09 Apr 2021

REVERSE TIME STOCHASTIC DIFFERENTIAL EQUATIONS [FOR GENERATIVE MODELLING]

'If I Could Turn Back Time' by Cher (1989)

$$dX_\tau = \left(-\mu(x_{1-\tau}) + \frac{1}{p(x_{1-\tau})}\partial_{x_{1-\tau}}\left[\sigma^2(x_{1-\tau})\,p(x_{1-\tau})\right]\right) d\tau + \sigma(x_{1-\tau})dW_\tau$$

What follows is a derivation of the main result of 'Reverse-Time Diffusion Equation Models' by Brian D.O. Anderson (1982). Earlier on this blog we learned that a stochastic differential equation of the form

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t \tag{1}$$

with the derivative of Wiener process W_t admits two types of equations, called the forward Kolmogorov or Fokker-Planck equation and the backward Kolmogorov equation. The details of the derivation of the forward and backward Kolmogorov equations via the Kramers-Moyal expansion can be found in the previous blog post. For notational brevity we will use the term $\mu(x_t)$ for the drift and $\sigma(x_t)$ as the diffusion parameter and omit the explicit time dependency.

The Kolmogorov forward equation is identical to the Fokker Planck equation and states

$$\partial_t p(x_t) = -\partial_{x_t}\left[\mu(x_t)p(x_t)\right] + \frac{1}{2}\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]. \tag{2}$$

It describes the evolution of a probability distribution $p(x_t)$ forward in time. We can quite frankly think of it as, for example, a Normal distribution being slowly transformed into an arbitrary complex distribution according to the drift and diffusion parameters $\mu(x_t)$ and $\sigma(x_t)$.

The Kolmogorov backward equation for $s \geq t$ is defined as

$$-\partial_t p(x_s|x_t) = \mu(x_t)\,\partial_{x_t} p(x_s|x_t) + \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t) \tag{3}$$

and it basically answers the question how the probability of x_s at a later point in time changes as we change x_t at an earlier point in time. The Kolmogorov backward equation is somewhat confounding with respect to time as we're taking the partial derivative with respect to the earlier time step t on which we are also coniditoning. But we can think of it as asking 'How does the probability of x_s at the later point in time s change, as we slowly evolve the probability distribution backwards through time and condition on x_t'.

Taking inspiration from our crude example earlier, the backward equation offers a partial differential equation which we can solve backward in time, which would correspond to evolving the arbitrarily complex distribution backwards to our original Normal distribution. Unfortunately there is no corresponding stochastic differential equation with a drift and diffusion term that describes the evolution of a random variable backwards through time in terms of a stochastic differential equation.

This is where the remarkable result from Anderson (1982) comes into play.

The granddaddy of all probabilistic equations, Bayes theorem, tells us that a joint distribution can be factorized by conditioning: $p(x_s, x_t) = p(x_s|x_t)p(x_t)$ with the time ordering $t \leq s$. Why do we invoke the joint probability $p(x_s, x_t)$ we might ask? What we're trying to achieve is to derive a stochastic differential equation that tells us from what values of x_t we can arrive at x_s. We can ask ourselves what the partial differential equation would be that describes the evolution of the joint distribution over time. First multiplying both sides of Bayes theorem with minus one and taking the derivative with respect to time t, we obtain via the product rule

$$\begin{aligned}
-\partial_t p(x_s, x_t) &= -\partial_t\left[p(x_s|x_t)p(x_t)\right] \qquad (4)\\
&= \underbrace{-\partial_t p(x_s|x_t)}_{\text{KBE}}p(x_t) - p(x_s|x_t)\underbrace{\partial_t p(x_t)}_{\text{KFE}} \qquad (5)
\end{aligned}$$

into which we can plug in the Kolmogorov forward (KFE) and Kolmogorov backward (KBE) equations,

$$\begin{aligned}
&-\partial_t p(x_s|x_t)p(x_t) - p(x_s|x_t)\partial_t p(x_t) \qquad (6)\\
&= \left(\mu(x_t)\,\partial_{x_t} p(x_s|x_t) + \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\right) p(x_t) \qquad (7)\\
&+ p(x_s|x_t)\left(\partial_{x_t}\left[\mu(x_t)p(x_t)\right] - \frac{1}{2}\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]\right) \qquad (8)
\end{aligned}$$

The derivative occuring in the backward Kolmogorov equation are

$$\begin{aligned}
\partial_{x_t} p(x_s|x_t) &= \partial_{x_t}\left[\frac{p(x_s, x_t)}{p(x_t)}\right] \qquad (9)\\
&= \frac{\partial_{x_t} p(x_s, x_t)p(x_t) - p(x_s, x_t)\partial_{x_t} p(x_t)}{p^2(x_t)} \qquad (10)\\
&= \frac{\partial_{x_t} p(x_s, x_t)}{p(x_t)} - \frac{p(x_s, x_t)\partial_{x_t} p(x_t)}{p^2(x_t)} \qquad (11)
\end{aligned}$$

The next step is to evaluate the derivative of the products in the forward Kolmogorov equation.

$$\begin{aligned}
\partial_{x_t}\left[\mu(x_t)p(x_t)\right] &= \partial_{x_t}\mu(x_t)\,p(x_t) + \mu(x_t)\,\partial_{x_t} p(x_t) \qquad (12)\\
\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] &= \partial_{x_t}^2\sigma^2(x_t)\,p(x_t) + 2\,\partial_{x_t}\sigma^2(x_t)\,\partial_{x_t} p(x_t) + \sigma^2(x_t)\,\partial_{x_t}^2 p(x_t) \qquad (13)
\end{aligned}$$

Substituting the derivatives of the probability distributions accordingly we obtain

$$\begin{aligned}
-\partial_t p(x_s, x_t) =& -\partial_t\left[p(x_s|x_t)p(x_t)\right] \qquad (14)\\
=& -\partial_t p(x_s|x_t)p(x_t) - p(x_s|x_t)\partial_t p(x_t) \qquad (15)\\
=& \left(\mu(x_t)\,\partial_{x_t} p(x_s|x_t) + \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\right) p(x_t) \qquad (16)\\
&+ p(x_s|x_t)\left(\partial_{x_t}\left[\mu(x_t)p(x_t)\right] - \frac{1}{2}\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]\right) \qquad (17)\\
=& \mu(x_t)\,\partial_{x_t} p(x_s|x_t)\,p(x_t) + \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) \qquad (18)\\
&+ p(x_s|x_t)\partial_{x_t}\mu(x_t)\,p(x_t) + p(x_s|x_t)\mu(x_t)\,\partial_{x_t} p(x_t) \qquad (19)\\
&- \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (20)\\
=& \mu(x_t)\left(\frac{\partial_{x_t} p(x_s, x_t)}{\cancel{p(x_t)}} - \frac{p(x_s, x_t)\partial_{x_t} p(x_t)}{p^{\cancel{2}}(x_t)}\right)\cancel{p(x_t)} \qquad (21)\\
&+ p(x_s|x_t)\partial_{x_t}\mu(x_t)\,p(x_t) + p(x_s|x_t)\mu(x_t)\,\partial_{x_t} p(x_t) \qquad (22)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) - \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (23)\\
=& \mu(x_t)\left(\partial_{x_t} p(x_s, x_t) - \frac{p(x_s, x_t)\partial_{x_t} p(x_t)}{p(x_t)}\right) \qquad (24)\\
&+ p(x_s|x_t)\partial_{x_t}\mu(x_t)\,p(x_t) + p(x_s|x_t)\mu(x_t)\,\partial_{x_t} p(x_t) \qquad (25)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) - \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (26)\\
=& \mu(x_t)\left(\partial_{x_t} p(x_s, x_t) - \cancel{p(x_s|x_t)\partial_{x_t} p(x_t)}\right) \qquad (27)\\
&+ p(x_s, x_t)\partial_{x_t}\mu(x_t) + \cancel{p(x_s|x_t)\mu(x_t)\,\partial_{x_t} p(x_t)} \qquad (28)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) - \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (29)\\
=& \underbrace{\mu(x_t)\,\partial_{x_t} p(x_s, x_t) + p(x_s, x_t)\partial_{x_t}\mu(x_t)}_{\text{product rule}} \qquad (30)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) - \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (31)\\
=& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] \qquad (32)\\
&+ \underbrace{\frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t)}_{(1)} - \underbrace{\frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]}_{(2)} \qquad (33)
\end{aligned}$$

In order to transform the partial differential equation above into a form from which we can deduce an equivalent stochastic differential equation, we match the terms of the second order derivatives with the following identity,

$$\begin{aligned}
&\frac{1}{2}\partial_{x_t}^2\left[p(x_s, x_t)\sigma^2(x_t)\right] \qquad (34)\\
=&\frac{1}{2}\partial_{x_t}^2\left[p(x_s|x_t)p(x_t)\sigma^2(x_t)\right] \qquad (35)\\
=&\frac{1}{2}\partial_{x_t}^2 p(x_s|x_t)p(x_t)\sigma^2(x_t) + \partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right]\partial_{x_t} p(x_s|x_t) + \frac{1}{2}\partial_{x_t}^2\left[p(x_t)\sigma^2(x_t)\right]p(x_s|x_t) \qquad (36)\\
=&\underbrace{\frac{1}{2}\sigma^2(x_t)\partial_{x_t}^2 p(x_s|x_t)p(x_t)}_{(1)} + \partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right]\partial_{x_t} p(x_s|x_t) + \underbrace{\frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[p(x_t)\sigma^2(x_t)\right]}_{(2)} \qquad (37)
\end{aligned}$$

by observing that the terms (1) and (2) occur in both equations. We can see from the expansion of the derivative above that we can combine the terms in our derivation if we expand the "center term". Furthermore we can employ the identity $-\frac{1}{2}X = -X + \frac{1}{2}X$ to obtain

$$\begin{aligned}
-\partial_t p(x_s, x_t) =& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] \qquad (38)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t) - \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] \qquad (39)\\
=& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] \qquad (40)\\
&+ \frac{1}{2}\,\sigma^2(x_t)\,p(x_t)\,\partial_{x_t}^2 p(x_s|x_t) \underbrace{- \frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]}_{-\frac{1}{2}X = -X + \frac{1}{2}X} \qquad (41)\\
&\underbrace{\pm\partial_{x_t} p(x_s|x_t)\partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right]}_{\text{complete the square}} \qquad (42)\\
=& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] + \color{red}{\frac{1}{2}\,\sigma^2(x_t)\,\partial_{x_t}^2 p(x_s|x_t)\,p(x_t)} \qquad (43)\\
&\underbrace{-p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] + \color{red}{\frac{1}{2}p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right]}}_{-\frac{1}{2}X = -X + \frac{1}{2}X} \qquad (44)\\
&\color{red}{\pm\partial_{x_t} p(x_s|x_t)\partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right]} \qquad (45)\\
=& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] + \color{red}{\frac{1}{2}\partial_{x_t}^2\left[p(x_s|x_t)p(x_t)\sigma^2(x_t)\right]} \qquad (46)\\
&\underbrace{-p(x_s|x_t)\partial_{x_t}^2\left[\sigma^2(x_t)\,p(x_t)\right] - \partial_{x_t} p(x_s|x_t)\partial_{x_t}\left[p(x_t)\sigma^2(x_t)\right]}_{-\partial_{x_t}\left[p(x_s|x_t)\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right]\text{ (product rule)}} \qquad (47)\\
=& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t)\right] + \frac{1}{2}\partial_{x_t}^2\left[p(x_s, x_t)\sigma^2(x_t)\right] \qquad (48)\\
&- \partial_{x_t}\left[p(x_s|x_t)\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right]. \qquad (49)
\end{aligned}$$

What remains to be done is to combine the joint probability and the conditional probability in the first order derivative terms to combine them,

$$\begin{aligned}
-\partial_t p(x_s, x_t) =& \partial_{x_t}\left[\mu(x_t)\,p(x_s, x_t) - p(x_s|x_t)\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right] \qquad (50)\\
&+ \frac{1}{2}\partial_{x_t}^2\left[p(x_s, x_t)\sigma^2(x_t)\right] \qquad (51)\\
=& \partial_{x_t}\left[p(x_s, x_t)\left(\mu(x_t) - \frac{1}{p(x_t)}\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right)\right] \qquad (52)\\
&+ \frac{1}{2}\partial_{x_t}^2\left[p(x_s, x_t)\sigma^2(x_t)\right] \qquad (53)\\
=& -\partial_{x_t}\left[p(x_s, x_t)\left(-\mu(x_t) + \frac{1}{p(x_t)}\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right)\right] \qquad (54)\\
&+ \frac{1}{2}\partial_{x_t}^2\left[p(x_s, x_t)\sigma^2(x_t)\right] \qquad (55)
\end{aligned}$$

the result of which is in the form of a Kolmogorov forward equation, although using the joint probability distribution $p(x_s, x_t)$. For the time ordering of $t \leq s$, we can observe that the term $-\partial_t p(x_s, x_t)$ describes the change of the probability distribution as we move backward in time. In accordance with Leibniz' rule we can marginalize over x_s without interferring with the partial derivative ∂_t, to obtain

$$\begin{aligned}
-\partial_t p(x_t) =& -\partial_{x_t}\left[p(x_t)\left(-\mu(x_t) + \frac{1}{p(x_t)}\partial_{x_t}\left[\sigma^2(x_t)\,p(x_t)\right]\right)\right] \qquad (56)\\
&+ \frac{1}{2}\partial_{x_t}^2\left[p(x_t)\sigma^2(x_t)\right] \qquad (57)
\end{aligned}$$

and introduce the time reversal $\tau \doteq 1 - t$ which, with respect to the integration with respect to the flow of time, yields

$$\begin{aligned}
-\partial_t p(x_t) =& \partial_\tau p(x_{1-\tau}) \qquad (58)\\
=& -\partial_{x_{1-\tau}}\left[p(x_{1-\tau})\left(-\mu(x_{1-\tau}) + \frac{1}{p(x_{1-\tau})}\partial_{x_{1-\tau}}\left[\sigma^2(x_{1-\tau})\,p(x_{1-\tau})\right]\right)\right] \qquad (59)\\
&+ \frac{1}{2}\partial_{x_{1-\tau}}^2\left[p(x_{1-\tau})\sigma^2(x_{1-\tau})\right] \qquad (60)
\end{aligned}$$

which finally gives us a stochastic differential equation analogous to the Fokker-Planck/forward Kolmogorov equation that we can solve backward in time:

$$dX_\tau = \left(-\mu(x_{1-\tau}) + \frac{1}{p(x_{1-\tau})}\partial_{x_{1-\tau}}\left[\sigma^2(x_{1-\tau})\,p(x_{1-\tau})\right]\right) d\tau + \sigma(x_{1-\tau})dW_\tau \tag{61}$$

where $\tilde{W}_t$ is a Wiener process that flows backward in time.

By keeping the $\sigma^2(x_t)$ constant and independent of x_t and applying the log-derivative trick, the drift simplifies to

$$\begin{aligned}
dX_\tau &= \left(-\mu(x_{1-\tau}) + \frac{1}{p(x_{1-\tau})}\partial_{x_{1-\tau}}\left[\overbrace{\sigma^2(x_{1-\tau})}^{=\sigma^2}\,p(x_{1-\tau})\right]\right)d\tau + \sigma(x_{1-\tau})dW_\tau \qquad (62)\\
&= \left(-\mu(x_{1-\tau}) + \frac{\sigma^2}{p(x_{1-\tau})}\partial_{x_{1-\tau}}\,p(x_{1-\tau})\right) d\tau + \sigma(x_{1-\tau})dW_\tau \qquad (63)\\
&= \left(-\mu(x_{1-\tau}) + \sigma^2\partial_{x_{1-\tau}}\,\log p(x_{1-\tau})\right)dt + \sigma(x_{1-\tau})d\tilde{W}_\tau \qquad (64)
\end{aligned}$$

MY BLOG