Stochastic Calculus, Courant Institute, Fall 2019
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2019/StochCalc.html
Jonathan Goodman, September, 2019

Class 3, Backward equations

1 Introduction

Backward equations are partial differential equations (PDEs) that are satisfied by conditional expectations. These conditional expectations may not be the expected values you're most interested in, but (1) they may determine the expectation you actually want, and (2) they may satisfy a partial differential equation (backward equations) that can solved. The relationship between diffusion processes and PDEs goes both ways. Sometimes a PDE is solved by simulating a related diffusion process, sometimes expectations of a diffusion process are found by solving a PDE.

As a simple example, suppose X_t is a Brownian motion path that will wander until time T. Then there will be a "payout" $v(X_t)$. The payout function is $v(x)$. The actual payout depends on the location of the Brownian motion at time T. The expected payout is

$$\mathrm{E}[\, v(X_T)] \ .$$

You can imagine that X_t is the score of a game and that your team wins if $X_T > 0$, and you want to know the probability that this happens. This fits into our framework by taking the payout to be the Heavyside function $v(x) = 1$ if $x > 0$ and $v(x) = 0$ if $x < 0$.

The expected value depends on the starting point of the Brownian motion, x_0. The probability of winning a game depends on the starting score. There are several notations for the dependence, including

$$f(x_0, 0) = \mathrm{E}_{x_0}[\, v(X_T)] = \mathrm{E}[\, v(X_T \mid X_0 = x_0)] \ .$$

The subscript x_0 in the first expectation means that the expectation is taken using the probability distribution of Brownian motion paths that start at x_0. We have used that kind of notation before, putting a subscript on the expectation to say which probability distribution is assumed. The expectation $E[\cdot|\cdot]$ is for conditional expectation. It represents the expected value of the first expression conditional on the second one. Here, it is the expected value of $v(X_T)$ conditional on $X_0 = x_0$.

The probability of winning a game can change while the game is in progress. Suppose $X_t = x$ for some $t < T$. The conditional expectation then is written as

$$f(x, t) = \mathrm{E}_{x,t}[\, v(X_T)] = \mathrm{E}[\, v(X_T) \mid X_t = x)] \ . \tag{1}$$

In sports betting, it may be possible to place bets during a game. The odds offered depend on the progress of the game up to that point. This applies to

general payout functions (not just the Heavyside function) and the corresponding conditional expectations (1). The conditional expectation (1) is the *value function*.

For Brownian motion, the value function satisfies a backward equation, which is a PDE related to the heat equation from last week. There are backward equations for other diffusion processes, which feature a partial differential operator (∂_x or ∂_x^2 are simple partial differential operators) called the *generator* of the process. If g is a function of x, we write the generator applied to g as Lg. The backward equation for the value function (1) is

$$\partial_t f + Lf = 0 \; . \tag{2}$$

The generator is defined by

$$(Lg)(x) = \lim_{t \downarrow 0} \frac{\mathrm{E}_x[\, g(X_t)] - g(x)}{t} \; . \tag{3}$$

To explain the left side, L is a linear operator. When L "operates" on a function g, it produces another function, which is Lg. The left side is the value of the function Lg at the point x. The expectation on the right is the expected value of $g(X_t)$ under the condition that $X_0 = x$. Since t is small, X_t will be close to x and $g(X_t)$ will be close to $g(x)$. If the limit exists, X_t is so close to x that the expected value differs by $O(t)$. The generator is a differential operator because $g(X_t) = g(x)$ may be approximated using a Taylor series around $X_t = x$ that uses only a few terms.

Backward equations "work" for diffusion processes because diffusions are Markov processes. A random process is a Markov process if "the future depends on the past only through the present". Diffusion processes are Markov processes, and there are others. This class will explain the Markov property, calculate generators, explain backward equations, and give some examples of how they can be used.

2 The Markov property and joint densities

To make that more clear, consider the joint probability density for values of a process X_t at various times. The *single time* PDF is $p(x,t)$, which is the PDF of X_t. We write the *two time* PDF as $p(x_1, x_2, t_1, t_2)$. This is the joint density for the two component random variable (X_{t_1}, X_{t_2}). For example

$$\mathrm{E}\big[\, (X_{t_2} - X_{t_1})^2\big] = \int\int (x_2 - x_1)^2 p(x_1, x_2, t_1, t_2)\, dx_1 dx_2 \; .$$

The two-time PDF is a probability density in the first two variables, so, for example,

$$\int\int p(x_1, x_2, t_1, t_2)\, dx_1 dx_2 = 1 \; .$$

The one-time density is the marginal of the two-time density. You can "integrate out" either x_1 or x_2, and you can do it at any time $t_1 \neq t$ or $t_2 \neq t$, so

$$p(x,t) = \int p(x, x_2, t, t_2)\, dx_2 = \int p(x_1, x, t_1, t)\, dx_1 \ .$$

These are basic facts about probability densities.

Some conditional probability densities are *normalized* versions of joint densities. To start, suppose (X, Y) is a two component random variable with PDF $p(x, y)$. The conditional density of X, conditional on knowing $Y = y$, may be written

$$p(x \mid Y = y) = \frac{1}{Z(y)} p(x, y) \ .$$

[This notation, using p for every probability density, is becoming common again in Bayesian statistics and machine learning.] The normalization constant may be found by requiring that the conditional density $p(x|y)$ should integrate to 1 in the x variable

$$\int p(x \mid Y = y) = \frac{1}{Z(y)} \int p(x, y)\, dx = 1 \implies Z(y) = \int p(x, y)\, dx \ . \tag{4}$$

The normalization constant usually depends on y, but it doesn't have to.

Suppose $t_2 > t_1$, which we didn't assume until now. The conditional density of X_{t_2}, conditional on $X_{t_1} = y$ was called, and still will be called, the *transition density* and written

$$G(y, x, t_1, t_2) = p(X_{t_2} = x \mid X_{t_1} = y) = \frac{1}{Z(y)} p(y, x, t_1, t_2) \ .$$

For standard Brownian motion without drift, this was (here y is a parameter and x is the variable)

$$G(y, x, t_1, t_2) = \mathcal{N}(y, (t_2 - t_1)) = \frac{1}{\sqrt{2\pi(t_2 - t_1)}} e^{-\frac{(x-y)^2}{2(t_2 - t_1)}} \ .$$

A Brownian motion with drift rate a has

$$G(y, x, t_1, t_2) = \mathcal{N}(y + a(t_2 - t_1), (t_2 - t_1)) = \frac{1}{\sqrt{2\pi(t_2 - t_1)}} e^{-\frac{(x-y-a(t_2-t_1))^2}{2(t_2 - t_1)}} \ .$$

A diffusion process is determined by its transition density.

Some formulas above, particularly the normalization constant formula (4) may make it seem that conditional densities are complicated. On the contrary, transition densities and other conditional probability densities are often simpler than joint densities. For example, suppose X_t is a standard Brownian motion starting at $X_0 = 0$. The joint density for $X_{t_1} = y$ and $X_{t_2} = x$ is

$$p(x, y, t_1, t_2) = \frac{1}{2\pi\sqrt{t_1(t_2 - t_1)}} e^{-\frac{y^2}{2t_1}} e^{-\frac{(x-y)^2}{2(t_2 - t_1}} \ .$$

The conditional density of x given y is

$$p(x|y,t_1,t_2) = \frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{(x-y)^2}{2(t_2-t_1)}} \ .$$

The second formula is simpler. You can get it from the first formula by integrating over y.

The Markov property concerns the joint distribution for three times, which may be called $t_1 < t_2 < t_3$. Think of t_1 as the past, t_2 as the present, and t_3 as the future. The Markov property from the introduction is that the distribution of X_{t_3} conditional on X_{t_1} and X_{t_2} is independent of X_{t_1}. This is what it means to say that the future (the distribution of X_{t_3} is independent of the past (the known value of X_{t_1}), once you know the present (the value of X_{t_2}). In terms of conditional probability densities, this is, for all x_1,

$$p(x_3 \mid x_1, x_2, t_1, t_2, t_3) = p(x_3 \mid x_2, t_2, t_3) \ . \tag{5}$$

The conditional density on the left side involves three times and three locations (t_1, t_2, t_3, $X_{t_1} = x_1$, etc.). The conditional density on the right side involves just two times $t_2 < t_3$ and the corresponding locations $X_{t_2} = x_2$, and $X_{t_3} = x_3$. The Markov property applies to conditioning on more than one past time. If the Markov property (5) is satisfied, then conditioning on more past information does not change the conditional density in the future. You can prove this using basic properties of conditional probability, or you can just believe me that it's true.

The value function (1) is particularly important if X_t satisfies the Markov property. The Markov property says that the value function at t_2 determines the value function at t_1 if $t_1 < t_2$. The first two quantities are equal because of the definition (1). The second and third are equal because of the Markov property.

$$f(x_1,t_1) = \mathrm{E}[\, v(X_T) \mid X_{t_1} = x_1)] = \mathrm{E}[\, f(X_{t_2}) \mid X_{t_1} = x_1)] \ . \tag{6}$$

We will use this formula every class from now on, probably many times per class.

This formula is an example of the *tower property*, which says that the expected value of the expected value is the expected value. Abstractly, suppose (Y, Z) is a two component random variable, $u(y, z)$ is some "payout function", and the overall expected value is

$$f = \mathrm{E}[\, u(Y,Z)] \ .$$

Define a "value function" to be the expected value of u, conditioned on $Y = y$, which is

$$f(y) = \mathrm{E}[\, u(Y,Z) \mid Y = y] \ .$$

The tower property says that the expected value of the conditional expectation is the overall expectation,

$$f = \mathrm{E}[\, f(Y)] \ .$$

You can prove this using the properties of conditional expectation and marginal probability.

We can apply this to derive our value function formula (6) by taking $Z = X_{t_3}$, and $Y = X_{t_2}$. The value $X_{t_1} = x_1$ is just a parameter. Without the Markov property, the conditional expectation that is called $f(y)$ in the general theory and $f(x_2, t_2)$ in our desired formula (6) would depend on x_1. But the Markov property says that the distribution of X_T conditioned on $X_{t_2} = x_2$ and $X_{t_1} = x_1$ is independent of x_1. Therefore we get jus $f(x_2, t_2)$ with no x_1 dependence.

3 Backward equation and generator

Much of stochastic calculus relies on being able to compute the generator of a process and understanding the relation between the generator and the backward equation. Here, we first figure out the generator for Brownian motion, then we use it to get the backward equation for Brownian motion, then we go through these steps again more generally. You will see a technical theme that is always important in analysis of diffusion processes. Suppose you have a small Δt and the corresponding change in the process is $\Delta X = X_{t+\Delta t} - X_t$. Suppose you want to compute the corresponding change $\Delta f = f(X + \Delta X, t + \Delta t) - f(X, t)$. You have to do Taylor expansion up to second order in x and first order in t. This is because ΔX is on the order of $\sqrt{\Delta t}$ so $(\Delta X)^2$ is on the same order as Δt.

Consider the generator formula (3) in the case where X_t is standard Brownian motion. Then $X_t = x + Z$, where $Z \sim \mathcal{N}(0, t)$. For small t, Z is small and we can use Taylor series

$$g(x + Z) = g(x) + \partial_x g(x) Z + \tfrac{1}{2} \partial_x^2 g(x) Z^2 + O(|Z|^3) \; .$$

Here, x and g are not random. Only Z is random. Therefore (Pay attention. We will do calculations like this a lot, but not with such small steps.)

$$\begin{aligned}
\mathrm{E}[g(X_t)] &= \mathrm{E}[g(x + Z)] \\
&= \mathrm{E}[g(x) + \partial_x g(x) Z + \tfrac{1}{2} \partial_x^2 g(x) Z^2 + O(|Z|^3)] \\
&= \mathrm{E}[g(x)] + \mathrm{E}[\partial_x g(x) Z] + \mathrm{E}[\tfrac{1}{2} \partial_x^2 g(x) Z^2] + \mathrm{E}[O(|Z|^3)] \\
&= g(x) + \partial_x g(x) \mathrm{E}[Z] + \tfrac{1}{2} \partial_x^2 g(x) \mathrm{E}[Z^2] + \mathrm{E}[O(|Z|^3)] \\
&= g(x) \; + \; \partial_x g(x) \cdot 0 \; + \; \tfrac{1}{2} \partial_x^2 g(x) \cdot t \; + \; O(t^{\frac{3}{2}}) \\
&= g(x) + \tfrac{1}{2} \partial_x^2 g(x) t + (t^{\frac{3}{2}}) \; .
\end{aligned}$$

The result, putting this back into the limit (3), is

$$Lg(x) = \frac{1}{2} \partial_x^2 g(x) \; . \tag{7}$$

The generator of Brownian motion is the second derivative operator (half of that).

The backward equation (2) is derived from the limit formula (3) using the tower property (6) in a way we will use again and again. We first put in a small Δt and then take the limit $\Delta t \to 0$. We use notation from before, $X_t = x + \Delta X$. We assume that $\mathrm{E}[\,|\Delta X^p|\,] = O(\Delta t^{\frac{p}{2}})$. This is true for Brownian motion (we have already seen). We will see later why it is true for other diffusion processes.

$$\begin{aligned} f(x,t) &= \mathrm{E}[\,f(x+\Delta X, t+\Delta t) \mid X_t = x] \\ 0 &= \mathrm{E}[\,f(x+\Delta X, t+\Delta t) - f(x,t) \mid X_t = x] \\ 0 &= \mathrm{E}[\,f(x+\Delta X, t+\Delta t) - f(x+\Delta X, t) \mid X_t = x] \\ &\quad + \mathrm{E}[\,f(x+\Delta X, t) - f(x,t) \mid X_t = x]\ . \end{aligned}$$

To proceed, divide both sides by Δt and figure out what happens as $\Delta t \downarrow 0$. First,

$$\frac{f(x+\Delta X, t+\Delta t) - f(x+\Delta X, t)}{\Delta t} = \partial_t f(x+\Delta X, t) + O(\Delta t)\ .$$

When $\Delta t \downarrow 0$, the first term on the right converges to $\partial_t f(x,t)$ and the second term converges to zero. Second, apply the definition (3) to the function $g(x) = f(x,t)$, and with Δt (here) for t (there), and you get

$$\lim_{\Delta t \downarrow 0} \frac{f(x+\Delta X, t) - f(x,t)}{\Delta t} = Lf(x,t)\ .$$

Together, we get the backward equation (2).

Here's a Brownian motion example. Suppose the payout function is $v(x) = x^2$. Suppose $t < T$. If $X_t = x$, then $X_T = x + Z$, where $Z \sim \mathcal{N}(0, T-t)$. Therefore,

$$f(x,t) = \mathrm{E}\big[\,(x+Z)^2\big] = x^2 + (T-t)\ .$$

The backward equation for Brownian motion, checking (7), is

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

In our case, $\partial_t f = \partial_t[x^2 + (T-t)] = -1$ and $\frac{1}{2}\partial_x^2[x^2 + (T-t)] = 1$, so the equation is satisfied.

If you know the value function $f(\cdot, t)$, then the backward equation determines f at earlier times $t' < t$. The "evolution" specified by the backward equation goes backward in time. One way to see this is the tower property formula (6), which gives a formula for $f(\cdot, t')$ in terms of $f(\cdot, t)$. You may not be able to calculate the expectation, but it does exist. Another way to see it will be the finite difference method in Assignment 3. There, you start with the final condition $f(x,T) = v(x)$ and work backwards in small time steps until you get to $t = 0$ (or whatever t is desired).

The generator of any diffusion process is determined by its infinitesimal mean and infinitesimal variance. We keep using the notation $\Delta X = X_{t+\Delta t} - X_t$. The infinitesimal mean will be called $a(x)$ and is defined by

$$\mathrm{E}[\,\Delta X] = a(x)\Delta t + O(\Delta t^2)\ .$$

The infinitesimal variance will be called $\mu(x)$ and is defined by

$$\mathrm{var}(\Delta X) = \mu(x)\Delta t + O(\Delta t^2)\ .$$

This definition is equivalent to

$$\mathrm{E}\big[\,(\Delta X)^2\big] = \mu(x)\Delta t + O(\Delta t^2)\ .$$

The difference between the variance and the expected square (the two definitions) is that the variance is the expected square, with the mean subtracted out. If Y is any random variable, then $\mathrm{var}(Y) = \mathrm{E}\big[\,Y^2\big] - \big(\mathrm{E}[Y]\big)^2$. Apply this with $Y = \Delta X$ and the approximate formula $\mathrm{E}[\,\Delta X] \approx a(x)\Delta t$, and you get

$$\mathrm{var}(\Delta X) = \mathrm{E}\big[\,(\Delta X)^2\big] - a(x)^2\Delta t^2\ .$$

This shows that the two definitions of μ differ by order Δt^2, which makes them equivalent.

The general generator calculation is almost the same as the one for Brownian motion. You use a Taylor expansion up to second order with an error term of order $|\Delta X|^3$. The difference is that you then have to evaluate the limit using infinitesimal mean and variance. If you look back, you will see that we did exactly this in the Brownian motion case. We knew that $a(x) = 0$ and $\mu(x) = 1$. The calculation is (use the definition (3) with Δt instead of t)

$$\begin{aligned}
Lg(x) &= \lim_{\Delta t\downarrow 0} \frac{\mathrm{E}[\,g(x+\Delta X)] - g(x)}{\Delta t} \\
&= \lim_{\Delta t\downarrow 0} \frac{\mathrm{E}\big[\,\partial_x g(x)\Delta X + \frac{1}{2}\partial_x^2 g(x)(\Delta X)^2 + O(|\Delta X|^3)\big]}{\Delta t} \\
&= \partial_x g(x) \lim_{\Delta t\downarrow 0} \frac{\mathrm{E}[\Delta X]}{\Delta t} + \frac{1}{2}\partial_x^2 g(x) \lim_{\Delta t\downarrow 0} \frac{\mathrm{E}\big[\,(\Delta X)^2\big]}{\Delta t} + \lim_{\Delta t\downarrow 0} \frac{O(\Delta t)^{3/2}}{\Delta t} \\
Lg(x) &= a(x)\partial_x g(x) + \frac{1}{2}\mu(x)\partial_x^2 g(x)\ . \qquad (8)
\end{aligned}$$

Part of this was the claim that $\mathrm{E}\big[\,(\Delta X)^3\big] = O(\Delta t^{\frac{3}{2}})$. We verified that this is true for Brownian motion. It "should be true" here too for the same reason, if the expected value of $(\Delta X)^2$ is on the order of Δt, then typical values of ΔX will be of the order of $\Delta t^{\frac{1}{2}}$, so typical values of $(\Delta X)^3$ will be of order $\Delta t^{\frac{3}{2}}$. This suggests that the expected value of $(\Delta X)^3$ is of the order of $\Delta t^{\frac{3}{2}}$. We will see in a later class that this argument can go wrong if ΔX has *fat tails*. We will also see that ΔX from a diffusion processes does not have fat tails. Look for this when you review the class at the end.

4 Finite difference solution

The method of *finite differences* is a numerical method for solving solve partial differential equations that has nothing to do with probability or simulation It has the advantage over simulation that there is no statistical error. The computing exercises in Assignment 1 and Assignment 2 both had statistical error coming from averaging many simulations. You probably noticed how long those calculations could take to get high accuracy. Finite differences can be faster and more accurate.

The backward equation, which is what we want to solve, is dynamical. It describes how the value function changes as you move further backward in time from the final time T. The $\partial_t f$ in the backward equation indicates that it is about dynamics. A class on ordinary differential equations offers methods such as the Euler method (see below) for computing (or simulating) dynamical systems. The partial differential equations of interest here have two related complicating features you don't see in a (typical) ODE class. One complicating feature is that the "state" at time t, from the point of view of the backward equation, is the function $f(\cdot, t)$. A function is described by infinitely many numbers (the values at infinitely many points, or fancier representations you might have seen such as Fourier series). Therefore, we must have an approximation to $f(\cdot, t)$ that involves only finitely many numbers. The other complicating feature is the "space derivatives" such as ∂_x^2. Since we don't know $f(\cdot, t)$ exactly (having approximated the function using finitely many numbers) we need a way to estimate the space derivatives using these numbers. The subtle thing is that differential operators are *unbounded*. The derivative of a function can be much larger than the function itself. For example, if $g(x) = \sin(kx)$ then $\partial_x g(x) = k\cos(kx)$, which is larger by a factor of k. Even worse, if $g(x)$ is discontinuous then $\partial_x g(x)$ either does not exist or is infinite, depending on your point of view. For this reason, some finite difference approximations that seem to make sense don't work because they are *unstable*. The computing exercise (task 4) has an example of this. The issue of stability is too technical and complex to be part of this course – take Numerical Methods II if you're interested.

Time stepping is an approach to computing solutions that change in time. To start, consider the ODE $\partial_t y(t) = F(y(t))$. Choose a small time step Δt and define the discrete times to be $t_k = k\Delta t$. Then $t_{k+1} = t_k + \Delta t$. The differential equation implies that

$$y(t + \Delta t) \approx y(t) + \Delta t \partial_t y(t) = y(t) + \Delta t F(y(t)) \ .$$

A finite difference strategy would be to define approximations $y_k \approx y(t_k)$ by

$$y_{k+1} = y_k + \Delta t F(y_k) \ .$$

This is the *forward Euler* method for approximate simulation of dynamical systems.

For the backward equation, of course, time goes in the other direction. More importantly, instead of a simple number or vector $y(t)$ we have a function $f(\cdot, t)$.

The computer cannot store $f(\cdot, t)$, so it stores an approximation. For our finite difference method, that approximation consists of the values of f at some *grid points* called x_j. Suppose the equation is solved on the interval $x_l \le x \le x_r$. Then the function $f(\cdot, t)$ (as a function of x), is represented (approximately) by its values $f_j(t) \approx f(x_j, t)$. We will use a *uniform grid* with *grid spacing* Δx, so $x_j = x_l + j\Delta x$. We choose Δx so that there are n grid points in $[x_l, x_r]$. This implies that $\Delta x = (x_r - x_l)/(n+1)$. The $(n+1)$ is because there are $n+1$ grid intervals of length Δx. The first is $[x_l, x_1]$, (in LaTeX, the letter l and the number 1 look almost the same). The last, which is interval (n+1), is $[x_n, x_r]$. The computer stores the estimates $f_j(t)$ only at discrete times t_k, as explained above for the ODE. We write f_{kj} for the approximation of $f(x_j, t_k)$. The n component vector $(f_{k1}, \ldots, f_{kn})$ will be called f_k. The "forward Euler" approximation for the backward equation will be to estimate f_{k+1} from f_k. This is a *time step*.

To do the time step, we need estimates of $\partial x f$ and $\partial_x^2 f$. This estimate has to use only the numbers f_{kj}. It seems natural to try the estimates based on (see Assignment 3 for more details)

$$\partial_x f(x,t) \approx \frac{f(x+\Delta x, t) - f(x-\Delta x, t)}{2\Delta x}$$
$$\partial_x^2 f(x,t) \approx \frac{f(x+\Delta x, t) - 2f(x,t) + f(x-\Delta x, t)}{\Delta x^2} \ .$$

If we take x to be grid point x_j, then $x - \Delta x$ is the grid point x_{j-1} and $x + \Delta x$ is x_{j+1}. Therefore, it seems reasonable to try

$$\partial_x f(x_j, t_k) \approx \frac{f_{k,j+1} - f_{k,j-1}}{2\Delta x}$$
$$\partial_x^2 f(x,t) \approx \frac{f_{k,j+1} - 2f_{kj} + f_{k,j-1}}{\Delta x^2} \ .$$

Using these approximations, we can estimate $\partial_t f(x_j, t_k)$ and use this to estimate $f(\cdot, t_{k+1})$.