Stochastic Calculus, Courant Institute, Fall 2022
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2022/index.html

Stochastic Calculus
Diffusions and Differential Equations

Jonathan Goodman, Fall, 2022

1 Introduction

A *differential equation* is an equation involving derivatives. A *partial differential equation*, or *PDE*, is an equation involving partial derivatives of a function of more than one variable. PDEs help us understand diffusion processes because quantities related to diffusions may satisfy PDEs. For example, suppose X_t is a one component process that satisfies the SDE

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

For each t, the value X_t is a one component random variable that has a probability density function, or *PDF*, which we write as $p(x,t)$. This satisfies the PDE

$$\partial_t p = \frac{1}{2}\partial_x^2\left(\,b(x)^2 p\right) - \partial_x\left(\,a(x)p\right) \ . \tag{2}$$

We use *operator* notation for partial derivatives. The partial derivatives of $f(x,t)$ are

$$\frac{\partial f}{\partial x} = \partial_x f \ , \ \ \frac{\partial^2 f}{\partial x^2} = \partial_x^2 f \ , \ \ \frac{\partial^2 f}{\partial x \partial t} = \partial_x\partial_t f \ , \ \ \text{etc.}$$

The operator notation is consistent with thinking of partial derivatives as "operating" on functions. These partial derivative operators have properties such as

$$\begin{aligned} \partial_x(\,\partial_x f) = \partial_x^2 f \ , & \ \ \text{applying an operator } \partial_x \text{ twice defines the square, } \partial_x^2 \\ \partial_x\partial_t f = \partial_t\partial_x f \ , & \ \ \text{derivatives with respect } x \text{ and } t \text{ "commute" .} \end{aligned}$$

The p equation (2) determines the *dynamics* of the probability density. You can think of $p(\cdot,t)$ as the "state" of the "evolving" probability distribution at time t. This "state" is a function of x. We write $p(\cdot,t)$ to emphasize the function, for fixed t, rather than the value at a specific point x. This function changes, as t increases, in a way that is determined by the PDE (2).

The dynamics alone do not determine p completely. You need more information, such as how p started at an *initial time*. We often take $t = 0$ to be the starting time. If you specify the distribution $p(\cdot,0)$, then the PDE (2) acts as an *evolution equation*, and determines the distribution $p(\cdot,t)$ for $t > 0$. The theory of PDEs of this "type" gives information about what $p(\cdot,t)$ can "look like", depending on $p(\cdot,0)$.

Partial differential equations that describe evolution can resemble ordinary differential equations, *ODEs*, that describe evolution. There is an important difference, which is the "direction of time". With an ODE, you can specify the state at time 0 and determine the state at time t, and you can determine the state at time 0 from the state at time t by "running the ODE backwards. As we will see, the PDE evolution equation (2) can be run forwards but not backwards. You can specify almost any probability density $p(\cdot, 0)$ and then there is a PDF $p(\cdot, t)$ defined for any $t > 0$ so that when you put in the x variable it satisfies the PDE (2). But you cannot go the other way. You cannot specify an arbitrary PDF $p(\cdot, T)$ for $T > 0$ and have have $p(\cdot, t)$ defined for $t < T$ in a way that the PDE (2) is satisfied.

The PDE (2) is useful in practice partly because there are computational methods for solving PDEs that can be applied. These lead to fast, accurate approximations to $p(\cdot, t)$ for $t > 0$ that do not require simulation, do not have statistical noise, and do not require large numbers of simulations to create a histogram (as was done in the first class). This "PDE solver" approach, unfortunately, is limited to diffusions in low dimensions, mainly $d = 1, 2, 3, 4$, with $d = 3$ being slow and $d = 4$ borderline infeasible. Simulation applies in any dimension, but gives noisy and less accurate results.

Value functions are other quantities related to diffusion processes that satisfy PDEs of evolution type. A *value function* is an expectation conditional on $X_t = x$. There are different kinds of value function, depending on what the conditional expectation is of. The *final time payout* is the case of the conditional expectation of a "payout" at a "final time" T, with the payout depending on X_T.

$$f(x,t) = \mathrm{E}[\, V(X_T) \mid X_t = x] \ . \tag{3}$$

For example, if you want to know the probability that $X_T > 2$, you take $V(x) = 0$ for $x < 2$ and $V(x) = 1$ for $x \geq 2$. Then $f(x,t)$ is the probability that $X_T \geq 2$ given that $X_t - x$. Value functions defined as (3) satisfy the PDE

$$\partial_t f = -\frac{1}{2} b(x)^2 \, \partial_x^2 f - a(x) \, \partial_x f \ . \tag{4}$$

This is an example of a *backward equation.* The name comes from the fact that the evolution goes backward in time. It is "obvious" that $f(x,T) = V(x)$. The PDE (4) evolves f backwards in time to determine $f(\cdot, t)$ for $t < T$. Like the *forward equation* (2) for p, this evolution equation has a direction of time constraint. This one goes only backwards while (2) goes only forwards.

One difference between the two PDEs is the sign of the term in involving ∂_x^2, which is positive for the forward equation (2) and negative for the backward equation (4). For PDEs like this (technical specifications omitted), the sign of the ∂_x^2 term determines the direction of time. Another difference is that the backward equation has the "coefficients" $b^2(x)$ and $a(x)$ "outside" the derivative, while the coefficients are "inside" for the forward equation. You can guess that the backward equation has coefficients outside the derivatives by realizing that the constant function $f(x,t) = C$ should be a solution. If $V(x) = C$ for all

x, then your "payout" is C no matter what X_T is. Therefore, any conditional expectation also should be equal to C, including the one that defines the value function (3). This should be true regardless of whether b and a are constants. If b and a are not constant, then $f(x,t) = C$ for all x, t satisfies the backward equation (4) but only because no derivatives of b or a are involved.

Generator

The similarities and differences between the forward and backward equations may be explained using the *generator* of the stochastic process (1). The generator is an operator that "acts on" functions of x, which may be called *test functions*. Suppose $g(x)$ is a smooth function.[1] We denote the generator by $\mathcal{L}$. It is defined by how it acts on functions,

$$\mathcal{L}g = h \iff h(x) = \lim_{t\to 0} \mathrm{E}\left[\frac{g(X_t) - g(x)}{t} \;\middle|\; X_0 - x \right] . \tag{5}$$

The generator is an *operator* that "acts on" functions. If g is a function, then $h = \mathcal{L}g$ is another function. It is like a matrix acting on vectors. The partial derivatives ∂_x and ∂_x^2 are other examples of operators. The generator for the diffusion process (1) is

$$\mathcal{L}g = \frac{1}{2} b(x)^2 \, \partial_x^2 g + a(x) \, \partial_x g \; . \tag{6}$$

Both sides of this formula are functions of x. You can get the value of $\mathcal{L}g$ at x by putting in $\partial_x^2 g(x)$ and $\partial_x g(x)$ on the right side. The value function evolution equation (4) may be expressed in terms of the generator as

$$\partial_t f + \mathcal{L} f = 0 \; . \tag{7}$$

This supposes f is a function of t and x. The ∂_t is the simple partial derivative of f with respect to t. The $\mathcal{L}f$ acts on f only in the x variables, with t treated as a parameter.

There is a sense of *adjoint* of an operator, which is denoted by a $*$. The adjoint of $\mathcal{L}$ is $\mathcal{L}^*$. The adjoint of the generator (6) is

$$\mathcal{L}^* q = \frac{1}{2} \partial_x^2 \left(b^2 q \right) - \partial_x \left(\, aq \, \right) \; . \tag{8}$$

For $\mathcal{L}^*$ it might be more confusing to leave out the x variable. To put it in, here is the formula for $(\mathcal{L}^* q)$ evaluated at a specific point x:

$$\left(\, \mathcal{L}^* q \, \right)(x) = \frac{1}{2} \partial_x^2 \left(b(x)^2 q(x) \right) - \partial_x \left(\, a(x) q(x) \, \right) \; .$$

[1] *Smooth* in this context means having enough derivatives for the calculations involved to make sense. For diffusions, "smooth" often means "has two continuous derivatives" but it might mean more. We say a function is "smooth" to avoid getting into a technical side-conversation about exactly which derivatives need to have what properties.

The forward equation (2) for p may be written as

$$\partial_t p = \mathcal{L}^* p \; . \tag{9}$$

We will use x and y for values X_t might have, but you have to keep track of which value is at time t and which is at time $t+s$.

Transition density

The forward equation (2) and backward equation (4) have solutions that may be expressed in terms of *transition density* for the corresponding stochastic process. The transition density is $G(x, y, s)$, which describes the evolution of the process over a positive time increment s. The process is assumed to start at $X_t = y$. As a function of x, G is the PDF of X_{t+s}.

The transition density determines the backward evolution of the value function because G is the probability density of X_T, conditioned on $X_{T-s} = y$. We write the expectation in (3) using the PDF of X_T:

$$f(y, T-s) = \int V(x)\, G(x, y, s)\, dx \; .$$

The forward evolution for the probability density comes from G and properties of conditional and marginal probability. Informally, if (X, Y) is a pair of random variables with joint probability distribution[2] $P(x, y)$, there is Bayes' rule, which gives $P(x, y) = P(x|y)P(y)$. The joint distribution, $P(x, y)$ is the product of the conditional $P(x|y)$ and the marginal $P(y)$. The marginal $P(x)$ is the sum or integral of the joint distribution over y. More formally, and using the probability densities and notation above,

$$p(x, t+s) = \int G(x, y, s)\, p(y, t)\, dy \; . \tag{10}$$

The density of X_{t+s} is given by the

Markov property

A stochastic process X_t is *Markov* (has the *Markov property* if the conditional probability of the future, conditional on the past, is the same as the conditional probability of the future, conditional on the present. The Markov property is that the state variable X has enough components so that X_t describes the state of the system being modeled completely, at time t. The probability distribution of $X_{[t,T]}$ depends on X_t, but otherwise not at all on X_s for $s < t$.

Informally, suppose there are times $t_1 < t_2 < \cdots < t_n$. The states at those times are $X_1, \cdots, X_n$. The joint distributions are given by Bayes' rule, as (we

[2] Many people in applied probability and statistics use $P(\cdots)$ for any probability distribution. Here, for example, $P(x)$ is the marginal distribution of X and $P(y)$ is the marginal distribution of Y, which are usually different. It seems strange to math people but it's easy to get used to.

use P for informal probabilities and p for probability densities associated to diffusion processes)

$$\begin{aligned}
P(x_1,x_2) &= P(x_1)P(x_2|x_1) \\
P(x_1,x_2,x_3) &= P(x_1)P(x_2|x_1)P(x_3|x_2,x_1) \\
&\vdots \\
P(x_1,\cdots,x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_2,x_1)\cdots P(x_n|x_{n-1},\cdots,x_1)\ .
\end{aligned}$$

Each new conditional probability depends on all the outcomes that have come before. For example, an expert watching a game of bridge knows that the card played next is random, but depends on the history of the game up to that point. Cards that were played early on give clues as to what cards each player still has.

The Markov property simplifies the complex conditional probabilities to probabilities depending only on the most recent outcome

$$P(x_n|x_{n-1},\cdots,x_1) = P(x_n|x_{n-1})\ .$$

A *Markov process* is one that knows the present but does not remember the past. If the process is Markov, then

$$\begin{aligned}
P(x_1,x_2) &= P(x_1)P(x_2|x_1) \\
P(x_1,x_2,x_3) &= P(x_1)P(x_2|x_1)P(x_3|x_2) \\
&\vdots \\
P(x_1,\cdots,x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_n|x_{n-1})\ .
\end{aligned}$$

The joint probability is still a product, but it is a product of simpler terms.

The SDE model (1) is Markov because the future evolution of X_t, which is given by dX_t, depends only on X_t through the noise and drift terms on the right, $b(X)t)$, and $a(X_t)$. The Markov property allows detailed probability densities involving the path $X_{[0,T]}$ to be specified as products of transition probabilities. As above, we take an increasing sequence of times $0 \le t_1 \le \cdots \le T$ and denote the corresponding states with the abbreviated notation $X_k = X_{t_k}$. The joint probability density for $X_1,\cdots,X_n$ at all these times is

$$p(x_1,\cdots,x_n,t_1,\cdots,t_n)\ .$$

Because of the Markov property, this may be expressed in terms of the starting density $p(x_1,t_1)$ and the transition densities.

$$\begin{aligned}
&p(x_1,\cdots,x_n,t_1,\cdots,t_n) \\
&\qquad = p(x_1,t_1)\,G(x_2,x_1,t_2-t_1)\ \cdots\ G(x_n,x_{n-1},t_n-t_{n-1})\ .
\end{aligned}$$

For example, the first G factor on the right is the PDF of X_{t_2}, conditioned on $X_{t_1} - x_1$. The time difference between t_1 and t_2 is $s = t_2 - t_1$.

2 Brownian motion, heat equation

Review of Brownian motion: Gaussian increments, etc.

This section uses dimensional Brownian motion to illustrate the ideas and formulas of Section 1. The variable X_t represents the total "input" starting at time 0. The input, of *innovation* between times t and $t+s>t$ is $X_{t+s}-X_t$. The input is "totally random" in that the mean of $X_{t+s}-X_t$ is zero. "Totally random" also means that innovations for distinct time intervals are independent. If $0 \leq t_1 \leq \cdots \leq t_n$ and $Y_1 = X_{t_1} - X_0$, $Y_n = X_{t_n} - X_{t_{n-1}}$, then the random innovations Y_k are a totally independent set.

Brownian motion is homogeneous in time, which means that the distribution of the innovation $Y_s = X_{t+s} - X_t$ depends only on s and not on t (or on anything that happens outside the interval $[t, t+s]$). Many properties of Brownian motion rely on the fact that the innovation Y_s may be thought of as the sum of two independent innovations ($\sim$ means that the random variables have the same distribution)

$$Y_s \sim Y^{(1)}_{\frac{s}{2}} + Y^{(2)}_{\frac{s}{2}} \; .$$

This is because

$$\begin{aligned} Y_s &= X_{t+s} - X_t \\ &= \left(X_{t+\frac{s}{2}} - X_t \right) + \left(X_{t+s} - X_{t+\frac{s}{2}} \right) \\ &= Y^{(1)}_{\frac{s}{2}} + Y^{(2)}_{\frac{s}{2}} \; . \end{aligned}$$

The last line has innovations $Y^{(1)}_{\frac{s}{2}} = X_{t+\frac{s}{2}} - X_t$ and $Y^{(2)}_{\frac{s}{2}} = X_{t+s} - X_{t+\frac{s}{2}}$. The innovations have the same distribution because the distribution depends only on the time increment $\frac{s}{2}$. They are independent because the time intervals are disjoint.

We can divide the interval into n pieces instead of two, which gives

$$Y_s \sim \sum_{k=1}^{n} Y^{(k)}_{\frac{s}{n}} \; . \tag{11}$$

The central limit theorem suggests that Y_s is approximately normal. The normal approximation gets better as $n \to \infty$. But the distribution of Y_s does not depend on n, which suggests that The distribution of Y_s is exactly normal. This is a theorem of probability. A random variable that can be represented as a sum of n i.i.d. (independent and identically distributed) pieces, for any integer $n>0$ is called *infinitely divisible*. An infinitely divisible random variable that has finite variance[3] is Gaussian. The variance of a sum of independent random variables is the sum of the variances. This implies that

$$\mathrm{var}(\, Y_s) = n \, \mathrm{var}\!\left(\, Y^{(1)}_{\frac{s}{n}} \right) \; .$$

[3] There are infinitely divisible distributions that have infinite variance and are not Gaussian. One example is the *Cauchy* distribution, with PDF $p(y) = C\frac{1}{1+y^2}$.

If n is even, you can take half the terms on the right of (11) to get a representation of $Y_{\frac{s}{2}}$. Therefore

$$\mathrm{var}(\,Y_s) = \frac{n}{2}\,\mathrm{var}\Big(\,Y^{(1)}_{\frac{s}{n}}\Big) + \frac{n}{2}\,\mathrm{var}\Big(\,Y^{(1)}_{\frac{s}{n}}\Big) = 2\,\mathrm{var}\big(\,Y_{\frac{s}{2}}\big)\ .$$

If you think about this more, you will see that the variance of Y_s must be proportional to s. The *standard* Brownian motion has the variance equal to s. To summarize, for standard Brownian motion

- Innovations for disjoint time intervals are independent, and have finite variance
- The distribution of an innovation depends only on the length of the time interval
- The distribution, therefore, is Gaussian
- The innovation for an interval of length s is Gaussian with mean zero and variance s.

The generator

There are many ways to find the forward and backward equations related to Brownian motion. Some are based on direct physical reasoning. The using the generator is simple mathematically but doesn't inspire as much physical intuition.

The generator of the standard Brownian motion process is identified from the generator definition (5) and Taylor series. This derivation also applies to diffusion processes that satisfy stochastic differential equations. For small t, which is the limit in (5), X_t is likely to be close to x. The value X_t may be written as $X_t = x + Y_t$, where Y_t is the innovation. The innovation is Gaussian with mean zero and variance t, which is small. A helpful mathematical trick is to write

$$Y_t = \sqrt{t}\,Z\ ,\quad Z \sim \mathcal{N}(0,1)\ .$$

A Gaussian with variance t is a "standard" normal (mean zero, variance 1) scaled by $\sqrt{t}$. Therefore $g(X_t)$ may be replaced by $g(x+\sqrt{t}\,Z)$ in the expectation (5). We want to divide by t and take the limit $t \to 0$. For that reason, we expand $g(x+\sqrt{t}\,Z)$ in Taylor series, keeping only terms that do not go to zero when divided by t:

$$g(x+\sqrt{t}\,Z) = g(x) + [\,\partial_x g(x)]\,\sqrt{t}\,Z\,Z + \frac{1}{2}\left[\,\partial_x^2 g(x)\right] tZ^2 + \frac{1}{6}\,[\cdots]\,t^{\frac{3}{2}}Z^3 + \cdots\ .$$

The last term on the right, when divided by t, still has $t^{\frac{1}{2}}$, which goes to zero in the limit $t \to 0$. Therefore we neglect it and all terms after it.

In the expectation of (5), x and t are not random and therefore may be taken outside the expectation. We use the Taylor expansion, dropping terms beyond tZ^2, to put the ratio of (5) into the form

$$\frac{g(X_t) - g(x)}{t} = t^{-\frac{1}{2}} \left[\partial_x g(x) \right] Z + \frac{1}{2} \left[\partial_x^2 g(x) \right] Z^2 + o(t) \; .$$

The "little oh" notation $o(t)$ is a way of denoting any quantity that goes to zero as t goes to zero. With expectations, we get

$$\mathrm{E}\left[\frac{g(X_t) - g(x)}{t} \right] = t^{-\frac{1}{2}} \left[\partial_x g(x) \right] \mathrm{E}[\, Z] + \frac{1}{2} \left[\partial_x^2 g(x) \right] \mathrm{E}\left[\, Z^2 \right] + o(t) \; .$$

In this expression, $\mathrm{E}[\, Z] = 0$ (i.e., Z has mean zero) and $\mathrm{E}\left[\, Z^2\right] = 1$ (i.e., Z has variance 1). The $o(t)$ goes to zero in the $t \to 0$ limit. This leaves

$$\mathcal{L} g(x) = \frac{1}{2} \partial_x^2 \, g(x) \tag{12}$$

This shows that, as an operator,

$$\mathcal{L} = \frac{1}{2} \partial_x^2 \; . \tag{13}$$

The generator for Brownian motion is simple.

The backward equation (7) for the final time payout value function (3), for Brownian motion, is

$$\partial_t f + \frac{1}{2} \partial_x^2 f = 0 \; . \tag{14}$$

The adjoint of the generator, in this case with $b = 1$, is

$$\mathcal{L}^* = \frac{1}{2} \partial_x^2 \; .$$

In this case, $\mathcal{L}$ is the same as $\mathcal{L}^*$. Operators with this property are called *self adjoint.* Self adjoint operators are common in mathematics, but not common as generators of diffusion processes. The forward equation (9) for Brownian motion is

$$\partial_t p = \frac{1}{2} \partial_x^2 p \; . \tag{15}$$

This equation is sometimes called "the" diffusion equation or (often without the $\frac{1}{2}$) the *heat equation.* The forward and backward equations involve ∂_x^2, but with opposite sign. It is traditional to write the backward equation in the form (14) to avoid the minus sign you would get by putting it on the other side of the equation.

Probability flux, probability current

There are ways to derive the forward equation (15) that make the equation itself seem more natural and explain its structure. One approach involves *probability* *flux*, also called probability *current*. You think of probability as a substance that moves around. If X_{t_1} is likely to be close to zero, then the probability for X_{t_1} is concentrated near zero. If X_{t_2} is likely to be further from zero, then the probability for X_{t_2} is less concentrated around zero. Between times t_1 and t_2, some of the probability moved away from zero. The forward equation (15) describes this movement of probability.

You can visualize probability by thinking of many independent Brownian motion paths. We use notation in which there are n paths in all, written $X_t^{(k)}$ for $k = 1, \cdots, n$. The number of paths in an interval $[a, , b]$ at time t is

$$N_{[a,b]}(t) = \#\left\{ k \mid a \leq X_t^{(k)} \leq b \right\} \; .$$

It is written with a capitol letter N instead of n because N is random. If Δx is small, then its probability density $p(x,t)$ does not change much in the interval, so

$$\Pr(x \leq X_t \leq x + \Delta x) = \int_x^{x+\Delta x} p(x', t)\, dx' \approx \Delta x\, p(x,t) \; . \tag{16}$$

Even if Δx is small, we may take n so large that

$$N_{[x, x+\Delta x]} \approx n\, \Delta x\, p(x,t) \; . \tag{17}$$

This is the number of particles (paths) multiplied by the probability that any given path is in $[x, x + \Delta x]$. The point is that you can observe the flow of probability by observing the flow of particles, if there are enough particles.

Brownian motion is a stochastic process that has no *jumps*. A Brownian motion particle cannot go from $x_{t_1} < a$ to $X_{t_2} > 0$ without crossing a at some intermediate time t. That is, there is at least one t_3 with $t_1 < t_3 < t_2$ so that $X_{t_3} = a$. This motivates that probability moves from one place to another by *flowing* but not by jumping. At any point a at any time t there is a *probability flux* or *probability current* $F(a,t)$ that tells you how much probability flows from $x < a$ to $x > a$ per unit time. This means, for example, that

$$\frac{d}{dt} \int_{-\infty}^{a} p(x,t)\, dx = -F(a,t) \; . \tag{18}$$

In terms of particles, F determines the net rate of particles “flowing” across the point a at time t. The particle motions themselves are random. If there are many particles near a, then many cross from $x < a$ to $x > a$ and many others cross the other way. The flux is determined by the difference between these numbers. If $F(a,t) > 0$, this means that more particles go from $x < a$ to $x > a$ in a small time interval $[t, t + \Delta t]$. That explains the minus sign in the flux formula (18). If $F(a,t) < 0$, then the derivative on the left of (18) is positive as the number of particles (or the probability) in $x < a$ is increasing. The quantity F is called *current* if you think of probability as something flowing, like electrical current.

Brownian motion, and all other diffusion processes governed by SDEs, have probability fluxes $F(a,t)$ that are determined by the particle distribution near

a at time t. Only particles near a at time t have a significant probability of crossing a in the time interval $[t, t + \Delta t]$. You might start by guessing that $F(a, t)$ is determined by $p(a, t)$. This guess is wrong for Brownian motion (but partly right for other diffusion processes), as you can see by asking what the flux would be if $p(x, t)$ were constant for x near our *control point*, a. If the particle distribution is the same on both sides of a, no matter what that density is, then the net flux should be zero. Brownian motion is symmetric. The probability of a particle at $a + \Delta x$ at time t having $X_{t+\Delta t} < a$ is the same as the probability of a particle $X_t = a - \Delta x$ moving to $X_{t+\Delta t} > a$.

It seems natural to guess that the particle flux is proportional to the gradient of the particle density at a. If $\partial_x p(a, t) > 0$, then the particle density is higher for $x > a$ than for $x < a$. That suggests that more particles go from $x > a$ to $x < a$ than cross the other way. This hypothesis is called *Fick's law* or the *Fourier law*. It says that there is a *diffusion coefficient*, called D, so that

$$F(a, t) = -D\, \partial_x p(a, t)\ . \tag{19}$$

The probability evolution equation (15) follows from the Fick's law flux model (19) combined with the local conservation of probability formula (18). If you put the t derivative inside the integral in (18), you get

$$\int_{-\infty}^{a} \partial_t p(x, t)\, dx = -F(a, t)\ .$$

Next, differentiate both sides with respect to a to find[4]

$$\partial_t p(a, t) = -\frac{d}{da} F(a, t)\ .$$

The derivative with respect to a on the right is the partial derivative of F with respect to its first argument, so it can be written as the derivative of F evaluated at $x = a$:

$$\partial_t p(a, t) = -\partial_x F(a, t)\ .$$

The final step, and the step that uses "physics" (properties of the Brownian motion diffusion process) is to differentiate Fick's law with respect to x, which leaves

$$\partial_x F(a, t) = -D\, \partial_x^2 p(a, t)\ .$$

The minus signs cancel when you put these together. The result is

$$\partial_t p(a, t) = D\, \partial_x^2 p(a, t)\ .$$

This is the forward equation (18), if the diffusion coefficient has the value $D = \frac{1}{2}$.

[4] The fundamental theorem of calculus says that if $u(x)$ is "any" integrand, then $\frac{d}{da} \int_{-\infty}^{a} u(x)\, dx = u(a)$.

Transition density, fundamental solution

Neither derivation of the forward equation (15) is completely convincing. The first relied on unjustified assertions relating to the generator and is adjoint. The second relied on “intuition” involving diffusion of Brownian motion particles and probability. Here is a derivation that is based on basic principles and calculations. It has the benefit of being clearly correct. The earlier incomplete derivations/justifications fill in intuition.

The transition density for any process represents the probability to go from y to x in time s. We write $G(x, y, s)$ for the probability density of X_{t+s}, at a point x, conditional on $X_t = y$. The integral formula (10) uses this definition and what is often called the “law of total probability”. The probability density to be at x at time $t + s$ is the “sum” (the integral) of all the ways X_{t+s} could get there, which is all the possible values X_t could have.

We now put this into a form where we can differentiate with respect to t and x and check directly that (15) is satisfied. Instead of starting with $X = y$ at time t and moving forward to time $t + s$, we start with $X = y$ at time 0 and move forward by a time t. The formula (10) becomes

$$p(x,t) = \int_{-\infty}^{\infty} G(x,y,t)\, p(y,0)\, dy \; . \tag{20}$$

It is clear in this form that differentiating $p(x, t)$ with respect to t of x results in differentiating $G(x, y, t)$ with respect to the same t or x. For this, we need a concrete formula for G. Brownian motion increments are mean zero Gaussians. Therefore, conditional on $X_0 = y$. we know X_t is Gaussian with mean y. The variance of the increment is equal to the time interval, which is t in this case. Therefore X_t is Gaussian with mean y and variance t. The density formula for this is

$$G(x,y,t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} \; . \tag{21}$$

We turn the idea (10) into an integral formula for $p(x, t)$.

The final check is “just” calculus. You have to be careful with signs to the the t derivative right:

$$\begin{aligned}
\partial_t \left[\frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{(x-y)^2}{2t}} \right] &= \frac{1}{\sqrt{2\pi}} \left[\left(\partial_t t^{-\frac{1}{2}} \right) e^{-\frac{(x-y)^2}{2t}} + t^{-\frac{1}{2}} e^{-\frac{(x-y)^2}{2t}} \left(\partial_t - \frac{(x-y)^2}{2t}, \right) \right] \\
&= \frac{1}{\sqrt{2\pi}} \left[-\frac{1}{2} t^{-\frac{3}{2}} e^{-\frac{(x-y)^2}{2t}} + t^{-\frac{1}{2}} e^{-\frac{(x-y)^2}{2t}} \frac{(x-y)^2}{2t^2} \right] \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{2} t^{-\frac{3}{2}} \left[\frac{(x-y)^2}{t} - 1 \right] e^{-\frac{(x-y)^2}{2t}}
\end{aligned}$$

$$\partial_t \left[\frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{(x-y)^2}{2t}} \right] = \frac{1}{2t} \left[\frac{(x-y)^2}{t} - 1 \right] G(x,y,t) \; . \tag{22}$$

For the x derivatives,

$$\begin{aligned}
\frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-y)^2}{2t}} &\xrightarrow{\partial_x} -\frac{x-y}{t}\frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-y)^2}{2t}} \\
&\xrightarrow{\partial_x} \frac{(x-y)^2}{t^2}\frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-y)^2}{2t}} - \frac{1}{t}\frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-y)^2}{2t}} \\
\partial_x^2 \frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-y)^2}{2t}} &= \frac{1}{t}\left[\frac{(x-y)^2}{t} - 1\right] G(x,y,t) \qquad (23)
\end{aligned}$$

When you compare (22) to (23), you see that

$$\partial_t G(x,y,t) = \frac{1}{2}\partial_x^2 G(x,y,t) \ . \qquad (24)$$

As explained, this shows that p in the integral representation (20) satisfies the forward equation (15).

People who solve PDEs call a function like G a *Green's function*, or the *fundamental solution*. The solution G is "fundamental" in that any other solution can be represented as an integral involving G. In probability, the Green's function has the interpretation as the transition density.

You can think of the integral (20) as an integral "operator" in that it "operates" on the function $p(\cdot,0)$ and produces a new function $p(\cdot,t)$. The G in the integral is the integral *kernel*. We have also seen differential operators, such as ∂_x and the generator $\mathcal{G}$. For Brownian motion, the kernel is the Gaussian transition density.

Smoothing, loss of information

The *integral representation* of $p(x,t)$ using the kernel (21) gives some information about how p "evolves" over time. The integral representation (20) represents $p(\cdot,t)$ as s sum (or *superposition*) of functions $G(x,y,t)$ for various values of y and "weights" $p(y,0)$. The kernel (21) is a smooth function of x, for any y. Therefore a weighted sum or integral of such functions also is a smooth function of x. This is true even if the starting distribution $p(y,0)$ is not a smooth function of y. Any non-smoothness (discontinuities, etc.) of the starting probability distribution $p(y,0)$ is lost at time $t > 0$. In this sense, the forward equation evolution loses information. It also suggests that $p(x,t)$ is a relatively simple function of x even if $p(y,0)$ is a complicated function of y.

The direction of time

Smoothing, or loss of information, is an inevitable consequence of "running" the forward equation forwards in time. Loss of information is the reason the forward equation cannot be run backwards in time. If you specify the function $p(y,0)$ and ask what is $p(x,t)$, you can find out using the forward equation dynamics – running it forwards in time. But if you specify $p(x,t)$, it is generally impossible to "run the forwards equation backwards" to find a corresponding function

$p(y,0)$. One way to see this is to specify that $p(x,t)$ is discontinuous. We know there is no corresponding starting probability $p(y,0)$ because the solution formula (20) gives a $p(x,t)$ that is not discontinuous.

But suppose there was a $p(y,0)$ which gave $p(x,t)$ via evolution operator (20). Even then it is generally impossible to reconstruct the starting distribution from the distribution at time t. This is because the backward evolution problem is *ill posed*[5] Arbitrarily small differences in the density $p(x,t)$ can correspond to arbitrarily large differences in the initial density $p(y,0)$. You can have two densities $p_1(x,t)$ and $p_2(x,t)$ that are within computer roundoff of each other but that correspond to very different initial densities $p_1(y,0)$ and $p_2(y,0)$ You know this because there are very different starting densities that give almost the same density at time t. That's the "loss of information". If you try to go backwards with the forward equation in the computer, things will go wrong for you. An example of this is coming with the backward equation.

An ODE (or a system of ODEs) can represent the dynamical evolution of a collection of n parameters. A PDE can represent the dynamical evolution of a function such as a probability density. A fundamental difference between ODE and PDE models is this – that a PDE model can have a distinguished direction of time. An ODE model, if it can be run forward, can be run backward.

The initial value problem

An *initial value problem*, when discussing quantities like probability distributions that change with time, means giving the quantity at some time and using the model to determine the quantity at later times. We are talking about the changing probability density $p(\cdot,t)$. The initial value problem is to compute $p(\cdot,t+s)$ from $p(\cdot,t)$ when $s>0$. This is to be done using the forward equation (9). The function $p(x,t)$, as a function of x, is the *initial condition* or *initial value*. The values of $p(x,t)$ are the *initial data*, or initial *values* or initial *conditions* for the initial value problem.

The integral representation (10) is one way to express the solution to the initial value problem. It is convenient to think of the initial time t as being $t=0$. That allows us to use t for the time variable in the forward equation. In practice it is not common that the Green's function (transition density) is known but the solution to the initial value problem is unknown. Computer methods such as finite difference methods are used to solve (approximately) the forward equation, once the initial conditions are given.

From the PDE point of view, the data initial value problem consists of a PDE such as the forward equation together with initial data and a starting time. The problem of the initial value problem ("task" would be a better term) is to find the solution for times.

[5] Here, "ill" means "bad" or "badly". "Posed" means "presented" you might "pose" a question or a problem to be solved. "Ill posed" means "badly presented".

The delta function, $\delta(x)$

The *delta function* is an "improper function" that is used to denote the PDF of a "random variable" that is not random. Specifically, if X is a one component random variable whose value is $X = 0$, then the PDF of X is written $\delta(x)$. The probability for $X \neq 0$ is zero, so $\delta(x) = 0$ when $x \neq 0$. On the other hand $\delta(x)$ represents probability, so

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1 \; .$$

The range of integration can be limited to any interval $[a, b]$ as long as $a < 0 < b$, because $\delta(x) = 0$ except when $x = 0$

$$\int_a^b \delta(x)\, dx = 1 \text{ if } \; a < 0 < b \; .$$

From a mathematical point of view, if δx were an actual function with $\delta(x) = 0$ for $x \neq 0$, then, no matter what value $\delta(0)$ has, the integral would be zero, not one. The delta function is an "improper" or "idealized" function rather than a function in the strict sense. There is a mathematical theory of *distributions*, not to be confused with probability distributions, where the delta function finds a mathematical definition. However, the theory of distributions is very technical.

Measure theory is a simpler setting for a mathematical definition of the delta function. A *measure* is a way to assign numbers to sets of numbers. Specifically, there are *probability measures* corresponding to random variables. If A is a set of numbers, then the corresponding probability measure P is

$$P(A) = \Pr(X \in A) \; .$$

A measure is a function, like $P(A)$, whose argument is a set, like A. The measure corresponding to the delta function is called the *point mass* or the *delta measure.* I will call it P_0 here, though it is written in many other ways. As a function, the definition of the "point mass at zero" measure is

$$P_0(X) = \begin{cases} 1 & \text{if } 0 \in A \; , \\ 0 & \text{if } 0 \notin A \; . \end{cases}$$

If we know that $X = 0$, then the probability of $X \in A$ is equal to zero if $0 \notin A$ and one if $0 \in A$.

The delta function is a convenient way to describe the point mass in integrals without invoking measure theory. If $f(x)$ is any function that is continuous at $x = 0$, then

$$\int_a^b f(x)\delta(x)\, dx = f(0) \; , \quad \text{if } a < 0 < b \; .$$

If the value of X is known to be some other value, such as $X = a$, then the PDF of X is $\delta(x - a)$, which you might see written as $\delta_a(x)$. If $f(x)$ is continuous at

$x = a$, then the function $f(x-a)$ is continuous at $x = 0$. Therefore, using a change of variables in the integral,

$$\int f(x)\delta(x-a)\,dx = \int f(x-a)\delta(x-a)\,dx = f(a)\ .$$

This corresponds to P_a, which is the point mass probability measure with "mass 1" at $x = a$.

The delta function is convenient for expressions involving diffusions with a known starting point, and for interpreting integral expressions such as the transition density representation integral (20). Suppose X_t is a Brownian motion starting from $X_0 = 0$. We saw that the probability density for X_t is

$$p(x,t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}\ .$$

When t is small (but positive), this probability density its mass closely concentrated around $x = 0$. In the $t \downarrow 0$ limit, this distribution converges to the delta function, at least in the sense of integrals

$$\lim_{t\downarrow 0} \int_{-\infty}^{\infty} f(x)p(x,t)\,dx = f(0) = \int_{-\infty}^{\infty} f(x)\delta(x)\,dx\ . \tag{25}$$

In this sense

$$\frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \ \to\ \delta(x)\ , \quad \text{as } t \to 0\ .$$

This convergence is not convergence of the values $p(x,t)$ to $\delta(x)$ as $t \to 0$. The number $p(0,t) = \frac{1}{\sqrt{2\pi t}}$ do not converge to $\delta(x)$ as $t \to 0$. The numbers $p(0,t)$ do not converge at all and $\delta(0)$ is not really defined. Nevertheless there is convergence "in the sense of distributions" or in the sense of "weak convergence" of probability measures. This means that the limit formula (25) is true whenever f is a suitable "test function". In probability, "suitable" often means "continuous and bounded". For convergence in distributions (the technical theory of distributions), "suitable" is more technical. The delta function is often drawn as a tall and narrow Gaussian.

The delta function is convenient for expressing initial conditions for the initial value problem for $p(x,t)$ if $X_0 = a$ is known. We write $p(x,0) = \delta(x-a)$. What a mathematician means by this is (25). The transition density is a solution to the initial value problem with delta function initial data.

3 Boundary conditions, hitting times

There are things about a diffusion process that you can want to know besides the evolution of probability densities and value functions. A *hitting time* is the first time X_t touches some set. By convention, we say that hitting time is infinite if the hitting event never happens. Hitting times are examples of *path dependent* functions of a process, which means functions of the process that depend on the who path rather than the value at some specific time.

We will study the specific hitting time for Brownian motion. It is traditional to use τ for a hitting time, with a subscript to indicate the specifics. Suppose $X_0 = a > 0$. The first hitting time at zero is

$$\tau = \min \{ \, t \mid X_t = 0 \, \} \; .$$

It is random. We will see that this specific hitting time has that $\tau < \infty$ almost surely, and yet $\mathrm{E}[\tau] = \infty$. The stopping time also defines the stopped process, which is

$$\widetilde{X}_t = \begin{cases} X_t & \text{if } t \leq \tau \\ X_\tau & \text{if } t \geq \tau \end{cases} \; .$$

The stopped process is the original process until the *stopping time* τ, after which it doesn't move. This may be written with "wedge" notation in which the wedge of two times is the smaller one

$$s \wedge t = \min(s,t) \; .$$

The stopped process is

$$\widetilde{X}_t = X_{t \wedge \tau} \; . \tag{26}$$

The stopped process is natural, for example, if you're modeling the diffusion of something that sticks when it touches a boundary.

Brownian motion has the property of being a *martingale*. This means that if $s < t$, then the expected value at time t is the known value at time s

$$\mathrm{E}[\, X_t \mid X_s = x] = x \; .$$

This is another way to say that the increment of Brownian motion is independent of the past and has mean zero. The stopped process (26) is also a martingale. This is an example of the important fact called *Doob's theorem*, which says that any stopped martingale is a martingale.

We want to calculate the PDF of the Brownian motion stopped at $x = 0$. This PDF has a concentration at $x = 0$ that represents the probability that $\tau \leq t$. It also has a regular density $p(x,t)$ defined for $x > 0$ that describes Brownian motion particles that have $\tau > t$. There are explicit formulas for all these quantities, which we find using the diffusion equation (15) together with a *boundary condition* for p that applies at the hitting boundary $x = 0$. The PDE plus boundary condition problem can be solved using a special trick called the *method of images*. The formulas confirm the probability flux picture. The *survival probability* os

$$\Pr(\tau \geq t) = \int_0^\infty p(x,t)\, dx \; .$$

We will see explicitly that the time derivative of the survival probability is given by the probability flux at the boundary $x = 0$.

You can argue that $p(x,t)$, which is the PDF or stopped Brownian motion with $x > 0$, satisfies the diffusion equation (15) by arguing that the representation formula (20) is approximately true if $x > 0$ and t is small enough. For any fixed x, if t is small enough there is very little chance to hit the $x = 0$ before time t. The formula (20) should apply to Brownian motion paths that do not hit the boundary before time t. It is possible to work this idea out in technical detail, but that would not be in the spirit of this class.

More subtle is the boundary condition

$$p(x,t) \to 0 \ , \quad \text{as } x \to 0 \ .$$

This is often written informally as a boundary condition that applies at the $x = 0$ boundary.

$$p(0,t) = 0 \ . \tag{27}$$

The boundary condition says that the probability density at x is small for x close to the boundary. This is because of the wildness of Brownian motion paths. If $X_t = x$, then it is likely that there is $s < t$ with $X_s < t$. Brownian motion paths go back and forth, so the particles near the boundary, most of them anyway, have touched the boundary at some earlier time. This argument can be justified, but at the cost of time this course does not have. People often give the argument that $p(0,t)$ is about survival probability at $x = 0$ at time t, which is zero because you're talking about $X_t = 0$. This argument sounds good, but we will see that it's wrong.

The method of images, reflection principle

The *method of images* is a clever way to find a formula for the density of surviving Brownian motion particles just described. Suppose X_t is a Brownian motion starting at $X_0 = a$ with $a > 0$, that is "absorbed" the first time it touches the boundary $x = 0$. This satisfies the following *initial boundary value problem* involving a PDE, initial conditions, and boundary conditions. The PDE is the forward equation for Brownian motion (15). The initial condition is $p(x,0) = \delta(x-a)$. The boundary condition is the *absorbing* boundary condition (27). The PDF $p(x,t)$ is defined for $t \geq 0$ and $x \geq 0$. The "interior" of the domain where p is defined is $x > 0$ and $t > 0$, and p is supposed to satisfy the PDE at every interior (x,t).

The *method of images* is a trick for finding functions that satisfy the absorbing boundary condition "by symmetry". The trick (the "method") is to "extend" the definition of p to points $x < 0$ "by symmetry" so that p becomes *odd* (or *skew symmetric* or *anti-symmetric*) in the sense that

$$p(-x,t) = -p(x,t) \ . \tag{28}$$

This is done by extending the initial values $p(x,t)$ to $x < 0$ by skew symmetry

$$p(x,0) = \delta(x-a) - \delta(x+a) \ . \tag{29}$$

The first term on the right is the original initial condition, which is a point mass at $x = a$. The second term term on the right is a negative point mass at $x = -a$. If you think of $\delta(x - a)$ as a narrow gaussian with unit area centered at $x = a$, then $-\delta(x + a)$ is the negative of a unit area Gaussian centered at $x = -a$. The negative mass at $-a$ is the *image* mass, after the reflection.

We now define $p(x, t)$ for all x and all $t \geq 0$ to be the solution of the pure initial value problem for the forward equation (15) and initial condition (28). The solution, like any other pure initial value solution, is given by the Green's function integral (20) with Gaussian Green's function (21). The solution is

$$p(x,t) = \frac{1}{\sqrt{2\pi t}} \left[e^{-\frac{(x-a)^2}{2t}} - e^{-\frac{(x+a)^2}{2t}} \right] . \qquad (30)$$

The positive term in brackets $[\cdots]$ is from the original delta mass corresponding to $X_0 = a$. The negative term is from the negative "image" at $x = -a$.

You can check that the method of images solution formula (30) satisfies all the requirements of the original initial boundary value problem, which are supposed to be satisfied for $x \geq 0$ and $t \geq 0$ only. It satisfies the initial condition $p(x, 0) = \delta(x - a)$ for $x \geq 0$ because the image $\delta(x + a)$ is equal to zero when $x \geq 0$. The image $-a$ is outside the x domain $x \geq 0$. It satisfies the PDE (15) for all x as long as $t > 0$. It satisfies the absorbing boundary condition (27) by symmetry. If $p(-x, t) = -p(x, t)$, then $p(-0, t) = -p(0, t)$, and a continuous function can only do that if $p(0, t) = 0$. If you don't believe that, you can put $x = 0$ into the right side of (30) and see that you get zero.

In PDE, this trick is sometimes called the *reflection principle*. You take as initial conditions the original conditions, minus the reflected conditions. Then the solution will be odd, by reflection symmetry, and therefore satisfy the absorbing boundary conditions. In probability, the term *reflection principle* is often used for a related fact ....

In probability, the *reflection principle* (or *Kolmogorov* reflection principle) is the fact that

$$\Pr(\, X_s \leq 0 \text{ for some } s \leq t\,) = 2 \Pr(\, X_t \leq 0\,) \ . \qquad (31)$$

This has a natural interpretation related to the symmetry of the random Brownian motion. If X_s "touches" the absorbing boundary at some time $s \leq t$, then $\tau \leq t$, where τ is the first hitting time described above. But Brownian motion is a Markov process, so the probability distribution of X_t with $t \geq \tau$ is symmetric about $x = 0$. This means that half of the Brownian motion particles that touch $x = 0$ before time t have $X_t > 0$ and the other half have $X_t < 0$. In case this seems too vague or unconvincing, you can check the reflection principle formula (31) directly by integration. The probability on the left is

$$\Pr(\, \tau \geq t\,) = 1 - \int_0^\infty p(x,t)\, dx \ .$$

On the other hand, we know that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-a)^2}{2t}}\, dx = 1 \ .$$

Therefore,

$$1 - \int_0^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-a)^2}{2t}}\, dx = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-a)^2}{2t}}\, dx$$

This is $\Pr(X_t \leq 0$. It is also equal to the contribution from the reflection part of (30), which is

$$\int_0^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x+a)^2}{2t}}\, dx$$

These identical integrals give the "2" on the right side of the reflection principle formula (31).