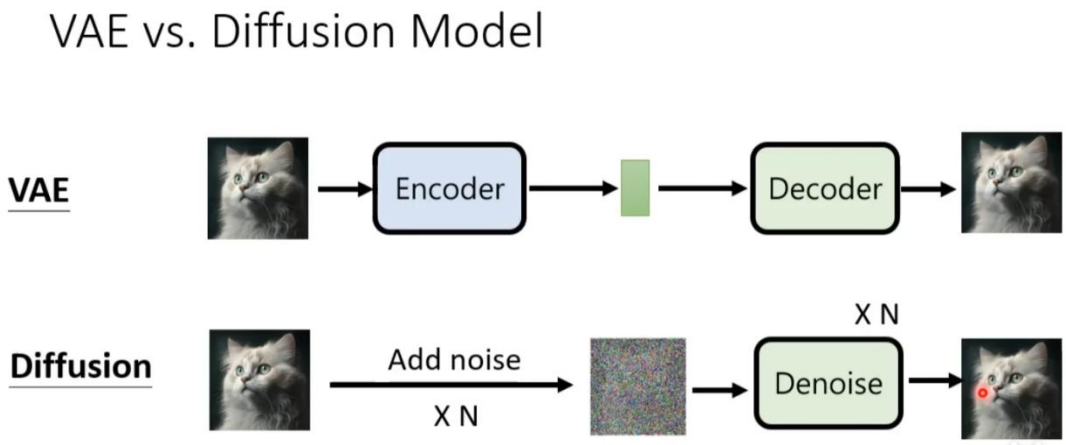
二. 模型背景与模型数学推导

我们刚刚已经介绍了图像生成这一问题，并且我们已经知道了扩散模型“图像生成”这一步骤是说的 denoise 过程。那他是怎么实现从一个随机的噪声中逐渐去噪变成想要的图像的呢？




主要演算法包含以下两个：

**Algorithm 1** Training

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

**Algorithm 2** Sampling

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

他看起来非常非常的简单

Repeat 就是说一直做第二行到第五行的工作，直到 converge 为止

$x_0$: clean image

$\varepsilon$: noise

我们需要首先在一个图像数据库中取样一个清晰的 x0 图像，随后取样一个小 t，小 t 是在 1 到大 T 中的一个整数。大 T 是一个比较大的数字，比如 1k

随后，我们在高斯分布中随机取样一个噪声，均值与方差是 0 和 1.

第五行算法相当复杂，我们一步一步拆分来看

第五行是说，我们需要把原图与噪声做一个加权，

***Training***

$\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$

$x_0$

$\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$

$\varepsilon$

Sample $t$

$\sqrt{\bar{\alpha}_t}$  $+\sqrt{1-\bar{\alpha}_t}$ $\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$ $=$ 

$x_0$ $\varepsilon$

这里的 α 1 到阿尔法 t 都是事先给好的，这样我们就得到了一个带噪声的图像，(这里的阿尔法 1 到阿尔法 t 是越来越小的)，也就是说，给的 t 越大，噪声越凶狠，原图像约不清晰。得到噪声图像以后，我们发现他和 t 一起被放入了一个伊普西隆 θ，这个就是 noise predictor

**Algorithm 1** Training

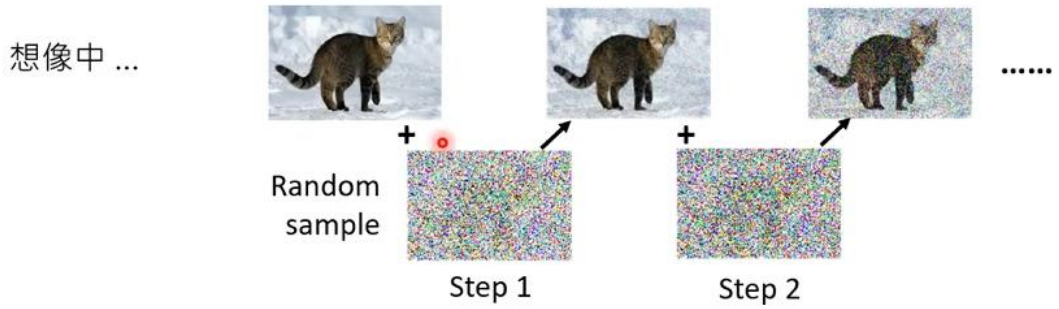
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ ⟵ sample clean image
3: $t \sim \text{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ⟵ sample a noise
5: Take gradient descent step on
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0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

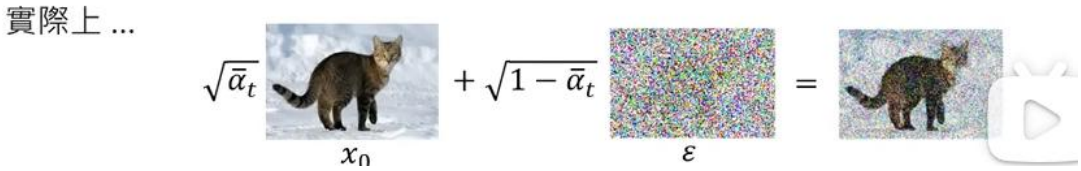
Noisy image

Noise predictor

这个说的就是我预测噪声与原噪声的范数平方。梯度就是在这上面计算并训练的。也就是，给定噪声图像与 t，我们相当于要去预测混入的噪音是怎样的。伊普西隆 θ 就

是用于预测这个噪音的。



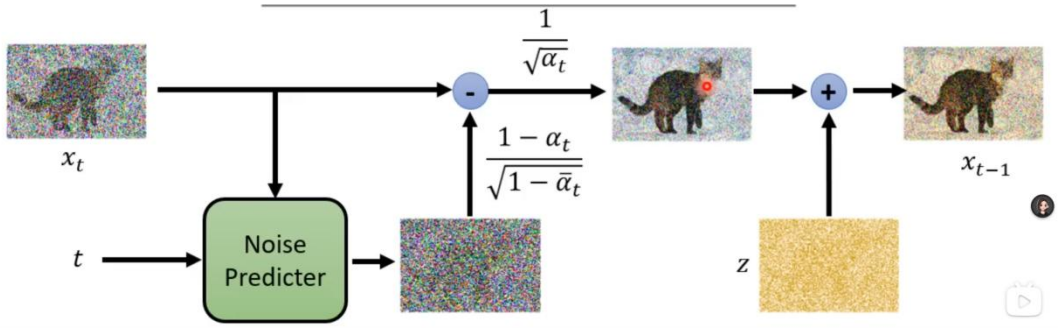

但是按照我们第一部分讲的 这个噪声应该是一步一步加上的啊！ 为什么训练过程是直接加上了一个噪声呢!

## 这是我们的第一个疑问

那 Sampling 过程， 也就是产生图像的过程是什么样的呢?

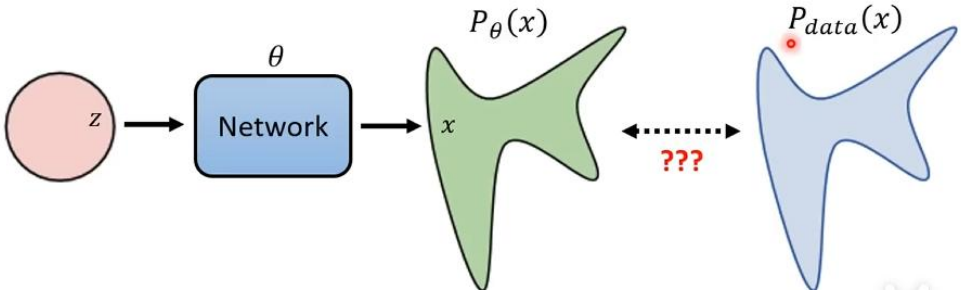
我们首先取一个纯碎都是噪声的图，这个噪声符合（0，1）正态分布（高斯分布），

随后循环大 T 步，我们发现这里很奇怪的现象，他这里又重新采样了一个高斯分布，在这里加权放进来图片中，也就是说，按照我们原本的思路，给定噪声图 xt 与 t，我们预测出当前的噪声，用 xt 减去后用这两个加权，会得到一个相对更清晰的图片，但是为什这里又要加上一个噪声呢?




## 这是我们的第二个疑问

想要解决这个问题，我们就必须要从长计议



回到我们刚刚讲的图像生成的任务，找到 network，让他能把 z 空间投影出来的空间与真实图像空间最一致。衡量标准是什么呢？就是我们之前所说的，找最大似然估计的方法。

给定训练训练数据，取样一些图片，去找一个 θ，让他能够生成的图片 p θ 是正确的图片的纪律最大，也就是：

$$\theta^* = arg\max_{\theta} \prod_{i=1}^{m} P_\theta(x^i)$$

这个步骤可以写成以下形式，取 log 不影响，换成求和形式，变换成求期望，随后减去一个与 θ 无关的值，最后推导出的是最小散度，这就是我们的最终目标（所有的影像生成模型）

Sample $\{x^1, x^2, \dots, x^m\}$ from $P_{data}(x)$

$$\theta^* = arg\max_{\theta} \prod_{i=1}^{m} P_\theta(x^i) = arg\max_{\theta} log \prod_{i=1}^{m} P_\theta(x^i)$$

$$= arg\max_{\theta} \sum_{i=1}^{m} logP_\theta(x^i) \approx arg\max_{\theta} E_{x\sim P_{data}}[logP_\theta(x)]$$

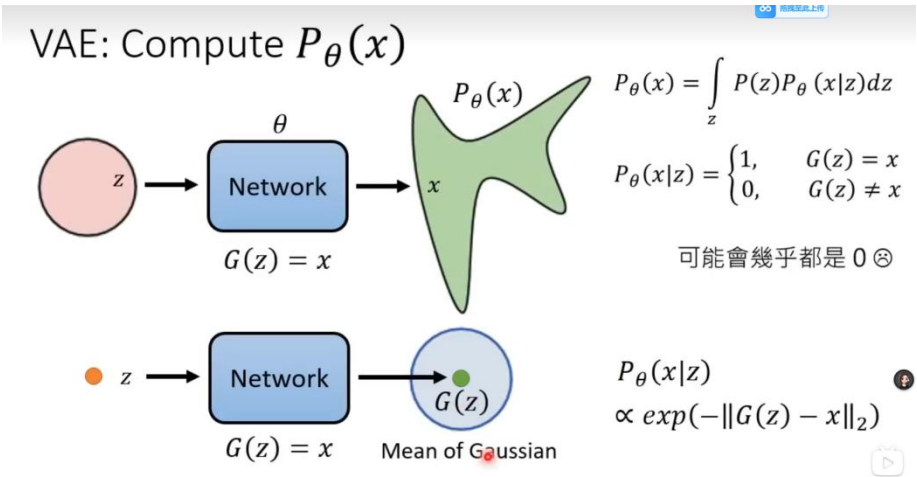
$$= arg\max_{\theta} \int_x P_{data}(x) logP_\theta(x)dx - \int_x P_{data}(x) logP_{data}(x)dx$$

(not related to $\theta$)

$$= arg\max_{\theta} \int_x P_{data}(x) log\frac{P_\theta(x)}{P_{data}(x)}dx = arg\min_{\theta} KL(P_{data}||P_\theta)$$

Difference between $P_{data}$ and $P_\theta$

那么 p θ 如何计算？



我们可以写成上面这个西塔 z 的形式，像计算 x 的产生概率，相当于要计算给定每一个 z 产生 x 的概率的积分嘛，但是这里的给定 z 的 p 西塔 x 怎么确定呢，我们当然可以选用这个规则，但是这样计算可能会让我们结果都是 0，因为几乎不可能让生成的图片和原图完全相等，所以 VAE 给出了这样的解决办法，network 预测的这个高斯分布的均值，和对应 x 点做差取二范数，随后取复数再取对数，这样能保证大于 0 的同时，举例越小，px 越大

虽然找到了计算原则，但实际上这里的 p 西塔还是不好优化的，但它有一个下届可以最大化

VAE: Lower bound of $logP(x)$

$$logP_\theta(x) = \int_z q(z|x)logP(x)dz \quad q(z|x) \text{ can be any distribution}$$

$$= \int_z q(z|x)log\left(\frac{P(z,x)}{P(z|x)}\right)dz = \int_z q(z|x)log\left(\frac{P(z,x)}{q(z|x)}\frac{q(z|x)}{P(z|x)}\right)dz$$

$$= \int_z q(z|x)log\left(\frac{P(z,x)}{q(z|x)}\right)dz + \underbrace{\int_z q(z|x)log\left(\frac{q(z|x)}{P(z|x)}\right)dz}_{KL(q(z|x)||P(z|x))} \quad \geq 0$$

$$\geq \int_z q(z|x)log\left(\frac{P(z,x)}{q(z|x)}\right)dz$$

也就是说，我们对 p 西塔取 log 以后，他一定可以写成这种这形式，这里的给定 xq 的条件概率可以是任何分布，这个等式都可以满足。具体原因比较复杂

随后呢我们都学过全概率公式对吧 全概率公式就可以把这个展开成这个形式，随后天才般的构造操作出现了，上下同城 q，把 log 相乘拆开，得到了这个式子

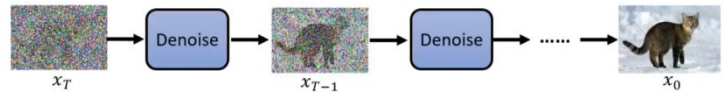
然后我们惊奇发现这个就是给定 x 的 qz 和给定 x 的 pz 的散度，散度是一定大于 0 的啊，前面这一项是可以优化的，那我们就可以优化这个下届。这一项可以写成

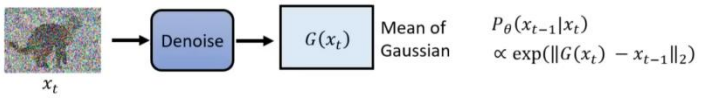
$$E_{q(z|x)}[log\left(\frac{P(x,z)}{q(z|x)}\right)]$$



而对于扩散模型。

DDPM: Compute $P_\theta(x)$




$$P_\theta(x_{t-1}|x_t) \propto \exp(\|G(x_t) - x_{t-1}\|_2)$$

$$P_\theta(x_0) = \int_{x_1:x_T} P(x_T)P_\theta(x_{T-1}|x_T)\dots P_\theta(x_{t-1}|x_t)\dots P_\theta(x_0|x_1)dx_1:x_T$$

p 西塔的标准和 VAE 一致，而 x 西塔，由于去噪是个马尔可夫链过程，p 西塔被写成这个形式，这样扩散模型和 vae 就被统一起来了，vae 的 q 是 decoder，而扩散模型则是 denoise

DDPM: Lower bound of $logP(x)$

**VAE** Maximize $logP_\theta(x)$ ⟶ Maximize $\mathrm{E}_{q(z|x)}[log\left(\frac{P(x,z)}{q(z|x)}\right)]$

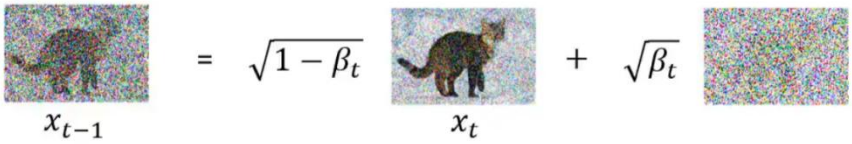
**DDPM** Maximize $logP_\theta(x_0)$ ⟶ Maximize $\mathrm{E}_{q(x_1:x_T|x_0)}[log\left(\frac{P(x_0:x_T)}{q(x_1:x_T|x_0)}\right)]$

但是

$$q(x_1:x_T|x_0) = q(x_1|x_0)q(x_2|x_1)\dots q(x_T|x_{T-1})$$

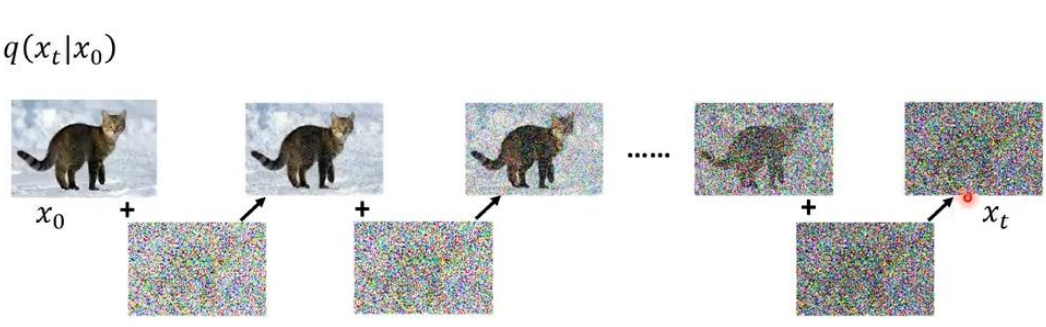
这个式子怎么被算出来呢?

我们知道 xt-1 和 xt 的关系是这样的， β t 是一串预先给定的超参数，我们实际上就是在做加权

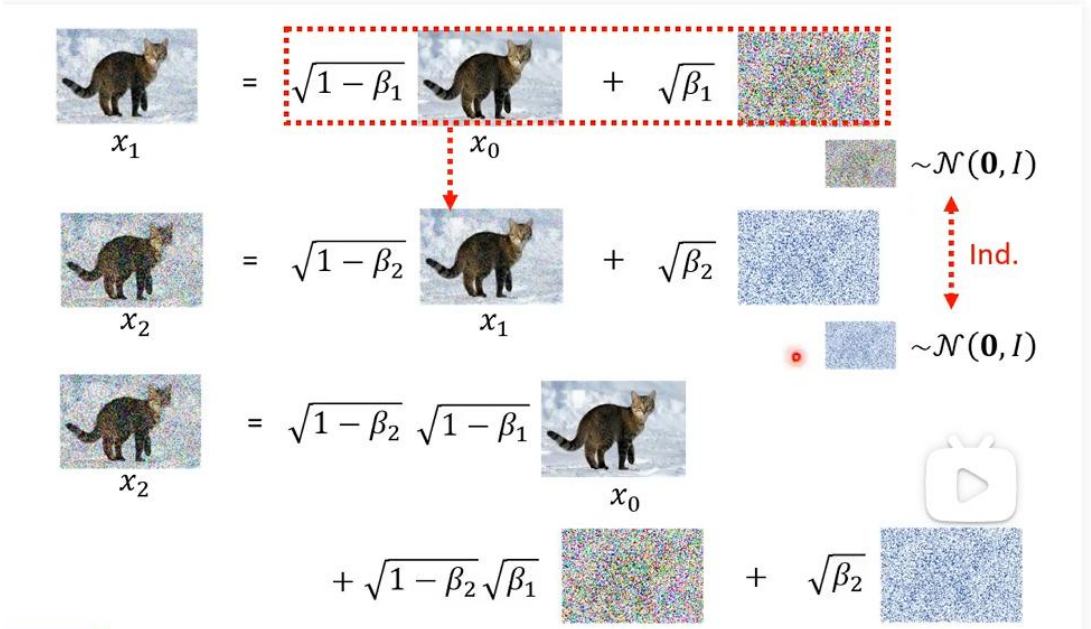



$x_{t-1}$ $x_t$ $\beta_1, \beta_2, \dots, \beta_T$

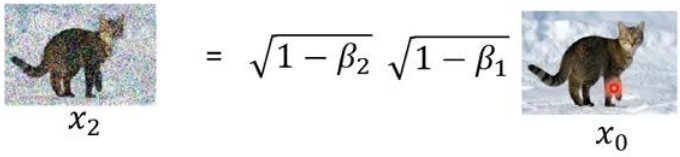
也就是

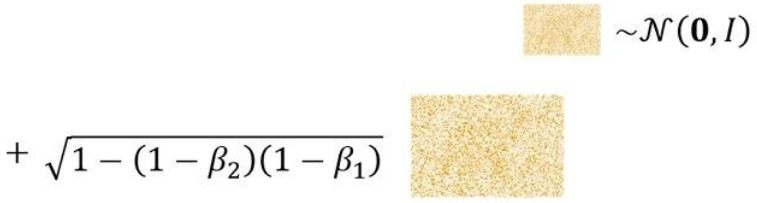



但实际上，这个是可以一步直接就被算出来的

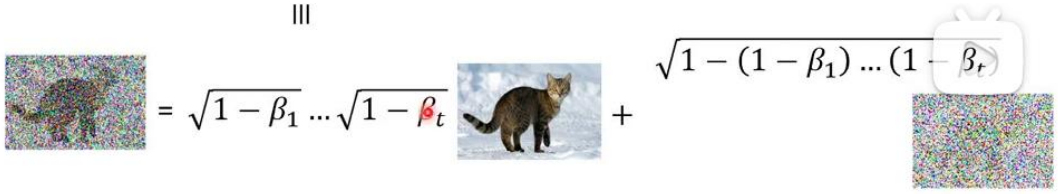



这里的两个噪声，其实可以简化成只去一次，并且加上这个权重




以此类推，整个过程可以被做成




那这里我们之前 train 算法里面的问题就被解决了

**Algorithm 1** Training

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

那我们的这个也就可以化简了，不过化简的过程有点不简单啊，我们给大家拉托大的：

$$\text{Maximize } \mathrm{E}_{q(x_1:x_T|x_0)}[log\left(\frac{P(x_0:x_T)}{q(x_1:x_T|x_0)}\right)]$$

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \tag{47}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)}\right] \tag{50}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)}\right] \tag{51}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right] \tag{52}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}}\right] \tag{53}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \tag{54}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \tag{55}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \tag{56}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}\right] \tag{57}$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}} \tag{58}$$

Maximize $\mathrm{E}_{q(x_1:x}$

化成了这个

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] \ -KL\big(q(x_T|x_0)||P(x_T)\big)$$

$$-\sum_{t=2}^{T}\mathrm{E}_{q(x_t|x_0)}\big[KL\big(q(x_{t-1}|x_t,x_0)||P(x_{t-1}|x_t)\big)\big]$$

第二项是不用管的，和网络的参数无关，第一项和第三项的处理方法是一样的

里面唯一不会算的东西就是给定 xt 和 x0，qxt-1 的条件概率


我们会算的是什么呢？给定加噪声的过程，我们一步一步迭代肯定可以知道 xt-1 吧, 但现在这一项相当于什么？完全不给加噪声的方式, 只给你加好噪声的图片和原图, 让你找 xt-1 的分布

通过第二托大的推导

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{71}$$

$$=\frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0,(1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I})} \tag{72}$$

$$\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \tag{73}$$

$$=\exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\right\} \tag{74}$$

$$=\exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}+\alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}}+C(\boldsymbol{x}_t,\boldsymbol{x}_0)\right]\right\} \tag{75}$$

$$\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t}+\frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t}+\frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}}-\frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\right\} \tag{76}$$

$$=\exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{77}$$

$$=\exp\left\{-\frac{1}{2}\left[\frac{\alpha_t(1-\bar{\alpha}_{t-1})+1-\alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{78}$$

$$=\exp\left\{-\frac{1}{2}\left[\frac{\alpha_t-\bar{\alpha}_t+1-\alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{79}$$

$$=\exp\left\{-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{80}$$

$$=\exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\boldsymbol{x}_{t-1}\right]\right\} \tag{81}$$

$$=\exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \tag{82}$$

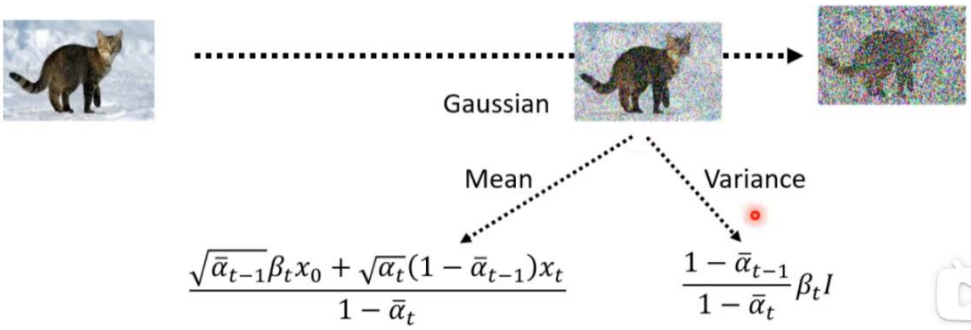
$$=\exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \tag{83}$$

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1};\underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)},\underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) \tag{84}$$

得到了这也是一个高斯分布，均值和方差如下

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] \; -KL\big(q(x_T|x_0)||P(x_T)\big)$$

$$-\sum_{t=2}^{T}\mathrm{E}_{q(x_t|x_0)}\big[KL\big(q(x_{t-1}|x_t,x_0)||P(x_{t-1}|x_t)\big)\big]$$

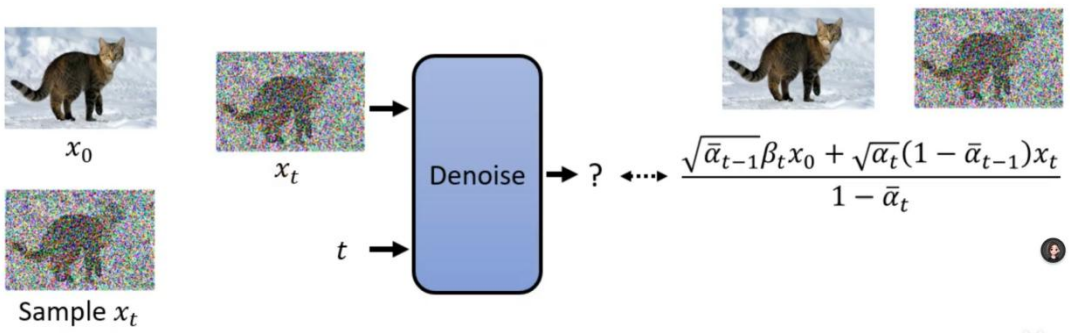



两个高斯分布的距离有具体的计算公式，也就可以极大化了

Recall that the KL Divergence between two Gaussian distributions is:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)\parallel\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y))=\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|}-d+\mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x)+(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)\right]$$

最后，也就是说我们给定 x0，xt 与 t，通过去噪的过程，我们可以预测一个这个，也就是两个的加权平均（这个本质是我们刚才找到高斯分布的均值），这个继续化简，也就得到了最下面的结果




$$\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)$$

也就是需要预测的结果

回到

**Algorithm 2** Sampling

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

取样这个算法，我们发现还有一个地方没有被解决，加这个 z 干什么，本来我直接预测概率最大的高斯分布的均值就可以了啊为啥还要加扰动呢?
对于这个说法，我是没想明白的，这里给大家李彦宏老师的观点，就是说对于一个生成模型而言，如果每次都去最大概率不加扰动，就会导致生成的结果一样，但实际上我们肯定不希望如此，所以我们给每一步去噪预测，都加一个扰动，这样就可以让我生成的结果有所差异。