How to Teach Your Cat DDPM

With Basic Mathematical Statistics

Catman Jr.

December 24, 2024

Introduction

As you can see, this is actually a note from my DDPM learning process. The beginning of everything comes from a piece of knowledge article, the blogger is 撒旦-cc, on Zhihu. You may find that the structure of this note is very similar to that of the original blog. I did 'reproduce' the famous Understanding Diffusion Models: A Unified Perspective, so I do not consider this article to be my original work. However, when I type my manuscript into latex, I combined the relevant lecture of Prof. Hung-Yi Lee and some contents of the original paper, and independently restored some derivation details and background knowledge omitted from the paper. Therefore, I named the article "How to Teach Your Cat DDPM". I hope my notes will be of considerable help to any non-math major who has only basic knowledge of calculus, linear algebra, and mathematical statistics (stochastic process).

The denoising diffusion probability model is a generative framework that relies on the progressive removal of noise from images. This model is grounded in the concept of a diffusion process, wherein noise is incrementally removed from an image over time. The architecture of this model comprises two key processes: a forward process, during which noise is systematically introduced into the image to generate training samples, and a reverse process, where noise is progressively eliminated to reconstruct the original image. In the original paper, the model is trained by optimizing a loss function that evaluates the quality of noise rather than directly assessing pixel-wise differences within the image. At the conclusion of this section, the original authors introduce two core algorithms. Therefore, it is essential to understand the step-by-step mechanism of how noise is added (forward process) and the rationale behind focusing on noise quality (reverse process).

In this paper, we will provide a detailed explanation of the denoise diffusion probability model, including the forward and reverse processes and the loss function used for training. I'm trying to add more details and some further explanation for non-math student like myself based on the famous work "Understanding the denoising diffusion probabilistic model". I hope this paper can help anyone with only basic probability,

statistics and ML knoledges understand the denoise diffusion probability model better. Out of the paper, I strongly recommend the original paper 'Understanding the denoising diffusion probabilistic model'(Calvin Luo, 2022) and 'Denoise diffusion probability models' (Ho et al., 2020) for more details.

Algorithm 1 Training

repeat
  $x_0 \sim q(x_0)$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  Take gradient descent step on
    $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2$
until converged

Algorithm 2 Sampling

$x_T \sim \mathcal{N}(0, I)$
for $t = T, \dots, 1$ **do**
  $z \sim \mathcal{N}(0, I)$ if$t > 1$ else $z = 0$
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$
end for
return x_0

I Forward Process(Sampling)

The forward process of the denoise diffusion probability model is a process of adding noises to the input image, where the original image and the noised image are linearly dependant. The original paper of DDPM used a piece of elegant but abstract pseudo code to show the sampling process for any timestamp t (Algorithm 2). And we try to find out why the writers could write out the sampling algorithm in the original paper by finishing this part.

Pf:

Let $\mathbf{x}_0$ be the original image.

Then $\mathbf{x}_{t-1} \to \mathbf{x}_t$ is a process of adding noises to the input image, where

$\mathbf{x}_{t-1}, \mathbf{x}_t$ are linearly dependant.

$$\text{Let } \mathbf{x}_t = a_t \mathbf{x}_{t-1} + b_t \varepsilon_t \text{ where } a_t, b_t \text{ are constant and the added noise } \varepsilon_t \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

Hence we added noises

$\implies$ $\mathbf{x}_{t-1}$ has more information than $\mathbf{x}_t$

$\implies$ a_t, b_t are attenuation coefficients.

$\implies$ $a_t, b_t \in (0, 1)$

$$\begin{aligned}
\text{Then } \mathbf{x}_t &= a_t \mathbf{x}_{t-1} + b_t \varepsilon_t \\
&= a_t(a_{t-1}\mathbf{x}_{t-2} + b_{t-1}\varepsilon_{t-1}) + b_t \varepsilon_t \\
&= a_t a_{t-1} \mathbf{x}_{t-2} + a_t b_{t-1} \varepsilon_{t-1} + b_t \varepsilon_t \\
&= (a_t \cdots a_1)\mathbf{x}_0 + (a_t \cdots a_2) b_1 \varepsilon_1 + (a_t \cdots a_3) b_2 \varepsilon_2 + \cdots + a_t b_{t-1} \varepsilon_{t-1} + b_t \varepsilon_t
\end{aligned} \tag{2}$$

while each term of the $b_i \varepsilon_i$ is independent normal noise.

Considering the superposition of Normal distribution

$$\text{Then } \mathbf{x}_t = (a_t \cdots a_1)\mathbf{x}_0 + \sqrt{(a_t \cdots a_2) b_1^2 + (a_t \cdots a_3) b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2} \cdot \overline{\varepsilon_t} \tag{3}$$

Noticing that we can sum up the square of the coefficients

Then $(a_t \cdots a_1)^2 + (a_t \cdots a_2)^2 b_1^2 + (a_t \cdots a_2)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2$

$= (a_1 \cdots a_2)^2 a_1^2 + (a_t \cdots a_2)^2 b_1^2 + (a_t \cdots a_2)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2$

$= (a_t \cdots a_2)^2 (a_1^2 + b_1^2) + (a_1 \cdots a_2)^2 b_2^2 + \cdots + a_t^2 b_{t-1}^2 + b_t^2$

$= (a_t \cdots a_3)^2 (a_t^2 (a_1^2 a_1^2) + b_2^2) + \cdots + a_t^2 b_{t-1}^2 + b_t^2$

$= \cdots$

$$= a_t^2 (a_{t-1}^2 (\cdots (a_2^2 (a_1^2 + b_1^2) + b_2^2) + \cdots b_{t-1}^2)_) + b_t^2 \tag{4}$$

Introduce a costraint $a^2 + b^2 = 1$, $s.t.\,(4) = 1$

let $\overline{a}_t = (a_t \cdots a_1)^2$,

Then (3) $\implies$ $\mathbf{x}_t = \sqrt{\overline{a_t}}\mathbf{x}_0 + \sqrt{1 - \overline{a_t}}\,\overline{\varepsilon_t}$, $\overline{\varepsilon_t} \sim \mathcal{N}(0, \mathbf{I})$ (5)

(1), (5) $\implies$ $\mathbf{x}_t = \sqrt{a_t}\mathbf{x}_{t-1} + \sqrt{1 - a_t}\,\varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$ (6)

wrt. $\mathcal{N}(0, \mathbf{I})$ could be seen as a standard normal distribution

$\implies \lim_{t \to \infty} a_t = 0$

So the original writer, Ho use $1 - a_t$ in the second term of (6)[1]

to make sure that variance is on the same scale for variance-preserving.

Noticing that the process can be regarded as sampling from a Gaussian distribution,

wrt. the Reparameterization Trick[2]

Now we can write out the complete foward noising process from (6):[3]

$$\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I}) \tag{7}$$

wrt. a_t is not a param that can be learnt by a DNN, and $\overline{\alpha_t} = 1 - \beta_t$

$$(5) \implies \mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I}),\ \overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s \tag{8}$$

Then the forward process is a posterior estimation based on joint probability density and

the Markov chain properties:

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

Q.E.D.

- *Notes for part I:*
- *1:* The original writer, Ho use $1 - a_t$ in the second term of (6) to make sure that variance is on the same scale for variance-preserving.
- *2:* The reparameterization trick is a technique used to enable gradient-based optimization of models with stochastic components. By expressing a random variable as a deterministic function of another random variable with a fixed distribution, it allows the gradients to be backpropagated through the stochastic

nodes. This is particularly useful in variational autoencoders (VAEs) and other models where sampling from a distribution is required during training.

- *3:* in $\mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$ the first term before ';' is the random variance, the second is the mean and the third term is the variance.

II Reverse Process(Training)

The reverse process of the denoise diffusion probability model is a process of removing noises from the noised image, where the noised image and the denoised image are linearly dependant. And we need to find a loss function for the training of the reverse process in this part (Algorithm 1).

Pf:

In this section, we are trying to find why the loss function used to traing the model is:

$$L_{simple}(\theta) := \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{\beta^2}{2\sigma^2_{q(t)}\alpha_t(1-\overline{\alpha}_t)}\left\|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon)\right\|^2\right] \tag{10}$$

At the very beginning, I would like to mention the VAE, which is the foundation of DDPM.

Let's define the encoder and decoder are: $x \to z$, $z \to x$, respectively.

Then the encoder can be expressed by: $q(z|x)$and decoder: $p(x|z)$

where $q(z)$ is the prior distribution

If you have read the original paper of DDPM, you will find the two models are very similar

So we need to find out how DDPM made these Markov process more complex and powerful.

The DDPM can be regarded as a HAE(Hierarchical Variational Autoencoder). And a HAE

devides the en- and decoder into T steps s.t.

encode proces: $\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{x}_2 \to \cdots \to \mathbf{x}_T$ and decode proces: $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_1 \to \mathbf{x}_0$

We can wirte these processes in a format that is similar to VAE:

encode proces: $q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and decode proces: $p(\mathbf{x}_{0:T}|\mathbf{x}_T) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})$

As we see, this change actually are trying to approximating a curve with a series of linear

functions. So the model is more powerful.

Considering that both HAE and VAE are likelihood-based models, with the target:

$\arg\min_{\mathbf{x}_0} p_\theta(\mathbf{x}_0)$,where θ is a neural network.

Propaedeutics:

Every steps of reverse process: $p_\theta(\mathbf{x}_{t-1})$ is similar to the forward one;

The linear relationship between steps can be modeled as:

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

Then the whole process can be written as a joint probability:

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

Based on the complete probability formular, we can write out the likelihood function of $p_\theta(\mathbf{x}_{0:T})$

$$\begin{aligned}
\log p_\theta(\mathbf{x}_0) &= \log \int p_\theta(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)\, d\mathbf{x}_0 d\mathbf{x}_1 d \ldots d\mathbf{x}_T \\
&= \log \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T} \\
&= \log \int \frac{p_\theta(\mathbf{x}_{0:T}) q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\, d\mathbf{x}_{1:T} \\
&= \log \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\frac{p_\theta(\mathbf{x}_{0:T})}{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\
&\geq \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1) \prod_{t=2}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1}) \prod_{t=1}^{T-1} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{T-1}, \mathbf{x}_T|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] - \mathbb{E}_{q_\theta(\mathbf{x}_{T-1}|\mathbf{x}_0)} \mathbb{E}_{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \left[\log \frac{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})}{p(\mathbf{x}_T)} \right] \\
&\quad - \sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0)} \mathbb{E}_{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \mathbf{x}_0)} \left[\log \frac{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] - \mathbb{E}_{q_\theta(\mathbf{x}_{T-1}|\mathbf{x}_0)} \mathbb{E}_{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \left[\log \frac{q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1})}{p(\mathbf{x}_T)} \right] \\
&\quad - \sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0)} \mathbb{E}_{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \left[\log \frac{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] - \mathbb{E}_{q_\theta(\mathbf{x}_{T-1}|\mathbf{x}_0)} \left[D_{KL}(q_\theta(\mathbf{x}_T|\mathbf{x}_{T-1}) \,||\, p(\mathbf{x}_T)) \right] \\
&\quad - \sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0)} \left[D_{KL}(q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) \,||\, p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \right]
\end{aligned} \tag{11}$$

where the first term is the reconstruction term, the second term is the prior matching term, and the third is the consistency term.[4]

Then we can solve $\arg\min_{\theta} p_\theta(\mathbf{x})$ by Monte-Carlo simulation

However, the last term we got is: $\sum_{t=1}^{T-1} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1},\mathbf{x}_{t+1}|\mathbf{x}_0)} \left[D_{KL}(q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) \,||\, p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \right]$

where both p and q will increase the variance.

Please call back that we will train a DNN to process from $\mathbf{x}_{t+1} \to \mathbf{x}_t$

And $\mathbf{x}_0$ is constant because it's the input image.

Following the principle of the Markov chain, we can re-write equation(10):

$$
\begin{aligned}
\log p_\theta(\mathbf{x}_0) &\geq \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1) \prod_{t=2}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})} \right]
\end{aligned}
$$

There we use a trick based on Markov chain: $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \dfrac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}$ (12)

$$
\begin{aligned}
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1) \prod_{t=2}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\frac{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q_\theta(\mathbf{x}_t|\mathbf{x}_0)}{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_0)}} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_t|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_2|\mathbf{x}_0)} \frac{p_\theta(\mathbf{x}_2|\mathbf{x}_0)}{p_\theta(\mathbf{x}_3|\mathbf{x}_0)} \cdots \frac{p_\theta(\mathbf{x}_{T-1}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T|\mathbf{x}_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] + \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] + \mathbb{E}_{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q_\theta(\mathbf{x}_t,\mathbf{x}_{t-1}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) - \mathbb{E}_{q_\theta(\mathbf{x}_T|\mathbf{x}_0)} \log \frac{q_\theta(\mathbf{x}_T|\mathbf{x}_0)}{p(\mathbf{x}_T)} \right] - \sum_{t=2}^{T} \mathbb{E}_{q_\theta(\mathbf{x}_t|\mathbf{x}_0)} \mathbb{E}_{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} \left[\log \frac{q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \right]
\end{aligned}
$$

$$= \mathbb{E}_{q_\theta(\mathbf{x}_1|\mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] - D_{KL}(q_\theta(\mathbf{x}_T|\mathbf{x}_0)\,||\,p(\mathbf{x}_T)) - \sum_{t=2}^{T}\mathbb{E}_{q_\theta(\mathbf{x}_t|\mathbf{x}_0)}\left[D_{KL}(q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\,||\,p(\mathbf{x}_{t-1}|\mathbf{x}_t))\right] \tag{13}$$

Then the denoising matching term(the third one) becomes 'Contrast Denoising', minimizing the loss of real and estimated process from adding noise to removing noise.

Then we only need to model $q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$ and $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to solve their D_{KL}.

Note that our goal is to construct a simple pdf, for example, a Gaussian pdf.

$$\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-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)} \\
&= \frac{\mathcal{N}(\mathbf{x}_t;\sqrt{\alpha_t}\mathbf{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\mathbf{x}_{t-1};\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0,(1-\overline{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\mathbf{x}_t;\sqrt{\overline{\alpha}_t}\mathbf{x}_0,(1-\overline{\alpha}_t)\mathbf{I})} \\
&\propto exp\left\{-\left[\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{2(1-\alpha_t)}+\frac{(\mathbf{x}_{t-1}-\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0)^2}{2(1-\overline{\alpha}_{t-1})}-\frac{(\mathbf{x}_t-\sqrt{\overline{\alpha}_t}\mathbf{x}_0)^2}{2(1-\overline{\alpha}_t)}\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[\frac{-2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}+\alpha_t\mathbf{x}_t^2}{(1-\alpha_t)}+\frac{-2\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_{t-1}\mathbf{x}_0+\mathbf{x}_{t-1}^2}{(1-\overline{\alpha}_{t-1})}+C(\mathbf{x}_t,\mathbf{x}_0)\right]\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(C is a constant function)} \\
&= exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}}{(1-\alpha_t)}+\frac{\alpha_t\mathbf{x}_t^2}{(1-\alpha_t)}-\frac{2\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_{t-1}\mathbf{x}_0}{(1-\overline{\alpha}_{t-1})}+\frac{\mathbf{x}_{t-1}^2}{(1-\overline{\alpha}_{t-1})}+C(\mathbf{x}_t,\mathbf{x}_0)\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}})\mathbf{x}_{t-1}^2-2(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}})\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[\frac{\alpha_t(1-\overline{\alpha}_{t-1})+1-\alpha_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\mathbf{x}_{t-1}^2-2(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}})\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[\frac{\alpha_t-\overline{\alpha}_t+1-\alpha_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\mathbf{x}_{t-1}^2-2(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}})\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[\frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\mathbf{x}_{t-1}^2-2(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}})\mathbf{x}_{t-1}\right]-\frac{1}{2}C(\mathbf{x}_t,\mathbf{x}_0)\right\} \\
&= exp\left\{-\frac{1}{2}\frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\left[\mathbf{x}_{t-1}^2-2\frac{\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}}}{\frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}}\mathbf{x}_{t-1}\right]-\frac{1}{2}C(\mathbf{x}_t,\mathbf{x}_0)\right\} \\
&= exp\left\{-\frac{1}{2}\frac{1-\overline{\alpha}_t}{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}\left[\mathbf{x}_{t-1}^2-2\frac{(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1-\alpha_t}+\frac{\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}_0}{1-\overline{\alpha}_{t-1}})(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_{t-1}\right]-\frac{1}{2}C(\mathbf{x}_t,\mathbf{x}_0)\right\} \\
&= exp\left\{-\frac{1}{2}\frac{1}{\frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}}\left[\mathbf{x}_{t-1}^2-2\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})\mathbf{x}_t+\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\overline{\alpha}_t}\mathbf{x}_{t-1}\right]-\frac{1}{2}C(\mathbf{x}_t,\mathbf{x}_0)\right\} \\
&= exp\left\{-\frac{(\mathbf{x}_{t-1}-\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})\mathbf{x}_t+\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\overline{\alpha}_t})^2}{2\frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}}\right\} \\
&\propto \mathcal{N}\left(\mathbf{x}_{t-1};\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})\mathbf{x}_t+\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\overline{\alpha}_t},\frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{I}\right)
\end{aligned} \tag{14}$$

Congrats! Now we have the posterior distribution of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$.

Then for all images $\mathbf{x}_{t-1}$, have:

$$\begin{aligned}\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}(\mu_q, \sigma_q^2) \\ &= \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\overline{\alpha}_t}, \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{I}\right) \\ &= \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})\mathbf{x}_t}{1-\overline{\alpha}_t} + \frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\overline{\alpha}_t}, \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\mathbf{I}\right) \end{aligned} \tag{15}$$

Note that $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta, \Sigma_\theta)$,where θ is the DNN.(Both μ and Σ are vectors(matrices))

Hence, $\mathcal{D}_{KL}(\mathcal{N}(\mathbf{x}; \mu_x, \Sigma_x) \,||\, \mathcal{N}(\mathbf{y}; \mu_y, \Sigma_y))$

$$= \frac{1}{2}\left[\log\frac{|\Sigma_y|}{|\Sigma_x|} + tr(\Sigma_y^{-1}\Sigma_x) + (\mu_y - \mu_x)^T\Sigma_y^{-1}(\mu_y - \mu_x)\right] \tag{16}$$

[5]Then $\arg\min_\theta \mathcal{D}_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||q_\theta(\mathbf{x}_{t-1}|\mathbf{x}_{t-1}))$

$$\begin{aligned} &= \arg\min_\theta \mathcal{D}_{KL}(\mathcal{N}(\mathbf{x}_{t-1}; \mu_q, \Sigma_{q(t)})||\mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta, \Sigma_{q(t)})) \\ &= \arg\min_\theta \frac{1}{2}\left[\log\frac{|\Sigma_{q(t)}|}{|\Sigma_{q(t)}|} + tr(\Sigma_{q(t)}^{-1}\Sigma_{q(t)}) + (\mu_\theta - \mu_{q(t)})^T\Sigma_{q(t)}^{-1}(\mu_\theta - \mu_q)\right] \\ &= \arg\min_\theta \frac{1}{2}\left[\log 1 - d + d + (\mu_\theta - \mu_{q(t)})^T\Sigma_{q(t)}^{-1}(\mu_\theta - \mu_q)\right] \\ &= \arg\min_\theta \frac{1}{2}\left[(\mu_\theta - \mu_{q(t)})^T\Sigma_{q(t)}^{-1}(\mu_\theta - \mu_q)\right] \\ &= \arg\min_\theta \frac{1}{2}\left[(\mu_\theta - \mu_{q(t)})^T(\sigma_{q(t)}^2\mathbf{I})(\mu_\theta - \mu_q)\right] \\ &= \arg\min_\theta \frac{1}{2\sigma_{q(t)}^2}\left[\left\|\mu_\theta - \mu_{q(t)}\right\|_2^2\right] \end{aligned} \tag{17}$$

Then our task become computing the μ_θ and Σ_θ, in (15) & (17)

Note that $\mathbf{x}_0$ & $\mathbf{x}_t$ are known because they are input image and output inmage, repectively.

And there exists a function mapping $\mathbf{x}_0$ to $\mathbf{x}_t$ (Yeah, the function is actually the DDPM.)

So one considerable method is compute the true μ_q by $(\mathbf{x}_t, t)$:

$$\mu_q = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \tag{18}$$

However, using equation (16), we have a more simple way to compute the μ_q.

Let μ_θ & μ_q in the same format. Then calculate $\left\|\mu_\theta - \mu_{q(t)}\right\|_2^2$

Note that our goal is to reconstruct image $\mathbf{x}_0$, where $\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t)$ based on μ_θ

Then we have: $\mu_\theta = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}f_\theta(\mathbf{x}_t, t) + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t$ (19)

Now continue our deduction in equation (17):

$$\begin{aligned} &\arg\min_\theta \mathcal{D}_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_{t-1})) \\ &= \arg\min_\theta \frac{1}{2\sigma_{q(t)}^2}\left[\left\|\mu_\theta - \mu_{q(t)}\right\|_2^2\right] \end{aligned}$$

$$= \arg\min_{\theta} \frac{1}{2\sigma_{q(t)}^2} \left[\left\| \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} f_\theta(\mathbf{x}_t, t) - \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathbf{x}_0 \right\|_2^2 \right]$$

Note that β_t here follows the difinition $\beta_t = 1 - \alpha_t$ in the paper rather than equation (7)

$$= \arg\min_{\theta} \frac{1}{2\sigma_{q(t)}^2} \left[\left\| \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} (f_\theta(\mathbf{x}_t, t) - \mathbf{x}_0) \right\|_2^2 \right]$$

$$= \arg\min_{\theta} \frac{1}{2\sigma_{q(t)}^2} \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} \left[\|f_\theta(\mathbf{x}_t, t) - \mathbf{x}_0)\|_2^2 \right] \tag{20}$$

Consider the meaning of our denoising process. we actually change the target function to minimize the loss between the real image and the estimated image.

Now we can try to handle equation (13), the Log-Likelihood Estimation : $\log p_\theta(\mathbf{x}_0)$

Firt, with equation(5), we can simplify (18) :

$$\begin{aligned}
\mu_q &= \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \\
&= \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} \frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\overline{\alpha}_t}} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}_t \\
&\quad **\epsilon_t \text{ is the noise we added by equation (5). Ignore that overline here, plz.} \\
&= \left[\frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} \right] \mathbf{x}_t - \frac{\sqrt{\alpha_{t-1}}\beta_t\sqrt{1-\overline{\alpha}_t}}{(1-\overline{\alpha}_t)\sqrt{\overline{\alpha}_t}}\epsilon_t \\
&= \left[\frac{\beta_t}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} \right] \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\overline{\alpha}_t}}\epsilon_t \\
&= \frac{1-\alpha_t+\alpha_t(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\overline{\alpha}_t}}\epsilon_t \\
&= \frac{1-\alpha_t+\alpha_t-\overline{\alpha}_t}{(1-\overline{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\overline{\alpha}_t}}\epsilon_t \\
&= \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\overline{\alpha}_t}}\epsilon_t
\end{aligned} \tag{21}$$

Congrats! Now we can estimate $\mathbf{x}_0$ from $\mathbf{x}_t$ and noise series ϵ_t

Note that $\hat{\epsilon}_t = f_\theta(\mathbf{x}_t, t)$

Then we can reconstruct equation (20) into:

$$\begin{aligned}
&\arg\min_{\theta} \mathcal{D}_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_{t-1})) \\
&= \arg\min_{\theta} \frac{1}{2\sigma_{q(t)}^2} \frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)\alpha_t} \left[\|f_\theta(\mathbf{x}_t, t) - \epsilon_t\|_2^2 \right]
\end{aligned} \tag{22}$$

Then we get the same equation with the paper:

$$\mathbb{E}_{\mathbf{x}_0,\epsilon} \left[\frac{\beta^2}{2\sigma_{q(t)}^2\alpha_t(1-\overline{\alpha}_t)} \left\| \epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon) \right\|^2 \right] \tag{23}$$

You may notice he difference between (20) and (22):

In (20), we directly let the DNN to optimize the loss between $\hat{\mathbf{x}_0}$ & $\mathbf{x}_0$.

Then based on the extra equations we made above, in (22) we let the DNN to optimize the loss

between the best noise series ϵ_t & the Gaussian noises $\hat{\epsilon_t}$.

Although in a mathematical view, they are both practical, the writers of DDPM find that the the denoising model is more effective.

THAT'S WHY THE MODEL IS CALLED THE DDPM!

Q.E.D.

- *Notes for part II:*
- *4:* $\mathcal{D}_{KL}$ is the Kullback-Leibler divergence, p and q are the forward and backward process. So as you can spot that these two D_{KL} are used to measure the loss between the two processes.
- *5:* In equation (17), we typically assume that the variances of the forward process (transition from data to noise) and the reverse process (transition from noise to data recovery) are identical. This assumption is based on the symmetry of the transformations at each time step within the DDPM framework. The assumption of symmetry simplifies the calculation of the Kullback-Leibler (KL) divergence, as it circumvents the need to deal with the complex differences between the covariance matrices of the two distributions. Although the variances may differ in the actual forward and reverse processes, this assumption is reasonable within the DDPM framework because it not only simplifies the training and computation of the model but also continues to capture the key characteristics of the data distribution. In this way, DDPM can effectively learn to recover high-quality data samples from noise while maintaining the coherence and consistency of the generation process.

References

[1] C. Luo, "Understanding the denoising diffusion probabilistic model," *arXiv preprint arXiv:2204.00283*, 2022.

[2] J. Ho, C. Meng, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, https://doi.org/10.48550/arXiv.2006.11239.

[3] Hung-Yi Lee. *Diffusion Model 原理剖析*. https://www.youtube.com/watch?v=ifCDXFdeaaM&list=PLJV_el3uVTsNi7PgekEUFsyVllAJXRsP-&index=4.

[4] 撒旦-cc. *一文解释 Diffusion Model (一) DDPM 理论推导*. https://zhuanlan.zhihu.com/p/565901160.