Notes on Path integral formulation of diffusion model

Haowu Duan[1]
[1] *Physics Department, University of Connecticut,*
2152 Hillside Road, Storrs, CT 06269, USA

CONTENTS

I. PATH INTEGRAL FORMULATION OF THE FORWARD AND TIME-REVERSED PROCESS

Flow model and diffusion model can be understood as the time evolution of the initial noise toward the target data distribution through ordinary and stochastic differential equation. In this note, I want to= demonstrate that Path integral formulation of the stochastic differential equation is the appropriate framework to intuitively understand diffusion model. I do not intend to give a full introduction to diffusion model since my goal is to clarify some of the conceptual confusion I had when I was learning diffusion model, and to provide a coherent narrative for the generative process of diffusion model.

We start with the forward process,

$$d\boldsymbol{x}(t) = \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\ dt + \sigma(t)\ d\boldsymbol{W}_t \tag{1.1}$$

where,

- $\boldsymbol{x}(t)$ is the random variable that corresponds to the data distribution at time t, denoted as $P_t\Big(t, \boldsymbol{x}(t)\Big)$.
- $f\Big(t, \boldsymbol{x}(t)\Big)$ is the drift term and $g(t)$ is the scale of the noise.
- dW_t is not differentiable, but continuously, we can write $\xi_t = g(t)\ dW_t$, and demand,
$$\langle \xi^i_t \xi^j_{t'} \rangle = \mathbb{E}[\xi_t \xi_{t'}] = \sigma^2(t)\delta^{ij}\delta(t-t') \tag{1.2}$$
since Dirac delta $\delta(t-t')$ is for continuous variable, upon discretization, we will have $\delta(t-t') \to \Delta t\ \delta_{mm'}$, with $t = m\Delta t$, and $t' = m'\Delta t$.

Now we ask the following question: given an initial distribution $P_0(x(t=0))$, if we draw samples from P_0 and update each data point based on Eq. (1.1), how will the data distribute at some later time T. To answer this question, the simplest way to do is to take a tiny time step ΔT. We will work in Itô scheme, $d\boldsymbol{x}(t) \to x_{t+\Delta} - x_t$, for discretization to avoid unnecessary subtleties since my goal is intuitive picture. I will add the discuss of scheme dependence in the appendices.

A. The forward process

$$P_{\Delta t}(\boldsymbol{x}_{\Delta t}) = \int d^d\boldsymbol{h} \int d^d\boldsymbol{x}_0\ P_0(\boldsymbol{x}_0)\delta\Big(\boldsymbol{x}_{\Delta t} - \boldsymbol{x}_0 - \boldsymbol{u}\Big(0, \boldsymbol{x}_0\Big)\ \Delta t - \sigma(0)\ \sqrt{\Delta t}\ \boldsymbol{h}\Big) \times \frac{1}{\sqrt{2\pi}} \exp\{-\frac{\boldsymbol{h}^2}{2}\} \tag{1.3}$$

all this equation does is to collect all the points that will be updated from the initial distribution to a single data point $\boldsymbol{x}_{\Delta t}$. The discretized noise is $d\boldsymbol{W}_t = \sqrt{\Delta t}\ \boldsymbol{h}$, with $\boldsymbol{h} \sim \mathcal{N}(0,1)$. If we take the integral with respect to the noise, the Dirac δ will set,

$$\begin{aligned}\boldsymbol{h} =& \frac{\Delta x - u(x_0)\Delta t}{\sqrt{\Delta t}\sigma(0)} \\ =& \sqrt{\Delta t}\frac{\Delta x - u(x_0)\Delta t}{\Delta t\ \sigma(0)}\end{aligned} \tag{1.4}$$

now we can repeat the same procedure for N steps, with $\Delta t = T/N$

$$\begin{aligned}P_T(\boldsymbol{x}_T) =& \prod_{i=0}^{N-1} \int d^d\boldsymbol{h}_i \int d^d\boldsymbol{x}_i\ \delta\Big(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i - \boldsymbol{u}\Big(i\Delta t, \boldsymbol{x}_i\Big)\ \Delta t - \sigma(\boldsymbol{x}_i)\ \sqrt{\Delta t}\ \boldsymbol{h}_i\Big) \times \frac{1}{\sqrt{2\pi}} \exp\{-\frac{\boldsymbol{h}_i^2}{2}\} P_0(\boldsymbol{x}_0) \\ =& \Big(\frac{1}{\sqrt{(2\pi)^N}} \prod_{i=0}^{N-1} \int d^d\boldsymbol{x}_i\Big) \quad \times \exp\{-\sum_{i=0}^{N} \Big[\sqrt{\Delta t}\frac{\Delta x(i\Delta t) - u(x_{i\Delta t})\Delta t}{2\Delta t\ \sigma^2(i\Delta t)}\Big]^2\} P_0(\boldsymbol{x}_0) \\ \to& \int [\mathcal{D}x(t)] \exp\{-\int_0^T dt \frac{\Big|\dot{x}(t) - u(x_t)\Big|^2}{2\ \sigma^2(t)}\} P_0(\boldsymbol{x}_0)\end{aligned} \tag{1.5}$$

where we took the continuous limit,

$$\begin{aligned}\lim_{\Delta t\to 0}\boldsymbol{h}^2 &= dt\Big(\frac{\dot{x}(t)-u(x_t)}{\sigma(t)}\Big)^2\\ \lim_{\Delta t\to 0}\prod_{i=0}^{N-1}\int d^d\boldsymbol{x}_i &\to \int[\mathcal{D}x(t)]\end{aligned}\tag{1.6}$$

Eq. (1.5) is the path integral representation of the forward process. For readers who are not familiar with path integral $\int[\mathcal{D}x(t)]$ is just an expression for integration of all possible path that update data point x_0 and reach x_T. $L(x(t),\dot{x}(t)) = |\dot{x}(t)-u(x_t)|^2/\ 2\sigma^2(t)$ is known as *Onsager-Machlup function*. Why do we go through this exercise? The main reason is that we now have an expression that maps distribution to distribution not point to point, and it is completely free of noise. But the forward process is only half the battle, that we only took our target distribution and added noise on it to reach distribution $P_T(X(T))$. We still need to reproduce the target distribution through the reverse process since the motivation is building generative models here.

B. The reverse process

The backward process is governed by Time-reversed stochastic differential equation. Or instead, we want something like,

$$P_0(x_0) = \int[\mathcal{D}x(t)]\exp\{-\int_0^T dt\ \tilde{L}(x)\}P_T(\boldsymbol{x}_T)\tag{1.7}$$

there are multiple ways to derive the path integral and time-reversed stochastic differential equation. Here we use the same strategy used in the forward process, though this is not the most clever approach, it is straightforward and intuitive. We start with one tiny step again, as matter of fact, let us first write,

$$\begin{aligned}P_{\Delta t}(\boldsymbol{x}_{\Delta t}) &= \int d^d\boldsymbol{h}\int d^d\boldsymbol{x}_0\ P_0(\boldsymbol{x}_0)\delta\Big(\boldsymbol{x}_{\Delta t}-\boldsymbol{x}_0-\boldsymbol{u}\Big(0,\boldsymbol{x}_0\Big)\ \Delta t-\sigma(0)\ \sqrt{\Delta t}\ \boldsymbol{h}\Big)\times\frac{1}{\sqrt{2\pi}}\exp\{-\frac{\boldsymbol{h}^2}{2}\}\\ &= \int d^d\boldsymbol{x}_0 P(x_{\Delta t}|x_0)P_0(x_0)\end{aligned}\tag{1.8}$$

our starting point would be,

$$P_0(\boldsymbol{x}_0) = \int d^d\boldsymbol{x}_{\Delta t}P(x_0|x_{\Delta t})P_{\Delta t}(x_{\Delta t})\tag{1.9}$$

Bayes's theorem tells us that,

$$P(A|B)P(B) = P(B|A)P(A)\tag{1.10}$$

as a result,

$$P(x_0|x_{\Delta t}) = \frac{P(x_{\Delta t}|x_0)P_0(x_0)}{P_\Delta(x_{\Delta t})}\tag{1.11}$$

we plug this result back into Eq. (1.9),

$$\begin{aligned}P_0(\boldsymbol{x}_0) &= \int d^d\boldsymbol{x}_{\Delta t}\frac{P(\boldsymbol{x}_{\Delta t}|\boldsymbol{x}_0)P_0(\boldsymbol{x}_0)}{P_\Delta(\boldsymbol{x}_{\Delta t})}P_{\Delta t}(\boldsymbol{x}_{\Delta t})\\ &= \int d^d\boldsymbol{h}\int d^d\boldsymbol{x}_{\Delta t}\delta\Big(\boldsymbol{x}_{\Delta t}-\boldsymbol{x}_0-\boldsymbol{u}\Big(0,\boldsymbol{x}_0\Big)\ \Delta t-\sigma(0)\ \sqrt{\Delta t}\ \boldsymbol{h}\Big)\times\frac{1}{\sqrt{2\pi}}\exp\{-\frac{\boldsymbol{h}^2}{2}\}\frac{P_0(\boldsymbol{x}_0)}{P_\Delta(\boldsymbol{x}_{\Delta t})}\end{aligned}\tag{1.12}$$

Take multiple iteration and $\Delta\to 0$ limit, we get,

$$P_0(\boldsymbol{x}_0) = \int[\mathcal{D}x(t)]\exp\{-\int_0^T dt\frac{\Big|\dot{x}(t)-u(\boldsymbol{x}_t)\Big|^2}{2\ \sigma^2(t)}-\int_0^T d\ln P_t(x_t)\}P_{\Delta t}(\boldsymbol{x}_{\Delta t})\tag{1.13}$$

To calculate the total derivative of $\ln P_t(x_t)$, we need to use *Itô's Lemma*,

Lemma. *Let $\boldsymbol{x}(t)$ be a stochastic process satisfying the stochastic differential equation,*

$$d\boldsymbol{x}(t) = \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\ dt + \sigma(t)\ d\boldsymbol{W}_t \tag{1.14}$$

where W_t is a Wiener process, $\boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)$ is the drift, and $\sigma(t)$ is the noise scale. If $f(t, \boldsymbol{x}(t))$ is a twice-differentiable function, then the differential of $f(t, \boldsymbol{x}(t))$ is given by

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\nabla_{\boldsymbol{x}} f + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} f\right) dt + \sigma(t)(\nabla_{\boldsymbol{x}} f)\cdot d\boldsymbol{W}_t. \tag{1.15}$$

as a result,

$$\begin{aligned}
\frac{d\ln P_t(x_t)}{dt} &= \frac{\partial \ln P_t(x_t)}{\partial t} + \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\nabla_{\boldsymbol{x}} \ln P_t(x_t) + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} \ln P_t(x_t) \\
&= \frac{1}{P_t(x_t)}\frac{\partial P_t(x_t)}{\partial t} + \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\frac{1}{P_t(x_t)}\nabla_{\boldsymbol{x}} P_t(x_t) + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} \ln P_t(x_t) \\
&= \frac{1}{P_t(x_t)}\Big[-\nabla_{\boldsymbol{x}}\boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)P_t(x_t) + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} P_t(x_t)\Big] + \boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big)\frac{1}{P_t(x_t)}\nabla_{\boldsymbol{x}} P_t(x_t) + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} \ln P_t(x_t) \\
&= -\nabla_{\boldsymbol{x}}\boldsymbol{u}\Big(t, \boldsymbol{x}(t)\Big) + \frac{\sigma^2(t)}{2}\frac{1}{P_t(x_t)}\nabla^2_{\boldsymbol{x}} P_t(x_t) + \frac{1}{2}\sigma(t)^2\nabla^2_{\boldsymbol{x}} \ln P_t(x_t) \\
&=
\end{aligned} \tag{1.16}$$

note that,

$$\begin{aligned}
\frac{1}{P_t(x_t)}\nabla^2_{\boldsymbol{x}} P_t(x_t) =& \nabla_{\boldsymbol{x}}\cdot\left(\frac{1}{P_t(x_t)}\nabla_{\boldsymbol{x}} P_t(x_t)\right) - \left(\nabla_{\boldsymbol{x}}\cdot\frac{1}{P_t(x_t)}\right)\Big(\nabla_{\boldsymbol{x}} P_t(x_t)\Big) \\
=& \nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}} \ln P_t(x_t) -
\end{aligned} \tag{1.17}$$