Introduction to Diffusion Models

Woncheol Shin

KAIST

2022.04.18

What are Diffusion Models

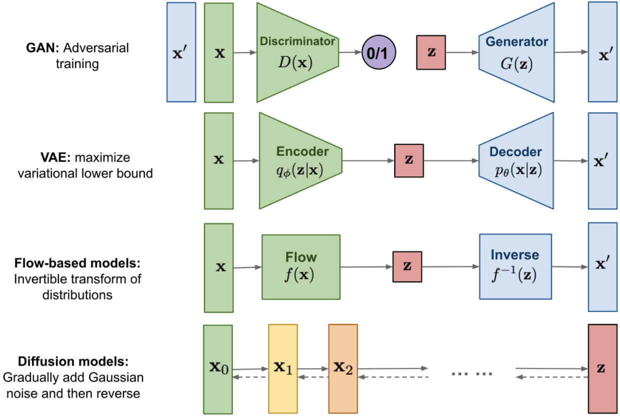

Figure: Overview of different types of generative models. (Source: [1])

Forward Diffusion Proces

We define a Markov chain of diffusion steps to slowly add small amount of Gaussian noise to a sample $\mathbf{x}_0$ in T steps, producing a sequence of noisy samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$.

Definition: Forward Diffusion Process

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

where $\mathbf{x}_0$ is a data point sampled from a real data distribution $q(\mathbf{x}_0)$ and $\{\beta_t \in (0,1)\}_{t=1}^T$ is a variance schedule.

Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1 < \beta_2 < \cdots < \beta_T$.
Ho et al. (2020) set the forward process variances to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.

Forward Diffusion Process

Property 1

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

where $\alpha_t=1-\beta_t$ and $\bar{\alpha}_t=\prod_{i=1}^t \alpha_i$.

Proof)

$$\begin{aligned}
\mathbf{x}_t & =\sqrt{\alpha_t} \mathbf{x}_{t-1}+\sqrt{1-\alpha_t} \mathbf{z}_{t-1} ; \text { where } \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
& =\sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2}+\sqrt{1-\alpha_t \alpha_{t-1}} \overline{\mathbf{z}}_{t-2} \\
& \quad \text { where } \overline{\mathbf{z}}_{t-2} \text { merges two Gaussians } \\
& =\cdots \\
& =\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_t ; \text { where } \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) & =\mathcal{N}\left(\mathbf{x}_t ; \sqrt{\bar{\alpha}_t} \mathbf{x}_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \quad \square
\end{aligned}$$

Eventually when $T \rightarrow \infty, \mathbf{x}_T$ is equivalent to an isotropic Gaussian distribution.

Idea: "If we can reverse the above process and sample from $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$, we will be able to recreate the true sample from a Gaussian noise input, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$."
However, we cannot easily estimate $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ because it needs to use the entire dataset.
Therefore, we need to learn a model p_θ to approximate these conditional probabilities!

Definition: Reverse Diffusion Process

Reverse Diffusion Process is defined as a Markov chain starting at $p_\theta(\mathbf{x}_T) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right)$$

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

* Note that if β_t is small enough, $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ is also Gaussian. Therefore, we define $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ as a Gaussian distribution.

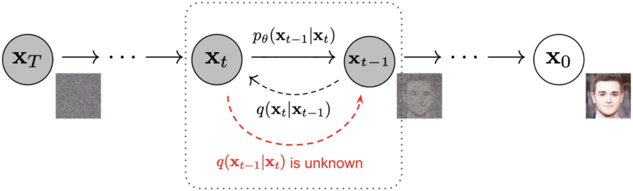

Figure: Forward and reverse diffusion process. (Source: [1] which is based on [2])

Reverse Diffusion Process

The reverse conditional probability $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is tractable when conditioned on x_0.

Property 2

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right)$$

where $\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right)$ and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$.

Proof) *Gaussian pdf: $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0\right) \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} = q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \frac{q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \because \text{Markov}$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1}\right)^2}{\beta_t} + \frac{\left(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\right)^2}{1-\bar{\alpha}_{t-1}} - \frac{\left(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0\right)^2}{1-\bar{\alpha}_t}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)\mathbf{x}_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right)\mathbf{x}_{t-1} + C\left(\mathbf{x}_t, \mathbf{x}_0\right)\right)\right)$$

where $C\left(\mathbf{x}_t, \mathbf{x}_0\right)$ is a function not involving $\mathbf{x}_{t-1}$.
(Continued on next slide)

$$\tilde{\beta}_t = 1/\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t$$

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) &= \left(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right)/\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)\\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0\\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right)\\ &\qquad\because \mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right) \text{ from Prop.1}\\ &= \frac{1}{\sqrt{\alpha_t}}\left(\mathrm{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right) \quad \square\end{aligned}$$

Learning Objective

Goal: We want to minimize the negative log-likelihood.

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right] \\
&\leq \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}\left[-\log p_\theta\left(\mathbf{x}_0\right) + D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)\right)\right] \\
&= \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}\left[-\log p_\theta\left(\mathbf{x}_0\right) + \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\left[\log \frac{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{p_\theta\left(\mathbf{x}_{0:T}\right) / p_\theta\left(\mathbf{x}_0\right)}\right]\right] \\
&= \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}\left[-\log p_\theta\left(\mathbf{x}_0\right) + \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\left[\log \frac{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{p_\theta\left(\mathbf{x}_{0:T}\right)} + \log p_\theta\left(\mathbf{x}_0\right)\right]\right] \\
&= \mathbb{E}_{\mathbf{x}_{0:T} \sim q(\mathbf{x}_{0:T})}\left[\log \frac{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{p_\theta\left(\mathbf{x}_{0:T}\right)}\right] := L_{\mathrm{VLB}}
\end{aligned}
$$

In other words, we can achieve the goal by minimizing L_{VLB}!

Learning Objective

We can convert L_{VLB} to be analytically computable.

Remark 1: L_{VLB}

$$
\begin{aligned}
L_{\mathrm{VLB}} = & \mathbb{E}_{q(\mathbf{x}_0)}[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_T\right)\right)}_{L_T}] \\
& + \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)}[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)}_{L_{t-1}}] \\
& + \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_1)}[\underbrace{-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0}]
\end{aligned}
$$

Proof)

$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log \frac{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}{p_\theta\left(\mathbf{x}_{0:T}\right)}\right] \\
&= \mathbb{E}_q\left[\log \frac{\prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}{p_\theta\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}\right] \\
&= \mathbb{E}_q\left[-\log p_\theta\left(\mathbf{x}_T\right) + \sum_{t=1}^{T} \log \frac{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}\right]
\end{aligned}
$$

(Continued on next slide)

Learning Objective

$$= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} + \log \frac{q(\mathbf{x}_1 \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)} \right]$$

$$= \mathbb{E}_q \left[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} \log \left(\frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \cdot \frac{q(\mathbf{x}_t \mid \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)} \right) + \log \frac{q(\mathbf{x}_1 \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)} \right]$$

$\because$ Markov property and Bayes' rule

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

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

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

$$= \mathbb{E}_{q(\mathbf{x}_0)} \underbrace{\left[D_{\mathrm{KL}}\left(q(\mathbf{x}_T \mid \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_T) \right) \right]}_{L_T} + \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)} \underbrace{\left[D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \right) \right]}_{L_{t-1}}$$

$$+ \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_1)} \underbrace{\left[-\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) \right]}_{L_0}$$

Definition: $L_T, L_{t-1},$ and L_0

$$(1)\ L_T = D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_T\right)\right)$$
$$(2)\ L_{t-1} = D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right) \text{ for } 2 \leq t \leq T$$
$$(3)\ L_0 = -\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)$$

1) L_T
- From Prop.1, $q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \rightarrow \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$ when $T \rightarrow \infty$.
- We assume that $p_\theta(\mathbf{x}_T) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$.
- L_T is constant and can be ignored during training.

2) L_{t-1}
- This term measures the difference between $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)$ and $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$.
- How do we optimize this term? (Next slide)

3) L_0
- This term reconstruct the original image from the slightly noised image.
- This term is optimized by MSE loss: $\left\|\mathbf{x}_0 - \boldsymbol{\mu}_\theta\left(\mathbf{x}_1, 1\right)\right\|^2$

$L_{t-1} = D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)$

$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_t\right), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \mathbf{I}\right)$

$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right)$

Let us set $\boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$ and $\boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right) = \sigma_t^2 \mathbf{I}$.

We have two options for σ_t^2: $\sigma_t^2 = \beta_t$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$.

According to Ho et al. (2020), both had similar results experimentally.

*$D_{KL}(p \| q) = \frac{1}{2}\left[\log \frac{|\Sigma_q|}{|\Sigma_p|} - k + \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right)^T \Sigma_q^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right) + \operatorname{tr}\left\{\Sigma_q^{-1} \Sigma_p\right\}\right]$

$$
\begin{aligned}
L_{t-1} & \propto \frac{1}{2 \sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right) - \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right)\right\|^2 \\
& = \frac{1}{2 \sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_t\right) - \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)\right\|^2 \\
& = \frac{\beta_t^2}{2 \sigma_t^2 \alpha_t\left(1-\bar{\alpha}_t\right)}\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right\|^2 \\
& = \frac{\beta_t^2}{2 \sigma_t^2 \alpha_t\left(1-\bar{\alpha}_t\right)}\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_t, t\right)\right\|^2
\end{aligned}
$$

Empirically, Ho et al. (2020) found that training the diffusion model works better with a simplified objective that ignores the weighting term:

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

Training and Sampling Algorithm

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathrm{Uniform}(\{1, \dots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  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 $t = T, \dots, 1$ **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 $\mathbf{x}_0$

Figure: Training and sampling algorithm. (Source: [2])

Generated Samples

Figure: Unconditional CIFAR10 progressive generation. (Source: [2])

References

[1] Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log.
https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.
[2] Jonathan Ho et al. "Denoising diffusion probabilistic models." arxiv Preprint
arxiv:2006.11239 (2020).