Denoising Diffusion Probabilistic Models (DDPM)

Mathematical Understanding

Kyoungman Kim

Sungkyunkwan University
Department of Mathematics

MIMIC Seminar

Table of Contents

Table of Contents

Original Paper

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at `https://github.com/hojonathanho/diffusion`.

- 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

What is Deep Learning?

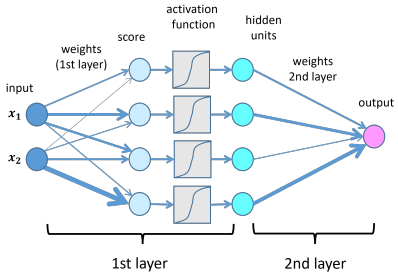

Figure: Deep Neural Network

- Deep learning designs a parameterized architecture (a neural network) and optimizes its parameters by minimizing a loss for a target task (e.g., image generation).
- Source: https://lamarr-institute.org/blog/deep-neural-networks/.

Kullback–Leibler Divergence

Definition(Kullback–Leibler divergence)

Let p and q be probability distributions of a random variable X.
The KL divergence of p from q is

$$D_{\mathrm{KL}}(p\|q) := \sum_x p(x)\ \log_2\left(\frac{p(x)}{q(x)}\right) \qquad \text{(discrete } X\text{)} \tag{1}$$

Continuous case and expectation form

$$D_{\mathrm{KL}}(p\|q) := \int p(x)\ \ln\left(\frac{p(x)}{q(x)}\right)\ dx \qquad \text{(continuous } X\text{)} \tag{2}$$

$$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_{X\sim p}\left[\log\left(\frac{p(X)}{q(X)}\right)\right] \tag{3}$$

Meaning of KL Divergence

Explanation

The quantity $D(p\|q)$ measures how well q approximates p: if we model X using q while the true distribution is p, then $D(p\|q)$ quantifies the degree of dissimilarity between p and q.

Remark

KL "divergence" is not a metric:

- Asymmetry: $D(p\|q) \neq D(q\|p)$
- No triangle inequality

Basic Idea of DDPM

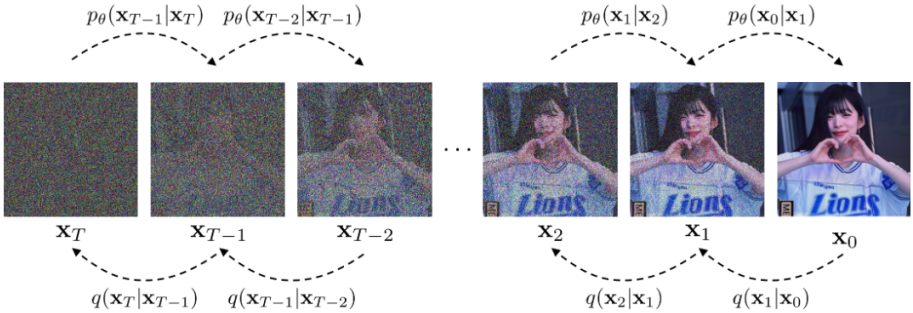

Figure: Noising (q) and denoising (p)

- By training a model to reverse a noising process, it learns how to generate high-quality data from pure noise.
- Source: https://freshrimpsushi.github.io/ko/posts/3638/.

Table of Contents

Forward Process

Forward Process (q)

A fixed process that gradually adds Gaussian noise to an image X_0 in T steps. We design this; the model does not learn it.

$$X_0 \xrightarrow{q} X_1 \xrightarrow{q} \dots \xrightarrow{q} X_T$$

- Starts from real data: $X_0 \sim p_{data}(X_0)$.
- Gaussian noise is used due to its convenient mathematical properties.

Forward Process

Forward Model(Fixed Markov Chain)

$$X_0 \sim p_{data}(X_0)$$

$$q(X_t|X_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}X_{t-1}, \beta_t I) \quad \text{for } t = 1, \ldots, T$$

Reparameterization Trick

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}Z_t, \quad \text{where } Z_t \sim \mathcal{N}(0, I)$$

Closed Form

$$q(X_t|X_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I) \approx \mathcal{N}(0, I)$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

Reverse Process

Reverse Process (p_θ)

A learnable process that denoises an image step-by-step. This is our generative model.

$$X_0 \xleftarrow{p_\theta} X_1 \xleftarrow{p_\theta} \dots \xleftarrow{p_\theta} X_T$$

- Generation starts from noise $X_T \sim \mathcal{N}(0, \mathbf{I})$.
- Goal: Learn $p_\theta(X_{t-1}|X_t)$ to approximate the true reverse $q(X_{t-1}|X_t)$.
- Since we don't know how to express $q(X_{t-1}|X_t)$ explicitly, instead we consider $q(X_{t-1}|X_t, X_0)$

Experimental Setup

DDPM Hyperparameters

- **Forward Process Setup: T=1000** step process is used, where the noise variance β_t increases **linearly** from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. This setup matches prior work for a fair comparison.

- **Resulting Prior Match:** This schedule ensures the final distribution $q(\mathbf{x}_T|\mathbf{x}_0)$ is almost indistinguishable from the target prior $\mathcal{N}(0, \mathbf{I})$.

$$L_T = D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel \mathcal{N}(0, \mathbf{I})) \approx 10^{-5} \text{ bits per dimension.}$$

Table of Contents

Understanding Long and Complicated Procedures

Remark

- Our goal is to trace the author's line of reasoning, trying to reconstruct their thinking process at each step of the calculation.
- Forcing every step to look "meaningful" is not ideal—some work is simply technical rather than intuitive.
- We do not need to memorize the entire process.
- In high-quality mathematical work, every assumption is applied carefully and appropriately.

Checkpoints

Key Assumptions & Properties

- **Gaussian parameterization:** We model the reverse step $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ as a Gaussian. In DDPM, the variance is typically fixed, so the network mainly learns the mean.
- **Markov structure and consecutive terms:** The forward process is a Markov chain, which enables simple factorizations and telescoping manipulations.
- **The role of $\mathbf{x}_0$:** The clean sample $\mathbf{x}_0$ serves as an anchor for deriving the posterior $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$.
- **Noise prediction viewpoint:** Learning the noise ϵ is the most important idea underlying the simplified objective.

- It is crucial to keep these points in mind throughout the derivation.

Approximating Gaussian : Result

Final Result

If $Y = \gamma X + \sigma Z$ with $\gamma \neq 0$, then in the limit of $\sigma \to 0$:

$$p_{X|Y}(x|y) \approx \mathcal{N}\left(\frac{1}{\gamma}(y + \sigma^2 \nabla \log p_Y(y)), \frac{\sigma^2}{\gamma^2} I\right)$$

Remark

- This suggests that, in the small-noise regime, the reverse conditional distribution can be approximated by a Gaussian.
- Hence, it is natural to model the reverse step in DDPM as Gaussian.

Approximating Gaussian : Tweedie's Formula(1)

Setup

Consider the random variable:

$$Y = \gamma X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

(We don't assume p_X is Gaussian.)

Conditional Expectation

If $Y = \gamma X + \sigma Z$ with $\gamma \neq 0$, then:

$$\begin{aligned} \mathbb{E}[X|Y] &= \frac{1}{\gamma}\mathbb{E}[\gamma X|Y] \\ &= \frac{1}{\gamma}(Y + \sigma^2 \nabla_Y \log p_Y(Y)) \end{aligned}$$

Approximating Gaussian : Tweedie's Formula(2)

Setup

Consider the random variable:

$$Y = \gamma X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

(We don't assume p_X is Gaussian.)

Conditional Variance

If $Y = \gamma X + \sigma Z$ with $\gamma \neq 0$, then:

$$\mathrm{Var}[X|Y] = \frac{\sigma^2}{\gamma^2}(I + \sigma^2 \nabla_Y^2 \log p_Y(Y))$$

Approximating Gaussian : Derivation(1)

Starting with Bayes' Rule

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \tag{1}$$

$$= \frac{p_{Y|X}(y|x)p_X(x)}{\int_{\mathbb{R}^d} p_{Y|X}(y|x')p_X(x')dx'} \tag{2}$$

$$= \frac{\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) p_X(x)}{\int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x'\|^2\right) p_X(x')dx'} \tag{3}$$

Approximating Gaussian : Derivation(2)

Final Gaussian Form

After algebraic manipulation and assuming small σ, the expression simplifies. $p_{X|Y}(x|y)$ takes the form of a Gaussian distribution:

$$\begin{aligned} p_{X|Y}(x|y) &\approx \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x - (y + \sigma^2 \nabla \log p_Y(y))\|^2\right) \\ &\approx \mathcal{N}(y + \sigma^2 \nabla \log p_Y(y), \sigma^2 I) \end{aligned}$$

- To obtain this form, we use a linear approximation based on the fact that $y \approx x$ and ignore higher-order terms.

- Source: Lecture Notes by Ernest K. Ryu (UCLA Math), https://ernestryu.com/courses/FM.html

Using two consecutive terms(1)

Relationship Between Terms

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

$$= \frac{q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0)\, q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \tag{2}$$

$$= \frac{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})\, q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \tag{3}$$

Remark

- This result is consistent with our intuition that the calculation becomes explicit under the assumption that $\mathbf{x}_0$ is known.

Using two consecutive terms(2)

The posterior is also a Gaussian

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

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\,\mathbf{x}_0$$

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

Using Markov property

Forward Process

$$q(\mathbf{x}_{0:T}) := q(\mathbf{x}_0) \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

Reverse Process

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

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\, \sigma_t^2 I\big).$$

- $\boldsymbol{\mu}_\theta$ is predicted by a neural net; σ_t is fixed.

Table of Contents

We need to maximize $p_\theta(\mathbf{x}_0)$!

Maximum Likelihood View

- The model probability of a clean image is

$$p_\theta(\mathbf{x}_0) = p(\mathbf{x}_0 \mid \theta).$$

- Therefore, training DDPM can be interpreted as a maximum likelihood estimation (MLE) problem:

$$\max_\theta \ \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \big[\log p_\theta(\mathbf{x}_0) \big].$$

- Equivalently, we minimize the negative log-likelihood:

$$\min_\theta \ \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \big[-\log p_\theta(\mathbf{x}_0) \big].$$

Designing Objective Function

Likelihood as an Expectation

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \tag{1}$$

$$= \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \tag{2}$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \tag{3}$$

Theorem (Jensen's inequality)

If f is a convex function and X is a random variable, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Designing Objective Function

- **Applying Jensen's Inequality**
  Since $-\log(\cdot)$ is convex,

$$\mathbb{E}_{q(\mathbf{x}_0)}[-\log p_\theta(\mathbf{x}_0)] = \mathbb{E}_{q(\mathbf{x}_0)}\left[-\log \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right]\right] \tag{1}$$

$$\leq \mathbb{E}_{q(\mathbf{x}_0)}\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] \tag{2}$$

$$= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] \tag{3}$$

- **Note that** $q(\mathbf{x}_{0:T}) = q(\mathbf{x}_0)\, q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ and

$$\mathbb{E}_{q(\mathbf{x}_0)}\left[\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[\cdot]\right] = \iint q(\mathbf{x}_0) q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)[\cdot] d\mathbf{x}_{1:T} d\mathbf{x}_0 \tag{4}$$

$$= \iint q(\mathbf{x}_{0:T})[\cdot] d\mathbf{x}_{1:T} d\mathbf{x}_0 = \mathbb{E}_{q(\mathbf{x}_{0:T})}[\cdot] \tag{5}$$

Designing Objective Function

Substituting Markov Factorizations

Using

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \qquad q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \tag{1}$$

we obtain

$$\mathbb{E}_{q(\mathbf{x}_0)}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log \frac{p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{\prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] \tag{2}$$

$$= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] =: L \tag{3}$$

- Note that $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ as $\mathbf{T}$ goes to infinity.

Our Objective Function

Loss as a Sum of KL-Divergence Terms

$$L = \mathbb{E}_q \Bigg[\underbrace{\sum_{t=2}^{T} D_{\mathrm{KL}}(q(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{X}_0) \parallel p_\theta(\mathbf{X}_{t-1} \mid \mathbf{X}_t))}_{L_{t-1}\text{: Denoising Process Matching}} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{X}_T \mid \mathbf{X}_0) \parallel p(\mathbf{X}_T))}_{L_T\text{: Prior Matching}} - \underbrace{\log p_\theta(\mathbf{X}_0 \mid \mathbf{X}_1)}_{L_0\text{: Reconstruction Term}} \Bigg]$$

- We arrive at a loss function expressed in terms of the KL divergence.
- Our primary focus will be on the L_{t-1} term.

IgnoringL_T Term

Forward Process and the L_T Term

- The forward process q is defined by a fixed variance schedule β_t, with no learnable parameters.
- The loss term L_T measures the KL divergence between the final noised distribution $q(\mathbf{x}_T|\mathbf{x}_0)$ and the prior $p(\mathbf{x}_T) \sim \mathcal{N}(0, \mathbf{I})$.

$$L_T = D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))$$

- Since this term contains no parameters θ to be trained, it can be treated as a constant and ignored during optimization.

Analyzing L_{t-1} Term (1)

Reverse Process Setup

The reverse process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is modeled as a Gaussian, where the main learnable quantity is the mean:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}\big).$$

Following the DDPM paper, the variance σ_t^2 is fixed rather than learned.

Parameterizing the Mean via Noise Prediction

Instead of predicting the mean directly, the network $\boldsymbol{\epsilon}_\theta$ is trained to predict the noise contained in the noisy sample $\mathbf{x}_t$.

$$\boldsymbol{\mu}_t(\mathbf{x}_t, \boldsymbol{\epsilon}) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon} \right), \tag{1}$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right). \tag{2}$$

Analyzing L_{t-1} Term(2)

The KL Divergence for L_{t-1}

The core of the loss is L_{t-1}, which aims to match our learned reverse step p_θ to the true posterior q.

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

Result of Expansion

Expanding the log-PDFs of the Gaussians leads to:

$$\mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \left(\log \frac{\sigma_t^2}{\tilde{\beta}_t} + D\tilde{\beta}_t + \frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2}{D} - D \right) \right] + C_1$$

Simplifying L_{t-1} to an MSE

Ignoring Constants

In the expanded form of L_{t-1}, the only term that depends on our learnable parameter θ is the difference between the means, $\boldsymbol{\mu}_\theta$. All other terms can be treated as constants during optimization.

Simplified Form

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right] + C_2$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) \approx \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$$

Deriving the Target Mean $\tilde{\mu}_t$

Key Steps and Intuition

- Recall the Closed-Form :

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

- Derivation Summary for $\tilde{\boldsymbol{\mu}}_t$:

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\mathbf{x}_0 \tag{2}$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \approx \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) \tag{3}$$

- Reverse process depends only on the initial point $\mathbf{x}_0$ and the sampled noise path $\boldsymbol{\epsilon}$ which fits our intuition.

Deriving the Target Mean $\tilde{\mu}_t$ (Detailed)

Make the learning target explicit

We rewrite $\mathbf{x}_t$ as a function of $(\mathbf{x}_0, \boldsymbol{\epsilon})$: $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\,\mathbf{x}_0 \tag{1}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \tag{2}$$

$$= \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\right)\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon} \tag{3}$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \tag{4}$$

where $\beta_t := 1 - \alpha_t$.

Final Derivation

Substituting the Means to Reach the Final Objective

$$\mathbb{E}_{\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon})}\left[\frac{1}{2\sigma_t^2}\left\|\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon})-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)\right)-\boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}),t)\right\|^2\right] \tag{1}$$

$$=\mathbb{E}_{\mathbf{x}_t}\left[\frac{1}{2\sigma_t^2}\left\|\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)\right)-\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right)\right)\right\|^2\right] \tag{2}$$

$$=\mathbb{E}_{\mathbf{x}_t}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right\|^2\right] \tag{3}$$

$$=\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\,t\right)\right\|^2\right] \tag{4}$$

Simplified Objective L_{simple}

Unweighted loss used in practice

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

Remark

- The original objective includes a weighting factor $\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}$.
- This weight is larger for small t (low-noise steps) and smaller for large t (high-noise steps), thus emphasizing denoising at low noise levels.

Pseudocode : Training

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

Pseudocode : Sampling

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** $\mathbf{x}_0$

Table of Contents

References

J. Ho, A. Jain, and P. Abbeel.
Denoising Diffusion Probabilistic Models.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Fresh Shrimp Sushi (Blog).
"[Paper Review] Denoising Diffusion Probabilistic Models (DDPM)"

Fresh Shrimp Sushi (Blog).
"Relative Entropy (Kullback-Leibler Divergence) in Classical Information Theory"

Ernest K. Ryu.
Lecture Notes for 'Generative AI and Foundation Models' (Ch. 3).
SNU, Spring 2024.

Table of Contents

Designing Objective Function

Separating the $t = 1$ Term

$$L = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})}\right] \tag{1}$$

$$= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})} - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)}\right] \tag{2}$$

Using the Posterior Identity

Since

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)}, \tag{3}$$

we have

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)\, q(\mathbf{x}_t \mid \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)} \tag{4}$$

Applying the Identity and Telescoping

$$L = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=2}^{T} \log \left(\frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \right) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} \right] \tag{1}$$

$$= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \log \left(\prod_{t=2}^{T} \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \right) - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} \right] \tag{2}$$

$$= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log p(\mathbf{x}_T) - \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} - \log \frac{q(\mathbf{x}_1 \mid \mathbf{x}_0)}{q(\mathbf{x}_T \mid \mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_1 \mid \mathbf{x}_0)} \right] \tag{3}$$

Designing Objective Function

Rewriting in KL Form

$$L = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\right] \tag{1}$$

$$= \mathbb{E}_{q(\mathbf{x}_0)}\left[D_{\mathrm{KL}}\big(q(\mathbf{x}_T \mid \mathbf{x}_0) \parallel p(\mathbf{x}_T)\big) + \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\Big[D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big)\Big]\right] \tag{2}$$
$$- \mathbb{E}_{q(\mathbf{x}_{0:T})}\big[\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)\big].$$

- This is the KL-based decomposition.

Designing Objective Function

The First KL Term

$$\mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\right] = \int q(\mathbf{x}_{0:T}) \log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\, d\mathbf{x}_{0:T} \tag{1}$$

$$= \int q(\mathbf{x}_0)\, q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\, d\mathbf{x}_{0:T} \tag{2}$$

$$= \iiint q(\mathbf{x}_0)\, q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\, d\mathbf{x}_{1:T-1}\, d\mathbf{x}_T\, d\mathbf{x}_0 \tag{3}$$

$$= \iint q(\mathbf{x}_0)\, q(\mathbf{x}_T \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\, d\mathbf{x}_T\, d\mathbf{x}_0 \tag{4}$$

$$= \mathbb{E}_{q(\mathbf{x}_0)}\left[\int q(\mathbf{x}_T \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_T \mid \mathbf{x}_0)}{p(\mathbf{x}_T)}\, d\mathbf{x}_T\right] \tag{5}$$

$$= \mathbb{E}_{q(\mathbf{x}_0)}\left[D_{\mathrm{KL}}\left(q(\mathbf{x}_T \mid \mathbf{x}_0) \,\|\, p(\mathbf{x}_T)\right)\right]. \tag{6}$$

The Second KL Term (1)

$$\sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \right] \tag{7}$$

$$= \sum_{t=2}^{T} \int q(\mathbf{x}_{0:T}) \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \, d\mathbf{x}_{0:T} \tag{8}$$

$$= \sum_{t=2}^{T} \int q(\mathbf{x}_0)\, q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \, d\mathbf{x}_{0:T} \tag{9}$$

$$= \sum_{t=2}^{T} \int q(\mathbf{x}_0)\, q(\mathbf{x}_{1:t-2}, \mathbf{x}_{t+1:T} \mid \mathbf{x}_t, \mathbf{x}_0)\, q(\mathbf{x}_{t-1}, \mathbf{x}_t \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \, d\mathbf{x}_{0:T} \tag{10}$$

$$= \sum_{t=2}^{T} \iiint q(\mathbf{x}_0)\, q(\mathbf{x}_{1:t-2}, \mathbf{x}_{t+1:T} \mid \mathbf{x}_t, \mathbf{x}_0)\, q(\mathbf{x}_t \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \, d\mathbf{x}_{1:t-2}\, d\mathbf{x}_{t+1:T}\, d\mathbf{x}_{t-1:t}\, d\mathbf{x}_0 \tag{11}$$

The Second KL Term (2)

$$= \sum_{t=2}^{T} \iint q(\mathbf{x}_0)\, q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)\, q(\mathbf{x}_t \mid \mathbf{x}_0) \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}\, d\mathbf{x}_{t-1:t}\, d\mathbf{x}_0 \tag{12}$$

$$= \sum_{t=2}^{T} \iint q(\mathbf{x}_0)\, q(\mathbf{x}_t \mid \mathbf{x}_0)\, D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big)\, d\mathbf{x}_t\, d\mathbf{x}_0 \tag{13}$$

$$= \int q(\mathbf{x}_0) \left[\sum_{t=2}^{T} \int q(\mathbf{x}_t \mid \mathbf{x}_0)\, D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big)\, d\mathbf{x}_t \right] d\mathbf{x}_0 \tag{14}$$

$$= \mathbb{E}_{q(\mathbf{x}_0)} \left[\sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) \right] \right]. \tag{15}$$

Thank you for your attention!

Questions or Comments?