1 DDPM

1.1 Diffusion

Diffusion 模型定义了一个概率分布转换模型 $\mathcal{T}$, 能将原始数据 x_0 构成的复杂分布

$$\mathbf{x}_0 \sim p_{\text{complex}} \Longrightarrow \mathcal{T}\left(\mathbf{x}_0\right) \sim p_{\text{prior}}$$

Diffusion 模型提出可以用马尔科夫链 (Markov Chain) 来构造 $\mathcal{T}$, 定义一系列条件概率分布 $q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \quad t \in \{1,2,3 \ldots T\}$, 将 $\mathbf{x}_0$ 依次转换为 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, 希望当 $T \rightarrow \inf$ 时, $\mathbf{x}_T \sim p_{\text{prior}}$

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathrm{I}\right)$$
$$q\left(\mathbf{x}_T\right)=p_{\text{prior }}\left(\mathbf{x}_T\right)=\mathcal{N}\left(\mathbf{x}_T ; \mathbf{0}, \mathrm{I}\right) \quad \text{ where } T \rightarrow \inf$$

即已知 $\mathbf{x}_{t-1}$ 时，$\mathbf{x}_t$ 的概率分布是一个平均值为 $\sqrt{1-\beta_t}\mathbf{x}_{t-1}$ ，协方差为 $\beta_t\mathbf{I}$ 的正态分布。其中 β 为扩散系数

根据数学技巧可得：（没有看的很懂推导的过程）

$$\mathbf{x}_t=\sqrt{1-\beta_t} \mathbf{x}_{t-1}+\sqrt{\beta_t} \mathbf{z}_{t-1} \quad \text{ where } \mathbf{z}_{t-1} \in \mathcal{N}(0, \mathbf{I})$$

将 x_t 一层一层迭代，可以得到以下式子：

假设 $\alpha_t=1-\beta_t, \bar{\alpha}_t=\prod_{i=1}^T \alpha_i$, 那么：

$$\begin{aligned}
\mathbf{x}_t & =\sqrt{\alpha_t} \mathbf{x}_{t-1}+\sqrt{1-\alpha_t} \mathbf{z}_{t-1} & & ; \text{ where } \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
& =\sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2}+\sqrt{\alpha_t\left(1-\alpha_{t-1}\right)} \mathbf{z}_{t-2}+\sqrt{1-\alpha_t} \mathbf{z}_{t-1} & & \\
& =\sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2}+\sqrt{1-\alpha_t \alpha_{t-1}} \overline{\mathbf{z}}_{t-2} & & ; \text{ where } \overline{\mathbf{z}}_{t-2}, \overline{\mathbf{z}}_{t-3}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
& =\ldots & & \\
& =\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \mathbf{z} & &
\end{aligned} \tag{3}$$

将公式 (3) 改写成条件概率的格式，可以得到：

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{\bar{\alpha}_t} \mathbf{x}_{t-1},(1-\bar{\alpha}_t) \mathrm{I}\right)$$

由于 $\beta_t \in (0,1)$, 那么 $\alpha_t \in (0,1)$ 。$t \rightarrow \inf$ 时, $\bar{\alpha}_t \rightarrow 0$ 。可以看出, $\sqrt{1-\beta_t}$ 和 $\sqrt{\beta_t}$ 作为系数保证了当 $T \rightarrow \inf$ 时, $q\left(\mathbf{x}_T\right)=p_{\text{prior }}\left(\mathbf{x}_T\right)=$

$\mathbf{N}(0, \mathbf{I})$ 。实际上，只要 T 取一个足够大的值，不需要无限次迭代，得到的分布就已经很接近于标准正态分布了。

下图为两个样本 (红蓝两颜色) 经过多次加噪最终得到 P_{prior}

1.2 逆 Diffusion

,$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ 分布未知，但数学上可证连续扩散过程的逆转具有与正向过程相同的分布形式。即即当扩散率 β_t 足够小, 扩散次数足够多时，离散扩散过程接近于连续扩散过程, $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ 的分布形式同 $q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$ 一致, 同样是高斯分布。但是很难直接写出 $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ 的分布参数。为此, 可以用分布 $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ 来近似 $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \ {}_\theta\left(\mathbf{x}_t, t\right)\right)$$

中 $\boldsymbol{\mu}_\theta$ 和 $\mathbf{\Sigma}_\theta$ 都是要学习的函数，接受 $\mathbf{x}_t, t$ 作为参数。这样，连续迭代多次后，可以得到近似的真实数据分布 $p_\theta\left(\mathbf{x}_0\right)$ 为：

$$p_\theta\left(\mathbf{x}_0\right)=\int p_\theta\left(\mathbf{x}_{0: T}\right) d \mathbf{x}_{1: T}$$

其中 $p_\theta\left(\mathbf{x}_{0: T}\right)$ 为 $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T$ 的联合概率分布。借助条件概率公式：

$$p_\theta\left(\mathbf{x}_{0: T}\right)=p\left(\mathbf{x}_T\right) \prod_{t=1}^T p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$$

在有了 μ_θ 和 ${}_\theta$ 之后，完成逆转过程。

实验效果如下：第一行为 Diffusion 前向，第二行为逆 Diffusion 后向。还有第三列表示漂移项，但我没有看明白

1.3 训练目标

用了交叉熵，但一堆公式没有搞明白 (QAQ)，就直接把结论给出来了。

1.4 改良

对于扩散模型来说，一个最大的缺点是需要设置较长的扩散步数才能得到好的效果，这导致了生成样本的速度较慢，比如扩散步数为 1000 的话，那么生成一个样本就要模型推理 1000 次。DDIM 改进：

不再限制扩散过程必须是一个马尔科夫链，这使得 DDIM 可以采用更小的采样步数来加速生成过程，DDIM 的另外是一个特点是从一个随机噪音生成样本的过程是一个确定的过程（中间没有加入随机噪音）。不再需要限制为马尔科夫链的原因是：推理分布满足边缘分布条件（扩散过程的特性）即可

通过 x_t 得到 x_{t-1} 的公式也改变了。

这里将生成过程分成三个部分: 一是由预测的 x_0 来产生的，二是由指向 x_t 的部分，三是随机噪音 (这里 ϵ_t 是与 $\mathbf{x}_t$ 无关的噪音) 。论文将 σ_t 近一步定义为:

$$\sigma_t^2 = \eta \cdot \tilde{\beta}_t = \eta \cdot \sqrt{(1-\alpha_{t-1}) / (1-\alpha_t)} \sqrt{(1-\alpha_t / \alpha_{t-1})}$$

这里考虑两种情况，一是 $\eta = 1$ ，此时 $\sigma_t^2 = \tilde{\beta}_t$ ，此时生成过程就和 DDPM 一样了。另外一种情况是 $\eta = 0$ ，这个时候生成过程就没有随机噪音了，是一个确定性的过程，论文将这种情况下的模型称为 DDIM (denoising diffusion implicit model)，一旦最初的随机噪音 $\mathbf{x}_T$ 确定了，那么 DDIM 的样本生成就变成了确定的过程。这一改进的作用有点搞不懂

关于数据泛化改良，只了解一个雏形，具体还没细看