读懂扩散生成模型

李扶洋

邮箱：951678201@qq.com

2022 年 10 月 28 日

§必备基础

高斯正态分布：

单变量正态高斯分布概率密度函数：

$$p(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2\big) \tag{1}$$

记作：$x\sim\mathcal{N}(\mu,\sigma)$

多随机变量正态高斯分布概率密度函数：

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big) \tag{2}$$

记作：$\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$

高斯分布性质：

#1:独立高斯随机变量之和仍是高斯分布

如果：$\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$，$\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu}^{'},\boldsymbol{\sigma}^{'})$，且$\mathbf{x},\mathbf{y}$同维度，则其和也是高斯分布：$\mathbf{x}+\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\mu}^{'},\boldsymbol{\sigma}+\boldsymbol{\sigma}^{'})$

为了介绍接下来的性质，再做以下假设：

随机变量$\mathbf{x}\in\mathbb{R}^n$，且$\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$，假设随机变量$\mathbf{x}$有两部分组成：

$\mathbf{x}_A=[x_1,\ldots,x_r]^T\in\mathbb{R}^r$，$\mathbf{x}_B=[x_{r+1},\ldots,x_n]^T\in\mathbb{R}^{n-r}$，因此：

$\mathbf{x}=\begin{bmatrix}\mathbf{x}_A\\ \mathbf{x}_B\end{bmatrix}$，$\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_A\\ \boldsymbol{\mu}_B\end{bmatrix}$，$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{AA}\boldsymbol{\Sigma}_{AB}\\ \boldsymbol{\mu}_{BA}\boldsymbol{\mu}_{BB}\end{bmatrix}$(其中，$\boldsymbol{\Sigma}_{AB}=\boldsymbol{\Sigma}_{BA}$)

#2:联合高斯随机分布的边际分布仍是高斯分布

$$p(\mathbf{x}_A)=\int_{\mathbf{x}_B}p(\mathbf{x}_A,\mathbf{x}_B;\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}_B \tag{3}$$

$$p(\mathbf{x}_B) = \int_{\mathbf{x}_A} p(\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}_A \tag{4}$$

以上两个边缘分布均为高斯分布：

$\mathbf{x}_A \sim \mathcal{N}(\boldsymbol{\mu_A}, \boldsymbol{\Sigma_{AA}})$，$\mathbf{x}_B \sim \mathcal{N}(\boldsymbol{\mu_B}, \boldsymbol{\Sigma_{BB}})$

#3:高斯随机分布的条件分布仍是高斯分布

$$p(\mathbf{x}_A|\mathbf{x}_B) = \frac{p(\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\mathbf{x}_A} p(\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}_A} \tag{5}$$

$$p(\mathbf{x}_B|\mathbf{x}_A) = \frac{p(\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\mathbf{x}_B} p(\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}_B} \tag{6}$$

以上两个条件分布均为高斯分布：

$\mathbf{x}_A|\mathbf{x}_B \sim \mathcal{N}(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B), \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}{}^{-1}\boldsymbol{\Sigma}_{BA})$

$\mathbf{x}_B|\mathbf{x}_A \sim \mathcal{N}(\boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}(\mathbf{x}_A - \boldsymbol{\mu}_A), \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}{}^{-1}\boldsymbol{\Sigma}_{AB})$

#4:高斯随机分布与标准正态分布的转换关系

$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$，则一定存在矩阵$\mathbf{B}(\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T)$，通过变换：

$\mathbf{Z} = \mathbf{B}^{-1}(\mathbf{X} - \boldsymbol{\mu})$，使得：$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

也就是说：任意多变量高斯分布都可以通过线下变换$\mathbf{X} = \mathbf{B}\mathbf{Z} + \boldsymbol{\mu}$来得到，

其中：$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

#5:两个高斯随机分布密度相乘(相除)得到一个高斯函数（而非高斯分布，因其积分非1）

$x_1 \sim \mathcal{N}(\mu_1, \sigma_1)$，$x_2 \sim \mathcal{N}(\mu_2, \sigma_2)$

乘积概率密度函数均值和方差为：$\mu = \frac{\mu_1{\sigma_2}^2 + \mu_2{\sigma_1}^2}{{\sigma_1}^2 + {\sigma_2}^2}$，$\sigma^2 = \frac{{\sigma_1}^2{\sigma_2}^2}{{\sigma_1}^2 + {\sigma_2}^2}$

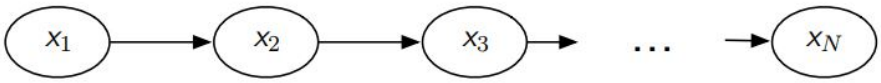
商的概率密度函数均值和方差为：$\mu = \frac{\mu_1{\sigma_2}^2 - \mu_2{\sigma_1}^2}{{\sigma_1}^2 - {\sigma_2}^2}$，$\sigma^2 = -\frac{{\sigma_1}^2{\sigma_2}^2}{{\sigma_1}^2 - {\sigma_2}^2}$

推广至高维情形下均值和方差为：

$\boldsymbol{\mu} = (\boldsymbol{\Sigma_1}^{-1} + \boldsymbol{\Sigma_2}^{-1})^{-1}(\boldsymbol{\Sigma_1}^{-1}\boldsymbol{\mu_1} + \boldsymbol{\Sigma_2}^{-1}\boldsymbol{\mu_2})$，$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma_1}^{-1} + \boldsymbol{\Sigma_2}^{-1})^{-1}$

高斯马尔可夫过程（链）：

马尔可夫链常被用来在微小状态变化前提下，从复杂副本中抽样；而高斯分布具有良好的解析形式和解



可将上图所示看作高斯马尔可夫过程，则对应的有：$\mathbf{x}_i = \mathbf{A}\mathbf{x}_{i-1} + \mathbf{B}\mathbf{v}_i$，

其中：$\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$，$\mathbf{v}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_v})$

§扩散模型的理论基础

下面将介绍的前向扩散和反向生成都使用了马尔可夫高斯过程，但为何前向扩散过程使用小步长变化的马尔可夫高斯过程时，反向生成过程也可以是马尔可夫高斯过程（且前向过程切分越细，这个越成立）？建议直接接受这个前人已经验证成立的结论：柯尔莫哥洛夫前向-后向方程（Kolmogorov forwardand backward equations）显示对于很多前向扩散过程，反向扩散过程能用同样的函数形式进行描述。

§DDPM：去噪扩散概率模型

前向扩散过程： 从未知的复杂的数据分布开始(记作：$q(\mathbf{x}_0)$,实际上我们不知道$q(\mathbf{x}_0)$的概率密度函数，但是我们有该分布的很多样本或者样本易得)，进行T（典型T值大于等于1000）步长的马尔可夫高斯扩散(使用方差参数序列：$\beta_1, \ldots, \beta_T$)，最终得到一个简单分布（如：$q(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$,表示为：

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{7}$$

前向过程方差参数β_t可以设定成参数来学习获得，或者当作超参数设置成固定小值（小于0.02）（这也是下面反向过程也具备高斯分布所需要的）；另外，一个值得注意的属性是前向过程在任意时间t处的抽样随机变量$\mathbf{x}_t$的概率分布具有解析解，即：

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\overline{\alpha}_t}\mathbf{x}_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{8}$$

其中：$\alpha_t := 1 - \beta_t$， $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$

反向生成过程： 反向生成过程将被训练以描述前向扩散使用的同一轨迹过程，即：

$$q(\mathbf{x}_{0:T}) = \underbrace{q(\mathbf{x}_0) \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})}_{Forward} = \underbrace{q(\mathbf{x}_T) \prod_{t=1}^{T} q(\mathbf{x}_{t-1}|\mathbf{x}_t)}_{Reverse} \tag{9}$$

上式中，【Forward：前向扩散过程】上面已详细说明，其马尔可夫高斯过程的所有参数均为已知；【Reverse：反向生成过程】按照之前说明，只知

其和前向过程有相同的函数形式：即马尔可夫高斯过程。但是每个反向过程的高斯分布参数是未知的,如何确定反向高斯分布的所有参数？

解决问题的思路：

使用参数化的概率密度函数$p_\theta(\mathbf{x}_{0:T})$,来逼近$q(\mathbf{x}_T)\prod_{t=1}^{T} q(\mathbf{x}_{t-1}|\mathbf{x}_t)$，即：

$$p_\theta(\mathbf{x}_{0:T}) = p_\theta(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_T)\prod_{t=1}^{T} q(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{10}$$

其中：　$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$ 由 (10)可以得到：

$$\begin{aligned}
p_\theta(\mathbf{x}_0) &= \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})p_\theta(\mathbf{x}_{1:T})d\mathbf{x}_{1:T}}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)p_\theta(\mathbf{x}_{1:T})d\mathbf{x}_{1:T}}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \int p_\theta(\mathbf{x}_0|\mathbf{x}_1)q(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T})d\mathbf{x}_{1:T}}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \int p_\theta(\mathbf{x}_0|\mathbf{x}_1)q(\mathbf{x}_{1:T-1}|\mathbf{x}_T,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{1:T-1}|\mathbf{x}_T,\mathbf{x}_0)}d\mathbf{x}_{1:T-1}\int q(\mathbf{x}_T|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}d\mathbf{x}_T \\
&= \underline{\int p_\theta(\mathbf{x}_0|\mathbf{x}_1)q(\mathbf{x}_{1:T-1}|\mathbf{x}_T,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{1:T-1}|\mathbf{x}_T,\mathbf{x}_0)}d\mathbf{x}_{1:T-1}}\int q(\mathbf{x}_T|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}d\mathbf{x}_T \\
&= \underline{\int p_\theta(\mathbf{x}_0|\mathbf{x}_1)q(\mathbf{x}_{1:T-2}|\mathbf{x}_{T-1},\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T-2}|\mathbf{x}_{T-1})}{q(\mathbf{x}_{1:T-2}|\mathbf{x}_{T-1},\mathbf{x}_0)}d\mathbf{x}_{1:T-2}} \\
&\quad * \int q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}d\mathbf{x}_{T-1} * \int q(\mathbf{x}_T|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}d\mathbf{x}_T \\
&= \quad \dots \\
&\vdots \\
&= \quad \dots \\
&= p_\theta(\mathbf{x}_0|\mathbf{x}_1)\int q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_1|\mathbf{x}_2)}{q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}d\mathbf{x}_1\int\dots\int q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}d\mathbf{x}_{T-1} \\
&\quad * \int q(\mathbf{x}_T|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}d\mathbf{x}_T
\end{aligned} \tag{11}$$

对上式 (11)两边取对数，得到：

$$
\begin{aligned}
\log p_\theta(\mathbf{x}_0) &= \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\
&\quad + \log \int q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_1|\mathbf{x}_2)}{q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}d\mathbf{x}_1 \\
&\quad + \dots \\
&\quad + \log \int q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}d\mathbf{x}_{T-1} \\
&\quad + \log \int q(\mathbf{x}_T|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}d\mathbf{x}_T \\
&= \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\
&\quad + \log E_{\mathbf{x}_1\sim q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}[\frac{p_\theta(\mathbf{x}_1|\mathbf{x}_2)}{q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}] \\
&\quad + \dots \\
&\quad + \log E_{\mathbf{x}_{T-1}\sim q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}[\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}] \\
&\quad + \log E_{\mathbf{x}_T\sim q(\mathbf{x}_T|\mathbf{x}_0)}[\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}] \\
&\geq \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\
&\quad + E_{\mathbf{x}_1\sim q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_1|\mathbf{x}_2)}{q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}] \\
&\quad + \dots \\
&\quad + E_{\mathbf{x}_{T-1}\sim q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}] \\
&\quad + E_{\mathbf{x}_T\sim q(\mathbf{x}_T|\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}]
\end{aligned}
\tag{12}
$$

即：

$$
\begin{aligned}
\log p_\theta(\mathbf{x}_0) &\geq \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\
&\quad + E_{\mathbf{x}_1\sim q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_1|\mathbf{x}_2)}{q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)}] \\
&\quad + \dots \\
&\quad + E_{\mathbf{x}_{T-1}\sim q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)}{q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)}] \\
&\quad + E_{\mathbf{x}_T\sim q(\mathbf{x}_T|\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}]
\end{aligned}
\tag{13}
$$

将 (13)进一步表示成Kullback-Leibler Divergence形式，为：

$$\begin{aligned}\log p_\theta(\mathbf{x}_0) \geq \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \\ - D_{KL}(q(\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_0)||p_\theta(\mathbf{x}_1|\mathbf{x}_2)) \\ - \dots \\ - D_{KL}(q(\mathbf{x}_{T-1}|\mathbf{x}_T,\mathbf{x}_0)||p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_T)) \\ - D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||p_\theta(\mathbf{x}_T))\end{aligned} \tag{14}$$

由 (15)可知，反向生成过程除了第一项（第一项可以通过后面描述的参数学习后反向抽样得到的$\mathbf{x}_1$，并特殊处理得到最终输出）和最后一项（最后一项可以从简单分布直接抽样得到，而不依赖参数）；其它项均可表示为（$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$），需要学习的参数需要逼近前向过程的后验概率($\mathbf{x}_0$给定的条件分布$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$)，接下来重点展开介绍。

实际上，利用贝叶斯公式，$\mathbf{x}_0$给定的条件分布$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$是具有解析解的高斯分布： $q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0),\widetilde{\beta}_t\mathbf{I})$，其中，$\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t$，$\widetilde{\beta}_t := \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$，由下面的推导过程可知，以上条件后验概率的均值和方差并非精确，而是取简化后一个大约解，过程如下：

$$\begin{aligned}q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) &= \frac{q(\mathbf{x}_t,\mathbf{x}_{t-1},\mathbf{x}_0)}{q(\mathbf{x}_t,\mathbf{x}_0)} \\ &= \frac{q(\mathbf{x}_t,\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\end{aligned} \tag{15}$$

要得到上述$\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)$的公式，这里需要忽略除数部分$q(\mathbf{x}_t|\mathbf{x}_0)$，利用$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t;\sqrt{1-\beta_t}\mathbf{x}_{t-1},\beta_t\mathbf{I})$、$q(\mathbf{x}_{t-1}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1};\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0,(1-\overline{\alpha}_{t-1})\mathbf{I})$和高斯分布性质#5[两个高斯随机分布密度相乘的结论]，并把分母中的β_t忽略且使$\mathbf{x}_t := \mathbf{x}_{t-1}$而得到。你也可以完全遵从高斯分布性质#5精确计算得到更为复杂而精确的均值和方差，但复杂的表达式没有简化版更方便用于后续的推导计算。

接下来聚焦讨论在$1 < t \leq T$时都适用的$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t))$的选择和参数学习策略问题。

DDPM原文和实验都设定$\mathbf{\Sigma}_\theta(\mathbf{x}_t,t))=\beta_t$复用前向扩散过程的高斯方差序列取值而无需通过训练参数学习得到；另外一种选择是，取值为前向扩散过程条件后验概率的方差$\mathbf{\Sigma}_\theta(\mathbf{x}_t,t)=\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$，不过实验表面，两种选择有相似的模型表现。

基于以上设定，$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t):=\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\mathbf{\Sigma}_\theta(\mathbf{x}_t,t))$进一步写为：$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t):=\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\beta_t\mathbf{I})$，更进一步：

$$
\begin{aligned}
D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) &= -E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\log[\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}] \\
&= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\log[\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}]
\end{aligned}
\tag{16}
$$

对 (16)代入：$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\sim\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0),\widetilde{\beta}_t\mathbf{I})$和
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\sim\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\beta_t\mathbf{I})$，且忽略$\widetilde{\beta}_t$与$\beta_t$的差异（或者使用$\mathbf{\Sigma}_\theta(\mathbf{x}_t,t))=\widetilde{\beta}_t$），得到：

$$
\begin{aligned}
E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\log[\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}] &= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\log[\exp \\
&\quad \frac{1}{2\widetilde{\beta}_t}((\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t))^T(\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)) \\
&\quad -(\mathbf{x}_{t-1}-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0))^T(\mathbf{x}_{t-1}-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)))] \\
&= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\
&\quad \frac{1}{2\widetilde{\beta}_t}((\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t))^T(\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)) \\
&\quad -(\mathbf{x}_{t-1}-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0))^T(\mathbf{x}_{t-1}-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)))] \\
&= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\
&\quad \frac{1}{2\widetilde{\beta}_t}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)^2-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)^2 \\
&\quad -2\mathbf{x}_{t-1}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)-\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)))]
\end{aligned}
\tag{17}
$$

对 (17)中$\mathbf{x}_{t-1}$，取值为$\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t, \mathbf{x}_0)$，代入得到：

$$\begin{aligned} E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} \log [\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}] &= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\ &\quad \frac{1}{2\widetilde{\beta}_t}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)^2 - \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)^2 \\ &\quad - 2\mathbf{x}_{t-1}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) - \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)))] \\ &= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\ &\quad \frac{1}{2\widetilde{\beta}_t}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)^2 + \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0)^2 \\ &\quad - 2\boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0))] \\ &= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\ &\quad \frac{1}{2\widetilde{\beta}_t}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) - \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0))^2] \end{aligned} \tag{18}$$

即：

$$\begin{aligned} D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) &= E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}[\\ &\quad \frac{1}{2\widetilde{\beta}_t}(\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) - \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0))^2] \end{aligned} \tag{19}$$

因此，我们可以对参数化的$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$进行训练去逼近$\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t, \mathbf{x}_0)$，由于后者已知为：

$$\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha}_t(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \tag{20}$$

得到一个回归拟合问题或者监督学习已知可解问题。
不过，更进一步的分析，可以得到更加简化的形式，因此，我们继续介绍。

对 (8)使用参数化技巧来直接得到：
$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\overline{\alpha}_t)}\boldsymbol{\epsilon}$，其中：$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，解出$\mathbf{x}_0$(也可用代入$\mathbf{x}_t$，这里不做此选择展开介绍)，代入 (20)得到（类似的需要一定的略去

和近似处理）：

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\frac{1}{\sqrt{\overline{\alpha}_{t-1}}\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{(1-\overline{\alpha}_t)}\boldsymbol{\epsilon}) + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \\
&= \frac{\beta_t}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{(1-\overline{\alpha}_t)}\boldsymbol{\epsilon}) + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \\
&= \frac{\beta_t + \alpha_t(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t + \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon} \\
&\approx \frac{0 + 1(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t + \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon} \\
&= \frac{(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t + \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon} \\
&\approx \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t + \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon} \\
&= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) + \frac{\beta_t}{\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon}(\mathbf{x}_0, t))
\end{aligned} \tag{21}$$

上式 (21)是$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$逼近的目标，由于训练阶段$\mathbf{x}_t$总是可以在$\mathbf{x}_0$的条件下抽样得到，所以是已知量，无需通过参数优化学习；这里令$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$具有和(21)一样的形式，即：

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) + \frac{\beta_t}{\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) \tag{22}$$

上式中$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$是参数化函数用来学习来预测实际$t$时刻的$\boldsymbol{\epsilon}(\mathbf{x}_0, t)$的取值的，是真正的需要参数化和训练学习的部分。

将 (22)， (21)代入忽略常量部分的 (19)，得到：

$$E_{\mathbf{x}_{t-1}\sim q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}[(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) - \widetilde{\boldsymbol{\mu}_t}(\mathbf{x}_t, \mathbf{x}_0))^2] = E_{\mathbf{x}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}, \mathbf{I})}[(\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \boldsymbol{\epsilon}(\mathbf{x}_0, t)))^2] \tag{23}$$

上式中$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\overline{\alpha}_t)}\boldsymbol{\epsilon}$，其中：$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$；

配合上式 (21)，有以下抽样和训练方法：选取一个$\mathbf{x}_0$，进行前向扩散抽样，记录每个时间节点的$\boldsymbol{\epsilon}$取值，计算并记录$\mathbf{x}_t$抽样值；对数据集中每个数据样本，进行前述操作；接着选取适当的Batch Size，进行多轮迭代优化，学习得到优化参数的$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$。

对于训练优化后的$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$，反向生成过程则可以描述为：

$$\begin{aligned}\mathbf{x}_{t-1} &= \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) + \sqrt{\widetilde{\beta}_t}\mathbf{z}, \;\; \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t + \frac{\beta_t}{\sqrt{(1-\overline{\alpha}_{t-1})}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) + \sqrt{\widetilde{\beta}_t}\mathbf{z}\end{aligned} \tag{24}$$

值得注意的是 (24)中的$\mathbf{x}_t$和 (21)中的$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon})$是有着本质不同的。前者和$\mathbf{x}_0$无关，来自简单分布直接抽样（例如：噪音），用来无条件生成$\mathbf{x}_0$；后者和$\mathbf{x}_0$有关，来自前向扩散抽样，基于$\mathbf{x}_0$给定和选定的前提

到这里，我想可用这样理解DDPM：

1,前向扩散过程是步步向对应数据样本中添加噪音的过程，信噪比持续降低直至噪音化；（作用于：训练阶段）

2,反向生成过程的学习训练就是要试图学习掌握每一步添加的噪音值，进而去除这些噪音，以逐步恢复原有数据样本信息；（先作用于：训练阶段；训练完成后，又作用于：使用模型阶段，用来从噪音反向生成数据样本实现对$p_{data}(\mathbf{x})$的抽样）

3，优化训练优化的模型是对对$p_{data}(\mathbf{x})$的近似，可用于对$p_{data}(\mathbf{x})$进行近似抽样，生成更多服从该分布的数据；

4，DDPM从$\mathbf{x}_0$到$\mathbf{x}_T$并没有类似于隐式模型VAE那样学习到有效压缩的特征学习，有的仅只是步步随机选择并添加的噪音值，因此，反向的生成过程也不是从特征到数据的映射过程，而只是如DDPM直观表达的求解噪音值和利用求解的噪音值不断去除噪音还原数据的过程；

5，DDPM生成过程典型的经过1000步以上，速度比较慢（相对于GAN和VAE类生成模型而言），但是其生成新数据的抽样实现非常容易直接（这是明显优势）。

下面,对像素和简单分布取样数据之间需要的变换进行分析解释：

我们知道，图像的每个像素点电脑中使用$\{0-255\}$的整数值表示，而简单正态分布的值往往在以0为中心的较小区间内；基于这个观察，为了让像素和简单概率分布函数的取值更加匹配和一致，需要对像素值进行缩放变换处理。

处理的思路是：

1,首先，将$\{0-255\}$的整数值线性映射到$[-1, 1]$，这解决了反向生成过程第一步从简单抽样取值的问题；

2,然后，我们可以通过反复迭代使用 (24)，直至得到$\mathbf{x}_0$；

3,最后，还需要输出得到的$\mathbf{x}_0$按位置线下恢复输出离散的像素值使得图像

得以呈现；

现在,回头聚焦 (15)右侧第一项，其解释如下：

由 (15)可知，右侧是一个浮动值，取决于参数$\boldsymbol{\theta}$的浮动值；学习优化的过程就是提升该浮动值，让其持续增大，为了让这个值具体话可以计算，除(15)右侧第一项外，其它项均为$Kullback-LeiblerDivergence$形式已有封闭解，我们现在就要解决该项的计算问题，以得到具体的边界值数值，方法如下（实质就是：在已知$\mathbf{x}_0$概率分布和抽样值的情况下，积分计算概率）：

$$p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D} \int_{\delta_(x_0{}^i)}^{\delta_+(x_0{}^i)} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\theta^i(\mathbf{x}_1, 1), \beta_1 \mathbf{I}) dx \tag{25}$$

其中，D是图像像素数，积分区间上下限为：

$$\delta_+(x) = \begin{cases} \infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1 \end{cases}$$

以上是对DDMP的介绍，下面一篇会介绍其具体的训练模型，敬请期待