Week 9
Generators, duality, change of measure

Jonathan Goodman

November 18, 2013

1 Generators

This section describes a common abstract way to describe many of the differential equations related to Markov processes. The forward or backward equation of a finite state Markov chain involves a matrix. The backward equation of a diffusion process involves partial derivatives in time and space. The space derivatives define a linear *operator*. This acts in a linear way on functions in the way a matrix acts in a linear way on vectors. There is a common abstract definition of the *generator* of a Markov process. This definition gives the backward equations we know in each of the cases we have studied so far.

The abstract point of view is convenient even when talking about a single example. The general point of view makes it clearer "what is going on" in examples. It helps us understand and remember formulas by making them seem simpler and more natural.

The backward equation of a Markov process involves the generator. The forward equation involves the *adjoint* of the generator. *Duality* is the relation between an operator and its adjoint. The *dual* of a matrix is its adjoint (transpose for real matrices). The adjoint of a differential operator is another differential operator.

1.1 Generator, discrete time

There are continuous and discrete time Markov processes. Suppose X_t is a discrete time Markov process and f is a function of the state x. The *generator* of the process is a linear *operator* defined by

$$Lf(x) = \mathrm{E}[\, f(X_1) \mid X_0 = x] \; . \tag{1}$$

Some examples will clarify this simple general definition.

Consider a finite state Markov chain with states $x_1, \ldots, x_n$. A function of x has n values, which we call $f_1, \ldots, f_n$. We abuse notation by writing f_k for $f(x_k)$. A column vector, also called f, has the numbers f_k as components. Computer programs make little distinction between vectors and function, writing `f[k]` for components of a vector and `f(k)` for a function of k. Some languages use `f(k)` for both.

Suppose $g = Lf$ by the definition (1). Then, as we saw in week 1,

$$g_j = \sum_k f_k \,\mathrm{P}(X_1 = k \mid X_0 = j) \\ = \sum_k \mathrm{P}(j \to k)\, f_k \; . \tag{2}$$

The transition probabilities are $P_{jk} = \mathrm{P}(j \to k)$. These are the entries of the transition matrix P. We recognize (2) as matrix multiplication $g = Pf$. This shows that in the case of finite state discrete time Markov chains, the generator is the same as the transition matrix. The generator is not a new object. Definition (1) is just an indirect way to describe transition probabilities.

Another example is a discrete time scalar linear Gaussian process $X_{t+1} = aX_t + \sigma Z_t$, where the Z_t are i.i.d. standard normals. A function of the state in this case is $f(x)$, a function of one continuous variable, x. If $g = Lf$, then (1) gives (recall the formula for a normal with mean ax and variance σ^2):

$$\begin{aligned} g(x) &= \mathrm{E}[\, f(X_t) \mid X_0 = x] \\ &= \mathrm{E}[\, f(ax + \sigma Z)] \\ &= \int_{-\infty}^{\infty} f(y)\, \frac{e^{-(ax-y)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\, dy \\ &= \int_{-\infty}^{\infty} P(x,y) f(y)\, dy \; . \end{aligned} \tag{3}$$

The *transition density* $P(x,y)$ in the last line is given by

$$P(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(ax-y)^2/(2\sigma^2)} \; . \tag{4}$$

We write this also as $P(x \to y)$. It is a probability density as a function of y, but not as a function of x.

The generator formalism (1) handles lots of more complicated examples. For example, suppose $X_t = (X_{1,t}, X_{2,t})$ is a two component continuous random variable, and at each time we take a Gaussian step either one of the components chosen at random with equal probabilities. Then a function of the state is a function of two variables $f(x_1, x_2)$, and (1) gives $g = Lf$ as

$$\begin{aligned} g(x_1,x_2) &= \mathrm{E}[\, f(X_{1,1}, X_{2,1}) \mid X_{1,0} = x_1 \text{ and } X_{2,0} = x_2] \\ &= \frac{1}{2}\Big(\mathrm{E}[\, f(X_{1,1}, X_{2,1}) \mid X_{1,0} = x_1 \text{ and } X_{2,0} = x_2 \text{ and moved } X_1] \\ &\quad + \mathrm{E}[\, f(X_{1,1}, X_{2,1}) \mid X_{1,0} = x_1 \text{ and } X_{2,0} = x_2 \text{ and moved } X_2]\Big) \\ &= \frac{1}{2}\Big(\int_{-\infty}^{\infty} \frac{e^{-(x_1-y)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \, f(y, x_2)\, dy \\ &\quad + \int_{-\infty}^{\infty} \frac{e^{-(x_2-y)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \, f(x_1, y)\, dy\Big) \; . \end{aligned}$$

This is an integral formula for the generator of the Markov chain defined by

$$(X_{1,t+1}, X_{2,t+1}) = \begin{cases} (X_{1,t} + \sigma Z_t, X_{2,t}) \;, & \text{Prob } = \frac{1}{2} \\ (X_{1,t}, X_{2,t} + \sigma Z_t) \;, & \text{Prob } = \frac{1}{2} \; . \end{cases}$$

1.2 Generator, continuous time

If X_t is a continuous time Markov process, then the generator is defined by

$$Lf(x) = \lim_{t \to 0} \frac{\mathrm{E}[\, f(X_t) - f(x) \mid X_0 = x]}{t} \; . \tag{5}$$

The definition presupposes that $\mathrm{E}[\, f(X_t) \mid X_0 = x] \to f(x)$ as $t \to 0$. Since we divide by t, the definition presupposes that $\mathrm{E}[\, f(X_t) \mid X_0 = x] = f(x) + O(t)$. There are continuous time Markov processes so that the limit (5) hardly exists. For diffusion processes, the limit may not exist if the function f is not in the *domain* of the generator. A theoretical PhD level class on stochastic processes would discuss these issues.

The general expression has many different specific realizations, as was the case for discrete time processes. Suppose X_t is a continuous time finite state space process with transition rate matrix $R_{jk} = \text{rate}(j \to k)$, defined by $\mathrm{P}(X_{t+dt} = k \mid X_t = j) = R_{jk} dt$, if $j \neq k$. To apply the generator definition (5), we record the short time transition probabilities:

$$\mathrm{P}(X_{dt} = j \mid X_0 = k) = \left\{ \begin{array}{l} R_{jk} dt \text{ if } j \neq k \\ 1 - \left(\sum_{l \neq j} P_{jl}\right) \text{ if } j = k \; . \end{array} \right\} \tag{6}$$

It is convenient to define the $j \to j$ "rate" as

$$R_{jj} = -\sum_{l \neq j} R_{jl} \; . \tag{7}$$

This is not really a rate because it a negative number. With this definition of R_{jj}, the whole rate matrix satisfies

$$\sum_{k=1}^{n} R_{jk} = 0 \; ,$$

for all j.

The rate matrix definition (7) simplifies the sort time approximation to the transition probability formula. The transition probability matrix at time t is

$$P_{jk}(t) = \mathrm{P}(j \to j \text{ in time } t) = \mathrm{P}(X_t = k \mid X_0 = j) \ . \tag{8}$$

With the R_{jj} definition (7), the short time approximate transition probability formulas (6) may be written in matrix form

$$P(dt) = I + R\,dt \quad , \quad P_{jk}(dt) = \delta_{jk} + R_{jk}dt \ . \tag{9}$$

A more precise mathematical statement of this is $P(t) = I + Rt + O(t^2)$.

The generator of a continuous time finite state space Markov process is the rate matrix. A function of x is a vector $f \in \mathbb{R}^n$. The definition (5) gives a formula for $g = Lf$:

$$\begin{aligned} g(j) &= \lim_{t\to 0} \frac{\mathrm{E}[\, f(X_t) - f(j) \mid X_0 = j]}{t} \\ &= \lim_{t\to 0} \frac{1}{t}\left(\sum_{k=1}^{n} \mathrm{P}(X_t = k \mid X_0 = j)\,[f(k) - f(j)] \right) \\ &= \lim_{t\to 0} \frac{1}{t}\left(\sum_{k=1}^{n} \left[\delta_{jk} + R_{jk}t + O(t^2)\right][f(k) - f(j)] \right) \ . \end{aligned}$$

It takes some cancellation for the limit to exist. We find that cancellation here by noting that $\sum \delta_{jk} f(k) = f(j)$. We apply this, then divide by t, and are left with

$$g(j) = \sum_{k=1}^{n} R_{jk} f(k) + O(t) \ .$$

In matrix form, this is

$$Lf = Rf \ , \tag{10}$$

which shows that the transition rate matrix is the generator.

Next consider a one dimensional diffusion process that satisfies

$$dX_t = a(X_t)dt + b(X_t)dW_t \ . \tag{11}$$

We have seen before that this is equivalent to three moment conditions

$$\left.\begin{aligned} \mathrm{E}[\,\Delta X_t \mid \mathcal{F}_t] &= a(X_t)\Delta t + o(\Delta t) \ , \\ \mathrm{E}\Big[\,(\Delta X_t)^2 \mid \mathcal{F}_t\Big] &= \sigma(X_t)^2 \Delta t + o(\Delta t) \ , \\ \mathrm{E}\Big[\,(\Delta X_t)^4 \mid \mathcal{F}_t\Big] &= o(\Delta t) \ . \end{aligned}\right\} \tag{12}$$

We apply these with $t = 0$, and Δt for t, and $\Delta X = X_t - x$. The Taylor expansion of f is

$$f(x + \Delta x) = f(t) + \partial_x f(x)\Delta X + \frac{1}{2}\partial_2 f(x)\Delta X^2 + O(|\Delta x|^3) \ .$$

We apply the moment conditions (12) and Cauchy Schwarz to get

$$
\begin{aligned}
\mathrm{E}[\, f(x+\Delta X) \mid X_0 = x] \; - \; f(x) \;&=\; \partial_x f(x)\mathrm{E}[\, \Delta x \mid X_0 = x] \; + \; \frac{1}{2}\partial_x^2 f(x)\mathrm{E}\Big[\, (\Delta x)^2 \mid X_0 = x\Big] \\
&\quad + \; O(|\Delta x|^3) \\
&=\; \partial_x f(x)a(x)t + \frac{1}{2}\sigma^2(x)\partial_x^2 f(x)\sigma(x)^2 t + o(t) \; .
\end{aligned}
$$

We substitute these into the generator definition (5), divide by t and let $t \to 0$, which leaves a formula for the generator of a diffusion process:

$$
Lf(x) = a(x)\partial_x f(x) + \frac{1}{2}\sigma^2(x)\partial_x^2 f(x) \; . \tag{13}
$$

Another example will illustrate the ease and simplicity of the generator approach. A *space dependent jump process* has a space dependent jump rate $\lambda(x)$. If $X_t = x$ then the probability of a jump in time dt is $\lambda(x)dt$. The jump distribution is given by a transition density function $R(x,y)$. This is the probability density to land at y if you jump from x. It is "easy" to see that the generator is given by

$$
Lf(x) = \lambda(x) \int R(x,y) f(y) dy \; .
$$

You recognize from (13) that the backward equation can be written in the abstract form

$$
0 = \partial_t f + Lf \; . \tag{14}
$$

We give a proof of this that holds for any continuous time Markov process with a generator. That proof then applies to diffusions, to finite state space processes, to jump processes, etc. The abstract proof has a third advantage beyond generality and clarity. It is simple. Consider a function $f(x,t)$, where x is in the state space of the Markov process. Suppose f is as differentiable as it needs to be for this derivation. Calculate

$$
\begin{aligned}
\mathrm{E}_{x,t}[\, f(X_{t+\Delta t}, t+\Delta t)] = \mathrm{E}_{x,t}[\, f(X_{t+\Delta t}, t)]& \\
+ \Delta t \, \mathrm{E}_{x,t}[\, \partial_t f(x,t)] + o(\Delta t)& \\
+ \Delta t \, \mathrm{E}_{x,t}[\, (\partial_t f(X_t,t) - \partial_t f(x,t))] &\; .
\end{aligned}
$$

The expectation in the last term on the right is $o(\Delta t)$ because it is the expectation of the kind of difference that appears in the definition (5) of L. We already showed that the first term on the right is

$$
\mathrm{E}_{x,t}[\, f(X_{t+\Delta t}, t)] = f(x,t) + Lf(x,t)\Delta t + o(\Delta t) \; .
$$

We see this by applying (5) to f as a function of x with t as a parameter. Substituting back, we find that

$$
\mathrm{E}_{x,t}[\, f(X_{t+\Delta t}, t+\Delta t)] = f(x,t) + \Big(Lf(x,t) + \partial_t f(x,t)\Big)\Delta t + o(\Delta t) \; . \tag{15}
$$

Finally, suppose $f(x,t) = \mathrm{E}_{x,t}[\,V(X_T)]$ is a value function. The tower property gives the equality

$$\mathrm{E}_{x,t}[\,f(X_{t+\Delta t}, t+\Delta t)] = f(x,t) \; .$$

The backward equation (14) follows from this by looking at the $O(\Delta t)$ terms that remain.

Ito's lemma has an interesting statement in terms of the generator. Suppose X_t is a diffusion process with a decomposition $X_t = Y_t + Z_t$ into the martingale part and the differentiable part. Then $\mathrm{E}\left[\,(dX_t)^2\right] = \mathrm{E}\left[\,(dY_t)^2\right]$ and $\mathrm{E}\left[\,(dZ_t)^2\right] = 0$. Ito's lemma is

$$df(X_t,t) = \partial_t f(X_t,t)dt + Lf(X_t,t)dt + \partial_x f(X_t,t)dY_t \; . \tag{16}$$

The first two terms on the right represent the deterministic part of the change in f. The last term is the unpredictable part. The definition (5) of the generator says the same thing, that Lf is the deterministic part of $df(X_t)$. The actual change in $f(X_t)$ is the expected change plus a part that has expected value zero, which is the *fluctuation*.

1.3 Duality and the forward equation in discrete time

The *forward equation* is the evolution equation that determines the probability distribution of X_t from the distribution of X_0. Forward equations are not as universal as backward equations because they concern probability measures rather than functions. The form of the forward equation depends more on the nature of the state space and probability measure for the forward equation. It is common to describe the generator and the backward equation instead of the forward equation for that reason.

Given the ease of writing the generator and the backward equation, it is natural to try to derive the evolution equation for the probability distribution of X_t from the backward equation rather than from "scratch" (a term in cooking for using unprocessed ingredients such as flour and yeast rather than, say, prepared pizza crust). If $u(t)$ represents the probability distribution of X_t, the forward equation in discrete time takes the form

$$u(t+1) = L^* u(t) \; . \tag{17}$$

The operator L^*, pronounced "ell star", is the *adjoint* of the generator, L.

The general notion of adjoint involves a *pairing* between functions and probability distributions. If X is a random variable with probability distribution u and f is a function, then

$$\langle u, f\rangle = \mathrm{E}[\,f(X)] \; . \tag{18}$$

Exactly what form this pairing takes depends on the situation. The general definition of *adjoint* is as follows. If L is an operator, then L^* is the adjoint of L if

$$\langle u, Lf\rangle = \langle L^* u, f\rangle \; , \tag{19}$$

for all functions f and distributions u for which the expressions make sense. We will see in all the examples that (19) determines L^*, which means that there is an L^* that satisfies the conditions, and that L^* is unique.

Now suppose the function $f(x)$ is the value function at some time: $f(x,t) = \mathrm{E}_{x,t}[\,V(X_T)]$. Let $f(t) = f(\cdot,t)$ be the vector that is the function $f(x,t)$ at time t. Let $u(t)$ be the probability distribution of X_t. The tower property for $\mathrm{E}[\,V(X_t)]$ without conditioning can be conditioned at time t or time $t+1$. The result is $\mathrm{E}[\,V(X_T)] = \mathrm{E}[\,f(X_t,t)] = \mathrm{E}[\,f(X_{t+1},t+1)]$. In terms of the pairings, we have $\mathrm{E}[\,f(X_t,t)] = \langle u(t), f(t)\rangle$ and $\mathrm{E}[\,f(X_{t+1},t+1)] = \langle u(t+1), f(t+1)\rangle$. The equality of pairings is

$$\langle u(t), f(t)\rangle = \langle u(t+1), f(t+1)\rangle \; .$$

The backward equation gives $f(t) = Lf(t+1)$. This allows the manipulations, which use the definition (19), and the fact that the pairing is linear in u and f,

$$\begin{aligned} \langle u(t), Lf(t+1)\rangle &= \langle u(t+1), f(t+1)\rangle \\ \langle L^* u(t), f(t+1)\rangle &= \langle u(t+1), f(t+1)\rangle \\ \langle \left(L^* u(t) - u(t+1)\right), f(t+1)\rangle &= 0 \; . \end{aligned}$$

The simplest way for this identity to be true is (17).

Here is how this abstract stuff works out in the case of a finite state space Markov chain. The probability distribution is given by the numbers $u_j(t) = \mathrm{P}(X_t = j)$. The pairing is

$$\langle u(t), f(t)\rangle = \mathrm{E}[\,f(X_t)] = \sum_{j=1}^{n} f(j)\mathrm{P}(X_t = j) = \sum_{j=1}^{n} u_j(t) f_j(t) \; .$$

This is just the vector inner product. Two different conventions are used commonly in this context. Some consider $u(t)$ to be an n component row vector, which is a $1 \times n$ matrix. Others consider $u(t)$ to be an n component column vector, which is an $n \times 1$ matrix. If $u(t)$ is a column vector, then $u(t)^t$ is the corresponding row vector. Depending on which convention is used, the pairing can be written in matrix/vector notation as

$$\langle u(t), f(t)\rangle = \begin{cases} u^t(t) f(t) & \text{if } u \text{ is a column vector} \\ u(t) f(t) & \text{if } u \text{ is a row vector .} \end{cases}$$

Both expressions are the product of an n component row vector (u or u^t depending on the convention) with an n component column vector, f.

The "u is a row vector" convention is very convenient for studying finite state space Markov chains. But it doesn't really work for fancier problems such as continuous time diffusions. This class will take the "u is a column vector" convention for that reason. In this convention, (19) becomes

$$u^t L f = \left(L^t u\right)^t f \; ,$$

so $L^* = L^t$. To check the algebra, note that the transpose of the transpose is the original matrix, $\left(L^t\right)^t = L$. Therefore $\left(L^t u\right)^t f = u^t \left(L^t\right)^t f = u^t L f$, as claimed. The adjoint in this pairing is just the matrix transpose. The forward equation is $u(t+1) = L^t u(t)$. We saw that the generator, L, is the transition probability matrix P. Therefore, the forward equation is

$$u(t+1) = P^t u(t) \; . \tag{20}$$

We repeat the less abstract derivation of the forward equation (20) from week 1.

$$\begin{aligned} u_k(t+1) &= \mathrm{P}(X_{t+1} = k) \\ &= \sum_{j=1}^{n} \mathrm{P}(X_{t+1} = k \mid X_t = j)\, \mathrm{P}(X_t = j) \\ &= \sum_{j=1}^{n} P_{j \to k} u_j(t) \; . \end{aligned}$$

Note that $P_{j\to k} = P_{jk}$ is the (k, j) entry of P^t. The abstract derivation is useful because it applies in situations where a more concrete derivation is "challenging".

We revisit another easy case. Suppose the Markov chain with state space $\mathbb{R}^n$ has a well defined transition density $P(x, y)$, which is the probability density for $X_{t+1} = y$ given that $X_t = x$. Let $u(x, t)$ be the PDF of X_t. We write $u(\cdot, t)$ to refer to u as a function of x for a fixed t. The pairing is

$$\langle u(\cdot, t), f\rangle = \mathrm{E}[\, f(X_t)] = \int f(x) u(x, t)\, dx \; . \tag{21}$$

The backward equation has the form (see above or go through the simple reasoning again)

$$Lf(x) = \int P(x, y) f(y)\, dy \; .$$

We substitute this into the pairing integral formula to find an expression for $L^* u$:

$$\begin{aligned} \langle u(\cdot, t), Lf\rangle &= \int u(x, t)\, (Lf)\, (x)\, dx \\ &= \int u(x, t) \left(\int P(x, y) f(y) dy \right)\, dx \\ &= \int \int u(x, t) P(x, y) f(y)\, dx dy \\ &= \int \left(\int P(x, y) u(x, t)\, dx \right)\, f(y)\, dy \; . \end{aligned}$$

This gives an integral formula for $L^* u$, which is

$$L^* u(\cdot, t)(y) = \int P(x, y) u(x, t)\, dx \; . \tag{22}$$

We take the adjoint of an integral operator, with respect to this pairing, by integrating over the first argument of P rather than the second. The is the integral analogue of the discrete case, where we sum over the first index of the matrix P rather than the second.

The forward equation comes with *initial conditions* that play the role of final conditions in the backward equation. The initial conditions give the probability distribution at a certain time, say, $t_0 = 0$. The forward equation then determines the probability distribution at all later times $t > t_0$. The probability distribution at earlier times $t < t_0$ is not determined by the forward equation because the operator that would do it,

$$u(t) = (L^*)^{-1} u(t+1) ,$$

might not exist.

1.4 The forward equation in continuous time

The pairings depend on the state space, not the nature of time (continuous vs. discrete). Therefore, the continuous time forward equation for a finite state space Markov process is

$$\partial_t u(t) = R^t u(t) ,$$

where R is the transition rate matrix, and $u(t) \in \mathbb{R}^n$ is the vector of probabilities $u_j(t) = \mathrm{P}(X_t = j)$.

The generic forward equation in continuous time is a consequence of the constancy of expected values, as it was in the discrete time case. Here, we have

$$\mathrm{E}[\,V(X_T)] = \int \mathrm{E}[\,V(X_T) \mid X_t = x]\, u(x,t)\, dx = \langle u(\cdot,t), f(\cdot,t)\rangle \ .$$

The left side is independent of t, so the right side is too. We differentiate with respect to t:

$$\begin{aligned}
0 &= \partial_t \langle u(t), f(t)\rangle \\
&= \langle \partial_t u(t), f\rangle + \langle u(t), \partial_t f(t)\rangle \\
&= \langle \partial_t u(t), f\rangle - \langle u(t), Lf(t)\rangle \\
&= \langle \partial_t u(t), f\rangle - \langle L^* u(t), f(t)\rangle \\
0 &= \langle \left[\partial_t u(t) - L^* u(t)\right], f(t)\rangle \ .
\end{aligned}$$

The natural way for this to be true always is for the quantity in square brackets to vanish. That gives

$$\partial_t u(t) = L^* u(t) \ . \tag{23}$$

This is the general forward equation. You also need initial conditions to determine $u(t)$. If $u(t_0)$ is known, then (23) determines $u(t)$ for $t > t_0$. The forward equation generally does not run "backwards". It is difficult or impossible to find $u(t)$ for $t < t_0$ from $u(t_0$, in general.

The main point of this section is to find the forward equation for diffusion processes. If the backward equation is

$$\partial_t f + \frac{1}{2}\mu(x)\partial_x^2 f + a(x)\partial_x f = 0 \ , \tag{24}$$

what is the forward equation. You can look at this PDE, or you look back at our derivation of the generator of a diffusion process (where we used σ^2 instead of μ) and see that the generator is given by

$$Lf(x) = a(x)\partial_x f + \frac{1}{2}\mu(x)\partial_x^2 f \ .$$

You get an expression for the generator itself by leaving f out of this formula

$$L = a(x)\partial_x + \frac{1}{2}\mu(x)\partial_x^2 \ .$$

A generator like this is called a linear *differential operator*.

We integrate by parts to calculate the adjoint of a linear differential operator with respect to the integral pairing (21). The boundary terms in this process are zero. We don't give a proof (surprise), but the reason is that the probability density should go to zero so rapidly as $|x| \to \infty$ that even if $f(x) \to \infty$, it will do so in a way that the boundary terms are zero in the limit. For example,

$$\begin{aligned}\int_{-\infty}^{\infty} u(x)\partial_x f(x)\, dx &= \lim_{R\to\infty}\int_{-R}^{R} u(x)\partial_x f(x)\, dx \\ &= \lim_{R\to\infty}\left[u(x)f(x)\Big|_{-R}^{R} - \int_{-R}^{R} \left(\partial_x u(x)\right) f(x) dx \right] .\end{aligned}$$

If $u(x)$ goes to zero exponentially as $|x| \to \infty$, then even if $f(x)$ grows linearly at infinity, still

$$u(R)f(R) \to 0 \text{ as } R \to \pm\infty \ .$$

Therefore,

$$\begin{aligned}\langle u, Lf\rangle &= \int_{-\infty}^{\infty} u(x)\left(a(x)\partial_x f(x) + \frac{1}{2}\mu(x)\partial_x^2 f(x)\right) dx \\ &= \int_{-\infty}^{\infty} \left(-\partial_x(a(x)u(x))\right) f(x)\, dx + \frac{1}{2}\int_{-\infty}^{\infty}\left(\partial_x^2(\mu(x)u(x))\right) f(x)\, dx \\ &= \int_{-\infty}^{\infty}\Big[-\partial_x(a(x)u(x)) + \frac{1}{2}\partial_x^2(\mu(x)u(x))\Big] f(x)\, dx \ .\end{aligned}$$

This is a pairing of the quantity in square braces $\big[\cdots\big]$ with f. That identifies the quantity in square braces as L^*u:

$$L^*u(x) = -\partial_x(a(x)u(x)) + \frac{1}{2}\partial_x^2(\mu(x)u(x)) \ . \tag{25}$$

This formula for L^* gives a specific form to the forward equation (23) for a diffusion equation:

$$\partial_t u(x,t) = -\partial_x\big(a(x)u(x,t)\big) + \frac{1}{2}\partial_x^2\big(\mu(x)u(x,t)\big) \; . \tag{26}$$

I don't think there is an easier derivation of this equation.

The backward and forward equations have similarities and differences. Both involve second derivatives in "space" and first derivatives in time. However, the relative signs of the second derivatives is reversed. If $\mu = 1$, the backward equation has $\partial_t f = -\frac{1}{2}\partial_x^2 f$, while the forward equation has $\partial_t u = \frac{1}{2}\partial_x^2 u$. The operator L has the same sign for the ∂_x and ∂_x^2 terms. The operator L^* has opposite signs. The backward equation does not differentiate coefficients, but only the unknown function. The forward equation has

$$\partial_x(a(x)u(x,t)) = (\partial_x a(x))u(x,t) + a(x)\partial_x u(x,t) \; .$$

We work through forward equation mechanics in the Ornstein Uhlenbeck example. The process is

$$dX_t = -\gamma X_t dt + \sigma dW_t \; . \tag{27}$$

The generator is $L = -\gamma\partial_x + \frac{\sigma^2}{2}\partial_x^2$. The adjoint is defined by $L^*u = \gamma\partial_x(xu(x)) + \frac{\sigma^2}{2}\partial_x^2 u(x)$. The forward equation (26) specializes to

$$\partial_t u(x,t) = \gamma\partial_x(xu(x,t)) + \frac{\sigma^2}{2}\partial_x^2 u(x,t) \; . \tag{28}$$

Since (27) is a linear Gaussian process, if $X_0 = 0$, then X_t will be a mean zero Gaussian for any $t > 0$. As a Gaussian, the density $u(x,t)$ is completely determined by the mean (which is zero in this example) and variance, $v(t) = \mathrm{E}\big[\,X_t^2\big]$. We compute the dynamics of the variance by Ito calculus:

$$\begin{aligned}
dv(t) &= d\mathrm{E}\big[\,X_t^2\big] \\
&= \mathrm{E}\big[\,d(X_t^2)\big] \\
&= \mathrm{E}[\,2X_t dX_t] + \mathrm{E}\Big[\,(dX_t)^2\Big] \\
&= -2\gamma\mathrm{E}\big[\,X_t^2\,dt\big] + \sigma^2 dt \\
&= -2\gamma v(t)dt + \sigma^2 dt \; .
\end{aligned}$$

In the notation of differential equations, this is written

$$\frac{d}{dt}\,v = -2\gamma v + \sigma^2 \; . \tag{29}$$

The steady state is $\dot{v} = 0$, which leads to $-2v_\infty + \sigma^2 = 0$, and then to $v_\infty = \frac{\sigma^2}{2\gamma}$. Dynamically, the decay coefficient being 2γ, the variance will approach its limiting variance with a rate constant 2γ. If $v(0) = 0$, that gives

$$v(t) = v_\infty\left(1 - e^{-2\gamma t}\right) = \frac{\sigma^2}{2\gamma}\left(1 - e^{-2\gamma t}\right) \; . \tag{30}$$

You should be able to plug the solution formula (30) into the differential equation (29) to see that it is a solution.

We can construct the corresponding solution to the forward equation using the ansatz method. A mean zero Gaussian with variance $v(t)$ has PDF

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

Since the PDE (28) is linear, we can leave out the common 2π factor for simplicity. The derivatives are

$$v(t)^{-1/2} e^{\frac{-1}{2}x^2 v(t)^{-1}} \xrightarrow{\partial_t} -\frac{1}{2}\dot{v}v^{-3/2}e^{--} + \frac{1}{2}v^{-5/2}x^2\dot{v}e^{--} \ ,$$

and

$$\begin{aligned} v(t)^{-1/2} e^{\frac{-1}{2}x^2 v(t)^{-1}} &\xrightarrow{\partial_x} v^{-3/2}(-x)e^{-x^2/2v} \\ &\xrightarrow{\partial_x} -v^{-3/2}e^{--} + v^{-5/2}x^2e^{--} \ , \end{aligned}$$

and

$$xv^{-1/2} e^{\frac{-1}{2}x^2 v(t)^{-1}} \xrightarrow{\partial_x} v^{-1/2}e^{--} - v^{-3/2}x^2e^{--} \ .$$

Substitute these back into the forward equation (28), leave out the common exponential factors, and you get

$$-\frac{1}{2}\dot{v}v^{-3/2} + \frac{1}{2}v^{-5/2}x^2\dot{v} = \gamma\left(v^{-1/2} - v^{-3/2}x^2\right) + \frac{\sigma^2}{2}\left(-v^{-3/2} + v^{-5/2}x^2\right) \ .$$

According to the ansatz recipe, we now collect the constant terms and the coefficients of x^2 from both sides. It turns out that these equations are the same, and both of them are (29). This shows that the variance calculation above actually constructs a Gaussian solution to the forward equation. As $t \to \infty$, the solution converges to a steady state

$$u(x,t) \to u_\infty(x) = \frac{1}{\sqrt{2\pi v_\infty}} e^{-x^2/2v_\infty} \quad \text{as} \quad t \to \infty \ .$$

For later reference, we describe the solution in case the initial condition has $v_0 \gg v_\infty$. The v dynamics (29) tells that $v(t)$ decreases from $v(0)$ to the limiting value v_∞. The maximum value of $u(x,t)$ is taken at $x = 0$, and is equal to $(2\pi v(t))^{-1/2}$. This starts at a small value and increases over time to $(2\pi v_\infty)^{-1/2}$. We can think of the density $u(x,t)$ as representing a cloud of many particles independently following the stochastic dynamics (27). At the starting time $t = 0$, this cloud is widely dispersed, because $v(0)$ is large. The probability density is small at any given point because the number of particles per unit distance is small. As time goes on, the variance decreases and the particles become more tightly concentrated near zero. The large time concentration is determined by the limiting variance v_∞. The probability density is larger because there are more particles per unit distance than before.

It is useful to formulate the forward equation in terms of a *probability flux*. We encountered this idea before when talking about hitting times. The probability flux in this case is

$$F(x,t) = a(x)u(x,t) - \frac{1}{2}\partial_x \left(\mu(x)u(x,t)\right) \ . \tag{31}$$

The forward equation takes the following form, which is called *conservation form*:

$$\partial_t u(x,t) + \partial_x F(x,t) = 0 \ . \tag{32}$$

The conservation equation (32) states that

$$\frac{d}{dt}\int_a^b u(x,t)dx = -F(b,t) + F(a,t) \ .$$

This says that $F(b,t)$ is the rate at which "probability" is leaving the interval $[a,b]$. It is a "flux" of particles, moving to the right if $F(b,t) > 0$, and to the left if $F(b,t) < 0$. This flux is made up of two parts, the advective, or drift, part $a(x)u(x,t)$, and the noise or fluctuation part $\frac{1}{2}\partial_x \left(\mu(x)u(x,t)\right)$.