Stochastic Calculus, Courant Institute, Fall 2018
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2018/index.html
Jonathan Goodman, September, 2018

Lesson 2, Diffusion processes

1 Introduction

A *diffusion process* is a kind of random process. We let X_t be the value of the process at time t. Most interesting diffusions have more than one component: $X_t = (X_{1,t}, \ldots, X_{d,t}) \in \mathbb{R}^d$. You can picture X_t as the location of a particle moving at random, but not completely randomly. The motion is governed by a deterministic component called *drift* or *infinitesimal mean*, and by a random *noise*. We specify a diffusion process by giving the drift and the noise level, both as a function of X_t. The model is a *stochastic differential equation*, or *SDE*. It becomes an ordinary differential equation (ODE) if the noise coefficient is zero. For an *Ito* type SDE, X_t is a *martingale* if the drift is zero.

This lesson discusses only the case $d = 1$, so X_t is just a number. A process is a diffusion (diffusion process) if it has these properties

1. The time variable t is continuous. For convenience we often imagine that the process starts at time $t = 0$ and is defined for every (real number) $t > 0$.

2. It is a *Markov process*. This means that X_{t_0} determines the distribution of X_t for $t > t_0$ "completely". The Markov property is explained in more detail below.

3. X_t is a continuous function of t. There are no "jumps".

Many random processes are modeled as diffusions, either exactly or approximately.

The Markov property has to do with conditional expectation and conditional probability of the future of a path given its present and its past. The path over an interval $[t_1, t_2]$ (with $t_2 > t_1$) is denoted by $X_{[t_1,t_2]}$. At time t_1, the *past* is the time interval $[0, t_1]$. The "information" from the past consists of the path $X_{[0,t_1]}$. The information of the present is the single value X_{t_1}. The Markov property concerns the path in the *future* of t_1. It is that the distribution of $X_{[t_1,t_2]}$ conditional on the past is the same as the distribution conditional on the present. The distribution conditional on $X_{[0,t_1]}$ is the same as the distribution conditional on X_{t_1}.

You specify a Markov process by giving its transition probability distributions. Suppose $t_2 > t_1$ and X_t is a Markov process. Write Y for X_{t_2} and X for X_{t_1}. The probability density for X_{t_2} conditional on X_{t_1} is $G(y, t_2, x, t_1)$. This is a probability density in the y variable in the sense that

$$\Pr(y \leq X_{t_2} \leq y + dy) = G(y, t_2, X_{t_1}, t_1)\, dy\ .$$

It may be thought of as the density for transitions

$$(X_{t_1} = x \text{ at time } t_1) \longrightarrow (Y = X_{t_2} \text{ at time } t_2 \text{ near } y) .$$

In Lesson 1 we used Bayes' rule for Brownian motion to write the joint density for X_t at two times. The same reasoning applies here, except that the transition density may depend on t_1 and t_2 separately, not just on the length of the time interval $t_2 - t_1$. The result is

$$u_2(x_1, t_1, x_2, t_2) = u(x_1, t_1)\, G(X_{t_1}, t_2, X_{t_1}, t_1) .$$

The joint density of X_{t_2} and X_{t_1} is equal to the marginal density of X_{t_1} multiplied by the conditional density of X_{t_2} given X_{t_1}. For times $t_1 < t_2 < t_3$, similar reasoning leads to

$$u_3(x_1, t_1, x_2, t_2, x_3, t_3) = u(x_1, t_1)\, G(x_2, t_2, x_1, t_1)\, G(x_3, t_3, x_2, t_2) .$$

This is the joint PDF for the three values $X_1 = X_{t_1}$. $X_2 = X_{t_2}$ and $X_3 = X_{t_3}$. If X is not a Markov process, this last formula is more complicated. The "transition density" for X_3 would depend on both X_2 and X_1.

In continuous time, we need something like the transition distribution for time dt. We will see in future lessons that this is related to what is called the *generator* of the process. For now, it suffices to say that for diffusion processes, the time dt transitions are determined by what is called here the *infinitesimal mean* and *infinitesimal variance*. The proper names for these are *drift* and *quadratic variation*. The infinitesimal mean is defined by (slightly more formal versions are given below)

$$\mathrm{E}[\, X_{t+dt} - x \mid X_t = x] = a(x,t)\, dt . \tag{1}$$

The infinitesimal variance is defined by

$$\mathrm{var}[\, X_{t+dt} \mid X_t = x] = v(x,t)\, dt . \tag{2}$$

There is a theorem like the central limit theorem that says that a diffusion process is completely determined by its infinitesimal mean and variance.

Some Markov processes are not diffusions. Just Brownian motion is the simplest diffusion, the *Poisson arrival process* is the simplest non-diffusion continuous time Markov process. The Poisson arrival process models random events called "arrivals". The number of arrivals from time zero to time $t > 0$ is N_t. The probability of an arrival in $(t, t+dt)$ is $\lambda\, dt$, with λ being the *arrival rate* parameter. Arrivals in disjoint time intervals are independent. The time to the first arrival has probability density $\lambda e^{-\lambda t}$ (we will see). This arrival process satisfies (1) and (8) with $a(x,t) = \lambda$ and $v(x,t) = \lambda$. The infinitesimal mean and variance do not determine the process completely unless we also know that the process is a diffusion.

A diffusion is a dynamic stochastic model of something. We interpret the Markov property as saying that the model is "complete" in that the state of the system at time t, which is X_t, contains all the information about the past that is relevant for predicting the future. The SDE that describes the diffusion process is written in *Ito form* as

$$dX_t = a(X_t)\, dt + b(X_t)\, dW_t \ . \tag{3}$$

In this SDE, $a(x)$ is the drift and $b(x)$ is the noise coefficient. If $b(x) = 0$ (no noise), this might be rewritten in more familiar ODE form as

$$\frac{dX_t}{dt} = a(X_t) \ . \tag{4}$$

More general diffusions are not written in ODE form because X_t is not a differentiable function of t even though X_t is a continuous function of t.

The dW in the noise term has two interpretations. One is as the "differential increment" of Brownian motion.[1] That is, $dW_t = W_{t+dt} - W_t$. We saw in Lesson 1 that dW_t is a Gaussian random that is independent of everything up to time t. The mean is zero and the variance is dt. Therefore, the mean of $b(X_t)dW_t$ is zero (because dW_t is independent of X_t) and the variance is

$$\mathrm{var}(\, b(X_t)dW_t \mid X_t) = \mathrm{E}\Big[\, b(X_t)^2 \, (dW_t)^2 \mid X_t \Big] = b^2(X_t)\, dt \ .$$

The infinitesimal variance of a diffusion is the square of the noise term in the SDE.

The other interpretation does not connect the noise to Brownian motion. In this interpretation, dW_t is just a convenient way to write "mean zero, variance dt, independent of whatever happened before time t". The first interpretation (the *strong form* is helpful in technical analysis of diffusion processes. The second (the *weak form*), is more useful for modeling. In modeling, you are interested in the process X_t on its own, and not in relation to some idealized Brownian motion that is not part of the system you are modeling. In the strong form, X is a function of W. In the weak form, X lives on its own.

Brownian motion itself is the simplest interesting diffusion. It was called X_t in Lesson 1, but it is often called W_t to distinguish it from other diffusion processes. Recall that *standard* Brownian motion is Gaussian with the properties

1. $W_0 = 0$.
2. If $t_2 > t_1$, then $\mathrm{E}[\, W_{t_2} | W_{t_1}] = W_{t_1}$ (the *martingale* property).
3. $\mathrm{var}[\, W_{t_2} - W_{t_1}] = t_2 - t_1$.

Brownian motion is a particularly simple Markov process in which the *increment* is independent of the present and the past. The increment between time t_1 and time t (with $t > t_0$) is $\Delta W = W_t - W_{t_1}$. The increment of Brownian motion is independent of $W_{[0,t_1]}$. The SDE that describes Brownian motion has zero drift ($a = 0$) and constant noise coefficient $b = 1$.

[1] Brownian motion is also called the *Wiener process*, after MIT mathematician Norbert Wiener. The notation W_t is for Wiener.

The *Ornstein Uhlenbeck* process has a linear drift that seeks to return X to zero and a constant noise:

$$dX_t = -\gamma X_t \, dt + \sigma dW_t \; . \tag{5}$$

The *mean reversion rate* coefficient γ, and the noise coefficient σ are constants of the model. The OU process is often a good model of a system with a stable equilibrium that is subject to small outside disturbances. It was used by Einstein as a model of the velocity of a small particle in a fluid, with X_t being the velocity at time t. A moving particle will slow because of friction with the fluid, which is modeled by $-\gamma X$, the friction force is proportional to the velocity. The particle also is subject to random forces caused by collisions from water molecules. In Einstein's simple model the amount of noise is constant, independent of time and the speed of the particle. This is σdW.

Geometric Brownian motion models exponential growth (or decay) in the presence of noise. It differs from the OU process in that the noise is proportional to the level. It is defined by a growth rate parameter μ and a *volitility* parameter σ.

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t \; . \tag{6}$$

Geometric Brownian motion is a simple model of the random price of a share of stock through time. If there is no noise, then the stock is a simple exponential. The noise is made proportional to the level so that the value of N "shared" of stock doesn't change under a "stock split". A stock split is replacing each share by two shares worth half the amount: $n \to 2n$, and $S \to \frac{1}{2}S$. This is our first model with *multiplicative noise*, which means that $b(x)$ is not a constant but varies with x. The generic X_t is replaced by S_t (for "stock") for geometric Brownian motion.

This Lesson begins the discussion of diffusion processes. The next section gives a more technical definition of the Markov property, drift and noise. The main goal of this lesson is the partial differential equations (PDEs) related to X_t, which are the *backward equation* and the *forward equation.*

2 Diffusions

It is a theorem (not proved in this course) that a diffusion process is determined by the *infinitesimal mean* and *infinitesimal variance.* Infinitesimal mean is often called *drift* and infinitesimal variance is called *quadratic variation.* "Information about the past" of t is denoted $\mathcal{F}_t$. The precise definition of $\mathcal{F}_t$ is not important yet. The important thing here is that if A is anything determined by the path, then

$$\mathrm{E}\big[A | X_{[0,t_1]}\big] = \mathrm{E}[A | \mathcal{F}_{t_1}] \; . \tag{7}$$

Suppose $dt > 0$ is an infinitesimal increment of time.[2] The corresponding *increment* of the diffusion process is $dX_t = X_{t+dt} - X_t$. The infinitesimal mean

[2]Mathematicians don't like dt because dt is supposed to be smaller than any positive number and yet not zero. As a mathematical fact, if $Q \geq 0$ and Q is less than any posi-

and variance defined above may be written

$$E[\, dX_t \mid \mathcal{F}_t] = a(X_t,t)\, dt \;, \quad E\Big[\, (dX_t)^2 \mid \mathcal{F}_t \Big] = v(X_t,t)\, dt \;.$$

The infinitesimal variance, $v(x,t)$ is defined by

$$\mathrm{var}[dX_t|\mathcal{F}_t] = v(X_t,t)dt \;. \tag{8}$$

For more careful (but still not completely rigorous) mathematical work, we define $\Delta t > 0$ to be a small but not infinitely small increment of time and $\Delta X_t = X_{t+\Delta t} - X_t$ the corresponding small increment of X. We will derive simple formal formulas involving differentials using more complicated formulas involving ΔX and Δt. For example, much of ordinary (deterministic) calculus may be summarized by saying $(dt)^2 = 0$ even though $dt > 0$. For example,

$$d(t^2) = (t+dt)^2 - t^2 = t^2 + 2t\, dt + dt^2 - t^2 = 2t\, dt \;\; \text{(because } dt^2 = 0\text{)}.$$

We might then divide by dt to get

$$\frac{d}{dt} t^2 = 2t \;.$$

Here's the same thing done less informally with Δt.

$$\Delta(t^2) = (t+\Delta t)^2 - t^2 = 2t\, \Delta t + O(\Delta t^2) \;.$$

Therefore

$$\begin{aligned} \frac{d}{dt} t^2 &= \lim_{\Delta t \to 0} \frac{\Delta(t^2)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \left[\frac{2t\, \Delta t + O(\Delta t^2)}{\Delta t} \right] \\ &= \lim_{\Delta t \to 0} \left[2t + O(\Delta t) \right] \\ &= 2t \;. \end{aligned}$$

The definition of $Q = O(\Delta t^p)$ is: there is an $\epsilon > 0$ and a $C > 0$ so that if $\Delta t \leq \epsilon$ then $|Q| < C \Delta t^p$. We say that the quantity Q is "on the order of" Δt^p. We used two facts about "big Oh" orders. One is that if $Q = O(\Delta t^p)$ and $p > 1$, then $Q/\Delta t = O(\Delta t^{p-1})$. The other is that if $R = O(\Delta t^{p'})$, with $p' > 0$ then $R \to 0$ as $\Delta t \to 0$. We write "$O(\Delta t^p)$" instead of "Q, with $Q = O(\Delta t^p)$". If you did not study mathematical analysis in college these definitions may be confusing at first. But they do not get more complicated than this and should quickly seem natural. You basically treat $O(\Delta t^p)$ as though it were Δt^p,

tive number, then $Q = 0$. Our dt is less formal – very small, positive, yet not zero. The English/Irish philosopher George (Bishop) Berkeley mocked Newton's infinitesimals as the "ghosts of departed quantities". "Departed" means dead; dt has died (gone to zero) and yet lives on (isn't zero).

even though the truth is more complicated. In the above calculation, we used $O(\Delta t^2)/\Delta t = O(\Delta t)$ and then $O(\Delta t) \to 0$ as $\Delta t \to 0$.

In the Δt framework, infinitesimal mean is

$$\mathrm{E}[\,\Delta X_t \mid \mathcal{F}_t] = a(X_t, t)\,\Delta t + O(\Delta t^2)\ . \tag{9}$$

the infinitesimal variance is

$$\mathrm{var}(\,\Delta X_t \mid \mathcal{F}_t) = v(X_t, t)\,\Delta t + O(\Delta t^2)\ . \tag{10}$$

There is a simpler expression

$$\mathrm{E}\Big[\,(\Delta X_t)^2 \mid \mathcal{F}_t\Big] = v(X_t, t)\,\Delta t + O(\Delta t^2)\ . \tag{11}$$

The formula (10) is equivalent to (11), because, if we assume 11), then

$$\begin{aligned}
\mathrm{var}(\,\Delta X_t \mid \mathcal{F}_t) &= \mathrm{E}\Big[\,(\Delta X_t)^2 \mid \mathcal{F}_t\Big] - \mathrm{E}[\,\Delta X_t \mid \mathcal{F}_t]^2 \\
&= \mathrm{E}\Big[\,(\Delta X_t)^2 \mid \mathcal{F}_t\Big] - \big[\,a(X_t, t)\Delta t + O(\Delta t^2)\big]^2 \\
&= v(X_t, t)\Delta t + O(\Delta t^2)\ .
\end{aligned}$$

Here we use more facts about "big Oh". One is that $\Delta t + O(\Delta t^2) = O(\Delta t)$. Another is that $(O(\Delta t))^2 = O(\Delta t^2)$. The above calculation should convince you that "big Oh" is a convenient way to reason about small quantities whose exact size isn't relevant.

The formula (11) may be interpreted as saying that ΔX_t is approximately on the order of $\sqrt{\Delta t}$, because ΔX^2 is, in the expected value, on the order of Δt. If ΔX is on the order of $\Delta t^{\frac{1}{2}}$, then ΔX^4 should be on the order of Δt^2. The Poisson arrival process shows that this reasoning is flawed. In fact,

$$\begin{aligned}
\mathrm{E}[\,\Delta N \mid \mathcal{F}_t] &= \lambda \Delta t + O(\Delta t^2)\ , \\
\mathrm{E}\Big[\,(\Delta N)^2 \mid \mathcal{F}_t\Big] &= \lambda \Delta t + O(\Delta t^2)\ , \\
\mathrm{E}\Big[\,(\Delta N)^4 \mid \mathcal{F}_t\Big] &= \lambda \Delta t + O(\Delta t^2)\ .
\end{aligned}$$

The first two seem fine, but the last one violated the reasoning. More on this in the exercises.

Most of the time you can tell a diffusion process from another Markov process fourth moments. A diffusion process has a fourth moment that follows the informal reasoning:

$$\mathrm{E}\Big[\,(\Delta X)^4 \mid \mathcal{F}_t\Big] = O(\Delta t^2)\ . \tag{12}$$

For most purposes, particularly in this class, you reason about diffusions using the infinitesimal mean (9), the infinitesimal square (11), which is equivalent to the infinitesimal variance (10), and the fourth moment bound (12). Future lessons will have more on the fourth moment bound.

2.1 Proofs with big Oh

The statement $Q = O(P)$ as $\Delta t \to 0$ literally means that Q and P are functions of Δt and there is an $\epsilon > 0$ and a $C < \infty$ so that $|Q| \leq CP$ if $\Delta t \leq \epsilon$. In a formula like $A = B + O(P)$ as $\Delta t \to 0$ the $O(P)$ is a quantity Q with $Q = O(P)$. There can be more than one big Oh quantity in a formula, for example $A = B + O(P) + O(R)$ as $\Delta t \to 0$. Here $O(P)$ represents a quantity Q_1 with $Q_1 = O(P)$ and $O(R)$ represents a Q_2 with $Q_2 = O(R)$. A typical application has P or R being powers of Δt. Someone who has understood a class in "Mathematical Analysis" (ϵ and δ proofs) will be able to reason with big Oh. These examples are for people who haven't taken such a class or are rusty.

Example 1. Show that $O(\Delta t^p)/\Delta t = O(\Delta t^{p-1})$. **A solution:** Let $P = O(\Delta t^p)$ be the quantity on the left and $R = P/\Delta t$. From the definition, there is an $\epsilon > 0$ and a $C > 0$ so that $|P| \leq C\Delta t^p$ if $\Delta t \leq \epsilon$. Therefore, if $\Delta t \leq \epsilon$, $R \leq C\Delta t^p/\Delta t = C\Delta t^{p-1}$.

Example 2. Show that $O(\Delta t^{p_1})O(\Delta t^{p_2}) = O(\Delta t^{p_1+p_2})$. **A solution:** Call the two functions $P_1 = O(\Delta t^{p_1})$ and $P_2 = O(\Delta t^{p_2})$. By the definitions, there is an $\epsilon_1 > 0$ and a C_1 so that $|P_1| \leq C_1\Delta t^{p_1}$ if $\Delta t \leq \epsilon_1$. There also is an ϵ_2 and C_2 for P_2. Take $\epsilon = \min(\epsilon_1, \epsilon_2)$. The min of two positive numbers is a positive number. If $\Delta t \leq \epsilon$ then $\Delta t \leq \epsilon_1$ and $\Delta t \leq \epsilon_2$. Therefore $|P_1| \leq C_1\Delta t^{p_1}$ and $|P_2| \leq C_2\Delta t^{p_2}$. Therefore $|P_1, P_2| \leq C_1C_2\Delta t^{p_1+p_2}$ if $\Delta t \leq \epsilon$. This proves that $|P_1, P_2| \leq C\Delta t^{p_1+p_2}$, with $C = C_1C_2$. *Comment:* Example 1 is a special case of Example 2, with $p_2 = -1$.

Example 3. Is it true that $O(\Delta t^{p_1})/O(\Delta t^{p_2}) = O(\Delta t^{p_1-p_2})$? **A solution:** It's not true. If $|P_2| \leq C_2\Delta t^{p_2}$ it might be that P_2 is much smaller than this. For example, $\Delta t^2 = O(\Delta t)$. If $P_2 = \Delta t^2$, and $P_1 = \Delta t^2$, then P_1/P_2 is not order $p_1 - p_2 = 2 - 1 = 1$. It is common to write $P = O(\Delta t^p)$ to imply that P is about that size and not very much smaller. But this is not part of the "big Oh" definition.

Example 4. Suppose that $f(x)$ and f' and f'' are continuous functions of x. Show that $f(x) = f(0) + xf'(0) + O(x^2)$ as $x \to 0$. *Comment: we use x instead of Δt as the variable that is going to zero.* **A solution:** One form of the Taylor series remainder formula is

$$f(x) = f(0) + xf'(0) + \frac{1}{2}x^2 f''(\xi)\;, \quad |\xi| \leq x\;.$$

Here, we know there is a ξ but do not know its value. Since f'' is continuous, there is a C so that

$$\max_{|x|\leq 1} |f''(x)| = C < \infty\;.$$

Take $\epsilon = 1$ and use the C given.

2.2 Cauchy Schwarz inequality

Suppose A and B are two random variables. The Cauchy Schwarz inequality is

$$\mathrm{E}[\,AB\,] \leq \sqrt{\mathrm{E}[\,A^2\,]\;\mathrm{E}[\,B^2\,]}\;. \tag{13}$$

The proof is a clever trick. For every real number t,

$$\mathrm{E}\Big[\,(A-tB)^2\Big] \geq 0\;.$$

If a random quantity is non-negative, then its expected value cannot be negative. Calculating, we get

$$0 \leq \mathrm{E}\big[\,A^2\big] - 2t\,\mathrm{E}[\,AB] + t^2\mathrm{E}\big[\,B^2\big]\;\;.$$

Since the right side is non-negative for every t, its minimum is non-negative. Minimizing over t (differentiate with respect to t, set the derivative to zero, solve for t), the minimum is achieved at

$$t_* = \frac{\mathrm{E}[\,AB]}{\mathrm{E}[\,B^2]}\;.$$

This gives

$$0 \leq \mathrm{E}\big[\,A^2\big] - \frac{\mathrm{E}[\,AB]^2}{\mathrm{E}[\,B^2]}\;.$$

Finally, multiply by the non-negative quantity $\mathrm{E}\big[\,B^2\big]$ and you get

$$\mathrm{E}[\,AB]^2 \leq \mathrm{E}\big[\,A^2\big]\,\mathrm{E}\big[\,B^2\big]\;\;.$$

The square root form of this is (13).

The Cauchy Schwarz inequality implies an inequality involving variance and covariance. Suppose

$$\overline{A} = \mathrm{E}[\,A]\;\;,\;\;\overline{B} = \mathrm{E}[\,B]\;\;.$$

Replace A with $A-\overline{A}$ and B with $\overline{B}$. The covariance is

$$\mathrm{cov}(A,B) = \mathrm{E}\big[\,(A-\overline{A})(B-\overline{B})\big]\;\;.$$

The Cauchy Schwarz gives

$$\mathrm{cov}(A,B) \leq \sqrt{\mathrm{var}(A)\,\mathrm{var}(B)}\;.$$

This may be re-written in terms of the *correlation coefficient* between two random variables:

$$\mathrm{corr}(A,B) = \frac{\mathrm{cov}(A,B)}{\sqrt{\mathrm{var}(A)\,\mathrm{var}(B)}}$$

This is a dimensionless measure of the statistical relationship between A and B. The Cauchy Schwarz inequality implies that

$$-1 \leq \mathrm{corr}(A,B) \leq 1\;.$$

Absolute number bounds make sense for correlation because it is dimensionless.

We are interested in the Cauchy Schwarz inequality here because of something technical we are about to do. There will soon be a Taylor series calculation up to order ΔX^2 with an error that is of order ΔX^3. We need to "bound" ΔX^3 in terms of the second and fourth moments. For this, apply Cauchy Schwarz with $A = |\Delta X|$ and $B = \left|\Delta X^2\right|$. The result is

$$\begin{aligned} \mathrm{E}\Big[\, |\Delta X|^3 \Big] &= \mathrm{E}\Big[\, |\Delta X| \, (\Delta X)^2 \Big] \\ &\leq \sqrt{\mathrm{E}\Big[\, (\Delta X)^2 \Big] \, \mathrm{E}\Big[\, (\Delta X)^4 \Big]} \end{aligned}$$

If X is a diffusion process, then we can use the variance bound (11) and the fourth moment bound (12), and some "big Oh calculations" to get

$$\mathrm{E}\Big[\, |\Delta X|^3 \Big] = \sqrt{O(\Delta t) O(\Delta t^2)} = O(\Delta t^{3/2}) \; . \tag{14}$$

3 Backward equation

Suppose $V(x)$ is a *payout* function. The corresponding *value function*, $f(x,t)$, is defined for $t \leq T$, by

$$f(x,t) = \mathrm{E}[\, V(X_T) \mid X_t = x] = \mathrm{E}_{x,t}[\, V(X_T) \mid X_t = x] \; . \tag{15}$$

The terms "payout" and "value" come from financial applications, but the ideas are more general than finance. The value function f depends on the payout function V and also on the diffusion process X. The goal of this section is to show that the value function satisfies the partial differential equation called the *backward equation*

$$\partial_t f + a(x)\partial_x f + \frac{1}{2} v(x) \partial_x^2 f = 0 \; . \tag{16}$$

Be careful not to confuse the payout function $V(x)$ with the infinitesimal variance $v(x)$.

The derivation of the backward equation has two steps. The first step is the *tower property*, also called the *law of total probability*. This allows us to express the values $f(x,t)$ in terms of the values $f(x,t_1)$ with $t_1 > t$. The Markov property also enters. It implies that

$$\mathrm{E}[\, V(X_T) \mid X_t = x \text{ and } X_{t_1} = y] = \mathrm{E}[\, V(X_T) \mid X_{t_1} = y] = f(y,t_1) \; . \tag{17}$$

This leads to the equation

$$f(x,t) = \mathrm{E}_{x,t}[\, f(X_{t_1}, t_1)] \; . \tag{18}$$

The value function at time t is represented as the expected value of the value function at a future time $t_1 > t$.

The second step is to apply the tower property with $t_1 = t + \Delta t$ and do Taylor series calculations in Δt. There is an increment ΔX corresponding to the time increment Δt. We will have to expand f to first order in Δt and to second order in ΔX. This is because $\mathrm{E}[\Delta X^2] = O(\Delta t)$.

The tower property is that the expected value of the expected value is the expected value. Suppose (Y, Z) is a two dimensional random variable with a joint PDF $u_2(y, z)$. Suppose $V(z)$ is a payout function and $g(y)$ is the conditional expectation of $V(Z)$ given that $Y = y$. Suppose f is the unconditional expectation of $V(Z)$. The tower property is

$$f = \mathrm{E}[\, g(Y)] \ . \tag{19}$$

The overall expectation is the expected value of the conditional expectation. this is a "tower" of expectations and conditioning.

Here are the formulas for (19). The marginal probability density for Y is

$$u_1(y) = \int u_2(y, z)\, dz \ .$$

The conditional density for Z given $Y = y$ is

$$u(z|y) = \frac{u_2(y, z)}{u_1(y)} \ .$$

You can check that $u(z, y)$ is a probability density in z for each y by integrating

$$\int u(z|y)\, dz = \frac{1}{u_1(y)} \int u_2(y, z)\, dz = \frac{u_1(y)}{u_1(y)} = 1 \ .$$

The conditional expectation may be written in several ways:

$$g(y) = \mathrm{E}_y[\, V(Z)] = \mathrm{E}[\, V(Z) \mid Y = y] = \int V(z) u(z|y)\, dz \ .$$

The overall expectation is

$$f = \mathrm{E}[\, V(Z)] = \int \int V(z) u_2(y, z)\, dydz \ .$$

The tower property is that this is the expected value of g:

$$f = \mathrm{E}[\, g(Y)] = \int g(y) u_1(y)\, dy \ .$$

We can verify this by substituting some of the above definitions:

$$\begin{aligned}
\int g(y) u_1(y)\, dy &= \int \int V(z) u(z_y) u_1(y)\, dydz \\
&= \int \int V(z) \frac{u_2(y, z)}{u_1(y)} u_1(y)\, dydz \\
&= \int \int V(z) u_2(y, z)\, dydz \ .
\end{aligned}$$

We apply the tower property with $Z = X_T$ and $Y = X_{t_1}$. The starting value $X_t = x$ is fixed throughout. All these calculations assume that (are conditioned on) $X_t = x$. The conditional expectation, called g in the abstract calculations above, is

$$g(y) = \mathrm{E}[\, V(X_T) \mid X_{t_1} = y \text{ and } X_t = x] \ .$$

The Markov property makes the second condition on the right irrelevant. Since T is in the future of t_1, which is in the future of t, once we know $X_{t_1} = y$, the value of X_t is irrelevant for expectations involving X_T. Therefore (this may be the main step in the whole thing)

$$g(y) = \mathrm{E}[\, V(X_T) \mid X_{t_1} = y] = f(y, t_1) \ .$$

The equation (18) follows from this when we substitute the definition $Y = X_{t_1}$.

The second step is the Taylor calculations. Take $t_1 = t + \Delta t$ and $X_{t_1} = x + \Delta x$. We expand $f(x + \Delta x, t + \Delta t)$ in Taylor series about x and t. Error "estimates"[3] in Taylor series are usually "the first neglected terms". Example 4 in the big Oh section is like that. It is also true, if you are careful, for functions of more than one variable like $f(x,t)$. If we assume that partial derivatives of f up to third order are continuous, then

$$f(x + \Delta x, t + \Delta t) = f + \Delta x \, \partial_x f + \frac{1}{2} \Delta x^2 \partial_x^2 f + \Delta t \, \partial_t f \qquad (20)$$

$$+ O(|\Delta x|^3) + O(\Delta t \, |\Delta x|) + O(\Delta t^2) \ . \qquad (21)$$

The arguments x, t are left out of every term on the right on the top line (20). We write, f for $f(x,t)$, $\partial_t f$ for $\partial_t f(x,t)$, etc. One of the "first neglected terms" is $\frac{1}{6} \Delta x^3 \partial_x^3 f$. This is on the order of $|\Delta x|^3$. The other lowest order neglected terms involve $\partial_t^2 f$ and $\partial_t \partial_x f$. They give the other two error contributions on the second line (21).

We put the expansion (20) (21) into the tower property formula (18) with $t_1 = t + \Delta t$ and $X_{t_1} = x + \Delta X$. We take expected values. Terms at time t with argument x come out of the expectation because they are determined (when $X_t = x$). The result is

$$f = f + \mathrm{E}_{x,t}[\, \Delta X] \, \partial_x f + \frac{1}{2} \mathrm{E}_{x,t}\left[\, \Delta X^2 \right] \partial_x^2 f + \Delta t \partial_t f$$

$$+ \mathrm{E}_{x,t}\left[O\left(|\Delta X|^3 \right) \right] + \Delta t \, \mathrm{E}_{x,t}[\, O\left(|\Delta X| \right)] + O\left(\Delta t^2 \right) \ .$$

We evaluate the expectations on the top line using the infinitesimal mean and variance formulas (9) and (11). The $O(\Delta t^2)$ error terms in those formulas contribute to the $O(\Delta t^2)$ on the second line.

$$0 = a(x) \Delta t \partial_x f + \frac{1}{2} v(x) \Delta t \partial_x^2 f + \Delta t \partial_t f$$

$$+ \mathrm{E}_{x,t}\left[O\left(|\Delta X|^3 \right) \right] + \Delta t \, \mathrm{E}_{x,t}[\, O\left(|\Delta X| \right)] + O\left(\Delta t^2 \right) \ .$$

[3] An *estimate* in mathematical proofs is not a guess at how large something is, but an upper bound. Any "big Oh" formula is an "estimate" in this sense.

We showed that $\mathrm{E}_{x,t}\left[\,|\Delta X|^3\right] = O(\Delta t^{\frac{3}{2}})$. It is an exercise to show that $\mathrm{E}_{x,t}[\,|\Delta X|] = O(\Delta t^{\frac{1}{2}})$. This gives (since $O(\Delta t^{\frac{3}{2}}) + O(\Delta t^2) = O(\Delta t^{\frac{3}{2}})$)

$$0 = a(x)\Delta t \partial_x f + \frac{1}{2} v(x) \Delta t \partial_x^2 f + \Delta t \partial_t f + O(\Delta t^{\frac{3}{2}})\ .$$

Finally, we divide both sides by Δt and take the limit $\Delta t \to 0$. The result is the backward equation (16).