Stochastic Calculus, Courant Institute, Fall 2021
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2021/index.html

Week 6

Jonathan Goodman, November, 2021

The two topics for this week both involve putting integrals in the exponential. For the *Feynman Kac formula*,[1] the integral is dt. For the *Girsanov change of measure formula*, the integral is dW_t. If X_t is an Ito process, then another Ito process is

$$Y_t = e^{X_t} \; .$$

We can apply Ito's lemma with $f(x) = e^x$, and $\partial_x f(x) = e^x = f(x)$, and $\partial_x^2 f(x) = f(x)$. The result is, using $Y_t = e^{X_t}$ on the right side

$$dY_t = Y_t \, dX_t + \frac{1}{2} Y_t \, (dX_t)^2 \; . \qquad (1) \qquad \boxed{\texttt{dexp}}$$

If X_t is a dt integral (Feynman Kac), then the second term on the right is zero.

The formula (1) is a differential expression that is linear in Y_t. You can make Y_t satisfy a linear differential relation by choosing X_t appropriately. Defining X_t by indefinite integrals might do this, for example. One possibility is:

$$X_t = \int_t^T a_s \, ds + \int_t^T b_s \, dW_s \; . \qquad (2) \qquad \boxed{\texttt{Xint}}$$

The differential is

$$dX_t = -a_t dt - b_t dW_t \; .$$

Integrating starting at t instead of ending at t as we have done up to now gives the minus signs. For example, if $a_t > 0$ then increasing t a little makes the ds integral in (2) a little smaller. For a more formal explanation, note that the integral from 0 to T does not depend on t and its differential is zero:

$$0 = d\left(\int_0^T a_s \, ds \right) = d\left(\int_0^t a_s \, ds \right) + d\left(\int_t^T a_s \, ds \right)$$

The differential of the first integral on the right is $a_t dt$. Therefore, the differential of the second integral must be $-a_t dt$. The same argument applies to the dW integrals, but reasoning about the sign of b_t seems sketchier. You might integrate from t as in (2) rather than to t as before so that $Y_t = e^{X_t}$ is more Markov like – not depending on the path from 0 to t.

[1] This is named for Richard Feynman, a physicist, and Marc Kac, a mathematician who was trying to understand what Feynman was saying. The name "Kac" is pronounced "cats", or maybe "cots". There is a restaurant on Houston street called the "Katz Deli". Both Marc Kac and the founder of the Katz Deli come from Poland, where their names were spelled the same.

Exponentials of integrals have various uses in stochastic calculus and modeling. An example from finance is then r_t is a random fluctuating interest rate. If you have M_t invested in an account that has this rate, then

$$dM_t = r_t M_t dt \; . \qquad (3) \quad \boxed{\texttt{dM}}$$

In time dt you add $r_t M_t dt$ to your account. The relation between M_t and r_t is given by an integral of the form ($\overset{\texttt{Xint}}{2}$) with $b_s = 0$. This is part of *Feynman Kac* theory. See Section $\overset{\texttt{sec:FK}}{1}$. Part of this is expanding the family of PDEs that are related to stochastic integrals. That is, it expands the family of PDEs that may be interpreted as backward equations. This expands the family of PDEs that may be solved by simulation.

A fancier application is the Girsanov *change of measure* formula of Sections $\overset{\texttt{sec:cm}}{3}$ and $\overset{\texttt{sec:G}}{4}$. Here, "measure" means probability distribution. Suppose some quantity a is defined in terms of a random variable X as an expectation, as we saw for expected utility last class:

$$a = \mathrm{E}_p[\, V(X)] = \int V(x)\, p(x)\, dx \; . \qquad (4) \quad \boxed{\texttt{a}}$$

We say "a is the expected value of $V(X)$ in the $p-$world", which refers to the "world" in which X is a random variable with p as its PDF. There is an expectation formula for the same number a in the "$q-$world", which is the world in which q is the PDF of X. For that, we first define the *likelihood ratio*

$$L(x) = \frac{p(x)}{q(x)} \; . \qquad (5) \quad \boxed{\texttt{L}}$$

(Probabilities may be called *likelihoods* when you use them as functions for some purpose other than integrating or summing to find expectation values.) Some algebra gives a formula for a in the "$q-$world", which is

$$a = \mathrm{E}_q[\, V(X)\, L(X)] = \int V(x)\, L(x)\, q(x)\, dx \; . \qquad (6) \quad \boxed{\texttt{aL}}$$

When you go from the $p-$ world to the $q-$world, you get the same a only if you compensate for changing measures (probability distributions) by putting in the likelihood ratio factor.

Changing measure for stochastic processes is *Girsanov theory.* In the $p-$world, X_t satisfies one SDE. In the $q - world$ is satisfies a different SDE. This relies on the point of view that an SDE is a way of defining a probability distribution on path space (the set of paths). Changing measure for SDE has various applications. One is that it might be easier or more desirable to simulate the $q-$process than the original $p-$process. Another is *rare event simulation.* That estimating expectation values where most paths make little contribution to the expectation. The "important" paths, the ones that contribute to the expectation, may be more likely in the $q-$world.

We will see that changing measure in path space is more subtle than changing measure for one dimensional random variables as in ($\overset{\texttt{aL}}{6}$). The change of measure formula (5) does not change zero probabilities to non-zero probabilities. For example, if $X > 0$ in the $p-$world, then $X > 0$ also in the $q-$world. This is easy to figure out in one dimension (or finite dimensions) by looking at where $p(x) = 0$ and $q(x) = 0$. The probability distributions in path space, which are genuine probability measures, are not given by probability densities. Girsanov's theorem tells you which worlds can be transformed into each other and which cannot.

1 Feynman Kac Theory

sec:FK

Feynman Kac theory, often called "the Feynman Kac formula", is a relationship between a multiplicative value function and its backward equation. Suppose X_t is a diffusion process that satisfies the SDE

$$dX_t = a(X_t)\,dt + b(X_t)\,dW_t\;. \tag{7}$$

SDE

A multiplicative functional is a function of the diffusion process path of the form

$$e^{\int_0^T V(X_s)\,ds}\;. \tag{8}$$

mf

It's called *multiplicative* because it's the exponential of an integral, which is a continuous sort of sum. If it were a discrete sum, the exponential would turn it into multiplication:

$$e^{a+b+c+\cdots} = e^a\, e^b\, e^c \cdots \;.$$

A "functional" is a function of a function. The Feynman Kac value function for a multiplicative functional is

$$f(x,t) = \mathrm{E}\Big[e^{\int_t^T V(X_s)\,ds} \mid X_t = x \Big] \;. \tag{9}$$

mvf

Section 2 explains how multiplicative functionals are used in finance to model fluctuating uncertain interest rates.

We can find the backward equation for f in (9) using the reasoning from Week 4 together with (1) from the Introduction. We look at the process starting at t but without conditional expectation

$$Y_t = e^{\int_t^T V(X_s)\,ds}\;. \tag{10}$$

Yt

Then (using reasoning from the Introduction)

$$dY_t = -V(X_t)\,Y_t\,dt\;.$$

There are no Ito terms here because $V(X_t)$ is a differentiable function of t (no dW part, quadratic variation equal to zero). On the other hand, looking at the definitions of $f(x,t)$ and Y_t, you see that

$$f(x,t) = \mathrm{E}[\, Y_t \mid X_t = x] \;.$$

Therefore,

$$\mathrm{E}[\, df(X_t,t) \mid X_t = x] = \mathrm{E}[\, dY_t \mid X_t = x] \ .$$

Ito's lemma applied to the left side while dY is described as above. As in Week 4, we expand f, then substitute the SDE $(\overset{\texttt{SDE}}{7})$ for dX and use the "Ito rule" $(dX_t)^2 = b^2 dt$. Conditioning on $X_t = x$ allows us to substitute x for a and b in the SDE. Finally, $E[dW] = 0$ leads to expressions involving only dt on both sides.

$$\begin{aligned}
\mathrm{E}[\, df(X_t,t) \mid X_t = x] &= \mathrm{E}[\, -V(X_t) Y_t \, dt \mid X_t = x] \\
\mathrm{E}\Big[\, f_x \, dX_t + \frac{1}{2} f_{xx} \, (dX_t)^2 + f_t \, dt \mid X_t = x \Big] &= -V(x) \, dt \, \mathrm{E}[\, Y_t \mid X_t = x] \\
f_x(x,t) \, a(x) \, dt + \frac{1}{2} f_{xx}(x,t) \, b^2(x) \, dt + f_t(x,t) \, dt &= -V(x) f(x,t) \, dt
\end{aligned}$$

The resulting equation may be written in the form

$$f_t + a(x) f_x + \frac{1}{2} b^2(x) f_{xx} + V(x) f = 0 \ . \qquad (11) \qquad \boxed{\texttt{beFK}}$$

This equation needs final conditions. This is "obvious", as with other backward equations. Setting $t = T$ in the definition of Y_t $(\overset{\texttt{Yt}}{10})$ gives $f(x,T) = 1$.

If that derivation seemed mysterious, here's a version that might seem more natural. This, too, is modeled on a derivation from Week 4. In this case, it's the one that uses the tower property to relate expectations at time t to expectations at time $d + dt$. With the multiplicative functional, the difference between t and $t + dt$ includes a piece of the integral. For that reason, we calculate

$$\begin{aligned}
e^{\int_t^T V(X_s)\, ds} &= e^{\int_t^{t+dt} V(X_s)\, ds + \int_{d+dt}^T V(X_s)\, ds} \\
&= e^{\int_t^{t+dt} V(X_s)\, ds} \, e^{\int_{d+dt}^T V(X_s)\, ds} \\
&= (\, 1 + V(x) \, dt \,) \, e^{\int_{d+dt}^T V(X_s)\, ds} \ .
\end{aligned}$$

This goes into the definition of the value function and allows us to separate out the contribution from "now" (between t and $t + dt$) from the rest. After that, you Taylor expand f and discard the dW terms (which have expected value zero). At the end, we multiply out expressions with dt and discard the dt^2 terms, as $(1 + u\, dt)(1 + v\, dt) = 1 + (u + v) dt$.

$$\begin{aligned}
f(x,t) &= \mathrm{E}\Big[\, e^{\int_t^{t+dt} V(X_s)\, ds} \, e^{\int_{d+dt}^T V(X_s)\, ds} \mid X_t = x \Big] \\
&= (\, 1 + V(x) \, dt \,) \, \mathrm{E}\Big[\, e^{\int_{d+dt}^T V(X_s)\, ds} \mid X_t = x \Big] \\
&= (\, 1 + V(x) \, dt \,) \, \mathrm{E}[\, f(X_{t+dt}, t + dt \mid X_t = x] \\
&= (\, 1 + V(x) \, dt \,) \left(f + f_x a(x) dt + \frac{1}{2} b^2 f_{xx} dt + f_t dt \right) \\
f &= f + \left(V(x) f + a(x) f_x + \frac{1}{2} b^2(x) f_{xx} + f_t \right) dt \ .
\end{aligned}$$

This confirms the backward equation (beFK 11) in the Feynman Kac theory.

The examples are all rather complicated, unfortunately. Take X_t to be Brownian motion and $V(x) = x$. Then the backward equation is

$$\partial_t f + \frac{1}{2}\partial_x^2 f + xf = 0 \; . \tag{12}$$

FKe

You can calculate the solution explicitly as in Exercise ex:FKe 2. Or you can guess by trial and error. Either way, we come to the ansatz

$$f(x,t) = e^{A(t)x+B(t)} \; .$$

We plug this into the backward equation (FKe 12) using the calculations

$$\begin{aligned}\partial_t f &= \left(\dot{A}x + \dot{B}\right) f \\ \partial_x^2 f &= \partial_x \left[Ae^{Ax+B}\right] = A^2 f \; .\end{aligned}$$

The backward equation takes the form

$$\dot{A}xf + \dot{B}f + \frac{1}{2}A^2 f + xf = 0 \; .$$

If you have an equation $cx + d = 0$, then $c = 0$ and $d = 0$. In this case, these become

$$\begin{aligned}\dot{A} + 1 &= 0 \\ \dot{B} + \frac{1}{2}A^2 &= 0 \; .\end{aligned}$$

The first equation, together with the final condition $A(T) = 0$ (why?) gives $A(t) = T - t$. The second equation becomes

$$\dot{B} = -\frac{1}{2}(T-t)^2 \; .$$

The solution is $B(t) = \frac{1}{6}(T-t)^3$. The full solution is

$$f(x,t) = e^{\frac{1}{6}(T-t)^3} e^{(T-t)x} \; . \tag{13}$$

exs

2 Interest rate models

sec:ir

The financial services industry is heavily focused on loans and financial instruments that depend on interest rates. There is a range of stochastic models of interest rates, but the simple models involve only the *short rate* (also called *overnight rate*), which is the interest rate (in %/year, say) that is paid on a loan where you "know" it will be repaid very soon (e.g., the next day). Let us call this rate R_t. One model is

$$dR_t = -\gamma(R_t - r_0)\, dt + \sigma dW_t \; . \tag{14}$$

sr

This is a mean-reverting process centered on an average long term average rate r_0.

A *floating rate* loan is a loan so that the interest rate at time t is related to an index of short rate loans at that time such as LIBOR (google this). We could use the model (sr 14) as a model of this fluctuating rate. Suppose you make a loan with floating interest rate R_t to be repaid at time T. The total repayment amount is a multiplicative function of the initial amount

$$M_T = M_0 e^{\int_0^T R_s\, ds} \; .$$

We may as well set $M_0 = 1$ for simplicity. The value function is

$$f(r,t) = \mathrm{E}[\, M_t \mid R_t = r] \;\; , \;\; M_t = e^{\int_t^T R_s -,ds} \; .$$

This satisfies the backward equation

$$\partial_t f - \gamma(r - r_0)\partial_r f + \frac{\sigma^2}{2}\partial_r^2 f - rf = 0 \; . \tag{15}$$

irf

This has an ansatz solution of the form

$$f(r,t) = e^{A(t)r + B(t)} \; . \tag{16}$$

ira

Try it and see. Models like this are called *affine* because the exponent is an affine function of r.

3 Change of measure

sec:cm

Change of measure for SDE models is called *Girsanov* theory. This consists of a simple formula and a mass of theory telling you when it applies. The formula is *Girsanov's formula* for the change of measure L relating two diffusion processes. It is defined in terms of exponential integrals. The theory is a theorem saying that two SDE processes are related to to each other by a change of measure formula (Girsanov's formula) if and only if they have the same quadratic variation. Suppose there are two diffusion processes specified using two SDEs as

$$\begin{aligned} dX_t &= a_1(X_t)\, dt + b_1(X_t)\, dW_t \\ dX_t &= a_2(X_t)\, dt + b_2(X_t)\, dW_t \;\; . \end{aligned}$$

These are related by a change of measure if $b_1 = b_2$ but not otherwise (the explanation will help you see the "fine print" missing from this simplified statement.). If $b_1(x) = b_2(x)$, then it is possible to calculate expected values involving the first process using paths from the second process – re-weighted by a likelihood factor – *Girsanov's formula.*

It is tricky to figure out when two diffusions are related by a change of measure, but it is easy to figure out for simple random variables. Nothing can go wrong that isn't obvious. As one example for single random variables, suppose q represents a standard Gaussian and p represents the standard *Cauchy* distribution. The density functions are

$$q(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad , \quad p(x) = \frac{1}{\pi}\frac{1}{x^2+1} \; .$$

These distributions are both symmetric and centered around $x = 0$, but they have vastly different "tail behavior". The Gaussian tails are very small, and large values are quite unlikely. For example, $\mathrm{var}_q(X) = \sigma_q^2 = 1$ and

$$\Pr_q(X > \sigma) \approx .4 \cdot .2 \cdot 3.7 \times 10^{-6} \approx 3 \times 10^{-7} \; .$$

[This is from approximations:

$$\begin{aligned}
\int_a^\infty e^{-\frac{1}{2}x^2}\, dx &\approx e^{-\frac{1}{2}a^2} \int_0^\infty e^{-ay}\, dy = \frac{1}{a} e^{-\frac{1}{2}a^2} \\
\Pr_q(X > a) &= \frac{1}{\sqrt{2\pi}} \frac{1}{a} e^{-\frac{1}{2}a^2} \\
\frac{1}{\sqrt{2\pi}} &\approx .4 \\
\Pr_q(X > 5) &\approx .4 \cdot \frac{1}{5} \cdot e^{-12.5} \; .]
\end{aligned}$$

For the Cauchy distribution the variance is infinite (the integral diverges), but you can calculate

$$\begin{aligned}
\Pr_p(X > 5) &= \frac{1}{\pi} \int_5^\infty \frac{1}{x^2+1}\, dx \\
&\approx \frac{1}{\pi}\left[\int_5^\infty \frac{1}{x^2}\, dx - \int_5^\infty \frac{1}{x^4}\, dx + \cdots \right] \\
&\approx \frac{1}{\pi}\left[\left.\frac{-1}{x}\right|_5^\infty + \frac{1}{3}\left.\frac{1}{x^3}\right|_5^\infty - \cdots \right] \\
&= \frac{1}{\pi}\left[\frac{1}{5} - \frac{1}{3}\frac{1}{125} + \cdots \right] \\
&\approx \frac{1}{3.142 \cdot 5} \\
&\approx \frac{1}{16 - \epsilon} \\
&\approx .0625 + \epsilon \; .
\end{aligned}$$

The Cauchy tail probability is a bit over 6% and is larger than the Gaussian by a factor of about $2 \cdot 10^5$.

Despite the qualitative differences, there is a likelihood ratio relative p and q. It is

$$L(x) = \frac{\frac{1}{\pi}\frac{1}{x^2+1}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}} = \sqrt{\frac{2}{\pi}} \frac{e^{\frac{1}{2}x^2}}{x^2+1} \; .$$

This might look bad from the point of view of probability densities because it grows exponentially at infinity. Nevertheless, the change of measure formula (aL 6) is true. If $V(x)$ is non-negative the two integrals are equal and one is infinite if and only if the other is infinite. The large $L(x)$ for large x is necessary to "inflate" the small Gaussian tails to the large Cauchy tails.

For single component random variables, a likelihood ratio can overcome anything except exact zeros. If $q(x) = 0$ for $x < 0$ (say), and $p(x) > 0$ for $x < 0$ (say, q represents an exponential distribution with $q(x) = e^{-x}$ for $x > 0$ and zero for $x < 0$, and p represents the Cauchy distribution). No $L(x)$ can turn a zero density into a positive density. You can tell when densities are related by looking at where they are equal to zero.

You can say this in a fancier way using probabilities and "events". An *event* is a set of outcomes that may or may not happen. For example, the probability that $X > 5$ is the probability of the *event* that $X > 5$. The event is a set of x values: $A = \{x \mid x > 5\}$. A probability density $p(x)$ determines the probabilities of events $P(A)$ by

$$P(A) = \int_{x \in A} p(x)\, dx \; .$$

The *probability space* is denoted by Ω and is the set of all possible outcomes. For a one component random variable, this is the set of all (real) numbers, $\Omega = \mathbb{R}$. An event is a subset of Ω. The PDF p assigns a number number $P(A)$ to any event A. We say that $P(A)$ is the *probability measure* of A.

The probability density $q(x)$ defines a different probability measure $Q(A)$. If p can be obtained from q by re-weighting, which is $p(x) = L(x)q(x)$, and if A is an event with $Q(A) = 0$ then $P(A) = 0$ also. I will not explain a proof of this, because it is more theoretical than we have time for. Still, I hope two things are clear: (1) for single component random variable distributions described by probability densities, you can tell which ones are related to which other ones by looking at where the probability densities are equal to zero, and (2) you can express this by saying which events have probability zero in the probability measures defined by the probability densities.

Change of measure is more subtle for diffusions because the probability distributions are not described by probability densities. The probability space Ω is the set of "paths", which are continuous functions of t defined for $0 \leq t \leq T$. An event is a subset of Ω, which is a set of paths. Warning: not every subset of Ω is an event. Even for one component random variables, not every subset of $\mathbb{R}$ is an event. An event must be *measurable.* The precise definition of *measurability* is subtle Informally, any event that is defined by a finite number or a sequence of conditions or inequalities is measurable. That's why it's possible to do so much probability in path space (diffusion processes) without worrying about events and measurability. If you want to know more about measurability and related issues, look in a book on *real variables* or a fancy probability book. It is possible to be an intelligent user of probability without understanding the details of probability measures just like it's possible to be an intelligent user of calculus without understanding the mathematical definition of real number and limit.

Suppose P and Q are two probability distributions (*probability measures* is the technical term I try to avoid). For example, they could be given by two SDEs with $a_1(x)$, $b_1(x)$, etc. We write X for the whole path $X_{[0,T]}$. A likelihood function $L(X)$ is a function of of the whole path. [Keep in mind that functions of functions may be called "functionals".] Suppose $V(X)$ is another functional. We write $\mathrm{E}_P[V(X)]$ or $\mathrm{E}_Q[V(X)]$ for the expected value of V in the two models. They are related by likelihood ratio L if, for "any" functional V,

$$\mathrm{E}_P[\,V(X)\,] = \mathrm{E}_Q[\,V(X)L(X)\,]\;. \qquad (17) \qquad \boxed{\texttt{lr}}$$

This looks like the earlier formula (6), but that one was just about integrals and probability densities. This one is abstract. The expected values $\mathrm{E}_P[\cdot]$ and $\mathrm{E}_Q[\cdot]$ are abstract probability integrals rather than integrals over a single real variable x.

The meaning of (17) is for practical purposes similar to the meaning of (6). For example, suppose you can create sample paths X_n by simulation Suppose you want to know

$$a = \mathrm{E}_P[\,V(X)\,]\;.$$

Suppose that you can simulate the Q process and know $L(X)$, the latter being *Girsanov's formula.* Then you can estimate a in two ways

$$\begin{aligned}
&\text{(simulate) } N \text{ paths } X_n \text{ using } P\;,\;\; a \approx \frac{1}{N}V(X_n)\\
&\text{(simulate) } N \text{ paths } X_n \text{ using } Q\;,\;\; a \approx \frac{1}{N}V(X_n)L(X_n)\;.
\end{aligned}$$

The random variables $Y_P = V(X)$ when $X \sim P$ and $Y_Q = V(X)L(X)$ with $X \sim Q$ have different distributions – that's often the point of change of measure. But they have the same expected values. The random variable Y_Q might be more complicated than Y_P but it might have less variance. This would make the more complicated evaluator of a more accurate. It might be that the Q process is a simplified version of the P process that is easier to simulate. You can get the same expectation if you compensate using the likelihood ratio.

The basic fact about abstract change of measure is the *Radon Nikodym theorem*. It gives a condition under which there ia a likelihood ratio relating Q to P as in (17). The condition is that P is *absolutely continuous* with respect to Q. This means that if A is any event with $Q(A) = 0$, then $P(A) = 0$ also. A likelihood ratio cannot turn probability zero events into events with positive probability.

With probability densities, probability zero is easy to diagnose – look where the probability density is equal to zero. Probability measures are usually defined by taking limits involving probability densities. It can be hard to tell when the limiting probability is zero. The expression *almost surely* often is used in this context. An event B happens *almost surely*, in the P probability measure, if $P(B) = 1$. This is the same as saying the event $A =$ "not B" has $P(A) = 0$. The Radon Nikodym condition can be stated: "If B happens almost surely in the

Q measure then B happens almost surely in the P measure." Two probability measures are called *equivalent* (for change of measure) if almost sure for either one implies almost sure for the other.

The opposite of "equivalent" is *completely singular* (sometimes called "orthogonal", which is a mis-use of that term). Probability measures P and Q are completely singular if there is an event A with $P(A) = 1$ and $Q(A) = 0$. This may seem unlikely at first, but we will see that it's common for distributions defined by different SDEs to be completely singular with respect to each other. Note that the definition is symmetric with respect to P and Q. If $P(A) = 1$ and $Q(A) = 0$, and if $B =$ not A, then $P(B) = 0$ and $Q(B) = 1$.

This is not true for "absolutely continuous with respect to". It is possible that P is absolutely continuous with respect to Q but Q is not absolutely continuous with respect to P. For example, suppose x is a one component random variable, $p(x)$ is the exponential distribution and $q(x)$ is the Gaussian. You can make q from p by making $L(x) = 0$ for $x < 0$, but you can't make p from q because q never gives negative numbers. If this example seems artificial, it is. Most of the time two natural probability measures either are absolutely continuous or completely singular with respect to each other.

3.1 Almost surely

sec:as

This subsection has examples of "almost surely" in fancier settings. They come from probability distributions defined by limits (like Brownian motion) or distributions that depend on infinitely many variables. As an example of an infinitely-many-variables distribution, consider an infinite sequence of coin tosses $U_1, U_2, \cdots, U_n, \cdots$. Suppose that $U_k = 1$ with probability p and $U_k = 0$ with probability $1 - p$, and all the U_k are independent. You can ask: what is the probability that $U_k = 1$ for all k. The answer is: If $p < 1$ then $\Pr(U_k = 1 \text{ for all } k) = 0$. You can see this with a calculation using the fact that the U_k are independent

$$\Pr(U_k = 1, k = 1, \cdots, n) = p^n \ .$$

Therefore

$$\Pr(U_k = 1 \text{ for all } k) \leq p^n \ , \quad \text{for all } n \ .$$

Zero is the only probability that is smaller than $\frac{1}{n}$ for all n. You can make this a positive statement too: If U_k are independent and $\Pr(U_k = 1) = p < 1$ for all k, then $U_k = 0$ for some k almost surely.

Here's a more subtle example that's closer to the issue of Section sec:G 4. Suppose you estimate p from the sequence U_n. A natural estimator from the first n samples is

$$\begin{aligned} \widehat{p}_n &= \frac{1}{n} \# \left\{ U_k = 1 \mid 1 \leq k \leq n \right\} \\ \widehat{p}_n &= \frac{1}{n} \sum_{k=1}^{n} U_k \ . \qquad (18) \end{aligned}$$

phn

It is a theorem, see Subsection $\overset{\texttt{sec:BC}}{\text{3.2 that}}$

$$\widehat{p}_n \to p \;, \quad \text{as } n \to \infty \;\text{ almost surely } . \qquad\qquad (19) \quad \boxed{\texttt{sc}}$$

This is an instance of the *strong law of large numbers*. Suppose A_p is the event that $\widehat{p}_n \to p$ as $n \to \infty$. Let B_p be the "complementary" event that the limit either does not exist or is not equal to p. Then *almost surely* means $\mathrm{Pr}_p(A_p) = 1$. The event B_p "never happens" in the $p-$world, which is $\mathrm{Pr}_p(B_p) = 0$.

The *probability space* of this example is the space of infinite sequences of zeros and ones. The $p-$measure is the probability distribution in which the U_k are independent and $U_k = 1$ with probability p. Let $q \neq p$ be a different probability between 0 and 1. We can define the $q-$measure in which the U_k are independent and $\mathrm{Pr}_q(U_k = 1) = q$. The strong law of large numbers for the $q-$measure says that $\widehat{p}_n \to q$ as $n \to \infty$, almost surely. This means that in the $q-$world, it is "almost sure" that $\lim_{n\to\infty} \widehat{p}_n \neq p$. Thus

$$\mathrm{Pr}_q(A_p) = 0 \;. \qquad\qquad (20) \quad \boxed{\texttt{Apq}}$$

In this example we have two measures and a set A_p so that the $p-$measure of A_p is one (The *measure* of an event is its probability.) and the $q-$measure of A_p is zero.

$\boxed{\texttt{sec:BC}}$

3.2 Borel Cantelli

The *Borel Cantelli lemma* is a clever way to prove that limits happen almost surely. Here is a slightly non-standard way to explain the idea. It uses three ideas. First, if $S_n \geq 0$ is a sequence of non-negative numbers and if the infinite sum is finite, then the S_n have a limit

$$\sum_{n=1}^{\infty} S_n < \infty \quad\Longrightarrow\quad \lim_{n\to\infty} S_n = 0 \;. \qquad\qquad (21) \quad \boxed{\texttt{BCsl}}$$

This statement is about any sequence of numbers. It is not probabilistic. Second, if $R \geq 0$ is a random variable that is allowed to take the value $R = \infty$, and if the expected value is finite, then R itself is finite almost surely

$$\mathrm{E}[R] < \infty \quad\Longrightarrow\quad R < \infty \;\text{ almost surely } . \qquad\qquad (22) \quad \boxed{\texttt{BCff}}$$

Third, you can exchange the order of summation and taking expected values for non-negative numbers

$$\sum_{n=1}^{\infty} \mathrm{E}[\,S_n] = \mathrm{E}\left[\,\sum_{n=1}^{\infty} S_n\right] \;. \qquad\qquad (23) \quad \boxed{\texttt{BCmc}}$$

There is a little more about these claims below, but first the consequence.

You can combine all three claims and get

$$S_n \geq 0 \;, \;\; \sum_{n=1}^{\infty} \mathrm{E}[\,S_n] < \infty \quad\Longrightarrow\quad \lim_{n\to\infty} S_n = 0 \;, \;\text{ almost surely } . \qquad (24) \quad \boxed{\texttt{BC}}$$

Define R to be the sum of the non-negative numbers S_k:

$$R = \sum_{n=1}^{\infty} S_n \ .$$

The claim (23) tells you that if $\sum \mathrm{E}[\,S_n] < \infty$, then $\mathrm{E}[R] < \infty$. This is the right side of (23). Then (22) tells you that $R = \sum S_n < \infty$ almost surely. Finally, (21) tells you that $S_n \to 0$ as $n \to \infty$ almost surely. This (24) is a version of the Borel Cantelli lemma.

I hope the three claims (21), (22) and (23) are plausible. The first, (21), could be an exercise in an "$\epsilon - \delta$" introduction to analysis class. The second, (22) is a version of the inequality, for any $m > 0$, if $R \geq 0$, $\mathrm{E}[R] \geq m \Pr(R \geq m)$. This is a basic fact about probability measures (which are not defined in these notes). The third, (23) is a version of the *monotone convergence theorem.* The random variables

$$R_N = \sum_{n=1}^{N} S_n$$

are a monotone increasing sequence, which means that $R_{n+1} \geq R_n$. For any finite n,

$$\sum_{n=1}^{N} \mathrm{E}[\,S_n] = \mathrm{E}\left[\sum_{n=1}^{N} S_n\right] = \mathrm{E}[\,R_N] \ .$$

When $N \to \infty$, the left side converges to the left side of (23). The monotone convergence theorem says that the right side converges to the right side of (23).

We apply the Borel Cantelli lemma (24) and a calculation to prove the claim (19). The trick is to define

$$S_n = (\widehat{p}_n - p)^4 \ . \qquad (25) \qquad \boxed{\texttt{p4}}$$

The power 4 on the right doesn't change the fact that $\widehat{p}_n - p \to 0$, but it does help the sum on the left of (24) converge. To see that, think about $a_n = \frac{1}{\sqrt{n}}$. The sum of a_n does not converge, but the sum of a_n^4 is the sum of $\frac{1}{n^2}$, which is finite.

$$a_n = \frac{1}{\sqrt{n}} \implies \sum_{n=1}^{\infty} a_n = \infty \ , \text{ but } \sum_{n=1}^{\infty} a_n^4 = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty \ .$$

The Borel Cantelli lemma is the conceptual framework of the proof. The technical part is the calculation which shows that

$$\mathrm{E}\Big[\ (\widehat{p}_n - p)^4\Big] \leq \frac{1}{n^2} \ . \qquad (26) \qquad \boxed{\texttt{p4c}}$$

The inequality (26) is the result of calculations that are "straightforward" (not subtle), but a little complicated. We start with

$$\widehat{p}_n - p = \frac{1}{n}\sum_{k=1}^{n} (U_k - p) \ .$$

The U_k on the right comes from the definition $(\overset{\texttt{phn}}{18})$ of $\widehat{p}_n$. The p on the right comes from the p on the left and $\frac{1}{n}p = p$. Note that $\mathrm{E}[(U_k - p)] = 0$. We need a power of $\widehat{p}_n - p$. The trick for that is to express the power as a multiple sum. For example, the square is a double sum

$$\begin{aligned}
(\widehat{p}_n - p)^2 &= (\widehat{p}_n - p)\,(\widehat{p}_n - p) \\
&= \left(\frac{1}{n}\sum_{j=1}^{n}(U_j - p)\right)\left(\frac{1}{n}\sum_{k=1}^{n}(U_k - p)\right) \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}(U_j - p)\,(U_k - p)\ .
\end{aligned}$$

Therefore

$$\mathrm{E}\Big[\,(\widehat{p}_n - p)^2\Big] = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathrm{E}[\,(U_j - p)\,(U_k - p)]\ . \qquad (27) \quad \boxed{\texttt{php2}}$$

If $j \neq k$, then $U_j - p$ is independent of $U_k - p$ and both have expected value zero. This gives

$$\mathrm{E}[\,(U_j - p)\,(U_k - p)] = 0\ , \quad \text{if } j \neq k\ .$$

The terms with $j = k$ have

$$\mathrm{E}\Big[\,(U_k - p)^2\Big] = \mathrm{var}(U_k - p) = p(1-p)\ .$$

There are n terms with $j = k$, so the sum is

$$\begin{aligned}
\mathrm{E}\Big[\,(\widehat{p}_n - p)^2\Big] &= \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{var}(U_k - p) \\
&= \frac{1}{n^2}\,n\,p(1-p) \\
\mathrm{E}\Big[\,(\widehat{p}_n - p)^2\Big] &= \frac{p(1-p)}{n}\ .
\end{aligned}$$

This calculation may be familiar from elementary probability because it is the calculation that shows that the variance of the sample mean is proportional to $\frac{1}{n}$. This rate of "decay to zero" is not fast enough to apply the Borel Cantelli lemma, because the sum of $\frac{1}{n}$ is infinite.

We get a better power of n using the fourth power instead of the square. The reasoning that led to $(\overset{\texttt{php2}}{27})$ leads to

$$\mathrm{E}\Big[\,(\widehat{p}_n - p)^4\Big] = \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{E}[\,(U_i - p)\,(U_j - p)\,(U_k - p)\,(U_l - p)]\ .$$

As for the case with two $U - p$ terms, the expected value is zero if any of the indices i, j, k, l is not "matched" (equal to one of the others). For example, if $i \neq j$ and $i \neq k$ and $i \neq l$, then $U_i - p$ is independent of the other three terms and the expected value of the product is zero. You get terms with non-zero expectations in one of four ways. The first three ways involve distinct pairs and the fourth involves all four being equal

$$
\begin{aligned}
i = j\ ,\ \ k = l\ ,\ \ i \neq k\ ,\ \ n(n-1) \text{ terms}\ ,\ \ \mathrm{E}\Big[\, (U_i - p)^2\, (U_k - p)^2 \Big] &= p^2(1-p)^2 \\
i = k\ ,\ \ j = l\ ,\ \ i \neq j\ ,\ \ n(n-1) \text{ terms}\ ,\ \ \mathrm{E}\Big[\, (U_i - p)^2\, (U_j - p)^2 \Big] &= p^2(1-p)^2 \\
i = l\ ,\ \ j = k\ ,\ \ i \neq j\ ,\ \ n(n-1) \text{ terms}\ ,\ \ \mathrm{E}\Big[\, (U_i - p)^2\, (U_j - p)^2 \Big] &= p^2(1-p)^2 \\
i = j = k = l\ \ ,\ \ n \text{ terms}\ ,\ \ \mathrm{E}\Big[\, (U_i - p)^4 \Big] &= p(1-p)(1-2p+2p^2)\ .
\end{aligned}
$$

This leads to

$$
\mathrm{E}\Big[\, (\widehat{p}_n - p)^4 \Big] = \frac{3p^2(1-p)^2(n^2-n)}{n^4} + \frac{p(1-p)(1-2p+2p^2)n}{n^4}\ .
$$

Most of the complexity of this expression is irrelevant. What matters here is whether the sum is finite. The first term on the right has n^2 in the numerator and n^4 in the denominator. The other term on the right has n "upstairs" and n^4 "downstairs". The sum of the first terms is, more or less, the sum of $\frac{1}{n^2}$ which does converge. This is a proof of the *strong law* of large numbers (sc 19) for this case.

3.3 The Radon Nikodym theorem

sec:RN

The Radon Nikodym theorem is stated above. One way to show two probability measures are absolutely continuous with respect to each other is to find a formula for the likelihood ratio. That's what Girsanov did for diffusions. The other side of this is finding ways to show that two probability measures are completely singular with respect to each other.

One approach to this is related to *hypothesis testing* in statistics. In statistics, a *hypothesis* is a belief that a random variable X came from a probability measure P. For example,

4 Girsanov theory

sec:G

Girsanov theory says which diffusion processes are equivalent to each other and which are not. When two diffusion processes are equivalent, it gives a formula for $L(X)$, which is *Girsanov's formula.* The answer is, skipping the fine print, that two diffusions are equivalent if they have the same infinitesimal variance and they are completely singular with respect to each other otherwise. You can change the infinitesimal mean (the drift) but not the infinitesimal variance (quadratic variation) with a likelihood function change of measure.

You can see these facts explicitly for Brownian motion. We define $dX_t = X_{t+dt} - X_t$. Suppose there is a $W-$world where X_t is just Brownian motion, with zero infinitesimal mean and unit infinitesimal variance:

$$\begin{aligned} \mathrm{E}_W\left[\, dX_t \mid X_{[0,t]} \right] &= 0 \\ \mathrm{E}_W\left[\, (dX_t)^2 \mid X_{[0,t]} \right] &= dt \; . \end{aligned}$$

In the $Z-$world, there is an adapted infinitesimal mean function a_t so that

$$\begin{aligned} \mathrm{E}_Z\left[\, dX_t \mid X_{[0,t]} \right] &= a_t dt \\ \mathrm{E}_Z\left[\, (dX_t)^2 \mid X_{[0,t]} \right] &= dt \; . \end{aligned}$$

The likelihood function that changes measure from W to Z is given by Girsanov's formula

$$L(X) = e^{\int_0^T a_t dX_t} e^{-\frac{1}{2}\int_0^T a_t^2 dt} \; . \qquad (28) \quad \boxed{\texttt{G}}$$

If $V(x)$ is any path functional defined on $[0,T]$, then

$$\mathrm{E}_Z\left[\, V(X_{[0,T]}) \right] = \mathrm{E}_W\left[\, V(X_{[0,T]}) L(X_{0,T]}) \right] \; . \qquad (29) \quad \boxed{\texttt{Gcm}}$$

5 Exercises

$\boxed{\texttt{ex:is}}$

1. *Importance sampling* is a major application of re-weighting. Even if we are interested in an expectation with respect to a density $p(x)$, the density $q(x)$ might give more accurate estimates (lower variance) by putting more samples in the region that contributes to the expectation. For this exercise, suppose X is a lognormal random variable of the form $X = e^{\sigma Z + \mu}$. We want to test simulation based ways to estimate $a = \mathrm{E}[X]$. We have seen that the lognormal can be very skewed, so that most samples are far below the mean. Therefore, an importance sampling strategy might be to simulate instead from a distribution $X = e^{sZ+m}$.

   (a) Suppose $X = e^{\sigma Z + \mu}$ has $p(x)$ as its pdf and $X = e^{sZ+m}$ has $q(x)$ as its pdf. Write a formula for the likelihood ratio and the expected value of X as an expectation over q with a likelihood ratio.

   (b) Let $\mathrm{var}_{s,m}$ be the variance of $XL(X)$ in the q density. Find examples (using analysis, not simulation) where $s \neq \sigma$ and $m \neq \mu$ has less variance.

   Note: it is cumbersome to work with the lognormal density. This exercise will be easier if you re-formulate it in terms of normal random variables. Likelihood ratios for normals are simpler.

$\boxed{\texttt{ex:FKe}}$

2. Consider the example functional that has backward equation ($\overset{\texttt{FKe}}{\boxed{12}}$). Suppose X_t is Brownian motion and consider the random variable

$$Z_{x,t} = \int_t^T X_s \, ds \; , \quad \text{conditioned on } X_t = x \; .$$

Show that $Z_{x,t}$ is Gaussian and identify the mean and variance. Use that information to get $f(x,t) = \mathrm{E}\left[\, e^{Z_{x,t}} \right]$. This should agree with the answer (13) found by the ansatz.

Apply one or both of these methods to solve the interest rate model (15) and (16). Either find the solution by the ansatz method or find the mean and variance of the quantity in the exponent directly.

ex:FK2

3. Consider the multiplicative functional in which X_t is Brownian motion and
$$Y_t = e^{\int_t^T X_s^2 \, ds} \; , \quad f(x,t) = \mathrm{E}[\, Y_t \mid X_t = x] \; .$$
The random variable in the exponent is not Gaussian, but the Ansatz method may still apply. Look for a solution using the ansatz method that involves $e^{A(t)x^2}$.

ex:cG

4. Consider the Brownian motion with positive constant drift and positive starting point
$$dX_t = adt + dW_t \; , \quad X_0 = x_0 \; , \quad a > 0 \; , \quad x_0 > 0 \; .$$
Let τ be the first hitting time when $X_t = 0$. Be aware that $\Pr(\tau < \infty) < 1$, which means it is possible (there is a non-zero probability) that $X_t > 0$ for all $t \geq 0$. Make a histogram (write Python code to make a histogram) of τ for $a = .5$ and $x_0 = .3$. You need to specify a final time T. Do this by trial and error, so that there are not many hits after time T. In the code, these must not be "hard wired", which means, for example, that the code should use a variable `a` that is assigned `a = .5`. Reproduce the histogram by simulating Brownian motion directly and putting in the Girsanov change of measure weight. For this, you have to work with a *weighted histogram*, where values of τ come with weights. Part of the exercise is to figure out how to do this. When you're computing the Girsanov factor, you need only integrate up to τ, not the final time T. Why?