随机微分方程 SDE

随机微分：在DDPM中扩散过程被划分为固定的T步，这种划分带有较大的人为性。我们可以将它们理解为在一个时间上连续的变化过程，用随机微分方程来描述前向过程（"拆楼"）：

$$dx = f_t(x)dt + g_t dw \tag{1}$$

可以将其看作是下述离散形式在 $\Delta t \to 0$ 时的极限：

$$x_{t+\Delta t} - x_t = f_t(x_t)\Delta t + g_t\sqrt{\Delta t}\,\varepsilon,\ \varepsilon \sim N(0, I) \tag{2}$$

假设拆楼需要1天，那么拆楼就是 x 从 $t=0$ 到 $t=1$ 的变化过程，每一小步的变化可以用上述方程描述。没有对时间间隔 Δt 作过分的限制，越小的 Δt 意味着对原始SDE越好的近似。若取 $\Delta t = 0.001$，则对应 $T=1000$，若 $\Delta t = 0.01$ 则 $T=100$，我们不需事先指定T，而是根据实际情况下的精确度来取适当T进行数值计算。

在扩散模型中引用SDE的好处是"将理论分析与代码实践分离"。可以借助连续性数学工具对它做分析，而实践时只需要用适当的离散方案对SDE进行数值计算。

逆向方程：式2的条件概率为：

$$P(x_{t+\Delta t} \mid x_t) = N(x_{t+\Delta t};\ x_t + f_t(x_t)\Delta t,\ g_t^2 \Delta t I)$$

$$\propto \exp\left(-\frac{\| x_{t+\Delta t} - x_t - f_t(x_t)\Delta t \|^2}{2 g_t^2 \Delta t}\right) \tag{3}$$

简单起见，这里省略了无关紧要的归一化因子

贝叶斯定理：

$$P(x_t \mid x_{t+\Delta t}) = \frac{P(x_{t+\Delta t} \mid x_t) P(x_t)}{P(x_{t+\Delta t})} = P(x_{t+\Delta t} \mid x_t) \exp(\log P(x_t) - \log P(x_{t+\Delta t}))$$

$$\propto \exp\left(-\frac{\| x_{t+\Delta t} - x_t - f_t(x_t)\Delta t \|^2}{2 g_t^2 \Delta t} + \log P(x_t) - \log P(x_{t+\Delta t})\right) \tag{4}$$

这个发现我不能理解

（可以发现，当 Δt 很小即 $x_{t+\Delta t}$ 与 x_t 足够接近时，$P(x_{t+\Delta t} \mid x_t)$ 才会明显不等于0，

反过来也只有在这种情况下$P(x_t|x_{t+\Delta t})$才明显不等于0。) 因此，我们需要对$x_{t+\Delta t}$与x_t足够接近时的情形做近似分析，用泰勒展开：

$$\log P(x_{t+\Delta t}) \approx \log P(x_t) + (x_{t+\Delta t} - x_t)\cdot \nabla_{x_t} \log P(x_t) + \Delta t \frac{\partial}{\partial t} \log p(x_t) \tag{5}$$

注意不要忽略$\frac{\partial}{\partial t}$项，因为$P(x_t)$是"$t$时刻随机变量等于$x_t$的概率密度"，而$P(x_{t+\Delta t})$实际上是"$t+\Delta t$时刻随机变量等于$x_{t+\Delta t}$的概率密度"，也就是说$P(x_t)$实际上同时是$t$和$x_t$的函数，所以要多一项$t$的偏导数，代入(4)，配方可得：

$$P(x_t|x_{t+\Delta t}) \propto \exp\left(-\frac{\|x_{t+\Delta t} - x_t - [f_t(x_t) - g_t^2 \nabla_{x_t} \log P(x_t)]\Delta t\|^2}{2g_t^2 \Delta t} + o(\Delta t)\right) \tag{6}$$

在式(6)中，当$\Delta t \to 0$时，$o(\Delta t) \to 0$不起作用，因此

$$P(x_t|x_{t+\Delta t}) \propto \exp\left(-\frac{\|x_{t+\Delta t} - x_t - [f_t(x_t) - g_t^2 \nabla_{x_t} \log P(x_t)]\Delta t\|^2}{2g_t^2 \Delta t}\right)$$

$$\approx \exp\left(-\frac{\|x_t - x_{t+\Delta t} + [f_{t+\Delta t}(x_{t+\Delta t}) - g_{t+\Delta t}^2 \nabla_{x_{t+\Delta t}} \log P(x_{t+\Delta t})]\Delta t\|^2}{2g_{t+\Delta t}^2 \Delta t}\right) \tag{7}$$

即$P(x_t|x_{t+\Delta t})$近似为$x_{t+\Delta t} - [f_{t+\Delta t}(x_{t+\Delta t}) - g_{t+\Delta t}^2 \nabla_{x_{t+\Delta t}} \log P(x_{t+\Delta t})]\Delta t$协方差为$g_{t+\Delta t}^2 \Delta t I$的正态分布，取$\Delta t \to 0$的极限，那么对应的SDE：

$$dx = [f_t(x) - g_t^2 \nabla_x \log P_t(x)]dt + g_t dw \tag{8}$$

这是反向过程的SDE，在P外标注了下标t，以突出这是t时刻的分布。

得分匹配：已知逆向SDE为式(8)，若进一步知道$\nabla_x \log P_t(x)$，则可以通过离散化格式

$$x_t - x_{t+\Delta t} = -[f_{t+\Delta t}(x_{t+\Delta t}) - g_{t+\Delta t}^2 \nabla_{x_{t+\Delta t}} \log P(x_{t+\Delta t})]\Delta t + g_{t+\Delta t}\sqrt{\Delta t}\,\varepsilon \tag{9}$$

来逐步完成"建楼"的生成过程[其中$\varepsilon \sim N(0, I)$]，从而完成一个生成扩散模型的构建。

t时刻的$P_t(x)$就是前面的$P(x_t)$，它的含义是t时刻的边缘分布。在实际应用时，我们一般会设计找到$P(x_t|x_0)$解析解的模型，这意味着

$$P(x_t|x_0) = \lim_{\Delta t \to 0} \int \cdots \iint P(x_t|x_{t-\Delta t}) P(x_{t-\Delta t}|x_{t-2\Delta t}) \cdots P(x_{\Delta t}|x_0)\, dx_{t-\Delta t}\, x_{t-2\Delta t} \cdots x_{\Delta t} \tag{10}$$

是可以直接求出的，比如当$f_t(x)$是关于x的线性函数时，$P(x_t|x_0)$就可以解析求解。在此前提下，有

$$P(x_t) = \int P(x_t|x_0)\tilde{p}(x_0)\, dx_0 = E_{x_0}[P(x_t|x_0)] \tag{11}$$

$$\nabla_{x_t} \log P(x_t) = \frac{E_{x_0}[\nabla_{x_t} P(x_t|x_0)]}{E_{x_0}[P(x_t|x_0)]} = \frac{E_{x_0}[P(x_t|x_0) \nabla_{x_t} \log P(x_t|x_0)]}{E_{x_0}[P(x_t|x_0)]} \tag{12}$$

最后式子具有"$\nabla_{x_t}\log p(x_t|x_0)$的加权平均"的形式，由于假设了$p(x_t|x_0)$有解析解，因此上式能够直接估算，然而涉及了对全体训练样本x_0的平均，一来计算量大，二是泛化能力不好。所以希望用一个神经网络来学一个函数$S_\theta(x_t,t)$使得能够计算$\nabla_{x_t}\log p(x_t)$

$E[x]=\arg\min_u E_x[\|u-x\|^2]$ ⇒ 有这样一个等式(能推导) (13)

要让u等于x的均值，只需最小化$\|u-x\|^2$的均值。

要让$S_\theta(x_t,t)$等于$\nabla_{x_t}\log p(x_t|x_0)$的加权平均(即$\nabla_{x_t}\log p(x_t)$)则只需最小化$\|S_\theta(x_t,t)-\nabla_{x_t}\log p(x_t|x_0)\|^2$的加权平均，即

$$\frac{E_{x_0}[p(x_t|x_0)\|S_\theta(x_t,t)-\nabla_{x_t}\log p(x_t|x_0)\|^2]}{E_{x_0}[p(x_t|x_0)]}$$

分母$E_{x_0}[p(x_t|x_0)]$只是调节Loss参数，不会影响最优解结果，可以直接去掉。最后再对x_t积分(相当于对每个x_t都要最小化上述损失)，得到最终的损失函数：

$$\begin{aligned}&\int E_{x_0}[p(x_t|x_0)\|S_\theta(x_t,t)-\nabla_{x_t}\log p(x_t|x_0)\|^2]\,dx_t\\&=E_{x_0,x_t\sim p(x_t|x_0)p(x_0)}[\|S_\theta(x_t,t)-\nabla_{x_t}\log p(x_t|x_0)\|^2]\end{aligned} \tag{15}$$

结果倒推：构建生成扩散模型的一般流程：

(1) 通过随机微分方程(1)定义"拆楼"(前向过程)；

(2) 求$p(x_t|x_0)$的表达式；

(3) 通过损失函数(15)训练$S_\theta(x_t,t)$(得分匹配)；

(4) 用$S_\theta(x_t,t)$替换式(8)的$\nabla_{x_t}\log p_t(x)$，完成"建楼"(反向过程)。