A Brief Review of

Stochastic Processes and Bayesian Inference

MVE550

Ruben Seyer <rubense@student.chalmers.se>

16th January 2020

Contents

Preface

A few short remarks about this document are in order. Of importance is the fact that it in no manner claims to be a complete treatise on the subjects contained within, but rather an overview of theory, results and methods for practical purposes. In many cases, a thorough understanding of the subject may already be required to make sense of the brief descriptions within. For that reason, frequent references to the course material that expand on mentioned subjects are contained within. The signs D and LN denote Dobrow, Robert P. *Introduction to Stochastic processes with R and Mathematica* and lecture notes for the course, respectively. References to section or page numbers may follow.

A brief explanation of notation follows. The author has elected to continue the usage of the π notation for probability mass functions and probability density functions in general, regardless of the distribution which must be implicitly inferred from arguments. Thus both continuous and discrete, univariate and joint, conditional and parameterized distributions are referred to in this manner, some overlapping; for example, some parameter ϑ may have a prior $\pi(\vartheta)$ and a posterior $\pi(\vartheta|\text{data})$ (with respect to data clear from context). If the distribution is not obvious, or some sort of evaluation is taking place, the distribution may be named and parameters expressed after a semicolon. This notation has a few great advantages: many statements simply generalize immediately, and Latin letters are freed up for better uses. In a similar manner, many integrals and sums are implicitly considered over the support of concerned distributions for brevity. Probabilities themselves are denoted Pr and expectations E.

Finally, no author is perfect and the present one constitutes no exception. This document is prone to revision and the reader should ensure through the date that it is in fact the most recent version. From the author's website, the most recent version is always available. Questions, corrections and remarks are always welcome and will be thankfully credited should they lead to valuable changes. We hope you find this document useful.

1 Background

1.1 Concepts from probability theory

Definition 1.1 (Stochastic process)**.** A *stochastic process* is a collection of random variables $\{X_t\}_{t\in I}$ defined on a common state space S with an index set I.

Common formulas include e.g. definitions of conditional probability and the variance formula with a thousand names:

$$\pi(x, y) = \pi(y|x)\pi(x) = \pi(x|y)\pi(y) \tag{1.1}$$

$$\mathrm{E}(f(X)) = \int_x f(x)\pi(x)\,dx \tag{1.2}$$

$$\mathrm{E}(Y|X) = \int_y y\pi(y|x)\,dy \tag{1.3}$$

$$\mathrm{Var}(Y) = \mathrm{E}((Y - \mathrm{E}(Y))^2) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 \tag{1.4}$$

Theorem 1.1 (Law of total probability)**.** *Let A be some event, and let $B_1, \ldots, B_n$ be a sequence of events that partitions S.*

$$\Pr(A) = \sum_k \Pr(A \cap B_k) = \sum_k \Pr(A|B_k)\Pr(B_k) \tag{1.5}$$

Definition 1.2 (Probability generating function)**.** For any random variable X with values in $\mathbb{N}_0$ its *probability generating function* (pgf) $G_X(s)$ is defined as

$$G_X(s) = \mathrm{E}(s^X) = \sum_{n=0}^{\infty} s^n \Pr(X = n) \tag{1.6}$$

Some properties of probability generating functions:

- $G(1) = 1$
- $G^{(j)}(0) = j!\Pr(X = j)$
- $G_{X+Y}(s) = G_X(s)G_Y(s)$, assuming X, Y independent
- $G^{(j)}(s) = \mathrm{E}[\prod_{i=0}^{j-1}(X - i)s^{X-j}]$
  $\implies G'(1) = \mathrm{E}(X),\ \mathrm{Var}(X) = G''(1) + G'(1) - G'(1)^2$

Definition 1.3 (Moment generating function)**.** For any random variable X its *moment generating function* (mgf) $M_X(t)$ is defined as

$$M_X(t) = \mathrm{E}(e^{tX}) = \int_x e^{tx}\Pr(X = x)\,dx \tag{1.7}$$

Both the pgf and mgf (where applicable) uniquely identify distributions.

1.2 Bayesian inference

Proposition 1.2 (Bayes' formula)**.**

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)} \tag{1.8}$$

We remark that it is simply a restatement of the definition in (1.1).

The general approach consists of updating models with observations of new data. Consider the case where we perform inference about a parameter ϑ: given a *prior* $\pi(\vartheta)$ and a *likelihood* $\pi(\text{data}|\vartheta)$ we can compute a *posterior* $\pi(\vartheta|\text{data})$ using the formula.

It may be difficult to compute the *prior predictive* $\pi(\text{data})$ directly. We can elect to compute the posterior using proportionalities, ignoring factors independent of ϑ, and afterwards infer the constant by the fact that all densities must integrate to 1 over their support. [LN 1.1]

Note that there is no requirement that the prior is entirely a valid distribution, as long as the posterior becomes one. Prior densities that integrate to infinity are called *improper priors*.

1.2.1 Conjugacies

Definition 1.4 (Conjugacy)**.** Given a likelihood model $\pi(x|\vartheta)$, a conjugate family of priors to this likelihood is a parametric family of distributions such that if the prior for ϑ is in this family, then the posterior is also in this family.

A short list of useful conjugacies follows, in an extremely concise notation with the pattern likelihood; prior $\to$ posterior. Recall that, for multiple data points, this process can be repeated. [LN 8]

- $X \sim \text{Binomial}(n, \vartheta)$; $\vartheta \sim \text{Beta}(\alpha, \beta) \to \text{Beta}(\alpha + x, \beta + n - x)$
- $X \sim \text{Exp}(\lambda)$; $\lambda \sim \text{Gamma}(\alpha, \beta) \to \text{Gamma}(\alpha + 1, \beta + x)$
- $(X_1, \dots, X_n) \sim \text{Multinomial}(n, \vartheta_1, \dots, \vartheta_k)$;
  $(\vartheta_1, \dots, \vartheta_k) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k) \to \text{Dirichlet}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
- $X \sim \text{Poisson}(\lambda)$; $\lambda \sim \text{Gamma}(\alpha, \beta) \to \text{Gamma}(\alpha + x, \beta + 1)$
- $X \sim \text{Normal}(\mu, \vartheta^{-1})$; $\vartheta \sim \text{Gamma}(\alpha, \beta) \to \text{Gamma}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$
- $X \sim \text{Normal}(\vartheta, \tau^{-1})$; $\vartheta \sim \text{Normal}(\mu, \tau_0^{-1}) \to \text{Normal}(\frac{\tau x + \tau_0 \mu}{\tau + \tau_0}, \frac{1}{\tau + \tau_0})$

Usage of these greatly simplifies computations. Some common priors reformulated into useful distributions:

- Uniform$(0, 1) \iff$ Beta$(1, 1)$
- $\pi(\vartheta) \propto 1 \iff$ Gamma$(1, 0)$
- $\pi(\vartheta) \propto 1/\vartheta \iff$ Gamma$(0, 0)$

1.2.2 Predictions

The end goal of inference is to make predictions. There are two expressions that are commonly used to compute the *posterior predictive*: [LN 1.4]

$$\pi(y_{\text{new}}|\text{data}) = \int_\vartheta \pi(y_{\text{new}}|\vartheta)\pi(\vartheta|\text{data})\,d\vartheta = \frac{\int_\vartheta \pi(y_{\text{new}}|\vartheta)\pi(\text{data}|\vartheta)\pi(\vartheta)\,d\vartheta}{\int_\vartheta \pi(\text{data}|\vartheta)\pi(\vartheta)\,d\vartheta} \tag{1.9}$$

$$\pi(y_{\text{new}}|\text{data}) = \frac{\pi(y_{\text{new}}|\vartheta)\pi(\vartheta|\text{data})}{\pi(\vartheta|y_{\text{new}}, \text{data})} \tag{1.10}$$

2 Markov chains

Definition 2.1 (Markov chain)**.** Let S be a discrete set (but not necessarily finite), called the *state space*. A *Markov chain* is a sequence of random variables $X_0, X_1, \ldots$ taking values in S with the property

$$\pi(X_{n+1}|X_0, X_1, \ldots, X_n) = \pi(X_{n+1}|X_n) \quad \forall n \geq 0 \tag{2.1}$$

Definition 2.2 (Time-homogeneity)**.** The chain is *time-homogenous* if, for any states,

$$\pi(X_{n+1}|X_n) = \pi(X_1|X_0) \quad \forall n \geq 0 \tag{2.2}$$

Definition 2.3 (Transition matrix)**.** A *stochastic matrix* is a real matrix P with non-negative entries satisfying $P\mathbf{1}^\mathsf{T} = \mathbf{1}^\mathsf{T}$ (i.e. the rows sum to one). The *transition matrix* is a stochastic matrix defined as $P_{ij} = \pi(X_1 = j|X_0 = i)$.

P is *positive* if all entries are positive, and *regular* if P^n is positive for some $n > 0$.

If v is a vector describing the distribution of states at step k, then vP is a vector describing the distribution of states at step $k + 1$, and likewise for vP^n at step $k + n$.

The probability of observing a specific sub-sequence $X_{n_1} = i_1, \ldots, X_{n_k} = i_k$ with $n_1 < \cdots n_k$ is

$$(p_0 P^{n_1})_{i_1}(P^{n_2-n_1})_{i_1 i_2}(P^{n_3-n_2})_{i_2 i_3} \cdots (P^{n_k-n_{k-1}})_{i_{k-1} i_k} \tag{2.3}$$

where p_0 is the distribution of states at X_0. [D 2.1-2.3]

2.1 Limiting and stationary distributions

The long-term behaviour of $P^n,\ n \gg 1$ varies depending on the chain. The matrix may stabilize with identical rows, with different rows (e.g. $P = I$), or not stabilize at all (e.g. odd cycles). A block-diagonal transition matrix may combine several behaviours.

Definition 2.4 (Limiting distribution)**.** A *limiting distribution* for a Markov chain with transition matrix P is a probability vector v such that

$$\lim_{n\to\infty} (P^n)_{ij} = v_j \quad \forall i, j \tag{2.4}$$

Markov chains have either no or one limiting distribution. If it exists, the probability corresponds to the proportion of time steps spent at each state. [D 3.1]

Definition 2.5 (Stationary distribution)**.** A *stationary distribution* for a Markov chain with transition matrix P is a probability vector v such that $vP = v$.

Markov chains have zero, one or many stationary distributions, and limiting distributions are stationary distributions (but not necessarily vice versa). [D 3.2]

We can find a v satisfying $vP = v$ by solving the system together with $v\mathbf{1}^\mathsf{T} = 1$, making an educated guess or computing an eigenvector to P^T with eigenvalue 1.

2.1.1 Random walk on an undirected graph

A (weighted) undirected graph defines a random walk Markov chain by at every time step following one of the edges out with a probability according to the weights. If the graph is finite, we have the stationary distribution

$$v_i = \frac{w(i)}{\sum_k w(k)} \tag{2.5}$$

where $w(i)$ is the sum of the weights of the edges into node i.

2.2 Properties

Definition 2.6 (Communication class)**.** State j is *accessible* from state i if $(P^n)_{ij} > 0$ for some $n \geq 0$. If i is accessible from j and vice versa we say they *communicate*. This property is in fact an equivalence relation, which defines *communication classes*.

2.2.1 Irreducibility

Definition 2.7 (Irreducibility)**.** A Markov chain is *irreducible* if it has exactly one communication class.

A regular transition matrix implies an irreducible chain (but not the other way around due to periodicity).

2.2.2 Recurrence and transience

Let $T_j := \min\{n > 0 : X_n = j\}$ be the first passage time to state j. Define $f_j := \Pr(T_j < \infty | X_0 = j)$ as the probability that a chain starting at j will return. A state j is *recurrent* if $f_j = 1$, otherwise *transient*.

The expected number of visits at j starting at i is given by $\sum_{n=0}^{\infty}(P^n)_{ij}$. Equivalently, a state j is *recurrent* if the series diverges to infinity, otherwise *transient*.

The states of a single communication class are either all recurrent or transient. If a state is recurrent, only states inside its communication class are accessible. The states of a *finite irreducible* Markov chain are all recurrent. (However, there are infinite irreducible Markov chains where all states are transient.) [D 3.3]

In general, states may be recurrent but the expected return time is infinite. Such states are called *null recurrent*. If the expected return time is finite we call the state *positive recurrent.*

2.2.3 Periodicity

Definition 2.8 (Periodicity)**.** The *period* of a state i is the gcd of all $n > 0$ such that $(P^n)_{ii} > 0$. A Markov chain is *periodic* if it is irreducible and all states have period > 1, otherwise *aperiodic*. [D 3.5]

Corollary 2.1. *All states of a communication class have the same period.*

2.2.4 Ergodicity

Definition 2.9 (Ergodicity)**.** A Markov chain is *ergodic* if

- it is irreducible,
- it is aperiodic and
- all states are positive recurrent (implied if finite state space). [D 3.6]

2.2.5 Time reversibility

Definition 2.10 (Time reversibility)**.** Let P be the transition matrix of an irreducible Markov chain with stationary distribution v. We say the chain is *time reversible* if, after reaching its stationary distribution, it looks the same both forwards and backwards:

$$\pi(X_k = i, X_{k+1} = j) = \pi(X_{k+1} = i, X_k = j) \tag{2.6}$$

$$\iff \Pr(X_k = i)P_{ij} = \Pr(X_k = j)P_{ji} \tag{2.7}$$

$$\iff v_i P_{ij} = v_j P_{ji} \tag{2.8}$$

If a probability vector satisifies this, then it must be the stationary distribution.

The random walk on a (weighted) undirected graph is time reversible, and any time reversible chain can be represented by a random walk on a weighted undirected graph with weights $w_{ij} = v_i P_{ij} = v_j P_{ji}$.

2.3 Absorbing chains

Definition 2.11 (Absorbing chain)**.** A state i is *absorbing* if $P_{ii} = 1$. A Markov chain is *absorbing* if it has at least one absorbing state.

By reordering the states, the transition matrix for an absorbing chain can be written in block form as

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} \tag{2.9}$$

where Q contains the transient states, R contains the recurrent (absorbing) states and I is an appropriate identity matrix. By induction

$$P^n = \begin{bmatrix} Q^n & (I + Q + Q^2 + \cdots + Q^{n-1})R \\ 0 & I \end{bmatrix} \tag{2.10}$$

and taking the limit, using $\lim_{n\to\infty} Q^n = 0$,

$$\lim_{n\to\infty} P^n = \lim_{n\to\infty} \begin{bmatrix} Q^n & (I - Q^{n-1})(I - Q)^{-1}R \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & (I - Q)^{-1}R \\ 0 & I \end{bmatrix} \tag{2.11}$$

Definition 2.12 (Fundamental matrix)**.** The matrix $F = (I - Q)^{-1}$ is called the *fundamental matrix*.

We list some common computations relating to absorbing chains. Note that the absorbing state indices k are counted separately from one. [D 3.8]

- The probability to be absorbed in absorbing state k starting from transient state i is given by $(FR)_{ik}$.
- The expected number of visits to state j starting in transient state i is given by F_{ij}.
- Thus, the expected number of steps until absorption starting in transient state i is given by $(F\mathbf{1}^\mathsf{T})_i$.

2.4 Limit theorems

Theorem 2.2 (Limit theorem for regular Markov chains)**.** *If the transition matrix P is regular, the limiting distribution exists, is positive, and it is the* unique *stationary distribution. [D 3.1 p.83–84]*

Theorem 2.3 (Limit theorem for finite irreducible Markov chains)**.** *Let $\mu_j = \mathrm{E}(T_j | X_0 = j)$ be the expected return time to j. Then $\mu_j < \infty$ and the vector $v_j = 1/\mu_j$ is a stationary distribution. [D 3.4 p.103]*

$$v_j = \lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} (P^m)_{ij} \tag{2.12}$$

Note that all finite regular Markov chains are finite irreducible Markov chains, although the latter theorem is weaker. The previous theorem also holds for infinite irreducible positive recurrent Markov chains:

Theorem 2.4 (Fundamental limit theorem for ergodic Markov chains)**.** *There exists a unique positive stationary distribution v which is the limiting distribution of the Markov chain. [D 3.6 p.109]*

3 Hidden Markov Models (HMM)

A hidden Markov model consists of a Markov chain $X_0, \ldots, X_n, \ldots$ and another chain (not necessarily Markov) $Y_0, \ldots, Y_n, \ldots$ such that

$$\Pr(Y_k | Y_0, \ldots, Y_{k-1},\ X_0, \ldots, X_k) = \Pr(Y_k | Y_{k-1}, X_k). \tag{3.1}$$

Generally, Y_k are observed while the corresponding X_k are hidden. We assume for our purposes that the X_k have a finite state space, Y_k are discrete and the stronger condition that Y_k only depends on the corresponding X_k and nothing else. If the HMM parameters are given and Y_i are observed, there are different ways to infer suitable values for X_i:

- Find the sequence $X_0, \ldots, X_n$ maximizing the probability of observing $Y_0, \ldots, Y_n$ (the Viterbi algorithm)
- Find the joint distribution of $X_0, \ldots, X_n$ given observed $Y_0, \ldots, Y_n$ and the model, although in practice you would find a sample
- Find the marginal for each X_i given observed $Y_0, \ldots, Y_n$ and the model (the Forward-Backward algorithm)

[LN 2]

3.1 The Forward-Backward algorithm

The Forward algorithm. For $i = 0, \ldots, T$ compute $\pi(X_i|Y_0, \ldots, Y_{i-1})$:

1. Obtain $\pi(X_i|Y_0, \ldots, Y_i)$ from $\pi(X_i|Y_0, \ldots, Y_{i-1})$ using Bayes' formula

$$\pi(X_i|Y_0, \ldots, Y_i) \propto_{X_i} \pi(Y_i|X_i)\pi(X_i|Y_0, \ldots, Y_{i-1}) \tag{3.2}$$

2. Obtain $\pi(X_{i+1}|Y_0, \ldots, Y_i)$ from $\pi(X_i|Y_0, \ldots, Y_i)$ using the transition matrix.

$$\pi(X_{i+1}|Y_0, \ldots, Y_i) \propto_{X_{i+1}} \int_{X_i} \pi(X_{i+1}|X_i)\pi(X_i|Y_0, \ldots, Y_i)\, dX_i \tag{3.3}$$

The Backward algorithm. For $i = T - 1, \ldots, 0$ compute $\pi(Y_i \ldots, Y_T|X_i)$ recursively:

$$\pi(Y_i, \ldots, Y_T|X_i) = \pi(Y_i|X_i) \int_{X_{i+1}} \pi(Y_{i+1}, \ldots, Y_T|X_{i+1})\pi(X_{i+1}|X_i)\, dX_{i+1} \tag{3.4}$$

Combining the two. We can now find a sample from $\pi(X_0, \ldots, X_T|Y_0, \ldots, Y_T)$.

$$\pi(X_i|Y_0, \ldots, Y_T) \propto_{X_i} \pi(Y_i, \ldots, Y_T|X_i)\pi(X_i|Y_0, \ldots, Y_{i-1}) \tag{3.5}$$

4 Branching processes

Definition 4.1 (Branching process)**.** A *branching process* is a discrete Markov chain $Z_0, Z_1, \ldots, Z_n, \ldots$ where

- the state space is the non-negative integers
- $Z_0 := 1$ by definition
- 0 is an absorbing state
- $Z_n = X_1 + X_2 + \cdots + X_{Z_{n-1}}$ where the X_j are iid random variables with the offspring distribution

Connecting the Z_n individuals in a generation n with their offspring in generation $n + 1$ we get a tree illustrating the process. To avoid trivial cases, we assume that $\Pr(X_j = 0) > 0$ (the process can die) and $\Pr(X_j \leq 1) < 1$ (the process can branch). [D 4.1]

Lemma 4.1. *All non-zero states are transient.*

4.1 Expectation and variance

Let $\mu := \mathrm{E}(X_j)$ and $\sigma^2 := \mathrm{Var}(X_j)$.

Through conditioning, we obtain the recursive formula $\mathrm{E}(Z_n) = \mu\,\mathrm{E}(Z_{n-1})$. Since $\mathrm{E}(Z_0) = 1$, by induction

$$\mathrm{E}(Z_n) = \mu^n. \tag{4.1}$$

We say that the process is *subcritical* if $\mu < 1$, *critical* if $\mu = 1$ and *supercritical* if $\mu > 1$.

Lemma 4.2. *If $\mu < 1$ the probability of extinction is* 1.

Using similar reasoning to solve a recurrence equation, we obtain

$$\mathrm{Var}(Z_n) = \sigma^2\mu^{n-1}\sum_{k=0}^{n-1}\mu^k = \begin{cases} n\sigma^2 & \text{if } \mu = 1, \\ \sigma^2\mu^{n-1}\frac{\mu^n-1}{\mu-1} & \text{otherwise} \end{cases} \tag{4.2}$$

[D 4.2]

4.2 Extinction probability theorem

Theorem 4.3 (Extinction probability theroem)**.** *Let G be the pgf for the offspring distribution of a branching process. The probability of eventual extinction is the smallest positive root of*

$$s = G(s) \tag{4.3}$$

and in the subcritical and critical cases, the extinction probability is 1.

This is based on the property that the pgf for Z_n, $G_n(s)$, is equal to repeated application of G such that $G_n = G \circ \cdots \circ G$ where G is applied n times. [D 4.4]

5 Inference theory for processes

5.1 Discrete state space Markov chains

The parameters we would like to infer are the transition matrix P and the probability vector p for the initial state X_0.

One possibility is to fix some p and choose the prior

$$\pi(P) = \prod_{i=1}^{s}\mathrm{Dirichlet}(P_i;\alpha_i) \tag{5.1}$$

where s is the size of the state space, P_i is the ith row of P and α_i is a vector of length s of positive parameters. Often $\alpha = (1,\ldots,1)$. This means every row is inferred independently.

The resulting posterior becomes

$$\pi(P|\text{data}) = \prod_{i=1}^{s} \text{Dirichlet}(P_i; \alpha_i + c_i) \tag{5.2}$$

where c_i are the corresponding observed transition counts.

We can now make predictions. Assume we have observed $k+1$ steps $x_0, \ldots, x_k$ and would like to predict the distribution for the next step. It follows that

$$\pi(x_{k+1}|x_k, \ldots, x_0) = \int_{P_{x_k}} P_{x_k,x_{k+1}}\, \pi(P_{x_k}|x_k, \ldots, x_0)\, dP_{x_k} \tag{5.3}$$

where for each possible value of x_{k+1} this is the expectation of the posterior for $P_{x_k,x_{k+1}}$. Simulation of the chain can be done iteratively by simulating the next step, updating counts, simulating again and so on.

5.2 Hidden Markov Models

Assume we observe $X_0, \ldots, X_n$ and $Y_0, \ldots, Y_n$. We model with parameters P, p as above for X and a matrix Q with elements $Q_{ij} = \Pr(Y_k = j|X_k = i)$ of *emittance probability*.

Inference can now be done separately for P, p and Q, in the same way as before.

5.3 Branching processes

The parameter we would like to infer is the probability vector a for the offspring distribution. Assume we observe counts $y_1, \ldots, y_N$ from N realizations of the process, with a total of S births. One possible choice is then $Y \sim \text{Binomial}(n, p)$ with $p \sim \text{Beta}(\alpha, \beta)$. We obtain the posterior $p|\text{data} \sim \text{Beta}(\alpha + S, \beta + nN - S)$.

6 Markov chain Monte Carlo (MCMC)

MCMC is a set of important algorithms used for inference based on simulation. The goal is to simulate from some distribution. We construct Markov chains with the posterior as the limiting distribution to achieve this. [LN 5]

6.1 Laws of large numbers

Theorem 6.1 (Strong law of large numbers for samples)**.** *If $Y_1, \ldots, Y_m$ and Y are independent random variables from a distribution with finite mean, and if r is a bounded function, then almost surely*

$$\lim_{m\to\infty} \frac{r(Y_1) + \cdots + r(Y_m)}{m} = \text{E}(r(Y)) \tag{6.1}$$

Theorem 6.2 (Strong law of large numbers for Markov chains)**.** *If $X_0, X_1, \ldots$ is an ergodic Markov chain with stationary distribution ν, and if r is a bounded function, then almost surely*

$$\lim_{m\to\infty} \frac{r(X_1) + \cdots + r(X_m)}{m} = \mathrm{E}(r(X)) \tag{6.2}$$

where X has the stationary distribution ν.

Note that this is also applicable to continuous state spaces. In practice, early values may be atypical and one might improve accuracy by throwing away the start of the sequence before averaging. This sequence is called the *burn-in*. In addition, we have little to no information about the size of m required for good approximations, and may need to resort to heuristics such as trace plots. [LN 5.2]

6.2 Inference using MCMC

Assume we know the likelihood $\pi(y|\vartheta)$ and the prior $\pi(\vartheta)$ so that we can compute the posterior $\pi(\vartheta|y)$ up to a constant. With a sample $\vartheta_1, \ldots, \vartheta_m$ from the posterior we can approximate predictions by the strong LLN:

$$\pi(y_{\text{new}}|y) = \int_\vartheta \pi(y_{\text{new}}|\vartheta)\pi(\vartheta|y)\,d\vartheta \approx \frac{1}{m}\sum_{i=1}^{m} \pi(y_{\text{new}}|\vartheta_i) \tag{6.3}$$

MCMC provides a sequence so that the above holds when $m \to \infty$ by the strong LLN for ergodic Markov chains. This is important! Without ergodicity, the approximations are garbage. The formula generalizes; for some function g,

$$\mathrm{E}(g(\vartheta)) = \int_\vartheta g(\vartheta)\pi(\vartheta|y)\,d\vartheta \approx \frac{1}{m}\sum_{i=1}^{m} g(\vartheta_i) \tag{6.4}$$

6.3 Metropolis-Hastings

Assume the prior $\pi(\vartheta)$ is known up to a constant. Given a *proposal function* $q(\vartheta_{\text{new}}|\vartheta)$ which for any ϑ provides a distribution for a new ϑ_{new}. Define for ϑ, ϑ_{new} the acceptance probability

$$a_{\vartheta,\vartheta_{\text{new}}} = \min\left\{1, \frac{\pi(\vartheta_{\text{new}})q(\vartheta|\vartheta_{\text{new}})}{\pi(\vartheta)q(\vartheta_{\text{new}}|\vartheta)}\right\} \tag{6.5}$$

The MH algorithm. From an initial ϑ_0, generate $\vartheta_1, \vartheta_2, \ldots$ by

1. proposing a new ϑ_{new} based on the old from the proposal function q, then
2. accepting with probability $a_{\vartheta,\vartheta_{\text{new}}}$ else repeating ϑ.

If this defines an ergodic Markov chain, its unique stationary distribution is $\pi(\vartheta)$. Note that if q is symmetric it disappears from the expression for the acceptance probability. The proof for the MH algorithm is based on proving time reversibility relative to the target distribution (which implies that the target distribution is in fact the limiting distribution). [D 5.2]

The theory is unchanged if the state space is continuous or even multivariate with a mix. The proposal distribution can be almost freely chosen as long as the chain becomes ergodic, but the choice will greatly influence the rate of convergence and the accuracy of the results.

Proposition 6.3 (Sufficient case for ergodic MH chains)**.** *If the target density is positive on the same set in which the proposal function generates proposals, and the proposal function is ergodic, then the Metropolis-Hastings chains is ergodic.*

6.3.1 Gibbs sampling

Gibbs sampling is a version of MH with a special type of proposal function. For each component of $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ use the conditional where all but one component are fixed (using the last value). This can be viewed as the sample iteratively updating each dimension separately. [D 5.3]

Lemma 6.4. *The acceptance probability in Gibbs sampling is always* 1.

6.3.2 Ising model

An application of Gibbs sampling, the Ising model is described by a simple set of postulates:

- The configurations $\sigma = (\sigma_{11}, \ldots, \sigma_{nn})$ consist of nodes in a grid, where in each node v the configuration can have two values such that $\sigma_v = \pm 1$.
- The energy of a configuration is defined as $E(\sigma) = -\sum_{v \sim w} \sigma_v \sigma_w$ where the sum is over all neighbour pairs v, w.
- The Gibbs distribution on the set of all configurations has pmf (where $\beta \in \mathbb{R}$)

$$\pi(\sigma) = \frac{\exp(-\beta E(\sigma))}{\sum_\tau \exp(-\beta E(\tau))} \tag{6.6}$$

We simulate by updating the components one at a time. To update, say, σ_k we determine the two possible states σ^+, σ^- where this node is plus or minus one respectively. The conditional distribution to update to, say σ^+, becomes

$$\Pr(\sigma^+ | \text{fixed comp.}) = \frac{\Pr(\sigma^+)}{\Pr(\sigma^+) + \Pr(\sigma^-)} = \frac{1}{1 + \exp(-\beta(E(\sigma^-) - E(\sigma^+)))} \tag{6.7}$$

However, the difference has a large number of common terms so that

$$E(\sigma^-) - E(\sigma^+) = \sum_{i \sim k} \sigma_i + \sum_{j \sim k} \sigma_j = 2 \sum_{i \sim k} \sigma_i \tag{6.8}$$

and therefore the conditional distribution only requires local information.

6.3.3 Perfect sampling

Given an ergodic Markov chain with finite sample space of size k and limiting distribution π, we would like to prove that given some n, X_n has the limiting distribution. The method is to prove that X_n is independent of X_0. We construct k dependent (coupled) Markov chains that are marginally chains as above, starting at the k possible starting states. If they all have identical values at X_n, we are done.

To couple the chains, at each step we simulate from the same unif(0, 1) when selecting the next state according to the transition matrix.

7 Poisson processes

Definition 7.1 (Counting process)**.** A *counting process* $\{N_t\}_{t \geq 0}$ is a stochastic process indexed by $\mathbb{R}_0^+$ where the state space is the non-negative integers and $0 \leq s \leq t \implies N_s \leq N_t$.

Definition 7.2 (Poisson process)**.** A *Poisson process* is a particular counting process $\{N_t\}_{t \geq 0}$ with rate parameter $\lambda > 0$ satsifying

- $N_0 = 0$
- $N_t \sim \text{Poisson}(\lambda t)\ \forall t > 0$
- Stationary increments: $N_{t+s} - N_s \overset{d}{=} N_t$
- Independent increments: $N_t - N_s$ and $N_r - N_q$ are independent when the time spans do not overlap, i.e. $0 \leq q < r \leq s < t$

There are several equivalent definitions for the Poisson process. Other possibilities are constraints on the distribution of N_t or instead of counting taking sums of inter-arrival times (seen below). [D 6.1]

7.1 Memorylessness

Definition 7.3 (Memorylessness)**.** A random variable X is *memoryless* if

$$\Pr(X > s + t | X > s) = \Pr(X > t) \tag{7.1}$$

Of note is that the Exponential distribution is the only continuous distribution on $\mathbb{R}^+$ that is memoryless. [D 6.2]

For the Poisson process, the definition implies memorylessness. In other words, it does not matter when we begin counting. Given a Poisson process $\{N_t\}_{t\geq 0}$ with rate λ and $s \geq 0$, let $M_t := N_{t+s} - N_s \; \forall t \geq 0$. Then $\{M_t\}_{t\geq 0}$ is a Poisson process with rate λ.

7.2 Properties

Let X_n be the *inter-arrival time* between arrival n and $n-1$. Then, for all $n \geq 1$:

$$X_n \sim \text{E}xp(\lambda) \tag{7.2}$$

Let $S_n = X_1 + \cdots + X_n$ be the *arrival time* of the nth arrival, where the X_i are independently as above:

$$S_n \sim \text{Gamma}(n.\lambda) \tag{7.3}$$

Let $M = \min\{X_1, \ldots, X_n\}$, where the X_i are independently as above:

$$M \sim \text{Exp}(\lambda_1 + \cdots + \lambda_k) \tag{7.4}$$

$$\Pr(M = X_k) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_k} \tag{7.5}$$

Let $S_1, S_2 \ldots$ be the arrival times as above. Conditional on $N_t = n$, we have that the joint pdf for $S_1, \ldots, S_n$ is uniform on $0 < s_1 < \cdots < s_n < t$. To simulate a Poisson process, we can simulate N_t, simulate arrival times on $[0, t]$ and then order them. [D 6.2] [D 6.5]

7.3 Variants

7.3.1 Thinning

Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate λ. Assume each arrival is "marked" as type k out of n possible types with probability p_k. (The law of total probability implies $p_1 + \cdots + p_n = 1$). Let $N_t^{(k)}$ be the count of arrivals of type k at time t. Then $\{N_t^{(k)}\}_{t\geq 0}$ is a Poisson process with rate $p_k\lambda$ and the processes $\{N_t^{(1)}\}_{t\geq 0}, \ldots, \{N_t^{(n)}\}_{t\geq 0}$ are independent. [D 6.4]

7.3.2 Superposition

Assume $\{N_t^{(1)}\}_{t\geq 0}, \ldots, \{N_t^{(n)}\}_{t\geq 0}$ are n independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_n$. Define, for $t > 0$, $N_t = N_t^{(1)} + \cdots + N_t^{(n)}$. Then $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda := \lambda_1 + \cdots + \lambda_n$. [D 6.4]

7.3.3 Spatial Poisson process

Definition 7.4 (Spatial Poisson process)**.** A collection of random variables $\{N_A\}_{A \subseteq \mathbb{R}^d}$ is a spatial Poisson process with rate λ if

- for each bounded set $A \subseteq \mathbb{R}^d$, $N_A \sim \text{Poisson}(\lambda|A|)$
- wherever $A \subset B$, $N_A \leq N_B$
- whenever $A \cap B = \varnothing$, N_A and N_B are independent

A spatial Poisson process is simulated by first simulating the total at time t and then simulating the points uniformly inside. [D 6.6]

7.3.4 Non-homogeneous Poisson processes

Definition 7.5 (Non-homogeneous Poisson process)**.** A counting process $\{N_t\}_{t \geq 0}$ is a non-homogeneous Poisson process with intensity function $\lambda(t)$ if

- $N_0 = 0$
- for $t > 0$, $N_t \sim \text{Poisson}(\int_0^t \lambda(x)\, dx)$
- it has independent increments

[D 6.7]

8 Continuous-time Markov chains

Definition 8.1 (Continuous-time Markov chain)**.** A *continuous-time* stochastic process $\{X_t\}_{t \geq 0}$ with discrete state space S is a continuous-time Markov chain if

$$\Pr(X_{t+s} = j | X_s = i, X_u, 0 \leq u \leq s) = \Pr(X_{t+s} = j | X_s = i) \tag{8.1}$$

Definition 8.2 (Time homogeneous)**.** The chain is *time homogeneous* if for all $s, t > 0$ and all $i, j \in S$

$$\Pr(X_{t+s} = j | X_s = i) = \Pr(X_t = j | X_0 = i) \tag{8.2}$$

Definition 8.3 (Transition function)**.** Let the *transition function* $P(t)$ be the matrix function such that $P(t)_{ij} = \Pr(X_t = j | X_0 = i)$.

Theorem 8.1 (Chapman-Kolmogorov equation)**.** *For the transition function we have*

$$P(s + t) = P(s)P(t) \tag{8.3}$$

In addition, we note that necessarily $P(0) = I$.

8.1 Properties

8.1.1 Holding times

Definition 8.4 (Holding times)**.** The holding time T_i is the time the continuous-time Markov chain started in i stays in i before moving to a different state, so that for all $s > 0$

$$\Pr(T_i > s) = \Pr(X_u = i, 0 \leq u \leq s) \tag{8.4}$$

The distribution of T_i is memoryless and therefore Exponential. Define q_i so that $T_i \sim \text{Exp}(q_i)$.

The expected holding time in i becomes $1/q_i$. Note that it is possible that $q_i = 0$ which implies the state is absorbing, or that $q_i = \infty$ for an explosive process.

8.1.2 The embedded chain

Definition 8.5 (Embedded chain)**.** The Markov chain obtained by listing visited states is called the *embedded chain* and its transition matrix denoted $\tilde{P}$.

Note that $\tilde{P}$ has zeros along the diagonal. A continuous-time Markov chain is completely determined by expected holding times $(1/q_1, \ldots, 1/q_k)$ and $\tilde{P}$.

8.1.3 Transition rates

One might consider a continuous-time Markov chain as a set of $k \times (k-1)$ independent "alarm clocks". When in state i, await the first alarm and move to the corresponding j.

For states i and j $(i \neq j)$ let q_{ij} be the rate of an Exponential random variable representing the time until an alarm. The time until the first alarm in state i is the minimum of the Exponential times, itself an Exponential random variable with rate $q_i = q_{i,1} + \cdots + q_{i,i-1} + q_{i,i+1} + \cdots + q_{i,k}$. That is, the rate of leaving a state is equal to the sum of the rates of moving out. The probability to move to a specific state (the entries in the embedded chain) are

$$\tilde{P}_{ij} = \frac{q_{ij}}{q_i} \tag{8.5}$$

8.2 The infinitesimal generator

Definition 8.6 (The infinitesimal generator)**.** To relate $P(t)$ to transition rates, the derivative at zero is used. Assuming $P(t)$ is differentiable we have

$$Q := P'(0) = \begin{bmatrix} -q_1 & q_{12} & \cdots & q_{1k} \\ q_{21} & -q_2 & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & -q_k \end{bmatrix} \tag{8.6}$$

where Q is the *infinitesimal generator* matrix. The state space need not be finite, only discrete. The rows of Q must sum to zero.

8.2.1 Kolmogorov Forward Backward

It follows that $\forall t \geq 0$, $P'(t) = P(t)Q = QP(t)$. We obtain the differential equations

$$P'(t)_{ij} = -P_{ij}q_j + \sum_{k \neq j} P_{ik}(t)q_{kj} \tag{8.7}$$

$$P'(t)_{ij} = -q_i P_{ij} + \sum_{k \neq i} q_{ik}P_{kj}(t) \tag{8.8}$$

Given the condition $P(0) = I$, the solution is

$$P(t) = e^{tQ} \tag{8.9}$$

where the matrix exponential is used. If Q is diagonalizable so that $Q = SDS^{-1}$ the right-hand side becomes the simpler expression $Se^{tD}S^{-1}$.

8.3 Limiting and stationary distributions

Recall the definition of limiting distribution v for all states i, j

$$\lim_{t \to \infty} P_{ij}(t) = v_j \tag{8.10}$$

Recall the definition of stationary distribution v: $\forall t \geq 0 : vP(t) = v$. A limiting distribution is a stationary distribution but not necessarily vice versa.

Unlike discrete time, no periodicity exists: if $P_{ij}(t) > 0$ for some $t > 0$ then $P_{ij}(t) > 0$ for all $t > 0$.

For a finite continuous-time Markov chain with finite holding time parameters, there are two possibilities: irreducible or absorbing.

8.3.1 Irreducible case

Recall that the chain is irreducible if $\forall i, j \; \exists t > 0 : P_{ij}(t) > 0$.

Theorem 8.2 (Fundamental limit theorem, continuous-time variant)**.** *Let $\{X_t\}_{t \geq 0}$ be a finite, irreducible continuous-time Markov chain with transition function $P(t)$. Then there exists a unique stationary distribution vector v which is also the limiting distribution.*

Corollary 8.3. *If v is a stationary distribution and Q is the infinitesimal generator*

$$vQ = 0 \tag{8.11}$$

8.3.2 Absorbing case

Assume $\{X_t\}_{t\geq 0}$ is a continuous-time Markov chain with k states. Assume further that the last state, called a is absorbing while the rest are transient.

The row corresponding to a in Q will be zero, as $q_a = 0$.

$$Q = \begin{bmatrix} V & * \\ 0 & 0 \end{bmatrix} \tag{8.12}$$

Let F be the $(k-1)\times(k-1)$-matrix so that F_{ij} is the expected time spent in state j when the process starts in i. It can be shown that

$$F = -V^{-1} \tag{8.13}$$

Note that, if the chain starts in i, the expected time until absorption is the sum of the ith row of F.

8.3.3 Embedded chain

Recall the embedded chain with transition matrix $\tilde{P}$. Stationary distributions for the embedded and continuous-time chains are generally not the same. However, there exists a simple relationship; a probability vector v is a stationary distribution for a continuous-time Markov chain iff ψ is a stationary distribution for the embedded chain where

$$\psi_j = Cv_jq_j \tag{8.14}$$

for an appropriate normalizing constant C.

8.4 Balance conditions

Definition 8.7 (Global balance)**.** For a continuous-time Markov chain, the long term rate of movement *into* a state must correspond to the long term rate of movement *out of* the state. This is called *global balance*, which we know holds for stationary distributions.

$$\sum_{i\neq j} \pi_i q_{ij} = \pi_j q_j \iff \pi Q = 0 \tag{8.15}$$

This generalizes; if $\mathcal{A}$ is a set of states, then the rates are conserved:

$$\sum_{i\in\mathcal{A}}\sum_{j\notin\mathcal{A}} \pi_i q_{ij} = \sum_{i\in\mathcal{A}}\sum_{j\notin\mathcal{A}} \pi_j q_{ji} \tag{8.16}$$

Definition 8.8 (Local balance)**.** The continuous-time Markov chain with unique stationary distribution π is said to be *time reversible* if, for all state pairs i, j

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{8.17}$$

i.e. the observed rate from i to j is equal to the one from j to i. If a probability vector satisifies this condition, it is the unique stationary distribution.

8.4.1 Markov processes as trees

Assume we have an irreducible continuous-time Markov chain with a *tree* (cycle-free undirected) transition graph. It follows from (generalized) global balance that the process is time reversible. Note that there exist time-reversible chains that are not trees.

8.5 Birth-and-death processes

Definition 8.9 (Birth-and-death process)**.** A *birth-and-death* process is a continuous-time Markov chain where the state space is the set of non-negative integers and transitions only occur to neighbouring integers. Note that it can be represented by a tree with a single branch and thus is time-reversible.

Let the rate of births be λ_i ($i \to i+1$) and the rate of deaths be μ_i ($i \to i-1$).

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -\mu_1 - \lambda_1 & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -\mu_2 - \lambda_2 & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{8.18}$$

The balance condition implies $\pi_i \lambda_i = \pi_{i+1}\mu_{i+1}$ which allows us to derive a recursive formula. Given that $\sum_{k=0}^{\infty} \prod_{i=1}^{k} \lambda_{i-1}/\mu_i < \infty$, the unique stationary distribution is

$$\pi_k = \pi_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \qquad \pi_0 = \left(\sum_{k=0}^{\infty} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \tag{8.19}$$

8.5.1 Queues

Definition 8.10 (Queue)**.** A *queue* is a stochastic process with the non-negative integers as state space. It is concisely described as A/B/c where A is the arrival process, B the service time process and c the number of servers.

Theorem 8.4 (Little's formula)**.**

$$L = \lambda W \tag{8.20}$$

where L is the long term average count of clients, λ is the long term rate of arrival and W is the long term average time in the system for clients.

A queue need not be Markov or memoryless, but the process notation M indicates exponential times. Consider a M/M/c queue, with arrival rate λ and service rate μ. If $\{X_t\}_{t \geq 0}$ is the number of clients at time t, this is a birth-and-death process. The death rate is equal to the number of active servers times the service rate.

In the M/M/1 case, we have a constant birth rate λ and death rate μ. The formulas above show that the limiting distribution is a Geometric distribution with parameter

$1 - \lambda/\mu$. Similarly, in the M/M/∞ case the limiting distribution is a Poisson distribution with parameter λ/μ. In general, for a M/M/c queue:

$$\pi_k = \begin{cases} \frac{\pi_0}{k!}\left(\frac{\lambda}{\mu}\right)^k & k = 1, \dots, c \\ \frac{\pi_0}{c^{k-c}c!}\left(\frac{\lambda}{\mu}\right)^k & k \geq c \end{cases} \tag{8.21}$$

8.5.2 Poisson subordination

Let $Y_0, Y_1, \dots$ be a discrete time discrete state space Markov chain with transition matrix R. Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate λ. Then $X_t = Y_{N_t}$ is a continuous-time Markov chain with

$$P(t) = \sum_{k=0}^{\infty} R^k \,\text{Poisson}(k; t\lambda) \tag{8.22}$$

Conversely, assume X_t is a continuous-time Markov chain with generator Q. Define

$$R = \frac{1}{\lambda}Q + I \tag{8.23}$$

where λ is chosen so R is positive. Then the above equation hols, so that X_t can be described as a Poisson subordination as above. In some cases, these calculations give better approximations. The discrete time and continuous-time chains have the same stationary distribution.

9 Brownian motion

Definition 9.1 (Standard Brownian motion)**.** *Standard Brownian motion* is a continuous-time stochastic process $\{B_t\}_{t\geq 0}$ with the following defining properties:

- $B_0 = 0$
- $B_t \sim \text{Normal}(0, t)$ for $t > 0$ (i.e. the standard deviation is $\sqrt{t}$)
- $B_{t+s} - B_s \sim \text{Normal}(0, t)$ for $t, s > 0$ (stationary increments)
- for $0 \leq q < r \leq s < t$, $B_t - B_s$ is independent from $B_r - B_q$ (independent increments)
- the function $t \mapsto B_t$ is almost surely continuous

9.1 Simulation

Given time points $t_1 < t_2 < \cdots < t_n$, we can write $B_{t_i} = B_{t_{i-1}} + (B_{t_i} - B_{t_{i-1}}) = B_{t_{i-1}} + Z_i$ where $Z_i \sim \text{Normal}(0, t_i - t_{i-1})$. Setting $t_0 = 0$, we get $B_{t_n} = \sum_{i=1}^{n} Z_i$. Thus, a

good way to simulate $t \mapsto B_t$ on $t \in [0, a]$ is to let $t_i = ai/n$, simulate independent $Z_i \sim \text{Normal}(0, a/n)$ and compute the sums as above. Note that each Z_i is a scaling of a standard normal by a factor $\sqrt{a/n}$.

Replacing these independent Normal random variables with differently distributed ones, with the same expectation and variance, will give approximately the same result by the Central limit theorem. This property can be used to study the limiting behaviour of a random walk. When the resolution goes to infinity the processes are exactly the same, an example of the *invariance principle*.

9.1.1 Zoom

Consider the first step of N total during t: $Z_1 \sim \text{Normal}(0, t/N)$. Now, we would like to zoom such that both Z_1' and $Z_1 - Z_1'$ are $\text{Normal}(0, t/(2N))$. (The second condition implies $Z_1 \sim \text{Normal}(Z_1', t/(2N))$.) The correct conditional distribution becomes

$$Z_1' | Z_1 \sim \text{Normal}\left(\frac{1}{2}Z_1, \frac{1}{\frac{1}{t/(2N)} + \frac{1}{t/(2N)}}\right) = \text{Normal}\left(\frac{1}{2}Z_1, \frac{t/2}{2N}\right) \tag{9.1}$$

A fractal behaviour is produced, such that the process is invariant under the zoom. These jagged paths, no matter the size of the intervals, lead to the intuition that the paths are *nowhere differentiable*.

9.2 Gaussian processes

Definition 9.2 (Gaussian process)**.** A *Gaussian process* is a continuous-time stochastic process $\{X_t\}_{t \geq 0}$ with the property that $\forall n \geq 1$ and $0 \leq t_1 < t_2 < \cdots < t_n$, $X_{t_1}, \ldots, X_{t_n}$ has a multivariate normal distribution. Thus a Gaussian process is completely determined by its mean function $\text{E}(X_t)$ and covariance function $\text{Cov}(X_s, X_t)$.

Brownian motion is a Gaussian process, as we can show that any linear combination of Brownian motion is normally distributed. A Gaussian process $\{X_t\}_{t \geq 0}$ is Brownian motion if

- $X_0 = 0$
- $E(X_t) = 0 \; \forall t$
- $\text{Cov}(X_s, X_t) = \min\{s, t\} \; \forall s, t \quad$ (implies $\text{Var}(X_t) = t$)
- $t \mapsto X_t$ is almost surely continuous.

It is necessary to prove that this process has stationary and independent increments. Note that if the covariance of two Normal variables is zero, then they are independent.

9.3 Transformations

Some useful transformations of Brownian motion, proven by the rules above:

- $\{-B_t\}_{t\geq 0}$
- $\{B_{t+s} - B_s\}_{t\geq 0}$ for any $s \geq 0$
- $\left\{\frac{1}{\sqrt{a}} B_{at}\right\}_{t\geq 0}$ for any $a > 0$
- the process $\{X_t\}_{t\geq 0}$ where $X_0 = 0$ and $X_t = tB_{1/t}$ for $t > 0$
- $\{x + B_t\}_{t\geq 0}$ "Brownian motion started at x"

9.4 Properties

9.4.1 First hitting time

Definition 9.3 (First hitting time)**.** The *first hitting time* T_a is defined as $T_a := \min\{t : B_t = a\}$.

It can be shown that B_{t+T_a} is Brownian motion, i.e. T_a is a stopping time. For $a > 0$:

$$\Pr(B_t > a | T_a < t) = \Pr(B_{t-Ta} > 0) = \frac{1}{2} \tag{9.2}$$

by the symmetry of the Normal distribution. By the definition of conditional probability:

$$\Pr(T_a < t) = 2\Pr(B_t > a) \tag{9.3}$$

Therefore, the density for T_a (for $a \neq 0$) is

$$\pi(t) = \frac{|a|}{\sqrt{2\pi t^3}} \exp\left(-\frac{a^2}{2t}\right) \tag{9.4}$$

and $T_a \sim$ Inverse-Gamma $\left(\frac{1}{2}, \frac{a^2}{2}\right) \iff \frac{1}{T_a} \sim$ Gamma $\left(\frac{1}{2}, \frac{a^2}{2}\right)$

9.4.2 Maximum

Define the maximum attained $M_t := \max_{0\leq s\leq t} B_t$. However, a certain maximum simply means the first hitting time for that value has been passed. For $a < 0$:

$$\Pr(M_t > a) = \Pr(T_a < t) = 2\Pr(B_t > a) = \Pr(|B_t| > a) \tag{9.5}$$

Thus M_t has the same distribution as $|B_t|$.

9.4.3 Zeros

Theorem 9.1 (Zeros of BM)**.** *The probability that Brownian motion has at least one zero in* (r, t) *with* $0 \leq r < t$ *is*

$$z_{r,t} = \frac{2}{\pi} \arccos \sqrt{\frac{r}{t}} \tag{9.6}$$

Let L_t be the last zero in $(0, t)$. Then

$$\Pr(L_t \leq x) = 1 - z_{x,t} = \frac{2}{\pi} \arcsin \sqrt{\frac{x}{t}} \tag{9.7}$$

which astonishingly means the last zero is distributed as $x/t \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$.

9.5 Extensions

BM with a drift. For real μ and $\sigma > 0$, define the Gaussian process X_t as $X_t := \mu t + \sigma B_t$.

Brownian bridge. Define a Gaussian process X_t by conditioning Brownian motion B_t on $B_1 = 0$. Then X_t is a Brownian bridge. In fact, a Brownian bridge can be expressed as $B_t - tB_1$ which makes it easy to simulate.

Geometric BM. The stochastic process $G_t = G_0 \exp(\mu t + \sigma B_t)$ where $G_0 > 0$ is called geometric BM with drift parameter μ and variance σ^2. One can show that $\text{E}(G_t) = G_0 e^{t(\mu+\sigma^2/2)}$ and $\text{Var}(G_t) = G_0^2 e^{2t(\mu+\sigma^2/2)}(e^{t\sigma^2} - 1)$

In addition, $\ln G_t = \ln G_0 + \mu t + \sigma B_t$ is a Gaussian process with expectation $\ln G_0 + \mu t$ and variance $t\sigma^2$.

9.6 Martingales

Definition 9.4 (Martingales)**.** A stochastic process $\{Y_t\}_{t \geq 0}$ is a *martingale* if for all $t \geq 0$:

- $\text{E}(Y_t | Y_r, 0 \leq r \leq s) = Y_s$ for $0 \leq s \leq t$
- $\text{E}(|Y_t|) < \infty$

Definition 9.5 (Martingales with respect to process)**.** $\{Y_t\}_{t \geq 0}$ is a *martingale with respect to* $\{X_t\}_{t \geq 0}$ if for all $t \geq 0$:

- $\text{E}(Y_t | X_r, 0 \leq r \leq s) = Y_s$ for $0 \leq s \leq t$
- $\text{E}(|Y_t|) < \infty$

Brownian motion is a martingale. Geometric BM is not, but with the correction $e^{-(\mu+\sigma^2/2)t} G_t$ is a martingale with respect to standard BM.

List of Theorems