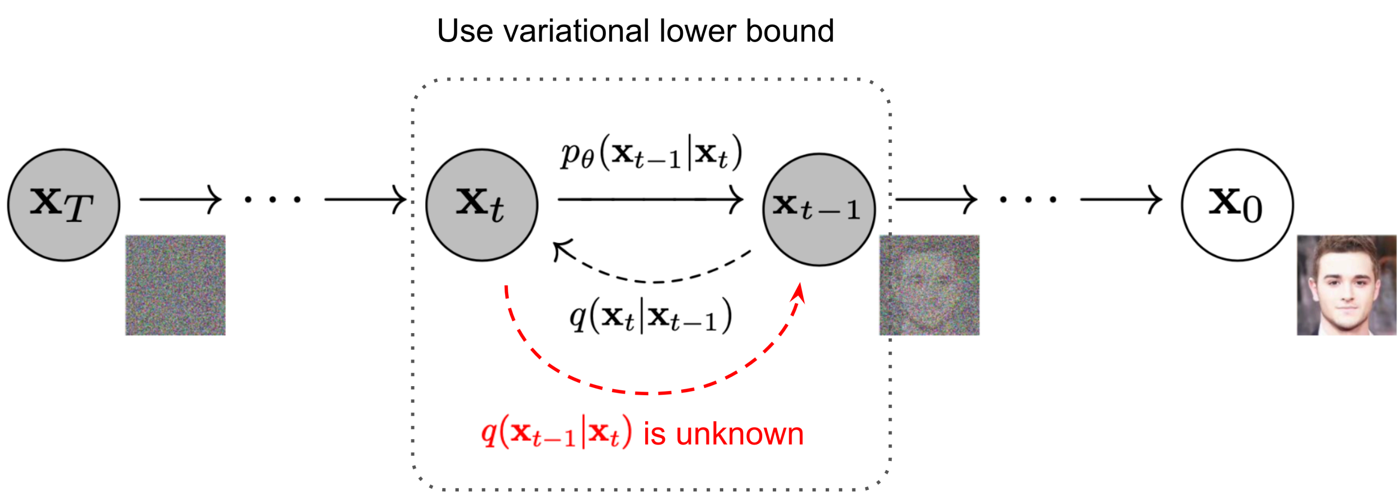
# 扩散模型之DDPM

> 一篇视角相似的论文：Denoising Diffusion Probabilistic Models in Six Simple Steps

生成模型包括GAN，VAE，Flow-based和Diffusion等

- GAN将数据生成这一无监督任务建模为一个有监督任务，但是问题在于不稳定的训练和生成结果多样性差
- VAE依赖于代理损失，通过最大化ELBO来间接最大化数据的似然
- Flow-based直接学习数据的分布，但是其结构需要精心设计来构建可逆变换
- Diffusion来源于非平衡热力学
- 与VAE和Flow-based不同，Diffusion通过一个固定的precedure学习（相较于Flow-based）并且中间变量具有高维度（应该是和VAE比较，Flow-based应该是不降维的？是不降维的）

正态分布的性质




## Forward diffusion process

参考：生成扩散模型漫谈（一）：DDPM = 拆楼 + 建楼，按照DDPM的记法重写了推导

首先，我们假定前向过程$q(\mathbf{x}_{1:T}|\mathbf{x}_0)$和反向过程$p_\theta(\mathbf{x}_{0:T})$都是一阶马尔可夫过程，即

$$
\begin{aligned}
q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \\
p_\theta(\mathbf{x}_{0:T}) &:= p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)
\end{aligned}
$$

给定真实数据$\mathbf{x}_0 \sim q(\mathbf{x})$，在第$t$步引入加性高斯噪声（Additive Gaussion noise，噪声直接加在原始信号上，原始信号不一定是高斯的）$\mathcal{N}(\mathbf{0}, \beta_t\mathbf{I})$的前向过程为一个线性高斯变换

$$
\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}
$$

- 多元高斯分布的线性高斯变换，来源

$$
\begin{aligned}
&\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma}) \\
&\Rightarrow \mathbf{A}\mathbf{x} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)
\end{aligned}
$$

- 重参数化技巧

$$
\begin{aligned}
\mathbf{z} &\sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{z}; \mu^{(i)}, \sigma^{2(i)}\mathbf{I}) \\
\mathbf{z} &= \mu + \sigma \odot \epsilon, \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}) \qquad ; \text{Reparameterization trick.}
\end{aligned}
$$

当前的限制条件只有$\alpha_t, \beta_t > 0$，$\epsilon_t$则是引入的噪声

- 可以描述为，$\mathbf{x}_t$通过将$\mathbf{x}_{t-1}$缩放并引入加性高斯噪声得到，$\alpha_t$也开根号是为了便于后续推导
- 概率的等价表述为

$$
q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})
$$

通过反复迭代这一分解，可以用$\mathbf{x}_0$噪声来表示所有任意时刻的$\mathbf{x}_t$

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t \\ &= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\beta_{t-1}}\epsilon_{t-1}\right) + \sqrt{\beta_t}\epsilon_t \\ &= \cdots \\ &= \sqrt{\alpha_t\cdots\alpha_1}\mathbf{x}_0 + \underbrace{\sqrt{\alpha_t\cdots\alpha_2}\sqrt{\beta_1}\epsilon_1 + \sqrt{\alpha_t\cdots\alpha_3}\sqrt{\beta_2}\epsilon_2 + \cdots + \sqrt{\alpha_t}\sqrt{\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t}_{\text{多个相互独立的正态噪声之和}}\end{aligned} \tag{2}$$

后一项多个相互独立的正态分布之和构成了均值为0，方差为$(\alpha_t\cdots\alpha_2)\beta_1 + (\alpha_t\cdots\alpha_3)\beta_2 + \cdots + \alpha_t\beta_{t-1} + \beta_t$的正态分布

- 两个**独立**多元正态随机向量的和分布仍然是正态分布，来源

$$\begin{gathered}\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}),\ \mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}}) \\ \Rightarrow \mathbf{x} + \mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{x}} + \mu_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{y}})\end{gathered}$$

- 概率的等价表述为

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

此时，只要$\alpha_t + \beta_t = 1$即$\alpha_t = 1 - \beta_t$，就有

$$(\alpha_t\cdots\alpha_1) + (\alpha_t\cdots\alpha_2)\beta_1 + (\alpha_t\cdots\alpha_3)\beta_2 + \cdots + \alpha_t\beta_{t-1} + \beta_t = 1 \tag{3}$$

这意味着满足上述条件就能够很容易地通过所有$\alpha_t$来计算后面形成的正态分布的方差，即

$$\mathbf{x}_t = \underbrace{\sqrt{\alpha_t\cdots\alpha_1}}_{\text{记为}\sqrt{\bar{\alpha}_t}}\mathbf{x}_0 + \underbrace{\sqrt{1 - (\alpha_t\cdots\alpha_1)}}_{\text{记为}\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_t, \quad \bar{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

- $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$
- 概率的等价表述为

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

总结而言，$\beta_t$表征了噪声的程度，而$\alpha_t$则是为了保持系数平方和为1，以便快速算出每个时间步所上生成样本而引入的

通常随着加噪声的进行，可以逐渐加入更多的噪声，即$\beta_1 < \beta_2 < \cdots < \beta_T$，也即$\bar{\alpha}_1 > \cdots > \bar{\alpha}_T$

- 这里源自lilianweng的post，但是有点问题，$\bar{\alpha}_t$本来就应该是单调递减的，这里本来想要表达的可能是$\alpha_1 > \cdots > \alpha_T$
- 当$\bar{\alpha}_t \approx 0$时就可以认为前向过程已经完成了

## Reverse diffusion process

为了从噪声$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$中还原出样本，需要知道分布$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$，但是$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$难以得到（需要整个数据集来统计），因此通过神经网络来拟合这些反向过程的条件概率

- 一开始就有个结论，如果$\beta_t$足够小，那么$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$也将是个高斯分布，这是为什么？
  - DPM论文中2.2节开头引用了经证明的结论，如果扩散过程是高斯过程或者二项过程，那么对于连续扩散过程（即扩散步长$\beta$足够小），逆向扩散过程和前向扩散过程有着同样的函数形式
  - 对DDPM而言，前向过程$q(\mathbf{x}_t|\mathbf{x}_{t-1})$服从高斯分布，因此如果$\beta_t$足够小，$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$就会是个高斯分布

我们可以用模型$p_\theta$来拟合条件分布$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$，即减小$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$和$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$的分布差异，由于$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$是个高斯分布，所以可以将$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$建模为高斯分布

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

这里回顾一下生成模型的目标，是要最大化$p_\theta(\mathbf{x}_0)$，我们的假设是，在Diffusion的稳态过程中，只要能够让每步的反向过程能够还原前向过程对应步的变换，就可以从噪声$\mathbf{x}_T$中还原出样本$\mathbf{x}_0$，而正向过程已经确定，所以只要让网络$p_\theta$拟合好$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$就足够了

因此，我们可以从分布差异出发，推导出它和样本分布的关系，寻找最大化$p_\theta(\mathbf{x}_0)$的变分下界

然而这里的$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$无法计算，从而无法算出分布差异

解决办法在于，由于Diffusion所假设的一阶马尔可夫性，当$t > 1$时，我们有

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}$$

$$\Rightarrow q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)=\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

从而分布差异可以使用$D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$来刻画

于是我们可以计算$t\geq 1$时的分布差异之和

$$D_{KL}=\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))+D_{KL}(q(\mathbf{x}_0|\mathbf{x}_1)\parallel p_\theta(\mathbf{x}_0|\mathbf{x}_1))$$

- 为什么这里是求和呢？一种解释是每一项是独立的，直接加起来不影响最优解；另一种技巧上的解释是可以裂项相消

后一项不可解，所以只能放弃这一约束，先看前一项的推导

$$\begin{aligned}
&\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))\\
&=\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathbb{E}_{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)}\\
&=\sum_{t=2}^{T}\int_{\mathbf{x}_t}q(\mathbf{x}_t|\mathbf{x}_0)\int_{\mathbf{x}_{t-1}}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)}\,d\mathbf{x}_t\,d\mathbf{x}_{t-1}\\
&=\sum_{t=2}^{T}\int_{\mathbf{x}_t}\int_{\mathbf{x}_{t-1}}q(\mathbf{x}_t|\mathbf{x}_0)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)}\,d\mathbf{x}_t\,d\mathbf{x}_{t-1}\\
&=\sum_{t=2}^{T}\int_{\mathbf{x}_t}\int_{\mathbf{x}_{t-1}}q(\mathbf{x}_t,\mathbf{x}_{t-1}|\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)}\,d\mathbf{x}_t\,d\mathbf{x}_{t-1} && (\star)\\
&=\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_t,\mathbf{x}_{t-1}|\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)}\\
&=\sum_{t=2}^{T}\int_{\mathbf{x}_t}\int_{\mathbf{x}_{t-1}}\left(\int_{\mathbf{x}_1}\cdots\int_{\mathbf{x}_{t-2}}\int_{\mathbf{x}_{t+1}}\cdots\int_{\mathbf{x}_T}q(\mathbf{x}_{1:T}|\mathbf{x}_0)\,d\mathbf{x}_1\cdots d\mathbf{x}_{t-2}\,d\mathbf{x}_{t+1}\cdots d\mathbf{x}_T\right)\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\,d\mathbf{x}_t\,d\mathbf{x}_{t-1} && \text{continue from }(\star)\\
&=\sum_{t=2}^{T}\int_{\mathbf{x}_{1:T}}q(\mathbf{x}_{1: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)}\,d\mathbf{x}_{1:T}\\
&=\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_{1: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)}\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)q(\mathbf{x}_t|\mathbf{x}_0)}\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\sum_{t=2}^{T}\left(\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}+\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}\right)\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left(\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{q(\mathbf{x}_T|\mathbf{x}_0)}\right)\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left(\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}-\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)p_\theta(\mathbf{x}_0|\mathbf{x}_1)}+\log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{q(\mathbf{x}_T|\mathbf{x}_0)}\right)\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left(\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}+\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)-\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right)\\
&=\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left(-\log p_\theta(\mathbf{x}_0)+\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}+\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)-\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\right)\\
&=-\log p_\theta(\mathbf{x}_0)+\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}+\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)-\mathbb{E}_{q(\mathbf{x}_T|\mathbf{x}_0)}\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)}\\
&=-\log p_\theta(\mathbf{x}_0)+D_{KL}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))+\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)-D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_T))
\end{aligned}$$

这个结果有四项

- 第一项是我们想要最小化的负对数似然
- 第二项是非负的KL散度
- 第三项是从$\mathbf{x}_1$重建$\mathbf{x}_0$的交叉熵损失
- 第四项是常量，约束后验$q(\mathbf{x}_T|\mathbf{x}_0)$和先验$p_\theta(\mathbf{x}_T)$保持一致

整理可得

$$\begin{aligned}\log p_\theta(\mathbf{x}_0) &\geq \log p_\theta(\mathbf{x}_0) - D_{KL}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)) \\ &= \underbrace{\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{\text{reconstruction term}, L_0} - \sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{\text{denoising matching term}, L_{t-1}} - \underbrace{D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_T))}_{\text{prior matching term}, L_T}\end{aligned}$$

此结果即为变分下界VLB (Variational Lower Bound ），也叫ELBO（Evidence Lower BOund）

- 一个等价（但推导反向相反）的证明可以参见lilianweng的整理，那一形式在论文中更为常见，直接从隐变量条件分布的变分下界开始推导，能得到一样的结论
- 需要注意的是，从变分下界开始的推导（即上述等价的反向推导）能够快速直接写出具体形式，也能够以简洁的形式表达出变分下界的意义，事实上很多论文会用它

这里最后一项$L_T$为常量，只要优化前两项，就能实现VLB的优化

接下来，我们分别看$L_0$和$L_{t-1}$的具体形式

## 计算denoising matching term

首先计算KL散度项，由于$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$是个高斯分布，所以可以定义

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \color{blue}{\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)}, \color{red}{\tilde{\beta}_t\mathbf{I}})$$

同时，我们还有参数化的高斯分布

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t,t), \mathbf{\Sigma}_\theta(\mathbf{x}_t,t))$$

这里DDPM将$\mathbf{\Sigma}_\theta(\mathbf{x}_t,t)$人为设置为一个和时间步相关的常量$\sigma_t^2$，这个常量可以是$\sigma_t^2=\beta_t$，也可以是下面求得的$\sigma_t^2=\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$，两者在实验上有着类似的结果，以下使用$\sigma_t^2=\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$推导

- 这两个值的选取实际上对应于DPM论文中reverse process entropy的上界和下界，即在论文2.6节所提的

$$\mathcal{H}_q\left(\mathbf{X}^{(t)}|\mathbf{X}^{(t-1)}\right)+\mathcal{H}_q\left(\mathbf{X}^{(t-1)}|\mathbf{X}^{(0)}\right)-\mathcal{H}_q\left(\mathbf{X}^{(t)}|\mathbf{X}^{(0)}\right)\leq\mathcal{H}_q\left(\mathbf{X}^{(t-1)}|\mathbf{X}^{(t)}\right)\leq\mathcal{H}_q\left(\mathbf{X}^{(t)}|\mathbf{X}^{(t-1)}\right)$$

- 这里下界是$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$的熵，而上界是$q(\mathbf{x}_t|\mathbf{x}_{t-1})$的熵，它们都是已知的高斯分布
- 对于任意$D$元高斯分布$p(\mathbf{x})=\mathcal{N}(\mathbf{x};\mu,\mathbf{\Sigma})$，可以求得其熵

$$\begin{aligned}\mathcal{H}_p(\mathbf{x}) &= -\int p(\mathbf{x})\log p(\mathbf{x})d\mathbf{x} \\ &= -\mathbb{E}_p\log\left((2\pi)^{-D/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right) \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\mathbb{E}_p\left((\mathbf{x}-\mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right) \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\mathbb{E}_p\left(\operatorname{tr}\left((\mathbf{x}-\mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right) && \text{; A real number can be viewed as a 1x1 matrix.} \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\mathbb{E}_p\left(\operatorname{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right)\right) && \text{; Cyclic property of the trace.} \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\operatorname{tr}\left(\mathbb{E}_p\left(\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right)\right) && \text{; Linearity of the trace.} \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbb{E}_p\left((\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right)\right) \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\right) \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{1}{2}\operatorname{tr}\mathbf{I} \\ &= \frac{D}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{\Sigma}|+\frac{D}{2}\end{aligned}$$

  - Cyclic property of the trace
    - $\operatorname{tr}(\mathbf{ABC})=\operatorname{tr}(\mathbf{BCA})=\operatorname{tr}(\mathbf{CAB})$
  - Linearity of the trace
    - $\mathrm{E}[\operatorname{tr}(A)]=\mathrm{E}\left[\sum_{i=1}^n a_{ii}\right]=\sum_{i=1}^n\mathrm{E}[a_{ii}]=\operatorname{tr}\left(\begin{bmatrix}\mathrm{E}[a_{11}] & \dots & \mathrm{E}[a_{1n}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[a_{n1}] & \dots & \mathrm{E}[a_{nn}]\end{bmatrix}\right)=\operatorname{tr}(\mathrm{E}[A])$ .
- 注意到结果只和协方差$\mathbf{\Sigma}$相关，从而到达上界时方差对应正向过程$q(\mathbf{x}_t|\mathbf{x}_{t-1})$方差$\sigma_t^2=\beta_t$，到达下界时方差对应条件为$\mathbf{x}_0$的反向过程$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$方差$\sigma_t^2=\tilde{\beta}_t$

接下来计算具体参数，由于$q(\mathbf{x}_t|\mathbf{x}_{t-1})=\mathcal{N}(\mathbf{x}_t;\sqrt{\alpha_t}\mathbf{x}_{t-1},\beta_t\mathbf{I})$， $q(\mathbf{x}_t|\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_t;\sqrt{\bar\alpha_t}\mathbf{x}_0,(1-\bar\alpha_t)\mathbf{I})$，假设所求高斯分布为$D$元高斯分布

$$
\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\
&= \frac{\frac{1}{(2\pi)^{D/2}(\beta_t)^{D/2}}\exp\left(-\frac{\left(\mathbf{x}_t-\sqrt{1-\beta_t}\mathbf{x}_{t-1}\right)^2}{2\beta_t}\right)\frac{1}{(2\pi)^{D/2}(1-\bar\alpha_{t-1})^{D/2}}\exp\left(-\frac{\left(\mathbf{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\mathbf{x}_0\right)^2}{2(1-\bar\alpha_{t-1})}\right)}{\frac{1}{(2\pi)^{D/2}(1-\bar\alpha_t)^{D/2}}\exp\left(-\frac{\left(\mathbf{x}_t-\sqrt{\bar\alpha_t}\mathbf{x}_0\right)^2}{2(1-\bar\alpha_t)}\right)} \\
&= \frac{1}{(2\pi)^{D/2}(\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t)^{D/2}}\exp\left(-\frac{1}{2}(\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t}+\frac{(\mathbf{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\mathbf{x}_0)^2}{1-\bar\alpha_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar\alpha_t}\mathbf{x}_0)^2}{1-\bar\alpha_t})\right) \\
&= \frac{1}{(2\pi)^{D/2}(\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t)^{D/2}}\exp\left(-\frac{1}{2}(\frac{\mathbf{x}_t^2-2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}+\alpha_t\mathbf{x}_{t-1}^2}{\beta_t}+\frac{\mathbf{x}_{t-1}^2-2\sqrt{\bar\alpha_{t-1}}\mathbf{x}_0\mathbf{x}_{t-1}+\bar\alpha_{t-1}\mathbf{x}_0^2}{1-\bar\alpha_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar\alpha_t}\mathbf{x}_0)^2}{1-\bar\alpha_t})\right) \\
&= \frac{1}{(2\pi)^{D/2}(\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t)^{D/2}}\exp\left(-\frac{1}{2}((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar\alpha_{t-1}})\mathbf{x}_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{2\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}\mathbf{x}_0)\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0))\right)
\end{aligned}
$$

其中$C(\mathbf{x}_t,\mathbf{x}_0)$在均值和方差的计算中不需要，故略去

整理成高斯分布概率密度函数可得

$$
\begin{aligned}
\tilde\beta_t &= 1/(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar\alpha_{t-1}}) = 1/(\frac{\alpha_t-\bar\alpha_t+\beta_t}{\beta_t(1-\bar\alpha_{t-1})}) = \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t \\
\tilde\mu_t(\mathbf{x}_t,\mathbf{x}_0) &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar\alpha_{t-1}}) \\
&= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}\mathbf{x}_0)\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t \\
&= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\mathbf{x}_t+\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}\mathbf{x}_0
\end{aligned}
$$

现在，我们有了$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_{t-1};\tilde\mu(\mathbf{x}_t,\mathbf{x}_0),\tilde\beta_t\mathbf{I}),p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)=\mathcal{N}(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\sigma_t^2\mathbf{I})=\mathcal{N}(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\tilde\beta_t\mathbf{I})$，所以可以直接计算KL散度了

- 对于多元$D$维高斯分布$p_1(\mathbf{x})=\mathcal{N}(\mathbf{x};\mu_1,\mathbf{\Sigma}_1),p_2(\mathbf{x})=\mathcal{N}(\mathbf{x};\mu_2,\mathbf{\Sigma}_2)$，它们之间的KL散度为

$$
\begin{aligned}
D_{KL}(p_1(\mathbf{x})\parallel p_2(\mathbf{x})) &= \int p_1(\mathbf{x})\frac{\log p_1(\mathbf{x})}{\log p_2(\mathbf{x})}d\mathbf{x} \\
&= \mathbb{E}_{p_1}\Big(\log p_1(\mathbf{x})-\log p_2(\mathbf{x})\Big) \\
&= \mathbb{E}_{p_1}\Big(-\frac{1}{2}\log|\mathbf{\Sigma}_1|-\frac{1}{2}(\mathbf{x}-\mu_1)^\top\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}\log|\mathbf{\Sigma}_2|+\frac{1}{2}(\mathbf{x}-\mu_2)^\top\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mu_2)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\mathbb{E}_{p_1}\Big((\mathbf{x}-\mu_1)^\top\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mu_1)\Big)+\frac{1}{2}\mathbb{E}_{p_1}\Big((\mathbf{x}-\mu_2)^\top\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mu_2)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\mathbb{E}_{p_1}\Big(\operatorname{tr}\big((\mathbf{x}-\mu_1)^\top\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mu_1)\big)\Big)+\frac{1}{2}\mathbb{E}_{p_1}\Big(\operatorname{tr}\big((\mathbf{x}-\mu_2)^\top\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mu_2)\big)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\mathbb{E}_{p_1}\Big(\operatorname{tr}\big(\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^\top\big)\Big)+\frac{1}{2}\mathbb{E}_{p_1}\Big(\operatorname{tr}\big(\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mu_2)(\mathbf{x}-\mu_2)^\top\big)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_1^{-1}\mathbb{E}_{p_1}\big((\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^\top\big)\Big)+\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_2^{-1}\mathbb{E}_{p_1}\big((\mathbf{x}-\mu_2)(\mathbf{x}-\mu_2)^\top\big)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_1\Big)+\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_2^{-1}\mathbb{E}_{p_1}\big(\mathbf{x}\mathbf{x}^\top-\mu_2\mathbf{x}^\top-\mathbf{x}\mu_2^\top+\mu_2\mu_2^\top\big)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{1}{2}\operatorname{tr}\mathbf{I}+\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_2^{-1}\big(\mathbf{\Sigma}_1+\mu_1\mu_1^\top-\mu_2\mu_1^\top-\mu_1\mu_2^\top+\mu_2\mu_2^\top\big)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{D}{2}+\frac{1}{2}\operatorname{tr}\big(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\big)+\frac{1}{2}\operatorname{tr}\Big(\mathbf{\Sigma}_2^{-1}\mu_1\mu_1^\top-\mathbf{\Sigma}_2^{-1}\mu_2\mu_1^\top-\mathbf{\Sigma}_2^{-1}\mu_1\mu_2^\top+\mathbf{\Sigma}_2^{-1}\mu_2\mu_2^\top\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{D}{2}+\frac{1}{2}\operatorname{tr}\big(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\big)+\frac{1}{2}\operatorname{tr}\Big(\mu_1^\top\mathbf{\Sigma}_2^{-1}\mu_1-\mu_1^\top\mathbf{\Sigma}_2^{-1}\mu_2-\mu_2^\top\mathbf{\Sigma}_2^{-1}\mu_1+\mu_2^\top\mathbf{\Sigma}_2^{-1}\mu_2\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{D}{2}+\frac{1}{2}\operatorname{tr}\big(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\big)+\frac{1}{2}\operatorname{tr}\Big((\mu_2-\mu_1)^\top\mathbf{\Sigma}_2^{-1}(\mu_2-\mu_1)\Big) \\
&= \frac{1}{2}\log\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}-\frac{D}{2}+\frac{1}{2}\operatorname{tr}\big(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\big)+\frac{1}{2}(\mu_2-\mu_1)^\top\mathbf{\Sigma}_2^{-1}(\mu_2-\mu_1)
\end{aligned}
$$

。

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}\big((\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\big) \\
&= \mathbb{E}\big(\mathbf{x}\mathbf{x}^\top - \mu\mathbf{x}^\top - \mathbf{x}\mu^\top + \mu\mu^\top\big) \\
&= \mathbb{E}(\mathbf{x}\mathbf{x}^\top) - \mu\mu^\top - \mu\mu^\top + \mu\mu^\top \\
&= \mathbb{E}(\mathbf{x}\mathbf{x}^\top) - \mu\mu^\top \\
\Rightarrow \mathbb{E}(\mathbf{x}\mathbf{x}^\top) &= \boldsymbol{\Sigma} + \mu\mu^\top
\end{aligned}
$$

因此，要求的KL散度为

$$
\begin{aligned}
L_{t-1} &= D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\
&= D_{KL}(\mathcal{N}(\mathbf{x}_{t-1};\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I}) \parallel \mathcal{N}(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\tilde{\beta}_t\mathbf{I})) \\
&= \frac{1}{2}\Big(\log\frac{|\tilde{\beta}_t\mathbf{I}|}{|\tilde{\beta}_t\mathbf{I}|} - D + \mathrm{tr}\,\big((\tilde{\beta}_t\mathbf{I})^{-1}(\tilde{\beta}_t\mathbf{I})\big) + (\mu_\theta-\tilde{\mu}_t)(\tilde{\beta}_t\mathbf{I})^{-1}(\mu_\theta-\tilde{\mu}_t)^\top\Big) \\
&= \frac{1}{2}\Big(-D + D + (\mu_\theta-\tilde{\mu}_t)^\top(\tilde{\beta}_t\mathbf{I})^{-1}(\mu_\theta-\tilde{\mu}_t)\Big) \\
&= \frac{1}{2\tilde{\beta}_t}\Big((\mu_\theta-\tilde{\mu}_t)^\top(\mu_\theta-\tilde{\mu}_t)\Big) \\
&= \frac{1}{2\tilde{\beta}_t}\|\mu_\theta-\tilde{\mu}_t\|^2
\end{aligned}
$$

由于我们已知了$\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)$的形式，所以我们可以假定$\mu_\theta(\mathbf{x}_t,t)$具有同样的形式，DPM中使用的是$\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0$，从而$\mu_\theta(\mathbf{x}_t,t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_\theta(\mathbf{x}_t,t)$，这种方式方差较大，生成效果较差；而直接学习一个$\mu_\theta$来估计$\tilde{\mu}_t$也是可行的；但是DDPM通过引入重参数化技巧进一步简化了优化目标

由于$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t,\ \bar{\epsilon}_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})$，所以有$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t)$，进而$\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)$可以进一步化简

$$
\begin{aligned}
\tilde{\mu}_t &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t) \\
&= \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_t\Big)
\end{aligned}
$$

于是DDPM使用了$\mu_\theta(\mathbf{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\Big)$，从而$L_{t-1}$可以进一步推导

$$
\begin{aligned}
L_{t-1} &= \frac{1}{2\tilde{\beta}_t}\|\mu_\theta-\tilde{\mu}_t\|^2 \\
&= \frac{1}{2\tilde{\beta}_t}\Big\|\frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\Big) - \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_t\Big)\Big\|^2 \\
&= \frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\Big\|\epsilon_\theta(\mathbf{x}_t,t) - \bar{\epsilon}_t\Big\|^2 \\
&= \frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\Big\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t, t) - \bar{\epsilon}_t\Big\|^2
\end{aligned}
$$

## 计算reconstruction term

接下来看看$L_0$如何计算

DDPM对$L_0$的设计着重于设计一种对于离散数据（0~255表示的图像）而言无损的压缩编码（lossless codelength）方式，并直接计算了似然

具体而言，DDPM首先假设图像被映射到了$[-1,1]$区间，因此将样本$\mathbf{x}_0$的似然使用$\mathbf{x}_0$附近的一个小区间内的积分来定义，每个维度（$D$维）认为是独立并单独计算

$$
p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D}\int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)}\mathcal{N}(x;\mu_\theta^i(\mathbf{x}_1,1),\delta_1^2)dx
$$

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

这一似然只需要在训练时计算，在采样时就不再需要了，采样时直接以$\mu_\theta(\mathbf{x}_1,1)$作为最终结果，这等价于采样最后一步不添加噪声

## Simplified training objective

虽然变分下界可以直接优化，但是DDPM选择了形式上和实现上都更加简单的优化目标，注意以下目标损失实际上包含了从$\mathbf{x}_0$到$\mathbf{x}_T$所有的$T$步，注意此处的$\epsilon$相当于上述推导中的$\bar{\epsilon}_t$

$$L_{simple}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right]$$

其中$t$从1到$T$均匀采样

$t=1$对应于$L_0$，但应该只是个简单的对应

$t>1$实际上是去掉系数后的$L_{t-1}$和NCSN denoising **score matching** model的loss weighting做法类似

$t=T$不需要优化，所以没出现

**Algorithm 1** Training

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

**Algorithm 2** Sampling

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

[image]

# 采样过程

采样过程实际上就是$\mathbf{x}_{t-1} = \tilde{\mu}_t + \sigma_t\mathbf{z}$

为什么采样过程不能$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t,t))$一步得到？一是为了保持采样过程的一阶马尔可夫性，防止生成的分布偏移；二是为了引入噪声，增大采样的多样性

# 实验相关

DDPM的实验在CIFAR10（Inception scores, FID scores, nll/lossless codelengths），CelebA-HQ 256 $\times$ 256，LSUM 256 $\times$ 256 上进行

直接优化VLB时codelength最好（显然），但是生成质量上用simplified objective好

可以在latent space插值，再denoise回样本空间！

The choice of the scheduling function can be arbitrary, as long as it provides a near-linear drop in the middle of the training process and subtle changes around $t=0$ and $t=T$.