1 Stochastic differential equations

A *stochastic differential equation*, usually called SDE, is a stochastic dynamical system of the form

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

A diffusion satisfies the SDE (1) if $G_t = a(X_t,t)$ and $F_t = b(X_t,t)$. The *coefficients* are $a(x,t)$, which is the *drift* coefficient, and $b(x,t)$, which is the *noise* coefficient. The SDE is *stationary* if the coefficients do not have explicit t dependence. Of course, even then they change with time as X_t changes. The noise if called *additive* if b is independent of x. Many problems in statistical physics and chemistry involve additive noise. The noise is *multiplicative* if b is genuinely x dependent.

If the noise coefficient vanishes, then (1) becomes the ordinary differential equation (ODE) $dX_t = a(X_t,t)dt$. This is more commonly expressed as

$$\frac{d}{dt}\,X(t) \;=\; a(X(t),t)\;.$$

ODE's are the way to express dynamics in continuous time with continuous paths. An SDE is a model with a deterministic part, which we call "drift" here, and noise, modeled by bdW.

A diffusion satisfies an SDE if it is a Markov process. To be a Markov process, the future must be independent of the past, conditional on the present. That means that the coefficients F_t and G_t must depend only on the state at time t, which is X_t. The formulas $G_t = a(X_t,t)$ and $F_t = B(X_t,t)$ express this dependence.

As an example, consider *geometric Brownian motion* (often called GBM):

$$dX_t \;=\; \mu X_t\,dt \;+\; \sigma X_t\,dW_t\;. \tag{2}$$

This is a common model in finance, where μ is the *rate of expected return* and σ is the *volatility*. Suppose $X_{1,t}$ and $X_{2,t}$ are geometric Brownian motions with parameters μ_1, σ_1 (for X_1), and μ_2 and σ_2 (for X_2). Let $Y_t = X_{1,t} + X_{2,t}$. Then Y is a diffusion, since

$$dY_t = G_t dt + F_t dW_t \quad \text{with} \quad G_t = \mu_1 X_{1,t} + \mu_2 X_{2,t} \quad \text{and} \quad F_t = \sigma_1 X_{1,t} + \sigma_2 X_{2,t}.$$

But Y_t by itself is not a Markov process because G_t and F_t are not determined by Y_t alone. If $\mu_1 \neq \mu_2$ and/or $\sigma_1 \neq \sigma_2$, then different values of $X_{1,t}$ and $X_{2,t}$ with the same $Y_t = X_{1,t} + X_{2,t}$ have different values of G_t and/or F_t. This example reinforces the philosophy that a stochastic process is a Markov process if the state has enough information. A stochastic process fails to be a Markov process if it does not keep enough information.

This example shows that we may be interested in SDE's that involve more than one component variable. To acommodate that, just suppose X_1 is the n

component column vector

$$X_t \;=\; \begin{pmatrix} X_{1,2} \\ X_{2,t} \\ \vdots \\ X_{n,t} \end{pmatrix} .$$

Then there n corresponding components of drift

$$a(X_t) \;=\; \begin{pmatrix} a_1(X_t) \\ a_2(X_t) \\ \vdots \\ a_n(X_t) \end{pmatrix} .$$

There are m *sources of noise*, modeled by the m component Brownian motion

$$W_t \;=\; \begin{pmatrix} W_{1,2} \\ W_{2,t} \\ \vdots \\ W_{m,t} \end{pmatrix} .$$

The $W_{k,t}$ are supposed to be independent standard Brownian motions, which means in particular that their $m \times m$ covariance matrix is $\operatorname{cov}(W_t) = tI_{m\times m}$. It is not required that $m = n$. The case $m < n$, which is called *degenerate diffusion*, is common in applications. The case $m = n$ with $\operatorname{rank}(b) = n$ is *non-degenerate*. We will see that there is no reason ever to take $m > n$. The noise "coefficient" is the $n \times m$ matrix

$$b(X_t) \;=\; \begin{pmatrix} b_{1,1}(X_t) & b_{1,2}(X_t) & \cdots & b_{1,m}(X_t) \\ b_{2,1}(X_t) & b_{2,2}(X_t) & & b_{2,m}(X_t) \\ & & & \\ \vdots & \vdots & & \vdots \\ b_{n,1}(X_t) & b_{n,2}(X_t) & & b_{n,m}(X_t) \end{pmatrix} .$$

Then bdW_t is the product of the $n \times m$ matrix b with the m component vector dW_t, which results in an n component vector that has the right dimension to influence X_t.

You need to be a little careful with the weak formulation of a multivariate SDE, particularly with the noise term. It is easy to see (at least to convince a reasonable person) that

$$E[\,\Delta X_j \mid \mathcal{F}_t] \;=\; a_j(X_t)\Delta t \,+\, o(\Delta t) \,.$$

There also is a covariance matrix $C(x)$ so that

$$\operatorname{cov}(\Delta X_t, \Delta X_k) \;=\; C_{jk}(X_t)\Delta t \,+\, o(\Delta t) \,. \tag{3}$$

We find $C(x)$ by writing

$$X_{j,t+\Delta t} \, - \, X_{j,t} \; = \; \Delta X_j \; = \; \int_t^{t+\Delta t} \sum_{i=1}^m b_{ji}(X_s) \, dW_{i,s} \; + \; \text{drift term.}$$

For small Δt, X_s is close to X_t, so to leading order we may replace $b_{ji}(X_s)$ with $b_{ji}(X_t)$. The result is (also ignoring the (smaller) drift term)

$$\Delta X_j \; \approx \; \sum_{i=1}^m b_{ji}(X_t) \, \Delta W_i \; .$$

This gives an approximation that is accurate enough to evaluate the covariance to leading order:

$$\Delta X_j \; \approx \; \sum_{i=1}^m b_{ji}(X_t) \, \Delta W_i \; .$$

We use the Kronecker δ to write

$$E[\, \Delta W_i \Delta W_l] \; = \; \delta_{il} \Delta t \; ,$$

and then

$$\begin{aligned} E[\, \Delta X_j \Delta X_k \mid \mathcal{F}_t] \; &\approx \; \sum_{i=1}^m \sum_{l=1}^m b_{ji}(X_t) b_{kl}(X_t) \, E[\Delta W_i \Delta W_l] \\ &= \; \sum_{i=1}^m b_{ji}(X_t) b_{ki}(X_t) \Delta t \; . \end{aligned}$$

We wrote $E[\, \Delta W_j \Delta W_k]$ instead of $E[\, \Delta W_j \Delta W_k \mid \mathcal{F}_t]$ because the conditioning does not change the answer (the independent increments property). Finally, note that $\sum_{i=1}^m b_{ji}(x) b_{ki}(x)$ is the (j,k) entry of the $n \times n$ matrix $b(x) b^t(x)$. This gives (3) with

$$C(x) \; = \; b(x) \, b^t(x) \; . \tag{4}$$

There is route to this result that does not muck with indices. Just write matrix/vector equation $\Delta X \approx b(X_t) \Delta t$ and then use[1]

$$\begin{aligned} \text{cov}(\Delta X) \; &= \; E\big[\Delta X \Delta X^t \big] \\ &= \; E\big[b(X_t) \Delta W \Delta W^t b^t(X_t) \mid \mathcal{F}_t \big] \\ &= \; b(X_t) \, E\big[\, \Delta W \Delta W^t \, \big] \, b^t(X_t) \\ &= \; b(X_t) \, I_{m \times m} \Delta t \, b^t(X_t) \\ &= \; b(X_t) b^t(X_t) b(X_t) \Delta t \; . \end{aligned}$$

This is the result (3) with (4).

[1] The transpose of AB is $B^t A^t$ for any matrix-matrix or matrix-vector product.

A weak solution to a multi-variate SDE is a process X_t with continuous sample paths that has the correct short time mean and variance above. Note that the weak formulation does not refer to noise coefficient $b(x)$, but only to the infinitesimal covariance (4). Any two processes that have the same $C(x)$ are identical from the weak point of view. But they may have SDE's with different coefficients b. As explained in the one dimensional context, one typically would derive an SDE in an application by finding a formula for $C(x)$. You then would then find any b that satisfies (4). You might use an analytic solution if you can find one. If you do it numerically, it may be convenient to use the Cholesky factorization of C.

This ambiguity – that more than one b is consistent with the same C – is not special to stochastic differential equations. Indeed, suppose $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}(0,1)$, and consider the models

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} , \tag{5}$$

and

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} . \tag{6}$$

From an observational point of view the models are indistinguishable Both have Gaussian output (X) with mean zero and covariance

$$\text{cov}(X) = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} .$$

But internally the models seem to tell different stories. The first (5) gives X_2 that depends only on the Z_2 factor while X_1 depends on both factors. The other model has these reversed: X_1 depends only on the first factor while X_2 depends on both. This is an instance of the warning statisticians are supposed to give: correlation (in this case, covariance) does not necessarily indicate causality.

2 Simulating SDEs and Monte Carlo

It is rare that you can solve a differential equation with a formula. It is even more rare for an SDE. Finding predictions from an SDE model may ask you to create approximate sample paths. We usually do this by time stepping. We have a small Δt, times $t_k = k\Delta t$ and seek approximate values

$$X_k \approx X(t_k) . \tag{7}$$

If you have X_k and you want X_{k+1}, the simplest thing is to seek something with approximately the correct mean and variance. That would be

$$X_{k+1} = X_k + a(X_k, t_k)\Delta t + b(X_k, t_k)\sqrt{\Delta t} Z_k . \tag{8}$$

If we define $\mathcal{F}_k$ as generated by $Z_0, \ldots, Z_k$, then (8) has the consequence that

$$E[\, X_{k+1} - X_k \mid \mathcal{F}_k] = a(X_k, t_k)\, \Delta t ,$$

and

$$E\Big[\left(X_{k+1} - X_k\right) \left(X_{k+1} - X_k\right)^t \mid \mathcal{F}_k \Big] \;=\; b(X_k, t_k) b^t(X_k, t_k)\, \Delta t \;.$$

exactly. Of course, the exact SDE process, by definition (the weak definition), satisfies these relations approximately.

One *time step* is the process of going from X_k to X_{k+1}. The time stepping method (8) is the *forward Euler* method. Without the noise term b, it is exactly the forward Euler method that you learn about in a class on ordinary differential equations. But there (non-noisy differential equations) there are whole families of more sophisticated, more accurate methods – Runge Kutta methods, linear multistep methods, implicit methods, etc. There are no known correct versions of most of these methods for stochastic differential equations. In an important sense, the plain Euler method (8) is as accurate as any other known method. This is not entirely true, as there are things like Milstein's method, which is better for some problems, and methods that treat the drift term in a fancier way. Still, the majority of actual computations of approximate sample paths for SDE's use Euler, and (unfortunately) rightly so.

When you generate an approximate sample path, you are trying to find a random path whose distribution is close to the distribution of the actual path X_t. It is possible for two random variables to have distributions that are close but not to be close themselves. One trivial example is that if $Z \sim \mathcal{N}(0,1)$ then Z and $-Z$ have the same distribution though they are not generally close. If you wanted to "simulate" a standard normal and returned $-Z$ instead of Z, nobody would know the difference. For the same reason, it usually is not important whether the approximate path value X_k is close to the exact path value $X(t_k)$. Instead, we should take (7) to be a hope that the distribution of X_k is close to the distribution of $X(t_k)$. This, in the approximate sample path business, is called *weak accuracy*. Weak accuracy is measured by how accurate expectations are. Suppose $V(X_{[0,T]})$ is some function of a random path[2], and $X^{\Delta t}$ is the approximate sample path produced, say, by (8), then weak accuracy asks how quickly $E\left[V(X) - V(X^{\Delta t})\right] \to 0$ as $\Delta t \to 0$. This applies to simple functions like $V(X) = V(X_T)$ (a function that depends only on the path value X_T) and to more complicated functions like $V(X) = \max_{t \leq T} X_t$.

There is a philosophical debate about what a random variable, or a random path, really means. How would you know if a single path were a sample solution of the SDE (1)? If you had a large number of such paths, you could ask whether they have the the correct distribution, but what would you ask about a single path (Warning, the Girsanov theorem is about things that are true for individual paths almost surely.). Most likely, if you are generating approximate sample paths, you are trying to learn $A = E[V(X)]$ for some function V. But A itself is not random.

Computing numbers like A by drawing random (approximate) random samples from some distribution is called *Monte Carlo*, named after the traditional gambling center of Europe (another place random numbers make a difference).

[2]Functions of paths are often called "functionals".

There is a helpful definition due to Mal Kalos:[3] Monte Carlo methods are computational methods that use random numbers as a tool to estimate some quantity that itself is not random. The benefit of this is that there may be more than one way to define A as the expected value of something involving a random variable. Some of these methods may be better than *direct simulation*. If you insist on generating (approximate) samples of the precise distribution (or path distribution) you started with, that is *simulation* rather than Monte Carlo. You might insist in direct simulation if you only want to get a feel for what sample paths look like.

The direct simulation method here would be to generate L independent (approximate) sample paths using independent forward Euler runs. Suppose i labels the path and k labels the time step. Then (8) gives

$$X^i_{k+1} \;=\; X^i_k + a(X^i_k\Delta t + b(X^i_k)\sqrt{\Delta t}Z^i_k \;.$$

Again, all the Z variables are independent standard normals. The estimate of A is

$$\widehat{A} \;=\; \frac{1}{L}\sum_{i=1}^{L} V(X^{i,\Delta t}) \;.$$

The error in Monte Carlo estimation has two components, *bias* and *statistical error* In the present context,

$$\text{bias} \;=\; A - E\left[\widehat{A}\right] \;=\; A - E\left[V\left(X^{i,\Delta t}\right)\right] \;.$$

This is not a random quantity although it is the expected value of a random quantity. Statistical error is random:

$$\text{statistical error} \;=\; E\left[\widehat{A}\right] \;-\; \widehat{A} \;.$$

Clearly, the total error is

$$\text{total error} \;=\; A - \widehat{A} \;=\; \text{(bias)} \;+\; \text{(statistical error)}.$$

You make the bias smaller by reducing Δt. If $T = t_n = n\Delta t$, this is the same as increasing the number of time steps, n, per sample path. You make the statistical error smaller by increasing L. In most applications, the bias is $O(\Delta t)$ or $O(\sqrt{\Delta t})$ (depending on V), and the statistical error is $O(L^{-1/2})$. The total work is proportional to the number of time steps in all, which is nL. Given the low accuracy and the large number of paths required for decent accuracy, Monte Carlo takes lots of computer time to give even not very accurate approximations.

The conclusion is that Monte Carlo estimation is almost never the best way to estimate A if there is a practical alternative. We will see that for low dimensional problems, a PDE approach usually is better.

[3] See, the excellent text *Monte Carlo Methods*, by Mal Kalos and Paula Whitlock.

3 Backward and forward equations

The backward and forward equations are a big part of the "calculus" part of stochastic calculus. There are methods for calculating things about diffusions.

Suppose X_t satisfies (1) for t in the range $0 \leq t \leq T$. Let $V(x)$ be a "payout" function and consider the *value function*

$$f(x,t) \;=\; E_{x,t}[V(X_T)] \;. \tag{9}$$

The notation $E_{x,t}[\cdot]$ refers to the expected value assuming that $X_t = x$. It is the same as saying

$$f(X_t,t) \;=\; E[\,V(X_T) \mid \mathcal{F}_t] \;. \tag{10}$$

We made this point earlier when talking about discrete state space Markov chains.

This value function satisfies a partial differential equation called the *backward equation* (or *Kolmogorov* or *Chapman-Kolmogorov* backward equation). One derivation works from the tower property applied over a short time interval Δt, together with the two definitions (9) and (10):

$$\begin{aligned} f(x,t) &= E_{x,t}[\,V(X_T)] \\ &= E_{x,t}\Big\{\,E[\,V(X_T) \mid \mathcal{F}_{t+\Delta t}]\Big\} \\ &= E_{x,t}[\,f(X_{t+\Delta t}, t+\Delta t)\,] \end{aligned}$$

The next step is to use a Taylor approximation of the right side. I write it for a scalar SDE then state the corresponding result for a system. The derivation for a system involves more writing.

$$f(X_{t+\Delta t}, t+\Delta t) \;\approx f(x,t) \;+\; \partial_x f(x,t)\Delta X \;+\; \frac{1}{2}\partial_x^2 f(x,t)\Delta X^2 \;+\; \partial_t f(x,t)\Delta t \;.$$

The terms that have been left out here lead, after taking expected values, to corrections that are $o(\Delta t)$, so they do not change the answer. Here, $\Delta X = X_{t+\Delta t} - x$. Since X satisfies the SDE in the weak sense, we have

$$E_{x,t}[\,\Delta X] \;=\; a(x)\Delta t + o(\Delta t) \;,$$

and

$$E_{x,t}\big[\,\Delta X^2\big] \;=\; b(x)^2\Delta t + o(\Delta t) \;.$$

With the Taylor approximation this gives

$$\begin{aligned} f(x,t) &= E_{x,t}\Big[\,f(x,t) \;+\; \partial_x f(x,t)\Delta X \;+\; \frac{1}{2}\partial_x^2 f(x,t)\Delta X^2 \;+\; \partial_t f(x,t)\Delta t\Big] \\ &= f(x,t) \;+\; \partial_x f(x,t)\, E_{x,t}[\,\Delta X] \;+\; \frac{1}{2}\partial_x^2 f(x,t)\, E_{x,t}\big[\,\Delta X^2\big] \;+\; \partial_t f(x,t)\Delta t \\ &= f(x,t) \;+\; \partial_x f(x,t)\, a(x)\Delta t \;+\; \frac{1}{2}\partial_x^2 f(x,t)\, b(x)^2\Delta t \;+\; \partial_t f(x,t)\Delta t \;+\; o(\Delta t) \;. \end{aligned}$$

The $o(\Delta t)$ on the end of the last line is a combination of the error in the Taylor approximation and the errors in the short time mean and variance of ΔX. The result (cancel f and divide by Δt and take the limit $\Delta t \to 0$) is

$$0 \;=\; \partial_t f(x,t) \;+\; \frac{1}{2}\partial_x^2 f(x,t)\, b(x)^2 \;+\; \partial_x f(x,t)\, a(x) \;. \tag{11}$$

This is the backward equation for a scalar SDE. For a system, you use more Taylor and the multivariate short time expectations. This replaces (11) with

$$0 \;=\; \partial_t f(x,t) \;+\; \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} C_{jk}(x)\, \partial_{x_j}\partial_{x_k} f(x,t) \;+\; \sum_{k=1}^{n} a_k(x)\partial_{x_k} f(x,t) \;. \tag{12}$$

Here (see above)

$$C_{jk}(x) \;=\; \sum_{i=1}^{m} b_{ji}(x) b_{ki}(x) \;,$$

or, in matrix form,

$$C(x) \;=\; b(x) b^t(x) \;.$$

The second derivative term may be expressed as the *trace* of a certain matrix. If M is an $n \times n$ matrix, then the trace of M is the sum of the diagonal entries of M:

$$\mathrm{Tr}(M) \;=\; \sum_{i=1}^{n} M_{ii} \;.$$

The trace has the mathematical property that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, where B is any other $n \times n$ matrix. Indeed, if $M = AB$, then a diagonal entry of M is

$$M_{ii} \;=\; \sum_{j=1}^{n} A_{ij} B_{ji} \;.$$

Therefore,

$$\mathrm{Tr}(AB) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} B_{ji} \;=\; \sum_{j=1}^{n}\sum_{i=1}^{n} B_{ji} A_{ij} \;=\; \mathrm{Tr}(BA) \;.$$

Note that the determinant also has this property, but unlike the determinant, $\mathrm{Tr}(AB) \neq \mathrm{Tr}(A)\mathrm{Tr}(B)$. If A or B (or both) is a symmetric matrix, then the order of the indices does not matter: $\mathrm{Tr}(AB) = \sum_{ij} A_{ij} B_{ij}$. There are different notations for derivatives you might see in the backward equation. One is a vector of first partials, written $\nabla f = Df$ with components $\nabla f_i = \partial_{x_i} f$. The other is the Hessian matrix of second partials $D^2 f$ with components $\left(D^2 f\right)_{ij} = \partial_{x_i}\partial_{x_j} f$. The backward equation then may be expressed as

$$0 \;=\; \partial_t f(x,t) \;+\; \frac{1}{2}\mathrm{Tr}\big(\, C(x) D^2 f(x,t)\big) \;+\; a(x) \cdot Df(x,t) \;. \tag{13}$$

The other equation in our pair is the *forward equation* (also called forward *Kolmogorov* equation or *Fokker Planck* equation). It concerns $u(x,t)$, which is the probability density of X_t. There is a direct derivation of the forward equation but we will see that it is not easy. An easier approach is to derive it from the backward equation using a *duality* argument. We argue that if the value function satisfies the backward equation for every payout function V and final time T, then u must satisfy a related PDE, the forward equation.

The derivation starts with the tower property written using the probability density. Let f be the unconditional expectation $f = E[V(X_T)]$, then we have

$$f \;=\; E\Big\{\, E[V(X_T) \mid \mathcal{F}_t]\Big\} \;=\; E[f(X_t,t)] \;=\; \int u(x,t) f(x,t)\, dx\ .$$

Notice that f does not depend on t, so if you differentiate with respect to t you get

$$0 \;=\; \frac{d}{dt}\int u(x,t) f(x,t)\, dx \;=\; \int \left(\partial_t u(x,t)\right) f(x,t)\, dx + \int u(x,t)\left(\partial_t f(x,t)\right)\, dx\ .$$

We continue from here in the scalar case first, then give the more general formula. Using (11), this becomes

$$0 \;=\; \int_{-\infty}^{\infty} \left(\partial_t u(x,t)\right) f(x,t)\, dx - \frac{1}{2}\int_{-\infty}^{\infty} u(x,t)\left(b^2(x)\partial_x^2 f(x,t)\right) dx - \int_{-\infty}^{\infty} u(x,t)\left(a(x)\partial_x f(x,t)\right) dx\ .$$

We integrate by parts to put all x derivatives on u. In doing this we ignore boundary terms at $x = \pm\infty$. This is because $u(x,t)$ is a probability density and therefore $u(x,t) \to 0$ as $x \to \pm\infty$. In doing this, we suppose that the other quantities, f, b, and b do not not grow as $x \to \pm\infty$ in a way that overpowers the decay of u. For the last term, we have

$$-\int_{-\infty}^{\infty} u(x,t)\left(a(x)\partial_x f(x,t)\right) dx \;=\; \int_{-\infty}^{\infty} \left[\,\partial_x\left(a(x)u(x,t)\right)\,\right] f(x,t) dx\ .$$

Notice that after integration by parts, the x derivative acts both on the u and on the drift coefficient a. For the diffusion term we integrate by parts twice to move both x derivatives off of f. Each time there is a change of sign, so there is no overall sign change:

$$\int_{-\infty}^{\infty} u(x,t)\left(b^2(x)\partial_x^2 f(x,t)\right) dx \;=\; \int_{-\infty}^{\infty} \left[\,\partial_x^2\left(b^2(x)u(x,t)\right)\,\right] f(x,t) dx\ .$$

Combining these results leads to

$$0 \;=\; \int_{-\infty}^{\infty} \left[\partial_t u(x,t) \;-\; \frac{1}{2}\partial_x^2\left(b(x)^2 u(x,t)\right) \;+\; \left(\partial_x a(x) u(x,t)\right)\right] f(x,t)\, dx \quad (14)$$

The stuff in square brackets depends only on the process X_t and not on the payout function $V(x)$ or the payout time T as long as $T > t$. Therefore, the

function $f(x,t)$ is more or less arbitrary. The only way the integral vanishes for every function $f(x,t)$ is if the thing multiplying f vanishes for all x. That implies:

$$\partial_t u(x,t) \;=\; \frac{1}{2}\partial_x^2 \left(b(x)^2 u(x,t)\right) \;+\; \left(\partial_x a(x) u(x,t)\right) \;. \tag{15}$$

This is the forward equation. The multi-dimensional version is derived in a similar way:

$$\partial_t u(x,t) \;=\; \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{x_i}\partial_{x_j}\left(C_{ij}(x)u(x,t)\right) \;-\; \sum_{i=1}^{n}\partial_{x_i}\left(a_i(x)u\right) \;. \tag{16}$$

4 Probability flux

The forward equation (16) can be formulated in terms of a *probability flux* $F(x,t)$ (also called *probability current* and written as $J(x,t)$). This is an n component object $F(x,t) = (F_1(x,t),\ldots,F_n(x,t))$. It represents the flow of probability through space at a given time. Suppose there were a large number of independent particles X_t^i, for $i = 1,\ldots,L$ moving according to the SDE (1). Imagine a hypersurface, Γ, in $\mathbb{R}^n$, or a curve in $\mathbb{R}^2$, or a point in $\mathbb{R}$. Each of these separates one part of space from another. The flux represents the net flow of particles across Γ per unit surface area (or per unit length for a curve in $\mathbb{R}^2$) per unit time. The individual particles move randomly so they cross Γ in both directions. Still there may be a net flux in one direction.

More precisely, suppose Γ is a simple closed surface that separates a finite volume interior from the exterior (a simple closed curve in two dimensions, two points in one dimension). "Simple" means that the surface or curve has no self intersections. Think of a circle or ellipse in two dimensions or the surface of a blob or donut in three dimensions. If Γ is smooth, there is a unit length normal vector, called $n(x)$, at each $x \in \Gamma$ that points out of the interior into the exterior. Let x be a point in Γ and $d\sigma$ a small patch of area (length for a curve in two dimensions) about x. Count the number of particle crossings from the interior to the exterior through $d\sigma$ and subtract the number of particle crossings from the exterior to the interior. Count over a small interval of time dt. The net difference is proportional to $d\sigma$ and dt. If there is a very large number of independent particles, the law of large numbers suggests that this net difference is large enough not to be very random. The net difference is proportional to $F(x,t)\cdot n(x)$. For example, if $n(x)$ is parallel to $F(x,t)$, then the flux per unit area per unit time is just $|F(x,t)|$ if F points outward, and $-|F|$ if F points inward. If F is parallel to Γ then there is no net flux, in keeping with $n\cdot F = 0$.

It may be easier to understand the probability flux when there is no noise. Then a particle at x simply moves with speed $a(x)$. If $u(x,t)$ represents the density of particles at point x at time t, then $n(x)\cdot a(x)u(x,t)d\sigma$ is the rate of particles crossing the piece of surface $d\sigma$ at the point x at time t. Therefore $F(x,t) = a(x)u(x,t)$ if there is no noise. We will soon replace this vague argument with something more mathematical.

The probability flux measures the net influx (flow into) or exflux (flow out of) the volume, V that is the interior of Γ. Since sample paths are continuous, a particle has to cross Γ and be counted in order to leave or enter V. This suggests that we can calculate $\frac{d}{dt}\Pr(X_t \in V)$ by integrating the flux over Γ. The formula is

$$\frac{d}{dt}\Pr(X_t \in V) \;=\; -\int_\Gamma F(x,t)\cdot n(x)\,d\sigma(x)\;. \tag{17}$$

The minus sign is because the integrand is positive if F points outward, in which case the probability of being inside Γ is decreasing.

The formula (17), together with the forward equation (16) gives a formula for $F(x,t)$. As we will check, this is

$$F_i(x,t) \;=\; -\frac{1}{2}\sum_{j=1}^{n}\partial_{x_j}\left(C_{ij}(x)u(x,t)\right) \;+\; a_i(x)u(x,t)\;. \tag{18}$$

Although we still have not derived this formula, you can check that it agrees with our idea of F in the case $C = 0$ (no diffusion). The backward equation may be written in the form

$$\partial_t u(x,t) \;=\; -\text{divergence}(F) \;=\; -\nabla F \;=\; -\sum_{i=1}^{n}\partial_{x_i}F_i(x,t)\;. \tag{19}$$

Finally, recall the *divergence theorem*, which is the identity[4]

$$\int_V \nabla F(x,t)\,dx \;=\; \int_\Gamma F(x,t)\cdot n(x)\,d\sigma(x)\;.$$

To put the pieces together, write the left side of (17) using the forward equation (16):

$$\begin{aligned}
\frac{d}{dt}\Pr(X_t\in V) &= \frac{d}{dt}\int_V u(x,t)\,dx && \text{(definition of probability density)}\\
&= \int_V \partial_t u(x,t)\,dx && (V \text{ is not changing in time})\\
&= -\int_V \nabla F(x,t)\,dx && \text{(forward equation in the form (19))}\\
&= -\int_\Gamma F(x,t)\cdot n(x)\,d\sigma(x)\;. && \text{(divergence theorem)}
\end{aligned}$$

[4]If you don't remember why this formula is true, you can verify it in simple cases. In one dimension, V is an interval that can be called (a,b). The outward normal points to the right at b and to the left at a. The flux has only one component, and the formula is $\int_a^b \partial_x F(x,t)dx = F(b) - F(a)$. In two or more dimensions you can verify the formula when V is a rectangle in the same way.

Now let us get the logic straight: This algebra shows that (17) is satisfied provided that F is defined by (18). This justifies calling (18) the probability flux.

The equations (19) and (18) are a very slightly different way to express the forward equation (16). They sometimes are called the *conservation form*, or *flux form* of the forward equation. The term conservation form comes from the idea that they express conservation of probability. That is, probability is not created or destroyed, only moved from place to place. Moreover, probability moves by "flowing" according to the probability flux. Probability cannot enter or leave the volume V without crossing Γ. This local conservation principle (the existence of the local probability flux) is not true, for example if X_t is a jump process with discontinuous sample paths.

It is possible to give a direct derivation of the forward equation in its conservation form (19) (18), but we do not have time to do it here.