8 生成扩散模型漫谈（六）：一般框架之ODE篇

Aug By 苏剑林 | 2022-08-08 | 105018位读者 引用

上一篇文章《生成扩散模型漫谈（五）：一般框架之SDE篇》中，我们对宋飏博士的论文《Score-Based Generative Modeling through Stochastic Differential Equations》做了基本的介绍和推导。然而，顾名思义，上一篇文章主要涉及的是原论文中SDE相关的部分，而遗留了被称为“概率流ODE（Probability flow ODE）”的部分内容，所以本文对此做个补充分享。

事实上，遗留的这部分内容在原论文的正文中只占了一小节的篇幅，但我们需要新开一篇文章来介绍它，因为笔者想了很久后发现，该结果的推导还是没办法绕开Fokker-Planck方程，所以我们需要一定的篇幅来介绍Fokker-Planck方程，然后才能请主角ODE登场。

再次反思

我们来大致总结一下上一篇文章的内容：首先，我们通过SDE来定义了一个前向过程（“拆楼”）：

$$d\boldsymbol{x} = \boldsymbol{f}_t(\boldsymbol{x})dt + g_t d\boldsymbol{w} \tag{1}$$

然后，我们推导了相应的逆向过程的SDE（“建楼”）：

$$d\boldsymbol{x} = \left[\boldsymbol{f}_t(\boldsymbol{x}) - g_t^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] dt + g_t d\boldsymbol{w} \tag{2}$$

最后，我们推导了用神经网络$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$来估计$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$的损失函数（得分匹配）：

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_0)}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\right\|^2\right] \tag{3}$$

至此，我们完成了扩散模型的训练、预测的一般框架，可以说，它是DDPM的非常一般化的推广了。但正如《生成扩散模型漫谈（四）：DDIM = 高观点DDPM》中介绍的

DDIM是DDPM的高观点反思结果，SDE作为DDPM的推广，有没有相应的“高观点反思结果”呢？有，其结果就是本文主题“概率流ODE”。

Dirac函数

DDIM做了什么反思呢？很简单，DDIM发现DDPM的训练目标主要跟$p(\boldsymbol{x}_t|\boldsymbol{x}_0)$有关，而跟$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$无关，所以它以$p(\boldsymbol{x}_t|\boldsymbol{x}_0)$为出发点，去推导更一般的$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$和$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)$。概率流ODE做的反思是类似的，它想知道在SDE框架中，对于固定的$p(\boldsymbol{x}_t)$，能找出哪些不同的$p(\boldsymbol{x}_{t+\Delta t}|\boldsymbol{x}_t)$（或者说找到不同的前向过程SDE）。

我们先写出前向过程(1)的离散形式

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \boldsymbol{f}_t(\boldsymbol{x}_t)\Delta t + g_t\sqrt{\Delta t}\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}) \tag{4}$$

这个等式描述的是随机变量$\boldsymbol{x}_{t+\Delta t},\boldsymbol{x}_t,\boldsymbol{\varepsilon}$之间的关系，我们可以方便地对两边求期望，然而我们并非想求期望，而是想求分布$p(\boldsymbol{y}_t)$（所满足的关系式）。怎么将分布转换成期望形式呢？答案是Dirac函数：

$$p(\boldsymbol{x}) = \int \delta(\boldsymbol{x}-\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y} = \mathbb{E}_{\boldsymbol{y}}[\delta(\boldsymbol{x}-\boldsymbol{y})] \tag{5}$$

Dirac函数严格定义是属于泛函分析的内容，但我们通常都是当它是普通函数来处理，一般都能得到正确的结果。由上式还可以得知，对于任意$f(\boldsymbol{x})$，成立

$$p(\boldsymbol{x})f(\boldsymbol{x}) = \int \delta(\boldsymbol{x}-\boldsymbol{y})p(\boldsymbol{y})f(\boldsymbol{y})d\boldsymbol{y} = \mathbb{E}_{\boldsymbol{y}}[\delta(\boldsymbol{x}-\boldsymbol{y})f(\boldsymbol{y})] \tag{6}$$

直接对上式两边求偏导数，得到

$$\nabla_{\boldsymbol{x}}[p(\boldsymbol{x})f(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{y}}\left[\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x}-\boldsymbol{y})f(\boldsymbol{y})\right] = \mathbb{E}_{\boldsymbol{y}}\left[f(\boldsymbol{y})\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x}-\boldsymbol{y})\right] \tag{7}$$

这是后面要用到的性质之一，可以发现它本质上是狄拉克函数的导数能够通过积分转移到所乘函数上去。

F-P方程

经过上述铺垫，现在我们根据式(4)写出

$$\begin{aligned}
&\delta(\boldsymbol{x} - \boldsymbol{x}_{t+\Delta t}) \\
=&\delta(\boldsymbol{x} - \boldsymbol{x}_t - \boldsymbol{f}_t(\boldsymbol{x}_t)\Delta t - g_t\sqrt{\Delta t}\boldsymbol{\varepsilon}) \\
\approx&\delta(\boldsymbol{x} - \boldsymbol{x}_t) - \left(\boldsymbol{f}_t(\boldsymbol{x}_t)\Delta t + g_t\sqrt{\Delta t}\boldsymbol{\varepsilon}\right)\cdot\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2}\left(g_t\sqrt{\Delta t}\boldsymbol{\varepsilon}\cdot\nabla_{\boldsymbol{x}}\right)^2\delta(\boldsymbol{x} -
\end{aligned}$$

这里当$\delta(\cdot)$是普通函数那样做了泰勒展开，只保留了不超过$\mathcal{O}(\Delta t)$的项。现在我们两边求期望：

$$\begin{aligned}
&p_{t+\Delta t}(\boldsymbol{x}) \\
=&\mathbb{E}_{\boldsymbol{x}_{t+\Delta t}}\left[\delta(\boldsymbol{x} - \boldsymbol{x}_{t+\Delta t})\right] \\
\approx&\mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{\varepsilon}}\left[\delta(\boldsymbol{x} - \boldsymbol{x}_t) - \left(\boldsymbol{f}_t(\boldsymbol{x}_t)\Delta t + g_t\sqrt{\Delta t}\boldsymbol{\varepsilon}\right)\cdot\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2}\left(g_t\sqrt{\Delta t}\boldsymbol{\varepsilon}\cdot\nabla_{\boldsymbol{x}}\right)^2\delta(\boldsymbol{x}\right. \\
=&\mathbb{E}_{\boldsymbol{x}_t}\left[\delta(\boldsymbol{x} - \boldsymbol{x}_t) - \boldsymbol{f}_t(\boldsymbol{x}_t)\Delta t\cdot\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2}g_t^2\Delta t\nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}}\delta(\boldsymbol{x} - \boldsymbol{x}_t)\right] \\
=&p_t(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}\cdot\left[\boldsymbol{f}_t(\boldsymbol{x})\Delta t\, p_t(\boldsymbol{x})\right] + \frac{1}{2}g_t^2\Delta t\nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}}p_t(\boldsymbol{x})
\end{aligned}$$

两边除以Δt，并取$\Delta t\to 0$的极限，结果是

$$\frac{\partial}{\partial t}p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}}\cdot\left[\boldsymbol{f}_t(\boldsymbol{x})p_t(\boldsymbol{x})\right] + \frac{1}{2}g_t^2\nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}}p_t(\boldsymbol{x}) \tag{10}$$

这就是式(1)所对应的“F-P方程”（Fokker-Planck方程），它是描述边际分布的偏微分方程。

等价变换

大家看到偏微分方程不用担心，因为这里并没有打算去研究怎么求解偏微分方程，只是借助它来引导一个等价变换而已。对于任意满足$\sigma_t^2 \le g_t^2$的函数σ_t，F-P方程(10)完全等价于

$$\begin{aligned}\frac{\partial}{\partial t}p_t(\boldsymbol{x}) =& -\nabla_{\boldsymbol{x}}\cdot\left[\boldsymbol{f}_t(\boldsymbol{x})p_t(\boldsymbol{x}) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_{\boldsymbol{x}}p_t(\boldsymbol{x})\right] + \frac{1}{2}\sigma_t^2\nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}}p_t(\boldsymbol{x})\\ =& -\nabla_{\boldsymbol{x}}\cdot\left[\left(\boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right)p_t(\boldsymbol{x})\right] + \frac{1}{2}\sigma_t^2\nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{x}}\end{aligned}$$

形式上该F-P方程又相当于原来的F-P的$\boldsymbol{f}_t(\boldsymbol{x})$换成了$\boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})$、$g_t$换成了$\sigma_t$，根据式$(10)$对应于式$(1)$，上式则对应于

$$d\boldsymbol{x} = \left(\boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right)dt + \sigma_t d\boldsymbol{w} \tag{12}$$

但是别忘了式(10)跟式(11)是完全等价的，所以这意味着式(1)和式(12)这两个随机微分方程所对应的边际分布$p_t(\boldsymbol{x})$是完全等价的！这个结果告诉我们存在不同方差的前向过程，它们产生的边际分布是一样的。这个结果相当于DDIM的升级版，后面我们还会证明，当$\boldsymbol{f}_t(\boldsymbol{x})$是关于$\boldsymbol{x}$的线性函数时，它就完全等价于DDIM。

特别地，根据上一篇SDE的结果，我们可以写出式(12)对应的反向SDE：

$$d\boldsymbol{x} = \left(\boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}(g_t^2 + \sigma_t^2)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right)dt + \sigma_t d\boldsymbol{w} \tag{13}$$

神经ODE

式(12)允许我们改变采样过程的方差，这里我们特别考虑$\sigma_t = 0$的极端情形，此时SDE退化为ODE（常微分方程）：

$$d\boldsymbol{x} = \left(\boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}g_t^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})\right)dt \tag{14}$$

这个ODE称为“概率流ODE（Probability flow ODE）”，由于实践中的$\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})$需要用神经网络$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$近似，所以上式也对应一个“神经ODE”。

为什么要特别研究方差为0的情形呢？因为此时传播过程不带噪声，从$\boldsymbol{x}_0$到$\boldsymbol{x}_T$是一个确定性变换，所以我们直接反向求解ODE就能得到由$\boldsymbol{x}_T$变换为$\boldsymbol{x}_0$的逆变换，这也是一个确定性变换（直接在式(13)中代入$\sigma_t = 0$也可以发现前向和反向的方程是一样的）。这个过程和flow模型是一致的（即通过一个可逆的变换将噪声变换成样本），所以概率流ODE允许我们将扩散模型的结果与flow模型相关结果对应起来，比如原论文提到概率流ODE允许我们做精确的似然计算、获得隐变量表征等，这些本质上都是flow模型的好处。由于flow模型的可逆性，它还允许我们在隐变量空间对原图做一些图片编辑等。

另一方面，从$\boldsymbol{x}_T$到$\boldsymbol{x}_0$的变换由一个ODE描述，这意味着我们可以通过各种高阶的ODE数值算法来加速从$\boldsymbol{x}_T$到$\boldsymbol{x}_0$的变换过程。当然，原则上SDE的求解也有一些加速方法，但SDE的加速研究远远不如ODE的容易和深入。总的来说，相比SDE，ODE在理论分析和实际求解中都显得更为简单直接。

回顾DDIM

在《生成扩散模型漫谈（四）：DDIM = 高观点DDPM》的最后，我们推导了DDIM的连续版本对应于ODE

$$\frac{d}{ds}\left(\frac{\boldsymbol{x}(s)}{\bar{\alpha}(s)}\right) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}(s), t(s)\right)\frac{d}{ds}\left(\frac{\bar{\beta}(s)}{\bar{\alpha}(s)}\right) \tag{15}$$

接下来我们可以看到，该结果其实就是本文的式(14)在$\boldsymbol{f}_t(\boldsymbol{x})$取线性函数$f_t\boldsymbol{x}$时的特例：在《生成扩散模型漫谈（五）：一般框架之SDE篇》的末尾，我们推导过对应的关系

$$\begin{cases} f_t = \dfrac{1}{\bar{\alpha}_t}\dfrac{d\bar{\alpha}_t}{dt} \\ g^2(t) = 2\bar{\alpha}_t\bar{\beta}_t\dfrac{d}{dt}\left(\dfrac{\bar{\beta}_t}{\bar{\alpha}_t}\right) \\ \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = -\dfrac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)}{\bar{\beta}_t} \end{cases} \tag{16}$$

将这些关系代入到式(14)【$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$替换为$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$】后，整理得到

$$\frac{1}{\bar{\alpha}_t}\frac{d\boldsymbol{x}}{dt} - \frac{\boldsymbol{x}}{\bar{\alpha}_t^2}\frac{d\bar{\alpha}_t}{dt} = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\frac{d}{dt}\left(\frac{\bar{\beta}_t}{\bar{\alpha}_t}\right) \tag{17}$$

左端可以进一步整理得到$\frac{d}{dt}\left(\frac{\boldsymbol{x}}{\bar{\alpha}_t}\right)$，因此上式跟式(15)完全等价。

文章小结

本文在SDE篇的基础上，借助F-P方程推导了更一般化的前向方程，继而推导出了“概率流ODE”，并证明了DDIM是它的一个特例。

转载到请包括本文地址：https://spaces.ac.cn/archives/9228

更详细的转载事宜请参考：《科学空间FAQ》

如果您需要引用本文，请参考：

苏剑林. (Aug. 08, 2022). 《生成扩散模型漫谈（六）：一般框架之ODE篇 》[Blog post]. Retrieved from https://spaces.ac.cn/archives/9228

```
@online{kexuefm-9228,
    title={生成扩散模型漫谈（六）：一般框架之ODE篇},
    author={苏剑林},
    year={2022},
    month={Aug},
    url={\url{https://spaces.ac.cn/archives/9228}},
}
```