1 Solving forward diffusion SDE

Forward diffusion SDE is given by

$$dX_t = \frac{1}{2}\Sigma^{-1}(\mu - X_t)\beta_t dt + \sqrt{\beta_t}dW_t, \quad t \in [0, T]$$

where X_t is n-dimensional stochastic process, W_t is standard n-dimensional Brownian motion, $\mu = (\mu_1 ... \mu_n)^{\mathbf{T}}$ is n-dimensional vector, Σ is $n \times n$ diagonal matrix with positive diagonal elements $\{\sigma_{ii}^2\}_1^n$ and noise schedule β_t is non-negative function $[0, T] \to \mathbb{R}^+$. Consider change of variables $Y_t = X_t - \mu$. Then we can rewrite forward diffusion SDE as

$$dY_t = -\frac{1}{2}\Sigma^{-1}Y_t\beta_t dt + \sqrt{\beta_t}dW_t$$

For every $i = 1, .., n$ we have

$$d\left(e^{\frac{1}{2\sigma_{ii}^2}\int_0^t \beta_s ds} Y_t^i\right) = e^{\frac{1}{2\sigma_{ii}^2}\int_0^t \beta_s ds} \cdot \frac{1}{2\sigma_{ii}^2}\beta_t Y_t^i dt + e^{\frac{1}{2\sigma_{ii}^2}\int_0^t \beta_s ds} \cdot \left(-\frac{1}{2\sigma_{ii}^2}Y_t^i\beta_t dt + \sqrt{\beta_t}dW_t^i\right) = e^{\frac{1}{2\sigma_{ii}^2}\int_0^t \beta_s ds}\sqrt{\beta_t}dW_t^i$$

Exponential of a diagonal matrix is just element-wise exponential, so we can rewrite it in multidimensional form as

$$d\left(e^{\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}Y_t\right) = \sqrt{\beta_t}e^{\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}dW_t \implies e^{\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}Y_t - Y_0 = \int_0^t \sqrt{\beta_s}e^{\frac{1}{2}\Sigma^{-1}\int_0^s \beta_u du}dW_s$$

or writing this down in terms of X_t

$$X_t = e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}X_0 + \left(I - e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}\right)\mu + \int_0^t \sqrt{\beta_s}e^{-\frac{1}{2}\Sigma^{-1}\int_s^t \beta_u du}dW_s \tag{1}$$

where I is $n \times n$ identity matrix.

2 Derivation of conditional distribution of $\mathrm{X_t}$

Let $A(s) = \sqrt{\beta_s}e^{-\frac{1}{2}\Sigma^{-1}\int_s^t \beta_u du}$. It is a diagonal matrix and its i-th diagonal element $a_{ii}(s)$ equals $\sqrt{\beta_s}e^{-\frac{1}{2\sigma_{ii}^2}\int_s^t \beta_u du}$. Assume $a_{ii}(s) \in L_2[0, T]$ for each i. Itô's integral $\int_0^t a_{ii}(s)dW_s^i$ is defined as the limit of integral sums when mesh of partition Δ tends to zero:

$$\int_0^t a_{ii}(s)dW_s^i = \lim_{\Delta \to 0}\sum_k a_{ii}(s_k)\Delta W_{s_k}^i \overset{d}{=} \lim_{\Delta \to 0}\mathcal{N}\left(0, \sum_k a_{ii}^2(s_k)\Delta s_k\right) \overset{d}{=} \mathcal{N}\left(0, \lim_{\Delta \to 0}\sum_k a_{ii}^2(s_k)\Delta s_k\right) = \mathcal{N}\left(0, \int_0^t a_{ii}^2(s)ds\right)$$

where the first equality in distribution holds due to the properties of Brownian motion and the fact that $a_{ii}(s_k)$ are deterministic (implying that $a_{ii}(s_k)\Delta W_{s_k}^i = a_{ii}(s_k)(W_{s_{k+1}}^i - W_{s_k}^i)$ are independent normal random variables with mean 0 and variance $a_{ii}^2(s_k)(s_{k+1} - s_k) = a_{ii}^2(s_k)\Delta s_k$) and the second equality in distribution follows from Lévy's continuity theorem (it's easy to check that the sequence of characteristic functions of random variables on the left-hand side converges point-wise to the characteristic function of the random variable on the right-hand side). Then, simple integration gives

$$\int_0^t a_{ii}^2(s)ds = \int_0^t \beta_s e^{-\frac{1}{\sigma_{ii}^2}\int_s^t \beta_u du}ds = \int_0^t \sigma_{ii}^2 d\left(e^{-\frac{1}{\sigma_{ii}^2}\int_s^t \beta_u du}\right) = \sigma_{ii}^2\left(1 - e^{-\frac{1}{\sigma_{ii}^2}\int_0^t \beta_s ds}\right)$$

It implies that in multidimensional case we have

$$\int_0^t \sqrt{\beta_s}e^{-\frac{1}{2}\Sigma^{-1}\int_s^t \beta_u du}dW_s = \int_0^t A(s)dW_s \sim \mathcal{N}\left(0, \lambda(\Sigma, t)\right), \quad \lambda(\Sigma, t) = \Sigma\left(I - e^{-\Sigma^{-1}\int_0^t \beta_s ds}\right)$$

and it follows from (1) that

$$Law(X_t|X_0) = \mathcal{N}(\rho(X_0, \Sigma, \mu, t), \lambda(\Sigma, t)), \quad \rho(X_0, \Sigma, \mu, t) = e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}X_0 + \left(I - e^{-\frac{1}{2}\Sigma^{-1}\int_0^t \beta_s ds}\right)\mu \tag{2}$$

3 Reverse dynamics

The result by Anderson (1982) implies that if n-dimensional process of the diffusion type X_t satisfies

$$dX_t = f(X_t, t)dt + g(t)dW_t, \quad t \in [0, T] \tag{3}$$

where $g(t)$ is a function $[0, T] \to \mathbb{R}$ then its reverse-time dynamics is given by

$$dX_t = (f(X_t, t) - g^2(t)\nabla \log p_t(X_t))dt + g(t)d\widetilde{W}_t, \quad t \in [0, T] \tag{4}$$

where $p_t(\cdot)$ is the probability density function of random variable X_t and $\widetilde{W}_t$ is a reverse-time standard Brownian motion such that X_t is independent of its past increments $\widetilde{W}_s - \widetilde{W}_t$ for $s < t$. Reverse-time dynamics means that all the integrals associated with reverse-time differentials have t as their lower limit (e.g. dX_t relates to $\int_t^T dX_s = X_T - X_t$). Anderson's result is obtained under the assumption that Kolmogorov equations (for probability density functions) associated with all considered processes have unique smooth solutions. On the other hand, Song et al. (2021) argued that SDE (3) has the same forward Kolmogorov equation as the following ODE:

$$dX_t = (f(X_t, t) - \frac{1}{2}g^2(t)\nabla \log p_t(X_t))dt, \quad t \in [0, T] \tag{5}$$

which means that processes following (3) and (5) are equal in distribution if they start from the same initial distribution $Law(X_0)$. In our case $f(X_t, t) = \frac{1}{2}\Sigma^{-1}(X_t - \mu)\beta_t$ and $g(t) = \sqrt{\beta_t}$, so we have two equivalent reverse diffusion dynamics:

$$dX_t = \left(\frac{1}{2}\Sigma^{-1}(X_t - \mu) - \nabla \log p_t(X_t)\right)\beta_t dt + \sqrt{\beta_t}d\widetilde{W}_t$$

and

$$dX_t = \frac{1}{2}\left(\Sigma^{-1}(X_t - \mu) - \nabla \log p_t(X_t)\right)\beta_t dt$$

where both differential equations are to be solved backwards.

4 Score estimation

If X_0 is known then (2) implies that

$$\log p_{0t}(X_t|X_0) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\det \lambda(\Sigma, t) - \frac{1}{2}(X_t - \rho(X_0, \Sigma, \mu, t))^{\mathbf{T}}\lambda(\Sigma, t)^{-1}(X_t - \rho(X_0, \Sigma, \mu, t)) \implies$$

$$\nabla \log p_{0t}(X_t|X_0) = -\lambda(\Sigma, t)^{-1}(X_t - \rho(X_0, \Sigma, \mu, t))$$

where $p_{0t}(\cdot|X_0)$ is the probability density function of conditional distribution $Law(X_t|X_0)$. So if we sample X_t by the formula $X_t = \rho(X_0, \Sigma, \mu, t) + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0, \lambda(\Sigma, t))$ then $\nabla \log p_{0t}(X_t|X_0) = -\lambda(\Sigma, t)^{-1}\epsilon_t$. In the simplified case when $\Sigma = I$ we have $\lambda(I, t) = \lambda_t I$ where $\lambda_t = 1 - e^{-\int_0^t \beta_s ds}$. In this case gradient of noisy data log-density reduces to $\nabla \log p_{0t}(X_t|X_0) = -\epsilon_t/\lambda_t$. If $\epsilon_t = \sqrt{\lambda_t}\xi_t$, then we have

$$X_t = \rho(X_0, I, \mu, t) + \sqrt{\lambda_t}\xi_t, \quad \xi_t \sim \mathcal{N}(0, I), \quad \nabla \log p_{0t}(X_t|X_0) = -\xi_t/\sqrt{\lambda_t}$$