Math 158, CME 298 Spring 2025 Notes

George Papanicolaou

May 4, 2025

Contents

1 Sums of independent identically distributed random variables

We will be dealing with sequences of independent identically distributed random variables $X_1, X_2, ..., X_n$ where $P\{X_j \leq x\} = F(x)$ is their common distribution function. We will also use the notation $F_X(x)$ when we want to identify the random variable whose distribution is $F(x)$. Independence means that the joint distribution of $X_1, X_2, ..., X_n$ is equal to the product of the marginals

$$P\{X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n\} = F(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} F(x_j)$$

This implies that the expectation of the product of any bounded functions of the random variables equals the product of the expectations: $E\{g_1(X_1)g_2(X_2)\cdots g_n(X_n)\} = E\{g_1(X_1)\}E\{g_2(X_2)\}\cdots E\{g_n(X_n)\}$.

We will be interested in the behavior of the sample or empirical mean

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

which is the simplest and most widely studied function of the random variables. We expect that it should be closely related to the theoretical mean $\mu = E\{X_j\}$. We denote by σ^2 the variance of X_j

$$\sigma^2 = \text{var}(X_j) = E\{(X_j - \mu)^2\} = \int (x - \mu)^2 dF(x)$$

1.1 The weak law of large numbers

The simplest large sample relation is the weak law of large numbers (WLLN) which says that $\bar{X}_n \to \mu$ in probability as $n \to \infty$. This is a consequence of the Chebyshev inequality (CI)

$$P\{|\overline{X}_n - \mu| > \delta\} \leq \frac{E\{(\overline{X}_n - \mu)^2\}}{\delta^2} = \frac{\sigma^2}{n\delta^2} \to 0$$

as $n \to \infty$ for all $\delta > 0$. We have used here the fact that $\overline{X}_n$ converges to μ also in mean square, $E\{(\overline{X}_n - \mu)^2\} \to 0$.

We note that neither independence of the X_j nor their finite variance are needed for the validity of WLLN. It is enough that the X_j be sufficiently uncorrelated, with finite variance, or that they be independent with finite $E\{|X_j|\} < \infty$ but may have infinite variance.

The Chebyshev inequality for any random variable X has the form

$$P\{|X - \mu| > \delta\} \leq \frac{E\{|X - \mu|^p\}}{\delta^p} \ , \quad \delta > 0 \ , \quad 0 < p < \infty$$

1.2 The strong law of large numbers

We would like to know when we can say that $\overline{X}_n \to \mu$ with probability one as $n \to \infty$, not only that we have convergence in probability

$$\lim_{n\to\infty} P\{|\overline{X}_n - \mu| > \delta\} = 0 \ , \quad \text{for all } \delta > 0$$

We would like to know when it is true that

$$P\{\lim_{n\to\infty} \overline{X}_n = \mu\} = 1$$

which is the strong law of large numbers (SLLN).

This is more involved because we need to calculate the probability of an event that depends on infinitely many random variables. It is necessary therefore that the infinite sequence of independent identically distributed (iid) random variables $X_1, X_2, ..., X_n, ...$ be defined on the same probability space Ω, that is $X_j = X_j(\omega),\ \omega \in \Omega$ are real valued functions on Ω. We may think of Ω as a set of elementary events on which a probability law P is defined, more precisely it is defined on subsets of Ω. We will not need a measure theoretic foundation for probability here except in special cases in which we will deal with the issues that arise without a general theory.

Let $A_n = \{\omega \in \Omega \mid |\overline{X}_n - \mu| \leq \delta\}$ for some $\delta > 0$. If we can show that, given $\delta > 0$, the set of all ω such that $|\overline{X}_n - \mu| \leq \delta$ for all n larger than some $N(\omega)$ has probability one, then we have proved the SLLN. But the event in question can be written as

$$\bar{A} = \cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_n$$

With $\bar{A}^c$, the complement of $\bar{A}$, we now have

$$P\{\bar{A}^c\} = P\{\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} A_n^c\} = \lim_{N\to\infty} P\{\cup_{n=N}^{\infty} A_n^c\} \leq \lim_{N\to\infty} \sum_{n=N}^{\infty} P\{A_n^c\}$$

We have used here some general properties of probability laws that we will not discuss in detail but which are rather intuitive. One is that the probability of the intersection of a sequence of decreasing sets is equal to the limit of the probability of these sets, and the other is that the probability of the union of sets is less than or equal to the sum of the probabilities. It is equal when the events are disjoint, that is, their pairwise intersections are empty. Suppose now that we can show that for any $\delta > 0$ fixed

$$P\{A_n^c\} = P\{|\overline{X}_n - \mu| > \delta\} \leq \frac{\text{constant}}{n^2}$$

Then we have that $P\{\bar{A}^c\} = 0$, since

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} P\{A_n^c\} = 0,$$

and hence $P\{\bar{A}\} = 1$, which proves the SLLN.

If the iid random variables $\{X_j\}$ have finite fourth order moments, $E\{|X_j|^4\} < \infty$ or $E\{(X_j - \mu)^4\} < \infty$, then an application of the Chebyshev inequality with $p = 4$ gives the needed estimate and we have the SLLN in this case. Of course this is only a sufficient condition for its validity. As with the WLLN it enough that $E\{|X|\} < \infty$.

The strong law of large numbers played an important role in the development of probability theory because it provides in a precise way an empirical interpretation for the assignment of probabilities (and expected values). A large enough statistical sample will have a sample mean that tends to the theoretical mean as the sample size tends to infinity. This is the frequency or frequentist interpretation of probability theory.

1.3 Weak convergence

Weak convergence or convergence in distribution of random variables is important because it is associated with the limit theorems of probability theory of which the central limit theorem discussed in the next section is the oldest. In weak convergence the random variables, that is, their realization as functions on a probability space, is, in principle, unimportant. What counts is their distribution function, which means that we are interested in probabilities rather than the values the random variables take. And we want to allow for some smoothing. All this is captured in the definition of weak convergence which says that a sequence (any sequence) of random variable $\{X_j\}$ converges weakly (in law, in distribution) to X if

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

for any bounded and continuous function $g(x)$. If $F_n(x)$ and $F(x)$ are the distribution functions of these random variables, the definition of weak convergence can be written as

$$\lim_{n\to\infty} \int g(x) dF_n(x) = \int g(X) dF(x) \ , \quad \text{for all bounded and continous } g(x),$$

which shows that only distribution functions are involved. The continuity of g is essential as it allows for a "coarse graining" of the features of the random variables involved, and leads as a consequence to results that can have "universal" behavior. The central limit theorem is the best and perhaps simplest example of all this, as we will see in the next section.

A basic theorem in weak convergence is the equivalence of three different forms of it. The following three statements are equivalent:

1. $\lim_{n\to\infty} E\{g(X_n)\} = E\{g(X)\}$, for all bounded and continous $g(x)$
2. $\lim F_n(x) = F(x)$, at all continuity points x of $F(x)$
3. $\lim_{n\to\infty} E\{e^{i\alpha X_n}\} = E\{e^{i\alpha X}\}$, pointwise for all $\alpha \in R$

The last form of weak convergence involves the characteristic functions of the random variables, which are the Fourier transforms of the distribution functions. Implicit in this theorem is the statement that knowledge of the expectations $E\{g(X)\}$ for all bounded and continuous g determines F uniquely, and knowledge of the characteristic function $E\{e^{i\alpha X}\}$ also determines F uniquely. The latter is not so surprising as it amounts to the uniqueness of the inverse Fourier transform, which provides also a formula for recovering F from its Fourier transform.

As an example of how the above equivalence is shown consider how the second statement follows from the first. The indicator function $\mathbb{1}_{(-\infty,x]}(y)$ is discontinuous but can be bounded above and below by continuous functions, $g_\epsilon(y) \le \mathbb{1}_{(-\infty,x]}(y) \le g^\epsilon(y)$, where $g^\epsilon(y)$ equals one for $y < x$, is linear between x and $x+\epsilon$ going from 1 to zero, and is zero for $y > x+\epsilon$. The lower function is defined similarly, being equal to 1 for $y < x-\epsilon$, linear between $x-\epsilon$ and x and zero for $y > x$. From the first statement we have that

$$E\{g_\epsilon(X)\} = \lim_{n\to\infty} E\{g_\epsilon(X_n)\} \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le \lim_{n\to\infty} E\{g^\epsilon(X_n)\} = E\{g^\epsilon(X)\}$$

From the definition of g_ϵ and g^ϵ it follows that

$$F(x-\epsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\epsilon)$$

If x is a continuity point of F then letting $\epsilon \to 0$ gives the second statement.

1.4 The central limit theorem (CLT)

The CLT states that if $X_1, X_2, ..., X_n$ are iid random variables with mean μ and variance σ^2 then $Z_n = \sqrt{n}(\overline{X}_n - \mu)$ converges weakly to a Gaussian random variable with mean zero and variance σ^2.

The proof uses characteristic functions. We also assume that we have finite third moments, $E\{|X|^3\} < \infty\}$, which is not necessary except to simplify the proof. We have that

$$E\{e^{i\alpha Z_n}\} = E\{e^{\frac{i\alpha}{\sqrt{n}}\sum_{j=1}^n (X_j-\mu)}\} = E\{\prod_{j=1}^n e^{\frac{i\alpha}{\sqrt{n}}(X_j-\mu)}\}$$

$$= \prod_{j=1}^n E\{e^{\frac{i\alpha}{\sqrt{n}}(X_j-\mu)}\} = (E\{e^{\frac{i\alpha}{\sqrt{n}}(X-\mu)}\})^n$$

The independence is used in going from the first to the second line above.

We now use the Taylor expansion with remainder for the exponential

$$|e^{ix} - 1 - ix - \frac{1}{2}(ix)^2| \leq \frac{1}{6}|x|^3$$

and note that derivatives of the characteristic function at zero exist and equal to moments if these are finite. This gives

$$(E\{e^{\frac{i\alpha}{\sqrt{n}}(X-\mu)}\})^n = (1 - \frac{\alpha^2\sigma^2}{2n} + O(\frac{1}{n^{3/2}}))^n \to e^{-\frac{\alpha^2\sigma^2}{2}}$$

as $n \to \infty$. But $e^{-\frac{\alpha^2\sigma^2}{2}}$ is the characteristic function of a Gaussian random variable with mean zero and variance σ^2, which is the well-known identnity

$$e^{-\frac{\alpha^2\sigma^2}{2}} = \int_{-\infty}^{\infty} e^{i\alpha x} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx \tag{1}$$

A proof of the CLT as above but with only second moments for the law of the $\{X_j\}$ can be given by noting that if $\phi(\alpha) = E\{e^{i\alpha(X-\mu)}\}$ then $\phi(0) = 1$, $\phi'(0) = 0$, $\phi''(0) = -\sigma^2$, and we have by Taylor's theorem with remainder

$$\phi^n(\frac{\alpha}{\sqrt{n}}) = \left(1 + \phi''(\alpha_n)\frac{\alpha^2}{2n}\right)^n \to e^{\frac{-\alpha^2\sigma^2}{2}}$$

where $0 < \alpha_n < \frac{\alpha}{\sqrt{n}}$ so that $\phi''(\alpha_n) \to \phi''(0) = -\sigma^2$, which implies the result.

The reason that the CLT plays such an important role in probability is that it is perhaps the simplest limit theorem in which the limit law is "universal", a Gaussian, regardless of what the distribution of the random variable X is so long as its mean and variance are given.

1.5 Characteristic functions and Fourier transforms

Let $\phi(\alpha) = E\{e^{i\alpha X}\}$ be the characteristic function of a random variable X. It is always defined when $\alpha \in \mathbb{R}$ and $\phi(0) = 1$, $|\phi(\alpha)| \leq 1$. The characteristic function of a Gaussian RV with mean zero and variance σ^2 is given by (1). Note that this characteristic function decays rapidly as $|\alpha| \to \infty$, and that the Gaussian density can be recovered with the inverse Fourier transform:

$$\phi(\alpha) = \int_{-\infty}^{\infty} e^{i\alpha x} dF(x), \quad F'(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha x} \phi(\alpha) d\alpha.$$

Here the density $F'(x)$ is a Gaussian and is, of course, continuous. These formulas hold when the density F' is, for example, a smooth function (has derivatives of all orders). They are the formulas for the Fourier transform and its inverse.

The distribution function F can be recovered uniquely from its characteristic function $\phi(x)$ in general, whether it has a density or not. This is done with a smoothing or regularization method that is of general interest so we will carry out here. Let $p_\sigma(x)$ be the Gaussian density with mean zero and variance σ^2, which is the density on the right side of (1). If Z is a Gaussian random variable with mean zero and variance one, and independent of X, then

$$\phi_\sigma(\alpha) = E\{e^{i\alpha(X+\sigma Z)}\} = \int_{-\infty}^{\infty} e^{i\alpha x} dF_\sigma(x) = e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha)$$

is the characteristic function of $X + \sigma Z$ with F_σ the associated distribution function. Adding a small amount of Gaussian noise to any random variable makes its characteristic function decay rapidly at infinity, which means that the distribution function F_σ not only has a density but the density is itself differentiable any number of times. Therefore we have that

$$\begin{aligned} F'_\sigma(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha x} \phi_\sigma(\alpha) d\alpha &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha x} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha x} e^{-\frac{\alpha^2\sigma^2}{2}} \int_{-\infty}^{\infty} e^{i\alpha y} dF(y) d\alpha \end{aligned}$$

Because of the Gaussian factor we can interchange the α and y integrations in the last integral and recognize the inverse Fourier transform of a Gaussian. We thus have

$$F'_\sigma(x) = \int_{-\infty}^{\infty} p_\sigma(x-y) dF(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha x} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha$$

Integrating we get that

$$F_\sigma(x) = \int_{-\infty}^{\infty} N_\sigma(x-y) dF(y) = \frac{1}{2\pi} \int_{-\infty}^{x} \int_{-\infty}^{\infty} e^{-i\alpha y} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha dy$$

where $N_\sigma(x)$ is the Gaussian distribution function

$$N_\sigma(x) = \int_{-\infty}^{x} p_\sigma(y)dy$$

From this expression for F_σ we can conclude immediately that if x is a point at which $F(x)$ is continuous then, since $N_\sigma(x)$ tends to zero or one, as $\sigma \to 0$, depending on x being negative or positive respectively, we have

$$F(x) = \lim_{\sigma\to 0} F_\sigma(x) = \lim_{\sigma\to 0} \frac{1}{2\pi} \int_{-\infty}^{x} \int_{-\infty}^{\infty} e^{-i\alpha y} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha dy$$

This limit expression on the right shows that at its continuity points the distribution function is recovered uniquely by the characteristic function.

At points of discontinuity, a closer inspection of how $N_\sigma(x)$ behaves near $x = 0$ as $\sigma \to 0$ shows that

$$\frac{F(x) + F(x-)}{2} = \lim_{\sigma\to 0} F_\sigma(x) = \lim_{\sigma\to 0} \frac{1}{2\pi} \int_{-\infty}^{x} \int_{-\infty}^{\infty} e^{-i\alpha y} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha dy$$

We recall here that $F(x)$ is by its definition continuous from the right and has a left hand limit, denoted here by $F(x-)$.

To recover the jump at points of discontinuity, assume that F has an isolated discontinuity at x. Then, omitting details, we have

$$F(x)-F(x-) = \lim_{\epsilon\to 0} \lim_{\sigma\to 0} (F_\sigma(x+\epsilon)-F_\sigma(x-\epsilon)) = \lim_{\epsilon\to 0} \lim_{\sigma\to 0} \frac{1}{2\pi} \int_{x-\epsilon}^{x+\epsilon} \int_{-\infty}^{\infty} e^{-i\alpha y} e^{-\frac{\alpha^2\sigma^2}{2}} \phi(\alpha) d\alpha dy.$$

Note the order in which the limits are taken is important and cannot be interchanged.

The point of this section is that characteristic functions carry all the information that probability distributions have. If we can get some property or some fact using characteristic functions, as we did in the CLT, then this property holds in all generality.

1.6 On convergence of random variables

We have already encountered several types of convergence of random variables which we now review and compare. More mathematical details are given in the last section of this chapter. In this section we let $X_1, X_2, \ldots$ be an arbitrary sequence of random variables and let X_∞ be another random variable.

We say that $\{X_n\}$ converges in probability to X_∞, $X_n \stackrel{P}{\to} X_\infty$, if for all $\delta > 0$, $P\{|X_n - X_\infty| > \delta\} \to 0$ as $n \to \infty$.

We say that $\{X_n\}$ converges in mean square to X_∞, $X_n \stackrel{MSQ}{\to} X_\infty$, if $E\{(X_n - X_\infty)^2\} \to 0$ as $n \to \infty$. Clearly convergence in mean square implies convergence in probability. This follows from the Chebyshev inequality. However, the converse is not true simply because

random variables can converge in probability and they may not even have finite second moments (variances), so mean square convergence does not make sense for them.

Convergence with probability one, as explained above, means that all random variables are defined on the same probability space and that $P\{\lim_{n\to\infty} X_n = X_\infty\} = 1$. This is a composite of three statements: First the limit exists, then the limit is equal to X_∞ and third this occurs with probability one. Convergence with probability one implies convergence in probability (you need the bounded convergence theorem to show this) but the converse is false. This is intuitively clear but somewhat technical to justify as are most statements that hold with probability one. The connection with mean square convergence is this: Convergence with probability one and boundedness of the variances of the X_n uniformly in n does not imply mean square convergence but it does imply that the limit random variable has finite variance. And mean square convergence does not imply convergence with probability one.

Regarding weak convergence, it is true that convergence in probability implies it. The converse is not true except when the weak limit of the random variables (distributions) X_n is deterministic, that is, when X_∞ takes only one value and its distribution has a jump of hight one at that one point. In that case weak convergence implies convergence in probability. Slutsky's theorem is a generalization of this statement that we use often in estimation theory and elsewhere. Its statement is that if $X_n \xrightarrow{\mathcal{L}} X_\infty$ and $Y_n \xrightarrow{P} c$, where c is a (deterministic) constant, then $X_n + Y_n \xrightarrow{\mathcal{L}} X_\infty + c$, $X_n Y_n \xrightarrow{\mathcal{L}} X_\infty c$ and if $c > 0$ then $X_n/Y_n \xrightarrow{\mathcal{L}} X_\infty/c$.

As an example of how these statements are shown, consider how convergence in probability implies weak convergence. For weak convergence it is sufficient that the test function g be bounded and uniformly continuous (bounded and continuous is enough) so that given and $\epsilon > 0$ there is a $\delta > 0$ so that $|g(x) - g(y)| < \epsilon$ if $|x - y| < \delta$. We then have that

$$\begin{aligned}
&|E\{g(X_n)\} - E\{g(X_\infty)\}| \\
&\le |E\{[g(X_n) - g(X_\infty)]\mathbb{1}_{(|X_n - X_\infty| \le \delta)}\}| + E\{|g(X_n) - g(X_\infty)|\mathbb{1}_{(|X_n - X_\infty| > \delta)}\} \\
&\le \epsilon(1 - P\{|X_n - X_\infty| > \delta\}) + 2\max_x |g(x)| P\{|X_n - X_\infty| > \delta\}
\end{aligned}$$

This implies that

$$\limsup_{n\to\infty} |E\{g(X_n)\} - E\{g(X_\infty)\}| \le \epsilon$$

and since ϵ is arbitrary we have the statement of weak convergence.

1.7 Confidence intervals for the empirical mean

An important application of the central limit theorem along with Slutsky's theorem is in getting confidence intervals in parameter estimation. Suppose that we use the empirical mean $\overline{X}_n$ to estimate the true mean μ, assumed unknown. First we introduce the sample

variance

$$s_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

It is easily seen that $E\{s_n^2\} = \sigma^2$, the theoretical variance, and assuming finite third moments we not only have that $\overline{X}_n \xrightarrow{P} \mu$ but also $s_n^2 \xrightarrow{P} \sigma^2$, which implies that $s_n \xrightarrow{P} \sigma$. Now the central limit theorem says that $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. By Slutsky's theorem we also have that $\frac{\sqrt{n}}{s_n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Given a confidence level α, for example $\alpha = .05$, we find ζ_α such that $P\{|Z| > \zeta_\alpha\} = \alpha$, where Z is a Gaussian random variable with mean zero and variance one. It then follows that for n large enough we have that $\frac{\sqrt{n}}{s_n}|\overline{X}_n - \mu| > \zeta_\alpha$ with probability α and hence

$$\overline{X}_n - \frac{s_n\zeta_\alpha}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{s_n\zeta_\alpha}{\sqrt{n}} \tag{2}$$

with probability $1-\alpha$. This is a confidence interval for the unknown mean μ in terms of a sample of size n, which we assume is large enough so that the central limit theorem can be used.

1.8 Large deviations

The weak law of large numbers states that $\overline{X}_n \xrightarrow{P} \mu$ and the central limit theorem that $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2)$. Let $\gamma > \mu$ and note that we must have

$$P\{\overline{X}_n > \gamma\} \to 0.$$

The question posed in large deviations is to estimate the rate at which this probability tends to zero. Assume that the underlying random variable X has a density $f(x)$ and a moment generating function. We denote the logarithm of the moment generating function by

$$L(\alpha) = \log E\{e^{\alpha X}\},\ \alpha \in \mathbb{R}$$

which we assume to be finite and differentiable in α. We have that $L(0) = 0$, $L'(0) = \mu$ and $L''(0) = \sigma^2$. We also assume that $L(\alpha)$, which is always convex, is in fact strictly convex.

The convexity is shown directly by noting that $L(\alpha)$ is differentiable for $\alpha \in \mathbb{R}$, and in any case in an open set if the range of α is bounded. We have

$$L'(\alpha) = \frac{E\{Xe^{\alpha X}\}}{E\{e^{\alpha X}\}},\quad L''(\alpha) = \frac{E\{X^2e^{\alpha X}\}}{E\{e^{\alpha X}\}} - \frac{(E\{Xe^{\alpha X}\})^2}{(E\{e^{\alpha X}\})^2} \geq 0$$

The inequality on the right comes from the Schwatz inequality:

$$|E\{Xe^{\alpha X}\}| \leq \sqrt{E\{X^2e^{\alpha X}\}E\{e^{\alpha X}\}}$$

We define the conjugate convex function of L by

$$H(\gamma) = \sup_{\alpha}(\alpha\gamma - L(\alpha)),\ \gamma \in \mathbb{R}$$

and note that L can be recovered from H by applying to the Legendre transform again

$$L(\alpha) = \sup_{\gamma}(\alpha\gamma - H(\gamma)),\ \alpha \in \mathbb{R}$$

We will show that

$$\lim_{n\to\infty} \frac{1}{n} \log P\{\overline{X}_n > \gamma\} = -H(\gamma),$$

which means that in the logarithmic sense we have

$$P\{\overline{X}_n > \gamma\} \approx e^{-nH(\gamma)}$$

For the proof we get first the upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\overline{X}_n > \gamma\} \leq -H(\gamma), \tag{3}$$

and then the lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{\overline{X}_n > \gamma\} \geq -H(\gamma), \tag{4}$$

from which the result follows. For the upper bound we note that

$$P\{\overline{X}_n > \gamma\} = P\{e^{\alpha n\overline{X}_n} > e^{\alpha n\gamma}\} \leq e^{-\alpha n\gamma}(E\{e^{\alpha X}\})^n = e^{-n(\alpha\gamma - L(\alpha))},\ \alpha \geq 0,$$

which after taking logs, dividing by n, taking the upper limit and then optimizing the right side over α we get the upper bound (3). If $\gamma > \mu$ then $H(\gamma) = \sup_{\alpha>0}(\alpha\gamma - L(\gamma))$ so the Markov inequality above is valid.

To get the lower bound we first introduce a transformation of law for X that plays an essential role here and in some other contexts such as in importance sampling. For any $\alpha \in \mathbb{R}$ let

$$f_\alpha(x) = \frac{e^{\alpha x} f(x)}{\int_{-\infty}^{\infty} e^{\alpha x} f(x)dx}$$

and note that we have the identity

$$f(x_1)f(x_2)\cdots f(x_n) = e^{-\alpha(x_1+x_2+\cdots+x_n)+nL(\alpha)} f_\alpha(x_1)f_\alpha(x_2)\cdots f_\alpha(x_n)$$

This implies that for any α real we have

$$P\{\overline{X}_n > \gamma\} = E\{\mathbb{1}_{(\overline{X}_n>\gamma)}\} = E_\alpha\{e^{-n(\alpha\overline{X}_n - L(\alpha))}\mathbb{1}_{(\overline{X}_n>\gamma)}\}$$

where E_α denotes expectation relative to the law with density f_α. Let ϵ be any small positive number and note that $\{\overline{X}_n > \gamma\} = \{\overline{X}_n - (\gamma+\epsilon) > -\epsilon\} \supset \{|\overline{X}_n - (\gamma+\epsilon)| < \epsilon\}$. If we chose $\alpha = \alpha^*(\gamma+\epsilon)$ so that $L'(\alpha^*) = \gamma+\epsilon$, which means that $E_{\alpha^*}\{X\} = \gamma+\epsilon$, the weak law of large numbers gives

$$P_{\alpha^*}\{|\overline{X}_n - (\gamma+\epsilon)| < \epsilon\} \to 1, \text{ as } n \to \infty$$

We now note that

$$E_{\alpha^*}\{e^{-n(\alpha^*\overline{X}_n - L(\alpha^*))}\mathbb{1}_{(\overline{X}_n > \gamma)}\} \geq e^{-n((\gamma+2\epsilon)\alpha^* - L(\alpha^*))}P_{\alpha^*}\{|\overline{X}_n - (\gamma+\epsilon)| < \epsilon\}$$

Taking logarithms, dividing by n and taking the lower limit as $n \to \infty$ gives

$$\liminf_{n\to\infty} \frac{1}{n}\log P\{\overline{X}_n > \gamma\} \geq -[(\gamma+2\epsilon)\alpha^*(\gamma+\epsilon) - L(\alpha^*(\gamma+\epsilon))]$$

Since ϵ is arbitrarily small we have

$$\liminf_{n\to\infty} \frac{1}{n}\log P\{\overline{X}_n > \gamma\} \geq -[\gamma\alpha^*(\gamma) - L(\alpha^*(\gamma))] = -H(\gamma)$$

which proves the lower bound and hence the large deviations theorem.

1.9 More on convergence of random variables

In this section we cover some useful convergence results and the mathematical level is more advanced. We have already discussed weak convergence is section 1.3 and its relation to convergence in probability in section 1.6 Here we consider in particular implications among various types of convergence. We begin with the following definitions, some of them in review.

If $\{X_n\}_{i=1}^\infty$ are a sequence of random variables.

- We say X_n converges to X_∞ in probability, denoted as $X_n \xrightarrow{P} X_\infty$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_\infty| > \epsilon) = 0.$$

- We say X_n converges to X_∞ in probability 1 (almost surely), denoted as $X_n \xrightarrow{a.s.} X_\infty$, if

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)\}) = 1.$$

  Here Ω is the underlying probability space on which all the random variables are defined. That is, there is a set Ω, a probability space, and the random variables are real valued functions defined on this set, $X_n(\omega)$. There is also a family $\mathcal{F}$ of subsets of Ω that is closed under complements, countable unions and intersections. In addition, the set of all $\omega \in \Omega$ such that $X_n(\omega) \leq x$ belongs to $\mathcal{F}$ for all real x (which says X_n is measurable). There is also a probability measure P on $(\Omega, \mathcal{F})$ such that $F_{X_n}(x) = P\{\omega \in \Omega \ : \ X_n(\omega) \leq x\}$.

- We say X_n converges to X_∞ in the mean of order p, L^p, $p \geq 1$, denoted as $X_n \xrightarrow{L^p} X_\infty$, if
$$\lim_{n\to\infty} \mathbb{E}|X_n - X_\infty|^p = 0.$$

- We say X_n converges to X_∞ in distribution (or weakly), denoted as $X_n \xrightarrow{\mathcal{D}} X_\infty$, if
$$\lim_{n\to\infty} F_n(x) = F_\infty(x)$$
for any continuous points x of F_∞. Here $F_n(\cdot), 1 \leq n \leq \infty$ is the cdf of X_n. In this case, we also say $F_n(\cdot)$ converges to $F_\infty(\cdot)$ weakly, denoted as $F_n \xrightarrow{w} F_\infty$.

We first list some facts about convergence, a few already introduced and discussed earlier in this chapter, in the form of problems to be solved, and then provide the proofs. There is some overlap with previous sections here.

1. Show that for $p \geq 1$, $X_n \xrightarrow{L^p} X_\infty$ implies $X_n \xrightarrow{P} X_\infty$. And give an example to show the converse statement doesn't hold.

2. Show that $X_n \xrightarrow{a.s.} X_\infty$ implies $X_n \xrightarrow{P} X_\infty$. Use the bounded convergence theorem which says that is if $|g(x)|$ is bounded uniformly for all real x and $g(X_n) \xrightarrow{a.s.} g(X_\infty)$ then $\mathbb{E}g(X_n) \longrightarrow \mathbb{E}g(X_\infty)$.

3. Show that $X_n \xrightarrow{L^1} X_\infty$ and X_n and X_∞ are integrable implies $\mathbb{E}X_n \longrightarrow \mathbb{E}X_\infty$.

4. Show that if $X_n \xrightarrow{P} X_\infty$ then $X_n \xrightarrow{\mathcal{D}} X_\infty$. Conversely, if $X_n \xrightarrow{\mathcal{D}} X_\infty$ and X_∞ is a non-random constant, then $X_n \xrightarrow{P} X_\infty$.

5. Show that if $X_n \xrightarrow{P} X_\infty$ and $f(\cdot)$ is a continuous function, then $f(X_n) \xrightarrow{P} f(X_\infty)$.

6. Show that X_n converges to X_∞ in distribution if and only if $E\{g(X_n)\}$ converges to $E\{g(X_\infty)\}$ for every bounded and continuous function g.

7. Show that if $X_n \xrightarrow{\mathcal{D}} X_\infty$ and $Y_n \xrightarrow{\mathcal{D}} c$ where c is a constant, then $(X_n, Y_n) \xrightarrow{\mathcal{D}} (X_\infty, c)$. More specifically, prove that for
$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x, Y_n \leq y) = \mathbb{P}(X_\infty \leq x)\mathbb{1}(c \leq y)$$
for any continouity points of $F_\infty(x, y) := \mathbb{P}(X_\infty \leq x)\mathbb{1}(c \leq y)$.

We now provide the proofs in the form of solutions to the problems stated.

1. Chebyshev inequality:

$$\mathbb{P}(|X_n - X_\infty| > \epsilon) = \mathbb{P}(|X_n - X_\infty|^p > \epsilon^p) \leq \frac{\mathbb{E}|X_n - X_\infty|^p}{\epsilon^p} \to 0 \ (n \to \infty)$$

   The converse does not hold, for example, define $X_n = n^{1/p}$ with probability $1/n$ and $X_n = 0$ with probability $1 - 1/n$, and $X_\infty = 0$ with probability 1, then $X_n \xrightarrow{P} X_\infty$ but X_n does not converge to X_∞ in L^p.

2. Almost sure convergence can be expressed in terms of the set: $\{\omega : \lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)\} = \bigcap_{\epsilon>0} \bigcup_{N\geq 1} \bigcap_{n\geq N} \{\omega : |X_n(\omega) - X_\infty(\omega)| < \epsilon\}$. Therefore if $X_n \xrightarrow{a.s.} X_\infty$, then $\forall \epsilon > 0$,

$$\mathbb{P}(\bigcup_{N\geq 1} \bigcap_{n\geq N} \{\omega : |X_n(\omega) - X_\infty(\omega)| < \epsilon\}) = 1$$

   Set $Y_n = X_n - X_\infty$. Then $Y_n \xrightarrow{a.s.} 0$ by assumption. Define $g(x) = \mathrm{I}_{|x|>\epsilon}$ which is uniformly bounded for all x. By the bounded convergence theorem,
   $\mathbb{E}g(Y_n) \to \mathbb{E}g(0) = 0$, that is, $\lim_n \mathbb{P}(\{\omega : |X_n(\omega) - X_\infty(\omega)| > \epsilon\}) = 0$

3. $|\mathbb{E}X_n - \mathbb{E}X_\infty| \leq \mathbb{E}|X_n - X_\infty| \to 0 \quad (n \to \infty)$

4. Denote F_n as the cdf for $X_n, 1 \leq n \leq \infty$. Then for any continuity point x of $F_\infty(x)$, we have

$$F_n(x) = \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X_\infty \leq x + \epsilon) + \mathbb{P}(|X_n - X_\infty| \geq \epsilon).$$

   Let $n \to \infty$ and then let $\epsilon \to 0$, we have $\limsup_{n\to\infty} F_n(x) \leq F_\infty(x)$. Similarly,

$$F_n(x) = \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X_\infty \leq x - \epsilon) - \mathbb{P}(|X_n - X_\infty| \geq \epsilon).$$

   Let $n \to \infty$ and then let $\epsilon \to 0$ we have $\liminf_{n\to\infty} F_n(x) \geq F_\infty(x)$. Hence $\lim_{n\to\infty}$ exists and equals to $F_\infty(x)$.

   On the other hand, if $X_n \xrightarrow{\mathcal{D}} X_\infty \equiv c$,

$$\mathbb{P}(|X_n - c| > \delta) = \mathbb{P}(X_n > c+\delta) + \mathbb{P}(X_n < c-\delta) = 1 - F_n(c+\delta) + F_n((c-\delta)-) \leq 1 - F_n(c+\delta) + F_n(c-\delta).$$

   Here $F_n((c-\delta)-)$ means $\lim_{\epsilon\to 0+} F_n(c-\delta-\epsilon)$. Taking $c+\delta, c-\delta$ as two continuity points of $F(x)$ with $F(c-\delta) = 0, F(c+\delta) = 1$, the limit yields

$$\limsup_{n\to\infty} \mathbb{P}(|X_n - c| > \delta) \leq 1 - 1 + 0 = 0.$$

   Hence $X_n \to X_\infty$ in probability, where $X_\infty \equiv c$.

5. For each $\epsilon > 0$, define

$$A_\delta = \{x : \exists y, \text{ s.t. } |x - y| < \delta, |f(x) - f(y)| > \epsilon\}.$$

Since $f(\cdot)$ is continuous everywhere, thus $A_\delta \to \emptyset$ as $\delta \to 0$.

Now $|f(X_n) - f(X_\infty)| > \epsilon$ implies that $|X_n - X_\infty| > \delta$ for some δ, or $X_\infty \in A_\delta$. Thus

$$\mathbb{P}(|f(X_n) - f(X_\infty)| > \epsilon) \leq \mathbb{P}(|X_n - X_\infty| > \delta) + \mathbb{P}(X_\infty \in A_\delta).$$

First take $n \to \infty$ we have

$$\limsup_{n\to\infty} \mathbb{P}(|f(X_n) - f(X_\infty)| > \epsilon) \leq \mathbb{P}(X_\infty \in A_\delta).$$

Take $\delta \to 0$ proves the problem.

6. We show that if X_n converges to X_∞ in distribution then $E\{g(X_n)\}$ converges to $E\{g(X_\infty)\}$ for every bounded and continuous function g. The converse is shown in section 1.3.

Given $\epsilon > 0$, choose points $x_0 < x_1 < \cdots < x_N$ such that (a) they are continuity points of F_∞, (b) $F_\infty(x_0) < \epsilon$, $(1 - F_\infty(x_N)) < \epsilon$, and (c) the given continuous function $g(x)$ can be approximated by a step function $g_N(x)$ in the interval $x_0 \leq x \leq x_N$ such that $\max_{x_0 \leq x \leq x_N} |g(x) - g_N(x)| < \epsilon$. Now we have that

$$|\int_{-\infty}^{\infty} g(x)dF_n(x) - \int_{-\infty}^{\infty} g(x)dF_\infty(x)| \tag{5}$$

$$\leq |\int_{-\infty}^{x_0} g(x)dF_n(x) - \int_{-\infty}^{x_0} g(x)dF_\infty(x)| \tag{6}$$

$$+|\int_{x_0}^{x_N} g(x)dF_n(x) - \int_{x_0}^{x_N} g(x)dF_\infty(x)| \tag{7}$$

$$+|\int_{x_N}^{\infty} g(x)dF_n(x) - \int_{x_N}^{\infty} g(x)dF_\infty(x)| \tag{8}$$

The first and third terms on the right are less than a constant times ϵ as $n \to \infty$ because of (a),(b) and the fact that g is bounded. In the middle term, we add and subtract g_N to get

$$|\int_{x_0}^{x_N} g(x)dF_n(x) - \int_{x_0}^{x_N} g(x)dF_\infty(x)| \tag{9}$$

$$\leq |\int_{x_0}^{x_N} g(x)dF_n(x) - \int_{x_0}^{x_N} g_N(x)dF_n(x)| \tag{10}$$

$$+|\int_{x_0}^{x_N} g_N(x)dF_n(x) - \int_{x_0}^{x_N} g_N(x)dF_\infty(x)| \tag{11}$$

$$+|\int_{x_0}^{x_N} g_N(x)dF_\infty(x) - \int_{x_0}^{x_N} g(x)dF_\infty(x)| \tag{12}$$

The first and third terms on the right side are less than a constant times ϵ because of (c). The middle term is now a finite sum of differences $F_n - F_\infty$ at continuity points of F_∞, so it goes to zero as $n \to \infty$ by the hypothesis.

7. For any continuity point x of $F_{X_\infty}(x)$ and $y < c$, we have

$$\mathbb{P}(X_n \leq x, Y_n \leq y) \leq \mathbb{P}(Y_n \leq y) \to 0 = \mathbb{P}(X_\infty \leq x, Y_\infty \leq y).$$

For any continuity point x of $F_{X_\infty}(x)$ and $y > c$, we have

$$\mathbb{P}(X_n \leq x, Y_n \leq y) = \mathbb{P}(X_n \leq x) - \mathbb{P}(X_n \leq x, Y_n > y).$$

As $\mathbb{P}(X_n \leq x, Y_n > y) \leq \mathbb{P}(Y_n > y) \to 0$, we obtain that $\mathbb{P}(X_n \leq x, Y_n \leq y) \to F_{X_\infty}(x) = \mathbb{P}(X_\infty \leq x, Y_\infty \leq y)$. This completes the proof.

2 Maximum likelihood estimation (MLE)

Let $X_1, X_2, ..., X_n$ be independent samples of a random variable whose distribution $F(x|\theta)$ depends on a real parameter θ with values in a bounded interval. How can we estimate the true value θ^* of this parameter from the observed sequence. In the maximum likelihood method we form the likelihood function, which is the joint density of the iid random variables $X_1, X_2, ..., X_n$ evaluated at the observed values

$$L_n(\theta) = L_n(\theta; X_1, X_2, ..., X_n) = \prod_{j=1}^{n} f(X_j|\theta). \tag{13}$$

Here $f(x|\theta) = \frac{d}{dx}F(x|\theta)$ is the density of the random variable, and we obtain an estimate of θ by maximizing it

$$\hat{\theta}_n = \text{argmax}_\theta L_n(\theta)$$

Clearly $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, ..., X_n)$ and the reason that this is an acceptable choice for an estimator is that the observed sample by the mere fact that it has occurred must have come from the distribution with the "most likely parameter". Of course, the reason that MLE estimators $\hat{\theta}_n$ are calculated and used is because they have very good large sample, $n \to \infty$, properties as we will see, which captures and articulates this intuition.

Assuming that the density $f(x|\theta) > 0$ is smooth in θ and that integrals we will be computing all exist, we introduce the scaled log likelihood function

$$l_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \log f(X_j|\theta) \tag{14}$$

and note that $l_n'(\hat{\theta}_n) = 0$, which is the usual first order condition for an extremum. Here primes denote derivatives with respect to θ and we have

$$l_n'(\theta) = \frac{1}{n}\sum_{j=1}^{n} \frac{f'(X_j|\theta)}{f(X_j|\theta)}$$

By the WLLN we have that

$$\lim_{n\to\infty} P_{\theta^*}\{|l_n(\theta) - l(\theta)| > \delta\} \to 0 \ , \quad \text{for all } \delta > 0$$

where

$$l(\theta) = E_{\theta^*}\{\log f(X|\theta)\} = \int \log(f(x|\theta)) f(x|\theta^*) dx.$$

Having assumed differentiability in θ we also have

$$\lim_{n\to\infty} P_{\theta^*}\{|l_n'(\theta) - l'(\theta)| > \delta\} \to 0 \ , \quad \text{for all } \delta > 0$$

where

$$l'(\theta) = E_{\theta^*}\{\frac{f'(X|\theta)}{f(X|\theta)}\} = \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta^*)dx.$$

Clearly we have that

$$l'(\theta^*) = \int \frac{f'(x|\theta^*)}{f(x|\theta^*)} f(x|\theta^*)dx = \int f'(x|\theta^*)dx = \frac{d}{d\theta}\int f(x|\theta)dx\Big|_{\theta=\theta^*} = 0$$

We also have that

$$l''(\theta^*) = -\int \frac{(f'(x|\theta^*)^2}{f(x|\theta^*)}dx < 0$$

which means that θ^* is, in general, a maximum of $l(\theta)$. We define the **Fisher information** by

$$I(\theta^*) = -l''(\theta^*) = \int \frac{(f'(x|\theta^*)^2}{f(x|\theta^*)}dx \tag{15}$$

and note that it is positive.

2.1 Large sample properties of MLE

An estimator is called asymptotically consistent if

$$\lim_{n\to\infty} P_{\theta^*}\{|\hat{\theta}_n - \theta^*| > \delta\} = 0 \ , \quad \text{for all } \delta > 0 \tag{16}$$

The fact that the MLE estimator is asymptotically consistent follows from the smoothness in θ of $f(x|\theta)$ and the existence the integrals that arise in the calculations above. This is not an obvious statement and is proved at the end of this chapter. It becomes more plausible when we invoke the uniform WLLN for $l_n(\theta)$ and $l_n'(\theta)$

$$\lim_{n\to\infty} P_{\theta^*}\{\max_{|\theta-\theta^*|\le a} |l_n(\theta) - l(\theta)| > \delta\} = 0 \ , \quad \text{for all } \delta > 0 \tag{17}$$

and similarly for l_n', where $a > 0$ is a fixed constant. This says that near θ^* the random curve $l_n(\theta)$ and its derivative $l_n'(\theta)$ are uniformly close to the deterministic curve $l(\theta)$ and its derivative, in probability. This then implies that the maximum of l_n, $\hat{\theta}_n$, is close to the maximum of l, which is θ^*, in probability.

An estimator is unbiased if $E_{\theta^*}\{\hat{\theta}_n\} = \theta^*$ and asymptotically unbiased if $\lim_{n\to\infty} E_{\theta^*}\{\hat{\theta}_n\} = \theta^*$. MLE estimators are in general biased but because of consistency they are, with additional assumptions on integrability, asymptotically unbiased.

An important application of the central limit theorem is in determining the fluctuations in the MLE. Let Z_n be the scaled error

$$\hat{\theta}_n = \theta^* + \frac{1}{\sqrt{n}} Z_n \ , \quad \text{or} \quad Z_n = \sqrt{n}(\hat{\theta}_n - \theta^*)$$

We will show that Z_n converges weakly to a Gaussian random variable with mean zero and variance equal to the reciprocal of the Fisher information I. It is not immediately clear how the CLT comes into the picture since $\hat{\theta}_n$ is not, in general, a sum of random variables. The way to represent it approximately as a ratio of sums of random variables is by using the delta method or expansion. We use a Taylor expansion with remainder (and the mean value theorem) to get

$$0 = l_n'(\hat{\theta}_n) = l_n'(\theta^*) + \frac{1}{\sqrt{n}} l_n''(\tilde{\theta}_n) Z_n$$

so that

$$Z_n = \frac{\sqrt{n} l_n'(\theta^*)}{-l_n''(\tilde{\theta}_n)}$$

where $\tilde{\theta}_n$ is between θ^* and $\hat{\theta}_n$ and tends to θ^* in probability, by consistency. Recall that

$$\sqrt{n} l_n'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{f'(X_j|\theta^*)}{f(X_j|\theta^*)}$$

and

$$E_{\theta^*}\{\sqrt{n} l_n'(\theta^*)\} = 0, \quad \mathrm{var}_{\theta^*}(\sqrt{n} l_n'(\theta^*)) = I(\theta^*)$$

.

We now note the following.

- The CLT can be applied directly to $\sqrt{n} l_n'(\theta^*)$ so as to show that it converges weakly or in distribution to a Gaussian random variable with mean zero and variance equal to the Fisher information $I(\theta^*)$.
- The uniform WLLN can be applied $-l_n''(\tilde{\theta}_n)$ to show that it converges in probability to the Fisher information[1]

Recall that

$$-l_n''(\theta) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{f'(X_j|\theta)}{f(X_j|\theta)} \right)^2 - \frac{1}{n} \sum_{j=1}^{n} \left(\frac{f''(X_j|\theta)}{f(X_j|\theta)} \right)$$

For θ replaced by $\tilde{\theta}_n \to \theta^*$, the first sum on the right tends by the uniform WLLN to the limit $I(\theta^*)$ and the second $E_{\theta^*}\{\frac{f''(X|\theta^*)}{f(X|\theta^*)}\}$, which is equal to zero, both in probability.

Therefore we have shown that $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges in distribution to a Gaussian random variable with mean zero and variance $1/I(\theta^*)$. From this result we can construct confidence intervals for the unknown parameter θ^* just as we did for the empirical mean

[1] T. S. Ferguson, "A Course in Large Sample Theory", CRC Press, 1996, page 122.

of sums of iid RVs. If ζ_α is defined as in (**??**) then we have that the unknown parameter θ^* satisfies

$$\hat{\theta}_n - \frac{\zeta_\alpha}{\sqrt{nI(\hat{\theta}_n)}} \leq \theta^* \leq \hat{\theta}_n + \frac{\zeta_\alpha}{\sqrt{nI(\hat{\theta}_n)}}$$

with probability $1-\alpha$ for n sufficiently large.

Slutsky's theorem says that if $U_n = V_n W_n + Y_n$ and (i) V_n converges weakly to V, (ii) W_n converges in probability to a constant w and (iii) Y_n converges in probability to zero, then U_n converges weakly to Vw.

2.2 Cramer-Rao lower bound and asymptotic efficiency of the MLE

We now want to show that the MLE is asymptotically efficient, which means that the variance of the limit fluctuation Z, which is the reciprocal of the Fisher information, is as small as it can be among all other consistent estimators. This is done by using the Cramer-Rao lower bound for the variance of estimators that we derive next.

For any random variables U, V we have by the Schwartz inequality

$$\mathrm{var}(U) \geq \frac{(\mathrm{cov}(U,V))^2}{\mathrm{var}(V)}$$

Let X be a vector-valued random variable with (multi-dimensional) density $f(x|\theta)$ depending on a parameter θ. Now apply the inequality with $U = w(X)$, a function of X, and $V = (\log f(X|\theta))'$. Since we have that $E_\theta\{V\} = 0$, we see that

$$\mathrm{cov}(U,V) = E_\theta\{w(X)(\log f(X|\theta))'\} = \frac{d}{d\theta} E_\theta\{w(X)\}$$

Therefore we have the inequality

$$\mathrm{var}_\theta(w(X)) \geq \frac{(\frac{d}{d\theta} E_\theta\{w(X)\})^2}{E_\theta\{(\frac{d}{d\theta}\log f(X|\theta))^2\}}$$

This is the Cramer-Rao inequality.

In the case that the random vector $\underline{X} = (X_1, X_2, \ldots, X_n)$ is a sample of size n from the one-dimensional density $f(x|\theta)$ we see easily that we have

$$\mathrm{var}_\theta(w(\underline{X})) \geq \frac{(\frac{d}{d\theta} E_\theta\{w(\underline{X})\})^2}{nE_\theta\{(\frac{d}{d\theta}\log f(X|\theta))^2\}}$$

This inequality is valid for any θ and any estimator $w(\underline{X})$.

We now apply this to the maximum likelihood estimator $\hat{\theta}_n$ of θ^*, that is, we let $w(\underline{X}) = \hat{\theta}_n(\underline{X})$. We then have

$$n\mathrm{var}_{\theta^*}(\hat{\theta}_n(\underline{X})) \geq \frac{(\frac{d}{d\theta} E_\theta\{\hat{\theta}_n(\underline{X})\}|_{\theta=\theta^*})^2}{I(\theta^*)}$$

where

$$I(\theta^*) = E_\theta\{(\frac{d}{d\theta}\log f(X|\theta))^2\}|_{\theta=\theta^*}$$

is the Fisher information. The numerator on the right involves the bias of the MLE. With enough assumptions we can show that for n large we have that

$$\frac{d}{d\theta}E_\theta\{\hat{\theta}_n(\underline{X})\}|_{\theta=\theta^*} \sim 1$$

so that we have the asymptotic inequality for large n:

$$\mathrm{var}_{\theta^*}(\sqrt{n}\hat{\theta}_n(\underline{X})) \geq \frac{1}{I(\theta^*)}. \tag{18}$$

We know that $\hat{\theta}_n \to \theta^*$ in probability and $\sqrt{n}(\hat{\theta}_n - \theta^*)$ tends in distribution to a Gaussian random variable with mean zero and variance one over the Fischer information. With mild integrability assumptions this implies that the MLE is asymptotically unbiased. It is biased in general. With further assumptions the derivative of the bias $(E_{\theta^*}\{\hat{\theta}_n\} - \theta^*)$ with respect to θ^* also tends to zero as $n \to \infty$. This shows that the MLE has the smallest asymptotic variance among all possible consistent and asymptotically unbiased (in the stronger sense involving the derivative) estimators.

As an example of how the Cramer-Rao inequality can be used, consider a sample $X_1, X_2, \ldots, X_n$ from the exponential density $f(x|\theta) = \theta e^{-\theta x},\ x > 0$, with parameter $\theta^* > 0$. We find easily from the log likelihood function that the MLE of θ is $\hat{\theta}_n = 1/\bar{X}_n$ where $\bar{X}_n$ is the sample mean. The MLE is biased because $E_{\theta^*}\{\hat{\theta}_n\} = \frac{n}{n-1}\theta^*$, but asymptotically unbiased and the derivative tends to one so that (18) can be used asymptotically for large n. The Fisher information is here $I = (\theta^*)^{-2}$. To see that $E_{\theta^*}\{\hat{\theta}_n\} = \frac{n}{n-1}\theta^*$, we note that

$$\begin{aligned} E_{\theta^*}\{\frac{1}{\bar{X}_n}\} &= \int_0^\infty \cdots \int_0^\infty \frac{n}{x_1 + x_2 + \cdots + x_n}(\theta^*)^n e^{-\theta^*(x_1+x_2\cdots+x_n)}dx_1\cdots dx_n \\ &= \theta^* \int_0^\infty \cdots \int_0^\infty \frac{n}{x_1 + x_2 + \cdots + x_n} e^{-(x_1+x_2\cdots+x_n)}dx_1\cdots dx_n \\ &= a_n\theta^* \end{aligned}$$

The constant a_n is evaluated by differentiating the first line with respect to θ^* and getting a differential equation for $E_{\theta^*}\{\frac{1}{\bar{X}_n}\}$, which has solution $a_n\theta^*$ with $a_n = \frac{n}{n-1}$.

2.3 Asymptotic normality of posterior densities

Instead of considering θ as a parameter upon which the density $f(x|\theta)$ depends, we may think of it as a random variable with a prior density $g(\theta)$. Once the sample $X_1, X_2, \ldots, X_n$ from $f(x|\theta^*)$ has been observed, the posterior density of θ given the sample is defined by

$$\pi_n(\theta) = \frac{L_n(\theta)g(\theta)}{\int L_n(\theta)g(\theta)d\theta} \tag{19}$$

where the likelihood function $L_n(\theta)$ is defined by (13).

We have used Bayes theorem for densities: The aposteriori density of the parameter(s) given the sample, $\pi_n(\theta)$, is equal to the density of the sample given the parameter(s), the likelihood $L_n(\theta)$, times the apriori density of the parameter, $g(\theta)$, divided by the marginal density of the sample. For events, Bayes theorem is the definition of conditional probability, essentially,

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}} = \frac{P\{B|A\}P\{A\}}{P\{B\}}$$

We will show that if we look at the posterior density in a neighborhood on the MLE $\hat{\theta}_n$ that is of order $1/\sqrt{n}$ it tends to a limit[2] that is a Gaussian density with mean zero and variance equal to the reciprocal of the Fisher information $I(\theta^*)$. The actual theorem states that if we let $\theta = \hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}$ and change variables in the posterior $\pi_n(\theta)$ so that it becomes $\tilde{\pi}_n(\tilde{\theta})$ then

$$\tilde{\pi}_n(\tilde{\theta}) \to \frac{e^{-\frac{I(\theta^*)\tilde{\theta}^2}{2}}}{\sqrt{2\pi(1/I(\theta^*))}} \tag{20}$$

in probability and for each $\tilde{\theta}$. Some hypotheses about the prior $g(\theta)$ are needed so that the limit can be taken inside the integral in the denominator in (19). We know by (16) that the MLE is consistent, $\hat{\theta}_n \to \theta^*$ in probability, so the change of variables is essentially centered about the true parameter value θ^*. However, the convergence of the posterior density is not valid if we do not center around $\hat{\theta}_n$. Note also that the limit posterior is independent of the prior density $g(\theta)$, assuming that the latter is positive for all θ as well as such that the limit can be taken inside the integral in (19).

For the proof we note that by (14), $L_n(\theta) = e^{nl_n(\theta)}$ and we have the expansion

$$l_n(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}) = l_n(\hat{\theta}_n) + \frac{1}{\sqrt{n}} l_n^{'}(\hat{\theta}_n)\tilde{\theta} + \frac{1}{2n} l_n^{''}(\tilde{\theta}_n)\tilde{\theta}^2$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_n$ and $\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}$ and so tends to θ^* in probability as $n \to \infty$. But $l_n^{'}(\hat{\theta}_n) = 0$, which is essential for centering, we have

$$l_n(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}) = l_n(\hat{\theta}_n) + \frac{1}{2n} l_n^{''}(\tilde{\theta}_n)\tilde{\theta}^2$$

Recall that

$$\pi_n(\theta) = \frac{L_n(\theta)g(\theta)}{\int L_n(\theta)g(\theta)d\theta}$$

[2] T. S. Ferguson, "A Course in Large Sample Theory", CRC Press, 1996, p. 141.

and we set $\theta = \hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}$ so that $\pi_n(\theta)d\theta$ goes to $\tilde{\pi}_n(\tilde{\theta})d\tilde{\theta}$ and

$$\tilde{\pi}_n(\tilde{\theta}) = \frac{e^{nl_n(\hat{\theta}_n)+\frac{1}{2}l_n''(\tilde{\theta}_n)\tilde{\theta}^2} g(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}})}{\int e^{nl_n(\hat{\theta}_n)+\frac{1}{2}l_n''(\tilde{\theta}_n)\tilde{\theta}^2} g(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}})d\tilde{\theta}}$$

or

$$\tilde{\pi}_n(\tilde{\theta}) \approx \frac{e^{\frac{1}{2}l_n''(\tilde{\theta}_n)\tilde{\theta}^2} g(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}})}{\int e^{\frac{1}{2}l_n''(\tilde{\theta}_n)\tilde{\theta}^2} g(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}})d\tilde{\theta}}$$

and expanding $g(\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}) \approx g(\hat{\theta}_n)$, cancelling it in numerator and denominator, and doing the Gaussian integral we almost get the result.

To complete the calculation, we note that from the uniform law of large numbers (17) for the second derivatives of the log likelihood function, and since $\tilde{\theta}_n$, which is between $\hat{\theta}_n$ and $\hat{\theta}_n + \frac{\tilde{\theta}}{\sqrt{n}}$, tends to θ^* in probability as $n \to \infty$, we have that

$$l_n''(\tilde{\theta}_n) \to -I(\theta^*)$$

in probability, as we saw in the CLT for the error in MLE. Using this we get the desired result (20). Note in particular that the g, the prior, cancels.

To summarize, this last relation does imply (20), in probability for each $\tilde{\theta}$. It can also be shown that this implies that

$$\int |\tilde{\pi}_n(\tilde{\theta}) - \frac{e^{-\frac{I(\theta^*)\tilde{\theta}^2}{2}}}{\sqrt{2\pi(1/I(\theta^*))}}| \, d\tilde{\theta} \to 0$$

in probability.

Classical large sample MLE is among the most useful method in data analysis. It is based on the classical limit theorems of probability, the WLLN, the uniform one in particular, and the CLT, along with Slutsky's theorem(s). The Bayesian version of the large sample MLE, the large sample maximum aposteriori probability (MAP), is essentially equivalent to it. There may, however, be advantages to using the one or the other formulation in specific applications.

2.4 Kullback-Leibler divergence and MLE consistency

We return to the consistency of the MLE discussed in Section 2.1 and provide a proof using the Kullback-Leibler divergence (KLD).

For two probability density functions $f_{\theta_1}, f_{\theta_0}$, the KLD is defined as

$$K(f_{\theta_1}, f_{\theta_0}) = \mathbb{E}_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(x)}.$$

Suppose we have data $X_1, \cdots, X_n$ that are i.i.d. from the distribution f_{θ_0}. We will prove consistency in two steps.

1. Assume that the parameter θ is in a finite set Θ and the densities corresponding to different θ are distinct and have the same support. We then show that the maximum likelihood estimator $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \xrightarrow{P} \theta_0$.

2. Assume that the parameter set Θ is compact and the densities corresponding to different $\theta \in \Theta$ are distinct and have the same support, and we have uniform continuity

$$\sup_{|\theta' - \theta| < \delta} \sup_x |\log f_{\theta'}(x) - \log f_\theta(x)| \to 0$$

as $\delta \to 0$. We then show that the maximum likelihood estimator $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \xrightarrow{P} \theta_0$.

We start with the first statement. Jensen's inequality says that

$$\phi(\mathbb{E}(Y)) \leq \mathbb{E}(\phi(Y))$$

if $\phi(y)$ is convex in the range of values of the RV Y. We now let $\phi(y) = -\log y$, and $Y = f_{\theta_1}(X)/f_{\theta_0}(X)$ where X has density f_{θ_0}. Then

$$0 = -\log(\mathbb{E}_{\theta_0}(Y)) \leq \mathbb{E}_{\theta_0}(-\log(Y)) = K(f_{\theta_1}, f_{\theta_0})$$

Equality holds if and only if $f_{\theta_0} = f_{\theta_1}$ $\mathbb{P}_{\theta_0}$ almost surely, i.e. $\theta_0 = \theta_1$.

For any $\theta \in \Theta$, by the weak law of large numbers,

$$l_n(\theta) - l_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} \mathbb{E}_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_\theta).$$

Hence,

$$\mathbb{P}_{\theta_0} \left(\frac{1}{n} \sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0 \right) \overset{n \to \infty}{\to} 0.$$

Recall that as shown above, $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$. Since $\hat{\theta}_n$ is the maximizer of the LHS, we have

$$\mathbb{P}_{\theta_0}(\hat{\theta_n} \neq \theta_0) = \mathbb{P}_{\theta_0} \left(\max_{\theta \neq \theta_0} \frac{1}{n} \sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0 \right) \leq \sum_{\theta \neq \theta_0} \mathbb{P}_{\theta_0} \left(\frac{1}{n} \sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0 \right).$$

Thus,

$$\mathbb{P}_{\theta_0}(\hat{\theta_n} \neq \theta_0) \overset{n \to \infty}{\to} 0.$$

For the second statement, we first show a uniform law of large numbers. For any $\eta > 0$, since Θ is compact and $\sup_{|\theta'-\theta|<\delta} \sup_x |\log f_{\theta'}(x) - \log f_\theta(x)| \to 0$ as $\delta \to 0$, we can find a finite subset Θ_η of Θ such that for any $\theta' \in \Theta$, there exists $\theta_1 \in \Theta_\eta$ such that $|\theta_1 - \theta'| < \eta$ and $\sup_x |\log f_{\theta_1}(x) - \log f_{\theta'}(x)| < \eta$. By compactness of Θ, for all $\epsilon > 0$, there exists $\eta > 0$ such that $K(f_{\theta_0}, f_\theta) > 2\eta$ for all $|\theta - \theta_0| > \epsilon$. Then

$$\begin{aligned}
\mathbb{P}_{\theta_0}(|\hat{\theta}_n - \theta_0| > \epsilon) &\leq \mathbb{P}_{\theta_0}\left(\exists \theta', |\theta' - \theta_0| > \epsilon : \frac{1}{n}\sum_{i=1}^n \log \frac{f_{\theta'}(X_i)}{f_{\theta_0}(X_i)} > 0\right) \\
&\leq \mathbb{P}_{\theta_0}\left(\exists \theta_1 \in \Theta_\eta, |\theta_1 - \theta_0| > \epsilon - \eta : \frac{1}{n}\sum_{i=1}^n \log \frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)} > -\eta\right) \\
&\leq \sum_{\theta_1 \in \Theta_\eta : |\theta_1 - \theta_0| > \epsilon - \eta} \mathbb{P}_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^n \log \frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)} > -\eta\right) \\
&\overset{n\to\infty}{\to} 0.
\end{aligned}$$

3 Basic Monte Carlo methods

The main problem in Monte Carlo simulation is to calculate using sampling methods expectations of complicated (multi-dimensional) functions of random variables (random vectors) that have also complicated distributions.

The most direct but often difficult to apply way of generating random variables with a given distribution $F(x)$ (density $f(x) = F'(x)$) is by noting that if U is a uniform random variable over $[0, 1]$ then $F^{-1}(U)$ has distribution $F(x)$. This is because $P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$. Of course generating i.i.d. uniform random variables $U_1, U_2, \ldots, U_n$ numerically is not easy and requites deep number theoretic methods, especially if n is large and rigorous statistical tests for independence are to be satisfied. But if we assume that this can be done then generation of a sample $X_1, X_2, \ldots, X_n$ with density f is "easy" except when the inverse distribution function F^{-1} is hard to obtain, and then the acceptance-rejection algorithm can be used.

Given a function $g(x)$ and assuming that we can generate an i.i.d. sample from $f(x)$ then we approximate $I = E(g(X))$ by

$$I_n = \frac{1}{n} \sum_{j=1}^{n} g(X_j)$$

This is basic Monte Carlo.

3.1 Properties of basic Monte Carlo

Clearly $E(I_n) = I$ and

$$\mathrm{var}(I_n) = E\left[\left(\frac{1}{n}\sum_{j=1}^{n} g(X_j) - I\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{j=1}^{n} (g(X_j) - I)\right)^2\right]$$

$$= E\left[\frac{1}{n^2}\sum_{j=1}^{n} (g(X_j) - I)^2\right] = \frac{1}{n}\mathrm{var}(g(X))$$

The fact that the variance of the Monte Carlo approximation I_n decays as one over the sample size is characteristic of this method and the main limiting factor for its applicability. It is therefore important to look for ways of reducing the multiplicative factor $\sigma^2 = \mathrm{var}(g(X))$ and this is what importance sampling tries to do.

We can also use the CLT to get confidence intervals for I using the approximations I_n. The CLT does apply since we assume that $\mathrm{var}(g(X)) < \infty$. Thus, $\sqrt{n}(I_n - I)$ converges in law to an $N(0, \sigma^2)$ random variable. Given the error level α (say $\alpha = 0.05$) and if ζ is such

that for a normal random variable Z with mean zero and variance one $P(|Z| > \zeta) = \alpha$, then for n large we have approximately

$$P(\frac{\sqrt{n}}{\sigma}|I_n - I| \leq \zeta) \sim 1 - \alpha$$

or

$$I_n - \frac{\zeta\sigma}{\sqrt{n}} \leq I \leq I_n + \frac{\zeta\sigma}{\sqrt{n}}$$

with probability $1 - \alpha$. Of course σ is not known and must be replaced by the sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(g(X_j) - I_n)^2$$

to get a realizable confidence interval. The justification for replacing σ by s_n in the CLT, that is, in still having $\frac{\sqrt{n}}{s_n}(I_n - I) \sim N(0,1)$ in law, requires the use of Slutsky's theorem, since $s_n \to \sigma > 0$ in probability and $\sqrt{n}(I_n - I) \to N(0, \sigma^2)$ in distribution and therefore $\frac{\sqrt{n}}{s_n}(I_n - I) \sim N(0,1)$ in distribution. The confidence intervals are now realizable

$$I_n - \frac{\zeta s_n}{\sqrt{n}} \leq I \leq I_n + \frac{\zeta s_n}{\sqrt{n}}$$

with probability $1 - \alpha$. The question of how large the number of realizations n must be for this to be reasonably accurate based on the asymptotic theory depends both on the integrand g and on the density f. Since we are not doing estimation here it is usually possible to increase the number of realizations and thus improve the accuracy of the confidence interval.

One can also use the continuous mapping theorem that can be stated in general as follows. Suppose that the pair of random variables (X_n, Y_n) converges in distribution to (X, Y). Let $h(x, y)$ be a function from $R^2 \to R$ and such that the set $\{(x, y) \mid h(x, y) \text{ is not continuous}\}$ has probability zero with respect to the limit law. Then $h(X_n, Y_n) \to h(X, Y)$ in distribution. This more general theorem can be applied here with $X_n = \sqrt{n}(I_n - I)$, $Y_n = s_n$ and $h(x, y) = x/y$.

3.2 Importance sampling

Importance sampling is used when the function $g(x)$ to be integrated and the density $f(x)$ overlap very little so that the product $g(x)f(x)$ is very small and therefore $I = E(g(X))$ is very small. Most of the samples drawn from f will not overlap significantly with regions where g is significant. This means that the relative error (standard deviation over mean) in direct Monte Carlo simulations will be very large.

The main idea in importance sampling is to introduce a reference density $\tilde{f}(x)$, let

$$M(x) = \frac{f(x)}{\tilde{f}(x)}$$

and then note that, assuming all integrals are well defined,

$$\tilde{E}(gM) = \int g(x)\frac{f(x)}{\tilde{f}(x)}\tilde{f}(x)dx = \int g(x)f(x)dx = E(g) = I$$

We can now generate an unbiased Monte Carlo approximation by using a sample from $\tilde{f}$, $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$, so that

$$I_n = \frac{1}{n}\sum_{j=1}^{n} g(\tilde{X}_j)M(\tilde{X}_j)$$

The variance of I_n is now

$$\tilde{E}((gM)^2) - (\tilde{E}(gM))^2 = \int (g(x))^2\frac{(f(x))^2}{\tilde{f}(x)}dx - I^2$$

The question is how to choose $\tilde{f}$ so as to reduce the variance of I_n. Assuming that g is positive we see right away that if we take

$$\tilde{f}(x) = \frac{g(x)f(x)}{E(g(X))}$$

then the variance of I_n is zero! The reference density puts all the weight just where it should, that is, where the product gf is significant. But this is hardly an improved Monte Carlo since in order to implement it we need to know $E(g(X))$, which is the very quantity that we are trying to approximate. A number of interesting algorithms can be developed, however, that use approximations of $E(g)$ to get $\tilde{f}$, which will then lead to improved approximations.

One possibility is to choose $\tilde{f}(x)$ as follows. Let $\{x_j\}$ be a partition of the real line and define

$$\tilde{f}(x) = \frac{\sum_j f(x_j^*)g(x_j^*)\mathbb{1}_{\{x_{j-1}<x\leq x_j\}}}{\sum_j f(x_j^*)g(x_j^*)(x_j - x_{j-1})}$$

where $x_j^* \in (x_{j-1}, x_j)$. Using sampling with this as a reference density tends to reduce the variance of the Monte Carlo simulation. If both $f(x)$ and $g(x)$ are differentiable and $\Delta x = max_j(x_j - x_{j-1})$ then $|I - \sum_j f(x_j^*)g(x_j^*)(x_j - x_{j-1})| = O(\Delta x)$ and the variance of the Monte Carlo approximation I_n, with this $\tilde{f}(x)$ as a reference density, is of order $\frac{\Delta x}{n}$ and therefore can be reduced by reducing Δx.

3.3 Acceptance-rejection

Suppose that we want to sample from a density $f(x)$ and we do not want to use $F^{-1}(U)$ where U is uniform in $[0, 1]$ because F^{-1} is too complicated. We may then be able to do

the following. Suppose there is another density $g(x)$ from which it is easy to sample and suppose that there is a constant $c > 1$ such that

$$\frac{f(x)}{g(x)} \leq c$$

for all x, and we may consider the smallest such constant. The acceptance-rejection algorithm consists of the following steps:
1. Generate Z from g
2. Generate an independent uniform $[0, 1]$ random variable U
3. Is $U \leq \frac{f(Z)}{cg(Z)}$? If yes then return Z (accept) and if not then repeat (go to step 1).

The output random variable, say X, will be in a set A if

$$\mathbb{1}_{\{X\in A\}} = \mathbb{1}_{\{Z_1\in A\}}\mathbb{1}_{\{U_1\leq \frac{f(Z_1)}{cg(Z_1)}\}} + \sum_{k=2}^{\infty}\left(\prod_{j=1}^{k-1}\mathbb{1}_{\{U_j>\frac{f(Z_j)}{cg(Z_j)}\}}\right)\mathbb{1}_{\{Z_k\in A\}}\mathbb{1}_{\{U_k\leq \frac{f(Z_k)}{cg(Z_k)}\}}$$

But

$$E(\mathbb{1}_{\{Z\in A\}}\mathbb{1}_{\{U\leq \frac{f(Z)}{cg(Z)}\}}) = \int \mathbb{1}_{\{z\in A\}}E(\mathbb{1}_{\{U\leq \frac{f(z)}{cg(z)}\}})g(z)dz = \frac{1}{c}\int_A f(z)dz$$

and using the independence of the random variables in the acceptance-rejection algorithm we see that

$$P(X \in A) = \frac{1}{c}\int_A f(z)dz + \frac{1}{c}\int_A f(z)dz\sum_{k=2}^{\infty}(1-\frac{1}{c})^{k-1} = \int_A f(z)dz$$

This shows that indeed the acceptance rejection algorithm does produce random variables with the correct density.

To understand better how the algorithm behaves, let N be the number of cycles needed to produce the desired random variable. Clearly

$$P(N = k) = \left[\int P(U > \frac{f(z)}{cg(z)})g(z)dz\right]^{k-1}\left[1 - \int P(U > \frac{f(z)}{cg(z)})g(z)dz\right] \ , \quad k = 1, 2, \ldots$$

$$= (1-\frac{1}{c})^{k-1}\frac{1}{c},$$

which is the geometric law, and therefore $E(N) = c$. This explains the role that the constant c plays.

Now acceptance-rejection can be combined with Monte Carlo somewhat in the way importance sampling is done as follows. Suppose that it is hard to sample from $f(x)$ and that there is another density $g(x)$ and a constant $c > 1$ such that $f(x)/g(x) \leq c$ for all

x. To compute $I = E(h(X))$ for some function $h(x)$ we generate $X_1, X_2, \ldots, X_n$ by the acceptance rejection algorithm and then compute

$$I_n = \frac{1}{n} \sum_{j=1}^{n} h(X_j)$$

This is a slightly less efficient approximation than when we generate the sample directly from $f(x)$ because if we count the steps needed in the acceptance rejection cycle then a total nc (with $c > 1$) steps are needed on average to compute I_n. Counting also the uniform random variables in the acceptance-rejection algorithm, we see that we need on average $2nc$ samples to generate I_n. The accuracy of I_n remains the same but the computational cost has increased.

3.4 Glivenko-Cantelli theorem and the Kolmogorov-Smirnov test

There is not much practical or theoretical methodological distinction between Monte Carlo and estimation methods so we discuss non-parametric estimation in this part of the notes. The connection between estimation and Monte Carlo is Bootstrap, discussed briefly in section 3.6.

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a distribution $F(x)$ and define the empirical distribution $F_n(x)$ by

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{X_j \leq x\}}$$

This is a random, since it depends on the sample, piece-wise constant distribution function that should serve as an estimate for the true $F(x)$, assumed unknown here. Since for each $x \in R$ the random variables $\mathbb{1}_{\{X_j \leq x\}}$ are bounded by one, have mean $F(x)$ and are independent, the weak law of large numbers tells us that $F_n(x) \to F(x)$ in probability. In fact this is true also with probability one and the CLT holds:

$$\sqrt{n}\,(F_n(x) - F(x)) \to N(0, F(x)(1 - F(x)) \text{ in distribution for each } x \in R$$

This is not so useful though because the variance of the limit Gaussian depends on the very distribution we want to estimate. Normalizing by F_n so that

$$\frac{\sqrt{n}}{\sqrt{F_n(x)(1 - F_n(x))}}\,(F_n(x) - F(x)) \to N(0, 1) \text{ in distribution for each } x \in R$$

is not as useful either because of the rather slow convergence that depends on x.

There are two basic results in non-parametric estimation that we now state and discuss briefly.

The first is the Glivenko-Cantelli limit theorem and the second the Kolmogorov-Smirnov limit theorem.

The Glivenko-Cantelli limit theorem states that

$$P\{\lim_{n\to\infty}\sup_{x\in R}|F_n(x)-F(x)|=0\}=1$$

The strong law of large numbers says that with probability one $\lim_{n\to\infty}|F_n(x)-F(x)|=0$ for each $x\in R$. So the main issue is to get the uniformity over $x\in R$. Here it can be shown easily because distribution functions are monotone and they are continuous except for at most a countable number of jumps. For the proof we note that given any $\epsilon>0$ we can choose a partiton of the real line $x_{-N}<x_{-N+1}<\ldots<x_j<\ldots x_N$ such that all the partition points are continuity points and $F(x_{j+1)}-F(x_j)\leq\epsilon$ for all indices $-N\leq j\leq N$, as well as having $F(x_{-N})\leq\epsilon$ and $1-F(x_N)\leq\epsilon$. We then have that for $x\in(x_{j-1},x_j)$

$$\begin{aligned}F_n(x)-F(x)\leq F_n(x_j)-F(x_{j-1})&\leq|F_n(x_j)-F(x_j)|+|F(x_j)-F(x_{j-1})|\\&\leq|F_n(x_j)-F(x_j)|+\epsilon\end{aligned}$$

and similarly for a lower bound, so that

$$\sup_x|F_n(x)-F(x)|\leq\max_{-N\leq j\leq N}|F_n(x_j)-F(x_j)|+\epsilon$$

Taking the limit we have that with probability one

$$\begin{aligned}\lim_{n\to\infty}\sup_x|F_n(x)-F(x)|&\leq\lim_{n\to\infty}\max_{-N\leq j\leq N}|F_n(x_j)-F(x_j)|+\epsilon\\&\leq\max_{-N\leq j\leq N}\lim_{n\to\infty}|F_n(x_j)-F(x_j)|+\epsilon=\epsilon\end{aligned}$$

which completes the proof. We have interchanged the limit with the maximum since it is only over a finite number of indices.

The Kolmogorov-Smirnov limit theorem states that if $F(x)$ is continuous then $D_n=\sup_{x\in R}|F_n(x)-F(x)|$ is a random variable whose law is independent of F and $\sqrt{n}D_n$ converges in distribution to a universal law,

$$\lim_{n\to\infty}P\{\sqrt{n}D_n\leq x\}=1-2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}$$

To see that the random variable does not, in fact, depend on the underlying distribution F we note that since the map $x=F^{-1}(y)$ is one to one from $[0,1])$ to the real line so we can replace x by $F^{-1}(y)$ in D_n to get $D_n=\sup_{y\in[0,1]}|F_n^U(y)-y|$ where the superscript U indicates that the empirical distribution function $F_n^U(x)$ is from an i.i.d. sample from the uniform distribution.

With the KS theorem we can formulate various statistical tests regarding the estimation of the unknown distribution $F(x)$. The limit distribution of $\sqrt{n}D_n$ is, for continuous $F(x)$,

the distribution of the maximum of the absolute value of the Brownian bridge because it reduces to dealing with uniform random variables. A Brownian bridge is a Gaussian stochastic process $\tilde{B}(t)$ defined on $[0,1]$, with mean zero and covariance $E\{\tilde{B}(t)\tilde{B}(s)\} = t(1-s)$ if $0 < t \leq s < 1$, symmetric in t and s. A Gaussian process indexed by $t \in [0,1]$ is a collection of Gaussian random variables such that for any set of indices $0 \leq t_0 < t_1 < \cdots < t_N \leq 1$ the vector $(\tilde{B}(t_0), \tilde{B}(t_1), \ldots, \tilde{B}(t_N))$ is Gaussian with mean zero and covariance matrix given as above. This limit theorem is an example of an invariance principle where the law of a function of the process before the limit, in this case the maximum absolute value, converges to the law of the same function of the limit process.

3.5 Density kernel estimation

A simpler non-parametric test when an unknown density is involved will be discussed in some detail because it shows explicitly the dependence of the rate of convergence on the smoothness of the density. In other words, one sees explicitly the slowing of the rate of convergence in a non-parametric estimation.

Suppose that we have a sample $X_1, X_2, \ldots, X_n$ from a density $f(x)$ which is not known and we want to estimate it. Since $F'(x) = f(x)$ an estimator for f can be obtained by differentiating the empirical distribution F_n. But this is a random step function so its derivative is a sum of delta functions. So we need some smoothing which we do with a smoothing kernel, a positive, infinitely differentiable function $\phi(x)$ such that

$$\phi(x) \geq 0 \ , \quad \int_R \phi(x)dx = 1 \ , \quad \int_R x\phi(x)dx = 0 \ , \quad \int_R x^2\phi(x)dx = 1$$

The estimate of the density f is now

$$f_{n,h}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h}\phi\left(\frac{x - X_j}{h}\right)$$

Clearly

$$E\{f_{n,h}(x)\} = \int_R \phi_h(x-y)f(y)dy \approx f(x) + \frac{h^2}{2}f''(x) = f(x) + \text{bias}$$

for h small, assuming differentiability of the density. Here $\phi_h(x) = \phi(x/h)/h$ and the small h expansion is just a Taylor expansion plus use of the normalization properties of the smoothing kernel ϕ.

The variance of the estimator is similarly given by

$$\text{var}(f_{n,h}(x)) = \frac{1}{n}\text{var}(\phi_h(x - X)) = \frac{1}{n}\left[\int_R \phi_h^2(x-y)f(y)dy - (\int_R \phi_h(x-y)f(y)dy)^2\right]$$

Assuming form here on the $\phi(x)$ in the Gaussian mean zero, variance one density we have that

$$\mathrm{var}(f_{n,h}(x)) \approx \frac{1}{n}\left[\frac{1}{h}\frac{1}{2\sqrt{\pi}}f(x) - f^2(x)\right]$$

to principal order as $h \to 0$. From this we conclude that the mean square error is

$$E\{(f_{n,h}(x)) - f(x))^2\} = \mathrm{var}(f_{n,h}(x)) + (\mathrm{bias})^2 \approx \frac{1}{h}\frac{1}{2n\sqrt{\pi}}f(x) + \frac{h^4}{4}(f''(x))^2$$

to principal order in $h \to 0$ for each term. Minimizing this error over h gives

$$h^*(n) = \frac{1}{n^{1/5}}\left(\frac{1}{2\sqrt{\pi}}\frac{f(x)}{(f''(x))^2}\right)^{1/5}$$

at points $x \in R$ such that $f''(x)$ is not zero.

It follows that with the optimal in MSQ sense smoothing we have

$$f_{n,h^*(n)}(x) \approx f(x) + O(n^{-2/5}) \quad \text{for each } x \text{ for which } f''(x) \neq 0,$$

in MSQ sense. Instead of an error $O(n^{-1/2})$, which is the usual one for parameter estimation, we have slower error decay because of the necessary smoothing. And, of course, the density $f(x)$ to be estimated must have at least two derivatives.

3.6 Bootstrap

Let $X_1, X_2, \ldots, X_n$ be a sample drawn from a distribution function $F(x)$ and let $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ be a statistic of interest, such as an estimator of an unknown parameter. We often want to calculate the standard deviation of this statistic, $SD = \sigma(F, n, \hat{\theta}) = \sigma(F)$, or some other measure of uncertainty, but the distribution function F is not known or it is not known with enough accuracy. The bootstrap[3] is a very effective way to do this.

We introduce the empirical distribution function of the sample, $\hat{F}_n$, which assigns mass $1/n$ at at the points x_i, $i = 1, 2, \ldots, n$, the values of the observed sample. We want to calculate $\hat{SD} = \sigma(\hat{F}_n, n, \hat{\theta}) = \sigma(\hat{F}_n)$ which, depending on the statistic $\hat{\theta}$ and the distribution F will be close to the theoretical SD for n large. The issue in bootstrap is primarily how to calculate $\hat{SD}$ for fixed but large n, since this is often a combinatorially complex problem for large n.

We do this by "Monte Carlo" using $\hat{F}_n$ as the basic distribution from which to sample. A bootstrap sample is denoted by $X_1^*, X_2^*, \ldots, X_n^*$, which is an i.i.d sequence drawn from $\hat{F}_n$. This is the same as drawing from $(x_1, x_2, \ldots, x_n)$ with replacement n times. Let $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$ be the bootstrap value of the statistic. The bootstrap standard

[3] See the lecture notes by B. Efron, "The Jackknife, the Bootstrap and Other Resampling Plans", CBMS-NSF Series in Applied Mathematics no. 38, 1982

deviation is simply a Monte Carlo approximation of $\hat{SD} = \sigma(\hat{F}_n)$. We do this by drawing repeated independent samples of size n from $\hat{F}_n$ and letting

$$\hat{SD}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2 \right)^{1/2}$$

where B is the number of bootstrap samples and it is assumed large enough to be a good Monte Carlo approximation of $\sigma(\hat{F}_n)$. Here $\hat{\theta}^{*\cdot}$ is the empirical mean of the bootstrap values of the statistic over the B Monte Carlo samples (re-samples). As B tends to infinity we have that $\hat{SD}_B \to \hat{SD}$, but since for finite n this is only an approximation of SD, which is what we want, we should ideally choose B so that the Monte Carlo error is comparable to the finite n error.

As an example, consider the sample mean $\bar{X}$ as the statistic of interest. Then the standard deviation is $\sigma(F) = (\frac{\mu_2}{n})^{1/2}$ where $\mu_2 = E_F(X - E_F(X))^2$ is the theoretical variance of F. The bootstrap standard deviation $\hat{SD} = (\frac{\hat{\mu}_2}{n})^{1/2}$ where $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, which is the sample variance. Thus $E_F(\hat{VAR}) = E_F(\frac{\hat{\mu}_2}{n}) = \frac{n-1}{n}\frac{\mu_2}{n} = \frac{n-1}{n} Var(\bar{X})$. The point of this example is that the bootstrap is consistent with what is expected in the same way that basic Monte Carlo is excepted to work, but now we only deal with the original sample of size n by resampling it.

4 Markov Chains

A time-homogeneous Markov chain $\{X_0, X_1, \ldots, X_n, \ldots\}$ taking values in a finite set S of size N is characterized by its transition probabilities

$$P(x,y) = P\{X_{n+1} = y | X_n = x\} \geq 0 \ , \quad x, y \in S \ , \quad \sum_y P(x,y) = 1,$$

which are independent of n because of time homogeneity. The Markov property means that conditional probabilities depend only on the latest information

$$P\{X_n = y | X_0, X_1, \ldots, X_{n-1}\} = P\{X_n = y | X_{n-1}\}$$

and therefore the joint probability of $\{X_n\}_{n\geq 0}$, in path space, is expressed as a product of transition probabilities

$$P\{X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0\} = \pi_0(x_0) P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

where $\pi_0(x) = P\{X_0 = x\}$. Path probabilities fully determine the S-valued process $\{X_n\}_{n\geq 0}$, and for Markov chains the initial probabilities π_0 and the transition probabilities P determine everything.

The n-step transition probability

$$P\{X_n = y | X_0 = x\} = P_x\{X_n = y\}$$

is obtained recursively by using the Markov property and time homogeneity. We have, introducing also some notation and using the law of iterated conditional expectation (or total probability), that

$$\begin{aligned} P_n(x,y) = P_x\{X_n = y\} = E_x\{P\{X_n = y | X_1, X_0 = x\}\} \quad & \text{Iterated conditional expectation} \\ = E_x\{P\{X_n = y | X_1\}\} \quad & \text{Markov property} \\ = E_x\{P_{X_1}\{X_{n-1} = y\}\} \quad & \text{time homogeneity} \\ = \sum_{z\in S} P(x,z) P_z\{X_{n-1} = y\} = \sum_{z\in S} P(x,z) P_{n-1}(z,y) & \end{aligned}$$

so that the matrix $(P_n(x,y)) = (P^n(x,y))_{x,y\in S}$ is identified as the n-th power of the $N \times N$ transition matrix $P = (P(x,y))_{x,y\in S}$.

If $\pi_0(x) = P\{X_0 = x\},\ x \in S$, is the initial probability of the chain then

$$P(X_n = x) = \pi_n(x) = \sum_{z\in S} \pi_{n-1}(z) P(z,x) = \sum_{z\in S} \pi_0(z) P^n(z,x) \ , \quad n = 1, 2, ...,\ x \in S.$$

Thus probability vectors $\pi_n = (\pi_n(x))_{x\in S}$ can be considered to be N-row vectors that get updated recursively by left multiplication with the transition matrix.

Expectations of functions of the state given the initial state

$$E_x\{f(X_n)\} = u_n(x) = \sum_{z\in S} P(x,z)u_{n-1}(z) = \sum_{z\in S} P^n(x,z)f(z)\ ,\quad n = 1,2,3,...,\ x \in S$$

can similarly be considered as N-column vectors $u_n = (u_n(x))_{x\in S}$ that are updated recursively by right multiplication with P, starting with the initial column vector $f = (f(x))$ when $X_0 = x$.

4.1 Exit times

Let $C \subset S$ and let $T = T_x$ be the first time to enter the complement of C, C^c, starting from $x \in C$, which is also the first exit time from C:

$$T = \min\{n \geq 1 \mid X_n \notin C\}$$

This is a random variable that takes integer values and is such that for all $n \geq 0$ the events $\{T > n\}$ depend only on the Markov chain up to time n, $\{X_0, X_1, X_2, \ldots, X_n\}$, that is, the random variable $\mathbb{1}_{\{T>n\}}$ is a function of the random variables $\{X_0, X_1, X_2, \ldots, X_n\}$.

$$\mathbb{1}_{\{T>n\}} = \mathbb{1}_{\{X_0\in C, X_1\in C, \ldots, X_n\in C\}}$$

It is called a stopping time. Clearly the exit time of the Markov chain from a subset C of the state space S is a stopping time.

We want to find a linear system of equations satisfied by $v_n(x) = P_x\{T > n\} = 1 - P_x\{T \leq n\}$, which is the probability distribution of T, starting from x and with $n = 0, 1, ...$ We use a **first transition or a renewal analysis** that relies on the Markov property and time homogeneity as follows. For $x \in C$ we have

$$v_n(x) = P_x\{T > n\} = E_x\{P\{T > n|X_1, X_0 = x\}\} = E_x\{P\{T > n|X_1\}\}$$

$$= \sum_{y\in C} P(x,y)P_y\{T > n-1\} = \sum_{y\in C} P(x,y)v_{n-1}(y)\ ,\quad n = 1,2,3,...\ ,\quad v_0(x) = \mathbb{1}_{\{x\in C\}}$$

To see why $P\{T > n|X_1\} = P_{X_1}\{T > n-1\}$, and to explain the notation, we first write $T = T(X_0, X_1, X_2, \ldots)$ to indicate that the exit time depends on the path of the Markov chain. By the definition of the exit time from C, $T > 0$ when $X_0 = x \in C$. With $n = 1, 2, \ldots,$ after one time unit has passed,

$$\mathbb{1}_{\{T(X_0,X_1,X_2,...)>n\}} = \mathbb{1}_{\{1+T(X_1,X_2,...)>n\}}$$

when $X_1 \in C$. In words, after one time step the Markov chain restarts from the state $X_1 \in C$ to which it went, and the exit time increases by one unit. Time homogeneity leads then to the result above.

If we let $P^C = (P(x,y),\ x,y \in C)$, which is a sub-stochastic transition matrix since in general its row sums are less than one, then for the vector $v_n = (v_n(x))_{x\in C}$ we have the linear recursion

$$v_n = P^C v_{n-1}\ ,\quad v_0 = \mathbf{1}$$

Note again that the column vectors v_n are restricted to elements x in C and $\mathbf{1}$ is the column vector of all ones. Alternatively we can write the recursion as an initial-boundary value problem:

$$v_n(x) = \sum_{y\in S} P(x,y)v_{n-1}(y)\ ,\quad n = 1,2,\dots\ ,\quad x \in S$$

with

$$v_0(x) = 1,\ x \in C\ ,\quad v_n(x) = 0,\ x \notin C,\ n = 0,1,\dots$$

Let us define the norm of row vectors to be the l^1 norm

$$||q|| = \sum_{y\in S} |q(y)|$$

and the norm of column vectors to be the maximum norm

$$||f|| = \max_x |f(x)|$$

Then the induced matrix norm is

$$||Q|| = \max_x \sum_y |Q(x,y)|$$

and we see that the norm of transition matrices P is one but the norm of P^C is in general less than one. This is the case if transitions from states in C to states in the complement occur with positive probability. Therefore, since $||Q^2|| \le ||Q||^2$, we conclude that, in general, $||v_n|| \to 0$ as $n \to \infty$, assuming that $||P^C|| < 1$. This means that the exit time form C, T, is finite with probability one as should be expected in this case.

Consider as another example the calculation of the mean exit time from C, $E_x\{T\}$. For some fixed $0 < s < 1$ define the moment generating function of the exit time

$$u(x;s) = E_x\{s^T\}$$

Using the first transition or renewal argument, as above, we find that u satisfies the linear system

$$u(x;s) = s\sum_{y\in C^c} P(x,y) + s\sum_{y\in C} P(x,y)u(y;s),\ x \in C$$

Assuming a finite expectation for the exit we have that

$$\bar{u}(x) = \frac{d}{ds}u(x;s)|_{s=1} = E_x\{T\},\ x \in C$$

By differentiating the equation for $u(x;s)$ with respect to s and setting $s=1$ we get, after some rearrangement

$$\bar{u}(x) = 1 + \sum_{y \in C} P(x,y)\bar{u}(y), \;\; x \in C$$

and in vector form

$$P^C \bar{u} - \bar{u} = -\mathbf{1}$$

Since the norm of P^C is less than one, as it is in general when the Markov chain can reach states in C^c from states in C, we have that

$$\bar{u} = (I - P^C)^{-1}\mathbf{1}$$

As one more example consider the calculation of

$$u(x) = u(x,z;y) = \mathbb{E}_x \left[\sum_{i=0}^{T_z - 1} \mathbb{1}_{\{X_i = y\}} \right]$$

where T_z is the first time to reach a fixed state z, y is some other state, and with x and y different from z. A complete analysis of u **as a function of** y is in section 4.7. Equations are here obtained for u as a function of x by first step analysis as follows. Let $f(x) = \mathbb{1}\{x = y\}$. We then have,

$$\begin{aligned}
u(x) = \mathbb{E}_x \left[\sum_{i=0}^{T_z - 1} f(X_i) \right] &= \mathbb{E}_x \left[\mathbb{E}_x \left[\sum_{i=0}^{T_z - 1} f(X_i) | X_1 \right] \right] \\
&= \mathbb{E}_x \left[f(x) + \mathbb{E}_x \left[\sum_{i=1}^{T_z - 1} f(X_i) | X_1 \right] \right] \\
&= f(x) + \mathbb{E}_x \left[\mathbb{E}_{X_1} \left[\sum_{i=0}^{T_z - 1} f(X_i) \right] \right] \\
&= f(x) + \sum_{w \neq z} P(x,w) \mathbb{E}_w \left[\sum_{i=0}^{T_z - 1} f(X_i) \right] \\
&= f(x) + (Qu)(x) \\
&= \mathbb{1}_{\{x=y\}} + (Qu)(x), \quad x, y \in S - \{z\},
\end{aligned}$$

where Q is the restriction of P to the continuation region $S - \{z\}$ of the Markov chain, since z is the (one point) stopping region, and Qu is the matrix-vector product. This system could also be solved as

$$u(x) = f(x) + Pu(x), \;\; x \in S, \text{ with the condition } u(z) = 0.$$

Either way, we have a linear system. Going from the second to the third line in the derivation above involves time homogeneity, adjusting $T \to T + 1$ and shifting down the summation index.

4.2 Transience and recurrence

In this section we consider a Markov chain with transition probabilities $P(x,y)$, $x,y \in S$ where the state space S need not be finite, and let $F_n(x,y)$ be the probability that y is reached for the first time at n, starting from x

$$F_n(x,y) = P\{X_n = y, X_1 \neq y, \ldots, X_{n-1} \neq y | X_0 = x\},\ n = 1,2,...$$

Note that $F_n(x,y) = P_x\{T_y = n\}$ where T_y is the first time that the state y is reached. We want to show that

$$P^n(x,y) = \sum_{m=1}^{n} F_m(x,y) P^{n-m}(y,y),\ n = 1,2,...$$

In terms of generating matrix functions,

$$P(x,y;s) = \sum_{n=0}^{\infty} s^n P^n(x,y), \quad F(x,y;s) = \sum_{n=1}^{\infty} s^n F_n(x,y),$$

we have that

$$P(x,y;s) = \delta_{x,y} + F(x,y;s)P(y,y;s)$$

for $0 \leq s < 1$ and where $\delta_{x,y}$, same as $\mathbb{1}_{\{x \neq y\}}$, is zero when $x \neq y$ and one otherwise.

A state y is said to be persistent or recurrent if $F(y,y;1) = \sum_{n=1}^{\infty} F_n(y,y) = 1$, which means that $T_y < \infty$ with probability one. We will show that a state y is persistent if and only if $P(y,y;1) = \infty$. Since

$$P(y,y;1) = \sum_{n=0}^{\infty} P^n(y,y) = E_y\{\sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = y\}}\}$$

we conclude that a state is persistent if and only if the mean number of returns to it is infinite.

Note that the event $\{T_y = n\} = \{X_0 = x, X_1 \neq y, \ldots, X_{n-1} \neq y, X_n = y\}$ and so it depends on the path only up to time n. We thus have

$$\begin{aligned}
P\{X_n = y | X_0 = x\} &= \sum_{m=1}^{n} P\{X_n = y, T_y = m | X_0 = x\} \\
&= \sum_{m=1}^{n} P\{X_n = y | T_y = m, X_0 = x\} P\{T_y = m | X_0 = x\} \\
&= \sum_{m=1}^{n} P\{X_n = y | X_m = y\} P\{T_y = m | X_0 = x\} \\
&= \sum_{m=1}^{n} F_m(x,y) P^{n-m}(y,y)
\end{aligned}$$

Then, write,

$$\begin{aligned} P(x,y;s) &= \sum_{n=0}^{\infty} s^n P^n(x,y) \\ &= \delta_{x,y} + \sum_{n=1}^{\infty} s^n P^n(x,y) \\ &= \delta_{x,y} + \sum_{n=1}^{\infty} s^n \left(\sum_{m=1}^{n} F_m(x,y) P^{n-m}(y,y) \right) \end{aligned}$$

by the previous result. Interchanging the summations we get the result

$$P(x,y;s) = \delta_{x,y} + F(x,y;s)P(y,y;s)$$

We note that everything is non-negative and so monotone increasing in $s \to 1$. From

$$P(y,y;s) = \frac{1}{1 - F(y,y;s)}$$

we see that as F increases to one, we have that P increases to infinity. Similarly, from

$$F(y,y;s) = \frac{P(y,y;s) - 1}{P(y,y;s)}$$

we see that as P increases to infinity, F increases to one.

4.3 Strong Markov property

Let $X_0, X_1, X_2, \ldots$ be a Markov chain on a finite state space S and let T be a stopping time, that is, a non-negative integer-valued random variable for which $\mathbb{1}\{T > n\}$ is a function of the path $X_0, X_1, X_2, \ldots, X_n$ up to time n. We want to show that

$$P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) = P(X_{T+m} = y | X_T = x), \ m \ge 1,$$

which is the strong Markov property, that is, the Markov property when conditioning on the path of the chain up to a stopping time.

The conditional probability $P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x)$ can be calculated by specifying the value of the stopping time

$$\begin{aligned} &P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) \\ &\qquad = \sum_{n=1}^{\infty} \mathbb{1}_{\{T=n\}} P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) \end{aligned}$$

$$= \sum_{n=1}^{\infty} \mathbb{1}_{\{T=n\}} P(X_{T+m} = y | X_k = x_k \ 0 \le k < n, \ X_n = x)$$

But T is a stopping time, which means that the event $\{T = n\}$ depends on, or is a function of, $X_k = x_k \ 0 \le k \le n$. Therefore we have that

$$\begin{aligned} P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) & \\ &= \sum_{n=1}^{\infty} E(\mathbb{1}_{\{X_{T+m}=y\}} \mathbb{1}_{\{T=n\}} | X_k = x_k \ 0 \le k < n, \ X_n = x) \\ &= \sum_{n=1}^{\infty} E(\mathbb{1}_{\{X_{n+m}=y\}} \mathbb{1}_{\{T=n\}} | X_k = x_k \ 0 \le k < n, \ X_n = x) \\ &= \sum_{n=1}^{\infty} \mathbb{1}_{\{T=n\}} P(X_{n+m} = y | X_k = x_k \ 0 \le k < n, \ X_n = x) \end{aligned}$$

The ordinary Markov property now tells us that

$$P(X_{n+m} = y | X_k = x_k \ 0 \le k < n, \ X_n = x) = P(X_{n+m} = y | X_n = x)$$

Therefore we have

$$P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) = \sum_{n=1}^{\infty} \mathbb{1}_{\{T=n\}} P(X_{n+m} = y | X_n = x)$$

and by reversing the steps we took to get here using the definition of stopping time we have the strong Markov property

$$\begin{aligned} P(X_{T+m} = y | X_k = x_k \ 0 \le k < T, \ X_T = x) &= \sum_{n=1}^{\infty} \mathbb{1}_{\{T=n\}} P(X_{T+m} = y | X_n = x) \\ &= P(X_{T+m} = y | X_T = x) \end{aligned}$$

4.4 Invariant probabilities

We now want to address the long time behavior of the Markov chain X_n and in particular to study its ergodic properties. We assume that the state space S is finite (although this is not necessary for the method used; compactness of S is) and the transition probabilities are uniformly positive:

$$P(x, y) \ge \frac{\delta}{N} \ , \quad x, y \in S$$

for some $\delta > 0$ and with $N = \#(S)$. With simple modifications the arguments below extend to the case where this condition holds for P to some fixed power. We are simply

requiring that the Markov chain be irreducible and aperiodic. **Irreducible** means that every state can be reached from every other state in a finite number of time steps (at most N) with positive probability. **Aperiodic** means that the greatest common divisor of return times (with positive probability) to states is one, for all states.

We will prove the following basic theorem. For any probability (row) vectors π_1 and π_2 we have that

$$||\pi_1 P - \pi_2 P|| \leq (1-\delta)||\pi_1 - \pi_2||$$

which means that P is a strict contraction when acting on differences of probability vectors. Using this theorem we will then prove by the contraction mapping iteration process that P has a unique invariant probability vector π, or a positive left eigenvector with eigenvalue 1,

$$\pi = \pi P, \text{ in components: } \pi(x) = \sum_{y \in S} \pi(y) P(y,x) \ , \quad x \in S$$

and that this invariant probability is approached geometrically (exponentially) fast for any starting probability π_0

$$||\pi_0 P^n - \pi|| \leq 2\rho^n \ , \quad 0 < \rho < 1 \quad (\rho = 1 - \delta)$$

Note also that since the row sums of the transition probability matrix P sum to one, the column vector $\mathbf{1}$, of all ones, is a right eigenvector with eigenvalue one, $P\mathbf{1} = \mathbf{1}$.

To prove the contraction property, let $q = \pi_1 - \pi_2$, which is a vector with positive and negative entries such that $\sum_{y \in S} q(y) = 0$. Let $q = q^+ + q^-$ where q^+ has all non-negative elements and q^- has all negative ones. Clearly, $\sum_{y \in S} q^+(y) = -\sum_{y \in S} q^-(y)$ and

$$||q|| = ||q^+|| + ||q^-|| = 2||q^+||.$$

We now have the following

$$\begin{aligned} ||qP|| &= \sum_y |qP(y)| \\ &= \sum_y |q^+ P(y) + q^- P(y)| \\ &= \sum_{y \in S^+} [q^+ P(y) + q^- P(y)] - \sum_{y \in S^-} [q^+ P(y) + q^- P(y)] \end{aligned}$$

and continuing

$$\begin{aligned}
&= q^+ \sum_{y\in S^+} P + q^- \sum_{y\in S^+} P - q^+ \sum_{y\in S^-} P - q^- \sum_{y\in S^-} P \\
&= q^+(\sum_{y\in S} P - 2\sum_{y\in S^-} P) - q^-(\sum_{y\in S} P - 2\sum_{y\in S^+} P) \\
&\leq ||q^+||(1-2\frac{\delta N^-}{N}) + ||q^+||(1-2\frac{\delta N^+}{N}) \\
&= ||q^+||(2-2\delta) = 2||q^+||(1-\delta) \\
&= ||q||(1-\delta)
\end{aligned}$$

which gives what we wanted. Here $S^\pm$ denotes the set of states where the entries in the sum are positive and negative, respectively, and $N^\pm$ is the size of these sets.

To show that $\pi = \pi P$ has a unique invariant probability vector solution, we define the sequence $\pi_n = \pi_{n-1}P$, with π_0 any probability vector. We have that

$$||\pi_n - \pi_{n-1}|| = ||(\pi_{n-1} - \pi_{n-2})P|| < \rho||\pi_{n-1} - \pi_{n-2}||$$

where $\rho = 1 - \delta < 1$. Iterating backwards we have that

$$||\pi_n - \pi_{n-1}|| < \rho^{n-1}||\pi_1 - \pi_0||$$

From this we conclude that the probability vectors π_n are a Cauchy sequence

$$\sup_m ||\pi_{n+m} - \pi_n|| \to 0$$

as $n \to \infty$ and therefore π_n has a limit π, which is a probability vector. Passing to the limit in $\pi_n = \pi_{n-1}P$ we see that π is an invariant probability vector and since it is the unique limit of a Cauchy sequence it must be the unique invariant probability vector. We also have exponential convergence

$$||\pi_0 P^n - \pi|| = ||\pi_0 P^n - \pi P^n|| < \rho^n||\pi_0 - \pi||$$

4.5 The ergodic theorem

For a Markov chain in a finite state space the ergodic theorem is valid under the hypotheses and results of the previous section, and gives an important interpretation of the invariant probabilities $\pi(x)$. We have that

$$\frac{1}{n}\sum_{j=0}^{n-1} \mathbb{1}\{X_j = z\} \to \pi(z)$$

in mean square as $n \to \infty$, for any $z \in S$ and any initial probability vector $P\{X_0 = x\} = \pi_0(x)$. In words, the invariant probability vector is the limit in mean square of the relative time spent in state z as $n \to \infty$, and it is independent of the initial state.

More generally, for any function $f(x)$ on the state space we have

$$\frac{1}{n}\sum_{j=0}^{n-1} f(X_j) \to \sum_{y\in S} \pi(y)f(y) = \pi f$$

in mean square as $n \to \infty$ for any initial probability vector $P\{X_0 = x\} = \pi_0(x)$.

If we take expectations on the left we have

$$E_x\{\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)\} = \frac{1}{n}\sum_{j=0}^{n-1} E_x\{f(X_j)\} = \frac{1}{n}\sum_{j=0}^{n-1} P^j f(x)$$

If we let $\pi_0 = \delta_x$, the probability vector concentrated at x, we have

$$E_x\{\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)\} = \frac{1}{n}\sum_{j=0}^{n-1} \pi_0 P^j f \to \pi f$$

by the results of the previous section. We want to show here that not only the means converge but the random time averages converge in mean square to the average with respect to the invariant probability.

The main step in proving the ergodic theorem is the introduction of a new function $\chi(x)$ that converts approximately the time average to an expression that is easy to handle because it is a martingale. This function is the solution of the Poisson equation (by analogy with PDEs)

$$(P - I)\chi = -f + \pi f$$

More explicitly, this system of equations for $\chi = (\chi(x))_{x\in S}$ has the form

$$\chi(x) = f(x) - \pi f + \sum_{y\in S} P(x, y)\chi(y), \quad x \in S$$

Note that the right hand side of the system $(P - I)\chi = -f + \pi f$ has mean zero, or inner product zero, with respect to π: $\pi(f - \pi f) = 0$. This then is a necessary condition for the solvability of the Poisson equation since the π average of the left side is always zero, for any χ:

$$\pi(P - I)\chi = 0\ , \quad \text{or} \quad \sum_{x\in S} \pi(x) \left(\sum_{y\in S} P(x, y)\chi(y) - \chi(x) \right) = 0$$

Of course, the Poisson equation does not have a unique solution since $\mathbf{1}$ is an invariant right vector: $(P - I)\mathbf{1} = 0$. Thus $\chi + c\mathbf{1}$ is a solution for any constant c.

Without loss of generality we may assume that f is such that $\pi f = 0$ as we may replace f by $f - \pi f$. We now show that the Poisson equation has a solution. If there is a solution, we can write

$$\chi = (I - P)^{-1} f = \sum_{n=0}^{\infty} P^n f$$

The sum is, however, convergent because

$$||P^n f|| = ||(P^n - \pi) f|| < \rho^n 2 ||f||$$

To see this, we let π_x be the probability row vector that is equal to one at x and zero elsewhere. We then have that $P^n f(x) = \pi_x P^n f$ and hence $||P^n f - \pi f|| = \max_x |\pi_x P^n f - \pi f| \leq \max_x ||\pi_x P^n - \pi|| ||f|| \leq \rho^n 2 ||f||$ by the results of the previous section.

Once we have a fixed solution $\chi(x)$ we form the collapsing (telescoping) sum

$$\begin{aligned}
\chi(X_n) - \chi(X_0) &= \sum_{j=0}^{n-1} (\chi(X_{j+1}) - \chi(X_j)) \\
&= \sum_{j=0}^{n-1} [\chi(X_{j+1}) - E\{\chi(X_{j+1})|X_j\} + (E\{\chi(X_{j+1})|X_j\} - \chi(X_j)] \\
&= \sum_{j=0}^{n-1} [\chi(X_{j+1}) - P\chi(X_j)] + \sum_{j=1}^{n} [P\chi(X_j) - \chi(X_j]
\end{aligned}$$

where we use the notation $E\{\chi(X_{j+1})|X_j\} = P\chi(X_j)$. Using the Poisson equation that χ satisfies, rearranging and dividing by n we have

$$\frac{1}{n} \sum_{j=0}^{n-1} f(X_j) = \frac{1}{n} (\chi(X_n) - \chi(X_0)) + \frac{1}{n} M_n \tag{21}$$

where

$$M_n = \sum_{j=0}^{n-1} [\chi(X_{j+1}) - P\chi(X_j)], \ n = 1, 2, \ldots \tag{22}$$

This representation of the time average whose limit we want is important because, up to the first term on the right that goes to zero as $n \to \infty$, the second term is a **martingale**, M_n, divided by n. The defining property of a martingale is

$$E\{M_n | X_0, X_1, ..., X_{n-1}\} = M_{n-1}, \ n = 1, 2, \ldots, \text{ with } M_0 = 0,$$

which is easily verified since the conditional expectation of the last term in the sum for M_n is zero. We also have that $E_x\{M_n\} = 0$.

Now the proof can be completed by noting that the variance of M_n has the form

$$E_x\{M_n^2\} = E_x\{\sum_{j=0}^{n-1}[\chi(X_{j+1}) - P\chi(X_j)]^2\}$$

because cross terms have mean zero, as they are for sums of zero mean independent (or only uncorrelated) random variables. This is the essential point for introducing and using χ to get M_n: we can calculate variances as if we were dealing with sums of independent, mean zero random variables. We now conclude that

$$\frac{1}{n^2}E_x\{M_n^2\} \leq \frac{\text{constant}}{n}$$

which implies the result we want:

$$E_x\{[\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)]^2\} \to 0$$

in the case $\pi f = 0$, which we have assumed as noted above.

4.6 The central limit theorem for Markov chains

Let

$$\sigma^2(x) = E_x\{(\chi(X_1) - P\chi(x))^2\} = P(\chi - P\chi)^2(x) \tag{23}$$

Under the same conditions for which we have proved the ergodic theorem in the previous section we also have a central limit theorem as follows. The scaled difference

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=0}^{n-1} f(X_j) - \pi f\right) \to N(0, \pi\sigma^2) \ , \quad n \to \infty \tag{24}$$

in distribution, where in particular

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\sigma^2(X_j) = \pi\sigma^2 = \sum_{x\in S}\pi(x)\sigma^2(x) \tag{25}$$

in mean square, by the ergodic theorem. In fact this CLT simply follows from the more general "martingale central limit theorem" in view of the representation of the time average of f that we obtained in the previous section in terms of an asymptotically small term and a scaled martingale.

Given the representation (21) with the martingale M_n defined by (22) it is enough to show that for any $\alpha \in \mathbb{R}$ we have that

$$\lim_{n\to\infty} E_x\{e^{i\frac{\alpha}{\sqrt{n}}M_n}\} = e^{-\frac{\alpha^2}{2}\pi\sigma^2}, \tag{26}$$

which proves the CLT (24) through the limit of characteristic functions. The ergodic theorem is used here in (25) as already noted. We also note that whereas the solution of the Poisson equation χ played a role only in streamlining the proof of the ergodic theorem, in the CLT it plays a basic role as it enters into the form of the limit variance. There are other, equivalent forms for the limit variance $\pi\sigma^2$, in the form of sums which correspond to expansions of χ in the previous Section.

We now prove the martingale CLT (26) by using a simple identity and a Taylor expansion. The identity is as follows. For any complex numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ we have

$$\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k = \sum_{j=1}^{n} \prod_{k=j+1}^{n} a_k (a_j - b_j) \prod_{k=1}^{j-1} b_k$$

with the convention that when the indices in the products are off then the product equals one. This identity holds not only for real or complex numbers but also for matrices or operators provided order in the products is respected. Now we note that

$$e^{i\frac{\alpha}{\sqrt{n}} M_n + \frac{\alpha^2}{2n}\sum_{k=1}^{n}\sigma^2(X_{k-1})} = \prod_{k=1}^{n} e^{i\frac{\alpha}{\sqrt{n}}(M_k - M_{k-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{k-1})}$$

Applying the identity we get

$$1 - e^{i\frac{\alpha}{\sqrt{n}} M_n + \frac{\alpha^2}{2n}\sum_{k=1}^{n}\sigma^2(X_{k-1})} = 1 - \prod_{k=1}^{n} e^{i\frac{\alpha}{\sqrt{n}}(M_k - M_{k-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{k-1})}$$

$$= \sum_{j=1}^{n} \left(1 - e^{i\frac{\alpha}{\sqrt{n}}(M_j - M_{j-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{j-1})}\right) \prod_{k=1}^{j-1} e^{i\frac{\alpha}{\sqrt{n}}(M_k - M_{k-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{k-1})}$$

Taking expectations first and then absolute values we have

$$\left| E_x \left\{ e^{i\frac{\alpha}{\sqrt{n}} M_n + \frac{\alpha^2}{2n}\sum_{k=1}^{n}\sigma^2(X_{k-1})} - 1 \right\} \right|$$

$$\leq \sum_{j=1}^{n} \left| E_x \left\{ \left(1 - e^{i\frac{\alpha}{\sqrt{n}}(M_j - M_{j-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{j-1})}\right) \prod_{k=1}^{j-1} e^{i\frac{\alpha}{\sqrt{n}}(M_k - M_{k-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{k-1})} \right\} \right|$$

Using an iterated conditional expectation inside the sum, given $\{X_0, X_1, \ldots X_{j-1}\}$, we get that the right hand side in the inequality becomes

$$\leq C_\alpha \sum_{j=1}^{n} E_x \left\{ \left| E \left\{ \left(1 - e^{i\frac{\alpha}{\sqrt{n}}(M_j - M_{j-1}) + \frac{\alpha^2}{2n}\sigma^2(X_{j-1})}\right) | X_0, X_1, \ldots X_{j-1} \right\} \right| \right\}$$

where C_α is a constant depending on α but not on n.

We now do a Taylor expansion of the terms in the conditional expectation. We note that

$$E\{e^{i\frac{\alpha}{\sqrt{n}}(M_j-M_{j-1})+\frac{\alpha^2}{2n}\sigma^2(X_{j-1})}|X_0,X_1,\dots X_{j-1}\}$$

$$=1-\frac{\alpha^2}{2n}E\{(M_j-M_{j-1})^2|X_0,X_1,\dots X_{j-1}\}+\frac{\alpha^2}{2n}\sigma^2(X_{j-1})+O(n^{-3/2})=O(n^{-3/2})$$

Therefore the error on the right side of the above inequality is of order $n^{-1/2}$ as $n\to\infty$ and we have shown that

$$\lim_{n\to\infty}E_x\left\{e^{i\frac{\alpha}{\sqrt{n}}M_n+\frac{\alpha^2}{2n}\sum_{k=1}^n\sigma^2(X_{k-1})}\right\}=1$$

This now combined with the ergodic theorem (25) for the variance proves the CLT for the martingale M_n and hence the CLT for the Markov chain (24). More explicitly, we have that $\frac{1}{n}\sum_{k=1}^n\sigma^2(X_{k-1})\to\pi\sigma^2$ in probability and

$$\left|E_x\left\{e^{i\frac{\alpha}{\sqrt{n}}M_n+\frac{\alpha^2}{2n}\sum_{k=1}^n\sigma^2(X_{k-1})}\right\}-E_x\left\{e^{i\frac{\alpha}{\sqrt{n}}M_n+\frac{\alpha^2}{2}\pi\sigma^2}\right\}\right|$$

$$=\left|E_x\left\{e^{i\frac{\alpha}{\sqrt{n}}M_n}\left(e^{\frac{\alpha^2}{2n}\sum_{k=1}^n\sigma^2(X_{k-1})}-e^{\frac{\alpha^2}{2}\pi\sigma^2}\right)\right\}\right|$$

$$\le E_x\left\{\left|e^{\frac{\alpha^2}{2n}\sum_{k=1}^n\sigma^2(X_{k-1})}-e^{\frac{\alpha^2}{2}\pi\sigma^2}\right|\right\}\to 0,\ n\to\infty$$

by the bounded convergence theorem. We have therefore shown that

$$\lim_{n\to\infty}E_x\left\{e^{i\frac{\alpha}{\sqrt{n}}M_n}\right\}=e^{-\frac{\alpha^2}{2}\pi\sigma^2}$$

which is the CLT via characteristic functions.

4.7 Expected number of visits to a state and the invariant probabilities

Let x,z and y be three states in S, not necessarily all different, and consider the expected number of visits to z before reaching y, starting from x:

$$u(x;z,y)=E_x\{\sum_{n=1}^{\infty}\mathbb{1}\{X_n=z\}\mathbb{1}\{T_y\ge n\}\}=E_x\{\sum_{n=1}^{T_y}\mathbb{1}\{X_n=z\}\},$$

where T_y is the first time to reach y

$$T_y=\inf\{n\ge 1\mid X_n=y\}$$

This is a stopping time and $\mathbb{1}_{\{T_y \geq n\}}$ depends, or is a function of, only $\{X_0, X_1, \ldots, X_{n-1}\}$. When we start at y then with the current definition T_y is the first time to return to y. Clearly

$$\sum_z u(x; z, y) = E_x\{T_y\}$$

which is the expected time to reach y. When $y = x$ then $\sum_z u(x; z, x)$ is the expected time to return to x, after starting from it. For any bounded function f on S we have that

$$u_f(x, y) = \sum_{z \in S} u(x; z, y) f(z) = E_x\{\sum_{n=1}^{T_y} f(X_n)\}$$

To get a recursion relation for u, **as a function of** z and not the starting point x as in Section 4.1, we write

$$\begin{aligned} u(x; z, y) &= E_x\{\sum_{n=1}^{\infty} \mathbb{1}\{X_n = z\} \sum_w \mathbb{1}\{X_{n-1} = w\} \mathbb{1}\{T_y \geq n\}\} \\ &= \sum_w \sum_{n=1}^{\infty} E_x\{E\{\mathbb{1}\{X_n = z\} \mathbb{1}\{X_{n-1} = w\} \mathbb{1}\{T_y \geq n\} | X_0, X_1, ..., X_{n-1}\}\} \end{aligned}$$

using iterated conditional expectation. Using the properties of the stopping time and the Markov property we have

$$\begin{aligned} &\sum_w \sum_{n=1}^{\infty} E_x\{E\{\mathbb{1}\{X_n = z\} \mathbb{1}\{X_{n-1} = w\} \mathbb{1}\{T_y \geq n\} | X_0, X_1, ..., X_{n-1}\}\} \qquad (27) \\ = &\sum_w P(w, z) \sum_{n=1}^{\infty} E_x\{\mathbb{1}\{X_{n-1} = w\} \mathbb{1}\{T_y \geq n\}\} \end{aligned}$$

We also have that

$$\begin{aligned} &\sum_w P(w, z) E_x\{\sum_{n=1}^{\infty} \mathbb{1}\{X_{n-1} = w\} \mathbb{1}\{T_y \geq n\}\} \\ = &\sum_w P(w, z) E_x\{\sum_{n=1}^{T_y} \mathbb{1}\{X_{n-1} = w\}\} \end{aligned} \qquad (28)$$

and by rearrangements we have

$$\begin{aligned}
& \sum_w P(w,z) E_x\{\sum_{n=1}^{T_y} \mathbb{1}\{X_{n-1}=w\}\} \\
= & \sum_w P(w,z) E_x\{\sum_{n=0}^{T_y-1} \mathbb{1}\{X_n=w\}\} \\
= & \sum_w P(w,z) \left[E_x\{\sum_{n=1}^{T_y} \mathbb{1}\{X_n=w\}\} + \delta_{xw} - \delta_{yw} \right] \\
= & \sum_w u(x;w,y) P(w,z) + P(x,z) - P(y,z) \qquad (29)
\end{aligned}$$

where δ_{xw} equals one when $x = w$ and zero otherwise. In vector-matrix form we have, for each fixed starting state x and ending state y, the following system for $u = u(x;z,y)$ as a row vector in z

$$u(I-P) = \pi_x P - \pi_y P$$

where π_x is the probability vector with component equal to one at $z = x$ and zero otherwise and similarly for π_y. This system is always solvable since $(\pi_x P - \pi_y P)\mathbf{1} = 0$, that is, the right hand side is orthogonal to the right null-vector of $I - P$, which is $\mathbf{1}$. When $x = y$, the right hand side is zero and we see that, in this case, u is proportional to the invariant probability vector π

$$\pi(z) = \frac{u(x;z,x)}{\sum_z u(x;z,x)} = \frac{u(x;z,x)}{E_x\{T_x\}} = \frac{E_x\{\sum_{n=1}^{T_x} \mathbb{1}\{(X_n=z\})\}}{E_x\{T_x\}}$$

In words, the expected number of visits to z between successive visits to x divided by the expected number of time steps between successive visits to x is the invariant probability $\pi(z)$. Furthermore, when $x = z$ then there is only one visit to x after leaving it, between successive visits, and so

$$\pi(x) = \frac{1}{E_x\{T_x\}}$$

4.8 Return times and the ergodic theorem

Consider again an ergodic Markov chain on a finite set S, with transition probabilities $P(x,y) > 0$. Fix a state x and let $T_1, T_2, T_3, \ldots$ be the successive return times to x, starting from it. We show first that they are independent, identically distributed random variables. We use the ergodic theorem for the Markov chain to show that the mean of the return time to x is equal to the reciprocal of the invariant probability $\pi(x)$, just as it was shown in the previous section without the ergodic theorem.

We show independence for the first two return times. Let

$$F_n(x) = P\{T_1 = n | X_0 = x\} = P\{X_n = x, X_{n-1} \neq x, \ldots X_1 \neq x | X_0 = x\}$$

We then have

$$\begin{aligned}
& \mathbb{P}\{T_2 = m, T_1 = n | X_0 = x\} \\
= \; & \mathbb{P}\{X_{n+m} = x, X_{n+m-1} \neq x, \ldots, X_{n+1} \neq x, X_n = x, X_{n-1} \neq x, \ldots X_1 \neq x | X_0 = x\} \\
= \; & \mathbb{P}\{X_{n+m} = x, X_{n+m-1} \neq x, \ldots, X_{n+1} \neq x | X_n = x, X_{n-1} \neq x, \ldots X_1 \neq x, X_0 = x\} \\
& \quad \times \mathbb{P}\{X_n = x, X_{n-1} \neq x, \ldots X_1 \neq x | X_0 = x\} \\
= \; & \mathbb{P}\{X_{n+m} = x, X_{n+m-1} \neq x, \ldots, X_{n+1} \neq x | X_n = x\} F_n(x) \text{ by the Markov property} \\
= \; & \mathbb{P}\{X_m = x, X_{m-1} \neq x, \ldots, X_1 \neq x | X_0 = x\} F_n(x) \text{ by time homogeneity} \\
= \; & F_m(x) F_n(x)
\end{aligned}$$

We can also write this as

$$\begin{aligned}
& \mathbb{P}\{T_2 = m, T_1 = n | X_0 = x\} \\
= \; & \mathbb{P}\{X_{n+m} = x, X_{n+m-1} \neq x, \ldots, X_{n+1} \neq x | X_n = x\} F_n(x) \text{ Markov property} \\
= \; & \mathbb{P}\{T_1 + T_2 = n + m | T_1 = n, X_n = x\} \mathbb{P}\{T_1 = n | X_0 = x\} \\
= \; & \mathbb{P}\{T_2 = m | X_{T_1} = x\} \mathbb{P}\{T_1 = n | X_0 = x\}
\end{aligned}$$

from where we see that $\mathbb{P}\{T_2 = m | X_{T_1} = x\} = F_m(x)$. Both the Markov property and the time homogeneity have been used, and it has been shown in this case that the strong Markov property holds since the conditioning involves a stopping time. Similarly, $T_1, T_2, T_3, \ldots$ are independent random variables. Their distribution is identical because of the time homogeneity of the Markov chain.

We argue that we can replace $n-1$ by $\sum_{k=1}^{\nu} T_k$ in the ergodic theorem. That is,

$$\lim_{\nu \to \infty} \frac{1}{1 + \sum_{k=1}^{\nu} T_k} \sum_{n=0}^{\sum_{k=1}^{\nu} T_k} f(X_n) = \pi f$$

in mean square. Setting $f(y) = \mathbb{1}_{\{y=x\}}$ we obtain

$$\lim_{\nu \to \infty} \frac{1}{\frac{1}{\nu} \sum_{k=1}^{\nu} T_k} = \pi(x).$$

But by the law of large numbers the left hand side is just $1/\mathbb{E}T_1$. To complete the proof we need to show that we can replace the index $\nu \to \infty$ in the ergodic theorem by $\tau_\nu = \sum_{k=1}^{\nu} T_k \to \infty$, in probability as $\nu \to \infty$. But $\frac{\tau_\nu}{\nu} - \mu \to 0$ in mean square (hence in

probability) where $\mu = E\{T_1\} > 0$ by the standard law of large numbers. In more detail, we need to show that

$$\left| \frac{1}{1+\tau_\nu} \sum_{n=0}^{\tau_\nu} f(X_n) - \frac{1}{1+[\mu\nu]} \sum_{n=0}^{[\mu\nu]} f(X_n) \right| \to 0$$

in probability as $\nu \to \infty$. Here $[a]$ denotes the integer part of a. But we have the estimate

$$\left| \frac{1}{1+\tau_\nu} \sum_{n=0}^{\tau_\nu} f(X_n) - \frac{1}{1+[\mu\nu]} \sum_{n=0}^{[\mu\nu]} f(X_n) \right| \leq 2 \left| \frac{1+[\mu\nu]}{1+\tau_\nu} - 1 \right| ||f||$$

and since $\frac{1+\tau_\nu}{1+[\mu\nu]} \to 1$ in probability, we also have that $\frac{1+[\mu\nu]}{1+\tau_\nu} \to 1$ in probability, which completes the proof.

4.9 MLE for Markov chains

Let $\{X_n, n \geq 0\}$ be a Markov chain in a finite state space S with transition probability $P(x, y; \theta)$, where θ is a real valued parameter. The goal is to estimate θ using observations $\{X_i, 0 \leq i \leq n\}$. We assume that the Markov chain is ergodic, uniformly in θ in some fixed interval of interest, and that we have a positive lower bound for $P(x, y; \theta)$. In this section we will carry out the following.

1. Let $Y_n = (X_n, X_{n-1})^T, n \geq 1$, which is also a Markov chain. We will show that the transition probability for Y_n is given by $Q(x_1, x_2; y_1, y_2) = P\{Y_n = (y_1, y_2)^T | Y_{n-1} = (x_1, x_2)^T\} = \mathbb{1}_{\{x_1 = y_2\}} P(y_2, y_1)$.

2. Let π be the invariant probability vector for X_n: $\sum_{x \in S} \pi(x) P(x, y) = \pi(y)$. We will show that an invariant probability vector for Y_n has components $\pi(x_2) P(x_2, x_1)$. By calculating Q^2, the two-step transition probabilities, we conclude that this invariant vector is unique and is approached exponentially fast. We in fact reduce this problem to the standard case analyzed in earlier sections.

3. Suppose the initial state x_0 is given. Moreover suppose θ^* is the true value of the underlying parameter. We use the ergodic theorem, to show that as $n \to \infty$ the normalized log likelihood function

$$l_n(\theta) = \frac{1}{n} \log[\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n; \theta)]$$

$$\to \sum_{x,y \in S} \pi(x; \theta^*) P(x, y; \theta^*) \log P(x, y; \theta) := \ell(\theta).$$

in probability.

4. Assume $P(x, y; \theta)$ is twice-differentiable with respect to θ. We show that $l'(\theta^*) = 0$ and $l''(\theta^*) < 0$.

5. We also use the delta method (as in the i.i.d. case) and the central limit theorem for Markov chains to get a CLT for the MLE $\hat{\theta}_n$ (where $l'_n(\hat{\theta}_n) = 0$).

We now go on to analyze the statements made above.

1. By the definition of conditional probability

$$Q(x_1, x_2; y_1, y_2) = P\{Y_n = (y_1, y_2)^T | Y_{n-1} = (x_1, x_2)^T\} = \frac{P\{Y_n = (y_1, y_2)^T,\ Y_{n-1} = (x_1, x_2)^T\}}{P\{Y_{n-1} = (x_1, x_2)^T\}}$$

$$= \frac{\pi_{n-2}(x_2)P(x_2, x_1)\mathbb{1}_{\{x_1=y_2\}}P(y_2, y_1)}{\pi_{n-2}(x_2)P(x_2, x_1)} = \mathbb{1}_{\{x_1=y_2\}}P(y_2, y_1)$$

2. We verify that

$$\pi(y_2)P(y_2, y_1) = \sum_{x_1, x_2} \pi(x_2)P(x_2, x_1)Q(x_1, x_2; y_1, y_2) = \sum_{x_1, x_2} \pi(x_2)P(x_2, x_1)\mathbb{1}_{\{x_1=y_2\}}P(y_2, y_1)$$

   which is a true relation because $\sum_x \pi(x)P(x, y) = \pi(y)$. We also have that

$$Q^2(x_1, x_2; y_1, y_2) = \sum_{z_1, z_2} Q(x_1, x_2; z_1, z_2)Q(z_1, z_2; y_1, y_2) = P(x_1, y_2)P(y_2, y_1) > 0$$

   the positivity being for all (finitely many) states and so the hypothesis for the ergodic theorem holds for $\{Y_n\}$.

3. Clearly we have by the ergodic theorem, which we have shown that it applies,

$$l_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \log P(X_{j-1}, X_j; \theta) + \frac{1}{n}\log \pi_0(X_0)$$

$$\rightarrow_{\{n\to\infty\}} \sum_{x,y\in S} \pi(x; \theta^*)P(x, y; \theta^*) \log P(x, y; \theta) = l(\theta)$$

   where we indicate dependence on the parameter(s) explicitly.

4. Differentiability of $l(\theta)$ with respect to θ follows from the uniformity of the ergodic limit with respect to θ. That is, the convergence is in mean square (or in probability), uniformly with respect to the parameter. We can then do differentiation. We have

$$l'(\theta) = \sum_{x,y\in S} \pi(x; \theta^*)P(x, y; \theta^*)\frac{P'(x, y; \theta)}{P(x, y; \theta)}$$

where prime denotes derivative with respect to θ, assumed a scalar parameter in $[\theta_1, \theta_2]$ and with $\theta^* \in (\theta_1, \theta_2)$. At $\theta = \theta^*$ we have

$$l'(\theta^*) = \sum_{x,y \in S} \pi(x;\theta^*) P'(x,y;\theta^*)$$

But for any θ we have $\sum_{x,y} \pi(x;\theta) P(x,y;\theta) = 1$ and so differentiating we get

$$0 = \sum_{x,y} \pi'(x;\theta) P(x,y;\theta) + \sum_{x,y} \pi(x;\theta) P'(x,y;\theta)$$

$$= \sum_{x} \pi'(x;\theta) + \sum_{x,y} \pi(x;\theta) P'(x,y;\theta) = \sum_{x,y} \pi(x;\theta) P'(x,y;\theta)$$

which implies that $l'(\theta^*) = 0$. In the same way we can calculate

$$l''(\theta) = \sum_{x,y \in S} \pi(x;\theta^*) P(x,y;\theta^*) \left(\frac{P''(x,y;\theta)}{P(x,y;\theta)} - \frac{(P'(x,y;\theta))^2}{P^2(x,y;\theta)} \right)$$

which leads to

$$l''(\theta^*) = - \sum_{x,y \in S} \pi(x;\theta^*) \frac{(P'(x,y;\theta^*))^2}{P(x,y;\theta^*)} = -I < 0$$

with I the Fisher information.

5. The maximum likelihood estimator of $\hat{\theta}_n$ of θ is defined as the maximizer of $l_n(\theta)$. From the ergodic theorem we know that $l_n(\theta)$ is close in mean square to $l(\theta)$, which is concave near the true value θ^*. As in the iid case in Section 2.4, what is needed (and not given here) is a uniform in θ ergodic theorem and then the iid argument extends to Markov chains. Intuitively, however, we expect that if θ is close enough to θ^* and n is large enough, then $l_n(\theta)$ will be concave with high probability. This means that $\hat{\theta}_n$ satisfies $l_n'(\hat{\theta}_n) = 0$ and this is close to $l(\theta^*)$ so that we have $P\{|\hat{\theta}_n - \theta^*| > \delta\} \to 0$ as $n \to \infty$ for any $\delta > 0$. Define the fluctuation error Z_n by

$$\hat{\theta}_n = \theta^* + \frac{1}{\sqrt{n}} Z_n$$

To get a CLT for the error, as in the iid case, we use the delta method

$$0 = l_n'(\hat{\theta}_n) = l_n'(\theta^* + \frac{1}{\sqrt{n}} Z_n) = l_n'(\theta^*) + l_n''(\theta^*) \frac{1}{\sqrt{n}} Z_n + \dots$$

where the dots signify terms that go to zero faster than $1/\sqrt{n}$ in mean square or in probability (but can be eliminated by using the mean value theorem as in section 2.1). Ignoring the higher order terms we get (approximately)

$$Z_n = \frac{\sqrt{n} l_n'(\theta^*)}{-l_n''(\theta^*)}$$

By differentiation we have that

$$l_n'(\theta^*) = \frac{1}{n}\sum_{j=1}^{n} \frac{P'(X_{j-1}, X_j; \theta^*)}{P(X_{j-1}, X_j; \theta^*)}$$

$$l_n''(\theta^*) = \frac{1}{n}\sum_{j=1}^{n} \left(\frac{P''(X_{j-1}, X_j; \theta^*)}{P(X_{j-1}, X_j; \theta^*)} - \frac{(P'(X_{j-1}, X_j; \theta^*))^2}{P^2(X_{j-1}, X_j; \theta^*)} \right)$$

where we have ignored the π_0 term that goes to zero in the limit. The ergodic theorem applies to $l_n''(\theta^*)$, which tends to $l''(\theta^*) = -I$ in mean square (or in probability). The next step is to apply the central limit theorem to $\sqrt{n} l_n'(\theta^*)$, if possible, where

$$\sqrt{n} l_n'(\theta^*) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{P'(X_{j-1}, X_j; \theta^*)}{P(X_{j-1}, X_j; \theta^*)}$$

As discussed earlier in these notes, we need to first transform the quantity of interest using the solution χ of a suitable Poisson equation. We then have this quantity expressed as a martingale plus a term that goes to zero in the limit. The key observation here, however, is that $l_n'(\theta^*)$ **is a martingale already**, with zero mean. In fact, we can verify the defining property of a martingale

$$E\{l_n'(\theta^*) | X_0, X_1 \cdots X_{n-1}\} = l_{n-1}'(\theta^*)$$

which clearly follows from the fact that

$$E_{\theta^*}\left\{ \frac{P'(X_{n-1}, X_n; \theta^*)}{P(X_{n-1}, X_n; \theta^*)} | X_{n-1} \right\} = \sum_y P'(X_{n-1}, y; \theta^*) = 0$$

Now for a zero mean martingale we have the key property, as is discussed in earlier sections,

$$\mathrm{var}_{\theta^*}(\sqrt{n} l_n'(\theta^*)) = E_{\theta^*}\left\{ \left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{P'(X_{j-1}, X_j; \theta^*)}{P(X_{j-1}, X_j; \theta^*)} \right)^2 \right\} = E_{\theta^*}\left\{ \frac{1}{n}\sum_{j=1}^{n} \left(\frac{P'(X_{j-1}, X_j; \theta^*)}{P(X_{j-1}, X_j; \theta^*)} \right)^2 \right\}$$

We can now apply the ergodic theorem (we only need convergence of the mean here) to the last sum and we see that $\mathrm{var}_{\theta^*}(\sqrt{n} l_n'(\theta^*)) \to I$, with I the Fisher information. The CLT now applied to the martingale $\sqrt{n} l_n'(\theta^*)$ tells us that it converges in distribution (in law, weakly) to an $N(0, I)$ random variable. Application of Slutsky's theorem leads to the CLT for the error of the MLE

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \to N(0, \frac{1}{I}) \text{ in distribution}$$

just as in the iid case. The Cramer-Rao lower bound applies here (as for any asymptotically consistent estimator) and we conclude that the MLE is asymptotically efficient, as we should expect. Thus all of the theory of MLE carries over to Markov chains. What is a lot more difficult now is the actual calculation of the MLE, $\hat{\theta}_n$, which can be done iteratively by the expectation minimization (EM) algorithm or by other optimization methods depending on the problem at hand.

4.10 Bayesian filtering

We want to determine recursive filtering equations in the Markov chain context. Let $X = \{X_n : n \geq 0\}$ be a Markov chain in a finite state-space S that is not observed directly, and let $Z = \{Z_n : n \geq 0\}$ be an observed, noise corrupted version of X defined via the path probability densities given the Markov chain:

$$\mathbb{P}\{Z_0 \in dz_0, Z_1 \in dz_1, \ldots, Z_n \in dz_n | X\} = \prod_{i=0}^{n} f(z_i; X_i) dz_i,$$

where $(f(\cdot; x) : x \in S)$ is a family of given density functions. The noisy observations are conditionally independent given the Markov chain. Let $\mu = (\mu(x) : x \in S)$ be a prior distribution of initial state X_0 (i.e $\mu(x) = \mathbb{P}(X_0 = x)$) and denote by

$$\mu_n(x) = \mathbb{P}(X_n = x | Z_0, \cdots, Z_n)$$

the posterior of the state at time n given the observations. We want to compute $\mu_{n+1}(x)$ recursively from $\mu_n(x)$ assuming that the transition probabilities $P(x, y),\ x, y \in S$, of the Markov chain are known.

Let $Z_{(n)} = \{Z_0, Z_1, \ldots, Z_n\}$ be the observed noisy Markov chain up to time n. From the properties of conditional probabilities we have that

$$\mu_n(x) = \mathbb{P}(X_n = x | Z_{(n)}) = \frac{P\{X_n = x, Z_{(n)}\}}{P\{Z_{(n)}\}} = \frac{P\{Z_n | X_n = x, Z_{(n-1)}\} P\{X_n = x, Z_{(n-1)}\}}{P\{Z_{(n)}\}}$$

where $P\{Z_{(n)}\}$ is the joint (unconditional) density of the noisy Markov chain evaluated at the observation $Z_{(n)}$. But by the law of total probability and the Markov property we have

$$\begin{aligned} P\{X_n = x, Z_{(n-1)}\} &= \sum_{z \in S} P\{X_n = x | X_{n-1} = z, Z_{(n-1)}\} P\{X_{n-1} = z, Z_{(n-1)}\} \\ &= \sum_{z \in S} P\{X_n = x | X_{n-1} = z\} P\{X_{n-1} = z, Z_{(n-1)}\} \end{aligned}$$

and

$$P\{X_{n-1} = z, Z_{(n-1)}\} = P\{X_{n-1} = z | Z_{(n-1)}\} P\{Z_{(n-1)}\} = \mu_{n-1}(z) P\{Z_{(n-1)}\}$$

Note also that

$$P\{Z_n|X_n = x, Z_{(n-1)}\} = P\{Z_n|X_n = x\} = f(Z_n; x)$$

by the independence of the noisy observations given the Markov chain.

We can now see how the recursion goes. We first update the state with the (assumed known) Markov chain transition probabilities

$$\mu_{n-1} \to \mu_{n|n-1}(x) = \sum_{z \in S} \mu_{n-1}(z) P(z, x) \ , \quad \mu_0(x) \ \text{ given.}$$

Then we update the observation

$$\mu_{n|n-1} \to \mu_n(x) = \frac{f(Z_n, x)\mu_{n|n-1}(x)}{\sum_{x \in S} f(Z_n; x)\mu_{n|n-1}(x)}$$

In applications there are two main limitations to this algorithm. First, the transition probabilities of the Markov chain are not known perfectly. They usually have unknown parameters in them, which must be estimated. Second, the updating-of-the-state step will be computationally intensive when the dimension of the Markov chain is large so Monte Carlo methods (particle methods) must be used. Combining filtering and maximum likelihood parameter estimation can be done with variants of the expectation-maximization (EM) algorithm.

5 Random walks and connections with differential equations

We begin with the simple random walk on equally spaced points in an interval. Let $\{X_n\}$ denote the symmetric random walk on $S = \{x_0 = 0, x_1 = \Delta x, \ldots, x_N = N\Delta x = a\}$ so that

$$P\{X_n = x_k | X_{n-1} = x_j\} = 1/2 \text{ when } k = j \pm 1 \text{ and 0 otherwise}$$

where x_j is an interior point, that is, it is not 0 or a. The process can be absorbed at 0 or a, in which case

$$P\{X_n = x_0 | X_{n-1} = x_0\} = 1 \ , \quad P\{X_n = x_N | X_{n-1} = x_N\} = 1$$

The matrix of transition probabilities $P(x, y)$ has the form

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1/2 & 0 & 1/2 & 0 & \cdots & 0 \\ & & & \cdots & & \\ 0 & \cdots & 0 & 1/2 & 0 & 1/2 \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix}$$

The process satisfies the stochastic difference equation $X_n = X_{n-1} + Z_n$ for $n = 1, 2, ...$, with $X_0 = x$ given, for example, and with $\{Z_n\}$ independent identically distributed random variables taking values $\pm\Delta x$ with probability 1/2. The process stops when it reaches the boundary points.

We can derive difference equations for probabilities of interest such as

$$u_j^n = P\{T > n | X_0 = x_j\}$$

where T is the first time to reach x_0 or x_N, the exit time from the interval. By the renewal method we find that

$$u_j^n = \frac{1}{2}(u_{j+1}^{n-1} + u_{j-1}^{n-1}) \ , \quad j = 1, 2, \ldots, N-1 \ , \quad n = 1, 2, 3, \ldots$$

with boundary conditions $u_0^n = u_N^n = 0$, $n = 0, 1, 2, \ldots$, and initial condition $u_j^0 = 1$, $j = 1, 2, \ldots, N-1$. In vector-matrix form this finite difference equation has the form

$$u^n = Pu^{n-1} \ , \quad n = 1, 2, 3, ...$$

where $u^n = (u_j^n)$ and where $u^0 = f$ with f equal to one at all interior points and zero at the two boundary points. The eigenvalues and eigenvectors of this symmetric tridiagonal matrix can be computed explicitly using, for example, the discrete Fourier transform.

The connection with partial differential equations comes from passing to the continuum limit, which we will do here formally, without detailed proofs that resemble the usual convergence proofs for finite difference methods or can be entirely probabilistic since X_n is

just a sum of iid random variables and the diffusion approximation is simply a restatement of the CLT.

We write the recursion relation for u^n as

$$\frac{1}{\Delta t}(u^n - u^{n-1}) = \sigma^2 \frac{1}{(\Delta x)^2}(P - I)u^{n-1}$$

with $\sigma^2 = (\Delta x)^2/\Delta t$ remaining fixed at Δt and Δx tend to zero. In the continuum limit we assume that $u_j^n \approx u(n\Delta t, j\Delta x)$ with $u(t,x)$ a smooth function that will satisfy a partial differential equation. It is easily seen that the j-th entry of

$$[\frac{1}{(\Delta x)^2}(P - I)u^{n-1}]_j \approx \frac{1}{2}u_{xx}(n\Delta t, j\Delta x)$$

and therefore the continuum limit of the difference equation in the interior becomes

$$u_t(t,x) = \frac{\sigma^2}{2}u_{xx}(t,x) \ , \quad t > 0 \ , \quad x \in (0,a)$$

with boundary conditions $u(t,0) = u(t,a) = 0$ and initial conditions $u(0,x) = 1$. This equation can be solved by Fourier sine series and the result is

$$u(t,x) = P_x\{T > t\} = \sum_{k=0}^{\infty} \frac{2\sqrt{2}}{(2k+1)\pi a} e^{-(\frac{2k+1)\pi}{a})^2 \frac{\sigma^2 t}{2}} \sin(\frac{2k+1)\pi x}{a})$$

Here T is the exit time of Brownian motion, the path limit in law of the random walk, from the interval $(0,a)$. The most interesting feature of the explicit solution is the rate of decay at $t \to \infty$

$$\frac{1}{t}\log P_x\{T > t\} \to -\frac{\pi^2\sigma^2}{2a^2}$$

When we consider this problem in the semi-infinite interval $(0,\infty)$ we simply have to solve the pde

$$u_t(t,x) = \frac{\sigma^2}{2}u_{xx}(t,x) \ , \quad t > 0 \ , \quad x > 0$$

with $u(t,0) = 0$ and $u(0,x) = 1$. It is interesting to compute here for $\lambda > 0$

$$E_x\{e^{-\lambda T}\} = -\int_0^{\infty} e^{-\lambda t} u_t(t,x)dt \tag{30}$$

which is the Laplace transform of the density $-u_t(t,x)$ of the exit time T from the origin. The Laplace transform of u

$$\hat{u}(\lambda, x) = \int_0^{\infty} e^{-\lambda t} u(t,x)dt$$

satisfies the ordinary differential equation

$$\lambda\hat{u} - 1 = \frac{\sigma^2}{2}\hat{u}_{xx} \ , \quad \hat{u}(\lambda, 0) = 0$$

for which we get the unique bounded solution

$$\hat{u} = \frac{1}{\lambda}(1 - e^{-\frac{\sqrt{2\lambda}}{\sigma}x})$$

And since (by integration by parts in (30)) $E_x\{e^{-\lambda T}\} = 1 - \lambda\hat{u}$ we see that

$$E_x\{e^{-\lambda T}\} = e^{-\frac{\sqrt{2\lambda}}{\sigma}x}$$

This Laplace transform can be inverted and the explicit form of the density of T can be obtained. But, contrary to what we have in a finite interval where the large t behavior of the distribution is exponentially small, in the semi-infinite interval the mean exit time is infinite, $E_x\{T\} = \infty$. This can be shown by differentiating the Laplace transform of T with respect to λ and then letting λ tend to zero.

5.1 Transience and recurrence

An irreducible and aperiodic Markov chain on an infinite space, a random walk for example, is recurrent if the expected number of visits to a state is infinite and transient otherwise. We will show that the one-dimensional random walk on the infinite lattice is recurrent while in three dimensions it is transient. It is also recurrent in two dimensions. We will use Fourier series for the analysis.

The one-dimensional random walk is $X_n = X_{n-1} + Z_n$ where $\{Z_n\}$ are iid random variables with values $\pm\Delta x$ with probability 1/2, Therefore, for $k \in (\frac{-\pi}{\Delta x}, \frac{\pi}{\Delta x})$ we have

$$\begin{aligned} E_x\{e^{ikX_n}\} &= E_x\{E\{e^{ikZ_n}e^{ikX_{n-1}}|X_{n-1}\}\} \\ &= e^{ikx}\left(E\{e^{ikZ}\}\right)^n \\ &= e^{ikx}\left(\frac{1}{2}(e^{ik\Delta x} + e^{-ik\Delta x})\right)^n \\ &= e^{ikx}(\cos k\Delta x)^n \end{aligned}$$

The values of X_n are $m\Delta x$ where $m = 0, \pm 1, \pm 2, ...$ so that if $x = 0$ then

$$E_0\{e^{ikX_n}\} = \sum_m e^{ikm\Delta x}p_m(n)$$

where $p_m(n)$ is the probability that $X_n = m\Delta x$, starting from 0. By the orthogonality of the complex exponentials we see that

$$p_m(n) = \frac{\Delta x}{2\pi}\int_{-\pi/\Delta x}^{\pi/\Delta x} E_0\{e^{ikX_n}\}e^{-ikm\Delta x}dk$$

which implies that

$$P_0(X_n = 0) = p_0(n) = \frac{\Delta x}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} (\cos k\Delta x)^n dk$$

We will use this Fourier representation of $p_0(n)$ to show that the one-dimensional symmetric random walk is recurrent.

We will show that the expected number of returns to zero is infinite. The expected number of returns by time n is

$$E_0\{\sum_{j=0}^{n} \mathbb{1}\{X_j = 0\}\} = \sum_{j=0}^{n} p_0(j) = \frac{\Delta x}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \frac{1 - (\cos k\Delta x)^{n+1}}{1 - \cos k\Delta x} dk$$

The integrand may be singular only at $k = 0$. Expanding near $k = 0$ we see that it behaves like $n+1$ and therefore in any fixed, small neighborhood of $k = 0$ we see that the integral is unbounded as $n \to \infty$. We conclude, omitting details, that the expected number of returns to the origin, $E_0\{\sum_{j=0}^{\infty} \mathbb{1}\{X_j = 0\}\} = \infty$, and therefore the random walk is recurrent.

In three dimensions the calculation leading to the Fourier representation of the probability of return to the origin in n steps is essentially the same as in one dimension except for notation. Now $m = (m_1, m_2, m_3)$ are points on the integer lattice in three dimensions, we denote the mesh size by Δx in all three directions, and we denote by $\mathbf{e}_j,\ j = 1, 2, 3$ the three unit coordinate vectors in $\mathbb{R}^3$. The Fourier variable is $k = (k_1, k_2, k_3)$ with each coordinate taking values in $(\frac{-\pi}{\Delta x}, \frac{-\pi}{\Delta x})$. The random walk is $X_n = X_{n-1} + Z_n$ where the increments are iid random variables taking values $\pm\mathbf{e}_j\Delta x,\ j = 1, 2, 3$, with probability $1/6$.

Repeating the above steps and using inner product notation for vectors we have that

$$E_0\{e^{ik\cdot X_n}\} = [\frac{1}{3}(\cos k_1\Delta x + \cos k_2\Delta x + \cos k_3\Delta x)]^n$$

Therefore

$$E_0\{\sum_{j=0}^{\infty} \mathbb{1}\{X_j = 0\}\} = (\frac{\Delta x}{2\pi})^3 \int_{-\pi/\Delta x}^{\pi/\Delta x} \int_{-\pi/\Delta x}^{\pi/\Delta x} \int_{-\pi/\Delta x}^{\pi/\Delta x} \frac{1}{1 - \frac{1}{3}(\cos k_1\Delta x + \cos k_2\Delta x + \cos k_3\Delta x)} dk$$

As in the one-dimensional case only $k = 0$ is a singular point and expanding near $k = 0$ we see that the denominator behaves like $|k|^2$. But the jacobian in polar coordinates in three dimensions is also proportional to $|k|^2$ so the singularity cancels and we have, in fact a convergent integral. This proves that the symmetric random walk in three dimensions is transient.

5.2 Connections with classical potential theory

The connections with classical potential theory come from the fact that the transition probability matrix for the random walk on the scaled lattice (scaled by Δx) converges to

the Laplace operator. Let as assume that we are in three dimensions and let $f(x)$ be a smooth and bounded function on $\mathbb{R}^3$. We can then define the transition operator of the random walk by

$$Pf(x) = \frac{1}{6}\sum_{j=1}^{3} f(x \pm e_j \Delta x) = \text{Average of nearest neighbors}$$

so that

$$\frac{1}{(\Delta x)^2}(Pf(x) - f(x)) \to \frac{1}{6}\Delta f(x) = \frac{1}{6}(f_{x_1x_1}(x) + f_{x_2x_2}(x) + f_{x_3x_3}(x))$$

as $\Delta x \to 0$, where Δ is the Laplace operator. For the expectation

$$u_m^n = E\{f(X_n)|X_0 = m\Delta x\},$$

where $m = (m_1, m_2, m_3)$, we have that

$$u_m^{n+1} = (Pu^n)(m\Delta x).$$

and we can rewrite this as

$$u_m^{n+1} - u_m^n = (P - I)u^n(m\Delta x)$$

Dividing by Δt we have

$$\frac{1}{\Delta t}(u_m^{n+1} - u_m^n) = \frac{(\Delta x)^2}{\Delta t} \cdot \frac{1}{(\Delta x)^2}(P - I)u^n(m\Delta x)$$

In the continuum limit $\Delta t \to 0$, $\Delta x \to 0$ with

$$\sigma^2 = \frac{1}{3}\frac{(\Delta x)^2}{\Delta t} = \text{constant}$$

we have that $u_m^n \approx u(n\Delta t, m\Delta x)$ with

$$u_t = \frac{\sigma^2}{2}\Delta u \ , \quad t > 0$$

with $u(0, x) = f(x)$. The factor $1/3$ is attached to the definition of σ^2 because it refers to coordinate-wise mean square displacement rather than overall mean square displacement.

We introduce the Laplace transform

$$\hat{u}(x, \lambda) = \int_0^\infty e^{-\lambda t} u(t, x) dt \ , \quad \lambda > 0$$

and note that the diffusion equation transforms to

$$\lambda\hat{u}(x,\lambda) - f(x) = \frac{\sigma^2}{2}\Delta\hat{u}(x,\lambda)$$

In terms of the Brownian motion process X_t, the continuous time analog of the random walk which is not considered in detail here, we have the probabilistic representation

$$u(t,x) = E_x\{f(X_t)\}$$

and for the Laplace transform

$$\hat{u}(x,\lambda) = \int_0^\infty e^{-\lambda t}E_x\{f(X_t)\}dt = E_x\{\int_0^\infty e^{-\lambda t}f(X_t)dt\}$$

When $f(x) = \mathbb{1}_A(x) = \mathbb{1}\{x \in A\}$ then

$$\hat{u}(x,\lambda) = E_x\{\int_0^\infty e^{-\lambda t}\mathbb{1}_A(X_t)dt\}$$

is the discounted, with rate λ, expected time spent by Brownian motion in the set A, starting from x. When $\lambda = 0$, $\hat{u}(x,0)$ is the expected time spent in A, which is the continuous analog of the quantity that characterizes transience and recurrence in random walks.

The continuum limit is interesting because it connects directly with potential theory, that is the theory of solutions of the Laplace equation. In three dimensions the Green's function for the equation

$$\Delta_x G(x,y) - \lambda G(x,y) = -\delta_y(x)$$

is explicitly given by

$$G(x,y) = \frac{e^{-\sqrt{\lambda}|x-y|}}{4\pi|x-y|}$$

Therefore we have the integral representation

$$\hat{u}(x,\lambda) = E_x\{\int_0^\infty e^{-\lambda t}f(X_t)dt\} = \int_A \frac{e^{-\frac{\sqrt{2\lambda}}{\sigma}|x-y|}}{4\pi|x-y|}dy$$

The expected time spent in A is thus given, when $f(x) = \mathbb{1}_A(x)$, by the Newtonian potential

$$E_x\{\int_0^\infty f(X_t)dt\} = \int_A \frac{1}{4\pi|x-y|}dy$$

For the random walk on the three dimensional lattice we do not have an explicit expression such as this one, which, in particular, shows that the time spent in any bounded set is finite.

The recurrence of the one-dimensional Brownian motion can be seen easily by noting the Green's function in one dimension has the form

$$G(x,y) = \frac{e^{-\sqrt{\lambda}|x-y|}}{2\lambda}$$

and therefore in one dimension we have

$$\hat{u}(x,\lambda) = E_x\{\int_0^\infty e^{-\lambda t} f(X_t)dt\} = \int_A \frac{e^{-\frac{\sqrt{2\lambda}}{\sigma}|x-y|}}{2\lambda} dy$$

This becomes infinite as $\lambda \to 0$, showing that the expected time spent in any set of positive volume is infinite. This is the analog of recurrence for Brownian motion in one dimension.

5.3 Random walk on a graph

Let $G = (V, E)$ be a finite undirected and connected graph where V denotes the set of vertices or nodes $x \in V$, the state space, and E denotes the set of edges connecting nodes $e_{x,y} \in E$. The adjacency matrix of the graph is a $|V| \times |V|$ matrix A whose entry at (x, y) is one if $e_{x,y} \in E$ and zero otherwise. Note that this matrix is symmetric for an undirected graph. Let $X = (X_n : n \geq 0)$ be a Markov chain that moves from vertex to vertex by choosing uniformly among the available edges. We will find the transition probability matrix P of this Markov chain and see its the relationship with A, and we will find the stationary or invariant probabilities $\pi(x)$.

The adjacency matrix is given by

$$A = \left(\mathbb{1}_{\{e_{x,y}\in E\}}\right),\ x, y \in V$$

The degree of a node is $\deg(x) = \sum_{y\in V} A(x, y)$. The transition probability matrix of the random walk is

$$P(x,y) = \left(\frac{\mathbb{1}_{\{e_{x,y}\in E\}}}{\deg(x)}\right)$$

The transition matrix is just the adjacency matrix with row sums normalized to one. The invariant vector is

$$\pi(x) = \frac{\deg(x)}{2|E|}$$

and note that $\sum_{x\in V} \deg(x) = 2|E|$ for the normalization. We use here the fact that the graph is undirected and the so the adjacency matrix is symmetric. We thus have

$$\sum_{x\in V} \pi(x)P(x,y) = \sum_{x\in V} \frac{\deg(x)}{2|E|} \frac{\mathbb{1}_{\{e_{x,y}\in E\}}}{\deg(x)} = \sum_{x\in V} \frac{\mathbb{1}_{\{e_{x,y}\in E\}}}{2|E|} = \frac{\deg(y)}{2|E|}$$

Here we use again the fact that the graph is undirected and that the adjacency matrix is symmetric, and hence $\deg(y) = \sum_{x\in V} \mathbb{1}_{\{e_{x,y}\in E\}}$.

5.4 Probabilistic representation of solutions of difference equations

Let $X_0, X_1, X_2, \ldots$ be the trajectory of the symmetric random walk on the two dimensional lattice $\mathbb{Z}^2 = \{x = (m_1\Delta x, m_2\Delta y),\ m_1, m_2 = 0, \pm 1, \pm 2, \ldots\}$ with span $(\Delta x, \Delta y)$. Let D be a bounded subset of the lattice and denote with ∂D its boundary, that is, the set of points in D that share a bond with points outside D. Let P denote the nearest neighbor averaging operator

$$Pf(x) = \frac{1}{4}\left(f(x + e_1\Delta x) + f(x - e_1\Delta x) + f(x + e_2\Delta y) + f(x - e_2\Delta y)\right)$$

where e_1, e_2 are unit vectors in the two coordinate directions. We want address the following questions:

1. Let $u(x)$ be the solution to the finite difference boundary value problem

   $$(I - P)u(x) = 0,\ x \in D, \quad u(x) = g(x),\ x \in \partial D$$

   where $g(x)$ is a given bounded function of the boundary. We want to find a probabilistic representation of the solution $u(x)$ as an expectation over trajectories of the random walk and show that it does solve the lattice boundary value problem. We can use this probabilistic representation for a Monte Carlo simulation of $u(x)$. The MC simulation desirable, or more efficient, than solution of the linear system by standard direct or iterative methods in high dimensions. The two dimensional lattice problem has an immediate analog in the $d-$dimensional lattice.

2. We want to show that the maximum principle holds for u. This is done in two ways: First from the difference equation and then from the probabilistic representation. The maximum principle says that $u(x)$ takes its maximum and minimum values on the boundary, and that if it does have an interior maximum or minimum then $u(x)$ must be a constant.

3. We also want to find a probabilistic representation for the solution of

   $$(I - P)u(x) = f(x),\ x \in D, \quad u(x) = 0,\ x \in \partial D$$

   where $f(x)$ is a given bounded function in D and show that this representation is well defined for all bounded f? We give, in particular, an upper bound for the solution of this partial difference equation.

We now analyze the questions posed.

1. Let $T \geq 1$ be the first exit time from D starting for $x \in D$. Then the desired probabilistic representation is $u(x) = E\{g(X_T)|X_0 = x\}$. Clearly the boundary condition holds and a standard first-step analysis gives $u(x) = Pu(x)$ in the interior.

This is the discrete analog of a harmonic function since $u(x)$ equals the average of the its values at the nearest neighbors to x, that is, it satisfies the mean value property. In a Monte Carlo simulation the expectation is replaced by the empirical mean of $g(X_T)$ over different, independent realizations of the random walk.

2. The maximum principle says that the maximum of $u(x)$ in the interior is less than or equal to the maximum of $g(x)$ over the boundary. This is obvious from the probabilistic representation, which shows how useful this representation can be. Analytically, we get the maximum principle from the mean value property. For suppose that $u(x)$ takes a maximum at an interior point x. Then it must be bigger than of equal to its values at the nearest neighbors. But it must be equal to their average, and this can only happen if $u(x)$ is identically a constant. Otherwise we have a contradiction and hence the maximum of $u(x)$ cannot be taken in the interior.

3. The probabilistic representation is $u(x) = E\{\sum_{j=0}^{T-1} f(X_j)|X_0 = x\}$ for x in the interior and we set $u(x) = 0$ at the boundary. First step analysis gives $u(x) = f(x) + Pu(x)$ along with the boundary condition. When $f(x) = 1$ we have $u(x) = E\{T|X_0 = x\}$ and $u(x) = 1 + Pu(x)$ plus the boundary condition. So if $E\{T|X_0 = x\} < \infty$ then u is well defined for any bounded f. How do we show that $E\{T|X_0 = x\} < \infty$? We note that $u(x)$ satisfies a linear system of equations for which the zero solution is the unique homogeneous one, by the maximum principle and the zero boundary conditions. Therefore the solution of $u(x) = 1+Pu(x)$ with $u(x) = 0$ at the boundary exists and is unique. This can be used to show that the expected exit time is finite but we need one additional result not covered (the optional stopping theorem). There is still another way to show that the expected exit time is finite. We enclose the finite lattice region D in a strip and calculate the mean exit time explicitly. The mean exit time from a strip depends on the dimension. In two dimensions the mean exit time from a strip of width N is $2i(N - i)$, where $x = (i, j)$. This is twice the mean exit time from the interval in one dimension. This strip solution gives an upper bound in two dimensions to the solution of $u(x) = E\{f(X_T)|X_0 = x\}$ with zero boundary conditions: $|u(x)| < ||f||2(N/2)^2$.

5.5 Discrete time mean reverting random walk

Let $X_n,\ n = 0, 1, 2, \ldots$, be a random walk on the real line $\mathbb{R}$ defined by

$$X_{n+1} = (1 - \mu)X_n + \sigma Z_{n+1},\ n = 0, 1, 2, \ldots,\ X_0 = 0,$$

where the random variables $\{Z_n\}_{n\geq 1}$ are independent and Gaussian with mean zero and variance one. This is a discrete time, continuous space random walk that is mean reverting to zero and is the discrete time analog of the Ornstein-Uhlenbeck process. The parameters μ and σ to be estimated are assumed to be in the range $0 < \mu < 1$ and $0 < \sigma < \infty$. We want to calculate:

1. The first two moments of the Gaussian-Markov process $\{X_n\}$, $E(X_n)$ and $E(X_n^2)$.
2. The transition probability density of the random walk $\{X_n\}$.
3. The maximum likelihood estimator $\hat{\mu}_n = \hat{\mu}_n(X_1, X_2, \cdots, X_n)$ for μ and also for σ^2.
4. We want to show that $\hat{\mu}_n$ converges in probability to μ as n tends to infinity, and similarly for $\hat{\sigma^2} \to \sigma^2$.
5. We want to calculate the asymptotic variance for $\sqrt{n}(\hat{\mu}_n - \mu)$ as n tends to infinity.
6. We note that the CLT holds for $\sqrt{n}(\hat{\mu}_n - \mu)$ as n tends to infinity.
7. We discuss the continuum limit, suitably defined.

We now begin the analysis of this random walk.

1. Taking expectations in the difference equation we get

$$E(X_{n+1}) = (1-\mu)E(X_n),\ n = 0, 1, 2, \ldots,\ X_0 = 0,$$

which implies that $E(X_n) = 0$. The $E(X_n^2)$ is calculated by squaring the difference equation and taking expectation:

$$E(X_{n+1}^2) = (1-\mu)^2 E(X_n^2) + \sigma^2,\ n = 0, 1, 2, \ldots,\ X_0 = 0,$$

Solving this we get

$$m_n = E(X_n^2) = \sigma^2 \frac{1-(1-\mu)^{2n}}{1-(1-\mu)^2} \to \frac{\sigma^2}{1-(1-\mu)^2} = m(\mu) = m$$

as $n \to \infty$ since $0 < \mu < 1$.

2. The transition density is

$$p(x_n, x_{n+1}) = \frac{e^{-\frac{(x_{n+1}-(1-\mu)x_n)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Note that if we start X_0 having a Gaussian density with mean zero and variance m then the random walk is stationary, that is, this is the density of X_n for any $n > 0$.

3. The likelihood function is

$$L_n(\mu, \sigma^2) = \prod_{j=0}^{n-1} p(X_j, X_{j+1}) = \frac{e^{-\frac{(X_{j+1}-(1-\mu)X_j)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

and the normalized log likelihood function is

$$l_n(\mu, \sigma^2) = -\frac{1}{n}\sum_{j=0}^{n-1} \frac{(X_j - (1-\mu)X_{j-1})^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

From this we get the MLE

$$\hat{\mu}_n = 1 - \frac{\frac{1}{n}\sum_{j=0}^{n-1} X_{j+1}X_j}{\frac{1}{n}\sum_{j=0}^{n-1} X_j^2}$$

and

$$\hat{\sigma^2}_n = \frac{1}{n}\sum_{j=0}^{n-1}(X_j - (1-\hat{\mu}_n)X_{j-1})^2$$

4. Using the difference equation we can rewrite the MLE as

$$\hat{\mu}_n = \mu - \sigma\frac{\frac{1}{n}\sum_{j=0}^{n-1} Z_{j+1}X_j}{\frac{1}{n}\sum_{j=0}^{n-1} X_j^2}$$

Let

$$M_n = \sum_{j=0}^{n-1} Z_{j+1}X_j \ , \ n \geq 1,$$

and define $M_0 = 0$. We see that $E(M_n|X_0, X_1, \ldots, X_{n-1}) = M_{n-1}$ which means that M_n is a martingale and that it also has mean zero. Because of the martingale property we have that

$$E(M_n^2) = \sum_{j=0}^{n} E(X_j^2) = \sum_{j=0}^{n} m_j$$

and this and Chebyshev's inequality imply that $n^{-1}M_n \to 0$ in mean square as $n \to \infty$. We also have that

$$\frac{1}{n}K_n = \frac{1}{n}\sum_{j=0}^{n} X_j^2 \to \frac{1}{n}\sum_{j=0}^{n} m_j \to m$$

in mean square. We discuss this further below. These facts imply that $\mu_n \to \mu$ in probability.

5. We note that $E\{(n^{-1/2}M_n)^2\} = \frac{1}{n}\sum_{j=0}^{n} m_j \to m$ and therefore $\mathrm{Var}\{\sqrt{n}(\hat{\mu}_n - \mu)\} \to \sigma^2 m^{-1} = 1 - (1-\mu)^2$. We also note that $-l_n''(\mu) = \frac{1}{n\sigma^2}\sum_{j=0}^{n} X_j^2 \to \frac{m}{\sigma^2}$ in mean square. This means that the Fisher information $I = m/\sigma^2$ and therefore the CR lower bound is achieved for the variance of the CLT for the MLE. The CLT for the MLE holds here as it is essentially the CLT for the martingale M_n, using also Slutsky's theorem.

6. It remains to show that $\frac{1}{n}\sum_{j=0}^{n} X_j^2 \to m$ in mean square. This is a form of the ergodic theorem for the discrete-time OU process but we do not have this theorem available here, even though it is true, because we only have the ergodic theorem for finite state (discrete state) Markov chains. However, we can get the result by direct computation using the explicit representation of $X_n = \sum_{j=1}^{n}(1-\mu)^{n-j}Z_j$ in terms of the i.i.d. N(0,1) random variables $\{Z_n\}$. An explicit calculation shows that $E\{[\frac{1}{n}\sum_{j=0}^{n-1}(X_j^2 - m_j)]^2\} \to 0$ as $n \to \infty$.

7. The continuum limit. Here we must change the basic recursion by letting $\mu = \gamma\Delta t$ and also changing the random part so as to have

$$X_{n+1} - X_n = -\gamma\Delta t X_n + \sigma\sqrt{\Delta t}Z_{n+1},\ n = 0, 1, 2, \ldots,\ X_0 = 0,$$

In the continuum limit $n \to \infty,\ \Delta t \to 0,\ n\Delta t = t$ we have $X_n \sim X_t$ in probability and X_t is the OU process

$$dX_t = -\gamma X_t dt + \sigma dW_t$$

Note that here we replace Z_n by $\sqrt{\Delta t}Z_n$, along with letting $\mu = \gamma\Delta t$. Here $W_t,\ t \geq 0$ is the Brownian motion process, X_t is the Ornstein-Uhlenbeck process, a Gaussian process with mean $e^{-\gamma t}X_0$ and variance $\frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma t})$.

The MLE estimator

$$\hat{\mu}_n = \mu - \frac{\frac{1}{n}\sum_{j=0}^{n-1} Z_{j+1}X_j}{\frac{1}{n}\sum_{j=0}^{n-1} X_j^2}$$

now gives an MLE estimator for γ, assuming that σ is known

$$\hat{\gamma}_n = \gamma - \frac{\sigma\sqrt{\Delta t}\sum_{j=0}^{n-1} Z_{j+1}X_j}{\Delta t\sum_{j=0}^{n-1} X_j^2}$$

In the continuum limit this becomes

$$\hat{\gamma}_t = \gamma - \frac{\sigma\int_0^t X_t dW_t}{\int_0^t X_t^2 dt}.$$

Notice that if in the original recursion we only let $\mu = \gamma\Delta t$ then the problem becomes singular since

$$m = \frac{\sigma^2}{1 - (1-\mu)^2}$$

and it will tend to infinity as Δt, and μ, goes to zero. It is essential to also scale the fluctuations so as to get a non-singular continuum limit.

However, in the continuum lime the MLE estimator $\hat{\gamma}_t$ of the rate of mean reversion requires that we let $t \to \infty$ so as to be a consistent estimator. We therefore need to consider the asymptotic behavior of

$$\frac{\sigma \int_0^t X_t dW_t}{\int_0^t X_t^2 dt}$$

as $t \to \infty$. First we look at the denominator and note that by the ergodic theorem for the OU process or by an explicit calculation

$$\frac{1}{t} \int_0^t X_t^2 dt \ \to \ \frac{\sigma^2}{2\gamma}$$

in probability. For the numerator we use the martingale CLT for the stochastic integral to get

$$\frac{1}{\sqrt{t}} \int_0^t X_t dW_t \ \to \ N(0, \frac{\sigma^2}{2\gamma})$$

in law, since the quadratic variation of this stochastic integral is the denominator whose (scaled) limit we already know. Therefore, using also Slutsky's theorem,

$$\sqrt{t}(\hat{\gamma}_t - \gamma) \ \to \ N(0, 2\gamma)$$

in law, which we can write as

$$\hat{\gamma}_t \sim \gamma \left(1 + Z\sqrt{\frac{2}{\gamma t}}\right)$$

where Z an $N(0,1)$ random variable.

Note that γt must be large for consistency, which means that the duration of the sample t must be much larger than the mean reversion time $\frac{1}{\gamma}$. This is intuitively clear, and we see that it comes out of the analysis of the MLE estimator for the rate of mean reversion as it should. Note also that σ does not appear in this asymptotic limit, which is perhaps surprising.

6 Brownian Motion

A Brownian motion $\{B_t, t \geq 0\}$, is a family of R.V.s indexed by $t \geq 0$ such that the following properties hold:

1. $B_0 = 0$.
2. B_t has independent increments
3. The increments $(B_t - B_s)$, $t > s$ are $N(0, t-s)$, that is, Gaussian with mean zero and variance $t - s$.

The independent increments property means that for all finite partitions of the time axis, $0 = t_0 < t_1 < t_2 < \ldots < t_n$, the corresponding Brownian motion increments $(B_{t_n} - B_{t_{n-1}}), (B_{t_{n-1}} - B_{t_{n-2}}), \ldots, (B_{t_1} - B_{t_0})$ are all independent random variables. Clearly Brownian motion is a Gaussian process in that for any times $0 = t_0 < t_1 < t_2 < \ldots < t_n$, the random variables $(B_{t_0}, B_{t_1}, \ldots, B_{t_n})$ are jointly Gaussian.

Before continuing with Brownian motion and its properties we define the martingale property of processes in discrete and continuous time.

Discrete time martingales

Given a sample space Ω and $\mathcal{F}$, a σ-algebra of subsets of Ω, along with an increasing family $\mathcal{F}_n$ of sub-σ-algebras of $\mathcal{F}$, $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$, consider a discrete time process $\{X_n\}_{n \geq 0}$ where the X_n are R.V.s on the probability space Ω, $X_n : \Omega \to \mathbb{R}$ $(X_n(\omega) \in \mathbb{R})$. We will assume that

$$\mathcal{F}_n = \sigma\{X_0, X_1, \ldots, X_n\} = \text{the sigma algebra generated by the R.V. up to time } n$$

and that $\mathbb{E}|X_n| < \infty$, for all n. If for all n

$$\mathbb{E}\{X_n | \mathcal{F}_{n-1}\} = X_{n-1}$$

then X_n has the martingale property or is a martingale.

Similarly for a continuous time process $\{X_t, t \geq 0\}$ with filtration

$$\mathcal{F}_t = \sigma\{X_s, s \leq t\} = \text{information about the process } X_s \text{ up to time } t$$

and $\mathbb{E}|X_t| < \infty$, the martingale property is

$$\mathbb{E}\{X_t | \mathcal{F}_s\} = X_s \ , \quad s \leq t.$$

6.1 Construction of a Brownian Motion

Consider the sample space

$$\Omega = \{\text{set of continuous functions on } \mathbb{R}_+\} = \{\omega(t) \in C_0 : \omega(t) : \mathbb{R}_+ \to \mathbb{R}\}$$

We will also use the notation:

$$B_t(\omega) = \omega(t) \in \mathbb{R}$$

to denote the coordinate of the continuous path ω at time t. Let us define the filtration

$$\mathcal{F}_t = \sigma\{B_s, s \leq t\} = \sigma\text{-algebra generated by all sets of continuous paths up to time } t$$

These are σ algebras of subsets of Ω, for $0 \leq t \leq T$ for any $T < \infty$. The mapping taking $(\Omega, \mathcal{F}_t) \to (\mathbb{R}, B(\mathbb{R}))$, so that $\omega \mapsto B_t(\omega)$, is measurable, i.e. B_t is a RV. Here $B(\mathbb{R})$ denotes the Borel subsets of the real line.

It is possible to construct a probability measure or law P on $(\Omega, \mathcal{F}_{t,(0\leq t\leq T)})$ such that $B_t(\omega)$ is a Brownian Motion, that is, a collection of random variables satisfying properties listed in the definition above. This is the canonical construction of the Brownian Motion (BM) process B_t. The main point in this construction is that the probability law P is defined on the sample space of continuous functions Ω, which means that the sample paths of BM are continuous.

The probability measure P on $(\Omega, \mathcal{F}_{t,(0\leq t\leq T)})$ is such that, for example, for $0 < t$ and $x \in \mathbb{R}$ fixed,

$$P\{\omega \in \Omega : B_t(\omega) - B_s(\omega) \leq x\} = \int_{-\infty}^{x} \frac{e^{-\frac{u^2}{2(t-s)}}}{\sqrt{2\pi(t-s)}} du$$

Kolmogorov Continuity Criterion: The definition of Brownian motion as a collection of random variables with independent, Gaussian increments, does not tell us that, in fact, it is possible to define these random variables as coordinates of continuous trajectories. What property in their definition is essential for having sample path continuity? The simplest and most often used tool for answering this question is the Kolmogorov criterion:

Suppose that there exist constants $\alpha, \beta > 0$ and $c \geq 0$ such that, for a stochastic process X_t, defined through its finite dimensional distributions,

$$\mathbb{E}\{|X_t - X_s|^\alpha\} \leq c|t-s|^{\beta+1}, \quad \forall t, s \leq T.$$

Then X_t has continuous sample paths, for $0 \leq t \leq T$, with probability one. This means that if at first the process X_t is suitably defined as a probability law on all functions from the positive real line to the real line, then this law assigns probability one on the subset of continuous functions.

To see that Brownian motion has continuous sample paths, according to the Kolmogorov criterion, we will use the following result:

The expectation of odd powers of B_t is zero:

$$\mathbb{E}\{B_t^{2p+1}\} = \int_{-\infty}^{\infty} u^{2p+1} \frac{e^{-\frac{u^2}{2t}}}{\sqrt{2\pi t}} du = 0$$

since u^{2p+1} is odd and the density is even.

Define $c_p = \frac{(2p)!}{2^p p!}$. The expectation of even powers of B_t is $c_p t^p$:

$$\mathbb{E}\{B_t^{2p}\} = \int_{-\infty}^{\infty} u^{2p} \frac{e^{-\frac{u^2}{2t}}}{\sqrt{2\pi t}} du$$

let $v = \frac{u}{\sqrt{t}}$

$$= \int_{-\infty}^{\infty} t^p v^{2p} \frac{e^{-\frac{v^2}{2}}}{\sqrt{2\pi}}$$

which is just the $(2p)^{\text{th}}$ moment of a mean zero Gaussian rv, thus

$$= \frac{(2p)!}{2^p p!} t^p$$

We can use this result to show that Brownian Motion has continuous paths. With $\alpha = 4$ and $\beta = 1$ and $c = c_4 = 3$, the Kolmogorov's continuity condition for B_t is

$$\mathbb{E}|B_t - B_s|^4 = \mathbb{E}B_{t-s}^4 = 3|t-s|^2$$

and therefore the Brownian motion process B_t has continuous paths. It can, in particular, be considered as the canonical process $B_t(\omega)$ on the sample space of continuous trajectories.

Although Brownian paths, $t \mapsto B_t(\omega)$, are continuous with probability one, they are not differentiable w.p.1 at any t. This is a consequence of the independent increments property.

6.2 Properties of Brownian Motion

Below we state some useful properties of Brownian Motion, B_t:

1. B_t is a martingale.

   Recall that by the Schwatz inequlity $\mathbb{E}|X| \leq \sqrt{\mathbb{E}|X|^2}$. Thus

$$\mathbb{E}\{|B_t|\} \leq \sqrt{\mathbb{E}\{B_t^2\}} = \sqrt{t} < \infty, \quad 0 \leq t \leq T < \infty.$$

and

$$\begin{aligned}\mathbb{E}\{B_t|\mathcal{F}_s\} &= \mathbb{E}\{B_t - B_s + B_s|\mathcal{F}_s\} \\ &= \mathbb{E}\{B_t - B_s|\mathcal{F}_s\} + \mathbb{E}\{B_s|\mathcal{F}_s\} \\ &= \mathbb{E}\{B_t - B_s\} + B_s \text{ by independent increments} \\ &= B_s\end{aligned}$$

2. $(B_t^2 - t)$ is a martingale.

   Clearly it is integrable.

$$\begin{aligned}\mathbb{E}\{B_t^2 - t|\mathcal{F}_s\} &= \mathbb{E}\{(B_t - B_s + B_s)^2 - t|\mathcal{F}_s\} \\ &= \mathbb{E}\{(B_t - B_s)^2 + 2B_s(B_t - B_s) + B_s^2 - t|\mathcal{F}_s\} \\ &= \mathbb{E}\{(B_t - B_s)^2\} + 2B_s\mathbb{E}\{B_t - B_s\} + B_s^2 - t \\ &= t - s + 2B_s \cdot 0 + B_s^2 - t \\ &= B_s^2 - s\end{aligned}$$

3. $X_t = e^{\lambda B_t - \frac{\lambda^2 t}{2}}$ is a martingale, for every real or complex λ.

   Recall that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathbb{E}e^{sX} = e^{\mu s + \sigma^2 s^2/2}$

$$\mathbb{E}\{e^{\lambda B_t - \frac{\lambda^2 t}{2}}\} = e^{-\frac{\lambda^2 t}{2}}\mathbb{E}\{e^{\lambda B_t}\} = e^{-\frac{\lambda^2 t}{2}} e^{\frac{\lambda^2 t}{2}} = 1 < \infty$$

   Note that the expectation is constant, which is a good sign for X_t being a martingale.

$$\begin{aligned}\mathbb{E}\{X_t|\mathcal{F}_s\} &= \mathbb{E}\{e^{\lambda B_t - \frac{\lambda^2 t}{2}}|\mathcal{F}_s\} \\ &= e^{-\frac{\lambda^2 t}{2}}\mathbb{E}\{e^{\lambda B_t}|\mathcal{F}_s\} \\ &= e^{-\frac{\lambda^2 t}{2}}\mathbb{E}\{e^{\lambda(B_t - B_s + B_s)}|\mathcal{F}_s\} \\ &= e^{-\frac{\lambda^2 t}{2}}\mathbb{E}\{e^{\lambda(B_t - B_s)}e^{\lambda B_s}|\mathcal{F}_s\} \\ &= e^{-\frac{\lambda^2 t}{2}} e^{\lambda B_s}\mathbb{E}\{e^{\lambda(B_t - B_s)}\} \\ &= e^{-\frac{\lambda^2 t}{2}} e^{\lambda B_s} e^{\lambda^2 \frac{t-s}{2}} \\ &= e^{\lambda B_s - \frac{\lambda^2 s}{2}} = X_s\end{aligned}$$

4. B_t is continuous in t a.s. (for almost all realizations).

   This was demonstrated above using Kolmogorov's criterion.

5. B_t is nowhere differentiable for any t and has infinite total variation. It is this that makes defining integrals with B_t difficult as we discuss next.

6.3 Total Variation and Quadratic Variation

For a function f on $[0,T] \to \mathbb{R}$ define its **total variation** as

$$\mathrm{TV}_T(f) = \lim_{N\to\infty} \sup_{\pi_N} \sum_{k=0}^{N-1} |f(t_{k+1}) - f(t_k)|,$$

where π_N is the set of all possible divisions of the interval $[0,T]$ into N segments $(0 = t_0 < t_t < \ldots < t_N = T)$ such that $\max_k(t_{k+1} - t_k) \to 0$ as $N \to \infty$. We say f has finite total variation if $\mathrm{TV}_T(f) < \infty$.

For a continuous and bounded function $g(t)$, $0 \le t \le T$ we can define the Riemann integral with respect to f:

$$\int_0^T g(s)df(s) = \lim_{N\to\infty} \sup_{\pi_N} \sum_{k=0}^{N-1} g(t_k^*)\left(f(t_{k+1}) - f(t_k)\right), \quad t_k^* \in [t_k, t_{k+1}].$$

This integral exists if g is continuous and f has finite total variation in $[0,T]$. It is independent of the choice of $t_k^* \in [t_k, t_{k+1}]$. It is not defined, in general, if f does not have finite total variation.

For BM: $\mathrm{TV}_T(B_t) = \infty$ w.p.1, and so

$$\int_0^T g(s)dB_s$$

does not exist in the classical sense. With a uniform partition of $[0,T]$, $\{t_k\}$ with $t_k = k\Delta t$, we have

$$\mathbb{E}\{\sum_{k=0}^{N-1} |B_{t_{k+1}}(\omega) - B_{t_k}(\omega)|\} = \sum_{k=0}^{N-1} \mathbb{E}\{|B_{t_{k+1}} - B_{t_k}|\} = \sum_{k=0}^{N-1} \tilde{c}_1\sqrt{\Delta t} = N\tilde{c}_1\sqrt{\Delta t} = \frac{\tilde{c}_1 T}{\sqrt{\Delta t}} \nearrow \infty$$

as $\Delta t \to 0$. Here $\tilde{c}_p$ is defined as

$$\tilde{c}_p = \int_{-\infty}^{\infty} \frac{|z|^p e^{-z^2/2}}{\sqrt{2\pi}} dz$$

and is obtained in a similar manner as for the even moments of Brownian motion above.

Thus, in the mean, the total variation of Brownian Motion goes to ∞. Since the integrand is positive, it can actually be shown that $\mathrm{TV}_T(B_t) \overset{\text{w.p.1}}{\to} \infty$. This is done using the Borel-Cantelli lemma but it is not presented here.

For a function $f : [0,T] \to \mathbb{R}$, its **quadratic variation** is defined as

$$\mathrm{QV}_T(f) = \lim_{N\to\infty} \sup_{\pi} \sum_{k=0}^{N-1} |f(t_{k+1}) - f(t_k)|^2$$

with the partition π defined as above and the limit being in mean square (or in probability).

We have the following fact.
$\mathrm{QV}_T(B_t) < \infty$ w.p.1 and in fact, $\mathrm{QV}_T(B_t) = T$.
The mean of QV is comuted as follows:

$$\mathbb{E}\{\mathrm{QV}_T(B_t)\} = \mathbb{E}\{\sum_{k=0}^{N-1} |B_{t_{k+1}}(\omega) - B_{t_k}(\omega)|^2\} = \sum_{k=0}^{N-1} \mathbb{E}\{|B_{t_{k+1}}(\omega) - B_{t_k}(\omega)|^2\} = \sum_{k=1}^{N} (t_k - t_{k-1}) = T$$

We show in the next section that the fluctuations (the fourth moments) also go to zero.

The fact that $\mathrm{TV}_T(B_t) = \infty$ and $\mathrm{QV}_T(B_t) = T$ has consequences when one tries to apply Taylor's Formula. Let $f(x)$ be a smooth function:

$$\Delta f(B_t) = f(B_{t+\Delta t}) - f(B_t) = f'(B_t)(B_{t+\Delta t} - B_t) + \frac{1}{2} f''(B_t)(B_{t+\Delta t} - B_t)^2 + \cdots$$

If we sum this and then use a general Riemann integral we would run into trouble since the sum over the first term tends to infinity and the sum over the second term is not small, since it tends to T. So we need a better definition of integration with respect to Brownian Motion.

6.4 The quadratic variation of Brownian motion

Let $(\Omega, \mathcal{F}, P)$ be the probability space of continuous functions with P the Brownian motion measure, $\mathcal{F}_t$ past events up to time t and $B_t(\omega)$ the coordinate of the path ω at time t. We show that if the interval $[0,t]$ is partitioned into N segments of length $\Delta = t/N$ then

$$\lim_{\Delta\to 0} \sup \sum_{j=0}^{N-1} (B_{t_{j+1}}(\omega) - B_{t_j}(\omega))^2 = t$$

in mean square. The limit on the left is the quadratic variation of Brownian motion and the sup is over any partition whose maximum interval goes to zero.

Let t_j for $0 \leq j \leq N$ be a partition of $[0,T]$ such that $\max_{0\leq j\leq N-1}(t_{j+1}-t_j) \to 0$. We write the difference

$$\sum_{j=0}^{N-1}[B(t_{j+1},\omega)-B(t_j,\omega)]^2 - t$$

as

$$\sum_{j=0}^{N-1}\{[B(t_{j+1},\omega)-B(t_j,\omega)]^2 - (t_{j+1}-t_j)\}$$

and notice that this is a sum of independent zero-mean random variables. Therefore

$$E\left(\sum_{j=0}^{N-1}\{[B(t_{j+1},\omega)-B(t_j,\omega)]^2 - (t_{j+1}-t_j)\}\right)^2$$

is equal to

$$E\left[\sum_{j=0}^{N-1}\{[B(t_{j+1},\omega)-B(t_j,\omega)]^2 - (t_{j+1}-t_j)\}^2\right]$$

$$= \sum_{j=0}^{N-1}\{3(t_{j+1}-t_j)^2 - 2(t_{j+1}-t_j)^2 + (t_{j+1}-t_j)^2\} = \sum_{j=0}^{N-1} 2(t_{j+1}-t_j)^2 \leq 2T \max_{0\leq j\leq N-1}(t_{j+1}-t_j)$$

where we used the fact that for B with distribution $N(0,t)$ we have $E[B^2] = t$ and $E[B^4] = 3t^2$. Thus we have

$$E\left(\sum_{j=0}^{N-1}[B(t_{j+1},\omega)-B(t_j,\omega)]^2 - t\right)^2 \xrightarrow{N\to\infty} 0$$

as required.

7 Stochastic Integral

We want to construct processes as solutions of differential equations which we write in the form

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t$$

with a corresponding integral form

$$X_t = x + \int_0^T \mu(X_t)dt + \int_0^T \sigma(X_t)dB_t$$

We need to define this integral with respect to Brownian Motion that enters. The problem is that Brownian motion has infinite total variation and therefore X_t is not expected to be differentiable because B_t is not differentiable. The Brownian integral must be constructed in a special way.

7.1 Class of Integrands for Stochastic Integrals

The integrands that will be integrated with respect to Brownian motion belong to a special class as follows.

A function $f(t, \omega)$

$$f(t, \omega) : [0, T] \times \Omega \to \mathbb{R}$$

is non-anticipating if for each $t \in [0, T]$ and $f(t, \cdot)$ is a random variable that is $\mathcal{F}_t$ measurable. This means that for any $A \subset \mathbb{R}$,

$$\{\omega : f(t, \omega) \in A\} \in \mathcal{F}_t.$$

Examples of non-anticipating f are: $f(t, \omega) = B_t(\omega)$ and $f(t, \omega) = \max_{0 \le s \le t} B_s(\omega)$, which clearly depend on the Brownian path only up to time t.

To define the stochastic integral we also want $\mathbb{E}\int_0^T f^2(t, \cdot)dt < \infty$, namely that $f(t, \omega)$ be square integrable as a process.

For $f(t, \omega)$ non-anticipating and square integrable, the stochastic integral $\int_0^T f(t, \omega)dB_t(\omega)$ is a well defined square integrable random variable defined as follows.

The stochastic integral can be defined as the mean square limit

$$\int_0^T f(s, \omega)dB_s(\omega) = \lim_{N \to \infty} \sum_{k=0}^{N-1} f(t_k, \omega) \left(B_{t_{k+1}}(\omega) - B_{t_k}(\omega) \right).$$

The limit does not depend on the partition $0 = t_0 < t_1 < t_2 < ... < t_N = T$. The way to do this is as follows.

The stochastic integral is first defined and analyzed for *simple* functions. The integrand f is a simple function if

$$f(t,\omega) = e_k(\omega), \quad t_k \le t < t_{k+1}, \quad k = 0, 1, \ldots, N-1$$

where $e_k(\omega)$ is $\mathcal{F}_{t_k}$ measurable and square integrable. Note that f can be written as

$$f(t,\omega) = \sum_{k=0}^{N-1} e_k(\omega)\chi_{[t_k,t_{k+1}]}(t).$$

For simple f the stochastic integral is

$$I_T(f) = \int_0^T f(s,\omega)dB_s(\omega) = \sum_{k=0}^{N-1} e_k(\omega)\left(B_{t_{k+1}}(\omega) - B_{t_k}(\omega)\right).$$

7.2 Properties

For simple integrands the following properties of the stochastic integral follow almost immediately.

1. Additivity
$$\int_0^T fdB = \int_0^u fdB + \int_u^T fdB \ , \quad 0 \le u \le T$$

2. Linearity with respect to the integrand
$$\int_0^T (f+g)dB = \int_0^T fdB + \int_0^T gdB$$

3. The expected value of $I_T(f)$ is zero. This is seen from the following calculation
$$\begin{aligned}
\mathbb{E}\{\int_0^T fdB\} &= \mathbb{E}\{\sum_{k=0}^{N-1} e_k(B_{t_{k+1}} - B_{t_k})\} \\
&= \sum_{k=0}^{N-1} \mathbb{E}\{e_k(B_{t_{k+1}} - B_{t_k})\} \\
&= \sum_{k=0}^{N-1} \mathbb{E}\{\mathbb{E}\{e_k\left(B_{t_{k+1}} - B_{t_k}\right)|\mathcal{F}_{t_k}\}\} \\
&= \sum_{k=0}^{N-1} \mathbb{E}\{e_k\mathbb{E}\{B_{t_{k+1}} - B_{t_k}\}\} \\
&= 0
\end{aligned}$$

4. Ito Isometry property holds

$$\mathbb{E}\{I_T^2(f)\} = \int_0^T \mathbb{E}f^2(s,\cdot)ds$$

as is seen by the following calculation.

$$\begin{aligned}
\mathbb{E}\{I_T^2(f)\} &= \mathbb{E}\{\left(\sum_{k=0}^{N-1} e_k(B_{t_{k+1}} - B_{t_k})\right)^2\} \\
&= \mathbb{E}\{\sum_{k=0}^{N-1}\sum_{\ell=0}^{N-1} e_k e_\ell (B_{t_{k+1}} - B_{t_k})(B_{t_{\ell+1}} - B_{t_\ell})\} \\
&= \mathbb{E}\{\sum_{k=0}^{N-1} e_k^2 (B_{t_{k+1}} - B_{t_k})^2\} + 2\sum_{k=0}^{N-1}\sum_{\ell=k+1}^{N-1} \mathbb{E}\{e_k e_\ell (B_{t_{k+1}} - B_{t_k})(B_{t_{\ell+1}} - B_{t_\ell})\}
\end{aligned}$$

The second term goes to zero due to the independence of the increments of Brownian motion

$$\begin{aligned}
&= \sum_{k=0}^{N-1} \mathbb{E}\{e_k^2 \mathbb{E}\{(B_{t_{k+1}} - B_{t_k})^2 | \mathcal{F}_{t_k}\}\} \\
&= \sum_{k=0}^{N-1} \mathbb{E}\{e_k^2\}(t_{k+1} - t_k) \\
&= \int_0^T \mathbb{E}\{f^2(s,\cdot)\}ds
\end{aligned}$$

5. $\int_0^t f dB$ is a continuous in t $\mathcal{F}_t$ martingale. The continuity in t is clear from the definition for simple integrands, inherited from the continuity of Brownian motion, and the martingale property is verified in the next section.

7.3 From Simple Functions to General (Non-Anticipating) Functions

Given $f(t,\omega)$, which is non-anticipating and square integrable, there exists a sequence of **simple** non-anticipating functions $f_n(t,\omega)$ such that

$$\mathbb{E}\int_0^T \left(f(s,\omega) - f_n(s,\omega)\right)^2 ds \to 0$$

then

$$\lim_{n\to\infty} \int_0^T f_n(t,\omega) dB_t(\omega)$$

exists in mean squared and defines the stochastic integral, denoted by

$$I_T(f,\omega) = \int_0^T f(t,\omega) dB_t(\omega).$$

Moreover, the properties in the previous section hold for this limit.

As an example, we now compute directly from the definition $\int_0^T B_s dB_s$, that is, the stochastic integral with $f(t,\omega) = B_t(\omega)$. With $T = \Delta N$ fixed and the uniform partition $t_k = k\Delta$ we have

$$\int_0^T B_s dB_s = \lim_{N\to\infty} \sum_{k=0}^{N-1} B_{t_k}(\omega)(B_{t_{k+1}}(\omega) - B_{t_k}(\omega)).$$

Using the identity $a(b-a) = \frac{1}{2}(b^2 - a^2 - (b-a)^2)$ we have

$$\begin{aligned}
&= \lim_{N\to\infty} \sum_{k=0}^{N-1} \frac{1}{2}(B^2_{t_{k+1}}(\omega) - B^2_{t_k}(\omega)) - \frac{1}{2} \lim_{N\to\infty} \sum_{k=0}^{N-1} (B_{t_{k+1}}(\omega) - B_{t_k}(\omega))^2 \\
&= \frac{1}{2}B_T^2 - \frac{1}{2}B_0^2 - \frac{1}{2}QV_T(B) \\
&= \frac{1}{2}\left(B_T^2 - T\right).
\end{aligned}$$

Note the extra $T/2$ term, which is perhaps unexpected at first. It is a direct consequence of the definition of the stochastic integral, with the Brownian increments pointing always forward in time, and the quadratic variation property of Brownian motion. If we apply the Ito Isometry to this example we have:

$$\mathbb{E}\{\left(\int_0^T B_s dB_s\right)^2\} = \mathbb{E}\{\int_0^T B_s^2 ds\} = \int_0^T s ds = \frac{1}{2}T^2$$

where we have interchanged the integral with the expectation. Now if we apply the result of the above computation we have

$$\mathbb{E}\left(\int_0^T B_s dB_s\right)^2 = \mathbb{E}\left(\frac{1}{2}B_T^2 - \frac{1}{2}T\right)^2 = \mathbb{E}\frac{1}{4}B_T^4 - \frac{1}{2}B_T^2 + \frac{1}{4}T^2 = \frac{3}{4}T^2 - \frac{1}{2}T^2 + \frac{1}{4}T^2 = \frac{1}{2}T^2,$$

which is what we got with the Ito isometry.

Let us introduce the notation

$$I_t = I(t,\omega) = \int_0^t f(s,\omega) dB_s(\omega), \quad 0 \le t \le T < \infty$$

and consider the family of RVs $\{I_t; 0 \le t \le T\}$.

For simple functions I_t is a continuous $\mathcal{F}_t$ martingale; i.e.

$$\mathbb{E}\{I_{t_2^*}|\mathcal{F}_{t_1^*}\} = I_{t_{1^*}}$$

for $0 < t_1^* < t_2^* < \infty$.

The continuity is clear and we have

$$\begin{aligned}
\mathbb{E}\{I_{t_2^*}|\mathcal{F}_{t_1^*}\} &= \mathbb{E}\{\sum_{k=0}^{n_2^*-1} e_k(B_{t_{k+1}} - B_{t_k})|\mathcal{F}_{t_1^*}\} \\
&= \mathbb{E}\{\sum_{k=0}^{n_1^*-1} e_k(B_{t_{k+1}} - B_{t_k}) + \sum_{k=n_1^*}^{n_2^*-1} e_k\left(B_{t_{k+1}} - B_{t_k}\right)|\mathcal{F}_{t_1^*}\} \\
&= I_{t_1^*} + \mathbb{E}\{\sum_{k=n_1^*}^{n_2^*-1} e_k\left(B_{t_{k+1}} - B_{t_k}\right)|\mathcal{F}_{t_1^*}\} \\
&= I_{t_1^*}
\end{aligned}$$

This proof is valid only when t_1^* and t_2^* are places where the simple function jumps, that is points in the partition of the time interval. An extension to the case where t_1^* and t_2^* are not in partition points can be given easily. We are still dealing with simple integrands.

We turn to establishing that the continuous martingale property remains true in the case when f is any non-anticipating square integrable function.

We will use the very useful Kolmogorov Inequality.

Kolmogorov/Doob Martingale Inequality

Let M_t be a p-integrable continuous martingale for $0 \le t \le T < \infty$,

$$\mathbb{E}|M_t|^p < \infty, \quad p \ge 1$$

and

$$\mathbb{E}\{M_t|\mathcal{F}_s\} = M_s, \quad t \ge s.$$

Then

$$\mathrm{P}\left\{\max_{0\le s\le t} |M_s| > \lambda\right\} \le \frac{\mathbb{E}|M_t|^p}{\lambda^p}$$

Kolmogorov's Inequality generalizes Chebychev's Inequality by saying that the maximum of $|M_s|$ is controlled by the mean of the **end point** "$|M_t|^p$".

Kolmogorov's Inequality in Discrete Time

Let $\{M_n, n \ge 0\}$ be a martingale with $\mathbb{E}|M_n|^p < \infty$ for some $p \ge 1$. Define the following

sets

$$\begin{aligned}
A_0 &= \{\omega \in \Omega : |M_0| > \lambda\} \\
A_1 &= \{\omega \in \Omega : |M_1| > \lambda, |M_0| \leq \lambda\} \\
A_2 &= \{\omega \in \Omega : |M_2| > \lambda, |M_0| \leq \lambda, |M_1| \leq \lambda\} \\
&\vdots = \vdots \\
A_i &= \{\omega \in \Omega : |M_i| > \lambda, |M_0| \leq \lambda, \ldots, |M_{i-1}| \leq \lambda\}
\end{aligned}$$

Note that $A_j \cap A_i = \emptyset$ for $i \neq j$. Thus

$$\{\omega \in \Omega : \max_{0 \leq k \leq n} |M_k| \geq \lambda\} = \cup_{k=0}^{n} A_k.$$

So

$$\mathrm{P}\left\{\max_{0 \leq k \leq n} |M_k| > \lambda\right\} = \sum_{k=0}^{n} \mathrm{P}\left\{A_k\right\} = \sum_{k=0}^{n} \mathbb{E}\{\chi_{A_k}\} \leq \sum_{k=0}^{n} \mathbb{E}\{\frac{|M_k|^p}{\lambda^p} \chi_{A_k}\}$$

since for each $\omega \in A_k$, $|M_k| > \lambda$, thus $(|M_k|/\lambda)^p > 1$.
We now use the Martingale property of M_n to bound $|M_k|$:

$$|M_k| = |\mathbb{E}\{M_n|\mathcal{F}_k\}| \leq \mathbb{E}\{|M_n||\mathcal{F}_k\} \leq \mathbb{E}\{|M_n|^p|\mathcal{F}_k\}^{\frac{1}{p}}$$

where the last inequality follows from Holder's Inequality:

$$|\mathbb{E}\{fg\}| \leq \mathbb{E}\{|f|^p\}^{\frac{1}{p}} \mathbb{E}\{|g|^q\}^{\frac{1}{q}}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Combining these two results we have

$$\begin{aligned}
\mathrm{P}\left\{\max_{0 \leq k \leq n} |M_k| > \lambda\right\} &\leq \sum_{k=0}^{n} \frac{1}{\lambda^p} \mathbb{E}\{\mathbb{E}\{|M_n|^p|\mathcal{F}_k\}\chi_{A_k}\} \\
&= \frac{1}{\lambda^p} \sum_{k=0}^{n} \mathbb{E}\{\mathbb{E}\{|M_n|^p \chi_{A_k}|\mathcal{F}_k\}\} \\
&= \frac{1}{\lambda^p} \sum_{k=0}^{n} \mathbb{E}\{|M_n|^p \chi_{A_k}\} \\
&= \frac{1}{\lambda^p} \mathbb{E}\{|M_n|^p \sum_{k=0}^{n} \chi_{A_k}\} \\
&= \frac{1}{\lambda^p} \mathbb{E}\{|M_n|^p\}
\end{aligned}$$

The continuous time version of the Kolmogorov Inequality is obtained from the discrete time case by a limit argument.

We now want to apply the general limit process given at the beginning of this chapter to see how the stochastic integrals converge. Take $f(t,\omega)$ to be a square integrable non-anticipating function and let f_n be a sequence of simple, square integrable and non-anticipating functions such that

$$\mathbb{E}\{\int_0^T (f_n - f)^2 dt\} \to 0$$

as $n \to 0$.

We know that $I_n(t,\omega)$ are continuous martingales and by the Ito Isometry we see that they are bounded:

$$\mathbb{E}\{I_n(t,\omega)^2\} = \mathbb{E}\{(\int_0^t f_n dB)^2\} = \mathbb{E}\{\int_0^t f_n^2 dt\} < \infty$$

Recall briefly the types of convergence for a sequence of rvs $\{X_n\}$:
Convergence in p^{th}-mean:

$$X_n \to X \quad \text{if} \quad \mathbb{E}|X_n - X|^p \to 0, \quad \text{as } n \to \infty$$

Convergence in Probability:

$$X_n \xrightarrow{P} X \quad \text{if} \quad \mathrm{P}\{|X_n - X| > \epsilon\} \to 0 \quad \text{as } n \to \infty, \quad \forall \delta > 0$$

Convergence Almost Surely:

$$X_n \xrightarrow{\text{a.s.}} X \quad \text{if} \quad X_n(\omega) \to X(\omega) \quad \text{as } n \to \infty \quad \text{, on any set of } \omega \in \Omega \text{ of probability 1}$$

We will show that the $I_n(t,\omega)$ converges as $n \to \infty$ in probability **uniformly in time**. Our approach is to show that $\{I_n\}$ is a Cauchy Sequence uniformly in t, in probability. That is, $\forall \delta > 0$, and $\epsilon \geq 0$ we have that:

$$\mathrm{P}\left\{\max_{0 \leq t \leq T} |I_n(t) - I_m(t)| > \delta\right\} \leq \epsilon \quad \forall m, n \geq N(T, \delta, \epsilon).$$

We note that

$$\mathrm{P}\left\{\max_{0 \leq t \leq T} |I_n(t) - I_m(t)| > \delta\right\} \leq \frac{\mathbb{E}(I_n(T) - I_m(T))^2}{\delta^2}$$

by Kolmogorov's Inequality, so that on the right we have

$$= \frac{1}{\delta^2}\mathbb{E}\{(\int_0^T (f_n - f_m)dB)^2\}$$

and applying Ito's Isometry

$$= \frac{1}{\delta^2}\mathbb{E}\{\int_0^T (f_n - f_m)^2 dt\}$$

Now, since f_n converges in mean square, it is a Cauchy sequence. Thus $\{I_n(t)\}$ is a Cauchy sequence uniformly in t, in probability.

By an application of the Borel-Cantelli lemma, for any Cauchy sequence in probability there is a subsequence, $\{I_{n_k}(t,\omega)\}$, which converges to a limit $I(t,\omega)$ uniformly in $t \in [0,T]$, with probability 1. This limit is, moreover, independent of the subsequence so it is the stochastic integral, that is, it is unique for otherwise the approximating sequence could not be Cauchy in probability.

So, $I(t,\omega)$ is continuous with probability 1, since it is the uniform limit of continuous functions.

We note therefore that

$$I_t = I(t,\omega) = \int_0^t f(s,\omega)dB_s(\omega),\ 0 \le t \le T$$

is continuous in t and square integrable in (t,ω). It is also non-anticipating. Thus the stochastic integral is a map from square integrable non-anticipating $f(t,\omega)$ to square integrable non-anticipating functions $I(t,\omega)$, which are also continuous in t and have the martingale property.

More generally, we can also consider **Ito processes** of the form

$$I(t,\omega) = I(0,\omega) + \int_0^t b(s,\omega)ds + \int_0^t \sigma(s,\omega)dB_s(\omega)$$

where b is a non-anticipating integrable function and σ is a non-anticipating square integrable function.

8 Ito's Formula

Recall the identity

$$g(t_2) = g(t_1) + \int_{t_1}^{t_2} g'(s)ds$$

where we assume that g is differentiable. We can write symbolically

$$dg = g'ds$$

This comes about by assuming the second order term $(ds)^2$ is small compared with ds. However suppose the argument of g is Brownian Motion. As we have seen $(dB_t)^2$ is not small compared to dt because it is equal dt in the sense that the Q.V. of Brownian motion is equal to the time interval over which it is taken. Thus, we may expect that

$$g(B_{t_2}) - g(B_{t_1}) = \int_{t_1}^{t_2} g'(B_s)dB_s + \frac{1}{2}\int_{t_1}^{t_2} g''(B_s)ds$$

This is a correct identity, and it is a simple version of Ito's formula. In differential form we can write it as

$$dg(B_t) = g'(B_t)dB_t + \frac{1}{2}g''(B_t)dt$$

When $g(t,x)$ is a function of two variables and the first and second derivatives exist then Ito's formula has the form

$$g(t,B_t) = g(0,0) + \int_0^t \left(\frac{\partial g}{\partial t}(s,B_s) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(s,B_s)\right)ds + \int_0^t \frac{\partial g}{\partial x}(s,B_s)dB_s$$

Note that when g satisfies the backward in time heat equation

$$g_t + \frac{1}{2}g_{xx} = 0$$

then $g(t,B_t)$ is a martingale because it is equal to a stochastic integral, up to an additive constant.

To see how Ito's formula arises, assume that g has three bounded derivatives and consider for a partition of the time interval $t_0 = 0 < t_1 < \ldots < t_N = T$ the difference

$$\begin{aligned} g(B_T) - g(0) &= \sum_{k=0}^{N-1} \big(g(B_{t_{k+1}}) - g(B_{t_k})\big) \\ &= \sum_{k=0}^{N-1} g'(B_{t_k})(B_{t_{k+1}} - B_{t_k}) + \frac{1}{2}\sum_{k=0}^{N-1} g''(B_{t_k})(B_{t_{k+1}} - B_{t_k})^2 + \frac{1}{6}\sum_{k=0}^{N-1} g'''(B_{t_k^*})(B_{t_{k+1}} - B_{t_k})^3 \\ &\approx \int_0^T g'(B_s)dB_s + \frac{1}{2}\int_0^T g''(B_s)ds + \frac{1}{6}\sum_{k=0}^{N-1} g'''(B_{t_k^*})(B_{t_{k+1}} - B_{t_k})^3 \end{aligned}$$

where we have used Taylor's expansion with remainder. In the integral with the second derivative we use the Q.V. property of Brownian motion, which is the key step in the analysis, which is similar to getting the quadratic variation of a stochastic integral in the previous chapter. And the third derivative argument is evaluated at an intermediate time point t_k^*. We would like to show also that the last, third derivative term goes to zero as $N \to \infty$. We have

$$\left| \frac{1}{6} \sum_{k=0}^{N-1} g'''(B_{t_k^*})(B_{t_{k+1}} - B_{t_k})^3 \right| \leq C \sum_{k=0}^{N-1} |B_{t_{k+1}} - B_{t_k}|^3 = C \left(\max_k |B_{t_{k+1}} - B_{t_k}| \right) \sum_{k=0}^{N-1} (B_{t_{k+1}} - B_{t_k})^2$$

where C is a constant bound for the third derivative of g. Note that $\sum_{k=0}^{N-1} (B_{t_{k+1}} - B_{t_k})^2$ is bounded in mean square since in the limit it goes to the quadratic variation of B. We also know that in the limit that $N \to \infty$, $\max_k |B_{t_{k+1}} - B_{t_k}| \to 0$ by the time continuity of Brownian motion. Therefore, the third order term goes to zero in mean square.

In the case where g is a function of two variables, $g = g(t, x)$, we have

$$g(T, B_T) - g(0, 0) = \sum_{k=0}^{N-1} \big(g(t_{k+1}, B_{t_{k+1}}) - g(t_k, B_{t_k}) \big)$$

$$= \sum_{k=0}^{N-1} \big(g(t_{k+1}, B_{t_{k+1}}) - g(t_k, B_{t_{k+1}}) + g(t_k, B_{t_{k+1}}) - g(t_k, B_{t_k}) \big)$$

Now we can recognize the first two terms in the sum as an approximation the integral of $g_t(B_s)$ and the last two terms give the same result as before. That only two bounded x derivatives of g are needed can be shown by an approximation argument.

We summarize the discussion of the previous section, which relies on (a) the Q.V. property and (b) the continuity property of Brownian motion.

Let $I_t(\omega)$ be the Ito Process

$$I_t(\omega) = I_0(\omega) + \int_0^t b(s, \omega) ds + \int_0^t \sigma(s, \omega) dB_s(\omega)$$

where $b(t, \omega)$ and $\sigma(s, \omega)$ are non-anticipating functions satisfying

$$\mathbb{E} \int_0^T \sigma^2(s, \cdot) ds < \infty \quad \text{and} \quad \mathbb{E} \int_0^T |b(s, \cdot)| ds < \infty$$

In differential form we have, omitting the ω,

$$dI_t = b(t) dt + \sigma(t) dB_t$$

Let $g(t, x)$ be a function with bounded derivatives up to first order in t and up to second order in x. Then using Ito's formula $Y_t(\omega) = g(t, I_t(\omega))$ is also an Ito Process given by

$$Y_t(\omega) = Y_0(\omega) + \int_0^t \left(\frac{\partial g}{\partial t}(s, I_s(\omega)) + \frac{1}{2} \sigma^2(s, \omega) \frac{\partial^2 g}{\partial x^2}(s, I_s(\omega)) + b(s, \omega) \frac{\partial g}{\partial x}(s, I_s(\omega)) \right) ds$$

$$+\int_0^t \sigma(s,\omega)\frac{\partial g}{\partial x}(s,I_s(\omega))dB_s(\omega)$$

If we define

$$\hat{b}(t,x,\omega) = g_t(t,x) + \frac{1}{2}\sigma^2(t,\omega)g_{xx}(t,x) + b(t,\omega)g_x(t,x) \quad \text{and} \quad \hat{\sigma}(t,x,\omega) = \sigma(t,\omega)g_x(t,x)$$

then $Y_t(\omega)$ is itself an Ito Process with $\hat{b} = \hat{b}(t, I_t(\omega), \omega)$ and $\hat{\sigma} = \hat{\sigma}(t, I_t(\omega), \omega)$. In differential form we write, omitting the ω,

$$dY_t = \hat{b}(t, I_t)dt + \hat{\sigma}(t, I_t)dB_t$$

8.1 Examples

Find $\mathbb{E}(\int_0^T \sigma(s,\cdot)dB_s(\cdot))^{2p}$ when σ is non-anticipating and $|\sigma(t,\omega)| \leq C$.

We use Ito's Formula with $g(x) = x^{2p}$, $g'(x) = 2px^{2p-1}$ and $g''(x) = 2p(2p-1)x^{2p-2}$. We disregard the issue of unboundedness at this time and will consider it in detail in the solved problems. We have that

$$g(I_t) = 2p(2p-1)\int_0^t \frac{1}{2}\sigma^2(s,\omega)I_s^{2p-2}(s,\omega)ds + \int_0^t 2p\sigma(s,\omega)I_s^{2p-1}(s,\omega)dB_s(\omega)$$

Taking expectations, the stochastic integral has mean zero (assuming it is well defined) and we have

$$u_p(t) = \mathbb{E}\{I_t^{2p}\} = p(2p-1)\mathbb{E}\{\int_0^t \sigma^2(s,\cdot)I_s^{2p-2}(s,\cdot)ds\} \leq p(2p-1)C^2\mathbb{E}\{\int_0^t I_s^{2p-2}(s,\cdot)ds\}$$

$$= C^2p(2p-1)\int_0^t u_{p-1}(s)ds$$

When $p = 1$ we have

$$\mathbb{E}\{I_t^2\} = \int_0^t \mathbb{E}\{\sigma^2\}ds \leq C^2t$$

Thus

$$u_2(t) < 6C^2\int_0^t C^2sds = 3C^4t^2$$

Therefore

$$\mathbb{E}\{I_t^4\} = \mathbb{E}\{\left(\int_0^t \sigma(s,\cdot)dB_s\right)^4\} \leq 3(C^2t)^2.$$

Recall that if $\sigma = 1$ then $\mathbb{E}\{B_t^4\} = 3t^2$.

Define $I(t,\omega) = \int_0^t \sigma(s,\omega)dB_s(\omega)$, with $|\sigma(t,\omega)| \leq C$, σ non-anticipating and $\alpha \in \mathbb{R}$. Define

$$M_\alpha(t) = e^{\alpha I(t,\omega) - \frac{\alpha^2}{2}\int_0^t \sigma^2(s,\omega)ds}$$

Note that when we take $\sigma = 1$, $M_\alpha(t) = \exp(\alpha B_t - \alpha^2 t/2)$, then $\mathbb{E}M_\alpha(t) = 1$, and it is a martingale, as was shown earlier. We would now like to generalize this to the case where σ is a bounded non-anticipating function. This is done below.

Let X_t and Y_t be two Ito Processes (with the same Brownian Motion). Then for a function $f(t,x,y)$ we extend Ito's formula in the form

$$df(t,X_t,Y_t) = f_t dt + f_x dX_t + f_y dY_t + \frac{1}{2}f_{xx}(dX_t)^2 + \frac{2}{2}f_{xy}dX_t dY_t + \frac{1}{2}f_{yy}(dY_t)^2$$

where $(dX_t)^2$, $(dY_t)^2$ and $dX_t dY_t$ are the infinitesimal form of the quadratic variations for X_t, Y_t and the cross-quadratic variation of X_t and Y_t. The latter is defined as the limit of products of differences of X_t and Y_t over partitions of the time interval.

We note that

$$d(X_tY_t) = Y_t dX_t + X_t dY_t + dX_t dX_t$$

and if

$$dX_t = b_1 dt + \sigma_1 dB_t \qquad \text{and} \qquad dY_t = b_2 dt + \sigma_2 dB_t$$

and $f(t,x,y) = x \cdot y$, then:

$$df(X_t,Y_t) = d(X_tY_t) = Y_t(b_1 dt + \sigma_1 dB_t) + X_t(b_2 dt + \sigma_2 dB_t) + \sigma_1\sigma_2 dt$$

$$= (b_1 Y_t + b_2 X_t + \sigma_1\sigma_2)dt + (\sigma_1 Y_t + \sigma_2 X_t)dB_t$$

Ito's formula generalizes to the case were we have n independent Brownian Motions $B_t^{(1)}(\omega), \cdots, B_t^{(n)}(\omega)$. If $b_j(t,\omega)$ and $\sigma_{jk}(t,\omega)$ are non-anticipating square integrable functions for $j,k = 1,2,\cdots,n$ let

$$X_t^{(j)}(\omega) = x_0^j + \int_0^t b_j(s,\omega)ds + \int_0^t \sum_{k=1}^n \sigma_{j,k}(s,\omega)dB_s^{(k)}(\omega), \quad j = 1,2,\cdots,n.$$

Then Ito's Formula has the form

$$df(t,X_t^{(1)},X_t^{(2)},\cdots,X_t^{(n)}) = df(t,\vec{X}_t) = f_t dt + \sum_{j=1}^n f_{x_j}dX_t^{(j)} + \frac{1}{2}\sum_{j,k=1}^n f_{x_j,x_k}dX_t^{(j)}dX_t^{(k)}$$

with

$$dX_t^{(j)} = b_j dt + \sum_{k=1}^n \sigma_{jk}dB_t^{(k)}$$

For independent, standard Brownian motions, $dB_t^{(j)} dB_t^{(k)} = \delta_{jk} dt$ and therefore

$$dX_t^{(j)} dX_t^{(k)} = \sum_{\ell=1}^{n} \sigma_{j\ell}\sigma_{k\ell} dt = (\Sigma\Sigma^T)_{jk} dt, \quad \Sigma = (\sigma_{jk}), \quad j,k = 1,2,\cdots,n$$

Note that Σ is not necessarily symmetric. Note also that we can rewrite the above in vector notation

$$d\vec{X}_t = \vec{b}dt + \Sigma d\vec{B}_t$$

$$df(t,\vec{X}_t) = f_t dt + \nabla f(\vec{X}_t)\cdot d\vec{X}_t + \frac{1}{2} d\vec{X}_t^T H(\vec{X}_t) d\vec{X}_t$$

where $H(\vec{X}_t)_{jk} = f_{x_j x_k}$. Therefore

$$df(t,\vec{X}_t) = \left(f_t + \nabla f(\vec{X}_t)\cdot\vec{b} + \frac{1}{2}\mathrm{Tr}\{H(\vec{X}_t)\Sigma\Sigma^T\} \right) dt + \nabla f(\vec{X}_t)\cdot\Sigma d\vec{B}_t$$

8.2 Stopping Times

A non-negative random variable $\tau(\omega) : \Omega \to [0,\infty)$ is a Brownian Stopping Time if

$$\{\omega \in \Omega : \tau(\omega) \le t\} \in \mathcal{F}_t, \forall\ 0 \le t \le T$$

In other words, the occurrence of the event $\{\tau \le t\}$ can be decided if we only know the Brownian path up to time t, that is, $B_s(\omega)$, $s \le t$.

$$\tau_N = \inf\{t > 0 : |I(t)| > N\}$$

is a stopping time when

$$I(t,\omega) = \int_0^t \sigma(s,\omega) dB_s(\omega)$$

Note that the two events

$$\{\tau_N \le t\} \quad \text{and} \quad \{\max_{0\le s\le t} |I(s)| \ge N\}$$

are equivalent because of the continuity of $I(t)$ with respect to t, and the second clearly is in $\mathcal{F}_t$.

We now estimate the probability that this stopping time is less than some given time t using the Kolmogorov inequality. We have

$$\mathrm{P}\{\tau_N \le t\} = \mathrm{P}\left\{ \max_{0\le s\le t} |I(s)| > N \right\}$$

and by the Kolmogorov Inequality (since $I(t)$ is a zero mean square integrable martingale)

$$\leq \frac{\mathbb{E}|I(t)|^2}{N^2}.$$

We now apply the Ito Isometry,

$$= \frac{\mathbb{E}\{\int_0^t \sigma^2(s,\omega)ds\}}{N^2}$$
$$\leq \frac{c^2 t}{N^2}.$$

Therefore, we see that $\mathrm{P}\{\tau_N \leq t\} \to 0$ as $N \to \infty$ for any fixed t and $\mathrm{P}\{\tau_N > t\} \to 1$ as $N \to \infty$ for all t. From this we now conclude that $I(t)$ remains bounded with probability one for any finite t, assuming that $|\sigma|$ is bounded.

Let $A_N = \{\omega \in \Omega : \tau_N \leq t\}$ and consider

$$\sum_{N=1}^{\infty} \mathrm{P}\{A_N\} \leq c^2 t \sum_{N=1}^{\infty} \frac{1}{N^2} < \infty.$$

We now apply the Borel-Cantelli lemma

$$\mathrm{P}\{\cap_{N=1}^{\infty} \cup_{n=N}^{\infty} \{\tau_n \leq t\}\} = \mathrm{P}\{\tau_N \leq t, \text{i.o.}\} = \lim_{N\to\infty} \mathrm{P}\{\cup_{n=N}^{\infty}\{\tau_n \leq t\}\}$$

$$\leq \lim_{N\to\infty} \sum_{n=N}^{\infty} \mathrm{P}\{\tau_n \leq t\} \leq \lim_{N\to\infty} C^2 t \sum_{n=N}^{\infty} \frac{1}{n^2} = 0$$

Thus we have that $\tau_N > t$ as $N \to \infty$ w.p.1 for any finite t, and therefore $\tau_N \to \infty$ w.p.1 as $N \to \infty$.

8.3 Optional Stopping Theorem

It is clear that it would be very useful if there was a lemma that says: the martingale property is preserved with t and s replaced by stopping times. For this purpose we introduce the notion of a σ-algebra associated with a stopping time. Let $\mathcal{F}$ be the smallest σ-algebra generated by $\mathcal{F}_t$ for all $t \geq 0$. $\mathcal{F}_\tau \subset \mathcal{F}$ is the σ-algebra generated by ω sets of the form

$$A \cap \{\tau \leq t\} \in \mathcal{F}_t, \ A \in \mathcal{F}, \ 0 \leq t \leq \infty.$$

Thus, we may say that $\mathcal{F}_\tau$ contains information about paths up to time τ.

Optional Stopping Theorem Let $M(t)$ be a continuous martingale and let $0 \leq \rho(\omega) \leq \tau(\omega) \leq T < \infty$ where $\rho(\omega)$ and $\tau(\omega)$ are two bounded stopping times. Suppose $\mathbb{E}|M(t)| \leq \infty$ for $0 \leq t \leq T$. Then

$$\mathbb{E}\{M(\tau)|\mathcal{F}_\rho\} = M(\rho)$$

For finite-valued τ and ρ) we consider the following.

Let τ take the values $\tau_1 < \tau_2 < \cdots < \tau_N \leq T$ and let $\tau_0 < \tau_1$, be fixed. Note that $M(\tau)$ can be written as

$$M(\tau) = M(\tau)\sum_{k=1}^{N} \chi_{\tau=\tau_k} = \sum_{k=1}^{N} M(\tau_k)\chi_{\tau=\tau_k} = \sum_{k=1}^{N} M(\tau_k)\left[\chi_{\tau>\tau_{k-1}} - \chi_{\tau>\tau_k}\right]$$

$$\begin{aligned}
\mathbb{E}\{M(\tau)|\mathcal{F}_{\tau_0}\} &= \mathbb{E}\{\sum_{k=1}^{N} M(\tau_k)\left[\chi_{\tau>\tau_{k-1}} - \chi_{\tau>\tau_k}\right]|\mathcal{F}_{\tau_0}\} \\
&= \sum_{k=1}^{N} \mathbb{E}\{\mathbb{E}\{M(\tau_k)(\chi_{\tau>\tau_{k-1}} - \tau_{\tau>\tau_k})|\mathcal{F}_{\tau_{k-1}}\}|\mathcal{F}_{\tau_0}\} \\
&= \sum_{k=1}^{N} \mathbb{E}\{M(\tau_{k-1})\chi_{\tau>\tau_{k-1}} - M(\tau_k)\chi_{\tau>\tau_k}|\mathcal{F}_{\tau_0}\} \\
&= \mathbb{E}\{M(\tau_0)\chi_{\tau>\tau_0}|\mathcal{F}_{\tau_0}\} \\
&= M(\tau_0)
\end{aligned}$$

Let $\phi(t,\omega)$ be any adapted r.v. and let $\phi(\rho,\omega)$ be $\mathcal{F}_\rho$ measurable. Consider

$$\begin{aligned}
\mathbb{E}\{M(\tau)\phi(\rho)\} &= \mathbb{E}\{\mathbb{E}\{M(\tau)|\mathcal{F}_\rho\}\phi(\rho)\} \\
&= \mathbb{E}\{M(\tau)\phi(\rho)\sum_{k=1}^{M} \chi_{\rho=\rho_k}\} \\
&= \sum_{k=1}^{M} \mathbb{E}\{M(\tau)\phi(\rho_k)\chi_{\rho=\rho_k}\} \\
&= \sum_{k=1}^{M} \mathbb{E}\{\mathbb{E}\{M(\tau)\phi(\rho_k)\chi_{\rho=\rho_k}|\mathcal{F}_{\rho_k}\}\} \\
&= \sum_{k=1}^{M} \mathbb{E}\{M(\rho_k)\phi(\rho_k)\chi_{\rho=\rho_k}\} \\
&= \mathbb{E}\{M(\rho)\phi(\rho)\sum_{k=1}^{M} \chi_{\rho=\rho_k}\} \\
&= \mathbb{E}\{M(\rho)\phi(\rho)\}
\end{aligned}$$

Thus

$$\mathbb{E}\{(\mathbb{E}\{M(\tau)|\mathcal{F}_\rho\} - M(\rho))\,\phi(\rho)\} = 0$$

for all bounded and $\mathcal{F}_\rho$ measurable r.v. ϕ. This implies that with probability one

$$\mathbb{E}\{M(\tau)|\mathcal{F}_\rho\} = M(\rho)$$

The OST as stated is obtained from the finite valued case by a limiting argument. The boundedness of the stopping times is an essential hypothesis which when violated invalidates the OST, as can be seen with simple counter-examples.

8.4 The Exponential Martingale

We return to Example 4.2:

$$M_\alpha(t) = e^{\alpha I(t) - \frac{\alpha^2}{2}\int_0^t \sigma^2(s)ds}.$$

We restrict the time interval of $M_\alpha(s)$ to $0 \le s \le t \wedge \tau_N = \min(\tau_N, t)$ where

$$\tau_N = \inf\{t > 0 : |I(t)| > N\}.$$

We can now apply Ito's Formula to $M_\alpha(s)$ since

$$M_\alpha(s) \le e^{N\alpha}$$

We take

$$f(t,x;\omega) = e^{\alpha x - \frac{\alpha^2}{2}\int_0^t \sigma^2(s,\omega)ds} \Rightarrow M_\alpha(t,\omega) = f(t, I(t,\omega);\omega)$$

Thus

$$\begin{aligned} dM_\alpha(t) &= \frac{df}{dt}dt + \frac{df}{dx}dI + \frac{1}{2}\frac{d^2f}{dx^2}(dI)^2 \\ &= -\frac{\alpha^2}{2}\sigma^2(t,\omega)M_\alpha(t)dt + \alpha\sigma(t,\omega)M_\alpha(t)dB_t + \frac{1}{2}\alpha^2 M_\alpha(t)\sigma^2(t,\omega)dt \\ &= \alpha\sigma(t,\omega)M_\alpha(t)dB_t \end{aligned}$$

This implies that

$$M_\alpha(s) = 1 + \int_0^s \alpha\sigma(s',\omega)M_\alpha(s',\omega)dB_{s'}(\omega).$$

In particular

$$M_\alpha(t\wedge\tau_N) = 1 + \int_0^{t\wedge\tau_N} \alpha\sigma(s,\omega)M_\alpha(s,\omega)dB_s(\omega).$$

We know that $\tau_N \to \infty$ w.p.1. Thus, using the OST

$$\mathbb{E}\{M_\alpha(t\wedge\tau_N)\} = 1 = \mathbb{E}\{M_\alpha(t)\chi_{\tau_N>t}\} + \mathbb{E}\{M_\alpha(t)\chi_{\tau_N\le t}\} \ge \mathbb{E}\{M_\alpha(t)\chi_{\tau_N>t}\}$$

Since $\tau_N \to \infty$ w.p.1 we have by Fatou's lemma (stated and proved at the end of this discussion)

$$\mathbb{E}\{M_\alpha(t)\} \leq 1$$

If we can show that $\mathbb{E}\{M_\alpha(t)\} = 1$ then we have essentially shown that $M_\alpha(t)$ is a continuous (non-negative) martingale with moments of all orders. Squaring and using the upper bound we have

$$\begin{aligned}
\mathbb{E}\{M_\alpha^2(t \wedge \tau_N)\} &= \mathbb{E}\{e^{2\alpha I(t\wedge\tau_N) - \alpha^2 \int_0^{t\wedge\tau_n} \sigma^2(s)ds}\} \\
&= \mathbb{E}\{M_{2\alpha}(t \wedge \tau_N) e^{\alpha^2 \int_0^{\tau\wedge\tau_N} \sigma^2(s)ds}\} \\
\leq e^{\alpha^2 t c^2} \mathbb{E}\{M_{2\alpha}(t \wedge \tau_N)\} & \\
&\leq e^{\alpha^2 t c^2}
\end{aligned}$$

Let $N \to \infty$ so that $\tau_N \to \infty$. Then we have by Fatou's lemma

$$\mathbb{E}\{M_\alpha^2(t)\} \leq e^{\alpha^2 t c^2}$$

Now note that

$$\mathbb{E}\{M_\alpha(t \wedge \tau_N)\} = 1 \leq \mathbb{E}\{M_\alpha^2(t)\}^{\frac{1}{2}} \mathbb{E}\{\chi_{\tau_N \leq t}\}^{\frac{1}{2}} + \mathbb{E}\{M_\alpha(t)\}$$

which implies that$1 \leq \mathbb{E}\{M_\alpha(t)\}$ and hence

$$\mathbb{E}\{M_\alpha(t)\} = 1$$

This almost completes the proof that $M_\alpha(t)$ is a (non-negative) martingale with moments of all orders and mean one. We need to verify also the martingale property in general: $E\{M_\alpha(t) - M_\alpha(s) \mid \mathcal{F}_s\} = 0$. This is equivalent to showing that $E\{M_\alpha(t,s) - 1 \mid \mathcal{F}_s\} = 0$, where

$$M_\alpha(t,s) = e^{\alpha(I(t)-I(s)) - \frac{\alpha^2}{2}\int_s^t \sigma^2(\gamma)d\gamma}$$

If we redefine $\tau_N = \inf\{t > s \mid |I(t) - I(s)| \geq N\}$ then we know that $E\{M_\alpha(t \wedge \tau_N, s) - 1 \mid \mathcal{F}_s\} = 0$, or equivalently for any bounded function $\phi(s)$ that is $\mathcal{F}_s$ measurable, $E\{(M_\alpha(t\wedge \tau_N, s) - 1)\phi(s)\} = 0$. We want to pass to the limit $N \to \infty$ inside the expectation since we do know (just shown above) that $\tau_N \to \infty$ with probability one. In order to do this we can use uniform integrability, which means that if $f_n \to f$ with probability one and $\lim_{M\to\infty} \sup_n E\{|f_n|\chi_{|f_n|>M}\} = 0$ then $E\{f_n\} \to E\{f\}$. We can use this here because $M_\alpha(t \wedge \tau_N, s)$ has uniformly in N bounded second moments. We also recall here Fatou's lemma, which we have already used.

Fatou's Lemma Let X_n be non-negative random variables that tend to X with probability one as $n \to \infty$. Since $X_n \wedge M \leq M$ for any constant M, we can pass to the limit under the expectation and get $\lim E\{X_n \wedge M\} = E\{X \wedge M\}$. But, $\liminf E\{X_n\} \geq \lim E\{X_n \wedge M\}$

and so $\liminf E\{X_n\} \geq E\{X \wedge M\}$. Letting M go to infinity in this last inequality we get Fatou's lemma:

$$\liminf E\{X_n\} \geq E\{X\}.$$

A somewhat more general statement is that $\liminf E\{X_n\} \geq E\{\liminf X_n\}$.

If $I(t,\omega) = \int_0^t \sigma(s,\omega) dB_s(\omega)$ with $|\sigma(t,\omega)| \leq C$ and non-anticipating, then for any α,

$$M_\alpha(t) = e^{\alpha I(t) - \frac{\alpha^2}{2}\int_0^t \sigma^2(s)ds}$$

is a non-negative, continuous martingale with moments of all orders (i.e. finite $\mathbb{E}M^p < \infty$). The moments can be bounded by

$$\mathbb{E}(M_\alpha(t))^p = E\{M_{p\alpha}(t) e^{\frac{\alpha^2 p(p-1)}{2}\int_0^t \sigma^2(s)ds}\} \leq e^{\frac{\alpha^2 p(p-1)c^2 t}{2}}$$

8.5 The Levy characterization of Brownian motion

We note first that all of the basic theory of stochastic integration and Ito's formula goes through when we replace Brownian motion B_t by a continuous, square integrable martingale X_t that has a given quadratic variation. We will use this fact to show that if in fact the quadratic variation of this martingale X_t is t then it is a Brownian motion, that is, its law is the Brownian motion law.

For any $\alpha \in \mathbb{R}$ let

$$M_t = e^{i\alpha X_t + \frac{1}{2}\alpha^2 t}$$

By Ito's formula we have

$$dM_t = i\alpha M_t dX_t + \frac{1}{2}\alpha^2 M_t dt - \frac{1}{2}\alpha^2 M_t d < X_t >$$

where $< X_t >$ is the quadratic variation of X. But by assumption $< X_t >= t$ and therefore we have that

$$dM_t = i\alpha M_t dX_t$$

which means that M_t is a continuous, integrable (bounded) martingale

$$\mathbb{E}\{M_t | \mathcal{F}_s\} = M_s$$

Therefore

$$\mathbb{E}\{e^{i\alpha X_t + \frac{1}{2}\alpha^2 t} | \mathcal{F}_s\} = e^{i\alpha X_s + \frac{1}{2}\alpha^2 s}$$

or

$$\mathbb{E}\{e^{i\alpha(X_t - X_s)} | \mathcal{F}_s\} = e^{-\frac{1}{2}\alpha^2(t-s)},\ 0 \leq s \leq t < \infty$$

This implies that X_t has Gaussian independent increments with mean zero and variance the length of the increment, which is the law of Brownian motion.

8.6 Moments of stochastic integrals

We will use Ito's formula to show that if $\sigma(t,\omega)$ is a bounded non-anticipating functional, $|\sigma| \le C$, then for the stochastic integral $I(t,\omega) = \int_0^t \sigma(s,\omega)dB_s(\omega)$ we have the moment estimates

$$E[I^{2p}(t)] \le 1 \cdot 3 \cdot 5 \cdots (2p-1)(C^2 t)^p$$

for $p = 1, 2, 3, ...$ How do we show that the mean of the martingale term in Ito's formula is zero? We will use stopping times as follows.

Let $\tau_N = \inf\{t > 0 \mid |I(t)| \ge N\}$. From Ito's formula we have that

$$I^{2p}(t \wedge \tau_N) = p(2p-1)\int_0^{t\wedge\tau_N} I^{2p-2}(s)\sigma^2(s)ds + 2p\int_0^{t\wedge\tau_N} I^{2p-1}(s)\sigma(s)dB(s)$$

The stochastic integral is well defined because for $0 \le s \le t \wedge \tau_N$ the integrand in bounded. More precisely, the stochastic integral

$$\int_0^t \chi_{\{s\le\tau_N\}} I^{2p-1}(s)\sigma(s)dB(s),\ t \ge 0,$$

has a bounded integrand and it is therefore an integrable martingale. Moreover,

$$\int_0^{t\wedge\tau_N} \chi_{\{s\le\tau_N\}} I^{2p-1}(s)\sigma(s)dB(s) = \int_0^{t\wedge\tau_N} I^{2p-1}(s)\sigma(s)dB(s).$$

We can now apply the optional stopping theorem so its expectation is zero. Therefore

$$E[I^{2p}(t \wedge \tau_N)] = p(2p-1)E[\int_0^{t\wedge\tau_N} I^{2p-2}(s)\sigma^2(s)ds]$$

or

$$E[I^{2p}(t \wedge \tau_N)] \le p(2p-1)\int_0^t C^2 E[I^{2p-2}(s)]ds$$

Since $\tau_N \to \infty$ as $N \to \infty$ we conclude from Fatou's lemma that

$$E[I^{2p}(t)] \le p(2p-1)\int_0^t C^2 E[I^{2p-2}(s)]ds$$

Starting with $p = 1$ in this inequality and iterating forward, we get the resulting estimate.

8.7 Martingales and PDEs

What equation must the assumed smooth and bounded function $u(t,x)$ satisfy so that

$$u(t, B_t)e^{\int_0^t V(B_s)ds}$$

is a martingale? Here $V(x)$ is a bounded function. We will use Ito's formula to address this as follows.

a) Since u is bounded and letting $X_t = x + B_t$, $(dX_t)(dX_t) = dt$, we can apply Ito's formula to get

$$du(t, X_t) = (u_t(t, X_t) + \frac{1}{2}u_{xx}(t, X_t))dt + u_x dB_t$$

and in integral form

$$u(t, X_t) - u(0, x) = \int_0^t (u_t(s, X_s) + \frac{1}{2}u_{xx}(s, X_s))ds + \int_0^t u_x(s, X_s)dB_s.$$

A sufficient condition for the above to be a Martingale is

$$u_t + \frac{1}{2}u_{xx} = 0.$$

b) We notice that u solves the backward heat equation so it is natural that it be provided with terminal conditions at some time T and then solved for $t < T$. If we let $u(T, x) = f(x)$ for some bounded function with two bounded derivatives, then from Ito's formula above we have

$$u(0, x) = E\{f(X_T)|X_0 = x\} = E_x\{f(X_T)\} = E\{f(x + B_T)\}$$

More generally, we have that

$$u(t, x) = u(t, x; T) = E\{f(X_T)|X_t = x\} = E_{t,x}\{f(X_T)\}$$

Note that because of time homogeneity, the solution is a function of $T - t$. If for $t \leq T$ we let $u(t, x; T) = v(T - t, x)$ then we see that $v(t, x)$ satisfies the ordinary heat equation

$$v_t = \frac{1}{2}v_{xx} \ , \quad t > 0$$

with initial condition $v(0, x) = f(x)$. The probabilistic representation of this solution is $v(t, x) = E_x\{f(X_t)\}$.

c) Let

$$Y_t = e^{\int_0^t V(s,X_s)ds}u(t, X_t);\ R_t = \int_0^t V(s, X_s)ds;\ ;\ U_t = u(t, X_t)$$

By definition we have $dR_t = V(t, X_t)dt$. Next using the product rule, which is equivalent to the 2-dim Ito formula applied to the function $f(x_1, x_2) = x_1x_2$, we have

$$dY_t = d(e^{R_t}U_t) = d(e^{R_t})U_t + e^{R_t}d(U_t) + d(e^{R_t})dU_t$$

Computing by Ito

$$dU_t = (u_t(t, X_t) + \frac{1}{2}u_{xx}(t, X_t))dt + u_x dB_t$$

$$d(e^{R_t}) = e^{R_t} d(R_t) + \frac{1}{2} e^{R_t} d(R_t)^2 = e^{R_t} d(R_t) = e^{\int_0^t V(s,X_s)ds} V(t,X_t)dt$$

and noting from the above $d(e^{R_t})dU_t = 0$ since $dtdt = 0$, and $dtdB_t = 0$ we finally get

$$dY_t = e^{\int_0^t V(s,X_s)ds} V(t,X_t)u(t,X_t)dt + e^{\int_0^t V(s,X_s)ds}[(u_t(t,X_t) + \frac{1}{2}u_{xx}(t,X_t))dt + u_x dB_t] =$$

$$= e^{\int_0^t V(s,X_s)ds} \left[\left(V(t,X_t)u(t,X_t) + u_t(t,X_t) + \frac{1}{2}u_{xx}(t,X_t) \right) dt + u_x(t,X_t)dB_t \right],$$

so we conclude that a sufficient condition for Y_t to be a martingale is

$$Vu + u_t + \frac{1}{2}u_{xx} = 0.$$

d) As in (b) above it is appropriate to prescribe terminal values for this PDE, $u(T,x) = f(x)$. From the integral form of the martingale Y_t we then find, after taking expectations, that

$$u(t,x;T) = E\{e^{\int_t^T V(s,X_s)ds} f(X_T)|X_t = x\} = E_{t,x}\{e^{\int_t^T V(s,X_s)ds} f(X_T)\}$$

Since $V = V(t,x)$ depends on t, the equation is not time homogeneous and so the solution depends on both the terminal time T and the starting time t. It is not a function of $T-t$ as in (b).

9 Stochastic control

We will formulate a stochastic control problem for a finite difference equation, or recursion, of the form

$$X_{n+1} = X_n + b(U_n, X_n)\Delta t + \sigma(U_n, X_n)\sqrt{\Delta t} Z_{n+1} \ , \quad n = 0, 1, 2, ...$$

where $X_0 = x$ is given and $\{Z_n\}$ are iid $N(0,1)$ random variables. This is a Markov chain with values in $\mathbb{R}$ and with coefficient functions $b(u,x)$ and $\sigma(u,x)$ that depend on both x and the control variable $u \in \mathbb{R}$. The transition probabilities are given by

$$P\{X_{n+1} \in A | X_n = x, U_n = u\} = \frac{1}{\sqrt{2\pi\sigma^2(u,x)\Delta t}} \int_A e^{-\frac{(y-x-b(u,x)\Delta t)^2}{2\sigma^2(u,x)\Delta t}} dy$$

For any bounded function $f(x)$ we define the transition operator

$$P_u f(x) = \frac{1}{\sqrt{2\pi\sigma^2(u,x)\Delta t}} \int e^{-\frac{(y-x-b(u,x)\Delta t)^2}{2\sigma^2(u,x)\Delta t}} f(y) dy$$

The reason the scaling with the time increment Δt is chosen this way is so that conditional increment of the Markov chain has the following mean and variance

$$E\{X_{n+1} - X_n | X_n = x, U_n = u\} = b(u,x)\Delta t$$

$$\mathrm{Var}\{X_{n+1} - X_n | X_n = x, U_n = u\} = \sigma^2(u,x)\Delta t$$

which implies that for any smooth and bounded function f we have that

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t}(P_u f(x) - f(x)) = \frac{\sigma^2(u,x)}{2}\frac{\partial^2 f(x)}{\partial x^2} + b(u,x)\frac{\partial f(x)}{\partial x} = \mathcal{L}_u f(x)$$

Thus, the continuum limit of the Markov chain exists and is a diffusion process. The above limit is analyzed in more detail in a later section.

The control of the process X_n is U_n, assumed to be a real valued random variable that depends only on $X_0, X_1, \ldots, X_n$, that is, it is nonanticipating, and such that

$$E_x\{\sum_{j=0}^{N-1} L(U_j, X_j)\Delta t + g(X_N)\}$$

is minimized over all control sequences $U_0, U_1, ..., U_{N-1}$, for some terminal time index N and for a given running cost function $L(u,x)$ and terminal cost $g(x)$. We introduce the value function

$$V^n(x) = \inf_{\mathcal{A}_n} E\{\sum_{j=n}^{N-1} L(U_j, X_j)\Delta t + g(X_N) | X_n = x\} \ , \quad n = N-1, N-2, ..., 0$$

where $\mathcal{A}_n$ is the set of all admissible control sequences starting at time n. Note that $V^N(x) = g(x)$. Clearly we want to find $V^0(x)$ and the associated optimal control sequence $U_0^*, U_1^*, ..., U_{N-1}^*$.

The main result in discrete time stochastic control is the derivation of the backward in time recursion for the determination of the value function V^0 and the associated optimal control. Using iterated conditional expectation we introduce heuristically the optimality principle

$$\begin{aligned}
V^n(x) &= \inf_{\mathcal{A}_n} E\{\sum_{j=n}^{N-1} L(U_j, X_j)\Delta t + g(X_N)|X_n = x\} \\
&= \inf_{\mathcal{A}_n} E\{E\{L(U_n, X_n)\Delta t + \sum_{j=n+1}^{N-1} L(U_j, X_j)\Delta t + g(X_N)|X_{n+1}, X_n = x\}|X_n = x\} \\
&= \inf_{u} E\{L(u, x)\Delta t + \inf_{\mathcal{A}_{n+1}} E\{\sum_{j=n+1}^{N-1} L(U_j, X_j)\Delta t + g(X_N)|X_{n+1}\}|X_n = x, U_n = u\} \\
&= \inf_{u}[L(u, x)\Delta t + E\{V^{n+1}(X_{n+1})|X_n = x, U_n = u\}] \\
&= \inf_{u}[L(u, x)\Delta t + P_u V^{n+1}(x)]
\end{aligned}$$

Therefore, the determination of the value function and the associated optimal control is reduced, heuristically so far, to solving the backward optimality recursion

$$V^n(x) = \inf_{u}[L(u, x)\Delta t + P_u V^{n+1}(x)] \ , \quad n = N-1, N-2, ..., 0$$

with $V^N(x) = g(x)$. This is the discrete time Hamilton-Jacobi-Bellman (HJB) equation.

Assuming that a unique minimum $u_n^*(x)$ exists at each step in the backward optimality recursion, which in general requires convexity and other assumptions on L, b, σ and g, we can then obtain the optimally controlled Markov chain recursively by

$$X_{n+1}^* = X_n^* + b(u_n^*(X_n^*), X_n^*)\Delta t + \sigma(u_n^*(X_n^*), X_n^*)\sqrt{\Delta t} Z_{n+1} \ , \quad n = 0, 1, 2, ...$$

with $X_0^* = x$. The optimal controls are given by $U_n^* = u_n^*(X_n^*)$ and they are Markovian, that is, they depend only on the current (optimal) state. The optimal value function satisfies

$$V^n(x) = L(u_n^*(x), x)\Delta t + P_{u_n^*(x)} V^{n+1}(x) \ , \quad n = N-1, N-2, ..., 0$$

with $V^N(x) = g(x)$. The optimal control problem is thus reduced to solving first the backward optimality recursion and then the forward recursion that determines the optimal state and the associated optimal controls.

9.1 The linear state, quadratic cost stochastic control

Let us consider a simple, linear, quadratic cost control problem in which $b(u,x) = bu,\ \sigma(u,x) = \sigma,\ L(u,x) = lu^2,\ g(x) = gx^2$, where b, σ, l, g are now constants and we take $\Delta t = 1$. In this case we may assume that the value function has the form

$$V^n(x) = a_n x^2 + b_n$$

and derive recursions for the sequences of constants $\{a_n\}$ and $\{b_n\}$, with $a_N = g$ and $b_N = 0$. Because of the Gaussian transition probability density for the Markov chain, the backward optimality recursion has the form

$$a_n x^2 + b_n = \inf_u [lu^2 + a_{n+1}(\sigma^2 + (x+bu)^2) + b_{n+1}]$$

The minimizing u is given by

$$u_n^*(x) = \frac{-ba_{n+1}}{l + b^2 a_{n+1}} \cdot x$$

and then the recursions for a_n and b_n have the form

$$a_n = \frac{l a_{n+1}}{l + b^2 a_{n+1}}\ ,\quad b_n = \sigma^2 a_{n+1} + b_{n+1}$$

with $a_N = g$ and $b_N = 0$. Once this sequence of constants is determined the optimally controlled Markov chain has the form

$$X_{n+1}^* = X_n^* + bu_n^*(X_n^*) + \sigma Z_{n+1}\ ,\quad n = 0, 1, 2, ...$$

with $X_0^* = x$. Note that the optimal control is a linear function of the state in this case.

9.2 The verification theorem

Let X_n satisfy as before the controlled recursion

$$X_{n+1} = X_n + b(U_n, X_n)\Delta t + \sigma(U_n, X_n)\Delta B_{n+1}\ ,\quad n = 0, 1, ...$$

with $X_0 = x$, and $U = (U_n)$ is a given non-anticipating sequence of controls. The random variables ΔB_{n+1} are just like before, i.i.d. Gaussian with mean zero and variance Δt (increments of Brownian motion B(t)). In the limit $n \to \infty,\ \Delta t \to 0$ with $n\Delta t = t$, X_n is close to $X(n\Delta t)$ in probability over any finite time trajectory, assuming the U_n converge also. The continuous time stochastic process $X(t)$ satisfies a controlled stochastic differential equation

$$dX(t) = b(U(t), X(t))dt + \sigma(U(t), X(t))dB(t)$$

for which the recursion above satisfied by X_n is the explicit, Euler, finite difference approximation. To indicate dependence on the control we write $X_n = X_n^U$. Let

$$V_n(x) = \inf_U E_{n,x}\{g(X_N^U)\} \ , \quad n \leq N,$$

where $N\Delta t = T$, be the discrete time value function of the optimal control problem. We assume for simplicity that the running cost $L(u, x)$ is zero here.

For a fixed control u, the transition probability density of X_n is as before given by

$$\pi_u(x_{n+1}|x_n) = \frac{1}{\sqrt{2\pi\sigma^2(u, x_n)\Delta t}} \exp\left\{\frac{-(x_{n+1} - x_n - b(u, x_n)\Delta t)^2}{2\sigma^2(u, x_n)\Delta t}\right\}.$$

Let us assume that the discrete HJB recursion of the previous section (with running cost L equal to zero)

$$v_{n-1}(x) = \inf_u \int \pi_u(y|x)v_n(y)dy \ , \quad 0 \leq n \leq N \ , \quad v_N(x) = g(x)$$

has a unique solution $v_n(x),\ n = N, N-1, \ldots$, with associated optimal controls $U^*(x) = (u_0^*(x), \ldots, u_{N-1}^*(x))$ given as functions of the state. With fixed controls $U = (u_0, u_1, \ldots, u_{N-1})$, with $v_n(x)$ the solution of the discrete HJB equation and using the identity

$$g(X_N^U) - v_n(X_n^U) = \sum_{k=n}^{N-1} \left(v_{k+1}(X_{k+1}^U) - v_k(X_k^U)\right)$$

we can show that

$$v_n(x) \leq E_{n,x}\{g(X_N^U)\}$$

and since the left does not depend on U,

$$v_n(x) \leq \inf_U E_{n,x}\{g(X_N^U)\}.$$

Repeating this calculation with the optimal control U^* we can show that equality holds, thus identifying the value function $V_n(x)$ with the solution of the discrete HJB equation $v_n(x)$. This is the verification theorem. It identifies the value function of the control problem with the solution of the HJB recursion and avoids the heuristic derivation using the principle of optimality.

From the identity given as well as the iterated conditional expectation and the Markov

property of X_n^U we have

$$\begin{aligned}
E_{n,x}\left[g(X_N^U)\right] &= E_{n,x}\left[v_n(X_n^U)\right] + \sum_{k=n}^{N-1} E_{n,x}\left[v_{k+1}(X_{k+1}^U)\right] - E_{n,x}\left[v_k(X_k^U)\right] \\
&= v_n(x) + \sum_{k=n}^{N-1} E_{n,x}\left[v_{k+1}(X_{k+1}^U)\right] - E_{n,x}\left[v_k(X_k^U)\right] \\
&= v_n(x) + \sum_{k=n}^{N-1} E_{n,x}\left[E_{n,x}\left[v_{k+1}(X_{k+1}^U)|X_k^U\right]\right] - E_{n,x}\left[v_k(X_k^U)\right] \\
&= v_n(x) + \sum_{k=n}^{N-1} E_{n,x}\left[E_{k,X_k^U}\left[v_{k+1}(X_{k+1}^U)\right]\right] - E_{n,x}\left[v_k(X_k^U)\right] \\
&\geq v_n(x) + \sum_{k=n}^{N-1} E_{n,x}\left[v_k(X_k^U)\right] - E_{n,x}\left[v_k(X_k^U)\right] \\
&= v_n(x).
\end{aligned}$$

The inequality comes from the definition of the recursive functions $v_n(x)$. This establishes the inequality

$$v_n(x) \leq \inf_U E_{n,x}\left[g(X_N^U)\right].$$

To establish equality, we can perform the same procedure but recognize that under U^*,

$$v_k(X_k^{U^*}) = E_{k,X_k^{U^*}}\left[v_{k+1}(X_{k+1}^{U^*})\right]$$

and therefore the cascading sum is identically zero and not just greater than or equal to zero. As noted already, this is the so-called verification theorem: starting from the solution of the discrete-time HJB recursion we identify its solution with the value function of the stochastic control problem.

9.3 The continuum limit

To see how we may pass to the continuum limit, we show first that for any smooth function $f(x)$ and for a fixed u we have that

$$\frac{1}{\Delta t}\left(\int \pi_u(y|x)f(y)dy - f(x)\right) \to \mathcal{L}_u f(x)$$

In more detail, we make a change of variable in the integral and then do a Taylor expansion for small Δt.

$$z = \frac{y - x - b(u,x)\Delta t}{\sigma(u,x)\sqrt{\Delta t}}.$$

The integral becomes

$$\int \pi_u(y|x)f(y)dy = \int \frac{1}{\sqrt{2\pi}} e^{-z^2/2} f(x + z\sigma(u,x)\sqrt{\Delta t} + b(u,x)\Delta t)dz,$$

and a Taylor expansion on f around x gives

$$\begin{aligned} f(x + z\sigma(u,x)\sqrt{\Delta t} + b(u,x)\Delta t) &= f(x) \\ &+ \left(z\sigma(u,x)\sqrt{\Delta t} + b(u,x)\Delta t\right)\frac{df}{dx} + \frac{1}{2}\left(z\sigma(u,x)\sqrt{\Delta t} + b(u,x)\Delta t\right)^2 \frac{d^2 f}{dx^2} \\ &+ \dots \end{aligned}$$

Rearranging gives

$$\begin{aligned} f(x + z\sigma(u,x)\sqrt{\Delta t} + b(u,x)\Delta t) &= f(x) + z\sigma(u,x)\sqrt{\Delta t}\frac{df}{dx} \\ &+ \Delta t\left(b(u,x)\frac{df}{dx} + z^2\frac{1}{2}\sigma(u,x)^2\frac{d^2 f}{dx^2}\right) + \mathcal{O}(\Delta t^{3/2}) \end{aligned}$$

Noting that

$$\int \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1, \quad \int \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} dz = 0, \quad \int \frac{1}{\sqrt{2\pi}} z^2 e^{-z^2/2} dz = 1,$$

we can substitute back into the integral to get

$$\begin{aligned} \int \pi_u(y|x)f(y)dy &= f(x) + \Delta t\left(b(u,x)\frac{df}{dx} + \frac{1}{2}\sigma(u,x)^2\frac{d^2 f}{dx^2}\right) + \mathcal{O}(\Delta t^{3/2}) \\ &= f(x) + \Delta t\mathcal{L}_u f(x) + \mathcal{O}(\Delta t^{3/2}) \end{aligned}$$

Finally, we get

$$\frac{1}{\Delta t}\left(\int \pi_u(y|x)f(y)dy - f(x)\right) = \mathcal{L}_u f(x) + \mathcal{O}(\Delta t^{1/2}).$$

We next argue heuristically that when $V_n(x) \approx V(t_n, x)$, where $V(t,x)$ is the solution of the continuous time HJB equation with $t_n = n\Delta t$. From the discrete HJB recursion satisfied by $V_n(x)$ we have that

$$\frac{1}{\Delta t}\left(V_n(x) - V_{n-1}(x)\right) + \frac{1}{\Delta t}\left(\inf_u \int \pi_u(y|x)V_n(y)dy - V_n(x)\right) = 0$$

This must be approximately zero also in the continuous time limit. Replacing $V_n(x)$ by the continuous time approximation $V(t_n, x)$ and from a Taylor expansion in Δt, we have

$$\begin{aligned} \frac{1}{\Delta t}\left(V(t_n,x) - V(t_{n-1},x)\right) + \frac{1}{\Delta t}&\left(\inf_u \int \pi_u(y|x)V(t_n,y)dy - V(t_n,x)\right) \\ &= V_t(t_n,x) + \inf_u \mathcal{L}_u V(t_n,x) + \mathcal{O}(\Delta t^{1/2}) \end{aligned}$$

assuming that the infimum can be interchanged with limit, which is in general a difficult problem to analyze. We thus have the HJB partial differential equation for $V(t,x)$

$$V_t(t,x) + \inf_u \mathcal{L}_u V(t,x) = 0,\ t < T,\ x \in \mathbb{R}$$

with the terminal condition $V(T,x) = g(x)$.

The rigorous passage of the discrete time control problem to the continuum limit can be technically difficult if we only have some minimal regularity of the limit HJB equation. One important point of the discrete approximation is that it is itself an optimal control problem. There are many discretizations of the continuum HJB equation that are not associated with any control problem. Such approximations may be problematic and may not be so stable, even though they may have formally higher order local truncation error.

10 The Poisson process

The Poisson process $N(t),\ t \geq 0$ with intensity λ is a pure jump process with integer values such that (i) $N(0) = 0$, (ii) $N(t)$ has independent increments, and (iii) for any $s \leq t$,

$$P(N(t) - N(s) = k) = e^{-\lambda(t-s)} \frac{[\lambda(t-s)]^k}{k!},\ k = 0, 1, \ldots \tag{31}$$

It is also assumed that $N(t)$ is normalized to be continuous from the right: $\lim_{h \downarrow 0} N(t+h) = N(t)$ with probability one. Independent increments means that the differences of the process over non-overlapping intervals are independent random variables. We denote by $\mathcal{F}_t$ the sigma algebra of events (collection of sets of paths closed under countable set operations) generated by cylinder sets of right-continuous paths with left hand limits, up to time t. It is also a right continuous family of increasing sigma algebras $\mathcal{F}_{t+} = \cap_{h>0} \mathcal{F}_{t+h} = \mathcal{F}_t$. By cylinder sets we mean sets of paths whose values are inside intervals for a finite but arbitrary number of times up to t. We denote by $\mathcal{F}_{t-}$ the smallest sigma algebra that contains the union of $\mathcal{F}_{t-h},\ h > 0$, and note that $\mathcal{F}_{t-} \subset \mathcal{F}_t$. The Poisson process is the simplest counting process as its paths are piecewise constant taking integer values. The jump times are denoted by $\tau_k,\ k = 0, 1, 2, \ldots$ with $\tau_0 = 0$ and $\tau_1 < \tau_2 < \cdots < \tau_k < \cdots$. The inter-jump times $\tau_{k+1} - \tau_k$ are independent random variables for different k that are exponentially distributed with parameter λ

$$\begin{aligned} P(\tau_{k+1} - \tau_k > t) &= P(N(\tau_k + t) - N(\tau_k) = 0) \\ &= E\{P(N(\tau_k + t) - N(\tau_k) = 0 | \tau_k)\} = e^{-\lambda t} \end{aligned}$$

Note that the Poisson process can be represented as

$$N(t) = \sum_{k=1}^{\infty} \mathbb{I}(\tau_k \leq t)$$

Direct calculation shows that $E(N(t)) = \lambda t$ so that the intensity λ is the average number of jumps per unit time. Moreover the difference $M(t) = N(t) - \lambda t$ is a mean zero right-continuous martingale, that is, for any $s \leq t$

$$E(M(t)|\mathcal{F}_s) = M(s)$$

Now we want to describe how a function of the Poisson process, $f(N(t))$, changes, where $f(n)$ is a bounded function on the integers. The difference $f(N(t)) - f(N(s))$ for $s < t$ is the sum of the jumps of this difference in the interval $(s, t]$

$$f(N(t)) - f(N(s)) = \sum_{s < \tau_k \leq t} \Delta f(N(\tau_k -))$$

where

$$\Delta f(N(\tau_k-)) = f(N(\tau_k)) - f(N(\tau_k-)) = f(N(\tau_k-)+1) - f(N(\tau_k-)) \tag{32}$$

We can write this sum as an integral, by definition,

$$f(N(t)) - f(N(s)) = \int_s^t \Delta f(N(\gamma-))dN(\gamma)$$

this being consistent with the sum since N jumps by a unit amount at the jump times. It is convenient to write this in terms of the martingale $M(t)$ by adding and subtracting λt,

$$f(N(t)) - f(N(s)) = \int_s^t \Delta f(N(\gamma-))\lambda d\gamma + \int_s^t \Delta f(N(\gamma-))dM(\gamma)$$

The reason this is convenient is because the expectation of the integral with respect to M is zero

$$E\left(\int_s^t \Delta f(N(\gamma-))dM(\gamma)|\mathcal{F}_s\right) = 0$$

This is seen clearly from the definition of the integral as a sum minus a Riemann integral with the same expectation. More explicitly, using the law of iterated conditional expectation we can write

$$E\left(\int_s^t \Delta f(N(\gamma-))dM(\gamma)|\mathcal{F}_s\right) = E\left(\int_s^t E\left(\Delta f(N(\gamma-))dM(\gamma)|\mathcal{F}_{\gamma-}\right)|\mathcal{F}_s\right)$$

But now in the inner expectation on the right $\Delta f(N(\gamma-))$ comes out of the expectation since it is known given $\mathcal{F}_{\gamma-}$. And then $E\left(dM(\gamma)|\mathcal{F}_{\gamma-}\right)$ is zero since this is the martingale property of M in its infinitesimal form. This gives the desired result. We have also shown, in particular, that

$$M_{\Delta f}(t) = \int_0^t \Delta f(N(\gamma-))dM(\gamma)$$

is a zero mean martingale

$$E(M_{\Delta f}(t)|\mathcal{F}_s) = M_{\Delta f}(s)$$

The above are easily extended to a function f that depends on time as well, $f = f(t,n)$. Then what amounts to **Ito's formula** for the Poisson process has the form

$$f(t,N(t)) = f(s,N(s)) + \int_s^t \left[f_t(\gamma,N(\gamma)) + \lambda\Delta f(\gamma,N(\gamma))\right]d\gamma + \int_s^t \Delta f(\gamma,N(\gamma-))dM(\gamma) \tag{33}$$

We do not need to evaluate the integrand at the left limit of the argument when dealing with Riemann integrals as instantaneous jumps to not contribute to the integral. Recall that Δf is defined by (32).

Now suppose we have a function $u(t,n)$ that satisfies the differential-difference equation

$$u_t(t,n) + \lambda[u(t,n+1) - u(t,n)] = 0,\ t < T,\ n \geq 0,\ u(T,n) = h(n) \tag{34}$$

which is the backward Kolmogorov equation for the Poisson process. Then using this instead of f in Ito's formula we see that

$$u(T,N(T)) - u(t,N(t)) = \int_t^T \Delta u(s,N(s-))dM(s)$$

Taking expectations and using the Markov property[4] to write $E(\ \cdot\ |\mathcal{F}_s) = E(\ \cdot\ |N(s))$ we have that

$$E(u(T,N(T))|N(t)) - u(t,N(t)) = 0$$

which says that $u(t,N(t))$ is a martingale when $u(t,n)$ is harmonic, that is, it satisfies (34). With $N(t) = n$ we have

$$u(t,n) = E(u(T,N(T))|N(t) = n) = E(h(N(T))|N(t) = n) \tag{35}$$

To summarize, we have derived Ito's formula (33), which we can also write as

$$df(t,N(t)) = [f_t(t,N(t)) + \lambda(f(t,N(t)+1) - f(t,N(t))]\,dt + \Delta f(t,N(t-))dM(t) \tag{36}$$

Note again that we do not need to specify $N(t-)$ in the dt terms but we need to do so with the multiplier of $dM(t)$ since it jumps as does the multiplier and we want $dM(t)$ to "point forward". This is often stated by saying that the function being integrated with $dM(t)$ as integrator is predictable or previsible. Using the solution of the backward Kolmogorov equation (34) in Ito's formula we obtain its representation as a conditional expectation in (35). When this argument is used with the Hamilton-Jacoby-Bellman equation it is called the verification lemma, as we do in Section 14. It is simply the use of Ito's formula as we did to connect the backward Kolmogorov equation (34) with conditional expectations of the Poisson process (35).

10.1 Time inhomogeneous Poisson process

If the Poisson process has intensity density $\lambda(t)$ then λ times $(t-s)$ in (31) should be replaced by $\int_s^t \lambda(\gamma)d\gamma$, and λ by its time-dependent version in all places where it comes up. If we define the time-dependent linear operator (generator)

$$\mathcal{L}_t f(t,n) = \lambda(t)[f(t,n+1) - f(t,n)]$$

then we can write Ito's formula (36) in the form

$$df(t,N(t)) = (f_t(t,N(t)) + \mathcal{L}_t f(t,N(t)))\,dt + \Delta f(t,N(t-))dM(t) \tag{37}$$

[4]Which follows from the independent increments property.

where $M(t)$ is the time-inhomogeneous Poisson martingale

$$M(t) = N(t) - \int_0^t \lambda(s)ds \tag{38}$$

It $u(t,n)$ satisfies the time-inhomogeneous backward Kolmogorov equation

$$u_t(t,n) + \mathcal{L}_t u(t,n) = 0,\ t < T,\ u(T,n) = h(n)$$

then the integrated form of Ito's formula gives

$$u(T,N(T)) - u(t,N(t)) = \int_t^T \Delta u(s,N(s-))dM(s)$$

which means that $u(t,N(t))$ is also a martingale. This is because we have

$$E\left(u(T,N(T))|\mathcal{F}_t\right) = u(t,N(t)) + E\left(\int_t^T \Delta u(s,N(s-))dM(s)|\mathcal{F}_t\right)$$

and the last term is zero, a property of Poisson stochastic integrals that we saw earlier. Therefore

$$E\left(u(T,N(T))|\mathcal{F}_t\right) = u(t,N(t))$$

which is the martingale property. Since $u(T,n) = h(n)$ and we have the Markov property for the Poisson process we get the probabilistic representation of the solution of the backward Kolmogorov equation

$$u(t,n) = E\left(h(N(T))|N(t) = n\right)$$

10.2 The exponential martingale

For any $\alpha \geq 0$ let

$$u(t,n) = e^{-\alpha n + \lambda t(1-e^{-\alpha})}$$

and note that

$$u_t(t,n) + \lambda[u(t,n+1) - u(t,n)] = 0$$

It follows from Ito's formula (36) that

$$M_\alpha(t) = u(t,N(t))$$

is a martingale when $M(t) = N(t) - \lambda t$ is a martingale, $E\{M_\alpha(t)|\mathcal{F}_s\} = M_\alpha(s)$. Therefore

$$E\{e^{-\alpha N(t) + \lambda t(1-e^{-\alpha})}|\mathcal{F}_s\} = e^{-\alpha N(s) + \lambda s(1-e^{-\alpha})}$$

Rearranging we have

$$E\{e^{-\alpha(N(t)-N(s)}|\mathcal{F}_s\} = e^{-\lambda(t-s)(1-e^{-\alpha})}$$

This shows that the conditional generating function of the increment $N(t) - N(s)$ is, independent of the conditioning, that of a Poisson process with intensity λ. It also implies that the increments over non-overlapping time intervals are independent. Therefore the process $N(t)$ is the Poisson process with intensity λ and we have deduced this from knowing that $N(t)$ is a counting process with Ito formula (36).

11 Compound Poisson processes

An example of how we can extend the basic theory of the Poisson process, consider the jump process

$$X(t) = x + \int_0^t g(s, N(s-))dN(s) = x + \sum_{0<\tau_k\le t} g(\tau_k, N(\tau_k-))$$

where $g(t,n)$ is a given bounded function on the integers and a continuous function of time with $g(t,0) = 0$. Let $f(x,n)$ be a bounded function. We note that

$$f(X(t),N(t)) - f(X(s),N(s)) = \sum_{s<\tau_k\le t} f(X(\tau_k),N(\tau_k)) - f(X(\tau_k-),N(\tau_k-))$$

We also note that

$$\begin{aligned} &f(X(\tau_k),N(\tau_k)) - f(X(\tau_k-),N(\tau_k-)) \\ &= f(X(\tau_k-) + g(\tau_k,N(\tau_k-)),N(\tau_k-)+1) - f(X(\tau_k-),N(\tau_k-)) \end{aligned}$$

so we have in integral form

$$\begin{aligned} &f(X(t),N(t)) - f(X(s),N(s)) \\ &= \int_s^t [f(X(\gamma-) + g(\gamma,N(\gamma-)),N(\gamma-)+1) - f(X(\gamma-),N(\gamma-))]dN(\gamma) \end{aligned}$$

In terms of the compensated Poisson process and martingale $M(t) = N(t) - \int_0^t \lambda(\gamma)d\gamma$ we have

$$\begin{aligned} &f(X(t),N(t)) - f(X(s),N(s)) \\ &= \int_s^t \lambda(\gamma)[f(X(\gamma) + g(\gamma,N(\gamma)),N(\gamma)+1) - f(X(\gamma),N(\gamma))]d\gamma \\ &+ \int_s^t [f(X(\gamma-) + g(\gamma,N(\gamma-)),N(\gamma-)+1) - f(X(\gamma-),N(\gamma-))]dM(\gamma) \end{aligned}$$

As before, in the Riemann integral we do need to specify that the argument is the left limit since that has no effect on the integral. When f depends on time as well and is differentiable we have Ito's formula:

$$\begin{aligned} &f(t,X(t),N(t)) - f(s,X(s),N(s)) \\ &= \int_s^t \lambda(\gamma)[f_t(\gamma,X(\gamma),N(\gamma)) + f(\gamma,X(\gamma) + g(\gamma,N(\gamma)),N(\gamma)+1) - f(\gamma,X(\gamma),N(\gamma))]d\gamma \\ &+ \int_s^t [f(\gamma,X(\gamma-) + g(\gamma,N(\gamma-)),N(\gamma-)+1) - f(\gamma,X(\gamma-),N(\gamma-))]dM(\gamma) \end{aligned}$$

11.1 Independent random jumps

An important example of a compound Poisson process is

$$X(t) = x + \int_0^t g(s, \eta_{N(s-)}) dN(s) = x + \sum_{0<\tau_k \le t} g(\tau_k, \eta_k) = x + \sum_{k=1}^{N(t)} g(\tau_k, \eta_k)$$

where $\eta_0 = 0$ and $\{\eta_1, \eta_2, \dots\}$ are independent identically distributed real valued random variable with distribution $F(y)$, independent of the Poisson process as well, and $g(t, y)$ is a bounded function of time and y a real variable with $g(t, 0) = 0$. The sum on the right is taken to be zero when $N(t) = 0$. As before, we now have

$$f(X(t), N(t)) - f(X(s), N(s))$$

$$= \int_s^t [f(X(\gamma-) + g(\gamma, \eta_{N(\gamma-)}), N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma)$$

The sigma algebra of sets of right-continuous paths with left hand limits $\mathcal{F}_t,\ 0 \le t \le T$ will now be generated by two-dimensional paths $(X(t), N(t))$, and note that there is additional randomness is the compound Poisson process $X(t)$ because of the random jumps. With this in mind, the conditional expectation of the right hand side is

$$E\left\{\int_s^t [f(X(\gamma-) + g(\gamma, \eta_{N(\gamma-)}), N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma) \mid \mathcal{F}_s\right\}$$

$$= E\left\{\int_s^t E\{[f(X(\gamma-) + g(\gamma, \eta_{N(\gamma-)}), N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma) \mid \mathcal{F}_{\gamma-}\} \mid \mathcal{F}_s\right\}$$

$$= E\left\{\int_s^t [\int_R f(X(\gamma-) + g(\gamma, y), N(\gamma-) + 1) dF(y) - f(X(\gamma-), N(\gamma-))] \lambda(\gamma) d\gamma \mid \mathcal{F}_s\right\}$$

We have used here the independence of the the random variables η_k from each other and from the Poisson process. Let $M_f(t)$ be defined by

$$M_f(t) = \int_0^t [f(X(\gamma-) + g(\gamma, \eta_{N(\gamma-)}), N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma)$$

$$- \int_0^t [\int_R f(X(\gamma-) + g(\gamma, y), N(\gamma-) + 1) dF(y) - f(X(\gamma-), N(\gamma-))] \lambda(\gamma) d\gamma$$

Then it is clear from the previous calculations that $M_f(t)$ is a zero mean martingale, $E\{M_f(t)|\mathcal{F}_s\} = M_f(s)$.

We can now write Ito's formula as follows

$$f(t, X(t), N(t)) - f(s, X(s), N(s))$$

$$= \int_s^t \lambda(\gamma) \left[f_t(\gamma, X(\gamma), N(\gamma)) + \int_R f(X(\gamma) + g(\gamma, y), N(\gamma) + 1) dF(y) - f(X(\gamma), N(\gamma)) \right] d\gamma$$
$$+ (M_f(t) - M_f(s))$$

As before, we have also allowed f to depend on time and we have removed the minus in the arguments of the Riemann integral. We note in particular that $(X(t), N(t))$ are jointly Markovian with infinitesimal generator

$$\mathcal{L}_t f(x, n) = \lambda(t) \left[\int_R f(x + g(t, y), n + 1) dF(y) - f(x, n) \right]$$

and the backward Kolmogorov equation is

$$u_t(t, x, n) + \mathcal{L}_t u(t, x, n) = 0, \ t < T$$

with $u(T, x, n) = v(x, n)$ for a given terminal function $v(x, n)$. As before, from Ito's formula we have the representation

$$u(t, x, n) = E_{t,x,n}\{v(X(T), N(T)\}$$

11.2 Markov chain random jumps

We consider now continuous time Markov chain with real values. The state of the process can be written as

$$X(t) = x + \int_0^t \eta_{N(s-)} dN(s) = x + \sum_{0 < \tau_k \leq t} \eta_k = x + \sum_{k=1}^{N(t)} \eta_k$$

where $\eta_0 = 0$, the sum on the right is zero when $N(t) = 0$, and now $\{\eta_1, \eta_2, \dots\}$ are the jumps of the real valued continuous time Markov chain $X(t)$ so that given that a jump occurs at time t when $X(t-) = x$ then the probability that the jump η is in set $A \subset R$ is given by the jump transition probability function $\pi(x, A)$. This jump transition probability is a continuous function of $x \in R$ and a probability measure on the Borel sets of the real line in the second argument. If it has a density then $\pi(x, dy) = \pi(x, y)dy$. We also assume that the intensity of the Poisson counting process λ is a function of the state $X(t)$, $\lambda = \lambda(X(t))$. For a bounded function $f(x, n)$ we have as before

$$f(X(t), N(t)) - f(X(s), N(s))$$
$$= \int_s^t [f(X(\gamma-) + \eta_{N(\gamma-)}, N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma)$$

We take the conditional expectation of the right side as before and we get

$$E \left\{ \int_s^t [f(X(\gamma-) + \eta_{N(\gamma-)}, N(\gamma-) + 1) - f(X(\gamma-), N(\gamma-))] dN(\gamma) \mid \mathcal{F}_s \right\}$$

$$= E\left\{\int_s^t E\{[f(X(\gamma-)+\eta_{N(\gamma-)}, N(\gamma-)+1) - f(X(\gamma-), N(\gamma-))]dN(\gamma) \mid \mathcal{F}_{\gamma-}\} \mid \mathcal{F}_s\right\}$$

$$= E\left\{\int_s^t \left[\int_R f(X(\gamma-)+y, N(\gamma-)+1)\pi(X(\gamma-), dy) - f(X(\gamma-), N(\gamma-))\right]\lambda(X(\gamma-))d\gamma \mid \mathcal{F}_s\right\}$$

We now conclude that if we define

$$M_f(t) = \int_0^t [f(X(\gamma-)+\eta_{N(\gamma-)}, N(\gamma-)+1) - f(X(\gamma-), N(\gamma-))]dN(\gamma)$$

$$-\int_0^t \left[\int_R f(X(\gamma-)+y, N(\gamma-)+1)\pi(X(\gamma-), dy) - f(X(\gamma-), N(\gamma-))\right]\lambda(X(\gamma-))d\gamma$$

then $M_f(t)$ is a zero mean martingale, $E\{M_f(t)|\mathcal{F}_s\} = M_f(s)$.

With this definition, Ito's formula has the form

$$f(t, X(t), N(t)) - f(s, X(s), N(s))$$

$$= \int_s^t \lambda(X(\gamma))\left[f_t(\gamma, X(\gamma), N(\gamma)) + \int_R f(X(\gamma)+y, N(\gamma)+1)\pi(X(\gamma), dy) - f(X(\gamma), N(\gamma))\right]d\gamma$$
$$+(M_f(t) - M_f(s))$$

We have also allowed again f to depend on time and we have removed the minus in the arguments of the Riemann integral. We note in particular that $(X(t), N(t))$ are jointly Markovian with infinitesimal generator

$$\mathcal{L}f(x,n) = \lambda(x)\left[\int_R f(x+y, n+1)\pi(x, dy) - f(x,n)\right]$$

and the backward Kolmogorov equation is

$$u_t(t,x,n) + \mathcal{L}u(t,x,n) = 0,\ t < T$$

with $u(T,x,n) = v(x,n)$ for a given terminal function $v(x,n)$. As before, from Ito's formula we have the representation

$$u(t,x,n) = E_{t,x,n}\{v(X(T), N(T)\}$$

We also note that $X(t)$, is a time homogeneous, continuous time Markov chain, that is, it is itself Markovian with generator $\mathcal{L}$ as above acting on functions only of x. And Ito's formula for $X(t)$ can be written more compactly as

$$f(t, X(t)) - f(0, X(0)) - \int_0^t [f_t(s, X(s)) + \mathcal{L}f(s, X(s))]ds = M_f(t)$$

with the martingale $M_f(t)$ defined above with f independent of n but depending on t and x.

12 Brownian motion, the Poisson process and Ito's formula

We very briefly review some basic facts about Brownian motion and Ito's formula. We start with the elementary definition of $W(t),\ t \geq 0$ as a collection of random variables such that (i) $W(0) = 0$, (ii) W has independent increments, and (iii) the increments $W(t) - W(s),\ s \leq t$ are Gaussian with mean zero and variance $t - s$. From the fact that $E\{(W(t) - W(s))^4\} = 3(t-s)^2$ it follows that $W(t)$ is continuous with probability, by the Kolmogorov continuity criterion. Contrast this with the Poisson process that does not have this moment property and is a pure jump process. A basic property of Brownian motion that leads to the special Ito calculus is its lack of differentiability. This is directly related to the independent increments property which says that the direction of change of $W(t)$ cannot be determined from knowledge of its current and past positions. This is, of course, only a heuristic explanation of the non-differentiability of Brownian motion.

The lack of differentiability is reflected in the behavior of the quadratic variation. If $0 = t_0 < t_1 < t_2 \ldots < t_N = T < \infty$ is a partition of the interval $[0, T]$ with $\max_k(t_{k+1} - t_k) \to 0,\ N \to \infty$ then

$$\lim_N \sum_{k=0}^{N-1} (W(t_{k+1}) - W(t_k))^2 = T$$

where the limit is in the mean square sense. If $W(t)$ was differentiable (or more exactly had finite total variation) then the quadratic variation would be zero.

Now we want to calculate the change of $f(W(t))$ where f is a twice differentiable function. We have

$$f(W(t)) - f(0) = \sum_{k=0}^{N-1} f(W(t_{k+1})) - f(W(t_k))$$

which is a telescoping sum. We can use Taylor's formula to get

$$f(W(t)) - f(0) = \sum_{k=0}^{N-1} \left[f_x(W(t_k))(W(t_{k+1}) - W(t_k)) + \frac{1}{2} f_{xx}(W(t_k^*))(W(t_{k+1}) - W(t_k))^2 \right]$$

where $t_k^* \in (t_k, t_{k+1})$. The first sum on the right converges in mean square to an Ito stochastic integral and the second to a Riemann integral. Skipping a lot of details we arrive at Ito's formula in the limit

$$f(W(t)) - f(W(s)) = \int_s^t \frac{1}{2} f_{xx}(W(\gamma))d\gamma + \int_s^t f_x(W(\gamma))dW(\gamma),\ t > s.$$

And if $f = f(t, x)$ depends on time then for $t > s$ we have

$$f(t, W(t)) - f(s, W(s)) = \int_s^t \left[f_\gamma(\gamma, W(\gamma)) + \frac{1}{2} f_{xx}(\gamma, W(\gamma)) \right] d\gamma + \int_s^t f_x(\gamma, W(\gamma))dW(\gamma). \tag{39}$$

The backward Kolmogorov equation for Brownian motion can be obtained from Ito's formula by letting $u(t,x)$ solve the backward heat, in this case, equation

$$u_t(t,x) + \frac{1}{2}u_{xx}(t,x) = 0,\ t < T,\ u(T,x) = h(x). \tag{40}$$

Then assuming that h is bounded and has two derivatives we can apply Ito's formula to $u(t,W(t))$ and get

$$u(T,W(T)) = u(t,W(t) + \int_t^T u_x(\gamma,W(\gamma))dW(\gamma)$$

Taking expectation conditional on $W(t) = x$ and noting that the stochastic integral is a zero mean martingale, we have

$$u(t,x) = E(h(W(T))|W(t) = x)$$

12.1 The Poisson and Brownian Ito's formula

We can combine the two Ito formulas (33) and (39) so that a suitably differentiable in $x \in \mathcal{R}$ and bounded function $f(t,x,n)$, with $n = 0,1,2,\ldots$, to get

$$\begin{aligned} &f(t,W(t),N(t)) = f(s,W(s),N(s)) \\ &+ \int_s^t \left[f_\gamma(\gamma,W(\gamma),N(\gamma)) + \frac{1}{2}f_{xx}(\gamma,W(\gamma),N(\gamma)) + \lambda\Delta f(\gamma,W(\gamma),N(\gamma)) \right] d\gamma \\ &+ \int_s^t f_x(\gamma,W(\gamma),N(\gamma))dW(\gamma) + \int_s^t \Delta f(\gamma,W(\gamma),N(\gamma-))dM(\gamma) \end{aligned}$$

This combined formula becomes a little more interesting when, for example, the intensity of the Poisson process λ is a function of the Brownian motion, that is, it is random, with the Brownian and the Poisson processes being independent. All that changes in the above Ito formula is the replacement of λ by $\lambda(W(\gamma))$ in the Riemann integral. The backward Kolmogorov equation can be derived as above when both a Poisson process and Brownian motion are involved and the sigma algebras $\mathcal{F}_t$ are generated by both W and N.

13 Applications of stochastic calculus of Poisson and Brownian motion

13.1 The Hawkes process

This is an interesting example because it shows how the Ito formula can connect efficiently a process with its backward Kolmogorov equation, that is, the paths of the process and the associated equation for calculating probabilities. The Hawkes process is a Poisson process with an intensity $\lambda(t)$ that is coupled to it. This models self-exciting behavior for the jump intensity because as the jumps occur their intensity is affected. A simple example is when the intensity satisfies for α, μ, β given and positive

$$d\lambda(t) = \alpha(\mu - \lambda(t))dt + \beta dN(t)$$

or, in terms of the martingale $M(t)$ defined by $M(0) = 0$ and $dM(t) = dN(t) - \lambda(t)dt$,

$$d\lambda(t) = (\alpha - \beta)\left(\frac{\mu}{1 - \frac{\beta}{\alpha}} - \lambda(t)\right)dt + \beta dM(t)$$

We assume that $0 \leq \beta < \alpha$ so that $E(\lambda(t)) = \overline{\lambda(t)}$ satisfies the differential equation

$$d\overline{\lambda(t)} = (\alpha - \beta)\left(\frac{\mu}{1 - \frac{\beta}{\alpha}} - \overline{\lambda(t)}\right)dt$$

which has solution

$$\overline{\lambda(t)} = \frac{\mu}{1 - \frac{\beta}{\alpha}} + \left(\lambda_0 - \frac{\mu}{1 - \frac{\beta}{\alpha}}\right)e^{-(\alpha-\beta)t}$$

Now the pair $(N(t), \lambda(t))$ is jointly a time-homegeneous Markov process and the associated Ito formula has the form

$$\begin{aligned} f(t, N(t), \lambda(t)) = f(s, N(s), \lambda(s)) &+ \int_s^t \left[f_\gamma(\gamma, N(\gamma), \lambda(\gamma)) + \mathcal{L}f(\gamma, N(\gamma), \lambda(\gamma))\right] d\gamma \\ &+ \int_s^t \left[f(\gamma, N(\gamma-) + 1, \lambda(\gamma-) + \beta) - f(\gamma, N(\gamma-), \lambda(\gamma-))\right] dM(\gamma). \end{aligned} \quad (41)$$

Here the infinitesimal generator $\mathcal{L}$ is given by

$$\mathcal{L}g(n, \lambda) = (\alpha - \beta)\left(\frac{\mu}{1 - \frac{\beta}{\alpha}} - \lambda\right)g_\lambda(n, \lambda) + \lambda\left[g(n + 1, \lambda + \beta) - g(n, \lambda)\right].$$

As we saw above, the backward Kolmogorov equation for

$$u(t, n, \lambda) = E(h(N(T), \lambda(T))|N(t) = n, \lambda(t) = \lambda)$$

has the again form

$$u_t + \mathcal{L}u = 0 \; t < T, \; u(T, n, \lambda) = h(n, \lambda).$$

As an application we can calculate the expectation of $N(t)$. Let $f = n$ in Ito's formula. Clearly $\mathcal{L}f = \lambda$ and so taking expectations in (41) we have

$$E(N(t)) = \int_0^t E(\lambda(s))ds = \int_0^t \overline{\lambda(s)}ds$$

since $N(0) = 0$. To calculate $E(N^2(t))$ we use Ito's formula (41) with three f, $f = n^2$, $f = \lambda n$ and $f = \lambda^2$ and after taking expectations we end up with a system of three coupled linear ordinary differential equations to solve for $E(N^2(t))$, $E(N(t)\lambda(t))$ and $E(\lambda^2(t))$.

13.2 The Avellaneda-Stoikov limit order trading model

This is a model that determines the optimal limit order spreads. Trading occurs at discrete times according to Poisson flows Q_t^a for ask or sell orders and Q_t^b for bid or buy orders executed up to time t. The order **inventory** up to time t is

$$Q_t = Q_t^b - Q_t^a + q$$

with q the initial inventory. The inventory changes by ± 1 when there is a buy or sell order executed, which is an acceptable assumption for a bulk model of limit order trading. The **cash process** X_t changes as

$$dX_t = p_t^a dQ_t^a - p_t^b dQ_t^b$$

where p_t^a and p_t^b are the ask and bid prices that are to be determined. In fact we determine the **spreads**

$$\delta_t^b = S_t - p_t^b, \quad \delta_t^a = p_t^a - S_t$$

where S_t is the observed mid-price. We assume here that

$$S_t = S_0 + \sigma W_t$$

with W_t the standard Brownian motion and σ the given volatility, which is an acceptable assumption for relatively high frequency trading.

The key assumption in this model is that Q_t^a and Q_t^b are independent standard Poisson processes, independent of W_t, and with intensities

$$\lambda^a(t) = Ae^{-\kappa\delta_t^a}, \quad \lambda^b(t) = Ae^{-\kappa\delta_t^b}$$

This assumption conveys the basic tradeoff to be resolved with the determination of the optimal spreads so as to maximize the utility of the wealth

$$\mathcal{W}_t = X_t + Q_t S_t$$

at a target time T: if the spreads are too wide then the order flows are too slow and if they are too narrow the order flows are as if there is no trading. The parameter κ is a measure of sensitivity of the flow rates to the spreads, that is, a measure of liquidity.

Given the intensities $\lambda^a(t),\ \lambda^b(t)$, we can write Ito's formula for the joint process (S_t, Q_t, X_t). If we define the generator

$$\begin{aligned}\mathcal{L}_t f(s,q,x) = \frac{\sigma^2}{2} f_{ss}(s,q,x) &+ \lambda^a(t)\left[f(s,q-1,x+s+\delta_t^a) - f(s,q,x)\right] \\ &+ \lambda^b(t)\left[f(s,q+1,x-s+\delta_t^b) - f(s,q,x)\right] \end{aligned} \tag{42}$$

then for $0 \le s < t$ we have

$$\begin{aligned} f(t,S_t,Q_t,X_t) =& f(s,S_s,Q_s,X_s) + \int_s^t \left[f_\gamma(\gamma,S_\gamma,Q_\gamma,X_\gamma) + \mathcal{L}_\gamma f(\gamma,S_\gamma,Q_\gamma,X_\gamma)\right] d\gamma \\ &+ \int_s^t \left[f(\gamma,S_\gamma,Q_{\gamma-}-1,X_{\gamma-}+S_\gamma+\delta_\gamma^a) - f(\gamma,S_\gamma,Q_{\gamma-},X_{\gamma-})\right] dM_\gamma^a \\ &+ \int_s^t \left[f(\gamma,S_\gamma,Q_{\gamma-}+1,X_{\gamma-}-S_\gamma+\delta_\gamma^b) - f(\gamma,S_\gamma,Q_{\gamma-},X_{\gamma-})\right] dM_\gamma^b \\ &+ \int_s^t \sigma f_x(\gamma,S_\gamma,Q_\gamma,X_\gamma) dW_\gamma. \end{aligned} \tag{43}$$

Here $dM_t^a = dQ_t^a - d\lambda^a(t)dt$ and $dM_t^b = dQ_t^b - d\lambda^b(t)dt$ are the martingales associated with the two Poisson flows. It is clear from Ito's formula that if $u(t,s,q,x)$ satisfies the backward Kolmogorov equation

$$u_t + \mathcal{L}_t u = 0,\ t < T, \quad u(T,s,q,x) = h(s,q,x)$$

then we have

$$u(t,s,q,x) = E(h(S_T,Q_T,X_T)|S_t = s, Q_t = q, X_t = x)$$

Up to now the spreads $\delta_t^{a,b}$ that control the intensities $\lambda^a(t) = Ae^{-\kappa\delta_t^a},\ \ \lambda^b(t) = Ae^{-\kappa\delta_t^b}$ are assumed to be given. We introduce now a form of the generator suitable for the Hamilton-Jacoby-Bellman equation,

$$\begin{aligned}\mathcal{L}_{\{\delta^a,\delta^b\}} f(s,q,x) = \frac{\sigma^2}{2} f_{ss}(s,q,x) &+ Ae^{-\kappa\delta^a}\left[f(s,q-1,x+s+\delta^a) - f(s,q,x)\right] \\ &+ Ae^{-\kappa\delta^b}\left[f(s,q+1,x-s+\delta^b) - f(s,q,x)\right]. \end{aligned} \tag{44}$$

The HJB equation is now the backward Kolmogorov equation with a pointwise supremum inserted

$$u_t + \sup_{\{\delta^a,\delta^b\}} \mathcal{L}_{\{\delta^a,\delta^b\}} u = 0,\ t < T, \quad u(T,s,q,x) = h(s,q,x). \tag{45}$$

This is a highly nonlinear equation but when it has a well defined solution, so that Ito's formula can be used, we have that

$$u(t,s,q,x) = \sup_{\{\delta^a_\cdot,\delta^b_\cdot\}} E(h(S_T,Q_T,X_T)|S_t=s,Q_t=q,X_t=x), \tag{46}$$

where the controls are the spreads $\delta^{a,b}_t$. The optimal controls $\delta^{*,a,b}_t$ are Marcovian when the payoff function is the exponential utility of the terminal wealth

$$h(s,q,x) = -e^{-\bar{\gamma}(x+qs)}.$$

The process $\mathcal{W}_t = X_t + Q_t S_t$ is the wealth of the limit order investor and the optimal control problem for the limit order spreads is to maximize the exponential utility of the terminal wealth, with $\bar{\gamma}$ a risk aversion parameter. The connection between the optimal control value function $u(t,s,q,x)$ in (46) and the solution of the HJB equation (45), when the latter is well behaved, is called the verification lemma and it is a direct application of Ito's formula (43). We consider this in the the next section for Ito diffusions. The calculations for jump processes are similar.

14 The Hamilton-Jacobi-Bellman equation for diffusions

The essential theoretical tool is Ito's formula for diffusions but the derivation extends to processes with jumps in a direct way. The theory of stochastic differential equations (SDE) or of the nonlinear Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) is not needed. What follows is just the so-called verification lemma without the theoretical details.

Consider the controlled diffusion process $X(t) = X^U(t)$ that satisfies the Ito stochastic differential equation

$$dX(t) = b(X(t), U(t))dt + \sigma(X(t), U(t))dB(t)$$

with $X(0) = x$. Here $U(t) \in \mathcal{U}$ is the control process, a non-anticipating function with values in the set $\mathcal{U}$ that is specified. We assume that the coefficients $b(x, u)$ and $\sigma(x, u)$ satisfy the Ito conditions as functions of x, uniformly in $u \in \mathcal{U}$. For $g(x)$ and $h(t, x, u)$ given, the Hamilton-Jacobi-Bellman (HJB) equation for the value function

$$V(t, x) = \inf_U E_{t,x}\{g(X(T)) + \int_t^T h(s, X(s), U(s))ds\} \ , \quad t \leq T$$

has the form

$$V_t(t, x) + \inf_u \{\mathcal{L}_u V(t, x) + h(t, x, u)\} = 0 \ , \quad t < T \tag{47}$$

with terminal conditions $V(T, x) = g(x)$. The infimum in this PDE is pointwise in (t, x), and $\mathcal{L}_u$ is the generator of the controlled diffusion with a fixed constant control u

$$\mathcal{L}_u = \frac{1}{2}\sigma^2(x, u)\frac{\partial^2}{\partial x^2} + b(x, u)\frac{\partial}{\partial x}$$

We will assume that the HJB equation has a classical solution, that is, with one time and two space derivatives, and denote the unique pointwise minimal control u by $u^* = u^*(t, x)$, assumed differentiable. The optimal, Markovian, control is then $U^*(t) = u^*(t, X^*(t))$ where $X^*(t)$ is the optimally controlled diffusion satisfying the Ito SDE

$$dX^*(t) = b(X^*(t), u^*(t, X^*(t)))dt + \sigma(X^*(t), u^*(t, X^*(t)))dB(t)$$

with $X^*(0) = x$, which is assumed to have a solution as an Ito diffusion.

Let $U(t)$ be any admissible control, let $X^U(t)$ be the solution of the Ito SDE (assuming it exists in the usual way). We will apply Ito's formula to $V(t, X^U(t))$ and deduce that

$$E_{t,x}\{g(X^U(T)) + \int_t^T h(s, X^U(s), U(s))ds\} \geq V(t, x),$$

and since the right side is independent of U,

$$\inf_U E_{t,x}\{g(X^U(T)) + \int_t^T h(s, X^U(s), U(s))ds\} \geq V(t, x)$$

From Itô's formula and after integrating we have

$$V(T, X^U(T)) = V(t, X^U(t)) + \int_t^T [V_t(s, X^U(s)) + \mathcal{L}_{U(s)} V(s, X^U(s))] ds$$
$$+ \int_t^T \sigma(X^U(s), U(s)) V_x(s, X^U(s)) dB_s$$

We add to both sides the integral in the objective to get

$$V(T, X^U(T)) + \int_t^T h(s, X^U(s), U(s)) ds = V(t, X^U(t))$$
$$+ \int_t^T [V_t(s, X^U(s)) + \mathcal{L}_{U(s)} V(s, X^U(s)) + h(s, X^U(s), U(s))] ds$$
$$+ \int_t^T \sigma(X^U(s), U(s)) V_x(s, X^U(s)) dB_s$$

Using the terminal condition and taking expectation given $X^U(t) = x$ we have further

$$E_{t,x}[g(X^U(T)) + \int_t^T h(s, X^U(s), U(s)) ds] = V(t, x)$$
$$+ E_{t,x}\left[\int_t^T [V_t(s, X^U(s)) + \mathcal{L}_{U(s)} V(s, X^U(s)) + h(s, X^U(s), U(s))] ds\right].$$

For suboptimal $U(t)$, $V_t(t, x) + \mathcal{L}_{U(t)} V(t, x) + h(t, x, u) \geq 0$, pointwise in (t, x) so

$$E_{t,x}[g(X^U(T)) + \int_t^T h(s, X^U(s), U(s)) ds] \geq V(t, x)$$

and, as noted, since V does not depend on U we have

$$\tilde{V}(t, x) = \inf_U E_{t,x}[g(X^U(T)) + \int_t^T h(s, X^U(s), U(s)) ds] \geq V(t, x).$$

This says that the solution of the HJB equation $V(t, x)$ is a lower bound for the function $\tilde{V}(t, x)$, the value function of the control problem.

In fact, $\tilde{V}(t, x) = V(t, x)$, so the solution of the HJB equation is the value function of the optimal control. To see this we apply Ito's formula to $V(t, X^*(t))$ and the optimal control $U^*(t) = u^*(t, X^*(t))$. We have

$$E_{t,x}[g(X^*(T))) + \int_t^T h(s, X^*(s), U^*(s)) ds] = V(t, x)$$

$$+E_{t,x}\left[\int_t^T [V_t(s,X^*(s)) + \mathcal{L}_{U^*(s)}V(s,X^*(s)) + h(s,X^*(s),U^*(s))]ds\right],$$

assuming that the martingale term is well defined so that its expectation is zero. Since the control $U^*(t) = u^*(t,X^*(t))$ is minimal for the generator we have $V_t(t,X^*(t)) + \mathcal{L}_{u^*(t,X^*(t))}V(t,X^*(t)) + h(t,X^*(t),U^*(t)) = 0$. Therefore,

$$E_{t,x}[g(X^*(T)) + \int_t^T h(s,X^*(s),U^*(s))ds] = V(t,x).$$

So, U^* is the optimal control since it attains the lower bound, and we have the desired identification

$$V(t,x) = \tilde{V}(t,x) = \inf_U E_{t,x}[g(X(T)) + \int_t^T h(s,X^U(s),U(s))ds].$$

Keep in mind that this identification or verification process (Lemma or Theorem) relies heavily on the regularity of the solution V of the HJB equation and then on the construction of the optimal process X^*. This can be a very difficult issue in practice.

14.1 HJB equation for processes with jumps

The verification lemma of the previous section, as it is called, depends on being able to apply Ito's formula to the solution of a nonlinear PDE, the HJB equation. The hard part here is being able to show that we can indeed apply Ito's formula to the solution of this nonlinear equation in some way. Once this has been understood and shown the rest is a relatively simple argument that, in fact, does not depend on dealing with controlled Ito diffusions. We could be dealing with controlled Markov chains as in Section 11.2, or with multi-dimensional controlled processes some of whose components are diffusions and some compound Poisson processes as with the Avellaneda-Stoikov model in Section 13.2.

For the verification lemma we need only know that if for example $X^U(t)$ is a controlled Markov chain then

$$V(t,X^U(t)) - V(0,X^U(0)) - \int_0^t [V_t(s,X^U(s)) + \mathcal{L}_{U(s)}V(s,X^U(s))]ds = M_f(t) \qquad (48)$$

is a martingale for functions $V(t,x)$ which are differentiable in t and the infinitesimal generator $\mathcal{L}_u$ can be applied to them. We do not need to know the detailed representation of this martingale which in the case of Ito diffusions is a Brownian stochastic integral, for example. As noted in Section 13.2, the HJB equation has the form of the backward Kolmogorov equation optimized pointwise over time and the state variable, and this follows from (48).

14.2 HJB equation for a controlled diffusion with boundary conditions

If there is a constraint, for example $X(t) \geq 0$ for all time then we need a stopping time and a boundary condition for the value function. For simplicity in writing we will assume here that $h = 0$ but the case h not zero works fine. Let τ be the exit time from zero, $\tau = \inf\{s > t \mid X(s) = 0\}$. Also, assume that the value function $V(t,x)$ is zero at $x = 0$. As before, we can apply Ito's formula but now up to the minimum of τ and T

$$V(\tau \wedge T, X^U(\tau \wedge T)) = V(t, X^U(t)) + \int_t^{\tau \wedge T} [V_t(s, X^U(s)) + \mathcal{L}_{U(s)} V(s, X^U(s))] ds$$
$$+ \int_t^{\tau \wedge T} \sigma(X^U(s), U(s)) V_x(s, X^U(s)) dB_s$$

Taking expectations (and using the optional stopping theorem) we have

$$E_{t,x}[V(\tau \wedge T, X^U(\tau \wedge T))] = V(t,x) + E_{t,x}\left[\int_t^{\tau \wedge T} [V_t(s, X^U(s)) + \mathcal{L}_{U(s)} V(s, X^U(s))] ds\right].$$

As above, for suboptimal U, we have

$$E_{t,x}[V(\tau \wedge T, X^U(\tau \wedge T))] \geq V(t,x)$$

We now note that

$$E_{t,x}[V(\tau \wedge T, X^U(\tau \wedge T))] = E_{t,x}[V(\tau, X^U(\tau)) \chi_{\tau \leq T}] + E_{t,x}[V(T, X^U(T)) \chi_{\tau > T}]$$

and using the terminal and boundary conditions we get that the left side is $E_{t,x}[g(X^U(T)) \chi_{\tau > T}]$ and therefore

$$E_{t,x}[g(X^U(T)) \chi_{\tau > T}] \geq V(t,x)$$

As above, again, we get equality when using the optimal control so that

$$\inf_U E_{t,x}[g(X^U(T)) \chi_{\tau > T}] = V(t,x)$$

If the h is not zero we have

$$\inf_U E_{t,x}[g(X^U(T)) \chi_{\tau > T} + \int_t^{\tau \wedge T} h(s, X^U(s), U(s)) ds] = V(t,x)$$

with $V(t,x)$ the solution of the HJB equation (47) for $x \geq 0$ and now with the boundary condition $V(t,0) = 0,\ t < T$, along with the terminal condition $V(T,x) = g(x),\ x > 0$.