Stochastic Calculus Notes, Lecture 8

Last modified December 9, 2004

1 Path space measures and change of measure

1.1. Introduction: We turn to a closer study of the probability measures on path space that represent solutions of stochastic differential equations. We do not have exact formulas for the probability densities, but there are approximate formulas that generalize the ones we used to derive the Feynman integral (not the Feynman Kac formula). In particular, these allow us to compare the measures for different SDEs so that we may use solutions of one to represent expected values of another. This is the Cameron Martin Girsanov formula. These changes of measure have many applications, including importance sampling in Monte Carlo and change of measure in finance.

1.2. Importance sampling: *Importance sampling* is a technique that can make Monte Carlo computations more accurate. In the simplest version, we have a random variable, X, with probability density $u(x)$. We want to estimate $A = E_u[\phi(X)]$. Here and below, we write $E_P[\cdot]$ to represent expecation using the P measure. To estimate A, we generate N (a large number) independent *samples* from the *population* u. That is, we generate random variables X_k for $k = 1, \ldots, N$ that are independent and have probability density u. Then we estimate A using

$$A \approx \widehat{A}_u = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \; . \tag{1}$$

The estimate is *unbiased* because the *bias*, $A - E_u[\widehat{A}_u]$, is zero. The error is determined by the variance $\text{var}(\widehat{A}_u) = \frac{1}{N}\text{var}_u(\phi(X))$.

Let $v(x)$ be another probability density so that $v(x) \neq 0$ for all x with $u(x) \neq 0$. Then clearly

$$A = \int \phi(x)u(x)dx = \int \phi(x)\frac{u(x)}{v(x)}v(x)dx \; .$$

We express this as

$$A = E_u[\phi(X)] = E_v[\phi(X)L(X)] \;\; , \quad \text{where } \; L(x) = \frac{u(x)}{v(x)} \;\; . \tag{2}$$

The ratio $L(x)$ is called the *score function* in Monte Carlo, the *likelihood ratio* in statistics, and the *Radon Nikodym derivative* by mathematicians. We get a different unbiased estimate of A by generating N independent samples of v and taking

$$A \approx \widehat{A}_v = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k)L(X_k) \; . \tag{3}$$

The accuracy of (3) is determined by

$$\mathrm{var}_v(\phi(X)L(X)) = E_v[(\phi(X)L(X) - A)^2] = \int (\phi(x)L(x) - A)^2 v(x)dx \ .$$

The goal is to improve the Monte Carlo accuracy by getting $\mathrm{var}(\widehat{A}_v) << \mathrm{var}(\widehat{A}_u)$.

1.3. A rare event example: Importance sampling is particularly helpful in estimating probabilities of rare events. As a simple example, consider the problem of estimating $P(X > a)$ (corresponding to $\phi(x) = \mathbf{1}_{x>a}$) when $X \sim \mathcal{N}(0,1)$ is a standard normal random variable and a is large. The naive Monte Carlo method would be to generate N sample standard normals, X_k, and take

$$\left\{ \begin{array}{l} X_k \sim \mathcal{N}(0,1), \ k = 1, \cdots, N \ , \\ A = P(X > a) \approx \widehat{A}_u = \dfrac{1}{N} \# \left\{ X_k > a \right\} = \dfrac{1}{N} \displaystyle\sum_{X_k > a} 1 \ . \end{array} \right. \tag{4}$$

For large a, the *hits*, $X_k > a$, would be a small fraction of the samples, with the rest being wasted.

One importance sampling strategy uses v corresponding to $\mathcal{N}(a,1)$. It seems natural to try to increase the number of hits by moving the mean from 0 to a. Since most hits are close to a, it would be a mistake to move the mean farther than a. Using the probability densities $u(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $v(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-a)^2/2}$, we find $L(x) = u(x)/v(x) = e^{a^2/2} e^{-ax}$. The importance sampling estimate is

$$\left\{ \begin{array}{l} X_k \sim \mathcal{N}(a,1), \ k = 1, \cdots, N \ , \\ A \approx \widehat{A}_v = \dfrac{1}{N} e^{a^2/2} \displaystyle\sum_{X_k > a} e^{-aX_k} \ . \end{array} \right.$$

Some calculations show that the variance of $\widehat{A}_v$ is smaller than the variance of of the naive estimator $\widehat{A}_u$ by a factor of roughly $e^{-a^2/2}$. A simple way to generate $\mathcal{N}(a,1)$ random variables is to start with mean zero standard normals $Y_k \sim \mathcal{N}(0,1)$ and add a: $X_k = Y_k + a$. In this form, $e^{a^2/2} e^{-aX_k} = e^{-a^2/2} e^{-aY_k}$, and $X_k > a$, is the same as $Y_k > 0$, so the variance reduced estimator becomes

$$\left\{ \begin{array}{l} Y_k \sim \mathcal{N}(0,1), \ k = 1, \cdots, N \ , \\ A \approx \widehat{A}_v = e^{-a^2/2} \dfrac{1}{N} \displaystyle\sum_{Y_k > 0} e^{-aY_k} \ . \end{array} \right. \tag{5}$$

The naive Monte Carlo method (4) produces a small $\widehat{A}$ by getting a small number of hits in many samples. The importance sampling method (5) gets roughly 50% hits but discounts each hit by a factor of at least $e^{-a^2/2}$ to get the same expected value as the naive estimator.

1.4. Radon Nikodym derivative: Suppose Ω is a measure space with $\sigma-$algebra $\mathcal{F}$ and probability measures P and Q. We say that $L(\omega)$ is the Radon Nikodym derivative of P with respect to Q if $dP(\omega) = L(\omega)dQ(\omega)$, or, more formally,

$$\int_\Omega V(\omega)dP(\omega) = \int_\Omega V(\omega)L(\omega)dQ(\omega) \ ,$$

which is to say

$$E_P[V] = E_Q[VL] \ , \tag{6}$$

for any V, say, with $E_P[|V|] < \infty$. People often write $L = \frac{dP}{dQ}$, and call it the *Radon Nikodym derivative* of P with respect to Q. If we know L, then the right side of (6) offers a different and possibly better way to estimate $E_P[V]$. Our goal will be a formula for L when P and Q are measures corresponding to different SDEs.

1.5. Absolute continuity: One obstacle to finding L is that it may not exist. If A is an event with $P(A) > 0$ but $Q(A) = 0$, L cannot exist because the formula (6) would become

$$P(A) = \int_A dP(\omega) = \int_\Omega \mathbf{1}_A(\omega)dP(\omega) = \int_\Omega \mathbf{1}_A(\omega)L(\omega)dQ(\omega) \ .$$

Looking back at our definition of the abstract integral, we see that if the event $A = \{f(\omega) \neq 0\}$ has $Q(A) = 0$, then all the approximations to $\int f(\omega)dQ(\omega)$ are zero, so $\int f(\omega)dQ(\omega) = 0$.

We say that measure P is *absolutely continuous* with respect to Q if $P(A) = 0 \Rightarrow Q(A) = 0$ for every[1] $A \in \mathcal{F}$. We just showed that L cannot exist unless P is absolutely continuous with respect to Q. On the other hand, the *Radon Nikodym theorem* states that an L satisfying (6) does exist if P is absolutely continuous with respect to Q.

In practical examples, if P is not absolutely continuous with respect to Q, then P and Q are completely singular with respect to each other. This means that there is an event, $A \in \mathcal{F}$ with $P(A) = 1$ and $Q(A) = 0$.

1.6. Discrete probability: In discrete probability, with a finite or countable state space, P is absolutely continuous with respect to Q if and only if $P(\omega) > 0$ whenever $Q(x) > 0$. In that case, $L(\omega) = P(\omega)/Q(\omega)$. If P and Q represent Markov chains on a discrete state space, then P is not absolutely continuous with respect to Q if the transition matrix for P (also called P) allows transitions that are not allowed in Q.

1.7. Finite dimensional spaces: If $\Omega = R^n$ and the probability measures are given by densities, then P may fail to be absolutely continuous with respect to

[1] This assumes that measures P and Q are defined on the same $\sigma-$algebra. It is useful for this reason always to use the algebra of Borel sets. It is common to imagine *completing* a measure by adding to $\mathcal{F}$ all subsets of events with $P(A) = 0$. It may seem better to have more measurable events, it makes the change of measure discussions more complicated.

Q if the densities are different from zero in different places. An example with $n = 1$ is P corresponding to a negative exponential random variable $u(x) = e^x$ for $x \leq 0$ and $u(x) = 0$ for $x > 0$, while Q corresponds to a positive exponential $v(x) = e^{-x}$ for $x \geq 0$ and $v(x) = 0$ for $x < 0$.

Another way to get singular probability measures is to have measures using δ functions concentrated on lower dimensional sets. An example with $\Omega = R^2$ has Q saying that X_1 and X_2 are independent standard normals while P says that $X_1 = X_2$. The probability "density" for P is $u(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \delta(x_2 - x_1)$. The event $A = \{X_1 = X_2\}$ has Q probability zero but P probability one.

1.8. Testing for singularity: It sometimes helps to think of complete singularity of measures in the following way. Suppose we learn the outcome, ω and we try to determine which probability measure produced it. If there is a set A with $P(A) = 1$ and $Q(A) = 0$, then we report P if $\omega \in A$ and Q if $\omega \notin A$. We will be correct 100% of the time. Conversely, if there is a way to determine whether P of Q was used to generate ω, then let A be the set of outcomes that you say came from P and you have $P(A) = 1$ because you always are correct in saying P if ω came from P. Also $Q(A) = 0$ because you never say Q when $\omega \in A$.

Common tests involve *statistics*, i.e. functions of ω. If there is a (measurable) statistic $F(\omega)$ with $F(\omega) = a$ almost surely with respect to P and $F(\omega) = b \neq a$ almost surely with respect to Q, then we take $A = \{\omega \in \Omega \mid F(\omega) = a\}$ and see that P and Q are completely singular with respect to each other.

1.9. Coin tossing: In common situations where this works, the function $F(\omega)$ is a limit that exists almost surely (but with different values) for both P and Q. If $\lim_{n\to\infty} F_n(\omega) = a$ almost surely with respect to P and $\lim_{n\to\infty} F_n(\omega) = b$ almost surely with respect to Q, then P and Q are completely singular.

Suppose we make an infinite sequence of coin tosses with the tosses being independent and having the same probability of heads. We describe this by taking ω to be infinite sequences $\omega = (Y_1, Y_2, \ldots)$, where the k^{th} toss Y_k equals one or zero, and the Y_k are independent. Let the measure P represent tossing with $Y_k = 1$ with probability p, and Q represent tossing with $Y_k = 1$ with probability $q \neq p$. Let $F_n(\omega) = \frac{1}{n}\sum_{k=1}^{n} Y_k$. The (Kolmogorov strong) law of large numbers states that $F_n \to p$ as $n \to \infty$ almost surely in P and $F_n \to q$ as $n \to \infty$ almost surely in Q. This shows that P and Q are completely singular with respect to each other. Note that this is not an example of discrete probability in our sense because the state space consists of infinite sequences. The set of infinite sequences is not countable (a theorem of Cantor).

1.10. The Cameron Martin formula: The *Cameron Martin formula* relates the measure, P, for Brownian motion with drift to the Wiener measure, W, for standard Brownian motion without drift. Wiener measure describes the process

$$dX(t) = dB(t)\ . \tag{7}$$

The P measure describes solutions of the SDE

$$dX(t) = a(X(t),t)dt + dB(t) \; . \tag{8}$$

For definiteness, suppose $X(0) = x_0$ is specified in both cases.

1.11. Approximate joint probability measures: We find the formula for $L(X) = dP(X)/dW(X)$ by taking a finite Δt approximation, directly computing $L_{\Delta t}$, and observing the limit of L as $\Delta t \to 0$. We use our standard notations $t_k = k\Delta t$, $X_k \approx X(t_k)$, $\Delta B_k = B(t_{k+1}) - B(t_k)$, and $\vec{X} = (X_1, \ldots, X_n) \in R^n$. The approximate solution of (8) is

$$X_{k+1} = X_k + \Delta t a(X_k, t_k) + \Delta B_k \; . \tag{9}$$

This is exact in the case $a = 0$. We write $V(\vec{x})$ for the joint density of $\vec{X}$ for W and $U(\vec{x})$ for teh joint density under (9). We calculate $L_{\Delta t}(\vec{x}) = U(\vec{x})/V(\vec{x})$ and observe the limit as $\Delta t \to 0$.

To carry this out, we again note that the joint density is the product of the transition probability densities. For (7), if we know x_k, then X_{k+1} is normal with mean x_k and variance Δt. This gives

$$G(x_k, x_{k+1}, \Delta t) = \frac{1}{\sqrt{2\pi\Delta t}} e^{-(x_{k+1}-x_k)^2/2\Delta t} \; ,$$

and

$$V(\vec{x}) = \left(2\pi \; \Delta t\right)^{-n/2} \exp\left(\frac{1}{2\Delta t} \sum_{k=0}^{n-1} (x_{k+1} - k_k)^2 \right) \; . \tag{10}$$

For (9), the approximation to (8), X_{k+1} is normal with mean $x_k + \Delta t a(x_k, t_k)$ and variance Δt. This makes its transition density

$$G(x_k, x_{k+1}, \Delta t) = \frac{1}{\sqrt{2\pi\Delta t}} e^{-(x_{k+1}-x_k-\Delta t a(x_k,t_k))^2/2\Delta t} \; ,$$

so that

$$U(\vec{x}) = \left(2\pi \; \Delta t\right)^{-n/2} \exp\left(\frac{1}{2\Delta t} \sum_{k=0}^{n-1} (x_{k+1} - k_k - \Delta t a(x_k, t_k))^2 \right) \; . \tag{11}$$

To calculate the ratio, we expand (using some obvious notation)

$$\left(\Delta X_k - \Delta t a_k\right)^2 = \Delta x_k^2 - 2\Delta t \Delta x_k + \Delta t^2 a_k^2 \; .$$

Dividing U by V removes the 2π factors and the Δx_k^2 in the exponents. What remains is

$$\begin{aligned} L_{\Delta t}(\vec{x}) &= U(\vec{x})/V(\vec{x}) \\ &= \exp\left(\sum_{k=0}^{n-1} (a(x_k), t_k)(x_{k+1} - x_k) - \frac{\Delta t}{2} \sum_{k=0}^{n-1} a(x_k), t_k)^2 \right) \; . \end{aligned}$$

The first term in the exponent converges to the Ito integral

$$\sum_{k=0}^{n-1}(a(x_k),t_k)(x_{k+1}-x_k) \to \int_0^T a(X(t),t)dX(t) \quad \text{as } \Delta t \to 0,$$

if $t_n = \max\{t_k < T\}$. The second term converges to the Riemann integral

$$\Delta t \sum_{k=0}^{n-1} a(x_k),t_k)^2 \to \int_0^T a^2(X(t),t)dt \quad \text{as } \Delta t \to 0.$$

Altogether, this suggests that if we fix T and let $\Delta t \to 0$, then

$$\frac{dP}{dW} = L(X) = \exp\left(\int_0^T a(X(t),t)dX(t) - \frac{1}{2}\int_0^T a^2(X(t),t)dt\right) \ . \tag{12}$$

This is the Cameron Martin formula.

2 Multidimensional diffusions

2.1. Introduction: Some of the most interesting examples, curious phenomena, and challenging problems come from diffusion processes with more than one state variable. The n state variables are arranged into an n dimensional state vector $X(t) = (X_1(t),\ldots,X_n(t))^t$. We will have a Markov process if the state vector contains all the information about the past that is helpful in predicting the future. At least in the beginning, the theory of multidimensional diffusions is a vector and matrix version of the one dimensional theory.

2.2. Strong solutions: The drift now is a drift for each component of X, $a(x,t) = (a_1(x,t),\ldots,a_n(x,t))^t$. Each component of a may depend on all components of X. The σ now is an $n \times m$ matrix, where m is the number of independent sources of noise. We let $B(t)$ be a column vector of m *independent* standard Brownian motion paths, $B(t) = (B_1(t),\ldots,B_m(t))^t$. The stochastic differential equation is

$$dX(t) = a(X(t),t)dt + \sigma(X(t),t)dB(t) \ . \tag{13}$$

A strong solution is a function $X(t,B)$ that is nonanticipating and satisfies

$$X(t) = X(0) + \int_0^t a(X(s),s)ds + \int_0^t \sigma(X(s),s)dB(s) \ .$$

The middle term on the right is a vector of Riemann integrals whose k^{th} component is the standard Riemann integral

$$\int_0^t a_k(X(s),s)ds \ .$$

The last term on the right is a collection of standard Ito integrals. The k^{th} component is

$$\sum_{j=1}^{m} \int_0^t \sigma_{kj}(X(s),s) dB_j(s) \; ,$$

with each summand on the right being a scalar Ito integral as defined in previous lectures.

2.3. Weak form: The weak form of a multidimensional diffusion problem asks for a probability measure, P, on the probability space $\Omega = C([0,T],R^n)$ with filtration $\mathcal{F}_t$ generated by $\{X(s) \text{ for } s \leq t\}$ so that $X(t)$ is a Markov process with

$$E\left[\Delta X \;\middle|\; \mathcal{F}_t\right] = a(X(t),t)\Delta t + o(\Delta t) \; , \tag{14}$$

and

$$E\left[\Delta X \Delta X^t \;\middle|\; \mathcal{F}_t\right] = \mu(X(t),t)\Delta t + o(\Delta t) \; . \tag{15}$$

Here $\Delta X = X(t+\Delta t) - X(t)$, we assume $\Delta t > 0$, and $\Delta X^t = (\Delta X_1, \ldots, \Delta X_n)$ is the transpose of the column vector ΔX. The matrix formula (15) is a convenient way to express the short time variances and covariances[2]

$$E\left[\Delta X_j \Delta X_k \;\middle|\; \mathcal{F}_t\right] = \mu_{jk}(X(t),t)\Delta t + o(\Delta t) \; . \tag{16}$$

As for one dimensional diffusions, it is easy to verify that a strong solution of (13) satisfies (14) and (15) with $\mu = \sigma\sigma^t$.

2.4. Backward equation: As for one dimensional diffusions, the weak form conditions (14) and (15) give a simple derivation of the backward equation for

$$f(x,t) = E_{x,t}\left[V(X(T))\right] \; .$$

We start with the tower property in the familiar form

$$f(x,t) = E_{x,t}\left[f(x+\Delta X, t+\Delta t)\right] \; , \tag{17}$$

and expand $f(x+\Delta X, t+\Delta t)$ about (x,t) to second order in ΔX and first order in Δt:

$$f(x+\Delta X, t+\Delta t) = f + \partial_{x_k} f \cdot \Delta X_k + \tfrac{1}{2}\partial_{x_j}\partial_{x_k} \cdot \Delta X_j \Delta X_k + \partial_t f \cdot \Delta t + R \; .$$

Here follow the *Einstein summation convention* by leaving out the sums over j and k on the right. We also omit arguments of f and its derivatives when the arguments are (x,t). For example, $\partial_{x_k} f \cdot \Delta X_k$ really means

$$\sum_{k=1}^{n} \partial_{x_k} f(x,t) \cdot \Delta X_k \; .$$

[2] The reader should check that the true covariances $E\left[(\Delta X_j - E[\Delta X_j])(\Delta X_k - E[\Delta X_k]) \;\middle|\; \mathcal{F}_t\right]$ also satisfy (16) when $E\left[\Delta X_j \;\middle|\; \mathcal{F}_t\right] = O(\Delta t)$.

As in one dimension, the error term R satisfies

$$|R| \leq C \cdot \left(|\Delta X| \, \Delta t + |\Delta X|^3 + \Delta t^2 \right) \, ,$$

so that, as before,

$$E\left[|R|\right] \leq C \cdot \Delta t^{3/2} \, .$$

Putting these back into (17) and using (14) and (15) gives (with the same shorthand)

$$f = f + a_k(x,t)\partial_{x_k} f \Delta t + \tfrac{1}{2}\mu_{jk}(x,t)\partial_{x_j}\partial_{x_k} f \Delta t + \partial_t f \Delta t + o(\Delta t) \, .$$

Again we cancel the f from both sides, divide by Δt and take $\Delta t \to 0$ to get

$$\partial_t f + a_k(x,t)\partial_{x_k} f + \tfrac{1}{2}\mu_{jk}(x,t)\partial_{x_j}\partial_{x_k} f = 0 \, , \tag{18}$$

which is the backward equation.

It sometimes is convenient to rewrite (18) in matrix vector form. For any function, f, we may consider its *gradient* to be the row vector $\bigtriangledown_x f = D_x f = (\partial_{x_1} f, \ldots, \partial_{x_n} f)$. The middle term on the left of (18) is the product of the row vector Df and the column vector x. We also have the *Hessian* matrix of second partials $(D^2 f)_{jk} = \partial_{x_j}\partial_{x_k} f$. Any symmertic matrix has a *trace* $\text{tr}(M) = \sum_k M_{kk}$. The summation convention makes this just $\text{tr}(M) = M_{kk}$. If A and B are symmetric matrices, then (as the reader should check) $\text{tr}(AB) = A_{jk}B_{jk}$ (with summation convention). With all this, the backward equation may be written

$$\partial_t f + D_x f \cdot a(x,t) + \tfrac{1}{2}\text{tr}(\mu(x,t) D_x^2 f) = 0 \, . \tag{19}$$

2.5. Generating correlated Gaussians: Suppose we observe the solution of (13) and want to reconstruct the matrix σ. A simpler version of this problem is to observe

$$Y = AZ \, , \tag{20}$$

and reconstruct A. Here $Z = (Z_1, \ldots, Z_m) \in R^m$, with $Z_k \sim \mathcal{N}(0,1)$ i.i.d., is an m dimensional *standard normal*. If $m < n$ or $\text{rank}(A) < n$ then Y is a degenerate Gaussian whose probability "density" (measure) is concentrated on the subspace of R^n consisting of vectors of the form $y = Az$ for some $z \in R^m$. The problem is to find A knowing the distribution of Y.

2.6. SVD and PCA: The *singular value decomposition* (SVD) of A is a factorization

$$A = U \Sigma V^t \, , \tag{21}$$

where U is an $n \times n$ orthogonal matrix ($U^t U = I_{n\times n}$, the $n \times n$ identity matrix), V is an $m \times m$ orthogonal matrix ($V^t V = I_{m\times m}$), and Σ is an $n \times m$ "diagonal" matrix ($\Sigma_{jk} = 0$ if $j \neq k$) with nonnegative *singular values* on the diagonal: $\Sigma_{kk} = \sigma_k \geq 0$. We assume the singular values are arranged in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots$. The singular values also are called *principal components* and the SVD is called *principal component analysis* (PCA). The columns of U and V (not V^t) are *left* and *right singular vectors* respectively, which also are called principal components or *principal component vectors*. The calculation

$$C = AA^t = (U\Sigma V^t)(V\Sigma^t U^t) = U\Sigma\Sigma^t U^t$$

shows that the diagonal $n \times n$ matrix $\Lambda = \Sigma\Sigma^t$ contains the eigenvalues of $C = AA^t$, which are real and nonnegative because C is symmetric and positive semidefinite. Therefore, left singular vectors, the columns of C, are the eigenvectors of the symmetric matrix C. The singular values are the nonnegative square roots of the eigenvalues of C: $\sigma_k = \sqrt{\lambda_k}$. Thus, the singular values and left singular vectors are determined by C. In a similar way, the right singular vectors are the eigenvectors of the $m \times m$ positive semidefinite matrix A^tA. If $n > m$, then the σ_k are defined only for $k \geq m$ (there is no $\Sigma_{m+1,m+1}$ in the $n \times m$ matrix Σ). Since the rank of C is at most m in this case, we have $\lambda_k = 0$ for $k > m$. Even when $n = m$, A may be rank deficient. The rank of A being l is the same as $\sigma_k = 0$ for $k > l$. When $m > n$, the rank of A is at most n.

2.7. The SVD and nonuniqueness of A: Because $Y = AZ$ is Gaussian with mean zero, its distribution is determined by its covariance $C = E[YY^t] = E[AZZ^tA^t] = AE[ZZ^t]A^t = AA^t$. This means that the distribution of A determines U and Σ but not V. We can see this directly by plugging (21) into (20) to get

$$Y = U\Sigma(V^tZ) = U\Sigma Z' \;\; , \quad \text{where } \; Z' = V^tZ \; .$$

Since Z' is a mean zero Gaussian with covariance $V^tV = I$, Z' has the same distribution as Z, which means that $Y' = U\Sigma Z$ has the same distribution as Y. Furthermore, if A has rank $l < m$, then we will have $\sigma_k = 0$ for $k > l$ and we need not bother with the Z'_k for $k > l$. That is, for generating Y, we never need to take $m > n$ or $m > \text{rank}(A)$.

For a simpler point of view, suppose we are given C and want to generate $Y \sim \mathcal{N}(0, C)$ in the form $Y = AZ$ with $Z \sim \mathcal{N}(0, I)$. The condition is that $C = AA^t$. This is a sort of square root of C. One solution is $A = U\Sigma$ as above. Another solution is the *Choleski decomposition* of C: $C = LL^t$ for a lower triangular matrix L. This is most often done in practice because the Choleski decomposition is easier to compute than the SVD. Any A that works has the same U and Σ in its SVD.

2.8. Choosing $\sigma(x,t)$: This non uniqueness of A carries over to non uniqueness of $\sigma(x,t)$ in the SDE (13). A diffusion process $X(t)$ defines $\mu(x,t)$ through (15), but any $\sigma(x,t)$ with $\sigma\sigma^t = \mu$ leads to the same distribution of X trajectories. In particular, if we have one $\sigma(x,t)$, we may choose any adapted matrix valued function $V(t)$ with $VV^t \equiv I_{m\times m}$, and use $\sigma' = \sigma V$. To say this another way, if we solve $dZ' = V(t)dZ(t)$ with $Z'(0) = 0$, then $Z'(t)$ also is a Brownian motion. (The Levi uniqueness theorem states that any continuous path process that is weakly Brownian motion in the sense that $a \equiv 0$ and $\mu \equiv I$ in (14) and (15) actually is Brownian motion in the sense that the measure on Ω is Wiener measure.) Therefore, using $dZ' = V(t)dZ$ gives the same measure on the space of paths $X(t)$.

The conclusion is that it is possible for SDEs wtih different $\sigma(x,t)$ to represent the same X distribution. This happens when $\sigma\sigma^t = \sigma'\sigma'^{\,t}$. If we have μ, we may represent the process $X(t)$ as the strong solution of an SDE (13). For this, we must choose with some arbtirariness a $\sigma(x,t)$ with $\sigma(x,t)\sigma(x,t)^t = \mu(x,t)$. The number of noise sources, m, is the number of non zero eigenvalues of μ. We never need to take $m > n$, but $m < n$ may be called for if μ has rank less than n.

2.9. Correlated Brownian motions: Sometimes we wish to use the SDE model (13) where the $B_k(t)$ are correlated. We can accomplish this with a change in σ. Let us see how to do this in the simpler case of generating correlated standard normals. In that case, we want $Z = (Z_1, \ldots, Z_m)^t \in R^m$ to be a multivariate mean zero normal with $\text{var}(Z_k) = 1$ and given correlation coefficients

$$\rho_{jk} = \frac{\text{cov}(Z_j, Z_k)}{\sqrt{\text{var}(Z_j)\text{var}(Z_k)}} = \text{cov}(Z_j, Z_k) \ .$$

This is the same as generating Z with covariance matrix C with ones on the diagonal and $C_{jk} = \rho_{jk}$ when $j \neq k$. We know how to do this: choose A with $AA^t = C$ and take $Z = AZ'$. This also works in the SDE. We solve

$$dX(t) = a(X(t),t)dt + \sigma(X(t),t)AdB(t) \ ,$$

with the B_k being independent standard Brownian motions. We get the effect of correlated Brownian motions by using independent ones and replacing $\sigma(x,t)$ by $\sigma(x,t)A$.

2.10. Normal copulas (a digression): Suppose we have a probability density $u(y)$ for a scalar random variable Y. We often want to generate families $Y_1, \ldots, Y_m$ so that each Y_k has the density $u(y)$ but different Y_k are correlated. A favorite heuristic for doing this[3] is the *normal copula*. Let $U(y) = P(Y < y)$ be the cumulative distribution function (CDF) for Y. Then the Y_k will have density $u(y)$ if and only if $U(Y_k) - T_k$ and the T_k are uniformly distributed in the interval $[0,1]$ (check this). In turn, the T_k are uniformly distributed in $[0,1]$ if $T_k = N(Z_k)$ where the Z_k are standard normals and $N(z)$ is the standard normal CDF. Now, rather than generating independent Z_k, we may use correlated ones as above. This in turn leads to correlated T_k and correlated Y_k. I do not know how to determine the Z correlations in order to get a specified set of Y correlations.

2.11. Degenerate diffusions: Many practical applications have fewer sources of noise than state variables. In the strong form (13) this is expressed as $m < n$ or $m = n$ and $\det(\sigma) = 0$. In the weak form μ is always $n \times n$ but it may be

[3] I hope this goes out of fashion in favor of more thoughtful methods that postulate some mechanism for the correlations.

rank deficient. In either case we call the stochastic process a *degenerate diffusion.* Nondegenerate diffusions have qualitative behavior like that of Brownian motion: every component has infinite total variation and finite quadratic variation, transition densities are smooth functions of x and t (for $t > 0$) and satisfy forward and backward equations (in different variables) in the usual sense, etc. Degenerate diffusions may lack some or all of these properties. The qualitative behavior of degenerate diffusions is subtle and problem dependent. There are some examples in the homework. Computational methods that work well for nondegenerate diffusions may fail for degenerate ones.

2.12. A degenerate diffusion for Asian options: An Asian option gives a payout that depends on some kind of time average of the price of the underlying security. The simplest form would have th eunderlier being a geometric Brownian motion in the risk neutral measure

$$dS(t) = rS(t)dt + \sigma S(t)dB(t) \; , \tag{22}$$

and a payout that depends on $\int_0^T S(t)dt$. This leads us to evaluate

$$E\left[V(Y(T))\right] \; ,$$

where

$$Y(T) = \int_0^T S(t)dt \; .$$

To get a backward equation for this, we need to identify a state space so that the state is a Markov process. We use the two dimensional vector

$$X(t) = \begin{pmatrix} S(t) \\ Y(t) \end{pmatrix} \; ,$$

where $S(t)$ satisfies (22) and $dY(t) = S(t)dt$. Then $X(t)$ satisfies (13) with

$$a = \begin{pmatrix} rS \\ S \end{pmatrix} \; ,$$

and $m = 1 < n = 2$ and (with the usual double meaning of σ)

$$\sigma = \begin{pmatrix} S\sigma \\ 0 \end{pmatrix} \; .$$

For the backward equation we have

$$\mu = \sigma\sigma^t = \begin{pmatrix} S^2\sigma^2 & 0 \\ 0 & 0 \end{pmatrix} \; ,$$

so the backward equation is

$$\partial_t f + rs\partial_s f + s\partial_y f + \frac{s^2\sigma^2}{2}\partial_s^2 f = 0 \; . \tag{23}$$

Note that this is a partial differential equation in two "space variables", $x = (s,y)^t$. Of course, we are interested in the answer at $t = 0$ only for $y = 0$. Still, we have include other y values in the computation. If we were to try the standard finite difference approximate solution of (23) we might use a *central difference* approximation $\partial_y f(s,y,t) \approx \frac{1}{2\Delta y}(f(s,y+\Delta y,t) - f(s,y-\Delta y,t))$. If $\sigma > 0$ it is fine to use a central difference approximation for $\partial_s f$, and this is what most people do. However, a central difference approximation for $\partial_y f$ leads to an unstable computation that does not produce anything like the right answer. The inherent instability of centeral differencing is masked in s by the strongly stabilizing second derivative term, but there is nothing to stabalize the unstable y differencing in this degenerate diffusion problem.

2.13. Integration with dX: We seek the anologue of the Ito integral and Ito's lemma for a more general diffusion. If we have a function $f(x,t)$, we seek a formula $df = adt + bdX$. This would mean that

$$f(X(T),T) = f(X(0),0) + \int_0^T a(t)dt + \int_0^T b(t)dX(t) \ . \tag{24}$$

The first integral on the right would be a Riemann integral that would be defined for any continuous function $a(t)$. The second would be like the Ito integral with Brownian motion, whose definition depends on $b(t)$ being an adapted process. The definition of the dX Ito integral should be so that Ito's lemma becomes true.

For small Δt we seek to approximate $\Delta f = f(X(t+\Delta t), t+\Delta t) - f(X(t),t)$. If this follows the usual pattern (partial justification below), we should expand to second order in ΔX and first order in Δt. This gives (wth summation convention)

$$\Delta f \approx (\partial_{x_j} f)\Delta X_j + \tfrac{1}{2}(\partial_{x_j}\partial_{x_k} f)\Delta X_j \Delta X_k + \partial_t f \Delta t \ . \tag{25}$$

As with the Ito lemma for Brownian motion, the key idea is to replace the products $\Delta X_j \Delta X_k$ by their expected values (conditional on $\mathcal{F}_t$). If this is true, (15) suggests the general Ito lemma

$$df = (\partial_{x_j} f)dX_k + \left(\tfrac{1}{2}(\partial_{x_j}\partial_{x_k} f)\mu_{jk} + \partial_t f\right)dt \ , \tag{26}$$

where all quantities are evaluated at $(X(t),t)$.

2.14. Ito's rule: One often finds this expressed in a slightly different way. A simpler way to represent the small time variance condition (15) is

$$E\left[dX_j dX_k\right] = \mu_{jk}(X(t),t)dt \ .$$

(Though it probably should be $E\left[dX_j dX_k \bigm| \mathcal{F}_t\right]$.) Then (26) becomes

$$df = (\partial_{x_j} f)dX_k + \tfrac{1}{2}(\partial_{x_j}\partial_{x_k} f)E[dX_j dX_k] + \partial_t f dt \ .$$

This has the advantage of displaying the main idea, which is that the fluctuations in dX_j are important but only the mean values of dX^2 are important, not the fluctuations. Ito's rule (never enumciated by Ito as far as I know) is the formula

$$dX_j dX_k = \mu_{jk} dt \; . \tag{27}$$

Although this leads to the correct formula (26), it is not structly true, since the standard defiation of the left side is as large as its mean.

In the derivation of (26) sketched below, the total change in f is represented as the sum of many small increments. As with the law of large numbers, the sum of many random numbers can be much closer to its mean (in relative terms) than the random summands.

2.15. Ito integral: The definition of the dX Ito integral follows the definition of the Ito integral with respect to Brownian motion. Here is a quick sketch with many details missing. Suppose $X(t)$ is a multidimensional diffusion process, $\mathcal{F}_t$ is the $\sigma-$algebra generated by the $X(s)$ for $0 \leq s \leq t$, and $b(t)$ is a possibly random function that is adapted to $\mathcal{F}_t$. There are n components of $b(t)$ corresponding to the n components of $X(t)$. The Ito integral is ($t_k = k\Delta t$ as usual):

$$\int_0^T b(t) dX(t) = \lim_{\Delta t \to 0} \sum_{t_k < T} b(t_k) \left(X(t_{k+1}) - X(t_k) \right) \; . \tag{28}$$

This definition makes sense because the limit exists (almost surely) for a rich enough family of integrands $b(t)$. Let $Y_{\Delta t} = \sum_{t_k < T} b(t_k) \left(X(t_{k+1}) - X(t_k) \right)$ and write (for appropriately chosen T)

$$Y_{\Delta t/2} - Y_{\Delta t} = \sum_{t_k < T} R_k \; ,$$

where

$$R_k = \bigl(b(t_{k+1/2}) - b(t_k) \bigr) \bigl(X(t_{k+1}) - X(t_{k+1/2}) \bigr) \; .$$

The bound

$$E\left[\bigl(Y_{\Delta t/2} - Y_{\Delta t} \bigr)^2 \right] = O(\Delta t^p) \; , \tag{29}$$

implies that the limit (28) exists almost surely if $\Delta t_l = 2^{-l}$.

As in the Brownian motion case, we assume that $b(t)$ has the (lack of) smoothness of Brownian motion: $E[(b(t + \Delta t) - b(t))^2] = O(\Delta t)$. In the martingale case (drift $= a \equiv 0$ in (14)), $E[R_j R_k] = 0$ if $j \neq k$. In evaluating $E[R_k^2]$, we get from (15) that

$$E\left[\bigl| X(t_{k+1}) - X(t_{k+1/2}) \bigr|^2 \mid \mathcal{F}_{t_{k+1/2}} \right] = O(\Delta t) \; .$$

Since $b(t_{t+1/2})$ is known in $\mathcal{F}_{t_{k+1/2}}$, we may use the tower property and our assumption on b to get

$$E[R_k^2] \leq E\left[\bigl| X(t_{k+1}) - X(t_{k+1/2}) \bigr|^2 \bigl| b(t_{k+1/2}) - b(t) \bigr|^2 \right] = O(\Delta t^2) \; .$$

This gives (29) with $p = 1$ (as for Brownian motion) for that case. For the general case, my best effort is too complicated for these notes and gives (29) with $p = 1/2$.

2.16. Ito's lemma: We give a half sketch of the proof of Ito's lemma for diffusions. We want to use k to represent the time index (as in $t_k = k\Delta t$) so we replace the index notation above with vector notation: $\partial_x f \Delta X$ instead of $\partial_{x_k} \Delta X_k$, $\partial_x^2(\Delta X_k, \Delta X_k)$ instead of $(\partial_{x_j}\partial_{x_k} f)\Delta X_j \Delta X_k$, and $\mathrm{tr}(\partial_x^2 f \mu)$ instead of $(\partial_{x_j}\partial_{x_k} f)\mu_{jk}$. Then ΔX_k will be the vector $X(t_{k+1}) - X(t_k)$ and $\partial_x^2 f_k$ the $n \times n$ matrix of second partial derivatives of f evaluated at $(X(t_k), t_k)$, etc.

Now it is easy to see who $f(X(T), T) - f(X(0), 0) = \sum_{t_k < T} \Delta F_k$ is given by the Riemann and Ito integrals of the right side of (26). We have

$$\begin{aligned} \Delta f_k &= \partial_t f_k \Delta t + \partial_x f_k \Delta X_k + \tfrac{1}{2}\partial_x^2 f_k(\Delta X_k, \Delta X_k) \\ &\quad + O(\Delta t^2) + O\left(\Delta t \left|\Delta X_k\right|\right) + O\left(\left|\Delta X_k^3\right|\right) \ . \end{aligned}$$

As $\Delta t \to 0$, the contribution from the second row terms vanishes (the third term takes some work, see below). The sum of the $\partial_t f_k \Delta t$ converges to the Riemann integral $\int_0^T \partial_t f(X(t), t) dt$. The sum of the $\partial_x f_k \Delta X_k$ converges to the Ito integral $\int_0^T \partial_x f(X(t), t) dX(t)$. The remaining term may be written as

$$\partial_x^2 f_k(\Delta X_k, \Delta X_k) = E\left[\partial_x^2 f_k(\Delta X_k, \Delta X_k) \mid \mathcal{F}_{t_k}\right] + U_k \ .$$

It can be shown that

$$E\left[\left|U_k\right|^2 \mid \mathcal{F}_{t_k}\right] \le CE\left[\left|\Delta X_k\right|^4 \mid \mathcal{F}_{t_k}\right] \le C\Delta t^2 \ ,$$

as it is for Brownian motion. This shows (with $E[U_j U_k] = 0$) that

$$E\left[\left|\sum_{t_k < T} U_k\right|^2\right] = \sum_{t_k < T} E\left[\left|U_k\right|^2\right] \le CT\Delta t \ ,$$

so $\sum_{t_k < T} U_k \to 0$ as $\Delta t \to 0$ almost surely (with $\Delta t = 2^{-l}$). Finally, the small time variance formula (15) gives

$$E\left[\partial_x^2 f_k(\Delta X_k, \Delta X_k) \mid \mathcal{F}_{t_k}\right] = \mathrm{tr}\left(\partial_x^2 f_k \mu_k\right) + o(\Delta t) \ ,$$

so

$$\sum_{t_k < T} E\left[\partial_x^2 f_k(\Delta X_k, \Delta X_k) \mid \mathcal{F}_{t_k}\right] \to \int_0^T \mathrm{tr}\left(\partial_x^2 f(X(t), t)\mu(X(t), t)\right) dt \ ,$$

(the Riemann integral) as $\Delta t \to 0$. This shows how the terms in the Ito lemma (26) are accounted for.

2.17. Theory left out: We did not show that there is a process satisfying (14) and (15) (existence) or that these conditions characterize the process (uniqueness). Even showing that a process satisfying (14) and (15) with zero drift and

$\mu = I$ is Brownian motion is a real theorem: the Levi uniqueness theorem. The construction of the stochastic process $X(t)$ (existence) also gives bounds on higher moments, such as $E\left[|\Delta X|^4\right] \leq C \cdot \Delta t^2$, that we used above. The higher moment estimates are true for Brownian motion because the increments are Gaussian.

2.18. Drift change of measure:

The anologue of the Cameron Martin formula for general diffusions is the Girsanov formula.