The Theory behinds Denoising Diffusion Probabilistic Models

Samy Vilhes

December 2024

1 Introduction

Denoising Diffusion Probabilistic Models (DDPM), introduced in the paper DDPM [2], are powerful generative models designed to rival state-of-the-art methods such as Variational Autoencoders (VAE) [3], Generative Adversarial Networks (GAN) [1], and regressive models like Normalizing Flows. These models achieve high-quality sample generation by iteratively denoising data from a Gaussian noise process, leveraging a diffusion-based framework that provides a more stable training process and better mode coverage compared to GANs while maintaining competitive generation quality. DDPMs are composed of a **forward process** and a **reverse process**.

2 The Forward Process

Let us consider x_0 as a sample, such as an image. The forward process involves progressively adding noise to x_0 over multiple steps, effectively transforming it into a noisy version through a series of stochastic operations:

$$x_0 \to x_1 \to x_2 \to \cdots \to x_T$$

where T denotes the total number of steps in the process. We choose T to be sufficiently large such that x_T is pure noise. This enables the transformation of a complex data distribution into a simple distribution.

Mathematically, this process can be expressed as:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t$$

where β_t represents a variance scheduler, and $\epsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise. When discussing distributions, if we denote $q(x_0)$ as the distribution of our data, we have:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I).$$

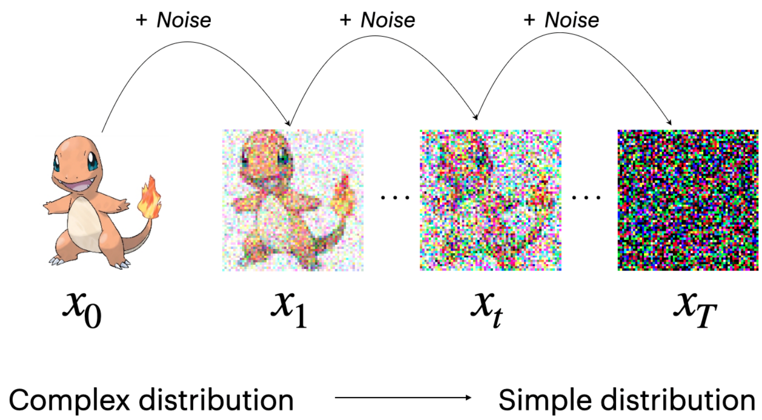

Figure 1: Forward Process

and:

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

While we choose T to be large, performing T sequential transformations is computationally inefficient. Fortunately, there exists a formula that allows us to directly transition from x_0 to x_t in a single step.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{1}$$

where:

- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$,
- $\alpha_t = 1 - \beta_t$,
- $\epsilon \sim \mathcal{N}(0, I)$.

(We let the proof in Section 8.1)

Thus:

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

With this equation, we conclude the theoretical explanation of the Forward Process. However, this process only transforms an observation into pure noise. The ultimate goal is to achieve the reverse: starting from pure noise, generate a realistic sample by applying T transformations in the opposite direction.

3 The Reverse Process

The joint distribution $p_\theta(x_{0:T})$ is called the **Reverse Process**. It is defined as a Markov Chain with learned Gaussian transition starting at $p(x_T) = \mathcal{N}(x_T; 0, I)$:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$$

with:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \text{ being the reverse distribution.}$$

We know that the reverse process follows a Gaussian distribution because β_t is chosen to be small enough, ensuring that the added noise in the forward process is minimal. Consequently, the reverse process also involves adding noise, albeit a different type, to reconstruct the original sample.
The core objective of diffusion models is to learn the parameters of the reverse distribution, $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$. Once these parameters are learned, we can iteratively transform a noisy image into a progressively less noisy one, ultimately reconstructing a realistic sample:

$$x_T \to x_{T-1} \to x_{T-2} \to \cdots \to x_0$$

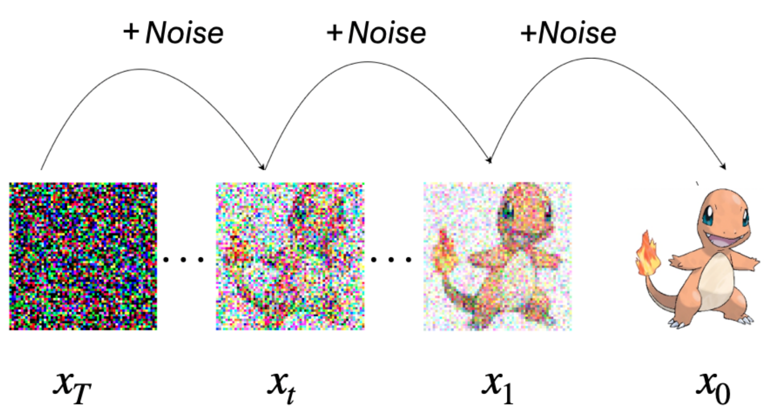

Figure 2: Reverse Process

To learn the reverse process, we aim to maximize the log-likelihood $\mathbb{E}_{q(x_0)}[\log p_\theta(x)]$. However, since this quantity does not have a closed form, we instead maximize its variational lower bound (VLB):

$$\mathbb{E}_{q(x_0)}\big[\log p_\theta(x)\big] \geq -\mathbb{E}_{q(x_{0:T})}\left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})}\right] = \text{VLB} \tag{2}$$

(We let the proof in Section 8.2)

The term on the right, VLB, represents the variational lower bound. In practice, we minimize its negative, i.e., $-\text{VLB} = L_{\text{VLB}}$.
Thus:

$$L_{\text{VLB}} = \mathbb{E}_{q(x_{0:T})}\left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})}\right]$$

$$= \mathbb{E}_{q(x_{0:T})}\left[D_{\text{KL}}\big(q(x_T \mid x_0) \,||\, p_\theta(x_T)\big) + \sum_{t=2}^{T} D_{\text{KL}}\big(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\big) - \log p_\theta(x_0 \mid x_1)\right] \tag{3}$$

(We let the proof in Section 8.3)

Furthemore:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\big(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\big) \tag{4}$$

With:

$$\tilde{\beta}_t = \beta_t \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}$$

and:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t$$

(We let the proof in Section 8.4)

In this way, both red and blue terms represent Kullback-Leibler divergences between Gaussian distributions. As a result, we can derive analytical expressions for these two terms.
We denote by:

- $L_T =$ red
- $L_{t-1} =$ blue
- $L_0 =$ green

Thus:

$$L_{\text{VLB}} = \mathbb{E}_{q(x_{0:T})}\Big[L_T + L_{t-1} + L_0\Big]$$

During the training, the term L_T can be ignored because it contains no learnable parameters ($p_\theta(x_T)$ is pure noise)

Taking a closer look at L_{t-1}, we leverage the form of $q(x_{t-1} \mid x_t, x_0)$ to make an assumption about the form of $p_\theta(x_{t-1} \mid x_t)$. Specifically, we assume that they follow a similar distribution. However, the mean parameter of q is the only term we cannot compute directly, as it requires knowledge of the original input

image x_0.
Thus, we suppose:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I\big)$$

Then, the primary objective is to learn $\mu_\theta(x_t, t)$, ensuring that it closely approximates $\tilde{\mu}_t(x_t, x_0)$ by minimizing the KL Divergence between these quantites:

$$\begin{aligned}
&D_{\mathrm{KL}}\big(q(x_{t-1} \mid x_t, x_0) \mid\mid p_\theta(x_{t-1} \mid x_t)\big) \\
&= D_{\mathrm{KL}}\Big(\mathcal{N}\big(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\big) \mid\mid \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I\big)\Big) \\
&= \frac{1}{2} \frac{1 - \bar{\alpha}_t}{\beta_t\big(1 - \bar{\alpha}_{t-1}\big)} ||\mu_\theta(x_t, t) - \tilde{\mu}_t(x_t, x_0)||_2^2 \qquad (5)
\end{aligned}$$

(We let the proof in Section 8.5)

Our goal is to estimate $\tilde{\mu}_t(x_t, x_0)$, and we know its form is given by:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t$$

During the denoising process the only thing we do not know in $\tilde{\mu}_t(x_t, x_0)$ is x_0. We will use the analytical form of $\tilde{\mu}_t(x_t, x_0)$ to suppose the form of $\mu_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_\theta + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t$$

Here, we estimate x_0 by x_θ the prediction of the input sample by the model. Then:

$$\begin{aligned}
&D_{\mathrm{KL}}\big(q(x_{t-1} \mid x_t, x_0) \mid\mid p_\theta(x_{t-1} \mid x_t)\big) \\
&= \frac{1}{2} \frac{1 - \bar{\alpha}_t}{\beta_t\big(1 - \bar{\alpha}_{t-1}\big)} ||\mu_\theta(x_t, t) - \tilde{\mu}_t(x_t, x_0)||_2^2 \\
&= \frac{1}{2} \frac{\beta_t \cdot \bar{\alpha}_{t-1}}{\big(1 - \bar{\alpha}_{t-1}\big)\big(1 - \bar{\alpha}_t\big)} ||x_\theta - x_0||_2^2 \qquad (6)
\end{aligned}$$

(We let the proof in Section 8.5)
Furthermore, using 1:

$$x_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}$$

But during the denoising process, we do not know the noise ϵ used to noise the model. Then we consider:

$$x_\theta = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}$$

Where ϵ_θ is the estimation by the model of the true initial noise ϵ. Thus:

$$\begin{aligned}
&D_{\mathrm{KL}}\big(q(x_{t-1} \mid x_t, x_0) \mid\mid p_\theta(x_{t-1} \mid x_t)\big) \\
&= \frac{1}{2} \frac{\beta_t \cdot \bar{\alpha}_{t-1}}{\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)} ||x_\theta - x_0||_2^2 \\
&= \frac{1}{2} \frac{\beta_t^2}{\tilde{\beta}_t\big(1-\bar{\alpha}_t\big)\alpha_t} ||\epsilon_\theta - \epsilon_0||_2^2 \qquad (7)
\end{aligned}$$

(We let the proof in Section 8.5)

During training, we focus solely on minimizing the simple term:

$$L = ||\epsilon_\theta - \epsilon_0||_2^2$$

Thus, the model, given x_t and t tries to estimate the input noise ϵ sampled.

4 The training

We will detail the training for a single sample:

Algorithm 1: Training Procedure for Diffusion Models

Input: Training dataset, number of timesteps T, model ϵ_θ

while *not converged* **do**

  1. Sample x_0 from the training set.
  2. Sample a timestamp $t \sim \text{Uniform}(\{1, \dots, T\})$.
  3. Sample noise $\epsilon \sim \mathcal{N}(0, \mathbb{I})$ with the same shape as x_0.
  4. Construct x_t using $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$.
  5. Feed x_t and t into the model $\epsilon_\theta(x_t, t)$ to predict ϵ.
  6. Compute the loss: $\mathcal{L} = ||\epsilon_\theta(x_t, t) - \epsilon||_2^2$.
  7. Perform backpropagation to update θ.

end

Output: Trained model parameters θ.

5 Details

5.1 Architecture

We will use for model the U-Net architecture [**unet**]. It will takes as input x_t and t and tries to predict ϵ the input noise.

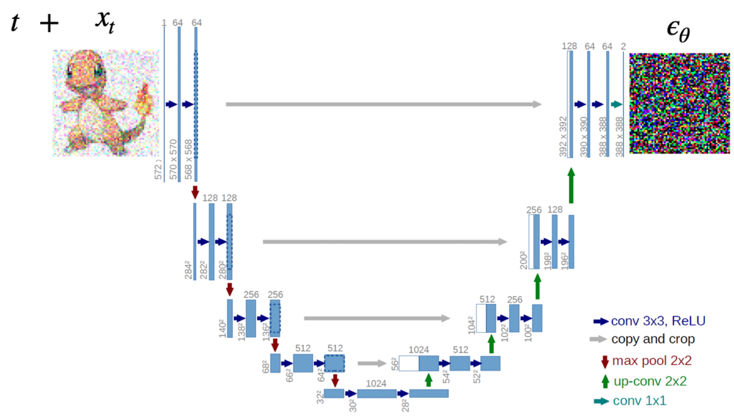

Figure 3: U-Net architecture

5.2 Hyperparameters

- We assume $T = 1000$ noise steps.
- β_t increases linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.
- The time information is provided to the model using Sinusoidal Positional Embeddings, as introduced in [4].
- The authors empirically found that replacing $\tilde{\beta}_t$ with β_t alone yields satisfactory results.

6 The Generation process

To generate new digits, we are interested the quantity:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \beta_t\big)$$

Where:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_\theta + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t$$

We saw that:

$$x_\theta = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}$$

Thus:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta\big) \tag{8}$$

(We let the proof in Section 8.6)

Then we will use the reparameterization trick to compute x_{t-1}, the denoised version of x_t, using the following formula:

$$x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\beta_t} z \quad \text{with } z \sim \mathcal{N}\big(0, I\big)$$

Algorithm 2: Generation Procedure for Diffusion Models

Input: Number of steps T, noise schedule $\{\beta_t\}_{t=1}^T$, and model $\mu_\theta(x_t, t)$.

1. **Initialize:** Sample $x_T \sim \mathcal{N}(0, I)$.
2. **For** $t = T, T-1, \ldots, 1$:
   - Sample $z \sim \mathcal{N}(0, I)$ if $t > 1$, else set $z = 0$.
   - Compute $x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\beta_t} z$.
3. **Return:** x_0.

Output: Generated sample x_0.

---

7 Experiments

We implement a basic DDPM with a U-Net based architecture for the Fashion-MNIST Dataset.

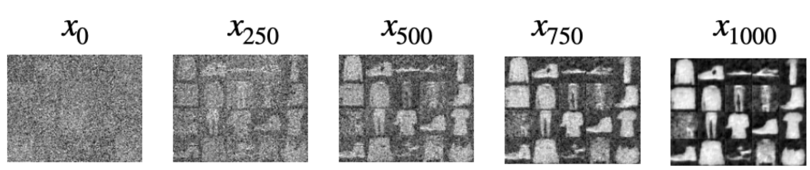

Figure 4: Fashion-MNIST Generation

The code can be found here:
`https://github.com/vilhess/codes/tree/main/ddpm`

8 Proofs

8.1 proof of 1

We will show by induction:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

where:

- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$,
- $\alpha_t = 1 - \beta_t$,
- $\epsilon \sim \mathcal{N}(0, I)$.

We have:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t$$

for $t = 1$:

$$\begin{aligned} x_1 &= \sqrt{1-\beta_1}x_0 + \sqrt{\beta_1}\epsilon_1 \\ &= \sqrt{\alpha_1}x_0 + \sqrt{1-\alpha_1}\epsilon_1 \\ &= \sqrt{\bar{\alpha}_1}x_0 + \sqrt{1-\bar{\alpha}_1}\epsilon_1 \end{aligned}$$

So it is true for $t = 1$
Let's suppose it is true for x_t. Then we need to show:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon$$

We know:

$$\begin{aligned} x_{t+1} &= \sqrt{1-\beta_{t+1}}x_t + \sqrt{\beta_{t+1}}\epsilon_{t+1} \\ &= \sqrt{1-\beta_{t+1}}\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right) + \sqrt{\beta_{t+1}}\epsilon_{t+1} \\ &= \sqrt{\alpha_{t+1}}\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right) + \sqrt{1-\alpha_{t+1}}\epsilon_{t+1} \\ &= \sqrt{\alpha_{t+1}}\sqrt{\bar{\alpha}_t}x_0 + \sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t}\epsilon + \sqrt{1-\alpha_{t+1}}\epsilon_{t+1} \\ &= \sqrt{\bar{\alpha}_{t+1}}x_0 + \color{blue}{\sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t}\epsilon} + \color{red}{\sqrt{1-\alpha_{t+1}}\epsilon_{t+1}} \end{aligned}$$

We have:

- $\color{blue}{\sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t}\epsilon} \sim \mathcal{N}(0, \alpha_{t+1}(1-\bar{\alpha}_t)I)$

- $\sqrt{1-\alpha_{t+1}}\epsilon_{t+1} \sim \mathcal{N}(0, (1-\alpha_{t+1})I)$

So summing blue + red:

$$
\begin{aligned}
&\sim \mathcal{N}(0, (\alpha_{t+1}(1-\bar{\alpha}_t) + 1 - \alpha_{t+1})I) \\
&\sim \mathcal{N}(0, (\alpha_{t+1} - \bar{\alpha}_{t+1} + 1 - \alpha_{t+1})I) \\
&\sim \mathcal{N}(0, (1-\bar{\alpha}_{t+1})I)
\end{aligned}
$$

Thus:

$$
\begin{aligned}
x_{t+1} &= \sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t}\epsilon + \sqrt{1-\alpha_{t+1}}\epsilon_{t+1} \\
&= \sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon
\end{aligned}
$$

So we finish the proof of 1 by induction.

8.2 proof of 2

$$
\begin{aligned}
\mathbb{E}_{q(x_0)}\big[\log p_\theta(x)\big] &= \mathbb{E}_{q(x_0)}\left[\log \int p_\theta(x_{0:T})\, dx_{1:T}\right] \\
&= \mathbb{E}_{q(x_0)}\left[\log \int q(x_{1:T} \mid x_0)\frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\, dx_{1:T}\right] \\
&= \mathbb{E}_{q(x_0)}\left[\log \mathbb{E}_{q(x_{1:T}|x_0)}\big[\frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\big]\right] \\
&\geq \mathbb{E}_{q(x_0)}\left[\mathbb{E}_{q(x_{1:T}|x_0)}\big[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\big]\right] \\
&\geq \mathbb{E}_{q(x_{0:T})}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\right] \\
&\geq -\mathbb{E}_{q(x_{0:T})}\left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})}\right]
\end{aligned}
$$

So we finish the proof of 2.

8.3 proof of 3

$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(x_{0:T})}\left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[\log \frac{\prod_{t=1}^T q(x_t \mid x_{t-1})}{p_\theta(x_T)\prod_{t=1}^T p_\theta(x_{t-1} \mid x_t)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \log \frac{\prod_{t=1}^T q(x_t \mid x_{t-1})}{\prod_{t=1}^T p_\theta(x_{t-1} \mid x_t)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \log \prod_{t=1}^T \frac{q(x_t \mid x_{t-1})}{p_\theta(x_{t-1} \mid x_t)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=1}^T \log \frac{q(x_t \mid x_{t-1})}{p_\theta(x_{t-1} \mid x_t)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^T \log \frac{q(x_t \mid x_{t-1})}{p_\theta(x_{t-1} \mid x_t)} + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right]
\end{aligned}
$$

Furthermore:

$$
\begin{aligned}
q(x_t \mid x_{t-1}) &= q(x_t \mid x_{t-1}, x_0) \quad \text{(because this is the noise processus)} \\
&= \frac{q(x_t, x_{t-1}, x_0)}{q(x_{t-1}, x_0)} \quad \text{(using Bayes Formula)} \\
&= \frac{q(x_t, x_{t-1}, x_0)}{q(x_t, x_0)} \frac{q(x_t, x_0)}{q(x_0)} \frac{q(x_0)}{q(x_{t-1}, x_0)} \\
&= q(x_{t-1} \mid x_t, x_0) \frac{q(x_t \mid x_0)}{q(x_{t-1} \mid x_0)} \quad \text{(using Bayes Formula)}
\end{aligned}
$$

Thus:

$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_t \mid x_{t-1})}{p_\theta(x_{t-1} \mid x_t)} + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_t, x_0)}{p_\theta(x_{t-1} \mid x_t)} \frac{q(x_t \mid x_0)}{q(x_{t-1} \mid x_0)} + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_t, x_0)}{p_\theta(x_{t-1} \mid x_t)} + \sum_{t=2}^{T} \log \frac{q(x_t \mid x_0)}{q(x_{t-1} \mid x_0)} + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_t, x_0)}{p_\theta(x_{t-1} \mid x_t)} + \log \frac{q(x_T \mid x_0)}{q(x_1 \mid x_0)} + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[\frac{q(x_T \mid x_0)}{\log p_\theta(x_T)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_t, x_0)}{p_\theta(x_{t-1} \mid x_t)} - \log q(x_1 \mid x_0) + \log \frac{q(x_1 \mid x_0)}{p_\theta(x_0 \mid x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[\frac{q(x_T \mid x_0)}{\log p_\theta(x_T)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_t, x_0)}{p_\theta(x_{t-1} \mid x_t)} - \log p_\theta(x_0 \mid x_1)\right]
\end{aligned}
$$

So we finish the proof of 3.

8.4 proof of 4

Firstly, we introduce:

$$
\begin{aligned}
\tilde{\beta}_t &= \left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)^{-1} \\
&= \left(\frac{\alpha_t - \alpha_t\bar{\alpha}_{t-1} + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right) = \left(\frac{1-\bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})}\right) \\
&= \beta_t \frac{(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}
\end{aligned}
$$

And:

$$\begin{aligned}
\tilde{\mu}(x_t, x_0) &= \left(\frac{\sqrt{\alpha_t}x_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}}\right)\tilde{\beta}_t \\
&= \left(\frac{\sqrt{\alpha_t}x_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}}\right)\beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \\
&= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \frac{\sqrt{\alpha_t}x_t}{1-\alpha_t} + \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}} \\
&= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\sqrt{\alpha_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_t}\beta_t \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \beta_t \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}x_0
\end{aligned}$$

Thus:

$$\begin{aligned}
q(x_{t-1} \mid x_t, x_0) &= \frac{q(x_{t-1}, x_t \mid x_0)}{q(x_t \mid x_0)} \\
&= \frac{q(x_t \mid x_{t-1}, x_0)q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} \\
&= {\color{red}q(x_t \mid x_{t-1})}\frac{{\color{blue}q(x_{t-1} \mid x_0)}}{{\color{green}q(x_t \mid x_0)}} \quad (\text{{\color{red}red} because noise process})
\end{aligned}$$

We know:

- ${\color{red}\text{red}} = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$.
- ${\color{blue}\text{blue}} = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})I)$
- ${\color{green}\text{green}} = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$

Then:

$$\begin{aligned}
q(x_{t-1} \mid x_t, x_0) &= {\color{red}q(x_t \mid x_{t-1})}\frac{{\color{blue}q(x_{t-1} \mid x_0)}}{{\color{green}q(x_t \mid x_0)}} \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{\left(x_t - \sqrt{\alpha_t}x_{t-1}\right)^2}{1-\alpha_t}\right) + \frac{\left(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\right)^2}{1-\bar{\alpha}_{t-1}} - \frac{\left(x_t - \sqrt{\bar{\alpha}_t}x_0\right)^2}{1-\bar{\alpha}_t}\right]
\end{aligned}$$

$$= \exp\left[-\frac{1}{2}\left(\frac{x_t^2 - 2x_t\sqrt{\alpha_t}x_{t-1} + \alpha_t x_{t-1}^2}{1-\alpha_t} + \frac{x_{t-1}^2 - 2x_{t-1}\sqrt{\bar{\alpha}_{t-1}}x_0 + \bar{\alpha}_{t-1}x_0^2}{1-\bar{\alpha}_{t-1}} - \frac{x_t^2 - 2x_t\sqrt{\bar{\alpha}_t}x_0 + \bar{\alpha}_t x_0^2}{1-\bar{\alpha}_t}\right)\right]$$

$$= \exp\left(-\frac{1}{2}\left[x_{t-1}^2\left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) - 2x_{t-1}\left(\frac{x_t\sqrt{\alpha_t}}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}}\right) + C(x_t, x_0)\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{x_{t-1}^2}{\tilde{\beta}_t} - 2x_{t-1}\frac{\tilde{\mu}_t(x_t, x_0)}{\tilde{\beta}_t}\right] + C(x_t, x_0)\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{x_{t-1}^2 - 2x_{t-1}\tilde{\mu}_t(x_t, x_0)}{\tilde{\beta}_t}\right] + C(x_t, x_0)\right)$$

This corresponds to the probability density function of a Gaussian distribution, given by:

$$\mathcal{N}\big(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\big),$$

Then we finish the proof of 4.

8.5 proofs of 5, 6, 7

$$D_{\text{KL}}\Big(\mathcal{N}\big(x_{t-1};\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I\big)\ ||\ \mathcal{N}\big(x_{t-1};\mu_\theta(x_t,t),\tilde{\beta}_t I\big)\Big)$$

$$= \mathbb{E}_{x_{t-1}\sim\mathcal{N}\big(\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I\big)}\Big[-\frac{N}{2}\log 2\pi - \frac{1}{2}\log\big(\det \tilde{\beta}_t I\big) - \frac{1}{2}\big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)$$

$$+ \frac{N}{2}\log 2\pi + \frac{1}{2}\log\big(\det \tilde{\beta}_t I\big) + \frac{1}{2}\big(x_{t-1}-\mu_\theta(x_t,t)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\mu_\theta(x_t,x_0)\big)\Big]$$

$$= \frac{1}{2}\mathbb{E}_{x_{t-1}\sim\mathcal{N}\big(\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I\big)}\Big[\big(x_{t-1}-\mu_\theta(x_t,t)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\mu_\theta(x_t,x_0)\big)$$

$$- \big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)\Big]$$

$$= \frac{1}{2}\mathbb{E}_{x_{t-1}\sim\mathcal{N}\big(\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I\big)}\Big[\big(x_{t-1}-\mu_\theta(x_t,t)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\mu_\theta(x_t,x_0)\big)\Big]$$

$$- \frac{1}{2}\mathbb{E}_{x_{t-1}\sim\mathcal{N}\big(\tilde{\mu}_t(x_t,x_0),\tilde{\beta}_t I\big)}\Big[\big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)^T \tilde{\beta}_t^{-1} I\big(x_{t-1}-\tilde{\mu}_t(x_t,x_0)\big)\Big]$$

$$= \frac{1}{2}\Big(\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big)^T \tilde{\beta}_t^{-1} I\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big) + tr\big(\tilde{\beta}_t^{-1}\tilde{\beta}_t I\big)\Big)$$

$$- \frac{1}{2}\Big(\big(\mu_\theta(x_t,t)-\mu_\theta(x_t,t)\big)^T \tilde{\beta}_t^{-1} I\big(\mu_\theta(x_t,t)-\mu_\theta(x_t,t)\big) + tr\big(\tilde{\beta}_t^{-1}\tilde{\beta}_t I\big)\Big) \quad \text{(using property 9)}$$

$$= \frac{1}{2}\Big(\frac{1}{\tilde{\beta}_t}\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big)^T\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big) + J\Big) - \frac{1}{2}J$$

$$= \frac{1}{2\tilde{\beta}_t}\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big)^T\big(\tilde{\mu}_t(x_t,x_0)-\mu_\theta(x_t,t)\big)$$

$$= \frac{1}{2}\frac{1-\bar{\alpha}_t}{\beta_t\big(1-\bar{\alpha}_{t-1}\big)}||\mu_\theta(x_t,t)-\tilde{\mu}_t(x_t,x_0)||_2^2 \quad \text{(proof of 5)}$$

$$= \frac{1}{2}\frac{1-\bar{\alpha}_t}{\beta_t\big(1-\bar{\alpha}_{t-1}\big)}||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_\theta + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t||_2^2$$

$$= \frac{1}{2}\frac{1-\bar{\alpha}_t}{\beta_t\big(1-\bar{\alpha}_{t-1}\big)}||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_\theta - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0||_2^2$$

$$= \frac{1}{2}\frac{1-\bar{\alpha}_t}{\beta_t\big(1-\bar{\alpha}_{t-1}\big)}||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\big(x_\theta - x_0\big)||_2^2$$

$$= \frac{1}{2}\frac{\bar{\alpha}_{t-1}\cdot\beta_t}{\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)}||x_\theta - x_0||_2^2 \quad \text{(proof of 6)}$$

Furthermore, using 1:

$$x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}$$

As we use x_θ as an estimation of x_0, we will consider:

$$x_\theta = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}$$

Where the only unknown term is ϵ_θ. Thus:

$$D_{\mathrm{KL}}\Big(\mathcal{N}\big(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\big) \;||\; \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t\big)\Big)$$

$$= \frac{1}{2}\frac{\bar{\alpha}_{t-1}\cdot\beta_t}{\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)}||x_\theta - x_0||_2^2$$

$$= \frac{1}{2}\frac{\bar{\alpha}_{t-1}\cdot\beta_t}{\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)}||\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}} - \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}||_2^2$$

$$= \frac{1}{2}\frac{\bar{\alpha}_{t-1}\cdot\beta_t}{\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)}||\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\big(\epsilon - \epsilon_\theta\big)||_2^2$$

$$= \frac{1}{2}\frac{\beta_t^2\cdot\bar{\alpha}_{t-1}\big(1-\bar{\alpha}_t\big)}{\beta_t\big(1-\bar{\alpha}_t\big)\big(1-\bar{\alpha}_{t-1}\big)\bar{\alpha}_t}||\epsilon - \epsilon_\theta||_2^2$$

$$= \frac{1}{2}\frac{\big(1-\bar{\alpha}_t\big)\cdot\beta_t^2}{\beta_t\big(1-\bar{\alpha}_{t-1}\big)\big(1-\bar{\alpha}_t\big)\alpha_t}||\epsilon - \epsilon_\theta||_2^2$$

$$= \frac{1}{2}\frac{\beta_t^2}{\tilde{\beta}_t\big(1-\bar{\alpha}_t\big)\alpha_t}||\epsilon - \epsilon_\theta||_2^2 \quad \text{(proof of 7)}$$

We recall this property:
Let: $X \sim \mathcal{N}(\mu, \Sigma)$ Then:

$$\mathbb{E}\Big[\big(X-u\big)^T A\big(X-u\big)\Big] = \big(\mu-u\big)^T A\big(\mu-u\big) + tr(A\Sigma) \tag{9}$$

8.6 proof of 8.6

We saw:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_\theta + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$$

We have:

$$x_\theta = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}$$

Thus:

$$\mu_\theta(x_t,t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_\theta + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$$

$$= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}\right) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$$

$$= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\left(1-\bar{\alpha}_t\right)\sqrt{\bar{\alpha}_t}}x_t + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}x_t - \frac{\sqrt{1-\bar{\alpha}_t}\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{\bar{\alpha}_t}\left(1-\bar{\alpha}_t\right)}\epsilon_\theta$$

$$= x_t\left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\left(1-\bar{\alpha}_t\right)\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\right) - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

$$= x_t\left(\frac{\beta_t}{\left(1-\bar{\alpha}_t\right)\sqrt{\alpha_t}} + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\right) - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

$$= x_t\left(\frac{\beta_t + \alpha_t\left(1-\bar{\alpha}_{t-1}\right)}{\left(1-\bar{\alpha}_t\right)\sqrt{\alpha_t}}\right) - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

$$= x_t\left(\frac{1-\alpha_t + \alpha_t - \bar{\alpha}_t)}{\left(1-\bar{\alpha}_t\right)\sqrt{\alpha_t}}\right) - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

$$= x_t\left(\frac{1}{\sqrt{\alpha_t}}\right) - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$$

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

So we finish the proof of 8.

References

[1] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *CoRR* abs/2006.11239 (2020). arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239.

[3] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: https://arxiv.org/abs/1312.6114.

[4] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: `1706.03762`. URL: `http://arxiv.org/abs/1706.03762`.