1 More on the probability flux

2 Properties of solutions of the forward and backward equations

3 Probability reweighting

Suppose you are interested in a particular probability measure P on a probability space Ω. Expected values are integrals with respect to the measure P

$$A_P \;=\; E[V] \;=\; \int_\Omega V(\omega)\, dP(\omega)\;.$$

Even of you are interested in A_P, there may be reason to take expectations with respect to a different probability measure Q. In order to do this, Q must be *absolutely continuous* with respect to P (definition below) and you have to know the *likelihood ratio* between P and Q. In that case

$$A_P \;=\; E_Q[VL] \;=\; \int_\Omega V(\omega)L(\omega)\, dQ(\omega)\;.$$

You get the answer A_P either by averaging in the "P measure" or by averaging in the "Q measure" and including the likelihood ratio.

Consider the finite dimensional case with probability densities. The probability space is $\Omega = \mathbb{R}^n$. The probability measures are $dP(x) = f(x)dx$ and $dQ(x) = g(x)dx$, where f and g are ordinary probability densities. Then

$$A_P \;=\; \int_{\mathbb{R}^n} V(x)\, f(x)dx \;=\; \int_{\mathbb{R}^n} V(x)\frac{f(x)}{g(x)}\, g(x)dx \;=\; \int_{\mathbb{R}^n} V(x)L(x)\, g(x)dx$$

This may be re-expressed as

$$E_f[V(X)] \;=\; E_g[V(X)L(X)] \quad \text{where } L(x) \;=\; \frac{f(x)}{g(x)}\;.$$

The ratio that defines L is called the *likelihood ratio* because the same ratio appears in the likelihood ratio test in statistics. Statisticians prefer to call probabilities and probability densities "likelihoods" in certain circumstances. *Reweighting* comes from the point of view that the probability density g is reweighted to become f. The reweighting function is the likelihood ratio, L.

One application of reweighting is the Monte Carlo technique of *importance sampling*. Suppose, for example, $A = \Pr(X > k)$, where X is a standard normal. This is quite small if k is large, but how small is it? The direct Monte Carlo method to estimate A would be to generate N independent $X_i \sim \mathcal{N}(0,1)$ and use

$$A \;\approx\; \frac{1}{N}\cdot \#\left\{X_i > k\right\}\;.$$

Each X_i with $X_i > k$ is a *hit*. For large enough N, the number of hits is approximately $N \cdot A$. But if $A = 10^{-3}$ (say), then only one in a thousand of the X_i will be hits. The rest are wasted.

Reweighting gives a more accurate estimate of A for the same number of samples. Suppose k is large. If you want to generate hits, you can sample a normal with mean k rather than mean zero. This will give about 50% hits. The two probability densities and the likelihood ratio are

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \;\;, \quad g(x) \;=\; \frac{1}{\sqrt{2\pi}} e^{-(x-k)^2/2} \;\;, \quad L(x) \;=\; e^{-kx+k^2/2} \;.$$

For $x > k$, $L(x) < e^{-k^2/2}$, which is small. The Monte Carlo method with reweighting is to generate $X_i \sim \mathcal{N}(k,1)$ and take

$$A \;\approx\; \frac{1}{N} \sum_{Z_i > k} L(X_i) \;=\; \frac{1}{N} \sum_{Z_i > k} e^{-kX_i + k^2/2} \;.$$

The estimator with reweighting (importance sampling) is more accurate because it estimates the tiny probability A as the average of N tiny numbers $L(X_i)$, with a large number of hits, rather than by counting very rare hits with each hit carrying weight 1, which is much larger. The reweighting method is called importance sampling because it tries to sample those parts of probability space that are important for determining the answer. In this case, that is $x > k$, but not too much larger.

In financial math, reweighting may be called "changing worlds", or changing *numeraire*. The picture is that in one world, $A = E_f[V(X)]$. In a different world, $A = E_g[V(X)L(X)]$. The number A is the same in both worlds, but the story behind A is different.

4 Absolute continuity

You might get the impression that you can re-weight any probability density to get any other one. But there are limits to this. For example, suppose X is positive in the f world and negative in the g world. You cannot reweight a positive number to become negative. More technically, suppose $f(x) = 0$ for $x < 0$ and $g(x) = 0$ for $x > 0$, then the supposed likelihood ratio is either infinite (if g is zero) or zero (if f is zero). The crucial formula $f(x) = L(x)g(x)$ is simply not true.

The general definition with respect to this issue is the following. Probability measure P is *absolutely continuous* with respect to probability measure Q if there is a function $L(\omega)$ so that $E_p[V(\omega)] = E_Q[V(\omega)L(\omega)]$ for more or less (OK, this is not the precise definition) any function V. The function L may be called the *Radon Nikodym derivative* because of the very informal manipulations

$$E_P[V] \;=\; \int V(\omega) dP(\omega) \;=\; \int V(\omega) \frac{dP(\omega)}{dQ(\omega)} \, dQ(\omega) \;=\; \int V(\omega) L(\omega) \, dQ(\omega) \;,$$

if you make the informal identification L as the quotient

$$L(\omega) \;=\; \frac{dP(\omega)}{dQ(\omega)}\;.$$

Of course, the quotient on the right may be hard to define rigorously.

If probability measures P and Q are given by probability densities in $\mathbb{R}^n$ it is obvious whether P is absolutely continuous with respect to Q. The only thing that can prevent this is $g(x) = 0$ in a region where $f(x) > 0$. Otherwise $L = \frac{g}{f}$ works. Something like this turns out to be true for any pair of probability measures on the same probability space Ω. Suppose that for any event $A \subset \Omega$,

$$Q(A) \;=\; \Pr_Q(A) \;=\; 0 \;\;\Longrightarrow\;\; P(A) \;=\; \Pr_P(A) \;=\; 0\;.$$

Then P is absolutely continuous with respect to Q and there is a reweighting function L that turns Q into P. In the finite dimensional example, let A be the event $A = \{x \mid f(x) > 0 \text{ and } g(x) = 0\}$. Since $g(x) = 0$ in A, it is clear that

$$\Pr_Q(A) \;=\; \int_A g(x)dx \;=\; 0\;.$$

If

$$\Pr_P(A) \;=\; \int_A f(x)dx > 0\;,$$

then there cannot be an $L(x)$ that turns g into f. You can find the proof of this theorem, called the *Radon Nikodym* theorem, in any good book that covers measure theory. I learned it from Walter Rudin's book *Real and Complex Analysis*.

Consider the probability densities $g(x)$ for the standard normal in one dimension and $f(x)$ for the rate one exponential. That is $g(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $f(x) = 0$ if $x < 0$ and $f(x) = e^{-x}$ if $x > 0$. Then f is absolutely continuous with respect to g because there are no x values where g vanishes. But g is not absolutely continuous with respect to f because $f = 0$ when $x < 0$. We say that probability measures P and Q are *equivalent* if P is absolutely continuous with respect to Q and Q is absolutely continuous with respect to P. The term is slightly unfortunate because the measures P and Q may give different answers to many questions: $E_P[V] \neq E_Q[V]$.

In the example above – g is gaussian and f is exponential – there is some "overlap" between f and g even though g is not absolutely continuous with respect to f. A more extreme case for probability measures P and Q is that there is some event A so that $\Pr_P(A) = 1$ and $\Pr_Q(A) = 0$. In this case, where there is absolutely no overlap between P and Q, the measures are called *completely singular* with respect to each other. Although there are obviously exceptions, as a general rule you should expect that if measures P and Q are not equivalent (each absolutely continuous with respect to the other) then they are completely singular with respect to each other.

You can look at this stuff from a statistician's point of view. Suppose you have an $\omega \in \Omega$ and you want to decide whether ω is a sample of P or Q. Usually

the goal is not to be right all the time, but to maximize the probability of being right. For example, suppose $\Omega = \mathbb{R}$, $P = \mathcal{N}(-1, 1)$ and $Q = \mathcal{N}(1, 1)$. It would make sense to guess P if $X < 0$ and Q if $X > 0$. More generally, if P and Q are equivalent and $L = \frac{dP}{dQ}$, then we could guess P if $L(\omega) > l_0$ and guess Q if $L(\omega) < l_0$. This is what statisticians call the *likelihood ratio test*. The *Neyman Pearson lemma* of statistics says that tests of this kind are essentially optimal.

If P and Q are completely singular with respect to each other, then this test is easy. If $\omega \in A$ you say P, otherwise Q. This will always be correct. That is, if $\omega \sim P$, then $\omega \in A$ almost surely, and if $\omega \sim Q$, then $\omega \notin A$ almost surely. So if you want to tell P samples from Q samples, you might look for properties of P and Q samples that hold almost surely, and are different.

A prime example of this is the "only if" part of *Girsanov's theorem*. This says that it is impossible to change the infinitesimal covariance of an SDE solution by reweighting. Suppose you have two SDE processes

$$dX_t \;=\; a(X_t)dt \,+\, b(X_t)dW_t \tag{1}$$

and

$$dX_t \;=\; f(X_t)dt \,+\, g(X_t)dW_t \;. \tag{2}$$

Both of these SDEs induce measures on the path space of continuous functions defined up to some time T: $\Omega = C[0, T]$. Let us suppose the a, b process (1) defines P and the f, g process (2) defines Q. The infinitesimal covariance is the $C(x)$ in the familiar formula (in the usual notation $\Delta X = X_{t+\Delta t} - X_t$)

$$E\big[\Delta X \Delta X^t \mid \mathcal{F}_t\big] \;=\; C(X_t)\Delta t \,+\, o(\Delta t) \;.$$

We saw that for (1) the infinitesimal covariance is $C_P(x) = b(x)b^t(x)$, and for (2) it is $C_Q(x) = g(x)g^t(x)$. We also studied the quadratic variation of an SDE solution and showed that, almost surely,

$$\lim_{\Delta t \to 0} \sum_{t_k < T} \big(X_{t_{k+1}} - X_{t_k}\big)\big(X_{t_{k+1}} - X_{t_k}\big)^t \;=\; \int_0^T C(X_t)\, dt \;. \tag{3}$$

For this sample space, a random outcome, ω, is the sample path from time 0 to time T, which is $\omega = X_{[0,T]}$. The formula (3) states that the infinitesimal covariance is a function of ω. Therefore if measures P and Q have $C_P(x) \neq C_Q(x)$ then you should be able to tell a P path from a Q path all the time.

The previous paragraph has some careful vague wording. The formula (3) is trying to say that if $C_P(x) \neq C_Q(x)$ then P and Q are completely singular. But it does not quite do that because it is possible, for example, that $C_P(x) = C_Q(x)$ when $x < 2$ (in one dimension) but not otherwise. In that contrived situation, telling P from Q would depend on the path. If the path ever crosses $x = 2$ you can tell, but there is a positive probability for that not to happen. A more typical example would be telling one diffusion coefficient from another. You can always tell $b = 1$ from $g = 2$ for example (if they are constant), or $b = 1$ from $g(x) = \sigma x$ (geometric Brownian motion).

5 Girsanov's theorem

The main part of Girsanov's theorem says that you can change one drift coefficient to another by re-weighting. This is possible for non-degenerate diffusions. The theorem is that if $b(b)b^t(x)$ is positive definite for every x (and sufficiently differentiable), and if $g(x)g(x)^t = b(x)b(x)^t$ for all x, then the measures P and Q are equivalent. Moreover, there is an explicit formula for the re-weighting function $L(X_{[0,T]}$ that takes Q to P. A slightly simpler statement is that any non-degenerate diffusion is equivalent to a martingale, which is a diffusion process with drift coefficient equal to zero. If you want to re-weight Q to P, you can do it two stages. You re-weight both Q and P to be a martingale. Call this *martingale measure* R. If L makes Q into R, and M makes P into R, then L/M makes Q into P. *Girsanov's formula* is the formula for $L(X)$ that changes the drift.

I explain the theorem and the re-weighting formula in the simple case of a one dimensional diffusion with constant noise

$$dX_t \;=\; a(X_t)dt \,+\, dW_t \;. \tag{4}$$

We want to re-weight this to set $a \equiv 0$, which is the case of standard Brownian motion. We suppose for simplicity that $X_0 = 0$. The argument here is quite simple in principle but it does involve some calculations.

Suppose you choose $\Delta t = T/n$ so that exactly n time steps take you to time T. Consider the n component vector $\vec{X}$ of n *observations* of the path X_t at the times $t_k = k\Delta x$: $\vec{X} = (X_{\Delta t}, X_{2\Delta t}, \ldots, X_T)$. Because $\vec{X}$ is a random element in $\mathbb{R}^n$, it has a probability density $u(\vec{x}) = u(x_1, \ldots, x_n)$. Here, x_k is the variable corresponding to X_{t_k}. We will find an approximate formula for u and see how it depends on a. Then it will be easy to find a factorization $u(\vec{x}) = L_n(\vec{X})v(\vec{x})$ where $v(\vec{x})$ is the probability density for $\vec{X}$ if $a \equiv 0$ in (4). The limit of $L_n(\vec{X})$ as $n \to \infty$ will be easy to identify. That will be Girsanov's formula in this example.

The joint probability density $u(\vec{x})$ is the product of the individual *transition densities* $u(x_{k+1} \mid x_k)$. By the Markov property,

$$u(x_{k+1} \mid (x_1, \ldots, x_k)) \;=\; u(x_{k+1} \mid x_k) \;.$$

Write $u(x_1, k \ldots, x_k)$ for the joint density of $(x_1, \ldots, x_k)$. Bayes' rule and the Markov property give

$$u(x_1, \ldots, x_{k+1}) \;=\; u(x_{k+1} \mid x_k)u(x_1, \ldots, x_k) \;=\; \cdots \;=\; \prod_{j=0}^{k} u(x_{j+1} \mid x_j) \;.$$

This gives

$$u(\vec{x}) \;=\; \prod_{j=1}^{n-1} u(x_{j+1} \mid x_j) \;. \tag{5}$$

This is an exact formula, but it is not so useful because we do not have a formula for the transition probability densities $u(x_{j+1} \mid x_j)$.

There is an approximate formula for the transition density that is accurate in the limit $\Delta t \to 0$. It comes from the forward Euler approximation to (4) from the last class. Applied to (4), this gives

$$X_{k+1} \;=\; X_k \,+\, a(X_k)\Delta t \,+\, \sqrt{\Delta t} Z_k \;,$$

where the Z_k are independent standard normals. In this approximation, the conditional density of X_{tk+1} conditional on $\mathcal{F}_{t_k}$ (i.e. conditioned on X_{t_k}) is normal with mean $X_{t_k} + a(X_{t_k})\Delta t$ and variance Δt. That is

$$u(x_{k+1} \mid x_k) \;\approx\; \frac{1}{\sqrt{2\pi\Delta t}} \exp\left[\frac{-\left(x_{k+1} - x_k - a(x_k)\Delta t\right)^2}{2\Delta t} \right] \tag{6}$$

Now multiply these together as (5) says to do, and you get

$$u(\vec{x}) \;\approx\; \frac{1}{(2\pi\Delta t)^{n/2}} \exp\left[\frac{-1}{2\Delta t} \sum_{j=1}^{n-1} \left(x_{j+1} - x_j - a(x_j)\Delta t\right)^2 \right] \;. \tag{7}$$

With $a \equiv 0$, (7) simplifies to

$$v(\vec{x}) \;=\; \frac{1}{(2\pi\Delta t)^{n/2}} \exp\left[\frac{-1}{2\Delta t} \sum_{j=1}^{n-1} \left(x_{j+1} - x_j\right)^2 \right] \;. \tag{8}$$

This formula is exact because with $a \equiv 0$, the process (4) is a standard Brownian motion. The transition density formula (6) is exact for Brownian motion because its transitions are exactly Gaussian.

It is relatively simple to get the formula for $L_n(\vec{x}) = u(\vec{x})/v(\vec{x})$. The $2\pi\Delta t$ factors cancel. Moreover

$$\left(x_{k+1} - x_k - a(x_k)\Delta t\right)^2 \;=\; \left(x_{k+1} - x_k\right)^2 - 2\left(x_{k+1} - x_k\right)\, a(x_k)\Delta t + a(x_k)^2 \Delta t^2 \;.$$

All this gives

$$L_n(\vec{x}) \;=\; \exp\left[\sum_{k=1}^{n-1} \left(x_{k+1} - x_k\right)\, a(x_k) \right] \exp\left[\frac{-1}{2} \sum_{k=1}^{n-1} a(x_k)^2 \Delta t \right] \;.$$

The problem is to find the limit as $n \to \infty$ when $X_k = X_{t_k}$ is the SDE process (4).

With the identification $x_k = X_{t_k}$, the exponent in the second factor contains

$$\sum_{j=1}^{n-1} a(x_k)^2 \Delta t \;\longrightarrow\; \int_0^T a(X_t)^2 dt \;\text{, as} \quad \Delta t \to 0 \;.$$

In the same way, the sum in the first factor converges to the Ito integral:

$$\sum_{k=1}^{n-1} a(X_{t_k}) \left(X_{t_{k+1}} - X_{t_k} \right) \longrightarrow \int_0^T a(X_t) dX_t \ .$$

Putting all this together gives Girsanov's formula for this case

$$L(X) \ = \ \exp\left[\int_0^T a(X_t) dX_t \ - \ \frac{1}{2} \int_0^T a(X_t)^2 dt \right] \ . \tag{9}$$

Weighting a Brownian motion by (9) turns it into a solution of the SDE (4), in the following sense. Suppose you want to evaluate $f = E[V(X_T)]$, where X_t satisfies the SDE (4). One way to do this is use the Euler method to generate a many (approximate) sample paths for (4). But our other formula for f is

$$f \ = \ E_{bm} \left\{ V(X_T) \exp\left[\int_0^T a(X_t) dX_t \ - \ \frac{1}{2} \int_0^T a(X_t)^2 dt \right] \right\} \tag{10}$$

In this formula, X_t is standard Brownian motion.