}{u}W_u$ from W_u. To show this, we use the fact that jointly gaussian distributed variables are only independent if their covariance is 0. We argue that $W_t - \frac{t}{u}W_u$ is jointly distributed with W_u and show the covariance is 0.

$$cov(W_t - \frac{t}{u}W_u, W_u) = cov(W_t, W_u) - \frac{t}{u}cov(W_u, W_u) = t - \frac{t}{u}\cdot u = 0$$

Now, we can write the conditional expectations

$$\mathbb{E}[W_t - \frac{t}{u}W_u|W_u] = \mathbb{E}[W_t - \frac{t}{u}W_u] = 0$$

Therefore,

$$\mathbb{E}[W_t|W_u] = \frac{t}{u}\mathbb{E}[W_u|W_u] = \frac{t}{u}W_u$$

We can perform a similar calculation for variance:

$$Var[W_t|W_u] = \frac{t(u-t)}{u}$$

Add in variance details

The best guess is the point on the line between $W(0) = 0$ and W_u

5.3 Brownian Bridge

Definition 20 (Brownian Bridge) *A process is a standard brownian bridge if it is standard brownian motion conditioned on* $W(1) = 0$

We think of Brownian Bridges as SBM on $[0,1]$. Using the above conditional distribution calculations, we can write

$$\mathbb{E}[W_t|W_1 = 0] = tW(1) = 0 \qquad cov(W(s), W(t)|W(1) = 0) = s(1-t)$$

This process is also gaussian so a brownian bridge is a gaussian process with mean and covariance functions as defined above. $X(t) = W(t) - tW(1)$ is a brownian bridge which allows us to compute many of the desired properties.

Boundary Crossing

This is an equivalent calculation to the reflection principle for a standard brownian motion. What is $\mathbb{P}\{\tau_b \leq t|W(1) = 0\}$?

Theorem 5.7

$$\mathbb{P}\{\tau_b \leq t | W(t) = x\} = e^{-\frac{2b(b-x)}{t}}$$

Skipped Proof

For a standard brownian bridge, this implies that

$$\mathbb{P}\{\tau_b \leq 1 | W(1) = 0\} = e^{-2b^2} \qquad \mathbb{P}\{\max X(s) \geq b\} = e^{-2b^2}$$

review second equality

6 Diffusions

We can generalize the notion of a Simple Brownian Motion with the following definition:

Definition 21 ((μ, σ^2)-Brownian Motion) *A Process is said to be (μ, σ^2)-Brownian Motion if it can be written as*

$$X(t) = X(0) + \mu t + \sigma W(t) \forall t$$

where $W(t)$ is a SBM.

μ is considered the drift of the random motion and σ is the variance. The (μ, σ^2)-Brownian Motion is a linear brownian motion; however, they can be used to construct a more general diffusion. Much like constructing a curve from tangent approximations, we say diffusions are generalized random motion defined by local approximations via (μ, σ^2)-Brownian Motion processes. Therefore, despite being constructed from linear structures, diffusions can model any for mof general continuous stochastic processes. Diffusions are then defined by functions $\mu(X_t)$ and $\sigma^2(X_t)$.

Definition 22 (Diffusion) *A stochastic process is a diffusion if it satisfies the strong Markov property and has continuous paths with probability 1.*

Diffusions are fully characterized by their drift and variance functions:

$$\mu(X) = \lim_{h \to 0} \frac{\mathbb{E}[X(t+h - X(t) | X(t) = x)]}{h} \tag{3}$$

$$\sigma^2(X) = \lim_{h \to 0} \frac{Var[X(t+h - X(t) | X(t) = x)]}{h} \tag{4}$$

$$\tag{5}$$

It can be shown that for every point $(t^*) = x$ the diffusion follows the brownian motion

$$(t - t^*)\mu(x) + \sigma^2(x) W(t - t^*)$$

This holds for an infinitesly small amount of time following t^*. This property is intuiitively identical to the taylor approximation of a curve. Notice that we can also write the variance expression as

$$\lim_{h \to 0} \frac{\mathbb{E}[(X(t+h - X(t))^2 | X(t) = x)]}{h} = \lim_{h \to 0} \frac{Var[X(t+h - X(t) | X(t) = x)]}{h} = \sigma^2(x)$$

This follows from $\mathbb{E}[X(t+h - X(t) | X(t) = x)] = 0$. We also assume that for $p \geq 3$

$$\lim_{h \to 0} \frac{\mathbb{E}[(X(t+h - X(t))^p | X(t) = x)]}{h} = 0$$

Geoemetric Brownian Motion

Suppose $X(t)$ is a (μ, σ^2)-Brownian Motion. Let $Y(t) = e^{X(t)}$. Since $X(t)$ satisfies the Markov property and has continuous paths, it follows that $Y(t)$ does as well. Therefore, it is a diffusion. The drift and variance functions are

$$\mu(x) = (\mu + \frac{\sigma^2}{2})x \tag{6}$$
$$\sigma^2(x) = \sigma^2 x^2 \tag{7}$$

This definition of a diffusion yields independent ratios. $\frac{Y(t_2)}{Y(t_1)} \perp \frac{Y(t_3)}{Y(t_2)}$ and each $Y(t) \geq 0$.

Ornstein-Uhlenbeck Process
This process was introduced by Einstein in his explanation of brownian motion. The process is defined by $\mu(x) = -x$ and $\sigma^2(x) = 2$. The $-x$ drift can be thought of as a frictional force that pushes the system back towards the 0.

Theorem 6.1 *For a diffusion $X(t)$ defined by $\mu(x)$ and $\sigma^2(x)$ consider the diffusion $Y(t) = f(X(t))$ where f is strinctly monotone and twice-differentiable. If $Y(t)$ is a diffusion (both abides by the strong Markov property and has continuous paths), then the drift and variance functions of $Y(t)$ are*

$$\mu_y(y) = \mu(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x) \tag{8}$$
$$\sigma_y^2(y) = f'(x)^2\sigma^2(x) \tag{9}$$

Proof:

Recall the Taylor expansion

$$g(x) \approx g(x^*) + (x - x^*)g'(x^*) + \frac{(x - x^*)^2}{2}g''(x^*)$$

We can apply this approximation to the conditional expectation of Y

$$\mathbb{E}[Y(t+h) - Y(t)|Y(t) = y] = \mathbb{E}[f(x(t+h)) - f(x(t))|x(t) = x] \tag{10}$$
$$= \mathbb{E}([x(t+h) - x(t))f'(x(t)) + \frac{1}{2}(x(t+h) - x(t))^2 f''(x(t))|x(t) = x] \tag{11}$$
$$= \mathbb{E}([x(t+h) - x(t))|x(t) = x]f'(x) + \frac{f''(x)}{2}\mathbb{E}[(x(t+h) - x(t))^2|x(t) = x] \tag{12}$$
$$\tag{13}$$

In the limit $h \to \infty$ we get

$$\lim_{h\to\infty} \frac{\mathbb{E}[Y(t+h) - Y(t)|Y(t) = y]}{h} = \lim_{h\to\infty} \frac{\mathbb{E}([x(t+h) - x(t))|x(t) = x]}{g} f'(x) + \frac{f''(x)}{2} \lim_{h\to\infty} \frac{\mathbb{E}[(x(t+h) - x(t))^2|x(t) = x]}{h} \tag{14}$$
$$= f'(x)\mu(x) + \frac{1}{2}f''(x)\sigma^2(x) \tag{15}$$

We can calculate something similar for the variance.

$$\mathbb{E}[(Y(t+h) - Y(t))^2|Y(t) = y] = \mathbb{E}[(f(x(t+h)) - f(x(t)))^2|x(t) = x] \tag{16}$$
$$= \mathbb{E}([(x(t+h) - x(t))f'(x(t)) + \frac{1}{2}(x(t+h) - x(t))^2 f''(x(t)))^2|x(t) = x] \tag{17}$$
$$= \mathbb{E}([x(t+h) - x(t))|x(t) = x]f'(x)^2 + \text{ terms involving } \mathbb{E}[(x(t+h) - x(t))^p|x(t) = x] \tag{18}$$
$$\tag{19}$$

Since $p \geq 3$ we know that these terms vanish.

$$\lim_{h\to 0} \frac{\mathbb{E}[(Y(t+h) - Y(t))^2|Y(t) = y]}{h} = f'(x)^2 \lim_{h\to\infty} \frac{\mathbb{E}([x(t+h) - x(t))^2|x(t) = x]}{h} = f'(x)^2\sigma(x)^2$$

6.1 Kolmogorov's Forward and Backward Equations

Our next step is to characterize the protability transition structure for a diffusion. In Markov Chains, we had the probability transition matrix provide us the transition probabilities from time t to $t+1$, but since diffusions are continuous, we need new machinery to speak on the rate of change of the probability density function for a given state. We can do this by defining a set of partial differential equations.
First, consider a fixed state y. We define the function $f(t,x)$ to be the density of X_t evaluated at y given that $X_0 = x$.

$$f(t,x) = f_{X_t}(y|X_0 = x)$$

What is the probability density of hitting state y, if we take the process X_t with $X_0 = 0$. The Kolmogorov backward equation states

$$\partial_t f(t,x) = \mu(x)\partial_x f(t,x) + \frac{1}{2}\sigma^2(x)\partial_{xx} f(t,x)$$

If we consider an initial probability density of X_0 and define $g(t,y)$ to be the density of X_t evaluated at y, $g(t,y) = f_{X_t}(y)$, the forward equation gives us how this probability density evolves over time

$$\partial_t g(t,y) = -\partial_y(\mu(y)g(t,y)) + \partial_{yy}\left(\frac{1}{2}\sigma^2(y)g(t,y)\right)$$

Driftless Brownian Motion

If we consider the simple SBM with $\mu(x) = 0$ and $\sigma^2(x) = 1$,

$$\partial_t g(t,y) = \frac{1}{2}\partial_{yy} g(t,y)$$

This equation is the heat equation since it models how the temperature of a material evolves over time.

Stationarity

Diffusions converge to a stationary distribution as $t \to \infty$.

Definition 23 (Stationary Distribution) *We say that $X(0) \sim \Pi$ is a stationary distribution for $X(t)$ if $\forall t \geq 0$, $X(t) \sim \Pi$*

Notice that the stationary definition implies $g(t,y) = \Pi(y)$ since $X_t \sim |Pi$. This means that the right hand side of the forward equation simplifies to 0.

$$0 = -\partial_y(\mu(y)\Pi(y)) + \partial_{yy}\left(\frac{1}{2}\sigma^2(y)\Pi(y)\right) \tag{20}$$

$$\tag{21}$$

This differential equation only has one variable and is much easier to work with.

Ornstein-Uhlenbeck Process
The OU process is defined by $\mu(x) = -x$ and $\sigma^2(x) = 2$. Our stationarity constraint now gives us

$$0 = -(y\Pi(y))' + \frac{1}{2}(2\Pi(y))'' \tag{22}$$

$$(\Pi(y)' + y\Pi(y))' = 0 \tag{23}$$

This implies that there is a constant such that $\Pi(y)' + y\Pi(y) = B$. We can write this as the equivalent expression

$$\left(\Pi(y)e^{y^2/2}\right)' = Be^{y^2/2}$$

We can then integrate both sides and write

$$\Pi(y)e^{y^2/2} = B\int_0^y e^{x^2/2}dx + C$$

If we consider $y \to -\infty$ this equation only holds true if $B = 0$
Therefore, we have

$$\Pi(y) = Ce^{-y^2/2}$$

If we include normalization, we see that this is simply the Gaussian distribution. Therefore, the process converges to the standard gaussian.

Langevin Dynamics

This can be thought of as the continuous analog of metropolis hastings. Essenentially, we can create a difussion that, in the limit $t \to \infty$, samples from the density:

$$\frac{e^{-\psi(x)}}{\int_{-\infty}^{\infty} e^{-\psi(x)} dx}$$

We use the diffusion $\mu(x) = -\psi'(x)$ and $\sigma^2(x) = 2$.

7 Stochastic Differential Equations

Stochastic Differential equations are methods to model randomness in the world. They also provide a clean definition of diffusions. Consider the rate of change equation

$$\frac{dX(t)}{dt} = \mu(x(t), t) + \sigma(x(t), t)N(t)$$

where the $N(t)$ term represents some sense of noise or randomness in the growth rate of our process. We make the following assumptions:

- $N(t)$ is a stationary Gaussian process ($N(t) - N(s)$ is a function of $t - s$)
- $\mathbb{E}[N(t)] = 0$
- $N(t) \perp N(s) \quad \forall s \neq t$

We can think of $N(t)$ as a SBM where $dW(t) = N(t)dt$. We can now think of our equation as

$$dX(t) = \mu(X(t), t)dt + \sigma(x(t), t)dW(t)$$

We interpret this stochastic differential equaiton as

$$X(t + dt) \approx X(t) + \mu(X(t), t)dt + \sigma(x(t), t)W(dt)$$

Notice that this expression is a generalization of the diffusion since the drift and variance now also depend on t as well as $X(t)$.

7.1 Ito's Formula

If $X(t)$ obeys $dX(t) = \mu(X(t), t)dt + \sigma(x(t), t)dW(t)$, then for all twice differnetiable functions f, if $Y(t) = f(t, X(t))$

$$dY(t) = \left(\frac{\partial f}{\partial t}(t, x(t)) + \mu(x(t), t) \cdot \frac{\partial f}{\partial x}(t, x(t)) + \frac{\sigma^2(t, x(t))}{2} \frac{\partial^2 f}{\partial^2 x}(t, x(t)) \right) dt \tag{24}$$

$$+ \sigma(t, x(t)) \frac{\partial f}{\partial x}(t, x(t))dW(t) \tag{25}$$

$Y(t)$ is a diffusion with the corresponding drift and variance functions. This expression can be cleverly simplified into

$$dY(t) = \frac{\partial f}{\partial t}(t, x(t))dt + \frac{\partial f}{\partial x}(t, x(t))dx(t) + \frac{1}{2} \frac{\partial^2 f}{\partial^2 x}(t, x(t))(dx(t))^2 \tag{26}$$

Notice that stochastic calculus introduce an additional $\frac{1}{2}\frac{\partial^2 f}{\partial^2 x}(t, x(t))(dx(t))^2$ term that wouldn't be expected if the system was deterministic. We can compute $(dx(t))^2$ by using the rules $dt(d\text{anything}) = 0$ and $(dW_t)^2 = dt$

Skipped showing the simplied form is equal to the first form

Geometric Brownian Motion

Let $Y(t) = e^{\mu t + \sigma W(t)}$. From above, we know that the drift and variance functions are $\mu(x) = \left(\mu + \frac{\sigma^2}{2}\right) x$ and $\sigma^2(x) = \sigma^2 x^2$. We can find the associated Stochastic Differential Equation as

$$dY(t) = (\mu + \frac{\sigma^2}{2})Y(t)dt + \sigma Y(t)dW(t)$$

How can we reocver $Y(t)$ given this SDE?