An Intuitive Introduction For Understanding and Solving Stochastic Differential Equations

Chris Rackauckas

May 28, 2017

Abstract

Stochastic differential equations (SDEs) are a generalization of deterministic differential equations that incorporate a "noise term". These equations can be useful in many applications where we assume that there are deterministic changes combined with noisy fluctuations. Ito's Calculus is the mathematics for handling such equations. In this article we introduce stochastic differential equations and Ito's calculus from an intuitive point of view, building the ideas from relatable probability theory and only straying into measure-theoretic probability (defining all concepts along the way) as necessary. All of the proofs are discussed intuitively and rigorously: step by step proofs are provided. We start by reviewing the relevant probability needed in order to develop the stochastic processes. We then develop the mathematics of stochastic processes in order to define the Poisson Counter Process. We then define Brownian Motion, or the Wiener Process, as a limit of the Poisson Counter Process. By doing the definition in this manner, we are able to solve for many of the major properties and theorems of the stochastic calculus without resorting to measure-theoretic approaches. Along the way, examples are given to show how the calculus is actually used to solve problems. After developing Ito's calculus for solving SDEs, we briefly discuss how these SDEs can be computationally simulated in case the analytical solutions are difficult or impossible. After this, we turn to defining some relevant terms in measure-theoretic probability in order to develop ideas such as conditional expectation and martingales. The conclusion to this article is a set of four applications. We show how the rules of the stochastic calculus and some basic martingale theory can be applied to solve problems such as option pricing, genetic drift, stochastic control, and stochastic filtering. The end of this article is a cheat sheet that details the fundamental rules for "doing" Ito's calculus, like one would find on the cover flap of a calculus book. These are the equations/properties/rules that one uses to solve stochastic differential equations that are explained and justified in the article but put together for convenience.

Contents

Acknowledgements

This article was based on the course notes of Math 271-C, stochastic differential equations, taught by Xiaohui Xie at University of California, Irvine. It is the compilation of notes from Shan Jiang, Anna LoPresti, Yu Liu, Alissa Klinzmann, Daniel Quang, Hannah Rubin, Jaleal Sanjek, Kathryn Scannell, Andrew Schaub, Jienian Yang, and Xinwen Zhang.

1 Introduction

Newton's calculus is about understanding and solving the following equation:

$$\frac{d}{dt}g(x) = g'(x)\frac{dx}{dt}$$

The purpose of this paper is to generalize these types of equations in order to include noise. Let W_t be the Wiener process (aka Brownian motion whose properties will be determined later). We write a stochastic differential equation (SDE) as

$$dx = f(x)dt + \sigma(x)dW_t$$

which can be interpreted as "the change in x is given by deterministic changes f with noise of variance σ". For these equations, we will need to develop a new calculus. We will show that Newton's rules of calculus will not hold:

$$dg(x) \neq g'(x)dx = g'(x)f(x)dt + g'(x)\sigma(x)dt$$

Instead, we can use *Ito's calculus* (or other systems of calculus designed to deal with SDEs). In Ito' s calculus, we use the following equation to find $dg(x)$:

$$\begin{aligned} dg(x) &= g'(x)dx + \frac{1}{2}g''(x)\sigma^2(x)dt \\ &= g'(x)f(x)dt + g'(x)\sigma(x)dt + \frac{1}{2}g''(x)\sigma^2(x)dt \end{aligned}$$

If we let $\rho(x,t)$ be the distribution of x at time t, we can describe the evolution of this distribution using the PDE known as the Kolmogorov equation:

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x)\rho(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma^2(x)\rho(x,t)\right].$$

The we can understand the time development of differential equations with noise terms by understanding their probability distributions and how to properly perform the algebra to arrive at solutions.

1.1 Outline

This article is structured as follows. We start out with a review of probability that would be encountered in a normal undergraduate course. These concepts are then used to build the basic theory of stochastic processes and importantly the Poisson counter process. We then define the Wiener process, Brownian motion, as a certain limit of the Poisson counter process. Using this definition, we derive the basic properties, theorems, and rules for solving SDEs, known as the stochastic calculus. After we have developed the stochastic calculus, we develop some measure-theoretic probability ideas that will be important for defining conditional expectation, an idea central to fucture estimation of stochastic processes and martingales. These properties are then applied to systems that may be of interest to the stochastic modeler. The first of which is the European option market where we use our tools to derive the Black-Scholes equation. Next we dabble in some continuous probability models for population genetics. Lastly, we look at stochastic control and filtering problems that are central to engineering and many other disciplines.

2 Probability Review

This chapter is a review of probability concepts from an undergraduate probability course. These ideas will be useful when doing the calculations for the stochastic calculus. If you feel as though you may need to review some probability before continuing, we recommend Durrett's *Elementary Probability for Applications*, or at a slightly higher level which is more mathematical, Grinstead and Snell's *Introduction to Probability*. Although a full grasp of probability is not required, it is recommended that you are comfortable with most of the concepts introduced in this chapter.

2.1 Discrete Random Variables

2.1.1 Example 1: Bernoulli Trial

A Bernoulli trial describes the experiment of a single coin toss. There are two possible outcomes in a Bernoulli trial: the coin can land on heads or it can land on tails. We can describe this using S, the set of all possible outcomes:

$$S = \{H, T\}$$

Let the probability of the coin landing on heads be $\Pr(H) = p$. Then, the probability of the coin landing on tails is $\Pr(T) = 1 - p$. Let$X$ be a random variable. This is a "function" that maps $S \to \mathbb{R}$. It describes the values associated with each possible outcome of the experiment. For example, let $X = 1$ if we get heads, and $X = 0$ if we get tails, then

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = H \\ 0 & \text{if } \omega = T \end{cases}$$

We define$\mathbb{E}[X]$ as the mean of X, or the *expectation* of X. To find $\mathbb{E}[X]$, we take all the possible values of X and weight them by the probability of X taking each of these values. So, for the

Bernoulli trial:

$$\begin{aligned}\mathbb{E}[X] &= \Pr(H) \cdot 1 + \Pr(T) \cdot 0 \\ &= p\end{aligned}$$

We use the probability distribution of X to describe all the probabilities of X taking each of its values. There are only two possible outcomes in the Bernoulli trial and thus we can write the probability distribution as

$$\begin{aligned}P(X = 1) &= P(H) = p \\ P(X = 0) &= P(T) = 1 - p\end{aligned}.$$

Define $\mathbb{V}[X]$ as the *variance* of X. This is a measure of how much the values of X diverge from the expectation of X on average and is defined as

$$\mathbb{V}[X] = \sigma_x^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

For the Bernoulli Trial, we get that

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X - p)^2\right] \\ &= \Pr(X = 1)(1 - p)^2 + \Pr(X = 0)(0 - p)^2 \\ &= p(1 - p)\end{aligned}$$

2.1.2 Example 2: Binomial Random Variable

This describes an experiment where we toss a coin n times. That is, it is a Bernoulli trial repeated n times. Each toss/repetition is assumed to be independent. The set of all possible outcomes includes every possible sequence of heads and tails:

$$S = \{HH \cdots H, TT \cdots T, HTHTT \cdots, ...\}$$

or equivalently

$$S = \{H, T\}^n$$

We may want to describe the how large the set S is, or the *cardinality* of the set S, represented by $|S|$, as the "number of things in S". For this example:

$$|S| = 2^n$$

For each particular string of heads and tails, since each coin flip is independent, we can calculate the probability of obtaining that particular string as

$$\Pr(s \in S) = p^k(1 - p)^{n-k}, \text{where } k\text{is the number of heads in } s$$

So for instance, $P(HH\cdots H) = p^n$. Say we want to talk about the probability of getting a certain number of heads in this experiment. Then let X be the random variable for the number of heads. We can describe the $range(X)$ as the possible values that X can take: $X \in \{0, 1, 2, ..., n\}$. Note: using "$\in$" is an abuse of notation since X is actually a function. Recall that the probability distribution of X is

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Using this probability distribution, we can calculate the expectation and variance of X. The expectation is

$$\begin{aligned}\mathbb{E}[X] = \mu_x &= \sum_{k=0}^{n} \Pr(X = k) \cdot k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot k \\ &= n \cdot p,\end{aligned}$$

while the variance is

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X-p)^2\right] \\ &= \sum_{k=0}^{n} \Pr(X = k)(k - np)^2 \\ &= n \cdot p(1-p).\end{aligned}$$

This is all solved using the Binomial theorem:

$$(a+b)^n = \sum_{k} \binom{n}{k} a^k b^{n-k}.$$

Note that if X_i is the Bernoulli random variable associated with the i^{th} coin toss, then

$$X = \sum_{i=1}^{n} X_i.$$

Using these indicator variables we can compute the expectation and variance for the Binomial trial more quickly. In order to do so, we use the following facts

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

and

$$\mathbb{V}[aX + bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] \text{ if } X \text{ and } Y \text{are independent}$$

to easily compute

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$$

and

$$\mathbb{V}[X] = \sum_{i=1}^{n} \mathbb{V}[X_i] = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

2.2 Probability Generating Functions.

A probability function $G(z)$ is a power series that can be used to determine the probability distribution of a random variable X. We define $G(z)$ as

$$G(z) := \mathbb{E}\left[z^X\right].$$

For example, the Bernoulli random variable has the generating function

$$\begin{aligned} G(z) = \mathbb{E}\left[z^X\right] &= \Pr(X=1)z^1 + \Pr(X=0)z^0 \\ &= pz + (1-p) \end{aligned}$$

while the Binomial random variable has the generating function

$$\begin{aligned} G(z) &= \sum_{k=0}^{n} \Pr(X=k) \cdot z^k \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot z^k \\ &= \sum_{k=0}^{n} \binom{n}{k} (pz)^k (1-p)^{n-k} \\ &= (pz + 1 - p)^n \end{aligned}$$

and thus by the Binomial thorem

$$(a+b)^n = \sum_{k} \binom{n}{k} a^k b^{n-k}.$$

we let $a = pz$ and $b = 1-p$ to get

$$G(z) = (pz + 1 - p)^n.$$

In general, where $X = \{x_1, x_2, ..., x_n\}$ and $\Pr(X = x_i) = p_i$:

$$G(z) = \sum_{i=1}^{n} z^{x_i} p_i$$

The probability generating function is special because it gives an easy way to solve for the probability that the random variable equals a specific value. To do so, notice that

$$G^{(k)}(0) = k!\, p_k,$$

and thus

$$\Pr(X = k) = \frac{G^{(k)}(z)|_{z=0}}{k!},$$

that is, the kth derivative of G evaluated at 0 gives a straight-forward way of finding p. Thus if we can solve for G then we can recover the probability distribution (which in some cases may be simple).

2.3 Moment Generating Functions

A moment generating function is a function that is used to recover the expectation of the powers (the moments) of X. A moment generating function may not exist for a given random variable, but when it does exist the moment generating function will be unique. We define the moment generating function $M(t)$ as

$$M(t) := \mathbb{E}\left[e^{tX}\right]$$

The *moments* of X are found by taking derivatives with respect to t of the moment generating function, evaluated at $t = 0$. The first derivative is the first moment, the second derivative is the second moment, etc. This gives us the expectations of X^k:

$$\begin{aligned} M^1(t)|_{t=0} &= \frac{d\mathbb{E}\left[e^{tX}\right]}{dt}|_{t=0} = \mathbb{E}\left[X\right] \\ &\vdots \\ M^{(k)}(t)|_{t=0} &= \frac{d^n\mathbb{E}\left[e^{tX}\right]}{dt^n}|_{t=0} = \mathbb{E}\left[X^{(k)}\right] \end{aligned}$$

For example, for the Binomial random variable, the moment generating function is:

$$\begin{aligned} M(t) = \mathbb{E}\left[e^{tX}\right] &= \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n}\binom{n}{k}\left(e^t p\right)^k(1-p)^{n-k} \\ &= (e^t p + (1-p))^n \end{aligned}$$

Thus we can set $t = 0$ and evaluate the derivative of the moment generating function at this point. This gives us the following:

$$M'(k)|_{t=0} = npe^t(e^t p + 1 - p)^{n-1}|_{t=0} = np = \mathbb{E}\left[X\right]$$

which is the the first moment of X, or just the expectation of X.

2.4 Continuous Time Discrete Space Random Variables

We can generalize these ideas to random variables with infinite time. To go from discrete to continuous time, start with a line segment divided into intervals of size Δt. Then let the interval size $\Delta t \to 0$. The number of intervals is $n = \frac{t}{\Delta t}$., so we can also think of this as letting the number of intervals $n \to \infty$.

Let's do this for the Binomial random variable. We can think of a continuous time random variable version in the following way: there is a coin toss within each interval, the probability of a "success" (the coin lands on heads) within each interval is $\lambda \Delta t$. Define X to be the number of successes within the interval (0,t). Thus the probability of k successes in the interval $(0, t)$ is

$$\Pr(X = k) = \binom{\frac{t}{\Delta t}}{k} (\lambda \Delta t)^k (1 - \lambda \Delta t)^{\frac{t}{\Delta t} - k}$$

The probability generating function can then be written as follows:

$$\begin{aligned} G(z) &= (zp + 1 - p)^n \\ &= (1 + p(1 - z))^n \\ &= (1 + \lambda(1 - z)\Delta t)^n \\ &= \left(1 + \frac{\lambda(1 - z)t}{n}\right)^n \end{aligned}$$

Recall that the definition of e^x is

$$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.$$

Thus we see that

$$\lim_{n \to \infty} G(z) = \lim_{n \to \infty} \left(1 + \frac{\lambda(1 - z)t}{n}\right)^n = e^{\lambda t(z-1)}$$

We can then evaluate the probability generating function to see that

$$\begin{aligned} &G(0) = \Pr(X = 0) = e^{-\lambda t} \\ &\vdots \\ &\Pr(X = k) = \frac{G^{(k)}(z)|_{z=0}}{k!} = \frac{\lambda t^k}{k!} e^{-\lambda t} \end{aligned}$$

So when $\Delta t \to 0$, $\Pr(X = k) = \frac{\lambda t^k}{k!} e^{-\lambda t}$ with $X \in \{0, 1, 2, ...\}$ (discrete but no upper limit). This is the Poisson distribution.

2.4.1 The Poisson Distribution

For the Poisson distribution, there is one parameter, λ, which tells the rate of success. Few things to note:

1. Events are independent of each other

   (a) Within a small interval Δt, the probability of seeing one event is $\lambda \Delta t$

2. The sum of all the probabilities is 1:

$$\sum_{k=0}^{\infty} \Pr(X = k) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k} e^{-\lambda t} = e^{\lambda t} e^{-\lambda t} = 1$$

3. A special property of the Poisson distribution is that the expectation and variance are the same:

   - The expectation is:

$$\mathbb{E}[X] = np = \frac{t}{\Delta t} \lambda \Delta t = \lambda t$$

   - and the variance is:

$$\mathbb{V}[X] = np(1-p) = \frac{t}{\Delta t} \lambda \Delta t (1 - \lambda \Delta t) = \lambda t (1 - \lambda t)$$

We might want to determine the probability of the next event occurring at some specified time. Suppose we start at time $t = 0$ and want to know the probability of the first event occurring at a specified time. In continuous time, we can determine this probability by considering the probability that an event will occur in some small interval starting at our specified time (and that no event occurs before this time interval). Let T be the time to the next event. Then,

$$P(T \in [t, t + \Delta t]) = \lambda e^{-\lambda t} \Delta t$$

where $e^{-\lambda t}$ is the probability that no event occurs before t and $\lambda \Delta t$ is the probability that the event occurs in the $(t, t + \Delta t)$ time window. This is an exponential distribution:

$$f(t) = \lambda e^{-\lambda t}$$

2.5 Continuous Random Variables

So far we have been talking about discrete random variables, first in discrete time and then in continuous time. We can also talk about continuous random variables. Let $X \in \mathbb{R}$ That is, X can take an infinite number of values. For example, consider the one-dimensional Gaussian density function.

2.5.1 The Gaussian Distribution

Take a random variable X. We denote that X is Gaussian distributed with mean μ and variance σ^2 squared by $X \sim N(\mu, \sigma^2)$. This distribution can be denoted by the following properties:

- Density function: $\rho(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Cumulative distribution function: $P(X \leq a) = \int_{-\infty}^{a} \rho(x)dx$
- Expectation: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\rho(x)dx = \mu$
- Variance: $\mathbb{V}[X] = \int_{-\infty}^{\infty} (x-\mu)^2\rho(x)dx = \sigma^2$

The way to read this is that the probability that X will take a value within a certain interval is given by $P\{x \in [x, x+dx]\} = \rho(x) \cdot dx$.

Recall that the nth moment of X is defined as

$$\mathbb{E}[x^p] = \int_{-\infty}^{\infty} x^p \rho(x)dx.$$

If we assume $\mu = 0$, the pth moment function of X can be simplified as

$$\begin{aligned} \mathbb{E}[x^p] &= \int_{-\infty}^{\infty} x^p \rho(x)dx \\ &= \int_{-\infty}^{\infty} x^p \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x^{p-1} x e^{-\frac{x^2}{2\sigma^2}} dx \end{aligned}$$

To solve this, we use integration by parts. We let $u = x^{p-1}$ and $dv = xe^{-\frac{x^2}{2\sigma^2}}$. Thus $du = (p-1)x^{p-2}$ and $v = -\sigma^2 e^{-\frac{x^2}{2\sigma^2}}$. Therefore we see that

$$\mathbb{E}[x^p] = \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} [(x^{p-1} e^{-\frac{x^2}{2\sigma^2}})\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} (p-1)x^{p-2} dx].$$

Notice that the constant term vanishes at both limits. Thus we get that

$$\begin{aligned} \mathbb{E}[x^p] &= \frac{\sigma^2(p-1)}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} x^{p-2} dx \\ &= \sigma^2(p-1)\mathbb{E}\left[x^{p-2}\right] \end{aligned}$$

Thus, using the base cases of the mean and the variance, we have a recursive algorithm for finding all of the further variances. Notice that since we assumed $\mu = 0$, we get that $\mathbb{E}[x^p] = 0$ for every odd p. For every even p, we can solve the recursive equation to get

$$\mathbb{E}[x^p] = \left(\frac{\sigma^2}{2}\right)^{\frac{p}{2}} \frac{p!}{(\frac{p}{2})!} = (p-1)!!\sigma^p$$

where the double factorial $a!!$ means to multiply only the odd numbers from 1 to a.

2.5.2 Generalization: The Multivariate Gaussian Distribution

We now generalize the Gaussian distribution to multiple dimensions. Let X be a vector of n variables. Thus we define $X \sim N(\mu, \Sigma)$ to be a random variable following the probability distribution

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

where $x, \mu \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$. Notice that $\mathbb{E}[x] = \mu$ and thus μ is a vector of the means while the variance is given by Σ:

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_1, X_2) & \ddots & & Cov(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ Cov(X_1, X_n) & Cov(X_2, X_n) & \dots & Var(X_n) \end{bmatrix} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \dots & \sigma_{X_1 X_n} \\ \sigma_{X_1 X_2} & \ddots & & \sigma_{X_2 X_n} \\ \vdots & & \ddots & \vdots \\ \sigma_{X_1 X_n} & \sigma_{X_2 X_n} & \dots & \sigma_{X_n}^2 \end{bmatrix}$$

and thus, since Σ gives the variance in each component and the covariance between components, it is known as the Variance-Covariance Matrix.

2.5.3 Gaussian in the Correlation-Free Coordinate System

Assume $\mu = \bar{0}$. Since Σ is positive definite and symmetric matrix, Σ has a guaranteed eigendecomposition

$$\Sigma = U \Lambda U^T$$

where Λ is a diagonal matrix of the eigenvalues and U is a matrix where the ith column is the ith eigenvector. Thus we notice that

$$\det(Q) = \det(U \Lambda U^T) = \det(U)^2 \det(\Lambda) = \lambda_1 \lambda_2 ... \lambda_n$$

since $\det(U) = 1$ (each eigenvector is of norm 1). Noting that the eigenvector matrix satisfies the property $U^T = U^{-1}$ we can define a new coordinate system $y = Ux$ and substitute to get

$$\rho(y) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2} x^T U^T U \Sigma^{-1} U U^T x} = \frac{1}{\sqrt{(2\pi)^n \det \Lambda}} e^{-\frac{1}{2} y^T \Lambda y} = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\lambda_i)}} e^{-\frac{y_i^2}{2\lambda_i}}$$

where λ_i is the i^{th} eigenvalue. Notice that in the y-coordinate system, each of the components are uncorrelated. Because this is a Gaussian distribution, this implies that each component of y is a Gaussian random variance $y_i \sim N(0, \lambda_i)$.

2.6 Gaussian Distribution in PDEs

Suppose we have the diffusion equation

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 p(x,t)}{\partial x^2}$$
$$p(x,0) = \psi(x)$$

The solution to this problem is a Gaussian distribution

$$p(x,t) = \frac{1}{2\pi t} e^{-\frac{x^2}{2t}}$$

Here$t$ acts like the variance, with $\sigma^2 = t$. Since Green's function for this PDE is simply the Gaussian distribution, the general solution can be written simply as the convolution

$$p(x,t) = \int \frac{1}{2\pi t} e^{-\frac{(x-y)^2}{2t}} p(y,0) \mathrm{d}y.$$

Suppose we are in n dimensional space and matrices $Q = Q^T$ are both positive and finite, and that $Q_{i,j}$ are the entries of Q. Thus look at the PDE

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= \frac{1}{2}\Sigma_{i,j=1}^{n} q_{i,j} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} p(x,t) \\ &= \frac{1}{2}\left(\nabla p(t,x)^T Q \nabla p(x,t)\right)\end{aligned}$$

The solution to this PDE is a multivariate Gaussian distribution

$$p(x,t) = \frac{1}{\sqrt{\det(Q)(2\pi t)^n}} \exp(-x^T (2Qt)^{-1} x)$$

where the covariance matrix, $\Sigma = tQ$ scales linearly with t. This shows a strong connection between the heat equation (diffusion) and the Gaussian distribution, a relation we will expand upon later.

2.7 Independence

Two events are defined to be independent if the measure of doing both events is their product. Mathematically we say that two events P_1 and P_2 are independent if

$$\mu(P_1 \cap P_2) = \mu(P_1)\mu(P_2).$$

This definition can be generalized to random variables. By definition X and Y are independent random variables if $\rho(x,y) = \rho_x(x)\rho_y(y)$ which is the same as saying that the joint probability density is equal to the product of the marginal probability densities

$$\rho_x(x) = \int_{-\infty}^{\infty} \rho(x,y)dy,$$

and ρ_y is defined likewise.

2.8 Conditional Probability

The probability of the event P_1 given that P_2 has occurred is defined as

$$\mu(P_1|P_2) = \frac{\mu(P_1 \cap P_2)}{\mu(P_2)}$$

An important theorem here is Bayes's rule. Bayes' rule can also be referred to as "the flip-flop theorem". Let's say we know the probability of P_1 given P_2. Bayes' theorem let's us flip this around and calculate the probability of P_2 given P_1 (note: these are not necessarily equal! The probability of having an umbrella given that it's raining is not the same as the probability of raining given that you have an umbrella!). Mathematically, Bayes' rule is the equation

$$\mu(P_2|P_1) = \frac{\mu(P_1|P_2)\mu(P_2)}{\mu(P_1)}$$

Bayes's rule also works for the probability density functions. For the probability density functions of random variables X and Y we get that

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

2.9 Change of Random Variables

Suppose we know a random variable $Y : P \to \mathbb{R}^n$ and we wish to change it via some function $\phi : \mathbb{R}^n \to \mathbb{R}^m$ to the random variable X as $X = \phi(Y)$. How do we calculate $\rho(x)$ from $\rho(y)$? Notice that if we define

$$Range(\phi) = \{x | x = \phi(y) \text{ for some } y\}$$

then $\rho(x) = 0$ if $x \notin Range(\phi)$ (since y will never happen and thus x will never happen). For the other values, we note that using calculus we get

$$\frac{dx}{dy} = \phi'(y)$$

Suppose $x \in S$. This means that we want to look for the relation such that

$$\Pr\left(x \in (x, x+dx)\right) = \Pr\left(y \in (y, y+dy)\right)$$

or

$$p(x)|dx| = \rho(y)|dy|$$

and thus

$$\rho(x) = \rho(y)|\frac{dy}{dx}| = \frac{\rho(y)}{|\phi'(y)|}$$

where $y = \phi^{-1}(x)$.

2.9.1 Multivariate Changes

If there are multiple variables, then this generalizes to

$$\rho(x) = \sum_{\{y|\phi(y)=x\}} \rho(y)\frac{1}{\nabla\phi(y)}\rho(\phi^{-1}(y))$$

2.10 Empirical Estimation of Densities

Take a probability density ρ with a vector of parameters $(\theta_1, \ldots, \theta_m)$. We write this together as $\rho_{(\theta_1,\ldots,\theta_m)}(x)$. Suppose that given data we wish to find the "best parameter set" that matches the data. This would mean we would want to find the parameters such that the probability of generating the data that we see is maximized. Assuming that each of n data points are from independent samples, the probability of getting all n data points is the probability of seeing each individual data point all multiplied together. Thus the likelihood of seeing a given set of n data points is

$$\mathcal{L}(x;\theta) = \prod_{i=1}^{n} \rho_{(\theta_1,\ldots,\theta_m)}(x_i)$$

where we interpret $\mathcal{L}(x;\theta)$ as the probability of seeing the data set of the vector x given that we have chosen the parameters θ. Maximum likelihood estimation simply means that we wish to choose the parameters θ such that the probability of seeing the data is maximized, and so our best estimate for the parameter set, $\hat{\theta}$, is calculated as

$$\hat{\theta} = \max_{\theta} \mathcal{L}(x;\theta).$$

Sometimes, it's more useful to use the Log-Likelihood

$$l(x;\theta) = \sum_{i=1}^{n} \rho_{(\theta_1,\ldots,\theta_m)}(x_i)$$

since computationally this uses addition of the probabilities instead of the product (which in some ways can be easier to manipulate and calculate). It can be proven that an equivalent estimator for θ is

$$\hat{\theta} = \max_{\theta} l(x;\theta).$$

3 Introduction to Stochastic Processes: Jump Processes

In this chapter we will develop the ideas of stochastic processes without "high-powered mathematics" (measure theory). We develop the ideas of Markov processes in order to intuitively develop the ideas of a jump process where the probability of jumps are Poisson distributed. We then use this theory of Poisson jump processes in order to define the Brownian motion and prove its properties. Thus by the end of this chapter we will have intuitively defined the SDE and elaborated its properties.

3.1 Stochastic Processes

A stochastic process is the generalization of a random variable to being a changing function in time. Formally, we define a stochastic process as a collection of random variables $\{X(t)\}_{t\geq 0}$ where $X(t)$ is a random variable for the value of X at the time t. This can also be written equivalently as X_t.

3.2 The Poisson Counter

Definition: The Poisson counter is the stochastic process $\{N(t)\}_{t\geq 0}$ where N is the number of events have happened by the time t and the probability of an event happening in a time interval is Poisson distributed. The Poisson counter satisfies the following properties:

- At time 0, no events have happened: $N(0) = 0$
- Independent increment: Give a time interval τ form time point t, the probability of k things happen in this interval does not depend on the time before:

$$\Pr\left(N(t+\tau) - N(t) | N(s), s < t\right) = \Pr\left(N(t+\tau) - N(t)\right)$$

- In a Poisson process, the probability k events happened before time t satisfies the Poisson distribution: $N(t) \sim Poisson(\lambda t)$,

$$\Pr(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

3.3 Markov Process

Definition: A Markov process $\{X(t)\}_{t\geq 0}$ as a process with the following property:

$$\Pr\left(X(t+\tau) = X | X(\delta), \delta \leq t)\right) = \Pr X(t+\tau) = X | X(t))$$

That is to say: one can make predictions for the future of the process based solely on its present state just as well as one could knowing the process's full history. For example, if weather is a Markov process, then the only thing that I need to know to predict the weather tomorrow is the weather today. Note that the Poisson counter is a Markov process. This is trivial given the independent increment part of the definition.

3.4 Time Evolution of Poisson Counter

To solve for the time evolution of the Poisson counter, we will instead of looking at a single trajectory look at an ensemble of trajectories. Think of the ensemble of trajectories as an "amount" or concentration of probability fluid that is flowing from one state to another. For a Poisson counter, the flow is the average rate λ. Thus look at state i is the particles that have jumped i times. The flow out of state i is λ times the amount of probability in state i, or $\lambda p_i(t)$. The flow into the state is simply the flow out of $i-1$, or $\lambda p_{i-1}(t)$. Thus the change in the amount of probability at state i is given by the differential equation

$$\frac{dp_i(t)}{dt} = -\lambda p_i(t) + \lambda p_{i-1}(t).$$

or

$$p_i(t+\Delta t) - p_i(t) = \lambda \Delta t p_{i-1}(t) - \lambda \Delta t p_i(t).$$

To Solve this, define $p(t)$ as the infinite vector (not necessarily a vector because it is countably infinite but the properties we use here of vectors hold for a rigged basis) where of $p_i(t)$. Thus we note that

$$\dot{p}(t) = Ap(t)$$

where

$$A = \begin{bmatrix} -\lambda & & \\ \lambda & -\lambda & \\ & \lambda & \ddots \\ & & \ddots \end{bmatrix}.$$

To solve this, we just need to solve the cascade of equations. Notice that

$$\dot{p}_0(t) = -\lambda p_0(t) \implies p_0(t) = e^{-\lambda t}$$

$$\begin{aligned} \dot{p}_1(t) &= -\lambda p_1(t) + \lambda p_0(t) \\ &= -\lambda p_1(t) + \lambda e^{-\lambda t} \end{aligned}$$

and thus we solve the linear differential equation to get

$$p_1(t) = \lambda t e^{-\lambda t}.$$

For all the others we note that

$$\dot{p}_i(t) = -\lambda p_i(t) + \lambda p_{i-1}(t)$$

which solves as

$$p_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}.$$

To see this is the general solution, simply plug it in to see that it satisfies the differential equation. Because

$$\dot{p}(t) = Ap(t),$$

is a linear system of differential equations, there is a unique solution. Thus our solution is the unique solution.

3.5 Bidirectional Poisson Counter

Now let us assume that we have two Poisson counter processes: one counting up with a rate λ and one counting down with a rate λ. Using the same flow argument, we get that

$$\frac{\dot{p}(t)}{dt} = -2\lambda p_i(t) + \lambda p_{i-1}(t) + \lambda p_{i+1}(t).$$

We assume that all of the probability starts at 0: $p_i(0) = \delta_{i0}$. Notice that this can be written as the system

$$\dot{p}(t) = \begin{bmatrix} \ddots & \ddots & \ddots & & \\ & \lambda & -2\lambda & \lambda & \\ & & \ddots & \ddots & \ddots \end{bmatrix} p(t) = Ap(t)$$

where A is a tridiagonal and infinite in both directions. To solve for this, we use the probability generating function. Define the probability generating function as

$$g(t,z) = \sum_{i=-\infty}^{\infty} z^i p_i(t) = \mathbb{E}\left[z^{x(t)}\right]$$

where the summation is the Laurent series, which is the sum from 0 to infinity added with the sum from -1 to negative infinity. Thus we use calculus and algebra to get

$$\begin{aligned} \frac{\partial g}{\partial t} &= \sum_{i=-\infty}^{\infty} z^i \dot{p}_i(t) \\ &= \sum_{i=-\infty}^{\infty} z^i \left[\lambda p_{i-1}(t) + \lambda p_{i+1} - 2\lambda p_i(t)\right] \\ &= \lambda \sum_{i=-\infty}^{\infty} z^i p_{i-1}(t) + \lambda \sum_{i=-\infty}^{\infty} z^i p_{i+1}(t) - 2\lambda \sum_{i=-\infty}^{\infty} z^i p_i(t). \end{aligned}$$

Notice that since the sum is infinite in both directions, we can trivially change the index and adjust the amount of z appropriately, that is

$$\sum_{i=-\infty}^{\infty} z^i p_{i-1}(t) = z \sum_{i=-\infty}^{\infty} z^i p_i(t) = zg(t,z)$$

$$\sum_{i=-\infty}^{\infty} z^i p_{i+1}(t) = \frac{1}{z}\sum_{i=-\infty}^{\infty} z^i p_i(t) = \frac{1}{z}g(t,z)$$

and thus

$$\frac{\partial g}{\partial t} = \left(\lambda z + \frac{\lambda}{z} - 2\lambda\right) g(t,z).$$

This is a simple linear differential equation which is solved as

$$g(t,z) = e^{\lambda(z+z^{-1}-2)t}.$$

However, to notice the importance of this, go back to the definition of g:

$$g(t,z) = \sum_{i=-\infty}^{\infty} z^i p_i(t).$$

Notice that, if we look at the nth derivative, only one term, the $i = n$ term, does not have a z. Thus if we take $z = 0$ (and discard all terms that blow up), we see that this singles out term which is equal to $n!p_i(t)$. This leads us to the fundamental property of the probability generating function:

$$p_i(t) = \frac{g^{(k)}(t,k)}{k!}|_{z=0}.$$

Thus we can show by induction using this formula with our closed form solution of the probability generating function that

$$p_n(t) = e^{-2\lambda t}\sum_{m=0}^{\infty} \frac{(2\lambda t)^{2m}}{2^{2m}m!(n+m)!} = e^{-2\lambda t}I_n(2\lambda t)$$

where $I_n(x)$ is the nth Bessel function.

3.6 Discrete-Time Discrete-Space Markov Process

Definition: a discrete-time stochastic process is a sequence of random variables $X = X_1, X_2, ..., X_n$. X_n is the state of the process X at time n and X_0 is the initial state. This process is called a discrete-time Markov Chain if for all m and all possible states $i_0, i_1, ...i, j \in X$,

$$\Pr\left(X_{n+1} = j | X_n = i_n, \ldots, X_0 = i_0\right) = \Pr\left(X_{n+1} = j | X_n = i_n\right) = P_{ij}.$$

This definition means that the probability of transition to another state simply depends on where you are right now. This can be represented as a graph where your current state is the node i. The probability of transition from state i to state j is simply P_{ij}, the one-step transition probability. Define

$$\overrightarrow{P(t)} = P(t) = \begin{bmatrix} P_1(t) \\ \vdots \\ P_n(t) \end{bmatrix}$$

as the vector of state probabilities. The way to interpret this is as though you were running many simulations, then the ith component of the vector is the percent of the simulations that are currently at state i. Since the transition probabilities only depend on the current state, we can write the iteration equation

$$P(t+1) = AP(t)$$

where

$$A = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}.$$

Notice that this is simply a concise way to write the idea that at every timestep, $P_{i,j}$ percent of the simulations at state i transfer to state j. Notice then that the probabilities of transitioning at any given time will add up to 1 (the probability of not moving is simply $P_{i,i}$).

Definition: A Regular Markov Chain as a Markov Chain for which some power n of its transition matrix A has only positive entries.

3.7 Continuous-Time Discrete-Space Markov Chain

Define a stochastic process in continuous time with discrete finite state space to be a Markov chain if

$$\Pr\left(X_{n+1} = j | X_n = i_n, ..., X_0 = i_0\right) = p_{ij}(t_{n+1} - t_n)$$

where p_{ij} is the solution of the forward equation

$$P'(t) = P(t)Q$$

where Q is a transition rate matrix which used to describe the flow of probability juices from state i to the state j. Notice that

$$\begin{aligned} P'(t) &= P(t)Q \\ \frac{P'(t+h) - P'(t)}{h} &= P'(t)Q \\ P(t+h) &= (I + Qh)P(t) \\ P(t+h) &= AP(t) \end{aligned}$$

and thus we can think of a continuous-time Markov chain as a discrete-time Markov chain with infinitely small timesteps and a transition matrix

$$A = I + Qh$$

Note that the transition rate matrix Q satisfies the following properties:

1. Transition flow between state i and j $q_{ij} > 0$ when $i \neq j$;
2. Transition probability from i and i $a_{ij} = q_{ij}h$;
3. $\sum_{j=1}^{n} q_{ij} = 0, q_{ii} = -\sum_{j \neq i} q_{ij}$

Property 1 is stating that the transition rate matrix is composed only "rates from i" and thus they are all positive values. Property 2 is restating the relation between Q and A. Property 3 is stating that the diagonal of Q is composed of the flows into the state i, and thus it will be a negative number. One last property to note is that since

$$P'(t) = P(t)Q$$

we get that

$$P(t) = P(0)e^{Qt}$$

where e^{Qt} is the matrix exponential (defined by its Taylor Series expansion being the same as the normal exponential). This means that there exists a unique solution to the time evolution of the probability densities. This is an interesting fact to note: even though any given trajectory evolves randomly, the way a large set of trajectories evolve together behaves deterministically.

3.7.1 Example: DNA Mutations

Look at a many bases of DNA. Each can take four states: $\{A, T, C, G\}$ Denote the percent that are in state i at the time t as $p_i(t)$. We write our system as the vector

$$P(t) = (p_A(t), p_G(t), p_C(t), p_T(t))^T$$

Define Q by the mutation rates from A to G, C to T, etc (yes, this is a very simplistic model. Chill bra.). We get that the evolution of the probability of being in state i at time t is given by the differential equation

$$P'(t) = QP(t).$$

Since we have mathematized this process, we can now use our familiar rules to investigate the system. For example, if we were to run this system for a very long time, what percent of the bases will be A? Since this is simply a differential equation, we find this by simply looking at the steady state: the $P(t)$ s.t. $P'(t) = 0$. Notice that means that we have

$$0 = QP(t) = \lambda$$

since 0 is just a constant. Thus the vector P that satisfies this property is an eigenvector of Q. This means that the eigenvector of Q corresponding to the eigenvalue of 0 gives the long run probabilities of being in state i respectively.

3.8 The Differential Poisson Counting Process and the Stochastic Integral

We can intuitively define the differential Poisson Counting Process is the process that describes the changes of a Poisson CounterN_t (or equivalently $N(t)$). Define the differential stochastic process dN_t by its some integral

$$N_t = \int_0^t dN_t.$$

To understand dN_t, let's investigate some of its properties. Since $N_t \sim Poisson(\lambda t)$, we know that

$$\mathbb{E}\left[N_t\right] = \lambda t = \mathbb{E}\left[\int_0^t dN_t\right].$$

Since $\mathbb{E}$ is defined as some kind of a summation or an integral, we assume that dN_t^2 is bounded which, at least in the normal calculus, lets us swap the ordering of integrations. Thus

$$\lambda t = \int_0^t \mathbb{E}\left[dN_t\right].$$

Since the expected value makes a constant, we intuit that

$$\lambda t = \int_0^t \alpha dt = \alpha t$$

and thus

$$\mathbb{E}\left[dN_t\right] = \lambda.$$

Notice this is simply saying that dN_t represents the "flow of probability" that is on average λ. Using a similar argument we also note that the variance of $dN_t = \lambda$. Thus we can think of the equation the term dN_t as a kind of a limit, where

$$N(t + dt) - N(t) = dN_t$$

is probability of jumping k times in the increment dt which is given by the probability distribution

$$P\left(k \text{ jumps in the interval } (t, t + dt)\right) = \frac{(\lambda dt)^k}{k!} e^{-\lambda dt}$$

Then how do we make sense of the integral? Well, think about writing the integral as a Riemann sum:

$$\int_0^t dN_t = \lim_{\Delta t \to 0} \sum_{i=0}^{n-1} \left(N(t_{i+1}) - N(t_i)\right)$$

where $t_i = i\Delta t$. One way to understand how this is done is algorithmically/computationally. Let's say you wanted to calculate one "stochastic trajectory" (one instantiation) of $N(t)$. What we can

do is pick a time interval dt. We can get $N(t)$ by, at each time step dt, sample a value from the probability distribution

$$P\left(k \text{ jumps in the interval } (t, t+dt)\right) = \frac{(\lambda dt)^k}{k!} e^{-\lambda dt}$$

and repeatedly add up the number of jumps such that $N(t)$ is the total number of jumps at time t. This will form our basis for defining and understanding stochastic differential equations.

3.9 Generalized Poisson Counting Processes

We now wish to generalize the counting processes. Define a counting process $X(t)$ as

$$dX(t) = f(X(t), t)dt + g(X(t), t)dN(t)$$

where f is some arbitrary function describing deterministic changes in time where g defines the "jumping" properties. The way to interpret this is as a time evolution equation for X. As we increment in time by Δt, we add $f(X(t), t)$ to X. If we jump in that interval, we also add $g(X(t), t)$. The probability of jumping in the interval is given by

$$P\left(k \text{ jumps in the interval } (t, t+dt)\right) = \frac{(\lambda dt)^k}{k!} e^{-\lambda dt}$$

Another way of thinking about this is to assume that the first jump happens at a time t_1. Then $X(t)$ evolves deterministically until t_1 where it jumps by $g(X(t), t)$, that is

$$\lim_{t \to t_1^+} X(t) = g\left(\lim_{t \to t_1^-} X(t), t\right) + \lim_{t \to t_1^-} X(t).$$

Notice that we calculate the jump using the left-sided limit. This is known as Ito's calculus and it is interpreted as the jump process "not knowing" any information about the future, and thus it jumps using only previous information.

We describe the solution to the SDE once again using an integral, this time we write it as

$$X(t) = X(0) + \int_0^t f(X(t), t)dt + \int_0^t g(X(t), t)dN(t).$$

Notice that the first integral is simply a deterministic integral. The second one is a stochastic integral. It can once again be understood as a Riemann summation, this time

$$\int_0^t g(X(t), t)dN_t = \lim_{\Delta t \to 0} \sum_{i=0}^{n-1} g(X(t_i), t_i)\left(N(t_{i+1}) - N(t_i)\right).$$

We can understand the stochastic part the same as before simply as the random amount of jumps that happen in the interval $(t, t+dt)$. However, now we multiply the number of jumps in the interval by $g(X(t_i), t_i)$, meaning "the jumps have changing amounts of power".

3.10 Important Note: The Defining Feature of Ito's Calculus

It is important to note that g is evaluated using the X and t before the jump. This is the defining principle of the Ito Calculus and corresponds to the "Left-Hand Rule" for Riemann sums. Unlike Newton's calculus, the left-handed, right-handed, and midpoint summations do not converge to the same value in the stochastic calculus. Thus all of these different ways of summing up intervals in order to solve the integral are completely different calculi. Notably, the summation principle which uses the midpoints

$$\int_0^t g(X(t),t)dN_t = \lim_{\Delta t\to 0}\sum_{i=0}^{n-1}\left(\frac{g(X(t_{i+1}),t_{i+1})+g(X(t_i),t_i)}{2}\right)(N(t_{i+1})-N(t_i))\,.$$

is known as the Stratonovich Calculus. You may ask, why choose Ito's calculus? In some sense, it is an arbitrary choice. However, it can be motivated theoretically by the fact that Ito's Calculus is the only stochastic calculus where the stochastic adder g does not "use information of the future", that is, the jump sizes do not adjust how far they will jump given the future information of knowing where it will land. Thus, in some sense, Ito's Calculus corresponds to the type of calculus we would believe matches the real-world. Ultimately, because these give different answers, which calculus best matches the real-world is an empirical question that could be investigated itself.

3.11 Example Counting Process

Define the counting process

$$dX_t = X_t dt + X_t dN_t.$$

Suppose that the jumps occur at $t_1 < t_2 < \dots$. Thus before t_1, we evolve deterministically as

$$dX_t = X_t dt$$

and thus $X(t) = e^t$ before t_1. At t_1, we take this value and jump by X_{t_1}. Thus since immediately before the jump we have a value e^{t_1}, immediately after the jump we have $e^{t_1} + e^{t_1} = 2e^{t_1}$. We once again it begins to evolve as

$$dX_t = X_t dt$$

but now with the initial condition $X(t_1) = 2e^{t_1}$. We see that in this interval the linear equation solves to $X(t) = 2e^t$. Now when we jump at t_2, we have the value $2e^{t_2}$ and jump by $2e^{t_2}$to get $4e^{t_2}$ directly after the jump. Seeing the pattern, we get that

$$X(t) = \begin{cases} e^t, & 0 \le t \le t_1 \\ 2e^t, & t_1 \le t \le t_2 \\ \vdots & \vdots \\ 2^n e^t, & t_n \le t \le t_{n+1} \end{cases}$$

3.12 Ito's Rules for Poisson Jump Process

Given the SDE

$$dx(t) = f(x(t),t)dt + \sum_{i=1}^{n} g_i(x(t),t)dN_i$$

where $x \in \mathbb{R}$, N_i is a Poisson counter with rate λ_i, $f : \mathbb{R}^n \to \mathbb{R}^n$, $g_i : \mathbb{R}^n \to \mathbb{R}^n$. Define $Y = \psi(X,t)$ as some random variable whose values are determined as a function of X and t. How do we find $dy(t)$, the time evolution of y? There are two parts: the determinsitic changes and the stochastic jumps. The deterministic changes are found using Newton's calculus. Notice using Newtonian calculus that

$$\Delta\text{Deterministic} = \frac{\partial\psi}{\partial t}\frac{\partial t}{\partial t} + \frac{\partial\psi}{\partial x}\frac{\partial x}{\partial t} = \frac{\partial\psi}{\partial t} + \frac{\partial\psi}{\partial x}(dx)_{deterministic} = \frac{\partial\psi}{\partial t} + \frac{\partial\psi}{\partial x}f(x).$$

The second part are the stochastic changes due to jumping. Notice that if the Poisson counter process i jumps in the interval, the jump will change x from x to $x + g_i$. This means that y will change from $\psi(x,t)$ to $\psi(x+g_i(x),t)$. Thus the change in y due to jumping is the difference between the two times the number of jumps, calculated as

$$\Delta\text{Jumps} = \sum_{i=1}^{n}[\psi(x + g_i(x)) - \psi(x,t)]dN_i,$$

where dN_i is the number of jumps in the interval. This approximation is not correct if some process jumps multiple times in the interval, but if the interval is of size dt ($[t, t+dt]$), then the probability that a Poisson process jumps twice goes to zero as $dt \to 0$. Thus this approximation is correct for infinitesimal changes. Putting these terms together we get

$$\begin{aligned} d\psi(x,t) &= \Delta\text{Deterministic} + \Delta\text{Jumps} \\ &= dy(t) = d\psi(x,t) = \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial x}f(x)dt + \sum_{i=1}^{m}[\psi(x + g_i(x)) - \psi(x,t)]dN_i. \end{aligned}$$

which is Ito's Rule for Poisson counter processes.

3.12.1 Example Problem

Let

$$dx(t) = -x(t)dt + dN_1(t) - dN_2(t)$$

where $N_1(t)$ is a Poisson counter with rate λ_1 and $N_2(t)$ is a Poisson counter with rate λ_2. Let us say we wanted to know the evolution of $Y = X^2$. Thus $\psi(X,t) = X^2$. Using Ito's Rules, we get

$$\begin{aligned} dx^2(t) &= 2x(-x(t))dt + ((x+1)^2 - x^2)dN_1 + ((x-1)^2 - x^2)dN_2 \\ &= -2x^2dt + (2x+1)dN_1 + (1-2x)dN_2 \end{aligned}$$

3.13 Dealing with Expectations of Poisson Counter SDEs

Take the SDE

$$dx = f(x,t)dt + \sum_{i=1}^{m} g_i(x,t)dN_i$$

where N_i is a Poisson counter with rate λ_i. Notice that since $N_i(t) \sim Poisson(\lambda t)$, we get

$$\mathbb{E}\left[N_i(t)\right] = \lambda t.$$

Also notice that because $N_i(t)$ is a Poisson process, the probability $N_i(t)$ will jump in interval $(t, t+h)$ is independent of $X(\sigma)$ for any $\sigma < t$. This mean that, since the current change is independent of the previous changes, $\mathbb{E}\left[g(x(t),t)dN_i(t)\right] = \mathbb{E}\left[g(x(t),t)\right]\mathbb{E}\left[dN_i\right] = \lambda_i \mathbb{E}\left[g(x(t),t)\right]$. Thus we get the fact that

$$\begin{aligned}
\mathbb{E}\left[x(t+h) - x(t)\right] &= \mathbb{E}\left[f(x,s)h\right] + \sum_{i=1}^{m} \mathbb{E}\left[g(x(t),t)dN_i(t)\right] \\
&= \mathbb{E}\left[f(x,t)\right]h + \sum_{i=1}^{m} \mathbb{E}\left[g_i(x,t)\right]\lambda_i h \\
\frac{\mathbb{E}\left[x(t+h) - x(t)\right]}{h} &= \mathbb{E}\left[f(x,t)\right] + \sum_{i=1}^{m} \lambda_i \mathbb{E}\left[g_i(x,t)\right]
\end{aligned}$$

and thus we take the limit as $h \to 0$ to get

$$\frac{d\mathbb{E}\left[x(t)\right]}{dt} = \mathbb{E}\left[f(x,t)\right] + \sum_{i=1}^{m} \lambda_i \mathbb{E}\left[g_i(x,t)\right].$$

3.13.1 Example Calculations

Given SDE

$$dx = -xdt + dN_1 - dN_2$$

we apply Ito's Rule to get

$$\frac{dE\left[x\right]}{dt} = -\mathbb{E}\left[x\right] + \lambda_1 - \lambda_2.$$

Notice that this is just an ODE in $\mathbb{E}\left[x\right]$. If it makes it easier to comprehend, let $Y = \mathbb{E}\left[x\right]$ to see that this is simply

$$Y' = -Y + \lambda_1 - \lambda_2.$$

which we can solve using our tools from ODEs. Notice the connection here: even though the trajectory is itself stochastic, its expectations change deterministically. Now we apply Ito's rules to get

$$dx^2 = -2x^2dt + (2x+1)dN_1 + (1-2x)dN_2$$

and thus we get that the expectation changes as

$$\frac{d\mathbb{E}\left[x^2\right]}{dt} = -2\mathbb{E}\left[x^2\right] + \left(2\mathbb{E}\left[x\right] + 1\right)\lambda_1 + \left(1 - 2\mathbb{E}\left[x\right]\right)\lambda_2.$$

To solve this, we would first need to complete solving the ODE for $\mathbb{E}[x]$, then plug that solution into this equation to get another ODE, which we solve. However, notice that this has all been changed into ODEs, something we know how to solve!

3.13.2 Another Example Calculation

Given SDEs

$$\begin{aligned} dx &= -xdt + zdt, \\ dz &= -2zdN. \end{aligned}$$

for $z \in \{-1, 1\}$. Thus by Ito's Rules

$$\begin{aligned} dx^2 &= 2x(-xdt + zdt), \\ &= -2x^2dt + 2xzdt. \end{aligned}$$

while

$$\begin{aligned} d(xz) &= zdx + xdz, \\ &= (z^2 - xz)dt - 2xzdN. \end{aligned}$$

Thus we get that

$$\begin{aligned} \frac{d\mathbb{E}\left[x^2\right]}{dt} &= -2\mathbb{E}\left[x^2\right] + 2\mathbb{E}\left[xz\right] \\ \frac{d\mathbb{E}\left[xz\right]}{dt} &= \mathbb{E}\left[x^2\right] - 2\mathbb{E}\left[xz\right]\lambda - \mathbb{E}\left[xz\right] \end{aligned}$$

as the system of ODEs that determine the evolution of certain expectations central to the solving of the variance and the covariance.

3.13.3 Important Example: Bidirectional Poisson Counter

Suppose we have two Poisson processes, $dN_1(t)$ and $dN_2(t)$ with rates $\frac{\lambda}{2}$ and define $y(t)$ by the SDE

$$dy(t) = dN_1(t) - dN_2(t)$$

We can rescale this equation s.t.

$$\begin{aligned} x_\lambda(t) &= \frac{1}{\sqrt{\lambda}}y(t) \\ dx_\lambda(t) &= \frac{1}{\sqrt{\lambda}}dN_1(t) - \frac{1}{\sqrt{\lambda}}dN_2(t) \end{aligned}$$

where the jump size is proportional to $\frac{1}{\sqrt{\lambda}}$. Then we have

$$\frac{d\mathbb{E}\left[x_\lambda(t)\right]}{dt} = \frac{\lambda}{2\sqrt{\lambda}} - \frac{\lambda}{2\sqrt{\lambda}} = 0$$

We use Ito's Rules with $g = \pm 1$ and the Binomial theorem to get

$$\begin{aligned}
\frac{d\mathbb{E}\left[x_\lambda^p(t)\right]}{dt} &= \mathbb{E}\left[\left(\left(x+\frac{1}{\sqrt{\lambda}}\right)^p - x^p\right)dN_1 + \left(\left(x-\frac{1}{\sqrt{\lambda}}\right)^p - x^p\right)dN_2\right] \\
&= \mathbb{E}\left[\left(\left(x^p + \binom{p}{1}\frac{1}{\sqrt{\lambda}}x^{p-1} + \binom{p}{2}\frac{1}{\lambda}x^{p-2} + \dots\right) - x^p\right)dN_1 + \left(\left(x^p - \binom{p}{1}\frac{1}{\sqrt{\lambda}}x^{p-1} + \binom{p}{2}\frac{1}{\lambda}x^{p-2} + \dots\right) - x^p\right)dN_2\right] \\
&= \mathbb{E}\left[\left(\binom{p}{2}\frac{1}{\lambda}x^{p-2} + \dots\right)(dN_1 + dN_2) + \left(\binom{p}{1}\frac{1}{\sqrt{\lambda}}x^{p-1} + \binom{p}{3}\frac{1}{\sqrt{\lambda^3}}x^{p-3} + \dots\right)(dN_1 - dN_2)\right] \\
&= \left(\binom{p}{2}\frac{1}{\lambda}\mathbb{E}\left[x^{p-2}\right]\right)\left(\frac{\lambda}{2} + \frac{\lambda}{2}\right) \\
&= \binom{p}{2}\mathbb{E}\left[x^{p-2}\right]
\end{aligned}$$

where we drop off all the higher order $\frac{1}{\lambda}$ terms since the will go to zero as $\lambda \to \infty$. This means that in the limit we get

$$\frac{dE[x^p(t)]}{dt} = \frac{p(p-1)}{2}E[x^{p-2}(t)]$$

Thus as $\lambda \to \infty$ we can ignore higher order terms to get all of the odd moments as 0 and the even moments as:

1. $\frac{d}{dt}E[x^2(t)] = 1$
2. $E[x^p(t)] = \frac{p!}{\frac{p}{2}!}\left(\frac{t}{2}\right)^{\frac{p}{2}}$

Let $\sigma^2 = t$ Then all moments match, so as $\lambda \to \infty$ the random variable $x_\infty(t)$ will be Gaussian with mean 0 and variance t. Thus we can think of $x_\infty(t)$ as a stochastic process whose probability distribution starts as a squished Gaussian distribution which progressively flattens linearly with time.

3.14 Poisson Jump Process Kolmogorov Forward Equation

Take the SDE

$$dx = f(x,t)dt + \sum_{i=1}^{m} g_i(x,t)dN_i$$

Assume that f and g_i are sufficiently smooth and the initial density is small. What we wish to find is the probability density function for X at a time t, $\rho(x,t)$. To derive this, take the arbitrary function $\psi(x)$. By the multivariate Ito's Rules,

$$d\psi(x) = \frac{\partial \psi}{\partial x}f(x,t)dt + \sum_{i=1}^{m}[\psi(x+g_i(x),t) - \psi(x,t)]dN_i$$

and thus

$$\frac{d\mathbb{E}\left[\psi\right]}{dt} = \mathbb{E}\left[\frac{\partial \psi}{\partial x} f(x,t)\right] + \sum_{i=1}^{m} \mathbb{E}\left[\psi(x+g_i(x),t) - \psi(x,t)\right] \lambda_i.$$

Recall that the definition of the expected value is

$$\mathbb{E}\left[\psi(x)\right] = \int_{-\infty}^{\infty} \rho(x,t)\psi(x)dx$$

and thus

$$\int_{-\infty}^{\infty} \frac{\partial \rho}{\partial t}\psi(x)dx \quad = \quad \int_{-\infty}^{\infty} \frac{\partial \psi}{\partial x} f(x,t)\rho(x,t)dx + \sum_{i=1}^{m} \lambda_i \int_{-\infty}^{\infty} [\psi(x+g_i(x),t) - \psi(x,t)]\rho(x,t)dx$$

We next simplify the equation term by term using integration by parts. What we want to get is every term having a $\psi(x)$ term so we can group all the integrals. Thus take the first integral

$$\int_{-\infty}^{\infty} \frac{\partial \psi}{\partial x} f\rho dx.$$

Here we let $u = f\rho$ and $dv = \frac{\partial \psi}{\partial x}$ in order to get that

$$\int_{-\infty}^{\infty} \frac{\partial \psi}{\partial x} f\rho dx = f\rho\psi|_{-\infty}^{\infty} - \int_{\infty}^{-\infty} \frac{\partial (f\rho)}{\partial x}\psi(x)dx.$$

Notice that in order for the probability density to integrate to 1 (and thus the integral be bounded), we must have ρ vanish at the infinities. Thus

$$\int_{-\infty}^{\infty} \frac{\partial \psi}{\partial x} f\rho dx = -\int_{\infty}^{-\infty} \frac{\partial (f\rho)}{\partial x}\psi dx.$$

Next we take the g term that does not have ψ in the same version. In order to solve the first g integral, we will need to change the variables to make the integrating variable simpler. Thus let $z = \tilde{g}_i(x) = x + g_i(x)$. Therefore $dz = (1 + g_i'(x))\, dx$ Using this, we can re-write

$$\int_{-\infty}^{\infty} \psi(x+g_i(x))\rho dx = \int_{-\infty}^{\infty} \psi(z) \frac{\rho(\tilde{g}_i^{-1}(z),t)}{|1+g_i'(\tilde{g}_i^{-1}(z)|} dz$$

However, notice that $\lim_{x\to\infty} z = \infty$ and $\lim_{x\to-\infty} z = -\infty$ and thus we do not need to change the bounds, making

$$\int_{-\infty}^{\infty} \psi(x+g_i(x))\rho dx = \int_{-\infty}^{\infty} \psi(x) \frac{\rho(\tilde{g}_i^{-1}(x),t)}{|1+g_i'(\tilde{g}_i^{-1}(x)|} dx.$$

Thus we plug these integrals back into our equation to get

$$\int_{-\infty}^{\infty} \frac{\partial \rho}{\partial t}\psi(x)dx = \int_{-\infty}^{\infty} \frac{\partial \psi}{\partial x}, f(x,t)\rho(x,t)dx + \sum_{i=1}^{m} \lambda_i \int_{-\infty}^{\infty} [\psi(x+g_i(x),t) - \psi(x,t)]\rho(x,t)dx$$

$$\int_{\infty}^{-\infty} \frac{\partial p}{\partial t}\psi dx = -\int_{\infty}^{-\infty} \frac{\partial (f\rho)}{\partial x}\psi dx + \sum_{i=1}^{m} \lambda_i [\int_{-\infty}^{\infty} \psi(x) \frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} dx - \int_{-\infty}^{\infty} \psi p dx].$$

We collect all of the terms to one side to get

$$\int_{-\infty}^{\infty} \psi(x) \left(\frac{\partial \rho}{\partial t} + \frac{\partial (f\rho)}{\partial x} - \sum_{i=1}^{m} \lambda_i \left[\frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} - \rho \right] \right) dx = 0$$

Since $\psi(x)$ we arbitrary, let ψ be the indicator for the arbitrary set A, that is

$$\psi(x) = I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0 & \text{o.w.} \end{cases}$$

Thus we get that

$$\int_{A} \left(\frac{\partial \rho}{\partial t} + \frac{\partial (f\rho)}{\partial x} - \sum_{i=1}^{m} \lambda_i \left[\frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} - \rho \right] \right) dx = 0$$

for any $A \subset \mathbb{R}$. Thus, in order for this to be satisfied for all subsets of the real numbers, the integrand must be identically zero. This means

$$\frac{\partial \rho}{\partial t} + \frac{\partial (f\rho)}{\partial x} - \sum_{i=1}^{m} \lambda_i \left[\frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} - \rho \right] = 0$$

which we arrange as

$$\frac{\partial p}{\partial t} = -\frac{\partial (fp)}{\partial x} + \sum_{i=1}^{m} \lambda_i \left[\frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} - \rho \right].$$

This equation describes the time evolution of the probability density function $\rho(x, t)$ via a deterministic PDE.

3.14.1 Example Kolmogorov Calculation

Take the SDE

$$dx = -xdt + dN_1 - dN_2$$

where N_1 and N_2 are Poisson counter with rate λ. To calculate the probability density function, we plug in the functions into the Kolmogorov equation to get

$$\frac{\partial p}{\partial t} = -\frac{\partial (-xp)}{\partial x} + \lambda[p(x-1, t) - p(x, t)] + \lambda[p(x+1, t) - p(x, t)].$$

If we are also given an initial density $\rho(x, 0)$ (say, a Dirac δ-function denoting that all trajectories start at the same spot), we can calculate the time evolution of the probability density using computational PDE solvers that we all know and love.

4 Introduction to Stochastic Processes: Brownian Motion

In this chapter we will use the properties of Poisson Counting Processes in order to define Brownian Motion and derive a calculus for dealing with stochastic differential equations written with differential Wiener/Brownian terms.

4.1 Brownian Motion / The Wiener Process

Now we finally define Brownian Motion. Robert Brown is credited for first describing Brownian motion. Brown observed pollen particles moving randomly and described the motion. Brownian motion is also sometimes called a Wiener process because Norbert Wiener was the first person to describe random motion mathematically. It is most commonly defined more abstractly using limits of random walks or abstract function space definitions. We will define it using the bidirectional Poisson counter. Recall that we define this using two Poisson processes, $dN_1(t)$ and $dN_2(t)$ with rates $\frac{\lambda}{2}$ and define $y(t)$ by the SDE

$$dy(t) = dN_1(t) - dN_2(t)$$

We can rescale this equation s.t.

$$\begin{aligned} x_\lambda(t) &= \frac{1}{\sqrt{\lambda}} y(t) \\ dx_\lambda(t) &= \frac{1}{\sqrt{\lambda}} dN_1(t) - \frac{1}{\sqrt{\lambda}} dN_2(t). \end{aligned}$$

Define the Wiener process, $W(t)$ (or equivalently, Brownian motion $B(t)$) as the limit as both of the rates go to infinity, that is

$$\lim_{\lambda \to \infty} X_\lambda(t) \to W(t).$$

4.2 Understanding the Wiener Process

Like before with the Poisson counter, we wish to understand what exactly dW_t is. We define $W(0) = 0$. Define dW_t by its integral

$$W_t = \int_0^t dW_t.$$

We can once again understand the integral by using the Riemann summation

$$\int_0^t g(X,t) dW_t = \lim_{\Delta t \to 0} \sum_{i=1}^{n-1} g(X_{t_i}, t_i) \left(dW_{t_{i+1}} - dW_{t_i} \right)$$

where $t_i = i\Delta t$ Notice once again that we are evaluating g using the Left-hand rule as this is the defining feature of Ito's Calculus: it does not use information of the future.

Recall that this definition is defined by the bidirectional Poisson Counter. Can we then understand the interval as a number of jumps? Since since the rate of jumps is infinitely high, we can think of this process as making infinitely many infinitely small jumps in every interval of time. Thus we cannot understand the interval as the "number of jumps" because infinitely many will occur! However, given the proof from 3.13.3, we get that $W(t) \sim N(0,t)$. Thus we can think of $(dW_{i+1} - dW_i) \sim N(0, dt)$, that is, the size of the increment is normally distributed. Algorithmically solving the integral by taking a normally distributed random number with variance dt and multiplying it by g to get the value of $g(X,t)dW_t$ over the next interval of time. Using Ito's Rules for Wiener Processes (which we will derive shortly) we can easily prove that

1. $\mathbb{E}\left[(W(t) - W(s))^2\right] = t - s$ for $t > s$.
2. $\mathbb{E}[W(t_1)W(t_2)] = \min(t_1, t_2)$.

Notice that this means $E[(W(t+\Delta t) - W(t))^2] = \Delta t$ and thus in the limit as $\Delta t \to 0$, then $E[(W(t+\Delta t) - W(t))^2] \to 0$ and thus the Wiener process is continuous almost surely (with probability 1). However, it can be proven that it is not differentiable with probability 1! Thus dW_t is some kind of abuse of notation because the derivative of W_t does not really exist. However, we can still use it to understand the solution to an arbitrary SDE

$$dX_t = f(X,t)dt + g(X,t)dW_i$$

as the integral

$$X_t = X_0 + \int_0^t f(X_t, t)dt + \int_0^t g(X_t, t)dW_t.$$

4.3 Ito's Rules for Wiener Processes

By the definition of the Wiener process, we can write

$$dW_i = \frac{1}{\sqrt{\lambda}}\left(dN_i - dN_{-i}\right)$$

in the limit where $\lambda \to \infty$. We can define the stochastic differential equation (SDE) in terms of the Wiener process

$$dX_t = f(X,t)dt + \sum_{i=1}^{n} g_i(x)dW_i = f(X,t)dt + \sum_{i=1}^{n} \frac{g_i(x)}{\sqrt{\lambda}}\left(dN_i - dN_{-i}\right)$$

We now define $Y_t = \psi(x,t)$. Using Ito's Rules for Poisson Jump Processes, we get that

$$dY_t = d\psi(x,t) = \frac{\partial \psi}{\partial t}dt + \frac{\partial \psi(x,t)}{\partial x}f(x,t)dt + \sum_{i=1}^{n}\left(\psi\left(x + \frac{g_i(x)}{\sqrt{\lambda}}\right) - \psi(x,t)\right)dN_i + \sum_{i=1}^{n}\left(\psi\left(x - \frac{g_i(x)}{\sqrt{\lambda}}\right) - \psi(x,t)\right)dN_{-i}.$$

To simplify this, we expand ψ by λ to get

$$\begin{aligned}\psi\left(x+\frac{g_i(x)}{\sqrt{\lambda}}\right) &= \psi(x)+\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}+\psi''(x)\frac{g_i^2(x)}{\lambda}+\mathcal{O}(\lambda^{-\frac{3}{2}})\\ \psi\left(x-\frac{g_i(x)}{\sqrt{\lambda}}\right) &= \psi(x)-\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}+\psi''(x)\frac{g_i^2(x)}{\lambda}+\mathcal{O}(\lambda^{-\frac{3}{2}})\end{aligned}$$

to simplify to (dropping off higher-order terms)

$$d\psi(x)=\frac{\partial\psi}{\partial t}dt+\frac{\partial\psi(x,t)}{\partial x}f(x,t)dt+\sum_{i=1}^{n}\left(\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}+\psi''(x)\frac{g_i^2(x)}{\lambda}\right)dN_i+\sum_{i=1}^{n}\left(-\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}+\psi''(x)\frac{g_i^2(x)}{\lambda}\right)dN_{-i},$$

and thus

$$d\psi(x)=\frac{\partial\psi}{\partial t}dt+\frac{\partial\psi(x,t)}{\partial x}f(x,t)dt+\sum_{i=1}^{n}\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}\left(dN_i-dN_{-i}\right)\sum_{i=1}^{n}\psi''(x)g_i^2(x)\left(\frac{1}{\lambda}\left(dN_i+dN_{-i}\right)\right).$$

Let us take a second to justify dropping off the higher order terms. Recall that the Wiener process looks at the limit as $\lambda\to\infty$. In the expansion, the terms dropped off are $\mathcal{O}(\lambda^{-\frac{3}{2}})$. Recall that

$$\mathbb{E}\left[dN_i\right]=\frac{\lambda}{2}.$$

Thus we see that the expected contribution of these higher order terms is

$$\lim_{\lambda\to\infty}\mathbb{E}\left[\mathcal{O}(\lambda^{-\frac{3}{2}})dN_i\right]=\lim_{\lambda\to\infty}\mathcal{O}(\lambda^{-\frac{1}{2}})=0.$$

However, we next note that

$$\mathbb{V}\left[dN_i\right]=\mathbb{E}\left[dN_i\right]=\frac{\lambda}{2}.$$

Thus we get that

$$\lim_{\lambda\to\infty}\mathbb{V}\left[\mathcal{O}(\lambda^{-\frac{3}{2}})dN_i\right]=\lim_{\lambda\to\infty}\mathcal{O}(\lambda^{-\frac{1}{2}})=0.$$

Therefore, since the variance goes to zero, the contribution of these terms are not stochastic. Thus the higher order terms deterministically make zero contribution (in more rigorous terms, they make no contribution to the change with probability 1). Therefore, although this at first glace looked like an approximation, we were actually justified in only taking the first two terms of the Taylor series expansion.

Sub-Calculation

To simplify this further, we need to do a small calculation. Define

$$dz_i=\frac{1}{\lambda}\left(dN_i+dN_{-i}\right).$$

Notice that

$$\frac{d\mathbb{E}\left[z_i\right]}{dt} = \frac{1}{\lambda}\left(\frac{\lambda}{2} + \frac{\lambda}{2}\right) = 1$$

which means

$$\mathbb{E}\left[z_i\right] = t.$$

Using Ito's rules

$$\begin{aligned} dz_i^2 &= \left(\left(z_i + \frac{1}{\lambda}\right)^2 - z_i^2\right)\left(dN_i + dN_{-i}\right) \\ &= \left(\frac{2z_i}{\lambda} + \frac{1}{\lambda^2}\right)\left(dN_i + dN_{-i}\right) \end{aligned}$$

and thus

$$\begin{aligned} \frac{d\mathbb{E}\left[z_i^2\right]}{dt} &= \left(\frac{2\mathbb{E}\left[z_i\right]}{\lambda} + \frac{1}{\lambda^2}\right)\left(\frac{\lambda}{2} + \frac{\lambda}{2}\right) \\ &= \left(\frac{2t}{\lambda} + \frac{1}{\lambda^2}\right)\lambda \\ &= 2t + \frac{1}{\lambda} \end{aligned}$$

to make

$$\mathbb{E}\left[z_i^2\right] = t^2 + \frac{t}{\lambda}.$$

This means that

$$\mathbb{V}\left[z_i\right] = \mathbb{E}\left[z_i^2\right] - \mathbb{E}\left[z_i\right]^2 = t^2 + \frac{t}{\lambda} - t^2 = \frac{t}{\lambda}.$$

Thus look at

$$Z = \lim_{\lambda \to \infty} z_i.$$

This means that

$$\mathbb{E}\left[Z\right] = t,$$

and

$$\mathbb{V}\left[Z\right] = 0.$$

Notice that since the variance goes to 0, in the limit as $\lambda \to \infty$, Z is a deterministic process that is equal to t, and thus $dZ = dt$.

Solution for Ito's Rules

Return to

$$d\psi(x) = \psi'(x)(f(X,t)dt + \sum_{i=1}^{n} \psi'(x)\frac{g_i(x)}{\sqrt{\lambda}}\left(dN_i - dN_{-i}\right) + \sum_{i=1}^{n} \psi''(x)g_i^2(x)\left(-\psi'(x)\frac{g_i(x)}{\sqrt{\lambda}} + \psi''(x)\frac{g_i^2(x)}{\lambda}\right)\left(\frac{1}{\lambda}\left(dN_i + dN_{-i}\right)\right).$$

Notice that the last term is simply z_i. Thus we take the limit as $\lambda \to \infty$ to get that

$$d\psi(x) = \psi'(x)(f(X,t)dt + \sum_{i=1}^{n} \psi'(x)g_i(x)dW_t + \sum_{i=1}^{n} \psi''(x)g_i^2(x)dt$$

which is Ito's Rules for an SDE. If x is a vector then we would arrive at

$$d\phi(x) = \langle \psi(x), f(x)\rangle dt + \sum_{i=1}^{n} \langle \nabla\psi(x), g_i(x)\rangle dW_i + \frac{1}{2}\sum_{i=1}^{n} g_i(x)^T \nabla^2 \psi(x) g_i(x) dt$$

where $\langle x, y\rangle$ is the dot product between x and y and ∇^2 is the Hessian.

4.4 A Heuristic Way of Looking at Ito's Rules

Take the SDE

$$dX_t = f(X,t)dt + \sum_{i=1}^{n} g_i(X,t)dW_i.$$

Define

$$Y_t = \psi(X_t, t).$$

By the normal rules of calculus, we get

$$\begin{aligned}
d\psi(X_t,t) &= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial X_t}(dX_t) + \frac{1}{2}\frac{\partial^2\psi}{\partial X_t^2}(dX_t)^2 \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial X_t}\left(f(X,t)dt + \sum_{i=1}^{n} g_i(X,t)dW_i\right) + \frac{1}{2}\frac{\partial^2\psi}{\partial X_t^2}\left(f(X,t)dt + \sum_{i=1}^{n} g_i(X,t)dW_i\right)^2 \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial X_t}\left(f(X,t)dt + \sum_{i=1}^{n} g_i(X,t)dW_i\right) + \frac{1}{2}\frac{\partial^2\psi}{\partial X_t^2}\left(f^2(X,t)dt^2 + f(X,t)\sum_{i=1}^{n} g_i(X,t)dW_i dt + \sum_{i=1}^{n} g_i^2(X,t)dW_i^2\right).
\end{aligned}$$

If we let

$$dt \times dt = 0$$

$$dW_i \times dt = 0$$

$$dW_i \times dW_j = \begin{cases} dt & : i = j \\ 0 & : i \neq j \end{cases}$$

then this simplifies to

$$d\psi(x,t) = \left(\frac{\partial\psi}{\partial t} + f(x,t)\frac{\partial\psi}{\partial x} + \frac{1}{2}\sum_{i=1}^{n} g_i^2(x,t)\frac{\partial^2\psi}{\partial x^2}\right)dt + \frac{\partial\psi}{\partial x}\sum_{i=1}^{n} g_i(x,t)dW_i$$

which is once again Ito's Rules. Thus we can think of Ito's Rules is saying that dt^2 is sufficiently small, dt and dW_i are uncorrelated, and $dW_i^2 = dt$ which means that the differential Wiener process squared is a deterministic process. In fact, we can formalize this idea as the defining property of Brownian motion. This is captured in Levy Theorem which will be stated in 6.7.

4.5 Wiener Process Calculus Summarized

Take the SDE

$$dx = f(x,t)dt + \sum_{i=1}^{n} g_i(x,t)dW_i$$

where $W_i(t)$ is a standard Brownian motion. We have showed that Ito's Rules could be interpreted as:

$$dt \times dt = 0$$

$$dW_i \times dt = 0$$

$$dW_i \times dW_j = \begin{cases} dt & : i = j \\ 0 & : i \neq j \end{cases}$$

and thus if $y = \psi(x,t)$, Ito's rules can be written as

$$dy = \frac{\partial \psi}{\partial t}dt + \frac{\partial \psi}{\partial x}dx + \frac{1}{2}\frac{\partial^2 \psi}{\partial x^2}(dx)^2$$

where, if we plug in dx, we get

$$dy = d\psi(x,t) = \left(\frac{\partial \psi}{\partial t} + f(x,t)\frac{\partial \psi}{\partial x} + \frac{1}{2}\sum_{i=1}^{n} g_i^2(x,t)\frac{\partial^2 \psi}{\partial x^2}\right)dt + \frac{\partial \psi}{\partial x}\sum_{i=1}^{n} g_i(x,t)dW_i$$

The solution is given by the integral form:

$$x(t) = x(0) + \int_0^t f(x(s))ds + \int_0^t \sum_{i=1}^{m} g_i(x(s))dW_i.$$

Note that we can also generalize Ito's lemma to the multidimensional $\mathbf{X} \in \mathbb{R}^n$ case:

$$d\psi(\mathbf{X}) = \left\langle \frac{\partial \psi}{\partial \mathbf{X}}, f(\mathbf{X}) \right\rangle dt + \sum_{i=1}^{m} \left\langle \frac{\partial \psi}{\partial \mathbf{X}}, g_i(\mathbf{X}) \right\rangle dW_i + \frac{1}{2}\sum_{i=1}^{m} g_i(\mathbf{X})^T \nabla^2 \psi(\mathbf{X}) g_i(\mathbf{X}) dt$$

There are many other facts that we will state but not prove. These are proven using Ito's Rules. They are as follows:

1. Product Rule: $d(X_tY_t) = X_tdY + Y_tdX + dXdY$.
2. Integration By Parts: $\int_0^t X_t dY_t = X_tY_t - X_0Y_0 - \int_0^t Y_t dX_t - \int_0^t dX_t dY_t$.
3. $\mathbb{E}\left[(W(t) - W(s))^2\right] = t - s$ for $t > s$.
4. $\mathbb{E}[W(t_1)W(t_2)] = \min(t_1, t_2)$.
5. Independent Increments: $\mathbb{E}\left[(W_{t_i} - W_{s_1})(W_{t_2} - W_{s_2})\right] = 0$ if $[t_1, s_1]$ does not overlap $[t_2, s_2]$.
6. $\mathbb{E}\left[\int_0^t h(t)dW_t\right] = \mathbb{E}\left[h(t)dW_t\right] = 0$.
7. Ito Isometry: $\mathbb{E}\left[\left(\int_0^T X_t dW_t\right)\right] = \mathbb{E}\left[\int_0^T X_t^2 dt\right]$

4.5.1 Example Problem: Geometric Brownian Motion

Look at the example stochastic process

$$dx = \alpha x dt + \sigma x dW.$$

To solve this, we start with our intuition from Newton's Calculus that this may be an exponential growth process. Thus we check Ito's equation on the logarithm $\psi(x,t) = \ln(x)$ for this process is,

$$\begin{aligned} d(\ln x) &= \left(0 + (\alpha x)\left(\frac{1}{x}\right) - \frac{1}{2}\left(\sigma^2 x^2\right)\left(\frac{1}{x^2}\right)\right)dt + \left(\frac{1}{x}\right)(\sigma x)\, dW \\ d(\ln x) &= \left(\alpha - \frac{1}{2}\sigma^2\right)dt + \sigma dW. \end{aligned}$$

Thus by taking the integral of both sides we get

$$\ln(x) = \left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W(t)$$

and then expnentiating both sides

$$x(t) = e^{\left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W(t)}.$$

Notice that since the Wiener process $W(t) \sim N(0,t)$, the log of x is distributed normally as $N(\left(\alpha - \frac{1}{2}\sigma^2\right)t, \sigma^2 t)$. Thus $x(t)$ is distributed as what is known as the log-normal distribution.

4.6 Kolmogorov Forward Equation Derivation

The Kolmogorov Forward Equation, also known to Physicists as the Fokker-Planck Equation, is important because it describes the time evolution of the probability density function. Whereas the stochastic differential equation describes how one trajectory of the stochastic processes evolves, the Kolmogorov Forward Equation describes how, if you were to be running many different simulations of the trajectory, the percent of trajectories that are around a given value evolves with time.

We will derive this for the arbitrary drift process:

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x)dW_i.$$

Let $\psi(x)$ be an arbitrary time-independent transformation of x. Applying Ito's lemma for an arbitrary function we get

$$d\psi(x,t) = \left(f(x,t)\frac{\partial \psi}{\partial x} + \frac{1}{2}\sum_{i=1}^{n} g_i^2(x,t)\frac{\partial^2 \psi}{\partial x^2} \right) dt + \frac{\partial \psi}{\partial x}\sum_{i=1}^{n} g_i(x,t)dW_i$$

Take the expectation of both sides. Because expectation is a linear operator, we can move it inside the derivative operator to get

$$\frac{d}{dt}\mathbb{E}\left[\psi(x)\right] = \mathbb{E}\left[\frac{\partial \psi}{\partial x}f(x)\right] + \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m}\frac{\partial^2 \psi}{\partial x^2}g_i^2(x)\right].$$

Notice that $\mathbb{E}\left[dW_i\right] = 0$ and thus the differential Wiener terms dropped out.

Recall the definition of expected value is

$$\mathbb{E}[\psi(x)] = \int_{-\infty}^{\infty} \rho(x,t)\psi(x)dx$$

where $\rho(x,t)$ is the probability density of equaling x at a time t. Thus we get that the first term as

$$\frac{d}{dt}\mathbb{E}\left[\psi(x)\right] = \int_{-\infty}^{\infty}\frac{\partial \rho}{\partial t}\psi(x)dx$$

For the others, notice

$$\mathbb{E}\left[\frac{\partial \psi}{\partial x}f(x)\right] = \int_{-\infty}^{\infty}\rho(x,t)\frac{\partial \psi}{\partial x}f(x)dx.$$

We rearrange terms doing integration by parts. Let $u = \rho f$ and $dv = \frac{\partial \psi}{\partial x}$. Thus $du = \frac{\partial(\rho f)}{\partial x}$ and $v = \psi$. Therefore we get that

$$\mathbb{E}\left[\frac{\partial \psi}{\partial x}f(x)\right] = \left[\rho f \psi\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\frac{\partial(\rho f)}{\partial x}\psi(x)dx.$$

In order for the probability distribution to be bounded (which it must be: it must integrate to 1), ρ must vanish at both infinities. Thus, assuming bounded expectation, $\mathbb{E}\left[\frac{\partial \psi}{\partial x} f(x)\right] < \infty$, we get that

$$\mathbb{E}\left[\frac{\partial \psi}{\partial x} f(x)\right] = -\int_{-\infty}^{\infty} \frac{\partial(\rho f)}{\partial x}\psi(x)dx.$$

The next term we manipulate similarly,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \psi^2}{\partial x^2} g_i^2(x)\right] &= \int_{-\infty}^{\infty} \frac{\partial^2 \psi}{\partial x^2} g_i^2(x)\rho(x,t)dx \\
&= \left[\rho g^2 \frac{\partial \psi}{\partial x}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{\partial\left(g_i^2(x)\rho(x,t)\right)}{\partial x}\frac{\partial \psi}{\partial x}dx \\
&= \left[\rho g^2 \frac{\partial \psi}{\partial x}\right]_{-\infty}^{\infty} - \left[\frac{\partial\left(\rho g^2\right)}{\partial x}\psi\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty} \frac{\partial^2\left(g_i^2(x)\rho(x,t)\right)}{\partial x^2}\psi(x)dx \\
&= \tfrac{1}{2} \int_{-\infty}^{\infty} \frac{\partial^2\left(g_i^2(x)\rho(x,t)\right)}{\partial x^2}\psi(x)dx
\end{aligned}$$

where we note that, at the edges, the derivative of ρ converges to zero since ρ converges to 0 and thus the constant terms vanish. Thus we get that

$$\int_{-\infty}^{\infty} \frac{\partial \rho}{\partial t}\psi(x)dx = -\int_{-\infty}^{\infty} \frac{\partial(\rho f)}{\partial x}\psi(x)dx + \frac{1}{2}\sum_i \int_{-\infty}^{\infty} \frac{\partial(g_i^2\rho)}{\partial x}\psi(x)dx$$

which we can re-write as

$$\int_{-\infty}^{\infty} \left(\frac{\partial \rho}{\partial t} + \frac{\partial(\rho f)}{\partial x} - \frac{1}{2}\sum_i \frac{\partial(g_i^2\rho)}{\partial x}\right)\psi(x)dx = 0.$$

Since $\psi(x)$ is arbitrary, let $\psi(x) = I_A(x)$, the indicator function for the set A:

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0 & o.w. \end{cases}$$

Thus we get that

$$\int_A \left(\frac{\partial \rho}{\partial t} + \frac{\partial(\rho f)}{\partial x} - \frac{1}{2}\sum_i \frac{\partial(g_i^2\rho)}{\partial x}\right)dx = 0$$

for any arbitrary $A \subseteq \mathbb{R}$. Notice that this implies that the integrand must be identically zero. Thus

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho f)}{\partial x} - \frac{1}{2}\sum_i \frac{\partial(g_i^2\rho)}{\partial x} = 0,$$

which we re-arrange as

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x)\rho(x,t)] + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2}{\partial x^2}\left[g_i^2(x)\rho(x,t)\right],$$

which is the Forward Kolmogorov or the Fokker-Planck equation.

4.6.1 Example Application: Ornstein–Uhlenbeck Process

Consider the stochastic process

$$dx = -xdt + dW_t,$$

where W_t is Brownian motion. The Forward Kolmogorov Equation for this SDE is thus

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}(x\rho) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\rho(x,t)$$

Assume that the initial conditions follow the distribution u to give

$$\rho(x,0) = u(x)$$

and the boundary conditions are absorbing at infinity. To solve this PDE, let $y = xe^t$ and apply Ito's lemma

$$dy = xe^t dt + e^t dx = e^t dW$$

and notice this follows has the Forward Kolmogorov Equation

$$\frac{\partial \rho}{\partial t} = \frac{e^{2t}}{2}\frac{\partial^2 \rho}{\partial y^2}.$$

which is a simple form of the Heat Equation. If $\rho(x,0) = \delta(x)$, the Dirac-δ function, then we know this solves as a Gaussian with diffusion constant $\frac{e^{2t}}{2}$ to give us

$$\rho(y,t) = \frac{1}{\sqrt{2\pi e^{2t}t}}e^{-\frac{y^2}{2e^{2t}t}}.$$

and thus $y \sim N(0, e^{2t}t)$. To get the probability density function in terms of x, we would simply do the pdf transformation as described in 2.9.

4.7 Stochastic Stability

The idea of stochastic stability is that linear stability analysis for deterministic systems generalizes to stochastic systems. To see this, look at the equation

$$dx = axdt + dW_t.$$

Notice that

$$\frac{d\mathbb{E}[x]}{dt} = a\mathbb{E}[x]$$

and thus

$$\mathbb{E}[x] = \mathbb{E}[x_0]\, e^{at}$$

which converges iff $a < 0$. Also notice that

$$\begin{aligned} dx^2 &= 2xdx + dt \\ &= 2x(axdt + dW_t) + dt \\ &= (2ax^2 + 1)dt + dW_t \end{aligned}$$

and thus

$$\frac{d\mathbb{E}\left[x^2\right]}{dt} = 2a\mathbb{E}\left[x^2\right] + 1$$

which gives

$$\mathbb{E}\left[x^2\right] = \left(\mathbb{E}\left[x_0^2\right] + \frac{1}{2a}\right) e^{2at}$$

which converges iff $a < 0$. This isn't a full proof, but it motivates the idea that if the deterministic coefficient is less than 0, then, just as in the deterministic case, the system converges and is thus stable.

4.8 Fluctuation-Dissipation Theorem

Take the SDE

$$dX = f(X,t)dt + g(X,t)dW_t.$$

If there exists a stable steady state, linearize the system around a steady state

$$dX = J_f(X_{ss},t)dt + g(X_{ss},t)dW_t,$$

where J_f is the Jacobian of f defined as

$$J_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

where $\frac{\partial f_i}{\partial x_j}$ is the partial derivative of the ith component of f by the jth component of x. The Fluctuation-Dissipation Theorem tells us that the variance-covariance matrix of the variables,

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_1,X_2) & \ddots & & Cov(X_2,X_n) \\ \vdots & & \ddots & \vdots \\ Cov(X_1,X_n) & Cov(X_2,X_n) & \dots & Var(X_n) \end{bmatrix} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \ddots & & \sigma_{X_2X_n} \\ \vdots & & \ddots & \vdots \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdots & \sigma_{X_n}^2 \end{bmatrix}$$

can be found at the steady state using the formula

$$J_f(X_{ss},t)\Sigma(X_{ss},t)+\Sigma(X_{ss},t)J_f^T(X_{ss},t)=-g^2(X_{ss},t).$$

5 Computational Simulation of SDEs

In this chapter we outline basic procedures for the computational simulation of SDEs. For a good introduction to computational simulation of SDEs, refer to Higham's *An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations.* For a more complete reference on higher-order methods, see Kloeden's *Numerical Solution of SDE Through Computer Experiments* and Iacus's *Simulation and Inference for Stochastic Differential Equations.* For the reference on Rossler's High Strong Order Runge-Kutta methods, see *Runge–Kutta Methods for the Strong Approximation of Solutions of Stochastic Differential Equations.* For adaptive timestepping and efficient implementation, see Rackauckas' *Adaptive Methods for Stochastic Differential Equations via Natural Embeddings and Rejection Sampling with Memory.* More efficient high order methods will be published soon.

5.1 The Stochastic Euler Method - Euler-Maruyama Method

An intuitive way to start the computational simulation of SDEs is to look at a straight-forward generalization of Euler's method for ODEs to SDEs. First, let us recap Euler's method for ODEs. Take the ODE

$$x'(t)=f(x,t).$$

Writing this out we get

$$\frac{dx}{dt}=f(x,t)$$

which we can represent as

$$dx=f(x,t)dt.$$

Euler's method is to approximate the solution using "discrete-sized dt's". Thus we take some fixed small constant Δt. We say that

$$x(t+\Delta t)-x(t)=f(x,t)\Delta t,$$

and thus

$$x(t+\Delta t)=x(t)+f(x,t)\Delta t$$

defines a recursive solution for the value of x at a time t given values of x at previous times. All that is left for the approximation is some initial condition needs to be started, such as $x(0)=y$ and thus problem is solved iteratively.

Now we take the stochastic differential equation

$$dX=f(X,t)dt+g(X,t)dW_t.$$

Once again, we define some small fixed constant Δt and write

$$\begin{aligned} X(t+\Delta t) - X(t) &= f(X,t)\Delta t + g(X,t)\Delta W_t, \\ X(t+\Delta t) &= X(t) + f(X,t)\Delta t + g(X,t)\Delta W_t. \end{aligned}$$

Recall from 4.2 that an interval Δt of the Wiener process is distributed as

$$\Delta W_t = (dW_{t+\Delta t} - dW_t) \sim N(0, dt).$$

Thus let η_i be a standard normal random variable: $\eta_i \sim N(0,1)$. Notice that $\Delta W_t = \eta_i \sqrt{dt}$. Thus we can write

$$X(t+\Delta t) = X(t) + f(X,t)\Delta t + \sqrt{\Delta t} g(X,t)\eta_i.$$

The way to interpret this is that, for each interval i, we sample a standard normal random variable η_i, and iterate one step using this equation, then sample another standard normal random variable and iterate again! So this shows the strong connection between Gaussian processes and the Wiener process. This method for iteratively solving a stochastic differential equation is known as the Euler-Maruyama Method.

5.2 A Quick Look at Accuracy

Recall that for Euler's method, when using a time discretization of size Δt, we have the the accuracy of the solution is $\mathcal{O}(\Delta t)$, that is, each step approximates the change in x well, give or take an amount Δt. This is justified using Taylor series expansions, since

$$x(t+\Delta t) = x(t) + x'(t)\Delta t + \mathcal{O}(\Delta t^2)$$

and thus, if x is the real solution and X is the approximation,

$$\begin{aligned} x(t+\Delta t) - x(t) &= x'(t)\Delta t + \mathcal{O}(\Delta t^2), \\ X(t+\Delta t) - X(t) &= X'(t)\Delta t. \end{aligned}$$

We can then write that

$$|x(t) - X(t)| \leq C\Delta t$$

and therefore we are confident that our approximation X converges to the real solution x linearly by Δt (we are not confident in our answer for variations as small as Δt^2 because we dropped those terms off!).

We may think that this simply generalizes. However, it gets more complicated quite quickly. First of all, how do we define the difference? There are two ways to define the difference. One way is known as the Strong Convergence. The strong convergence is the expected difference between "the real solution" and the approximation solution, that is if x is the real solution and X is our approximation, then the strong convergence is the factor γ defining

$$\mathbb{E}\left[|x - X|\right] \leq C\Delta t^{\gamma}.$$

We can think of a different type of convergence, known as the weak convergence, as

$$|\mathbb{E}[x] - \mathbb{E}[X]| \leq C\Delta t^{\beta}.$$

Notice how different these ideas of convergence are. Weak convergence means that the average trajectory we computationally simulate does Δt^{β} good, where as strong convergence means that every trajectory does Δt^{γ} good. Thus if we are looking at properties defined by ensembles of trajectories, then weak convergence is what we are looking at. However, if we want to know how good a single trajectory is, we have to look at the strong convergence.

Here comes the kicker. For the Euler-Maruyama method, it has a strong convergence of order $\frac{1}{2}$ and a weak convergence of order 1. That's to say it has a slower convergence than the Euler method, and the average properties only converge as fast as the Euler method! Thus, in practice, this method is not very practical given its extremely slow convergence.

5.3 Milstein's Method

To understand how to make the algorithms better, we have to understand what went wrong. Why was the convergence of the stochastic method slower than that of the deterministic method? The answer comes from the Taylor series expansion. Notice that in the stochastic Taylor series (expansion using Ito's Rules), the second order terms matter for the first-order effects because $(dW_t)^2 = dt$! Thus, accounting for the second order Taylor series term instead write $\frac{\partial g}{\partial x} = g_x$ to get the method

$$X(t+\Delta t) = X(t) + \left(f(X,t) - \frac{1}{2}g(X,t)g_x(X,t)\right)\Delta t + \sqrt{\Delta t}g(X,t)\eta_i + \frac{\Delta t}{2}g(X,t)g_x(X,t)\eta_i^2.$$

Notice that we have added a $-\frac{1}{2}g(X,t)g_x(X,t)dt$ term in order to cancel out the $\frac{1}{2}g(X,t)g_x(X,t)\,(dW_t)^2$ term in the stochastic Taylor series expansion. Notice too that we only sample one random number, but the second order term uses that random number squared. This method is known as Milstein's method. As you may have guessed, since it now accounts for all order 1 effects, it has order 1 strong and weak convergence.

5.4 KPS Method

As Milstein's method is still only as good as the deterministic Euler method, I wish to give one last method for the purpose of simulating SDEs for applications. For most applications, you want you convergence order greater than 1. The simplest of such methods is the KPS method which as a strong order of convergence of 1.5 and a weak order of convergence 2. The method can be written as follows:

$$
\begin{aligned}
X(t+\Delta t) &= X(t) + f\Delta t + g\Delta W_t + \frac{\Delta t}{2} g g_x \left((\Delta W_t)^2 - \Delta t \right) \\
&+ g f_x \Delta U_t + \frac{1}{2}\left(f f_x + \frac{1}{2} g^2 f_{xx} \right) \Delta t^2 \\
&+ \left(f g_x + \frac{1}{2} g^2 g_{xx} \right) (\Delta W_t \Delta t - \Delta U_t) \\
&+ \frac{1}{2} g \left(g g_x \right)_x \left(\frac{1}{3} (\Delta W_t)^2 - \Delta t \right) \Delta W_t
\end{aligned}
$$

where

$$\Delta W_t = \sqrt{\Delta t}\eta_i$$

and

$$\Delta U_t = \int_t^{t+\Delta t} \int_t^s dW_s ds$$

can be written as

$$\Delta U_t = \frac{1}{3}\Delta t^3 \lambda_i$$

where $\lambda_i \sim N(0,1)$, a standard normal random variable (that is not η_i!).

5.5 High Strong Order Runge-Kutta Methods

Using a colored root tree analysis, Rößler was able to develop a systematic method for developing order 1.5 multi-step stochastic Runge-Kutta schemes. These resulted in less computational steps than the KPS schemes, and the number of steps grows much slower as the Ito dimension increases than in the KPS schemes. They also have the advantage of being more structurally simple, making them the faster method in both implementation and runtime, and have slower growth in the number of coefficients as the number of Ito dimensions grows. Generalizations of the Rößler methods Stratanovich integration are also derived in his paper, and our methods will trivially generalize as well. He showed that the the Runge-Kutta methods

$$
\begin{aligned}
U_{n+1} &= U_n + \sum_{i=1}^{s} \alpha_i f\left(t_n + c_i^{(0)} \Delta t, H_i^{(0)} \right) \Delta t + \\
&\sum_{i=1}^{s} \left(\beta_i^{(1)} I_{(1)} + \beta_i^{(2)} \frac{I_{(1,1)}}{\sqrt{\Delta t}} + \beta_i^{(3)} \frac{I_{(1,0)}}{\Delta t} + \beta_i^{(4)} \frac{I_{(1,1,1)}}{\Delta t} \right) g\left(t_n + c_i^{(1)} \Delta t, H_i^{(1)} \right)
\end{aligned}
\tag{5.1}
$$

with stages

$$H_i^{(0)} = U_n + \sum_{j=1}^{s} A_{ij}^{(0)} f\left(t_n + c_j^{(0)}\Delta t, H_j^{(0)}\right)\Delta t + \sum_{j=1}^{s} B_{ij}^{(0)} g\left(t_n + c_j^{(1)}\Delta t, H_j^{(1)}\right)\frac{I_{(1,0)}}{\Delta t}, \tag{5.2}$$

$$H_i^{(1)} = U_n + \sum_{j=1}^{s} A_{ij}^{(1)} f\left(t_n + c_j^{(0)}\Delta t, H_j^{(0)}\right)\Delta t + \sum_{j=1}^{s} B_{ij}^{(1)} g\left(t_n + c_j^{(1)}\Delta t, H_j^{(1)}\right)\sqrt{\Delta t} \tag{5.3}$$

if they satisfy a set of order conditions. The coefficients $\left(A_0, B_0, \beta^{(i)}, \alpha\right)$ must satisfy the following order conditions to achieve order .5:

1. $\alpha^T e = 1$
2. ${\beta^{(1)}}^T e = 1$
3. ${\beta^{(2)}}^T e = 0$
4. ${\beta^{(3)}}^T e = 0$
5. ${\beta^{(4)}}^T e = 0$

additionally, for order 1:

1. ${\beta^{(1)}}^T B^{(1)} e = 0$
2. ${\beta^{(2)}}^T B^{(1)} e = 1$
3. ${\beta^{(3)}}^T B^{(1)} e = 0$
4. ${\beta^{(4)}}^T B^{(1)} e = 0$

and lastly for order 1.5:

1. $\alpha^T A^{(0)} e = \frac{1}{2}$
2. $\alpha^T B^{(0)} e = 1$
3. $\alpha^T \left(B^{(0)} e\right)^2 = \frac{3}{2}$
4. ${\beta^{(1)}}^T A^{(1)} e = 1$
5. ${\beta^{(2)}}^T A^{(1)} e = 0$
6. ${\beta^{(3)}}^T A^{(1)} e = -1$
7. ${\beta^{(4)}}^T A^{(1)} e = 0$
8. ${\beta^{(1)}}^T \left(B^{(1)} e\right)^2 = 1$
9. ${\beta^{(2)}}^T \left(B^{(1)} e\right)^2 = 0$
10. ${\beta^{(3)}}^T \left(B^{(1)} e\right)^2 = -1$
11. ${\beta^{(4)}}^T \left(B^{(1)} e\right)^2 = 2$
12. ${\beta^{(1)}}^T \left(B^{(1)} \left(B^{(1)} e\right)\right) = 0$
13. ${\beta^{(2)}}^T \left(B^{(1)} \left(B^{(1)} e\right)\right) = 0$
14. ${\beta^{(3)}}^T \left(B^{(1)} \left(B^{(1)} e\right)\right) = 0$
15. ${\beta^{(4)}}^T \left(B^{(1)} \left(B^{(1)} e\right)\right) = 1$

16. $\frac{1}{2}\beta^{(1)^T}\left(A^{(1)}\left(B^{(0)}e\right)\right)+\frac{1}{3}\beta^{(3)^T}\left(A^{(1)}\left(B^{(0)}e\right)\right)=0$

These methods are the (Rößler) SRI methods. We will refer to the algorithms by the tuple of 44 coefficients $\left(A_0, B_0, \beta^{(i)}, \alpha\right)$. Note that this method can be easily extended to multiple Ito dimensions in the case of diagonal noise with similar results. We only focus on a single Ito dimension for simplicity of notation (though our results will extend to higher Ito dimensions as well in the trivial manner). To satisfy the conditions, Rößler proposed the following scheme known as SRIW1:

$c^{(0)}$	$A^{(0)}$	$B^{(0)}$
$c^{(1)}$	$A^{(1)}$	$B^{(1)}$
	α^T	$\beta^{(1)^T}$ $\beta^{(2)^T}$
		$\beta^{(3)^T}$ $\beta^{(4)^T}$
	$\tilde{\alpha}^T$	$\tilde{\beta}^{(3)^T}$ $\tilde{\beta}^{(4)^T}$

0												
$\frac{3}{4}$	$\frac{3}{4}$				$\frac{3}{2}$							
0	0	0			0	0						
0	0	0	0		0	0	0					
0												
$\frac{1}{4}$	$\frac{1}{4}$				$\frac{1}{2}$							
1	1	0			-1	0						
$\frac{1}{4}$	0	0	$\frac{1}{4}$		-5	3	$\frac{1}{2}$					
	$\frac{1}{3}$	$\frac{2}{3}$	0	0	-1	$\frac{4}{3}$	$\frac{2}{3}$	0	-1	$\frac{4}{3}$	$-\frac{1}{3}$	0
					2	$-\frac{4}{3}$	$-\frac{2}{3}$	0	2	$\frac{5}{3}$	$-\frac{2}{3}$	1
	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0

In the case where noise is additive, the methods can be vastly simplified to

$$U_{n+1} = U_n + \sum_{i=1}^{s} \alpha_i f\left(t_n + c_i^{(0)}\Delta t, H_i^{(0)}\right)\Delta t + \sum_{i=1}^{s}\left(\beta_i^{(1)} I_{(1)} + \beta_i^{(2)}\frac{I_{(1,0)}}{\Delta t}\right) g(t_n + c_i^{(1)}\Delta t) \tag{5.4}$$

with stages

$$H_i^{(0)} = U_n + \sum_{j=1}^{s} A_{ij}^{(0)} f\left(t_n + c_j^{(0)}\Delta t, H_j^{(0)}\right)\Delta t + \sum_{j=1}^{s} B_{ij}^{(0)} g\left(t_n + c_j^{(1)}\Delta t\right)\frac{I_{(1,0)}}{\Delta t} \tag{5.5}$$

The coefficients $\left(A_0, B_0, \beta^{(i)}, \alpha\right)$ must satisfy the conditions for order 1:

1. $\alpha^T e = 1$
2. $\beta^{(1)^T} e = 1$
3. $\beta^{(2)^T} e = 0$

and the additional conditions for order 1.5:

1. $\alpha^T B^{(0)} e = 1$
2. $\alpha^T A^{(0)} e = \frac{1}{2}$
3. $\alpha^T \left(B^{(0)} e\right)^2 = \frac{3}{2}$
4. $\beta^{(1)^T} c^{(1)} = 1$
5. $\beta^{(2)^T} c^{(1)} = -1$

where $c^{(0)} = A^{(0)} e$ with $f \in C^{1,3}(\mathcal{I} \times \mathbb{R}^d, \mathbb{R}^d)$ and $g \in C^1(\mathcal{I}, \mathbb{R}^d)$. These are the (Rößler) SRA methods. From these conditions he proposed the following Strong Order 1.5 scheme known as SRA1:

$c^{(0)}$	$A^{(0)}$	$B^{(0)}$	
	α^T	$\beta^{(1)^T}$	$\beta^{(2)^T}$

0					
$\frac{3}{4}$	$\frac{3}{4}$		$\frac{1}{2}$		
	$\frac{1}{3}$	$\frac{2}{3}$	1 0	-1	1

5.6 Timestep Adaptivity

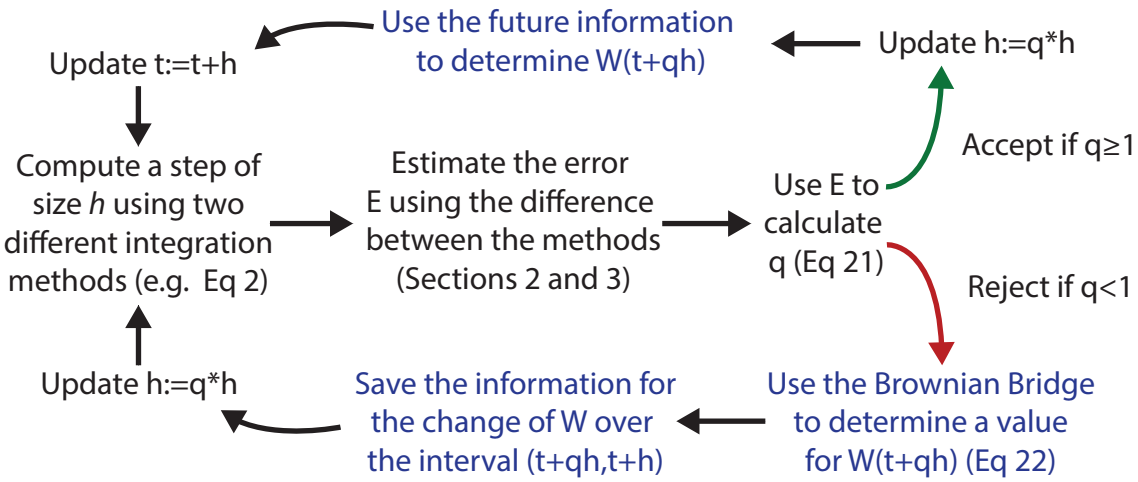
The efficient methods for timestep adaptivity are discussed in Rackauckas & Nie 2017. This section is pulled almost entirely from the paper *Adaptive Methods for Stochastic Differential Equations via Natural Embeddings and Rejection Sampling with Memory.* Two steps are required for building an adaptive method. First, and error estimate has to be derived. Then, from that error estimate, one has to choose the accept the step or reject the current step (and change Δt). For the error

estimate, it was shown that a natural error estimator exists for any high-order SRK method. A simplified version is simply:

$$\begin{aligned} E &= |\delta E_D + E_N| \\ &\leq \delta \left| \Delta t \sum_{i=1}^{s} f\left(t_n + c_i^{(0)} \Delta t, H_i^{(0)}\right) \right| \\ &\quad + \left| \sum_{i=1}^{s} \left(\beta_i^{(3)} \frac{I_{(1,0)}}{\Delta t} + \beta_i^{(4)} \frac{I_{(1,1,1)}}{\Delta t} \right) g\left(t_n + c_i^{(1)} \Delta t, H_i^{(1)}\right) \right| \end{aligned} \tag{5.6}$$

where s is the number of stages and δ is a user-chosen balance between determinsitic and noise error in the error estimate. A similar summation gives the estimate for additive noise equations.

With the error estimate, the overall algorithm is depicted as:



As in deterministic adaptive stepping algorithms, we define

$$q = \left(\frac{\epsilon h}{\gamma E} \right)^2 \tag{5.7}$$

where ϵ is the chosen suggested error, γ is a penalty factor (in deterministic methods it is often taken as $\gamma = 2$), and E is an error estimate. The step logic is the following:

1. If $q < 1$, reject the initial choice of h and repeat the calculation with qh
2. If $q \geq 1$, then accept the computed step and change h to $\min(h_{max}, qh)$ where h_{max} is chosen by the user.

For the acceptance/rejectance of the step, care must be taken to not bias the Wiener process. If extreme values of ΔW are always thrown out then the sample properties are no longer valid. Thus we must always keep any calculated value of ΔW. The procedure has to be enhanced as follows. First, propose a step with ΔW^P and ΔZ^P for a timestep h. If these are rejected, we wish to instead attempt a step of size qh. Thus we need to sample a new value at $W(t_n + qh)$ using the known values of $W(t_n)$ and $W(t_n+h)$. To do so, we use the result that if $W(0) = 0$ and $W(h) = L$, then by the properties of the Brownian Bridge we calculate that for $q \in (0, 1)$, $W(qh) \sim N(qL, (1 - q)qh)$. We then propose to step by qh and take the random numbers $\Delta W = W(qh)$ and $\Delta Z = Z(qh)$ found via their appropriate distribution from the Brownian bridge. We then store the modified versions of ΔW^P and ΔZ^p. Notice that since we have moved ΔW in the qh timestep, what remains is $\overline{\Delta W} = \Delta W^P - \Delta W$ and $\overline{\Delta Z} = \Delta Z^P - \Delta Z$ as the change in the Brownian path from qh to h. We then store the values $L = \left(1 - qh, \overline{\Delta W}, \overline{\Delta Z}\right)$ as a 3-tuple in a stack to represent that after our current calculations, over the next interval of size L_1, the Brownian process W will change by L_2 and the process Z will change by L_3. Thus when we finally get to $t_n + qh$, we look at these values to tell us how the Brownian path changes over the next $1 - qh$ time units. By doing so, we will effectively keep the properties of the Brownian path while taking arbitrary steps. This leads to the RSwM1 algorithm. More complex handling of the timestep, but using the same general idea, leads to the more efficient RSwM3 algorithm. Note that these stepping routines are compatible with any high order SDE method as long as some error estimator exists.

5.7 Simulation Via Probability Density Functions

Another method for simulating SDEs is to use the Forward Kolmogorov Equation. Recall that the SDE

$$dX = f(X, t)dt + g(X, t)dW_t$$

has a probability distribution function ρ which satisfies the PDE

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x)\rho(x,t)] + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2}{\partial x^2}\left[g_i^2(x)\rho(x,t)\right].$$

This can thus be a way to solve for the probability density using computational PDE solvers. If we have an initial condition, $X(0) = x_0$, then this corresponds to the initial condition $\rho(x, 0) = \delta(x-x_0)$ where δ is the Dirac-δ function. This can be particularly useful for first-passage time problems, where we can set the boundary conditions as absorbing: $\rho(a, t) = \rho(b, t) = 0$, and thus ρ is the total probability distribution of trajectories that have not been absorbed. Likewise, if we want a boundary condition where we say "trajectories reflect off the point a", then we simply use a condition $\frac{\partial \rho}{\partial x}|_{x=a} = 0$.

However, though this method may be enticing to those who are experienced with computational PDE solvers, this method is not the holy grail because it cannot simulate trajectories. If you need

the trajectory of a single instantiation of the process, for example, "what does the stock price over time look like?", you cannot use this method.

6 Measure-Theoretic Probability for SDE Applications

Here we will get down and dirty with some measure-theoretic probability in order to define conditional expectations, martingales, and get further into stochastic calculus.

6.1 Probability Spaces and σ-algebras

The idea is that the probability space is simply a set of all of the events, a collection of subsets of events (where if two events are in a subset, it is the event that both happen), and a measure which is a function that gives the number of how probable a collection of subsets are. These are thus define in set-theoretic terms so that they correspond to the normal rules of calculus that you would expect.

Definition: Let Ω be a nonempty set, and let $\mathcal{F}$ be a collection of subsets of Ω. We say $\mathcal{F}$ is a σ-algebra if

1. $\emptyset \in \mathcal{F}$,
2. $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$,
3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Propositions:

1. $\Omega \in \mathcal{F}$.
2. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$.

To prove proposition 1, notice that since $\emptyset \in \mathcal{F}$, by property 2 the compliment of the empty set, Ω, must be in $\mathcal{F}$. Proposition two follows by DeMorgan's Law. By property 3 the union is in $\mathcal{F}$, and so the union of the compliments are in $\mathcal{F}$ by applying property 2 to each component. Thus the compliment of the union of the complements is in $\mathcal{F}$ by property 2. By DeMorgan's Law, the intersection of the complement of the complements, or simply the intersection, must be in $\mathcal{F}$ proving the proposition.

Definition: Given Ω and $\mathcal{F}$, a probability measure P is a function $\mathcal{F} \to [0, 1]$ such that:

1. $P(\Omega) = 1$
2. $A_1, A_2, , \ldots$ is a sequence of disjoint subsets in $\mathcal{F} \Rightarrow P(\bigcup_{n=1}^{\infty} A_n) = \sum_{i=1}^{\infty} P(A_n)$ (Countable additivity)

Definition: $(\Omega, \mathcal{F}, P)$ is a probability space.

Proposition: $A_1 \subset A_2 \ldots, \subset A_i, \ldots \in \mathcal{F}$, then $P(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$ and $P(\bigcap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$.

These propositions come directly from the properties of measures. See a measure theory text for more information if you do not know and you're Curious George. There are many possibilities for what kind of sample space, Ω, we may want to consider:

1. Ω can be finite.
2. Ω can be countably infinite.
3. Ω can be uncountably infinite.

To determine $\mathcal{F}$ we might want to consider the *powerset* of Ω, which is composed of all possible subsets of Ω

1. If Ω is finite: the number of subsets is $2^{|\Omega|}$
2. If Ω is infinite: the powerset of S is uncountable. Thus, in order to avoid paradoxes such as the Banach-Tarski paradox, we use a σ-algebra.

Suppose $\Omega = \mathbb{R} = (-\infty, \infty)$ and for each $a, b \in \mathbb{R}$ where $a < b$, $(a, b) \in \mathcal{F}$. The *Borel Set* is the smallest σ-algebra containing open intervals. We almost always assume this set when working with σ-algebras.

6.1.1 Example: Uniform Measure

Suppose we are given the uniform measure on $[0, 1]$

$$P[a, b] = b - a, \ 0 \leq a \leq b \leq 1.$$

where$\Omega = [0, 1]$ and $\mathcal{F}$ is the set of all closed intervals of Ω. We note here that the Borel σ-algebra, $\mathcal{B}[0, 1]$, which is $\mathcal{F}$, also contains all open intervals as constructed by

$$(a, b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n} \right].$$

Thus the Borel sets on the real line are the closed and opened sets! Thus, all of the sets you would ever use in practice are in the Borel set σ-algebra.

6.1.2 Coin Toss Example

Take the coin tossing example. Let Ω_∞ be the set of infinite sequence of H's or T's; $p \in (0, 1)$ be the probability of H in any coin toss. Let ω_i be the ith event, which can take the values H and T. We can construct a probability measure $P > 0$ defined as the probability of seeing at least one heads, assuming each toss is independent. We can also construct a series of σ-algebra $\mathcal{F}$ as follows

- $\mathcal{F}_0 = \{\phi, \Omega_\infty\}$, $P(\phi) = 0$, $P(\Omega_\infty) = 1$
- $\mathcal{F}_1 = \{\phi, \Omega_\infty, A_H, A_T\}$ where
  - $A_H = \{w \in \Omega_\infty : w_1 = H\}$
  - $A_T = \{w \in \Omega_\infty : w_1 = T\}$
- $\mathcal{F}_1 = \{\phi, \Omega_\infty, A_H, A_T, A_{HH}, A_{TT}, A_{HT}, A_{TH}, A_{HH}^C, A_{TT}^C, A_{HT}^C, A_{TH}^C, A_{HH} \bigcup A_{TT}, \ldots, A_{HT} \bigcup A_{TH}\}$ where
  - $A_{HH} = \{w \in \Omega_\infty : w_1 = H, w_2 = H\}$
  - $A_{HT} = \{w \in \Omega_\infty : w_1 = T, w_2 = T\}$
  - $\vdots$

We can see that $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \ldots$ and that the cardinality can grow very quickly. That is, $|\mathcal{F}_n| = 2^{2^n}$. We can define

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{i=1}^{\infty} \mathcal{F}_i\right)$$

Notice that

1. Let $A = \{\omega | \omega_n = H, \forall n = 1, 2, ...\}$, then $A = A_H \cap A_{HH} \cap ... \cap A_{HHH...} \Rightarrow A \in \mathcal{F}_\infty$.
2. $P(A) = \lim_{n\to\infty} P(\text{see n H's in a row}) = \lim_{n\to\infty} p^n = 0$.

Let $A = \{\omega | \lim_{n\to\infty} \frac{H_n(\omega)}{n} = 0.5\}$, where $H_n(\omega) =$ the number of H's in the first n tosses. We have a few questions:

- Question 1: Is $A \in \mathcal{F}_\infty$?
- Question 2: If $A \in \mathcal{F}_\infty$, what's $P(A)$?

Question 1: For a given $m \in \mathcal{N}$, define $A_{n,m} = \{\omega : |\frac{H_n(\omega)}{n} - 0.5| \leq \frac{1}{m}\}$, then $A_{n,m} \in \mathcal{F}_n \subset \mathcal{F}_\infty$. If $\omega \in A$, then for all $\forall \epsilon > 0$, there exists an N such that for all $n > N$, $|\frac{H_n(\omega)}{n} - 0.5| < \epsilon$. Let $\epsilon = \frac{1}{m}$. Thus there exists an N such that for every $\omega \in A_{n,m}$ and for all $n > N$,

$$\omega \in \cap_{m=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_{n,m} \Rightarrow A = \cap_{m=1}^{\infty} \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_{n,m}$$

Thus $A \in \mathcal{F}_\infty$.

Question 2: By Strong Law of Large Numbers, $P(A) = 1$.

6.2 Random Variables and Expectation

Definition: A *random variable* is a real valued function $X : \Omega \to \mathcal{R} \cup \{\infty, -\infty\}$ with the property that for Borel subsets $\mathcal{B}$ of $\mathcal{R}, \{X \in \mathcal{B}\} = \{\omega \in \omega : \ X(\omega) \in \mathcal{B}\}$.

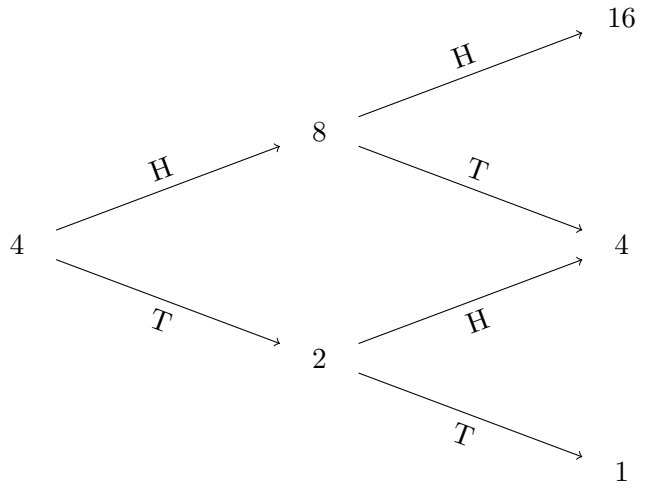
Definition: A *measure* is a nonnegative countably additive set function; that is a function $\mu : \mathcal{F} \to \mathcal{R}$ with

1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$, and
2. if $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets, then $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.

If $\mu(\Omega) = 1$, we call μ a *probability measure*. In particular, $\mu_X(B) = P\{X \in B\}$.

6.2.1 Example: Coin Toss Experiment

Look once again at the infinite coin toss experiment. Take the probability space of the coin toss experiment $(\Omega_\infty, \mathcal{F}_\infty, p)$. Assume that we start with 4 dollars, and every time we a head we double our money, and if we get a tail we half our money. Let S_n be the random variable that denotes our value after n tosses.



Thus

- $S_0(\omega) = 4$, for all ω;
- $S_1(\omega) = 8$ if $\omega_1 = H$; 2 otherwise;
- ...
- $S_{n+1}(\omega) = 2S_n(\omega)$ if $\omega_{n+1} = H$; $\frac{1}{2}S_n(\omega)$ otherwise.

6.2.2 Example: Uniform Random Variable

A random variable X uniformly distributed on $[0,1]$ can be simulated based on the example of infinite independent coin toss with $p = 0.5$. To do so, let $Y_n(\omega)$ be the indicator that the nth coin toss is a heads, that is$Y_n(\omega) = 1$ if $\omega_n = H$ and 0 otherwise. Thus we define the random variable

$$X = \sum_{n=1}^{\infty} \frac{Y_n(\omega)}{2^n}.$$

Notice if we look at the base-2 decimal expansion of X, that would be the sequence of heads and tails where a 1 in the ith digit means the ith toss was a heads. Thus we see that the range of X is $[0,1]$ and every decimal expansion has equal probability of occurring, meaning that X is uniformly distributed on $[0,1]$. Thus we get that the probability of being between a and b is

$$\mu_X(a,b) = b - a, \;\; 0 \leq a \leq b \leq 1.$$

6.2.3 Expectation of a Random Variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space, we define

1. If Ω is finite, then $E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega)$.
2. If Ω is countably infinite, then $E(X) = \sum_{k=1}^{\infty} X(\omega_k)P(\omega_k)$.
3. If Ω is uncountable, then $\int_\Omega X(\omega)dP(\omega)$, or $\int_\Omega X dP$.

Intuition on Lebesgue Integration The integration in case 3 is based on *Lebesgue Integral*. Think of the Lebesgue integral as using a "Reimann sum on the y-axis", meaning you cut up increments of the y axis instead of the x. The reason this is done is because it can do better for functions defined on weird domains. For example, if we let

$$f(x) = \begin{cases} 1, & x \text{ is irrational} \\ 0, & o.w. \end{cases}$$

we know that almost all of the time $x = 1$ (since the rationals are a countable set while the irrationals are an uncountable set (a larger infinity)). This is however not computable using the Reimann integral. But, using the Lebesgue integral we get

$$\int_0^1 df(x) = 1\mu(A)$$

where $\mu(A)$ means the measure (the length) of the set where $f(x) = 1$. Since there are only countably many holes, $\mu(A) = \mu([0,1]) = 1$. Thus

$$\int_0^1 df(x) = 1$$

which matches our intuition. This may sound scary at first, but if you're not comfortable with these integrals you can understand the following theory by just reminding yourself it's simply another way to integrate that works on weirder sets than the Reimannian integral.

Construction of Expectation/Lebesgue Integral:

1. Let X be a characteristic function, i.e. $X(\omega) = 1$ if $\omega \in A; 0$ otherwise for some $A \in \mathcal{F}$, then we define $\int_\Omega X dP = P(A)$.
2. If X is a simple function, i.e. $X(\omega) = \sum_{k=1}^n c_k X_{A_k}(\omega)\ \forall A_k \in \mathcal{F}$, then we define $\int_\Omega X dP = \sum_{k=1}^n c_k P(A_k)$.
3. If X is nonnegative, then we define
$$\int_\Omega X dP = sup\{\int_\Omega Y dP : Y\ is\ simple, Y(\omega) \leq X(\omega),\ \forall \omega \in \Omega\}.$$
4. For any general X, we define $\int_\Omega X dP = \int_\Omega X^+ dP - \int_\Omega X^- dP$, where $X^+ = max\{0, X\}$ and $X^- = max\{-X, 0\}$. X is *integrable* if $\int_\Omega |X| dP < \infty$.

6.2.4 Properties of Expectations:

- $\int_A X dP = \int_\Omega I_A X dP$, where I_A is the indicator function on A.
- $\int_\Omega (X + Y) dP = \int_\Omega X dP + \int_\Omega Y dP$.
- $\int_\Omega cX dP = c \int_\Omega X dP$.
- If $X(\omega) \leq Y(\omega)$ a.s., then $\int_\Omega X dP \leq \int_\Omega Y dP$.
- $\int_{A \cup B} X dP = \int_A X dP + \int_B X dP$ if $A \cap B = \emptyset$.

Note that the definition of almost surely (a.s.) will be explained in soon.

6.2.5 Convergence of Expectation

Let's start with an example. Let $X_n \sim N(0, \frac{1}{n})$. Thus the probability density of X_n is

$$f_n(x) = \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}}$$

Moreover, $E(X_n) = 0\ \forall n$. Also notice that

$$\int_{-\infty}^{\infty} f_n(x) dx = 1.$$

Define

$$f(x) = \lim_{n\to\infty} f_n(x).$$

Notice that this converges point-wise to the function $f(x) = 0$. Thus

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(x)dx \neq \int_{-\infty}^{\infty} f(x)dx.$$

6.2.6 Convergence Theorems

We will take these theorem without proof. They are based off of famous measure theory theorems and are useful in proving properties later in this chapter.

For this theorem, we will need a few definitions. Take two random variables X and Y.

Definition: Define the idea that $X = Y$ almost surely (a.s). if $P\{\omega \in \Omega : X(\omega) = Y(\omega)\} = 1$. Intuitively, this means that X and Y agree on everything except for at most a measure 0 set, but since no event from a measure 0 set will almost surely occur, they are basically always equivalent.

Proposition: Take $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. If $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all ω except on a set of measure 0, then X_n converges a.s. to X.

Proposition: If $0 \leq X_1(\omega) \leq X_2(\omega) \leq \ldots$ for all ω except on a set of measure 0, then $0 \leq X_1 \leq X_2 \leq ...$ a.s.

With these, we can write a few convergence theorems. Note that these results can be directly obtained from well-known measure-theory theorems. These three theorems are also true for a more general class of functions: *measurable functions*.

Theorem: Monotone Convergence Theorem. Take $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables converging almost surely (a.s.) to a random variable X. Assume $0 \leq X_1 \leq X_2 \leq ...$ a.s.. Thus

$$\int_{\Omega} X dP = \lim_{n\to\infty} \int_{\Omega} X_n dP,$$

or equivalently

$$E(X) = \lim_{n\to\infty} E(X_n).$$

Theorem: Dominated Convergence Theorem**.** Take $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables converging almost surely (a.s.) to a random variable X. Assume that there exists a random variable Y s.t. $|X_n| \leq Y$ a.s. for all n, and $E(Y) < \infty$, then $E(X) = \lim_{n\to\infty} E(X_n)$.

Theorem: Fatou's Lemma. Take $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables converging almost surely (a.s.) to a random variable X. Thus $E(X) \leq \liminf_{n\to\infty} E(X_n)$.

6.3 Filtrations, Conditional Expectations, and Martingales

6.3.1 Filtration Definitions

For the following definitions, take the probability space (Ω, f, p).

Definition: Let $\Omega \neq \emptyset$. Let T be a fixed positive number. Assume that for all $t \in [0,T]$, there exists a σ-algebra $f(t)$. Assume that if $s < t$, $f(s) \subset f(t)$. We define the collection of σ-algebras $\mathcal{F}_T = \{f(t)\}_{0 \leq t \leq T}$ as the filtration at time T. It is understood intuitively as the complete set of information about the stochastic process up to time T.

Now we need to bring random variable definitions up to our measure-theoretic ideas.

Definition: Let X be a random variable in (Ω, f, p). The σ-algebra generated by X is the set

$$\sigma(X) = \left\{X^{-1}(A) : A \in f\right\}$$

Definition: Given (Ω, f, p) with a filtration $\mathcal{G}_t$ and a random variable X, we call X $\mathcal{G}_t$-measurable if $\sigma(X) \subseteq \mathcal{G}_t$.

Definition: Let $X(t)$ be a stochastic process of (Ω, f, p) and let $\mathcal{F}_t$ be a filtration defined on $t \in [0,T]$. $X(t)$ is an adapted stochastic process if for each $t \in [0,T]$, $X(t)$ is $\mathcal{F}_t$-measurable.

What these definitions have setup is the idea of "information about a random process". A random variable X is $\mathcal{G}_t$-measurable if, given the information of $\mathcal{G}_t$, X is already known. A stochastic process is adapted if a filtration $\mathcal{F}_s$ gives us all of the information about the stochastic process up to the time s, and thus $X(t)$ is known for all $t \in [0,s]$.

6.3.2 Independence

Definition: Take two events A,$B \in f$. The events A and B are independent if $P(A \cap B) = P(A)P(B)$.

Definition: Let $G, H \subset f$ be two sub-σ-algebras of f. G and H are independent if, for all $A \in G$ and $B \in H$, $P(A \cap B) = P(A)P(B)$.

Definition: Take the random variables X and Y of the probability space (Ω, f, p). X and Y are independent if $\sigma(X)$ and $\sigma(Y)$are independent.

These are straight-forward and obvious to make yourself feel better about all this and make it feel easy.

6.3.3 Conditional Expectation

Definition: Let G be a sub-σ-algebra of f ($G \subset f$). Let X be a random variable that is either non-negative or integrable. The conditional expectation of X given G, $\mathbb{E}[X|G]$, is a random variable that satisfies the following properties:

1. $\mathbb{E}[X|G]$ is G-measurable.
2. Partial Averaging Property: For all $A \in G$,

$$\int_A \mathbb{E}[X|G]dp = \int_A X dp.$$

Let us understand what this definition means. We can interpret $\mathbb{E}[X|G]$ to be the random variable that is the "best guess" for the values of X given the information in G. Since the only information that we have is the information in G, this means that $\mathbb{E}[X|G]$ is G-measurable. Our best guess for the value of X is the expectation of X. Notice that in measure-theoretic probability, the expectation of $\mathbb{E}[X|G]$ for the event A is defined as $\int_A \mathbb{E}[X|G]dp$. Thus the partial averaging property is simply saying that we have adapted the random variable $\mathbb{E}[X|G]$ such that its expectation for every event that has happened in G is the same as the expectation of X, which means that for any thing that has happened, our best guess is simply what X was!

Definition: Take the random variables X and Y. We define $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$.

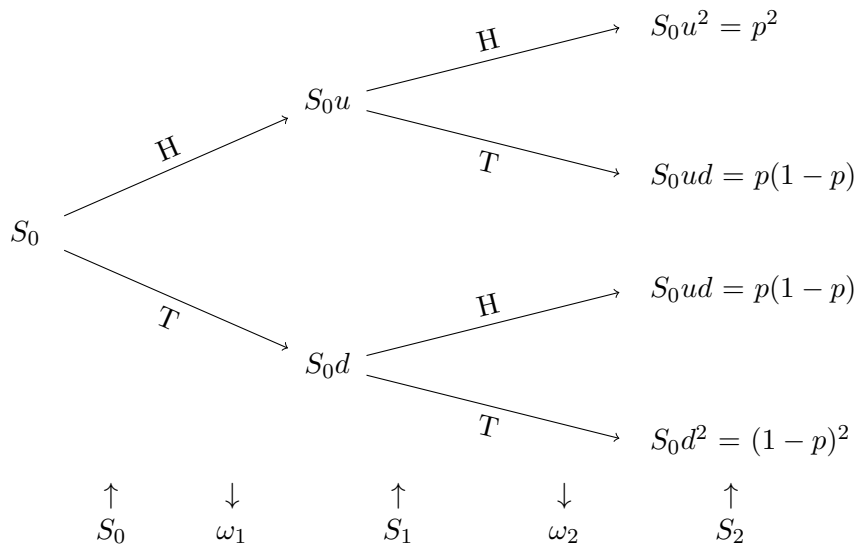
This gives us our link back to the traditional definition of conditional expectations using random variables. Notice that this is how we formally define functions of random variables. Writing the conditional expectation as $\mathbb{E}[Y|X = x_i] = y_i$, we can think of this as a mapping from x_i to y_i. Thus we can measure-theoretically interpret $f(Y) = \mathbb{E}[X|Y]$.

6.3.4 Properties of Conditional Expectation

1. **Theorem:** $\mathbb{E}[X|G]$ exists and is unique except on a set of measure 0. This is a direct result of the Radon-Nikodym theorem.
2. $\mathbb{E}\left[\mathbb{E}[X|G]\right] = \mathbb{E}[X]$. Our best estimate of $\mathbb{E}[X|G]$ is $\mathbb{E}[X]$ if we have no information.
3. Linearity: $\mathbb{E}\left[aX + bY|G\right] = a\mathbb{E}\left[X|G\right] + b\mathbb{E}[Y|G]$.
4. Positivity: If $X \geq 0$ a.s., then $\mathbb{E}[X|G] \geq 0$ a.s.
   (a) This can be generalized: For all $A \subset \mathbb{R}$, if $X \in A$ a.s., then $\mathbb{E}[X|G] \in A$ a.s.
5. Taking out what is known: If X is G-measurable, then $\mathbb{E}[XY|G] = X\mathbb{E}[Y|G]$.Notice that this is because if X is G-measurable, it is known given the information of G and thus can be treated as a constant.
6. Iterated Conditioning: If $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then $\mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}\right] = \mathbb{E}[X|\mathcal{H}]$.
7. Independence of $\mathcal{G}$: If X is independent of $\mathcal{G}$, then $\mathbb{E}[X|\mathcal{G}] = E[X]$. If $\mathcal{G}$ gives no information about X, then the expectation condition on the information of $\mathcal{G}$ is simply the expectation of X.

6.3.5 Example: Infinite Coin-Flipping Experiment

Consider the infinite coin-flipping experiment $(\Omega_\infty, \mathcal{F}, p)$. Recall that this has a probability p of being heads and $1-p$ of being tails. You start with S_0 money, and for each heads you multiply your money by u, and for each tails you multiply your money by d. The experiment can be explained by the following tree:



Let's say we have the information given by the filtration $\mathcal{G}$ and we let X be our earnings from the game after two coin-flips. We explore the properties in the following scenarios.

1. $\mathcal{G} = \mathcal{F}_0$**.** Recall that $\mathcal{F}_0 = \{\Omega_\infty, \emptyset\}$ and thus it literally contains no information. Thus we have that X is independent of the information in $\mathcal{F}_0$. $\mathbb{E}[X|\mathcal{F}_0] = \mathbb{E}[X]$. We can calculate this using the traditional measures:

$$\mathbb{E}[X] = \sum_i x_i \Pr(X = x_i) = S_0 \left(p^2u^2 + 2p(1-p)ud + (1-p)^2d^2\right).$$

2. $\mathcal{G} = \mathcal{F}_1$**.** Recall that $\mathcal{F}_1$ is the information contained after the first event. Thus we know the result of the first event, ω_1. Thus we can calculate what we would expect X to be given what we know about the first event. Thus

$$E[X|\mathcal{F}_1] = \begin{cases} S_0\left(pu^2 + (1-p)ud\right), & \omega_1 = H \\ S_0\left(pud + (1-p)d^2\right), & \omega_1 = T \end{cases}.$$

3. $\mathcal{G} = \mathcal{F}_2$**.** Notice that if we have the information of $\mathcal{F}_2$, then X is determined (X is $\mathcal{G}$-measurable). Thus

$$E[X|\mathcal{F}_2] = \begin{cases} S_0u^2, & \omega_1\omega_2 = HH \\ S_0ud, & \omega_1\omega_2 = HT \text{ or } TH \\ S_0d^2 & \omega_1\omega_2 = TT \end{cases}.$$

6.4 Martingales

Definition: Take the probability space $(\Omega, \mathcal{F}, p)$ with the filtration $\{\mathcal{F}_t\}_{0\leq t\leq T}$, and let $\{M_t\}_{0\leq t\leq T}$ is an adaptive stochastic process w.r.t. $\{\mathcal{F}_t\}$. M_t is a martingale if $\mathbb{E}[M_t|\mathcal{F}_s] = M_s$ for all $0 \leq s \leq t$. If $\mathbb{E}[M_t|\mathcal{F}_s] \geq M_s$ then we call M_t a sub-martingale, while if $\mathbb{E}[M_t|\mathcal{F}_s] \leq M_s$ then we call M_t a super-martingale.

These definitions can be interpreted as follows. If, knowing the value at time s, our best guess for M_t is that it is the value M_s, then M_t is a martingale. If we assume that it will grow in expectation in time, it's a sub-martingale while if we assume that its value will shrink in time, it is a super-martingale.

6.4.1 Example: Infinite Coin-Flipping Experiment Martingale Properties

Looking back at the infinite coin-toss experiment, if S_n is the money we have after event n, then

$$S_{n+1} = \begin{cases} uS_n, & \omega_{k+1} = H \\ dS_n, & \omega_{k+1} = T \end{cases}.$$

Thus we have that

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = puS_n + (1-p)dS_n = S_n\left(pu + (1-p)d\right) = \alpha S_n.$$

Recalling that S_n is a martingale if we would expect that our value in the future is the same as now, then

1. If $\alpha = 1$, then S_n is a martingale.
2. If $\alpha \geq 1$, then S_n is a sub-martingale.
3. If $\alpha \leq 1$, then S_n is a super-martingale.

6.4.2 Example: Brownian Motion

Theorem: A Brownian motion B_t is a martingale.

Proof:

$$\mathbb{E}[B_t|\mathcal{F}_s] = \mathbb{E}[B_t - B_s + B_s|\mathcal{F}_s] = \mathbb{E}[B_t - B_s|\mathcal{F}_s] + \mathbb{E}[B_s|\mathcal{F}_s] = 0 + B_s = B_s$$

since on average $B_t - B_s = 0$.

6.5 Martingale SDEs

Theorem: Take the SDE

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t.$$

If $a(X_t, t) = 0$, then X_t is a martingale.

Proof: If we take the expectation of the SDE, then we see that

$$\frac{dE[X_t]}{dt} = a(X_t, t)$$

and thus if $a(X_t, t) = 0$, the expectation does not change and thus X_t is a martingale.

6.5.1 Example: Geometric Brownian Motion

Take the stochastic process

$$Z_t = f(W_t, t) = e^{-\theta W_t - \frac{1}{2}\theta^2 t}.$$

Notice using Ito's Rules that

$$\begin{aligned} dZ_t &= \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial W_t}dW_t + \frac{1}{2}\frac{\partial^2 f}{\partial W_t^2}(dW_t)^2 \\ &= -\frac{1}{2}\theta^2 Z_t dt - \theta Z_t dW_t + \frac{1}{2}\theta^2 Z_t dt \\ &= -\theta Z_t dW_t \end{aligned}$$

since $(dW_t)^2 = dt$. Thus since there is no deterministic part, Z_t is a martingale. Thus since $Z_0 = 1$, we get that

$$\mathbb{E}[Z_t|\mathcal{F}_s] = Z_s,$$

and

$$\mathbb{E}[Z_t|\mathcal{F}_0] = 1.$$

6.6 Application of Martingale Theory: First-Passage Time Theory

Take the SDE

$$dX_t = f(X_t, t)dt + \sum_{i=1}^{n} g_i(X_t, t)dW_i.$$

Here we will consider first-passage time problems. These are problems where we start in a set and want to know at what time we he the boundary. Fix $x > 0$. Let $\tau = \inf\{t \geq 0 : \ X_t = x\}$ which is simply the first time that X_t hits the point x.

6.6.1 Kolmogorov Solution to First-Passage Time

Notice that we can solve for the probability definition using the Forward Kolmogorov Equation

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x)\rho(x,t)] + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2}{\partial x^2}\left[g_i^2(x)\rho(x,t)\right].$$

Given the problem, we will also have some initial condition $\rho(x,0) = \rho_0(x)$. If we are looking for first passage to some point x_0, then we put an absorbing condition there $\rho(x_0,t) = 0$. Thus the probability that you have not hit x_0 by the time t is given by the total probability that has not been absorbed by the time t, and thus

$$\Pr(t \geq \tau) = \int_{-\infty}^{\infty} \rho(x,t)dx$$

to get the cumulative probability distribution

$$\Pr(t \leq \tau) = 1 - \int_{-\infty}^{\infty} \rho(x,t)dx.$$

Thus the probability density function for τ is simply

$$f_\tau(t) = \frac{\partial}{\partial t}\left[\int_{-\infty}^{\infty} \rho(x,t)dx\right].$$

This method will always work, though many times the PDE will not have an analytical solution. However, this can always be solved using computational PDE solvers.

6.6.2 Stopping Processes

If M_t is a martingale, then the stopped martingale is defined as

$$M(t \wedge \tau) = \begin{cases} M(t), & t \leq \tau \\ M(\tau), & t \geq \tau \end{cases}$$

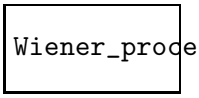
where $M(t \wedge \tau) = \min(t, \tau)$. This simply the process that we stop once it hits x.

Theorem: If M_t is a martingale, then $M_{t\wedge\tau}$ is a martingale.

The proof is straight-forward. Since the stopped martingale does not change after τ, then the expectation definitely will not change after τ. Since it is a martingale before τ, the expectation does not change before τ. Thus $M_{t\wedge\tau}$ is a martingale.

6.6.3 Reflection Principle

Another useful property is known as the Reflection Principle. Basically, if we look at any Brownian motion at the time T, there is just as much of a probability of it going up and there is of it going down. Thus the trajectory that is reflected after a time τ is just as probable as the non-reflected path.



(Picture taken from Oksendal *Stochastic Differential Equations: An Introduction with Applications*) We can formally write the Reflection Principle as

$$P(\tau \leq t,\ B_t < x) = P(B_t \geq x).$$

This leads to the relation

$$\begin{aligned} P(\tau \leq t) &= P(\tau \leq t, B_t < x) + P(\tau \leq t, B_t \geq x) \\ &= 2P(B_t \geq x) \\ &= 2\int_x^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{u^2}{2t}}\, du. \end{aligned}$$

6.7 Levy Theorem

Theorem: Levy Theorem. Take $(\Omega, \mathcal{F}, p)$. If $W(t)$ is a martingale under p and $dW(t) \times dW(t) = dt$, then $W(t)$ is Brownian Motion.

We state this without proving it since this takes a lot more detailed treatment of Brownian motion.

6.8 Markov Processes and the Backward Kolmogorov Equation

6.8.1 Markov Processes

Take $0 \leq t_0 < t$ and let $h(y)$ be a function. Define

$$\mathbb{E}^{t_0,x}\left[h(x(t)\right] = v(t_0, x)$$

as the expectation of $h(x(t))$ given that $x(t_0) = x$.

Definition: Take the random variable x. If

$$\mathbb{E}^{0,\zeta}\left[h(x(t)|\mathcal{F}_{t_0}\right] = \mathbb{E}^{t_0,x}\left[h(x(t)\right] = v(t_0, x),$$

then we say x satisfies the Markov property and is thus a Markov process. Another way of stating this is that

$$\mathbb{E}\left[f(X_t)|\mathcal{F}_s\right] = \mathbb{E}\left[f(X_t)|\sigma(X_s)\right],$$

or

$$P(x_t \in A|\mathcal{F}_s) = P(x_t \in A|x_s).$$

Intuitively, this property means that the total information for predicting the future of x_t is completely encapsulated by the current state x_t.

6.8.2 Martingales by Markov Processes

Take the SDE $dX_t = a(X_t)dt + b(X_t)dW_t$.

Theorem: X_t is a Markov Process.

Theorem: Given $\mathbb{E}^{t_0,x}[h(x(t)] = v(t_0, x)$, $\{v(x,t)\}_{0\leq t\leq T}$ is a martingale w.r.t. $\{\mathcal{F}_t\}$.

Proof: Look at $\mathbb{E}[v(x,t)|\mathcal{F}_s]$ for $0 < s \leq t$. Notice

$$\mathbb{E}[v(x(t),t)|\mathcal{F}_s] = \mathbb{E}\left[\mathbb{E}^{t_0,x}[h(x(t))|\mathcal{F}_t]\,|\mathcal{F}_s\right] = \mathbb{E}[h(x(t))|\mathcal{F}_s] = v(x(s),s)$$

since $\mathcal{F}_s \subset \mathcal{F}_t$. Thus v is a martingale.

6.8.3 Transition Densities and the Backward Kolmogorov

We can use this in the following proof. Notice that

$$\begin{aligned} dv &= \frac{\partial v}{\partial t}dt + \frac{\partial v}{\partial x}dx + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}(dx)^2 \\ &= \frac{\partial v}{\partial t}dt + \frac{\partial v}{\partial x}(a(X_t)dt + b(X_t)dW_t) + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}b^2(x(t),t)dt \\ &= \left(\frac{\partial v}{\partial t} + a(X_t)\frac{\partial v}{\partial x} + \frac{1}{2}b^2(x(t),t)\frac{\partial^2 v}{\partial x^2}\right)dt + b(X_t)\frac{\partial v}{\partial x}dW_t. \end{aligned}$$

Since v is a martingale, the deterministic part must be zero. Thus we get the equation

$$\frac{\partial v}{\partial t} + a(X_t)\frac{\partial v}{\partial x} + \frac{1}{2}b^2(x(t),t)\frac{\partial^2 v}{\partial x^2} = 0.$$

We can solve this using the initial condition $v(x,T) = h(x)$ to solve for $v(x(t),t)$.

Define the transition density function as

$$\rho(t_0,t,x,y)dy = P\{x(t) \in (y, y+dy)|x(t_0) = x\}$$

which is the probability of transitioning from x to y between t_0 and t. Notice that the probability of transitioning from $h(x)$ to $h(y)$ from the time t to T is

$$v(x,t) = \mathbb{E}^{t,x}[h(x(T))] = \int_{-\infty}^{\infty} \rho(t,T,x,y)h(y)dy.$$

Now suppose we know $h(z) = \delta(x - y)$, meaning that we know that the trajectory ends at y. Thus

$$\begin{aligned} v(x,t) &= \int_{-\infty}^{\infty} \rho(t,T,x,y)\delta(x-y)dy \\ &= \rho(t,T,x,y). \end{aligned}$$

Thus we plug this into the PDE for v to get

$$\frac{\partial \rho}{\partial t} + a(x(t),t)\frac{\partial \rho}{\partial x} + \frac{1}{2}b^2(x(t),t)\frac{\partial^2 \rho}{\partial x^2} = 0.$$

with the terminal condition $\rho(T,x) = \delta(x-y)$ where y is the place the trajectory ends. Thus this equation, the Kolmogorov Backward Equation, tells us that if we know the trajectory ends at y at a time T, what is the probability that it was x and a time $t < T$. Notice that this is not the same as the Kolmogorov Forward Equation. This is because diffusion forward does not equal diffusion backwards. For example, say you place dye in water. It will diffuse to spread out uniformly around the water. However, if we were to play that process in reverse it would look distinctly different.

6.9 Change of Measure

The purpose of the change of measure is that is can greatly simplify equations by taking out deterministic parts and making random variables into martingales. This will be crucial when understanding the solution to the Black-Scholes equation.

6.9.1 Definition of Change of Measure

Take the probability space $(\Omega, \mathcal{F}, p)$. Let z be a non-negative random variable s.t. $\mathbb{E}[z] = 1$. Define the new measure $\tilde{p}$ as

$$\tilde{p}(A) = \int_A z dp.$$

Thus for any random variable x in the probability space $(\Omega, \mathcal{F}, \tilde{p})$ we get that

$$\tilde{\mathbb{E}}[x] = \mathbb{E}[xz]$$

and if $z > 0$ a.s. we get

$$\mathbb{E}[x] = \tilde{\mathbb{E}}\left[\frac{x}{z}\right]$$

where $\tilde{\mathbb{E}}$ is the expectation using the measure $\tilde{p}$.

Definition: p and $\tilde{p}$ are equivalent if they agree on which sets have probability zero.

Corollary: They agree on which events will not occur a.s.

Theorem: Radon-Nikodym Theorem. Let p and $\tilde{p}$ be equivalent measures on $(\Omega, \mathcal{F})$. There exists a random variable $z > 0$ a.s. such that $\hat{p}(A) = \int_A z dp$ where $z = \frac{d\hat{p}}{dp}$. We call $\frac{d\hat{p}}{dp}$ the Radon-Nikodym derivative.

6.9.2 Simple Change of Measure Example

Take $(\Omega, \mathcal{F}, p)$ and let X be a standard normal random variable under p. Define $y = x + \theta$. Notice that under p, $y \sim N(\theta, 1)$. However, if we define

$$z(\omega) = e^{-\theta X(\omega) - \frac{\theta^2}{2}}$$

and define

$$\tilde{p}(A) = \int_A z dp,$$

in this reweighing of the probability space y is standard normal. This is because by definition the pdf of y using the measure $\tilde{p}$ is

$$\begin{aligned}
\widetilde{f}(y) &= z(\theta) N(x + \theta, 1) \\
&= e^{\theta x - \theta^2/2} e^{-(x+\theta)^2/2} \frac{1}{\sqrt{2\pi}} \\
&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \sim N(0, 1)
\end{aligned}$$

6.9.3 Radon-Nikodym Derivative Process

Take any random variable X. Using this Radon-Nikodym derivative and a filtration, we can use it to define a unique process with no deterministic part Z, and use this random variable to define a measure $\tilde{p}$ to transform one stochastic process to another.

Definition: Take(Ω, F, P) with F_t as a filtration of this space up to t with $0 \leq t \leq T$ where T is some fixed finish time. Let $\zeta = \frac{d\widetilde{P}}{dP}$ and satisfy the conditions required to be a Radon-Nikodym derivative. We can define a Radon-Nikodym derivative process (RNDP) Z_t as

$$Z_t = \mathbb{E}[\zeta | \mathcal{F}_t].$$

Notice that

1. $Z(t)$ is a martingale.
2. If we let y be an $\mathcal{F}_t$-meausrable random variable, then

$$\tilde{\mathbb{E}}[y] = \mathbb{E}[yZ_t] = \mathbb{E}\left[y\mathbb{E}[\zeta | F_t]\right] = \mathbb{E}\left[\mathbb{E}[y\zeta | F_t]\right] = \mathbb{E}[y\zeta],$$

and

$$\tilde{\mathbb{E}}[y | F_s] = \frac{\mathbb{E}[yZ_t | F_s]}{Z_s},$$

for $0 \leq s \leq t \leq T$.

6.9.4 Girsanov Theorem

Using the change of measure determined by the RNDP, we can always construct a Brownian motion from an adapted stochastic process. This is known as Girsanov Theorem.

Theorem: Girsanov Theorem. Let W_t be a Brownian motion in (Ω, F, P) and F_t be a filtration with $0 \leq t \leq T$. Let $\theta(t)$ be an adapted stochastic process on F_t. Define

$$\begin{aligned} Z(t) &= e^{-\int_0^t \theta(u)dW(u) - \frac{1}{2}\int_0^t \theta^2(u)du} \\ \tilde{W}_t &= W_t + \int_0^t \theta(u)du \\ d\tilde{W}_t &= dW_t + \theta(t)dt \end{aligned}$$

Define the probability measure $\tilde{p}$ using Z as

$$\tilde{p}(A) = \int_A Z dp.$$

Then the following statements hold:

1. $\mathbb{E}[Z] = 1$
2. $\tilde{W}_t$ is a Brownian motion under $\tilde{p}$.

Proof: First we show that Z_t is a martingale. Notice

$$dlnZ_t = -\theta(t)dW_t - \frac{1}{2}\theta^2(t)dt$$

and thus if we let $\psi(Z) = e^Z$, then we use Ito's Rules to get

$$\begin{aligned} dZ_t &= \psi(\ln Z_t) \\ &= Z_t d(\ln Z_t) + \frac{1}{2} Z_t \left(d(\ln Z_t)\right)^2 \\ &= Z_t[-\theta(t)dW - \frac{1}{2}\theta^2(t)dt + \frac{1}{2}\theta^2(t)dt] \\ &= -Z_t\theta(t)dW \end{aligned}$$

and thus since the deterministic changes are zero, Z_t is a martingale. Notice that since $Z(0) = 1$, $\mathbb{E}[Z_t] = Z_0 = 1$. Thus we have proven the first property.

In order to prove the second property, we employ the Levy Theorem from 6.7. By the theorem we have that if $\tilde{W}_t$ is a martingale under $\tilde{p}$ and $d\tilde{W}_t \times d\tilde{W}_t = dt$, then W_t is Brownian Motion under $\tilde{p}$. Notice that

$$d\tilde{W}_t \times d\tilde{W}_t = (dW_t + \theta(t)dt)^2 = dW_t^2 = dt.$$

Thus in order to show $\tilde{W}_t$ is a Brownian motion under $\tilde{p}$, we simply need to show that $\tilde{W}_t$ is a martingale. Since Z_t is an RNDP, we use property 2 of RNDPs to see

$$\tilde{\mathbb{E}}[\tilde{W}_t|F_s] = \frac{\mathbb{E}\left[\tilde{W}_t Z_t|F_s\right]}{Z_s}.$$

Thus we use Ito's Rules to get

$$\begin{aligned} d\left(\tilde{W}_t Z_t\right) &= \tilde{W}_t dZ_t + Z_t d\tilde{W}_t + d\tilde{W}_t dZ \\ &= -\tilde{W}_t Z_t \theta(t) dW_t + Z_t dW_t + Z_t \theta(t) dt \\ &= Z_t\left(Z_t - \tilde{W}_t\theta(t)\right) dW_t \end{aligned}$$

and thus, since the deterministic changes are zero, $\tilde{W}_t Z_t$ is a martingale under p. Thus we use the definition of a martingale to get

$$\tilde{\mathbb{E}}[\tilde{W}_t|F_s] = \frac{\tilde{W}_s Z_s}{Z_s} = \tilde{W}_s$$

and thus $\tilde{W}_s$ is a martingale under $\tilde{p}$. Therefore by Levy Theorem we get that $\tilde{W}_t$ is a Brownian motion under $\tilde{p}$, completing the proof.

7 Applications of SDEs

Using the theory we have developed, we will now look into some applications of SDEs.

7.1 European Call Options

Let $S(t)$ be the value of the stocks at the time t. Assume that it follows Geometric Brownian Motion:

$$dS(t) = \mu S(t)dt + \sigma S(t)dW.$$

The European call option is the right to buy the stock at a set price k at a future time T. Thus the option only has value if $S(T) > k$, whereas if $S(T) \leq k$ the call option has no value. Thus let

$$g(S(T)) = \begin{cases} S(T) - k, & S(T) > k \\ 0, & S(T) \leq k \end{cases} = [S(T) - k]^+$$

be the value of the option at the time T. Let $v(t, S(t))$ be the value of the European call option at the time t. What we wish to do is find out how to evaluate v.

7.1.1 Solution Technique: Self-Financing Portfolio

Let $X(t)$ be the value of our portfolio. Let we assume there are only two things: this stock and the money market. By finding out the optimal amount of money we should be investing into the stock we can uncover its intrinsic value and thus determine v. Note that if we put our money in the money market, the it occurs interest at a rate r. Thus we can write the value of our portfolio as

$$X(t) = \underbrace{\Delta(t)S(t)}_{\text{Total amount invested in the stock}} + \underbrace{X(t) - \Delta(t)S(t)}_{\text{Money market}}$$

and thus by Ito's Rules

$$\begin{aligned} dX(t) &= \Delta(t)dS(t) + (X(t) - \Delta(t)S(t))rdt \\ &= (rX(t) + (\mu - r)\Delta(t)S(t))\,dt + \sigma\Delta(t)S(t)dW. \end{aligned}$$

Now we expand using Ito's Rules to get

$$\begin{aligned} dv(t, S(t)) &= \frac{\partial v}{\partial t}dt + \frac{\partial v}{\partial x}dS(t) + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}(dS)^2 \\ &= \frac{\partial v}{\partial t}dt + \mu S(t)\frac{\partial v}{\partial x}dt + \sigma\frac{\partial v}{\partial x}S(t)dW + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}\sigma^2 S^2(t)dt \\ &= \left(\frac{\partial v}{\partial t} + \mu S(t)\frac{\partial v}{\partial x} + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}\sigma^2 S^2(t)\right)dt + \sigma\frac{\partial v}{\partial x}S(t)dW \end{aligned}$$

We now assume that there is no arbitrage. This can be rooted in what is known as the "Efficient Market Hypothesis" which states that a free-market running with complete information and rational individuals will operate with no "arbitrage" where arbitrage is the "ability to beat the system". For example, if one store is selling a certain candy bar for \$1 and another store is buying it for \$2, there is an arbitrage here of \$1 and you can make money! But, if these people had complete information and were rational, we assume that they will have worked this out and no opportunity like this will be available. In stock market terms, this means that the price of a good equates with the value of the good. This means that we can assume that the value of the option at the time t, $v(X, t)$, will equate with its market price. Notice that, since the only way we could have made money with our portfolio is that the value of the invested stock has increased in order to make our option more valuable, the value of the option at time t is equal to the value of our portfolio. Since price equates with value we get the condition

$$v(X, t) = X(t)$$

and thus

$$dv = dX.$$

Thus we solve for v by equating the coefficients between the dX and the dv equations. Notice that the noise term gives us that

$$\begin{aligned}\sigma\frac{\partial v}{\partial x}S(t) &= \sigma\Delta(t)S(t)\\ \frac{\partial v}{\partial x} &= \Delta(t)\end{aligned}$$

where we interpret the differential value of v to be associated with the stock price at t

$$\frac{\partial v}{\partial x} = \frac{\partial v}{\partial x}\bigg|_{S(t)}.$$

This is known as the hedging. The matching the dt coefficients we receive the equation

$$rX(t) + (\mu - r)\Delta(t)S(t) = \frac{\partial v}{\partial t} + \mu S(t)\frac{\partial v}{\partial x} + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}\sigma^2 S^2$$

where we replace $X(t) = v(X,t)$ to get

$$\begin{aligned}rv + (\mu - r)\Delta(t)S(t) &= \frac{\partial v}{\partial t} + \mu S(t)\frac{\partial v}{\partial x} + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}\sigma^2 S^2\\ rv - r\frac{\partial v}{\partial x}x &= \frac{\partial v}{\partial t} + \frac{1}{2}\sigma^2 x^2\frac{\partial^2 v}{\partial x^2}\end{aligned}$$

as a PDE for solving for v. Since we have no arbitrage, the value of the option, v, equals the price of the option, $g(T, S(T)$, at the time T. This gives us the system

$$\begin{cases}\dfrac{\partial v}{\partial t} + rx\dfrac{\partial v}{\partial x} + \dfrac{1}{2}\sigma^2 x^2\dfrac{\partial^2 v}{\partial x^2} = rv & \text{linear parabolic PDE}\\ v(T,x) = [x-k]^+\end{cases}$$

whose solution gives us the evaluation of the option at the time t. This is known as the Black-Scholes-Merdin Equation. This equation can be solved using a change of variables, though we will use another method.

7.1.2 Solution Technique: Conditional Expectation

Note that in the Black-Scholes equation, v is independent of μ. This is because the current evaluation of the stock will have already incorporated its long term expected returns. Thus we notice

$$v(t, S(t)) \neq E[[S(t) - k]^+ | \mathcal{F}_t].$$

Instead we notice that, since v is independent of r, we can let μ take any value to arrive at the same conclusion for v. Thus let $\mu = r$. Thus we would get

$$\begin{aligned}v(t, S(t)) &= e^{-r(T-t)}E[[S(t) - k]^+ | \mathcal{F}_t]\\ v(t, S(t)) &= e^{-r(T-t)}E[v(T, S(T))|\mathcal{F}_t]\\ e^{-rt}v(t, S(t)) &= E[e^{-rT}(v(T, S(T))|\mathcal{F}_t]\end{aligned}$$

and thus by definition v is a martingale. Thus we use Ito's Rules on $e^{-rt}v$ to get

$$d(e^{-rt}v(t,S(t))) = e^{-rt}\left(-rv + \frac{\partial v}{\partial t}]dt + e^{-rt}\frac{\partial v}{\partial x}dS + e^{-rt}\frac{1}{2}\frac{\partial^2 v}{\partial x^2}(dS)^2\right)$$

and use the martingale property to equate the deterministic changes with zero to once again uncover the Black-Scholes equation.

Thus, since we know that it's equivalent to say $\mu = r$ and evaluate the equation

$$v(t,S(t)) = e^{-r(T-t)}E[v(T,S(T))|\mathcal{F}_t],$$

we note that

$$v(t,S(t)) = e^{-r(T-t)}E^{t,S(t)}[[S(t)-k]^+|\mathcal{F}_t].$$

Recall from the SDE that

$$dS(t) = \mu S(t)dt + \sigma S(t)dW$$

which is Geometric Brownian Motion whose solution is

$$S(T) = S(t)e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma(W(T)-W(t)}.$$

Thus, noting that $W(T) - W(t) = \Delta W \sim N(0, T-t)$, we use the definition of expected value to get

$$\begin{aligned} v(t,S(t)) &= \int_{-\infty}^{\infty}[S(t)e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma u} - k]^+ \frac{1}{\sqrt{2\pi(T-t)}}e^{-\frac{u^2}{2(T-t)}}du \\ &= \frac{1}{\sqrt{2\pi(T-t)}}\int_{\alpha}^{\infty}\left(S(t)e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma u} - k\right)e^{-\frac{u^2}{2(T-t)}}du \end{aligned}$$

where

$$\alpha = \frac{\ln\left(\frac{k}{S(t)}\right) - \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma}.$$

We can solve this to get

$$v(t,x) = xN(d_+) - ke^{-r\tau}N(d_-)$$

where $\tau = T - t$ and

$$d_\pm(\tau,x) = \frac{1}{\sigma\sqrt{\tau}}\left[\ln\frac{x}{k} + \left(r \pm \frac{\sigma^2}{2}\right)\tau\right]$$

and

$$N(y) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{u^2}{2}}du = -erfc(y).$$

7.1.3 Justification of $\mu = r$ via Girsanov Theorem

Take the generalized stock price SDE

$$dS(t) = \mu(t)S(t)dt + \sigma(t)S(t)dW$$

and let

$$D(t) = e^{\int_0^t r(s)ds}.$$

Thus we use Ito's Rules to get

$$d\left(D(t)S(t)\right) = \sigma(t)D(t)S(t)\left[dW_t + \theta(t)dt\right]$$

where

$$\theta(t) = \frac{\mu(t) - r(t)}{\sigma(t)}.$$

Thus we define $d\tilde{W}_t = dW_t + \theta(t)dt$ to get

$$d\left(D(t)S(t)\right) = \sigma(t)D(t)d\tilde{W}_t.$$

Therefore if we define

$$Z_t = e^{-\int_0^t \theta(u)dW(u) - \frac{1}{2}\int_0^t \theta^2(u)du},$$

and

$$\tilde{p}(A) = \int_A Z_t dp,$$

then $d\tilde{W}_t$ is a Brownian motion under $\tilde{p}$. Therefore, since there is no deterministic change, $D(t)S(t)$ is a martingale under $\tilde{p}$. We note that $D(t)X(t)$ is also a martingale under $\tilde{p}$. We call $\tilde{p}$ the Risk-Neutral World. Note that in the Risk-Neutral World that, we can set the price of the option, discounted by $D(t)$, as the expectation conditioned on the totality of information that we have. Thus for $V(S(T), T)$ as the payoff of a derivative for a security for stock $S(T)$ at time T, we get

$$D(t)V(S(t), t) = \tilde{E}\left[D(t)V(S(t), t)|\mathcal{F}_t\right].$$

This is an equivalent expression as the conditional expectation from before, saying that we can let $\mu = r$ because this is simply a change of measure into the Risk-Neutral World.

7.2 Population Genetics

We consider a new case study involving population genetics under the Wright-Fisher model. The Wright-Fisher model is a discrete stochastic model of genetic variance which was first introduced to formally develop the ideas of genetic drift. We will start by considering the classical Wright-Fisher model and show how it can be approximated using SDEs. This will give us an intuitive way of generalizing and simulating Wright-Fisher models in a way that is useful for applications. This type of analysis is taken from Richard Durrett's *Probability Models for DNA*.

7.2.1 Definitions from Biology

First we will introduce some definitions from biology. A gene locus is simply a location on a chromosome or a portion of a chromosome. It can represent a gene, a SNP (single base pair polymorphism) or simply a location. An allele is one of a number of alternative forms of the locus. A dominant allele is usually capitalized. For human and other diploid animals, there are typically two alleles of paternal and maternal origin. A genotype is the genetic makeup of an individual. In this case it will denote the types of alleles an individual individual carries. For example, if the individual has one dominant allele and one recessive allele, its genotype is Aa. Having the same pair of alleles is called homozygous while having different alleles is called heterozygous. A mutation is a random genetic change. Here we refer to it as the random spontaneous change of one allele to another. Fitness is the measure of survivability and ability to reproduce of an individual possessing a certain genotype (this is wrong for many philosophical / scientific reasons, but we can take this as a working definition since this is a major topic worth its own book) . Neutral evolution is the case where all genotypes of interest have the same fitness. This implies that there is no selection from such genetic variations and all change is inherently random.

7.2.2 Introduction to Genetic Drift and the Wright-Fisher Model

Here we introduce the Wright-Fisher model. Consider a model with the following assumptions:

1. There is a population with finite size N.
2. The size of the population is constant throughout evolution.
3. There are discrete, non-overlapping generations.
4. Mating is completely random and is determined by random sampling with replacements. This means any individual can randomly give rise to 0, 1, ... many offsprings.

From this, the time evolution of the Wright-Fisher model is defined as:

1. At generation 0, there are $2N$ alleles some A, some a.
2. At generation 1, each A or a allele from generation 0 may result in one, more or zero copies of that allele.

The following theorem holds:

Theorem: Genetic Drift. Because the population size is finite and fixed, due to uneven passing of alleles by chance, eventually there will be only one allele, A or a, left. This is known as Genetic Drift.

Notice the implication of the theorem. This means that, even in the absence of selection, you can have alleles fix in the population and thus have a form of evolution occur! We can intuitively understand this as simply due to first-passage problems.

7.2.3 Formalization of the Wright-Fisher Model

Let us characterize the Wright-Fisher model in probabilistic terms. Let X_n be the total number of A alleles at generation n. Considering that there are a total of $2N$ alleles in any given generation, we can define $P_n = \frac{X_n}{2N}$ to be the percentage of alleles that are A. Let X_0 be some known initial condition. Because there are $2N$ independent alleles that could be passed and the probability of generating an A allele at generating n is P_{n-1} (because of sampling with replacement), we can derive the distribution for X_n using indicator functions. Order the alleles in the population. Let $X_{n,i}$ be 1 if the ith allele is an A and 0 if the ith allele is a 0. Since we are sampling with replacement, the choice of allele i is independent of the choice of allele j for any $i \neq j$. Thus each $X_{n,i} \sim Bern(P_{n-1})$, that is each is a Bernoulli random variable with probability of heads equal to P_{n-1}. Therefore, since $X_n = \sum_i X_{n,i}$ and each $X_{n,i}$ are independent, X_n is modeled by the result of $2N$ coin-flips each with a probability P_{n-1} of heads. Thus

$$X_i \sim Bin(2N, P_{i-1}),$$

which makes

$$\Pr(X_k = k, 0 \leq k \leq 2N) = \binom{2N}{k} P_{n-1}^k \left(1 - P_{n-1}\right)^{2N-k}.$$

Thus we can show that

$$\mathbb{E}\left[X_n | X_{n-1}\right] = 2NP_{n-1} = X_{n-1}$$

which implies X_n is a martingale. Also

$$\mathbb{V}\left[X_n\right] = 2NP_{n-1}(1 - P_{n-1}).$$

Notice that we can think of this process as some form of a random walk for X_n with probability of going right as P_{n-1}. One can rigorous show then that for the 1-dimension random walk of this sort

$$\lim_{n\to\infty} P_n = 0 \text{ or } 1$$

which means that the process will fix to one of the endpoints in a finite time. Since X_n is a martingale, we note that

$$\lim_{n\to\infty} \mathbb{E}\left[X_n\right] = X_0,$$

which means

$$\Pr(X_\infty = 2N) * (2N) + \Pr(X_\infty = 0) * 0 = X_0$$

and thus by dividing by $2N$ we get

$$\Pr(X_\infty = 2N) = P_0.$$

7.2.4 The Diffusion Generator

If we assume N is large, then it is reasonable to assume that the gene frequencies P_n will change in a manner that is almost continuous and thus can be approximated by an SDE. Consider an SDE of the form

$$dX_t = a(x)dt + \sqrt{b(x)}dW_t$$

with the initial condition $X_0 = x$. Applying Ito's rule gives us

$$\begin{aligned} df(X_t) &= \frac{df}{dX}dX_t + \frac{1}{2}\frac{d^2f}{dX^2}(dX_t)^2, \\ &= \frac{df}{dX}a(x)dt + \frac{df}{dX}\sqrt{b(x)}dW + \frac{1}{2}\frac{d^2f}{dX^2}b(x)dt. \end{aligned}$$

Writing $f_X = \frac{\partial f}{\partial X}$ and $f_{XX} = \frac{\partial^2 f}{\partial X^2}$, we get that

$$\frac{\mathbb{E}[f(X_t)]}{dt} = f_X a(x) + \frac{1}{2}f_{XX}b(x).$$

Thus we define $\mathcal{L}$ to be the operator

$$\mathcal{L}f = \frac{\mathbb{E}[f(X_t)]}{dt} = f_X a(x) + \frac{1}{2}f_{XX}b(x).$$

$\mathcal{L}$ is known as the diffusion generator. If we let $f(x) = x$ then

$$\mathcal{L}f = a(x) = \frac{d\mathbb{E}[f(x)]}{dt}$$

where $a(x)$ defines the infinitesimal mean changes. If we define $f(y) = (y-x)^2$ for a fixed x, then

$$b(x) = \frac{d}{dt}\mathbb{E}\left[(X_t - x))^2\right]$$

makes $b(x)$ define the infinitesimal variance. Therefore, we can generally relate discrete stochastic processes to continuous stochastic processes by defining the SDE with the proper a and b such that the mean and variance match up. This can formally be shown to be the best continuous approximation to a discrete Markov process using the Kolmogorov Forward Equation.

7.2.5 SDE Approximation of the Wright-Fisher Model

Recall that in the Wright-Fisher Model we have that

$$\mathbb{V}[X_n] = 2NP_{n-1}(1 - P_{n-1}).$$

Notice too then that

$$\mathbb{V}[P_n] = P_{n-1}(1 - P_{n-1}).$$

Because

$$\mathbb{E}[X_n | X_{n-1}] = X_{n-1}$$

we get that

$$\frac{d\mathbb{E}[X_n]}{dt} = 0.$$

This means that we define the SDE approximation of the Wright-Fisher model to be the one that matches the mean and variances, that is

$$dX_t = \sqrt{X_t(1 - X_t)}dW_t.$$

Notice that this approximation converges as $N \to \infty$.

7.2.6 Extensions to the Wright-Fisher Model: Selection

Now we extend the Wright-Fisher model to incorporate more of the population dynamics. For example, assume that there is selection. Assume that the fitness of A is 1 and that the fitness of a is $1 - s$ for some constant s. We define this using a re-sampling idea. We say that if A is sampled, will accept it with probability 1, while if a is sampled, we will accept it with a probability $1 - s$. This means that the probability of choosing A is

$$\frac{X_{n-1}}{X_{n-1} + (1 - X_{n-1})(1 - s)}.$$

Assuming s is small, the probability of choosing A is

$$\approx \frac{X_{n-1}}{1 - (1 - X_{n-1})s} = X_{n-1}\left(1 + (1 - X_{n-1})\, s + \mathcal{O}(s^2)\right)$$

and thus

$$\Delta X_n = X_n - X_{n-1} \approx X_{n-1}(1 - X_{n-1})s.$$

Therefore we make

$$\frac{d\mathbb{E}[X_n]}{dt} = sX_{n-1}(1 - X_{n-1}).$$

To make this continuous we note

$$\begin{aligned} \frac{d\mathbb{E}[P_n]}{dt} &= \frac{s}{2N} P_{n-1}(1 - P_{n-1}). \\ &= \gamma P_{n-1}(1 - P_{n-1}) \end{aligned}$$

where $\gamma = \frac{s}{2N}$. Therefore we make this continuous by matching the diffusion operator as

$$\mathcal{L} = \gamma x(1-x)\frac{d}{dx} + \frac{1}{2}x(1-x)\frac{d^2}{dx^2}$$

and thus we approximate the Wright-Fisher model with selection as the SDE

$$dX_t = \gamma X_t\left(1-X_t\right)dt + \sqrt{X_t\left(1-X_t\right)}dW_t$$

when N is large.

7.2.7 Extensions to the Wright-Fisher Model: Mutation

Let $\mu_1 = \frac{\beta_1}{2N}$ be the chance of a allele randomly changing to A allele, and $\mu_2 = \frac{\beta_2}{2N}$ be the chance of the reverse mutation. One can derive the differential generator for this process to be

$$\mathcal{L} = \left(\gamma x(1-x) + \beta_1\left(1-x\right) + \beta_2 x\right)\frac{d}{dx} + \frac{1}{2}x(1-x)\frac{d^2}{dx^2}.$$

Therefore the best SDE approximation to the Wright-Fisher Model with selection and mutation is

$$dX_t = \left(\gamma x(1-x) + \beta_1\left(1-x\right) + \beta_2 x\right)dt + \sqrt{X_t\left(1-X_t\right)}dW_t.$$

7.2.8 Hitting Probability (Without Mutation)

Let $h(t) = \Pr(\tau \leq t)$ where $\tau = \inf_t\left\{X(t) = a\right\}$. For example, $a = 1$ would mean that τ is the time for fixation by A. If there is no selection in the model, then $h(t)$ is a martingale and thus

$$\mathbb{E}\left[h(t)|\mathcal{F}_s\right] = h(s)$$

and therefore, in the limit as $t \to 0$,

$$\begin{aligned} h(x) &= \int_0^1 \Pr(X_t = y|X_0 = x)h(y)dy \\ \int_0^1 \delta(x-y)h(y)dy &= \int_0^1 \Pr(X_t = y|X_0 = x)h(y)dy \\ 0 &= \int_0^1 \left[\Pr(X_t = y|X_0 = x) - \delta(x-y)\right]h(y)dy. \end{aligned}$$

This, in the limit as $t \to 0$,

$$\Pr(X_t = y|X_0 = x) - \delta(x-y) \to \frac{\partial \rho}{\partial t}$$

Recall that from the Kolmogorov Forward Equation that

$$\frac{\partial \rho}{\partial t} = \frac{\partial(a\rho)}{\partial x} + \frac{1}{2}b^2\frac{\partial^2 \rho}{\partial x^2}.$$

Thus

$$\begin{aligned} 0 &= \int_0^1 \frac{\partial \rho}{\partial t} h(y) dy \\ 0 &= \int_0^1 \left[-\frac{\partial (a\rho)}{\partial x} + \frac{1}{2} b^2 \frac{\partial^2 \rho}{\partial x^2} \right] h(y) dy \end{aligned}$$

To solve this we use integration by parts to get the PDE

$$\begin{aligned} \mathcal{L}h &= 0 \\ h(a) &= 1 \\ h(b) &= 0 \end{aligned}$$

where

$$\mathcal{L} = a(x)\frac{d}{dx} + \frac{1}{2} b(x) \frac{d^2}{dx}.$$

To solve for the probability density function of the first-passage time, define $\psi = \frac{dh}{dx}$ and thus $\psi' = -\frac{a(x)}{b(x)}$. Therefore we get

$$\psi(x) = \psi(0) e^{-\int_0^x \frac{a(y)}{b(y)} dy}$$

and

$$h(x) = \int_0^x \psi(y) dy.$$

This is expanded in the book by Kimora and Crow.

7.2.9 Understanding Using Kolmogorov

Notice that, given the SDE, we can understand the dynamics using the Kolmogorov equation. The fixation time distribution can be calculated by using the forward Kolmogorov with absorbing conditions at 0 and 1, that is $\rho(0,t) = \rho(1,t)$ and thus we can solve for the fixation times using the first-passage time distributions as examined in 6.6.1. Also note that when we have mutation, there exists a non-trivial steady state distribution. Using the Kolmogorov equation, we can solve for this steady state distribution as the distribution π s.t. $\frac{\partial \pi}{\partial t} = 0$ and thus

$$0 = -\frac{\partial}{\partial x}[a(x)\pi(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[b^2(x)\pi(x,t)\right].$$

For the Wright-Fisher model with selection and mutation, we note that this can be solved so that

$$\pi(x) = x^{2\beta_1 - 1}(1-x)^{2\beta_2 - 1} = \Gamma(2\beta - 1).$$

Notice that this solution is only valued when $\beta_1, \beta_2 > \frac{1}{2}$, giving us the necessary and sufficient conditions for a non-trivial steady state.

7.3 Stochastic Control

The focus of this lecture will be on stochastic control. We begin by looking at optimal control in the deterministic case.

7.3.1 Deterministic Optimal Control

Suppose you have a deterministic ODE

$$\dot{x} = F(x(t), u(t))$$

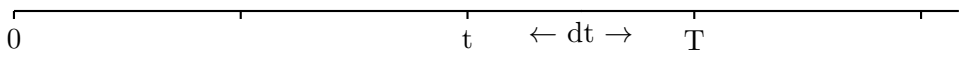
where $x(t)$ is the state variable and $u(t)$ is the control variable. We wish to apply optimal control over a fixed time period $[0, T]$ such that

$$V(x(t), t) = \min_u \left\{ \int_0^T C[x(t), u(t)]dt + D[x(T)] \right\}$$

where C is the cost function, D is the terminal cost, and $V(x(t), t)$ is the optimal or minimal cost at a time t. Thus what we are looking to do is, given some cost function C, we are looking to minimize the total cost. Let's say for example that we want to keep x at a value c, but sending control signals costs some amount α. Thus an example could be that $C(x(t), u(t)) = (x(t) - c)^2 + \alpha u(t)$. What we want to do is solve for the best $u(t)$.

7.3.2 Dynamic Programming

The typical way of trying to solve this general problem is using a dynamic programming technique.



At point t there is an optimal value $V(x(t), t)$. One way to try to control the system is by taking small steps and observing how the system evolves. We will do this by stepping backwards. Divide the cost into two components. Notice that

$$V(x(t), t) = \min_u \{C(x(t), u(t))dt + V(x(t + dt), t + dt)\}$$

Essentially the goal is to find the $u(t)$ for the increment dt that minimizes the growth of V. Do a Taylor expansion on V to get

$$V(x(t + dt), t + dt) \approx V(x(t), t) + \frac{\partial V}{\partial t} + \frac{\partial V}{\partial X} \cdot \dot{x}(t)dt$$

The third part of the above equation $\frac{\partial V}{\partial X} \cdot \dot{x}(t)dt$ is not necessarily scalar, but will be the dot product between two scalars. The importance of this is because it is recursive, and this will lead to the minimal solution. By plugging this in we get

$$\min_u \left\{ C(x(t), u(t)) + \frac{\partial V}{\partial t} + \left\langle \frac{\partial V}{\partial X}, F(x(t), u(t)) \right\rangle \right\} = 0.$$

since $V(x(t), t)$ does not depend on $u(t)$ in the range $(t, t + dt)$. We move $\frac{\partial V}{\partial t}$ outside to get

$$\frac{\partial V}{\partial t} + \min_u \left\{ C(x, u) + \left\langle \frac{\partial V}{\partial x}, F(x, u) \right\rangle \right\} = 0$$

The initial condition is the at the cost at the end must equal the terminal cost:

$$V(x, T) = D(x).$$

This defines an ODE where, counting back from T, the terminal cost is the initial condition and we solve backwards using the stepping equation. However, the PDE that needs to be solved to find the optimal control of the system which may be hard because the minimization may be nontrivial. This is known as deterministic optimal control, where the best u will be found. These types of equations are known as Hamilton-Jacobian-Bellman (HJB) equations, and are famous if you are studying optimal control.

7.3.3 Stochastic Optimal Control

Now we look at the same problem where the underlying process is the stochastic process

$$dX_t = b(x_t, u_t)dt + \sigma(x_t, u_t)dW_t.$$

Let $X_t \in \mathbb{R}^n$ and W_t be an m-dimensional Brownian motion. For this to work, we need $\sigma \in \mathbb{R}^{n \times m}$. Define the value equation as the minimum expected cost, that is

$$V(x(t), t) = \min_u E\left[\int_t^T C(X(t), u(t))dt + D[X(T)] \right]$$

What type of control is being used? That is the question that needs to be addressed, because there are many types of controls:

1. Open Loop Control (Deterministic Control).

   Suppose $u(t, \omega) = u(t)$. This case has no random events and thus it will be deterministic control (open looped control).

2. Open Looped Control (Feedback Control).

   Suppose U_t is M_t-adapted, where M_t is the σ-algebra generated by $X_s, 0 \leq s \leq t$. Essentially for this σ-algebra you have all the information about the trajectory from 0 up to a certain time point, the history must be known.

3. Markov Control

   $U(t,\omega) = u_0(t, x_t(\omega))$. Markov control uses less control than the open looped control. It only uses what is current, there is nothing beyond that. This is commonly used in programming applications because only the last state needs to be saved, leading to iterative solutions that do not require much working memory or RAM. An example of where this would be used is a control theory about robots. The robot has to decide to walk or stop, and this decision only depends on the current state.

These types of controls explain what type of information is allowed to be used. Regardless, we solve the problem using dynamic programming but this time we use the stochastic Taylor series expansion:

$$V(x(t+dt), t+dt) = \frac{\partial V}{\partial t} \cdot dt + \frac{\partial V}{\partial x} \cdot dX(t) + \frac{1}{2}(dx_t)^T \frac{\partial^2 V}{\partial x^2}(dx_t)$$

where $\frac{\partial^2 V}{\partial x^2}$ is the Hessian of V. Thus

$$V(x(t+dt), t+dt) = V(x(t), t) + \frac{\partial V}{\partial t}dt + \left\langle \frac{\partial V}{\partial x}, b(x_t, u_t) \right\rangle dt + \left\langle \frac{\partial V}{\partial x}, \sigma(x_t, u_t) dB_t \right\rangle + \frac{1}{2}\sum_{i,j} a_{ij} \frac{\partial^2 V}{\partial x_i \partial x_j}$$

where

$$a_{ij} = (\sigma\sigma^T)_{ij}.$$

After obtaining the expectation we solve as before to get

$$\frac{\partial V}{\partial t} + \min_u \left\{ C(x,u) + \langle \frac{\partial V}{\partial x}, b(x_t, u_t) \rangle + \frac{1}{2}\sum_{ij} a_{ij} \frac{\partial^2 V}{\partial x_i \partial x_j} \right\} = 0$$

with the same initial (final) condition

$$V(x, T) = D(x).$$

Whether or not the solution exists, or is unique are hard questions. In general the goal is a practical solution to the HJB equations. This type of equation does not necessarily have an analytical solution. Though caution needs to be exercised as numerical solutions have issues as well. For example, how could one find the u that minimize this? Some form of the calculus of variations? There is no clear way of how to do this.

7.3.4 Example: Linear Stochastic Control

Suppose

$$dX_t = (H_t X_t + M_t U_t)dt + \sigma_t dW_t$$

where

$$x_0 = x, \quad t \geq 0. \;\; H_t \in R^{n/n}, U_t \in R^k,$$

$$\sigma_t R^{n\times m}, \;\; M_t \in R^{n\times k}.$$

This minimizes u over the expectation,

$$V^a(x,0) = E^{x,0}\left\{\int_0^T (x_t^T C_t X_t + u_t^T D_t u_t)dt + X_T^T R X_T\right\}$$

where C_t are the costs of controlling at t and D_t is the terminal cost. Try to make the u small. We denote

$$\psi(t,x) = \min_u V^u$$

Using s as the time stand-in variable, we plug the equation into the HJB equation to get

$$\frac{\partial \psi}{\partial s} + \min_u \left\{ x^T C_s X + u^T P_s v + \sum_{i=1}^n (H_s x + M_s)_i \frac{\partial \psi}{\partial x_i} + \frac{1}{2}\sum_{ij} (\sigma_s \sigma_s^T)_{ij} \frac{\partial^2 \psi}{\partial x_i \partial x_j} \right\} = 0$$

To solve this, try the solution

$$\psi(x,t) = X_t^T S_t x_t + a_t$$

where S_t is positive semi-definite. Thus we get

$$x^T \dot{S}_t x + \dot{a}_t + \min_u \left\{ x^T C_t x + V^T D_t v + v^T + < H_t x + M_t v, 2S_t x > + \sum_{ij} (\sigma_s \sigma_s^T)_{ij} (S_t)_{ij} = 0 \right\}$$

$$x\dot{S}_t x + \dot{a}_t + \min_u \left\{ x^T c_t x + V^T D_t v + < H_t x + M_t v m 2S) t x > \mathrm{tr}[(\sigma_t \sigma_t^T) S_t] \right\} = 0$$

We note that the optimal u is

$$u* = -D_t^{-1} M_t^T S_t x.$$

Plugging in this we solve to

$$x^T(\dot{s} + C_t - S_t M_t D_t^{-1} M_t^T S_t + 2H_t^T S_t)x + \dot{a}_t + tr(\sigma\sigma^T S)_t = 0.$$

By matching coefficients, we get the equations

$$\dot{S}_t + C_t - S_t M_t D_t^{-1} M_t^T S_t + 2H_t^T S_t = 0 \;\; S_T = R$$
$$\dot{a}_t = -tr(\sigma\sigma^T s)_t, \;\; a_T = 0$$

The first equation is called a Ricarti equation.

$$\dot{S}_t + S_t A_t S_t + B_t S_t + C_t = 0$$

These two equations give you S_t and a_t such that we arrive at the optimal value.

7.4 Stochastic Filtering

The general problem involves a system of equations

$$\text{(System)} \quad dX_t = b(t, x_t)dt + \sigma(t, x_t)dU_t$$

$$\text{(Observations)} \quad dZ_t = C_t, x_t)dt + \gamma(t, x_t)dV_t$$

$$U_t : p\text{-dim Brownian motion}$$

$$V_t : r\text{-dim Brownian motion}$$

where X_t is the state, Z_t is the observation, and U_t and V_t are two independent Brownian motions. Assume F, G, C, D are bounded on bounded intervals, $Z_0 = 0$.

In this problem, we seek to estimate the value of the system X at a future time t based on the observations U up to the present time $s < t$, that is, conditional on $\mathcal{G}_t$, the σalgebra generated by $\{Z_s\}_{0\leq s\leq t}$.

7.4.1 The Best Estimate: $\mathbb{E}[X_t|\mathcal{G}_t]$

First you cannot disrespect the intuitive argument. The intuitive argument is that the best prediction is to find the value you would expect given all of the information you have. The information you have is $\mathcal{G}_t$, and so the best prediction given the totality of information is $\mathbb{E}[X_t|\mathcal{G}_t]$.

More formally, we take that the best estimate based on the observations would be a function that only needs to use the information, implying that $\hat{X}(\circ)$ is $\mathcal{G}_t$-measurable. We define the best estimate as the one that minimizes the Euclidean distance, that is

$$\hat{X}(\circ) = \inf_{Y\in\mathcal{K}} \left\{\mathbb{E}\left[(X_t - Y)^2\right]\right\}$$

where

$$\mathcal{K}_t := \left\{Y : \Omega \to \mathbb{R}^n : Y \in L^2(P) \text{ and } Y \text{ is } \mathcal{G}_t\text{-measurable}\right\}.$$

with $L^2(P)$ being the set of L_2integrable functions by the measure P.

Theorem: Let $\mathcal{G}_t \subset \mathcal{F}_t$ be a sub-σ-algebra and let $X \in L^2(p)$ be $\mathcal{F}_t$-measurable. Let $\mathcal{N} = \left\{Y \in L^2(P) : Y \text{ is}\mathcal{G}_t\text{-measurable}\right\}$. It follows that

$$\mathcal{P}_{\mathcal{N}}(X_t) = \mathbb{E}[X_t|\mathcal{G}_t]$$

where $\mathcal{P}_{\mathcal{N}}$ is the orthogonal projection of X onto $\mathcal{N}$.

Proof: To prove that $\mathcal{P}_{\mathcal{N}}(X_t) = \mathbb{E}[X_t|\mathcal{G}_t]$, we simply need to show it satisfies the two properties of the conditional expectation. Notice that it is trivial that $\mathcal{P}_{\mathcal{N}}(X_t)$ is $\mathcal{G}_t$-measurable since every $X \in \mathcal{N}$ is $\mathcal{G}_t$-measurable. Thus we just need to check the Partial Averaging Property. Since $\mathcal{P}_{\mathcal{N}}$ is an orthogonal projection onto N, we get that $X - \mathcal{P}_{\mathcal{N}}(X)$ is orthogonal to N. Now take $I_A \in \mathcal{N}$ as an arbitrary indicator function for $A \in \mathcal{G}_t$. This means that we define I_A as

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & o.w. \end{cases}.$$

Since $I_A \in \mathcal{N}$, $X - \mathcal{P}_{\mathcal{N}}(X)$ is orthogonal to I_A. Thus we from the Hilbert-space dot product that

$$\langle X - \mathcal{P}_{\mathcal{N}}(X), I_A \rangle = 0 = \int_\Omega (X - \mathcal{P}_{\mathcal{N}}(X))\, I_A dp = \int_A (X - \mathcal{P}_{\mathcal{N}}(X))\, dp$$

and thus, since I_A was arbitrary, for all $A \in \mathcal{G}_t$,

$$\int_A X dp = \int_A \mathcal{P}_{\mathcal{N}}(X) dp.$$

Thus the partial averaging property is satisfied completed our proof.

However, the Hilbert Projection Theorem (from Wikipedia or Rudin) states that there is a unique Y in the projection space $\mathcal{N}$ such that $(X_t - Y)^2$ is minimized, which gives the necessary and sufficient condition that the vector $x - y$ is orthogonal to $\mathcal{N}$. This means that the vector Y which minimizes $(X_t - Y)^2$ is also the projection of X_t onto the space $\mathcal{N}$! Thus we have that

$$\inf_{Y \in \mathcal{K}} \left\{ \mathbb{E}\left[(X_t - Y)^2 \right] \right\} = \mathcal{P}_{\mathcal{N}}(X_t)$$

to get the relation

$$\hat{X}_t = \mathbb{E}\left[X_t | \mathcal{G} \right].$$

7.4.2 Linear Filtering Problem

In order to obtain a solution, we will look at an easier case: the Linear Filtering Problem. Consider the following 1-dimensional linear system with linear observations:

$$\begin{aligned} dX_t &= F(t)X_t dt + C(t)dU_t; \quad F(t), C(t) \in \mathbb{R} \\ dZ_t &= G(t)X_t dt + D(t)dV_t; \quad G(t), D(t) \in \mathbb{R} \end{aligned}$$

where X_t is the state, Z_t is the observation, and U_t and V_t are two independent Brownian motions. Assume F, G, C, D are bounded on bounded intervals, $Z_0 = 0$, and X_0 is normally distributed.

For this problem, we will simply outline the derivation for the Kalman-Bucy Filter and provide intuition for what the derivation means (for the full proof, see Oksendal). The derivation proceeds as follows:

Step 1: Show It's A Gaussian Process Let $\mathcal{L}$ be the closure (the set including its limit points) of the set $L^2(p)$ of random variables that are linear combinations of the form

$$c_0 + c_1 Z_{s_1}(\omega_1) + c_2 Z_{s_2}(\omega) + \ldots + c_k Z_{s_k}(\omega)$$

with $s_j \leq t$ and each $c_j \in \mathbb{R}$. Let $\mathcal{P}_{\mathcal{L}}$ be the projection from $L^2(p)$ onto $\mathcal{L}$. It follows that

$$\hat{X}_t = \mathcal{P}_{\mathcal{L}}(X_t).$$

We can interpret this step as saying that the best estimate for X_t can be written as a linear combination of past values of Z_t. Notice that since the variance term in the SDE is not dependent on Z and X, the solution will be a Gaussian distribution. Since the sum of Gaussian distributions is a Gaussian distribution, this implies that $\hat{X}$ is Gaussian distributed! This gives the grounding for our connection between estimating X_t from Z_t by using Brownian motions and Gaussian processes. Because this step is so important, we include a proof.

Theorem: Take X, Z_s $s \leq t$ be random variables in $L^2(p)$ and assume that $(X, Z_{s_1}, \ldots, Z_{s_n}) \in \mathbb{R}^{n+1}$ has a normal distribution. For all $s_1, \ldots, s_n \leq t$ with $n \geq 1$, it follows that

$$\mathcal{P}_{\mathcal{L}}(X) = \mathbb{E}[X|\mathcal{G}_t] = \mathcal{P}_{\mathcal{K}}(X).$$

Proof: Define $\check{X} = \mathcal{P}_{\mathcal{L}}(X)$ and $\tilde{X} = X - \check{X}$. This means that $\tilde{X} \perp \mathcal{L}$. Thus we can conclude $\mathcal{P}_{\mathcal{L}}(X) = \mathcal{P}_{\mathcal{K}}(X)$ if $\tilde{X}$ must be trivial, that is $\tilde{X} = 0$. We do this in steps:

1. If $(y_1, \ldots, y_k) \in \mathbb{R}^n$ is normally distributed, then $c_1 y_1 + \ldots + c_k y_k$ is normally distributed. We leave out the proof that this means that in the limit as $k \to \infty$ this is still normally distributed. Thus, since $(X, Z_{s_1}, \ldots, Z_{s_n})$ is normally distributed, $\check{X}$ is normally distributed.

2. Since $\tilde{X} \perp \check{X}$ and each $Z_{s_j} \in \mathcal{L}$, $\tilde{X}$ is orthogonal to each Z_{s_j}. Thus $\mathbb{E}\left[\tilde{X} Z_{s_j}\right] = 0$. Since the Z_s's are jointly Gaussian, non-correlation implies independence. $\tilde{X}$ is independent of $Z_{s_i}, \ldots, Z_{s_n}$. Thus denoting $\mathcal{G}$ as the σ-algebra generated by the Z_s, we get $\tilde{X}$ is independent of $\mathcal{G}_t$.

3. Take $I_{\mathcal{G}}$ to be the indicator function for events ω in the any arbitrary set $G \subset \mathcal{G}_t$. Since $\check{X} = X - \check{X}$, we multiply both sides by the indicator and take expectations to get

$$\mathbb{E}\left[\left(X - \check{X}\right) I_G\right] = \mathbb{E}\left[I_G \tilde{X}\right].$$

Since $\tilde{X}$ is independent of $\mathcal{G}$,

$$\begin{aligned} \mathbb{E}\left[\left(X - \check{X}\right) I_G\right] &= \mathbb{E}[I_G]\,\mathbb{E}\left[\tilde{X}\right] \\ &= \mathbb{E}[I_G]\,\mathbb{E}\left[X - \check{X}\right]. \end{aligned}$$

Since the probability of any individual event is measure zero, $\mathbb{E}[I_G] = 0$. Thus

$$\mathbb{E}\left[\left(X - \check{X}\right) I_G\right] = 0$$

which gives

$$\int_G X dp = \int_G \check{X} dp$$

for any $G \subset \mathcal{G}_t$. Thus the partial averaging property is satisfied, meaning

$$\check{X} = \mathbb{E}[X|\mathcal{G}_t]$$

completing the proof.

Step 2: Estimate Using a Gram-Schmidt-Like Procedure To understand this step, recall the Gram-Schmidt Procedure. What the Gram-Schmidt procedure does is, given a countable set of vectors, it finds a basis set of orthogonal vectors that will span the same space. It does this by iteration. First, it takes the first vector as the first basis vector. Then it recursively does the following. You take the next vector v, find its projection of this vector onto the current basis space by using the dot product (call it v_p), and then, knowing that this implies $v - v_p$ is orthogonal to the basis space, we add $v - v_P$ as a basis vector.

Here we do a similar procedure. Since there are countably many Z_t that are used to span $\mathcal{P}_{\mathcal{L}}$, order them. We replace Z_t by the *innovation process* N_t defined as

$$N_t = Z_t - \int_0^t (GX)_s^{\wedge} \, ds$$

where

$$(GX)_s^{\wedge} = \mathcal{P}_{\mathcal{L}(X,s)}(G(s)X(s) = G(s)\hat{X}(s)$$

or equivalently

$$\begin{aligned} dN_t &= dZ_t - G(t)\hat{X}_t dt \\ &= G(t)\left(X - \hat{X}\right) dt + D(t) dV_t. \end{aligned}$$

Note that $(GX)_s^{\wedge}$ basically the "basis set spanned by the $t < s$" and thus the next set we add to the basis set is kind of Z_t minus that the dot product with that value. Thus this is a type of continuous version of the Gram-Schmidt procedure. We can prove that the following properties hold:

1. N_t has orthogonal increments: $\mathbb{E}\left[(N_{t_1} - N_{s_1})(N_{t_2} - N_{s_2})\right] = 0$ for every non-overlapping $[s_1, t_1]$ and $[s_2, t_2]$. So each time increment is orthogonal (since time is the basis of this Gram-Schmidt-Like procedure).

2. $\mathcal{L}(N, t) = \mathcal{L}(Z, t)$. That is simply that N and Z span the same space, as guaranteed by the Gram-Schmidt process.

Step 3: Find the Brownian Motion Define $dR_t = \frac{dN_t}{D(t)}$. Using the non-overlapping independence property, we can show that R_t is actually a Brownian motion. Notice trivially that $\mathcal{L}(N, t) = \mathcal{L}(R, t)$ and thus the space spanned by this Brownian motion is sufficient for the estimation of $\hat{X}$. Therefore we get that

$$\begin{aligned} \hat{X} &= \mathcal{P}_{\mathcal{L}(R,t)}(X(t)) \\ &= \mathbb{E}\left[X_t\right] + \int_0^t \frac{\partial}{\partial s} \mathbb{E}\left[X_t R_s\right] dR_s. \end{aligned}$$

To solve this, simply derive the SDE

$$d\hat{X}_t = \left(F(t) - \frac{G^2(t)S(t)}{D^2(t)}\right)\hat{X}_t dt + \frac{G(t)S(t)}{D^2(t)} dR_t$$

where

$$\hat{X}_0 = \mathbb{E}[X_0]$$

and $S(t) = \mathbb{E}\left[\left(X - \hat{X}\right)^2\right]$ satisfies the Riccardi equation

$$\begin{aligned} S'(t) &= 2F(t)S(t) - \frac{G^2(t)}{D^2(t)}S^2(t) + C^2(t) \\ S(0) &= \mathbb{E}\left[(X_0 - \mathbb{E}[X_0])^2\right]. \end{aligned}$$

This solution is known as the Kalman-Bucy Filter.

7.5 Discussion About the Kalman-Bucy Filter

Notice that the Kalman Filter can be though of as the best way to estimate an unobservable process $X(t)$ from observable variables $Z(t)$. In robotics, this could be like estimating the location you are currently at given your measurements of how much your wheels have turned (the connection between this and actual location is stochastic because of jittering, skidding, etc.). Notice that in the linear equation

$$\begin{aligned} dX_t &= F(t)X_t dt + C(t)dU_t; \quad F(t), C(t) \in \mathbb{R} \\ dZ_t &= G(t)X_t dt + D(t)dV_t; \quad G(t), D(t) \in \mathbb{R} \end{aligned}$$

that, if we do not know the form of the equation, we can think of F, C, G, and D as unobservables as well. Thus using a multidimensional version of the Kalman filter, we can iteratively estimate the "constants" too! Thus, if we discretize this problem, we can estimate the future values by estimating G, D, F, and C, and then using these to uncover X. Notice that since these constants are changing, this type of a linear solution can approximate any non-linear interaction by simply making the time-step small enough. Thus even though this is just the "linear approximation", the linear approximation can computationally solve the non-linear problem! For a detailed assessment of how this is done, refer to Hamilton's *Time Series Analysis*.

8 Stochastic Calculus Cheat-Sheet

Basic Probability Definitions

Binomial Distribution $\sim Bin(n, p)$

- Distribution function:$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Cumulative Distribution: $P(X \leq k) = \sum_{i \leq k} P(X = i)$
- Expectation: $\mathbb{E}[X] = \sum_{i=1}^{n} kP(X = k) = np$
- Variance: $\mathbb{V}[X] = \mathbb{E}[X - \mathbb{E}[X]] = np(1-p)$

Poisson Distribution $\sim Poisson(\lambda)$

- Density function: $\rho(X) = \lambda e^{-\lambda t}$
- Cumulative Distribution: $P(X \leq a) = \int_{-\infty}^{a} \rho(x)dx$
- Expectation: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\rho(x)dx = \lambda$
- Variance: $\mathbb{V}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[x])^2 \rho(x)dx = \lambda$

Gaussian Distribution $\sim N(\mu, \sigma^2)$

- Density function: $\rho(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Cumulative Distribution: $P(X \leq a) = \int_{-\infty}^{a} \rho(x)dx$
- Expectation: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\rho(x)dx = \mu$
- Variance: $\mathbb{V}[X] = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \rho(x)dx = \sigma^2$

Useful Properties

- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X, Y are independent
- $\mathbb{V}[aX + bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$ if X, Y are independent

Poisson Counter SDEs

$$dx = f(x,t)dt + \sum_{i=1}^{m} g_i(x,t)dN_i$$

$$N_i(t) \sim Poisson(\lambda_i t)$$

$$\mathbb{E}\left[N_i(t)\right] = \lambda t$$

$$\int_0^t dN_t = \lim_{\Delta t \to 0} \sum_{i=0}^{n-1} \left(N(t_{i+1}) - N(t_i)\right), \quad t_i = i\Delta t$$

$$P\left(k \text{ jumps in the interval } (t, t+dt)\right) = \frac{(\lambda dt)^k}{k!} e^{-\lambda dt}$$

Ito's Rule

$$dY_t = d\psi(x,t) = \frac{\partial \psi}{\partial t}dt + \frac{\partial \psi}{\partial x} f(x)dt + \sum_{i=1}^{m} [\psi(x + g_i(x), t - \psi(x,t)]dN_i$$

Forward Kolmogorov Equation

$$\frac{\partial p}{\partial t} = -\frac{\partial (fp)}{\partial x} + \sum_{i=1}^{m} \lambda_i \left[\frac{\rho(\tilde{g}_i^{-1}(x), t)}{|1 + g_i'(\tilde{g}_i^{-1}(x)|} - \rho\right] \quad \tilde{g}_i(x) = x + g_i(x)$$

Wiener Process SDEs

$$dx = f(x,t)dt + \sum_{i=1}^{n} g_i(x,t)dW_i$$

$$W_t \sim N(0,t)$$

$$(dt)^n = \begin{cases} dt, & n = 1 \\ 0, & o.w \end{cases}$$

$$dW_i^n dt^m = 0$$

$$dW_i \times dW_j = \begin{cases} dt & : i = j \\ 0 & : i \neq j \end{cases}$$

$$\int_0^t g(X,t)dW_t = \lim_{\Delta t \to 0} \sum_{i=1}^{n-1} g(X_{t_i}, t_i)\left(dW_{t_{i+1}} - dW_{t_i}\right), \quad t_i = i\Delta t$$

$$dy = d\psi(x,t) = \frac{\partial \psi}{\partial t}dt + \frac{\partial \psi}{\partial x}dx + \frac{1}{2}\frac{\partial^2 \psi}{\partial x^2}(dx)^2$$

Ito's Rule

$$dy = d\psi(x,t) = \left(\frac{\partial \psi}{\partial t} + f(x,t)\frac{\partial \psi}{\partial x} + \frac{1}{2}\sum_{i=1}^{n} g_i^2(x,t)\frac{\partial^2 \psi}{\partial x^2}\right)dt + \frac{\partial \psi}{\partial x}\sum_{i=1}^{n} g_i(x,t)dW_i$$

Multidimensional $\mathbf{X} \in \mathbb{R}^n$ Ito's Rule

$$d\psi(\mathbf{X}) = \left\langle \frac{\partial \psi}{\partial \mathbf{X}}, f(\mathbf{X}) \right\rangle dt + \sum_{i=1}^{m} \left\langle \frac{\partial \psi}{\partial \mathbf{X}}, g_i(\mathbf{X}) \right\rangle dW_i + \frac{1}{2}\sum_{i=1}^{m} g_i(\mathbf{X})^T \nabla^2 \psi(\mathbf{X}) g_i(\mathbf{X}) dt$$

Forward Kolmogorov Equation (Fokker-Planck Equation)

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[f(x)\rho(x,t)] + \frac{1}{2}\sum_{i=1}^{m} \frac{\partial^2}{\partial x^2}\left[g_i^2(x)\rho(x,t)\right]$$

Backward Kolmogorov Equation

$$\frac{\partial \rho}{\partial t} + a(x(t),t)\frac{\partial \rho}{\partial x} + \frac{1}{2}b^2(x(t),t)\frac{\partial^2 \rho}{\partial x^2} = 0$$

Fluctuation-Dissipation Theorem

$$J_f(X_{ss},t)\Sigma(X_{ss},t) + \Sigma(X_{ss},t)J_f^T(X_{ss},t) = -g^2(X_{ss},t).$$

Useful Properties

1. Product Rule: $d(X_tY_t) = X_t dY + Y_t dX + dX dY$.
2. Integration By Parts: $\int_0^t X_t dY_t = X_tY_t - X_0Y_0 - \int_0^t Y_t dX_t - \int_0^t dX_t dY_t$.
3. $\mathbb{E}\left[(W(t) - W(s))^2\right] = t - s$ for $t > s$.
4. $\mathbb{E}[W(t_1)W(t_2)] = \min(t_1, t_2)$.
5. Independent Increments: $\mathbb{E}\left[(W_{t_i} - W_{s_1})(W_{t_2} - W_{s_2})\right] = 0$ if $[t_1, s_1]$ does not overlap $[t_2, s_2]$.
6. $\mathbb{E}\left[\int_0^t h(t)dW_t\right] = \mathbb{E}\left[h(t)dW_t\right] = 0$.
7. Ito Isometry: $\mathbb{E}\left[\left(\int_0^T X_t dW_t\right)\right] = \mathbb{E}\left[\int_0^T X_t^2 dt\right]$

Simulation Methods

$$dx = f(x,t)dt + \sum_{i=1}^{n} g_i(x,t)dW_i \quad \eta_i, \lambda_i \sim N(0, \Delta t) \quad \eta_i \text{ and } \lambda_i \text{ are independent}$$

Euler-Maruyama (Strong Order 1/2, Weak Order 1)

$$X(t+\Delta t) = X(t) + f(X,t)\Delta t + \sqrt{\Delta t} g(X,t)\eta_i.$$

Milstein's Method (Strong Order 1, Weak Order 1)

$$X(t+\Delta t) = X(t) + \left(f(X,t) - \frac{1}{2}g(X,t)g_x(X,t)\right)\Delta t + \sqrt{\Delta t} g(X,t)\eta_i + \frac{\Delta t}{2} g(X,t)g_x(X,t)\eta_i^2.$$

KPS Method (Strong Order 1.5, Weak Order 1.5)

$$\begin{aligned} X(t+\Delta t) &= X(t) + f\Delta t + g\Delta W_t + \frac{\Delta t}{2} g g_x \left((\Delta W_t)^2 - \Delta t\right) \\ &+ g f_x \Delta U_t + \frac{1}{2}\left(f f_x + \frac{1}{2} g^2 f_{xx}\right)\Delta t^2 \\ &+ \left(f g_x + \frac{1}{2} g^2 g_{xx}\right)(\Delta W_t \Delta t - \Delta U_t) \\ &+ \frac{1}{2} g \left(g g_x\right)_x \left(\frac{1}{3}(\Delta W_t)^2 - \Delta t\right)\Delta W_t \end{aligned}$$

$$\Delta W_t = \sqrt{\Delta t}\eta_i \quad \Delta U_t = \frac{1}{3}\Delta t^3 \lambda_i$$

Other Properties

Properties of Conditional Expectation

1. $\mathbb{E}[X|G]$ exists and is unique except on a set of measure 0.
2. $\mathbb{E}\left[\mathbb{E}[X|G]\right] = \mathbb{E}[X]$. Our best estimate of $\mathbb{E}[X|G]$ is $\mathbb{E}[X]$ if we have no information.
3. Linearity: $\mathbb{E}\left[aX + bY|G\right] = a\mathbb{E}\left[X|G\right] + b\mathbb{E}[Y|G]$.
4. Positivity: If $X \geq 0$ a.s., then $\mathbb{E}[X|G] \geq 0$ a.s.
   (a) This can be generalized: For all $A \subset \mathbb{R}$, if $X \in A$ a.s., then $\mathbb{E}[X|G] \in A$ a.s.
5. Taking out what is known: If X is G-measurable, then $\mathbb{E}[XY|G] = X\mathbb{E}[Y|G]$.Notice that this is because if X is G-measurable, it is known given the information of G and thus can be treated as a constant.

6. Iterated Conditioning: If $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then $\mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}\right] = \mathbb{E}[X|\mathcal{H}]$.

7. Independence of $\mathcal{G}$: If X is independent of $\mathcal{G}$, then $\mathbb{E}[X|\mathcal{G}] = E[X]$. If $\mathcal{G}$ gives no information about X, then the expectation condition on the information of $\mathcal{G}$ is simply the expectation of X.

8. If M_t is a martingale, then $\mathbb{E}[M_t|\mathcal{F}_s] = M_s$.