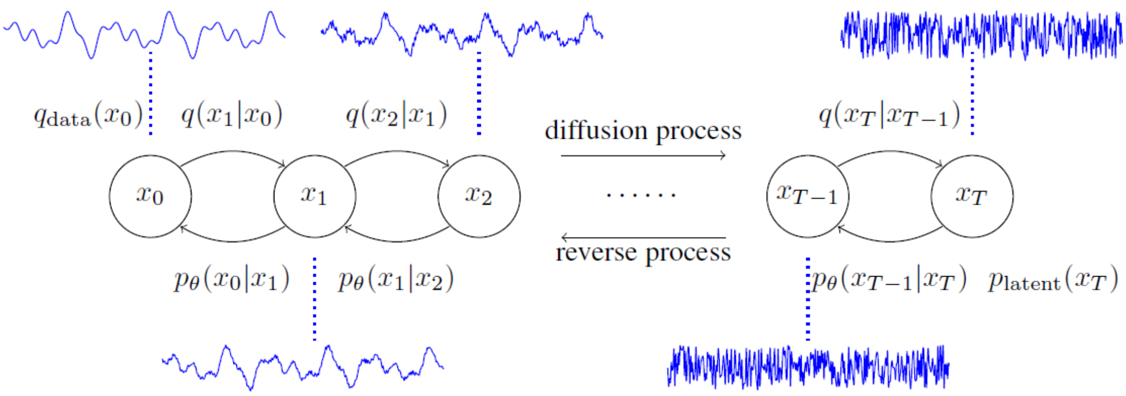
目录

DDPM完全解读

名词解析



- diffusion process: also called forward process, training process, represented by $q(x_t)$
- reverse process: also called sampling process, inference process, represented by $p(x_t)$

> 注意：
>
> - diffusion process：is fixed to a Markov chain that gradually adds Gaussian noise to the data according to variance schedule $\beta_1, \ldots \beta_T$, 即变换前后满足高斯分布，当前状态只与前一时刻有关; 下一小节将给出diffusion过程的分布预定义形式，即**variance schedule**是自定义的constant；且从上图可以看出diffusion过程与θ无关，只是为了求loss，将θ作用在forward input上，后续将具体介绍
> - reverse process: is defined as a Markov chain with learned Gaussian transition starting at $P(x_T) := N(x_T; 0, I)$，但reverse过程的mean and std是与θ相关的函数，为了使$p_\theta(x_0)$尽量接近$q_{data}(x_0)$，需要找到mean和std的最佳定义，使likelihood of $p_\theta(x_0)$最大，这也是diffusion model的loss定义，后续将具体介绍

预定义

Diffusion Process

根据Diffusion Model[1]的定义，定义了diffusion process is fixed to a Markov chain that gradually adds Gaussian noise to the data according to variance schedule $\beta_1, \ldots \beta_T$:

$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$where\ \beta_t\ is\ variance\ schedule,\ also\ called\ diffusion\ rate.$$

> 注意此处$\prod$不是阶乘，代表嵌套的意思，类似normalizing flow里的表达

Reverse Process

根据定义，sampling/reverse process is defined as a Markov chain with learned Gaussian transition starting at $p(x_T) := N(x_T; 0, I)$:

$$
\begin{gathered}
p(x_T) := N(x_T; 0, I) \\
p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)I) \\
p_\theta(x_{0:T}) = p_\theta(x_0, x_1, \ldots, x_T) = p(x_T) \cdot \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)
\end{gathered}
$$

> 注意此处$\prod$不是阶乘，代表嵌套的意思，类似normalizing flow里的表达

Loss

Reverse process is θ parameterization, 为了使reverse process尽可能的得到fidelity result，则需要找到使$p_\theta(x_0)$最接近$q_{data}(x_0)$分布的参数θ，即采用maximize log likelihood estimation，等价于minimize negative log likelihood。

下式中的$p_\theta(x_0)$通常很难准确求解，Sampling process start from $p_{latent}(x_T) := N(x_T; 0, I)$, 则:

$$
\begin{gathered}
p_\theta(x_0, x_1, \ldots, x_{T-1}|x_T) \cdot p(x_T) = p_\theta(x_0, x_1, \ldots, x_T) \\
p_\theta(x_0) = \int p_\theta(x_0, x_1, \ldots, x_T) dx_1 dx_2 \ldots dx_T = \int p_\theta(x_{0:T}) dx_{1:T}
\end{gathered}
$$

故θ通过minimize variational bound on log likelihood求解，log likelihood定义为:

$$
\begin{gathered}
-\mathbb{E}_{q_{data}(x_0)} log p_\theta(x_0) = -\mathbb{E}_{q_{data}(x_0)}(log \mathbb{E}_{q(x_1,\ldots,x_T|x_0)}[\frac{p_\theta(x_0, x_1, \ldots, x_{T-1}|x_T) \cdot p(x_T)}{q(x_1, \ldots, x_T|x_0)}]) \\
\text{计算其}variational\ bound\text{，将}\mathbb{E}\text{都提前，并合并下标概率，可得:} \\
-\mathbb{E}_{q_{data}(x_0)} log p_\theta(x_0) \leq -E_{q(x_0,\ldots,x_T)} log[\frac{p_\theta(x_0, x_1, \ldots, x_{T-1}|x_T) \cdot p(x_T)}{q(x_1, \ldots, x_T|x_0)}] \\
\text{简写为} \\
-\mathbb{E}_{q_{data}(x_0)} log p_\theta(x_0) \leq -\mathbb{E}_{q(x_{0:T})} log[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] =: L
\end{gathered}
$$

将上式L进一步可化简为:

$$
\mathbb{E}_q\Big[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Big] \quad (5)
$$

具体推导如下（注以下推导没加负号[2]，加负号的推导见[3]-Extra information-A Extended derivations）：

Proof. We expand the ELBO in Eq. (3) into the sum of a sequence of tractable KL divergences below.

$$
\begin{aligned}
\mathrm{ELBO} &= \mathbb{E}_q \log \frac{p_\theta(x_0, \cdots, x_{T-1}|x_T) \times p_{\text{latent}}(x_T)}{q(x_1, \cdots, x_T|x_0)} \\
&= \mathbb{E}_q \left(\log p_{\text{latent}}(x_T) - \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \\
&= \mathbb{E}_q \left(\log p_{\text{latent}}(x_T) - \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} - \sum_{t=2}^{T} \left(\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \right) \right) \\
&= \mathbb{E}_q \left(\log \frac{p_{\text{latent}}(x_T)}{q(x_T|x_0)} - \log p_\theta(x_0|x_1) - \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right) \\
&= -\mathbb{E}_q \left(\mathrm{KL}\left(q(x_T|x_0) \| p_{\text{latent}}(x_T)\right) + \sum_{t=2}^{T} \mathrm{KL}\left(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)\right) - \log p_\theta(x_0|x_1) \right)
\end{aligned} \quad (9)
$$

注意，loss的计算需要用到

$$q(x_T|x_0)$$
$$q(x_{t-1}|x_t,x_0)$$
$$p_\theta(x_{t-1}|x_t),p_\theta(x_0|x_1)$$

前两个分布以下将分别推导，第三个分布即DDPM[3] proposed distribution，which makes DDPM resembling denoising score matching[4].

推导

$q(x_T|x_0)$

根据diffusion model预定义，从1至T的任意时刻，x_t相对x_0的后验：$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$

以$q(x_3|x_0)$为例，即$t=3$时：

$$q(x_1|x_0) := N(x_1; \sqrt{1-\beta_1}x_0, \beta_1 I_1), so\ x_1 = \sqrt{1-\beta_1}x_0 + \sqrt{\beta_1}I_1$$
$$q(x_2|x_1) := N(x_2; \sqrt{1-\beta_2}x_1, \beta_2 I_2), so\ x_2 = \sqrt{1-\beta_2}(\sqrt{1-\beta_1}x_0 + \sqrt{\beta_1}I_1) + \sqrt{\beta_2}I_2$$
$$q(x_3|x_2) := N(x_3; \sqrt{1-\beta_3}x_2, \beta_3 I_3), so\ x_3 = \sqrt{1-\beta_3}[\sqrt{1-\beta_2}(\sqrt{1-\beta_1}x_0 + \sqrt{\beta_1}I_1) + \sqrt{\beta_2}I_2] + \sqrt{\beta_3}I_3$$

令$\alpha_t = 1-\beta_t, \overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$，则$q(x_3|x_0)$的均值，方差为

$$mean = \sqrt{1-\beta_3}\sqrt{1-\beta_2}\sqrt{1-\beta_1} = \sqrt{\overline{\alpha_3}}$$

$$std^2 = (\sqrt{1-\beta_3}\sqrt{1-\beta_2}\sqrt{\beta_1})^2 + (\sqrt{1-\beta_3}\sqrt{\beta_2})^2 + (\sqrt{\beta_3})^2 = \alpha_3\alpha_2(1-\alpha_1) + \alpha_3(1-\alpha_2) + (1-\alpha_3) = 1-\alpha_1\alpha_2\alpha_3 = 1-\overline{\alpha_3}$$

因此，$q(x_3|x_0) := N(x_3; \sqrt{\overline{\alpha_3}} \cdot x_0, (1-\overline{\alpha_3})I$，同理，推广到所有的t可得：

$$q(x_t|x_0) := N(x_t; \sqrt{\overline{\alpha_t}} \cdot x_0, (1-\overline{\alpha_t})I)$$

或者，迭代的理论推导如下：

$$\begin{aligned}
x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-1}}x_{t-3} + \sqrt{\alpha_t\alpha_{t-1}\beta_{t-2}}\epsilon_{t-2} + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_2\beta_1}\epsilon_1 + \cdots + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t
\end{aligned}$$

Note that $q(x_t|x_0)$ is still Gaussian, and the mean of x_t is $\sqrt{\bar{\alpha}_t}x_0$, and the variance matrix is $(\alpha_t\alpha_{t-1}\cdots\alpha_2\beta_1 + \cdots + \alpha_t\beta_{t-1} + \beta_t)I = (1-\bar{\alpha}_t)I$. Therefore,

$$q(x_t|x_0) = \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t}x_0,\ (1-\bar{\alpha}_t)I). \tag{10}$$

It is worth mentioning that,

$$q(x_T|x_0) = \mathcal{N}(x_T;\ \sqrt{\bar{\alpha}_T}x_0,\ (1-\bar{\alpha}_T)I), \tag{11}$$

where $\bar{\alpha}_T = \prod_{t=1}^{T}(1-\beta_t)$ approaches zero with large T.

$q(x_{t-1}|x_t,x_0)$

根据贝叶斯公式，全概率公式等，可得：

$$\begin{aligned}
q(x_{t-1}|x_t,x_0) &= \frac{q(x_{t-1},x_t,x_0)}{q(x_t,x_0)} = \frac{q(x_t|x_{t-1},x_0)\cdot q(x_0,x_{t-1})}{q(x_t,x_0)} = \frac{q(x_t|x_{t-1},x_0)\cdot q(x_0,x_{t-1})}{q(x_t,x_0)} \\
&= \frac{q(x_t|x_{t-1},x_0)\cdot q(x_{t-1}|x_0)\cdot q(x_0)}{q(x_t|x_0)\cdot q(x_0)} = \frac{q(x_t|x_{t-1},x_0)\cdot q(x_{t-1}|x_0)}{q(x_t|x_0)}
\end{aligned}$$

根据各维独立高斯分布，将x_{t-1}关于x_t,x_0的分布可进一步化简得：

Next, by Bayes rule and Markov chain property,

$$
\begin{aligned}
q(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1})\, q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t I)\, \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})I)}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)} \\
&= (2\pi\beta_t)^{-\frac{d}{2}}(2\pi(1-\bar{\alpha}_{t-1}))^{-\frac{d}{2}}(2\pi(1-\bar{\alpha}_t))^{\frac{d}{2}} \times \\
&\quad \exp\left(-\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|^2}{2\beta_t} - \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\|^2}{2(1-\bar{\alpha}_{t-1})} + \frac{\|x_t - \sqrt{\bar{\alpha}_t}x_0\|^2}{2(1-\bar{\alpha}_t)}\right) \\
&= (2\pi\tilde{\beta}_t)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\tilde{\beta}_t}\left\|x_{t-1} - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t\right\|^2\right)
\end{aligned}
$$

Therefore,

$$
q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \tilde{\beta}_t I). \tag{12}
$$

其中，$\overline{\beta}_t$代表：

$$
\overline{\beta}_t = \frac{1-\overline{\alpha_{t-1}}}{1-\overline{\alpha_t}} \cdot \beta_t
$$

$p_\theta(x_{t-1}|x_t)$

由于L_{t-1}是$q(x_{t-1}|x_t, x_0)$与$p_\theta(x_{t-1}|x_t)$的KL散度，$q(x_{t-1}|x_t, x_0)$已求得，因此需要确定$p_\theta(x_{t-1}|x_t)$的分布，使loss有closed-form calculation，因此DDPM[3]给出了$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)I)$中$\mu_\theta(x_t, t), \sum_\theta(x_t, t)$的形式，实现了closed-form expression，并与denoising score matching对应起来，以下将具体介绍。

根据前文reverse process预定义，

$$
p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)I)
$$

Loss中与其相关的项为：

$$
L_{t-1} = \sum_{t>1} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))
$$

其中，根据上文推导，将$q(x_{t-1}|x_t, x_0)$简写为：

$$
q(x_{t-1}|x_t, x_0) := N(x_{t-1}; \overline{\mu_t}(x_t, x_0), \overline{\beta}_t I)
$$
$$
where\ \overline{\mu_t}(x_t, x_0) = \frac{\sqrt{\overline{\alpha_{t-1}}} \cdot \beta_t}{1-\overline{\alpha_t}} \cdot x_0 + \frac{\sqrt{\alpha_t} \cdot (1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}} \cdot x_t
$$

首先考虑std

Ho通过实验发现，令

- $\sum_\theta(x_t, t) = \sigma_t^2 = \beta_t$，$\beta_t$趋向1
- $\sum_\theta(x_t, t) = \sigma_t^2 = \overline{\beta}_t$，$\overline{\beta}_t < \beta_t$

有相似的结果，The first choice is optimal for $x_0 := N(x_0; 0, I)$, and thes econd is optimal for x_0 deterministically set to one point. These are the two extreme choicesc orresponding to upper and lower bounds on reverse process entropy for data with coordinatewise unit variance[1].

无论第一种还是第二种方式，$\sum_\theta(x_t, t)$都与θ无关，由于$q(x_{t-1}|x_t, x_0)$的variance与θ也无关，因此两个分布的std项带入KL divergence中计算得到常数C

> 因此，实验中采用第二种方式。

其次考虑mean

将C带入L_{t-1}可化简得：

$$
L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C \tag{8}
$$

根据上文：

$$q(x_t|x_0) := N(x_t; \sqrt{\overline{\alpha_t}} \cdot x_0, (1-\overline{\alpha_t})I) \to x_t = \sqrt{\overline{\alpha_t}} \cdot x_0 + \sqrt{1-\overline{\alpha_t}} \cdot \epsilon \to x_0 = \frac{1}{\sqrt{\overline{\alpha_t}}} \cdot (x_t - \sqrt{1-\overline{\alpha_t}} \cdot \epsilon)$$

$$\overline{\mu_t}(x_t, x_0) = \frac{\sqrt{\overline{\alpha_{t-1}}} \cdot \beta_t}{1-\overline{\alpha_t}} \cdot x_0 + \frac{\sqrt{\alpha_t} \cdot (1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}} \cdot x_t$$

带入上式得：

$$L_{t-1} - C = \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t\left(\mathbf{x}_t(\mathbf{x}_0,\epsilon), \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0,\epsilon) - \sqrt{1-\bar{\alpha}_t}\epsilon)\right) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0,\epsilon),t)\right\|^2\right] \tag{9}$$

$$= \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0,\epsilon),t)\right\|^2\right] \tag{10}$$

根据上式可知，μ_θ需要预测（尽可能近似）$\frac{1}{\sqrt{\alpha_t}}(x_t(x_0,\epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon)$，由于其中$x_t(x_0,\epsilon)$是reverse input，是已知的（gaussian noise），因此，μ_θ的proposion将参数作用于ϵ，同时将x_t也作为其输入（因为μ_θ是关于x_t, t的函数，不会引入新的自变量），因此Ho将mean定义为：

$$\mu_\theta(\mathbf{x}_t,t) = \tilde{\mu}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t))\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right) \tag{11}$$

其中，ϵ_θ is a function approximator,用来根据x_t估计ϵ,即guassion noise.

整合

有了std和mean的定义，即可给出x_{t-1}的解析：

$$x_{t-1} = mean + std \cdot z = \frac{1}{\sqrt{\alpha_t}} \cdot (x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t,t)) + \sigma_t z$$
$$where\ z := N(0,I)$$

x_{t-1}既可以视为reverse所得，也可视为diffusion所得，因此training过程，ϵ_θ可作用于diffusion variable

The complete sampling procedure resembles Langevin dynamics with ϵ_θ as a learned gradient of the data density. 同时，Eq10.可简化为：

$$\mathbb{E}_{x_0,\epsilon}[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\overline{\alpha_t})}||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha_t}} \cdot x_0 + \sqrt{1-\overline{\alpha_t}} \cdot \epsilon, t)||]$$

which resembles denoising score matching over multiple noise scales indexed by t. 上式 is equal to (one term of) the variational bound for the Langevin-like reverse process, we see
that optimizing an objective resembling denoising score matching is equivalent to using variational inference to fit the finite-time marginal of a sampling chain resembling Langevin dynamics. 即与denoising score matching对应起来了。

此外，通过不同的化简形式，也可以约掉x_t，使μ_θ化简为关于x_0，但作者通过实验发现这种方式生成的图片质量更差。

因此，DDPM的diffusion process和reverse process可概括如下：

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1,\ldots,T\})$
4: $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
2: **for** $t = T,\ldots,1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

> 注意，training过程中，取$t \in Uniform1,\ldots,T$, 不一样的的t对应不一样的β, 实际实现上，每次循环，不同的样本的t都不同；对同一样本，每次loop随机取t，即相当于每次训练了不同长度的markov chain，足够多次的iteration之后，遍历多次训练了整个链；不同长度的链都尽量优化到最小loss。
>
> eg. DDPM论文中，T取1000，即链的长度为1000

Reference

[1]. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, 2015.

[2]. DIFFWAVE: A VERSATILE DIFFUSION MODEL FOR AUDIO SYNTHESIS, 2021.

[3]. Denoising Diffusion Probabilistic Models, 2020.

[4]. Generative Modeling by Estimating Gradients of the Data Distribution, 2019.