Review of "Denoising Diffusion Probabilistic Models" by Jonathan Ho, Ajay Jain, and Pieter Abbeel

Summary by Camden Kitowski

Contents

1 Full Citation

Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851

2 Paper Summary

- **Diffusion Probabilistic Model**: A model trained using variational inference using a parameterized Markov chain to generate high quality samples after a finite time.
- **Forward Process**: This process gradually adds noise to data samples until the data is a normal distribution of noise. The closed form of the forward process:

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

- **Reverse Process**: This process gradually removes noise to obtain a new data sample. The reverse process follows:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right)\prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right), \quad p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right).$$

3 Concept Review

3.1 Math Derivations

3.1.1 Forward Process

The forward process, denoted with q, is a fixed Markov chain. The forward process gradually adds noise from a Gaussian distribution to the data over t time steps with a variance schedule from $\beta_1, \ldots, \beta_T$. Let x_0 be the original image and let x_t be the final image with isotropic Gaussian noise.

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \quad q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{0}$$

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

$$\begin{aligned}
q\left(x_t \mid x_{t-1}\right) &= \mathcal{N}\left(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) && (1)\\
&= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon && (2)\\
&= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon && (3)\\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon && (4)\\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}x_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\epsilon && (5)\\
&= \sqrt{\alpha_t\alpha_{t-1}\ldots\alpha_1\alpha_0}x_0 + \sqrt{1-\alpha_t\alpha_{t-1}\ldots\alpha_1\alpha_0}\epsilon && (6)\\
&= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon && (7)
\end{aligned}$$

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right) \tag{8}$$

The following derivations break down steps 3 and 4.

$$\begin{aligned}
x_t &= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon_{t-1}\\
&= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1}\\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\varepsilon_{t-1}\\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t\left(1-\alpha_{t-1}\right)}\varepsilon_{t-2} + \sqrt{1-\alpha_t}\varepsilon_{t-1}
\end{aligned}$$

Let $X = \sqrt{\alpha_t\left(1-\alpha_{t-1}\right)}\varepsilon_{t-2}$
Let $Y = \sqrt{1-\alpha_t}\varepsilon_{t-1}$

Using reparameterization trick

$$\begin{aligned}
0 + \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-2} \cdots\cdots\cdots\cdots \to X \sim \mathcal{N}(0, \alpha_t(1-\alpha_{t-1})I)\\
0 + \sqrt{1-\alpha_t}\epsilon_{t-1} \cdots\cdots\cdots\cdots \to Y \sim \mathcal{N}(0, (1-\alpha_t)I)
\end{aligned}$$

$$\begin{aligned}
&\text{If } X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \quad Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)\\
&\quad Z = X + Y\\
&\text{Then } Z \sim \mathcal{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)
\end{aligned}$$

$$\begin{aligned}
Z \sim \mathcal{N}\left(0, \sigma_X^2 + \sigma_Y^2\right) \cdots\cdots\cdots\cdots \to \sigma_X^2 + \sigma_Y^2 &= \alpha_t\left(1-\alpha_{t-1}\right) + \left(1-\alpha_t\right)\\
&= \alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t\\
Z \sim \mathcal{N}\left(0, \left(1-\alpha_t\alpha_{t-1}\right)I\right) \leftarrow \cdots\cdots\cdots\cdots\cdots\cdots &= 1 - \alpha_t\alpha_{t-1}
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \bar{\varepsilon}_{t-2} \\
&: \\
&= \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_1} x_0 + \sqrt{1 - \alpha_t \alpha_{t-1} \ldots \alpha_1} \varepsilon \\
&= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon \qquad \text{same result in line 7 above}
\end{aligned}
$$

3.1.2 ELBO and Reverse Process

The loss function is derived below. The loss function is the $-\log\left(p_\theta\left(x_0\right)\right)$. However, this is intractable because it depends on all other time steps before x_0 starting at x_T. This would mean keeping track of $t-1$ random variables, which is not possible in practice. The solution is to compute the variational lower bound for this objective. The overall setup of diffusion models is close to the setup of variational autoencoders.

$$
-\log\left(p_\theta\left(x_0\right)\right) \leq -\log\left(p_\theta\left(x_0\right)\right) + \frac{D_{KL}\left(q\left(x_{1:T} \mid x_0\right) || p_\theta\left(x_{1:T} \mid x_0\right)\right)}{\downarrow}
$$

$$
\log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{1:T} \mid x_0\right)}\right)
$$

$$
p_\theta(x_{1:T} \mid x_0)
$$

$$
\frac{p_\theta\left(x_0 \mid x_{1:T}\right) p_\theta\left(x_{1:T}\right)}{p_\theta\left(x_0\right)} \quad \text{Bayes' rule}
$$

$$
\frac{p_\theta\left(x_0, x_{1:T}\right)}{p_\theta\left(x_0\right)} \quad \text{Joint probability}
$$

$$
\frac{p_\theta\left(x_{0:T}\right)}{p_\theta\left(x_0\right)} \quad \text{Combine probability}
$$

$$
\log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{\frac{p_\theta\left(x_{0:T}\right)}{p_\theta\left(x_0\right)}}\right) \rightarrow \log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{0:T}\right)}\right) + \log\left(p_\theta\left(x_0\right)\right)
$$

$$
-\log\left(p_\theta\left(x_0\right)\right) \leq -\log\left(p_\theta\left(x_0\right)\right) + \log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{0:T}\right)}\right) + \log\left(p_\theta\left(x_0\right)\right)
$$

$$
-\log\left(p_\theta\left(x_0\right)\right) \leq \log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{0:T}\right)}\right)
$$

$$
\mathbb{E}\left[-\log p_\theta\left(x_0\right)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta\left(x_{0:T}\right)}{q\left(x_{1:T} \mid x_0\right)}\right] =: L
$$

ELBO Derivation Continued

$$
\begin{aligned}
L &= \mathbb{E}_q\left[\log\left(\frac{q\left(x_{1:T} \mid x_0\right)}{p_\theta\left(x_{0:T}\right)}\right)\right] \\
&= \mathbb{E}_q\left[\log\left(\frac{\prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right)}{p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right)}\right)\right] \quad \text{Product form} \\
&= \mathbb{E}_q\left[-\log\left(\mathrm{p}\left(\mathrm{x}_T\right)\right)+\sum_{t=1}^{T} \log\left(\frac{q\left(x_t \mid x_{t-1}\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)\right] \quad \text{Turn products into sums} \\
&= \mathbb{E}_q\left[-\log(p\left(x_T\right))+\sum_{t=2}^{T} \log\left(\frac{q\left(x_t \mid x_{t-1}\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right)\right] \quad \text{Split up summation} \\
&= \mathbb{E}_q\left[-\log\left(\mathrm{p}\left(\mathrm{x}_T\right)\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right) q\left(x_t \mid x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right) q\left(x_{t-1} \mid x_0\right)}\right)+\log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right)\right] \quad \text{Bayes' rule and condition on } x_0 \\
&= \mathbb{E}_q\left[-\log\left(\mathrm{p}\left(\mathrm{x}_T\right)\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_t \mid x_0\right)}{q\left(x_{t-1} \mid x_0\right)}\right)+\log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right)\right] \quad \text{Split up summation} \\
&= \mathbb{E}_q\left[-\log\left(\mathrm{p}\left(\mathrm{x}_T\right)\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\log\left(\frac{q\left(x_T \mid x_0\right)}{q\left(x_1 \mid x_0\right)}\right)+\log\left(\frac{q\left(x_1 \mid x_0\right)}{p_\theta\left(x_0 \mid x_1\right)}\right)\right] \quad \text{Simplify summation} \\
&= \mathbb{E}_q\left[-\log\left(\mathrm{p}\left(\mathrm{x}_T\right)\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)+\log q\left(x_T \mid x_0\right)-\log p_\theta\left(x_0 \mid x_1\right)\right] \quad \text{Log rules and cancel} \\
&= \mathbb{E}_q\left[\log\left(\frac{q\left(x_T \mid x_0\right)}{\mathrm{p}\left(\mathrm{x}_T\right)}\right)+\sum_{t=2}^{T} \log\left(\frac{q\left(x_{t-1} \mid x_t, x_0\right)}{p_\theta\left(x_{t-1} \mid x_t\right)}\right)-\log p_\theta\left(x_0 \mid x_1\right)\right] \quad \text{Combine terms} \\
&= \mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)}_{L_T}+\sum_{t=2}^{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}} \underbrace{-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0}\right] \quad \text{KL divergence notation}
\end{aligned}
$$

$\underline{L_T}$ **- Constant Term** - Since q does not have any learnable parameters, this term will be constant during training. This term can be ignored.

$\underline{L_{t-1}}$ **- Denoising Term - Reverse Process** - The reverse process starts with an isotropic Gaussian image, x_t, and denoises the image to make a sample that resembles an image in the training data, x_0. The reverse process is an iterative process to denoise the image with each new image dependent on the last image within the sequence. This term compares the actual noise (q) and predicted noise (p).

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

$$
q\left(x_{t-1} \mid x_t, x_0\right)=\mathcal{N}\left(x_{t-1} ; \tilde{\mu}_t\left(x_t, x_0\right), \tilde{\beta}_t I\right) \quad \text{Actual Mean}
$$

$$
\text{where} \quad \tilde{\mu}_t\left(x_t, x_0\right):=\frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1-\bar{\alpha}_t} x_0+\frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t \quad \text{and} \quad \tilde{\beta}_t:=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t, \quad x_0=\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t-\sqrt{1-\bar{\alpha}_t} \varepsilon_t\right)
$$

$$
\begin{aligned}
\tilde{\mu}_t\left(x_t, x_0\right) &= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 \\
\tilde{\mu}_t\left(x_t\right) &= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_t\right) \quad \text{Substitute } x_0 \\
&= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \frac{1}{\sqrt{\bar{\alpha}_{t-1} \cdot \alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_t\right) \quad \text{Break up } \alpha_t \\
&= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\beta_t}{\sqrt{\alpha_t\left(1-\bar{\alpha}_t\right)}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_t\right) \quad \text{Cancel terms} \\
&= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\beta_t}{\sqrt{\alpha_t\left(1-\bar{\alpha}_t\right)}} x_t - \frac{\beta_t}{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)}\sqrt{1-\bar{\alpha}_t}\varepsilon_t \quad \text{Break up terms} \\
&= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t\left(1-\bar{\alpha}_t\right)}} x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t\left(1-\bar{\alpha}_t\right)}}\sqrt{1-\bar{\alpha}_t}\varepsilon_t \quad \text{Substitute } \beta_t \\
&= \frac{\alpha_t}{\sqrt{\alpha_t}} \cdot \frac{\left(1-\bar{\alpha}_{t-1}\right)}{\left(1-\bar{\alpha}_t\right)} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)} x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \cdot \frac{\sqrt{1-\bar{\alpha}_t}}{\left(1-\bar{\alpha}_t\right)} \cdot \varepsilon_t \quad \text{Reformulate } \sqrt{\alpha_t} \\
&= \frac{\alpha_t - \bar{\alpha}_{t-1} \cdot \alpha_t}{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)} x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}} \cdot \frac{1}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_t \quad \text{Reformulate } \frac{\sqrt{1-\bar{\alpha}_t}}{\left(1-\bar{\alpha}_t\right)} \\
&= \frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)} x_t - \frac{1}{\sqrt{\alpha_t}} \cdot \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_t \quad \text{Combine terms} \\
&= \frac{1-\overline{\alpha}_t}{\sqrt{\alpha_t}\left(1-\overline{\alpha}_t\right)} x_t - \frac{1}{\sqrt{\alpha}} \cdot \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_t \quad \text{Cancel terms} \\
&= \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1}{\sqrt{\alpha_t}} \cdot \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_t \quad \text{Cancel terms} \\
\tilde{\mu}_t\left(x_t\right) &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right)
\end{aligned}
$$

$$
\begin{aligned}
p_\theta\left(x_{t-1} \mid x_t\right) &:= \mathcal{N}\left(x_{t-1}; \boldsymbol{\mu}_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \quad \text{Predicted mean with neural network} \\
p_\theta\left(x_{t-1} \mid x_t\right) &:= \mathcal{N}\left(x_{t-1}; \boldsymbol{\mu}_\theta\left(x_t, t\right), \beta I\right) \quad \text{Beta is fixed}
\end{aligned}
$$

Initially, the neural network approximated the mean.

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

$$
\boldsymbol{\mu}_\theta\left(x_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(x_t, t\right)\right)
$$

The authors use a simple Mean Squared Error to compare the actual mean and predicted mean.

$$
\begin{aligned}
L_t &= \mathbb{E}_{x_0,\varepsilon}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(x_t\right)-\boldsymbol{\mu}_\theta\left(x_t,t\right)\right\|^2\right] \\
&= \mathbb{E}_{x_0,\varepsilon}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right)-\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(x_t,t\right)\right)\right\|^2\right] \\
&= \mathbb{E}_{x_0,\varepsilon}\left[\frac{\left(1-\alpha_t\right)^2}{2\alpha_t\left(1-\bar{\alpha}_t\right)\sigma_t^2}\left\|\varepsilon_t-\varepsilon_\theta\left(x_t,t\right)\right\|^2\right]
\end{aligned}
$$

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

Authors use simple MSE without the weighted term because the quality of samples is better. It is unnecessary to use the weighted equation. Standard L2 loss with predicted noise and actual noise works better.

L_0 **- Reconstruction Term** - This term is thrown away by the authors. This term can be approximated by the same neural network in the previous term. Ignoring this term makes sampling better.

$$
p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right) = \prod_{i=1}^{D}\int_{\delta_-\left(x_0^i\right)}^{\delta_+\left(x_0^i\right)} \mathcal{N}\left(x;\mu_\theta^i\left(\mathbf{x}_1,1\right),\sigma_1^2\right)dx \tag{13}
$$

$$
\delta_+(x) = \begin{cases} \infty & \text{if } x=1 \\ x+\frac{1}{255} & \text{if } x<1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x=-1 \\ x-\frac{1}{255} & \text{if } x>-1 \end{cases}
$$

3.2 GitHub Repository Link

https://github.com/CamdenKitowski/DenoisingDiffusionModels

4 Math Terms and Definitions

Variational Inference - the method/approach to approximate complex probability distribution P with a simpler one Q. This is used when exact inference is intractable. Usually done by minimizing KL divergence between Q and P.

Evidence Lower Bound (ELBO) - objective function (function that is either maximized or minimized to solve a problem) used in variational inference. It represents a lower bound on the log-likelihood of the observed data.
Two part objective:

- Maximize expected log-likelihood of data under the approximate distribution Q
- Minimize KL divergence between Q and prior distribution over the latent variables

Kullback-Leibler (KL) divergence - known as relative entropy - measure of how one probability distribution diverges from another one.
For discrete variables:

$$
D_{KL}(P||Q) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right)
$$

For continuous variables

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

where $p(x)$ and $q(x)$ are probability density functions of P and Q, respectively. The divergence is always non-negative. If it is zero, then P and Q are the same distribution.

Likelihood and Log-Likelihood -
can you explain why we start with negative log likelihood?

Likelihoood - the likelihood of a set of parameters θ given observations X is the probability of X given θ, denoted $P(X|\theta)$. The likelihood function $L(\theta|X)$ is defined as the joint probability/density of the observed data as a function of the parameters θ. Use this to find the parameters values that maximize the probability. Known as Maximum Likelihood Estimation

Log-Likelihood - Use the natural logarithm of likelihood function, denoted as $\log(L(\theta|X))$. Turns products into sums, making it easier to differentiate and more stable when looking for the maximum likelihood estimates.

More Notes - Both serve the same purpose. Maximizing the likelihood is equivalent to maxmizing the log-likelihood because log is strictly increasing function. So the set of parameters that maximizes the likelihood also maximizes the log-likehood.

Markov Chain - model systems that transition from one state to another, where the probability of each subsequent state depends only on the current state and not on the sequence of events that preceded it

Expected Value - measure of central tendency of its probability distribution. Calculated by summing all possible values of the variable, each multiplied by its probability of occurrence

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

Conditional Mean/Expectation - it's the expected value of a random variable given that a certain condition(s) holds true.

$$E[Y|X = x] = \sum_{y} y \cdot P(Y = y|X = x)$$

Probability vs. Expected Value -
- $P(y|x)$ is used when we're interested in the chance of an event happening.
- $E(y|x)$ is used in situations where we want to predict or estimate the average outcome of a random variable given some condition or information.

Forward Process Equation vs. Forward Process Posteriors -

Forward Process Equation - tells you how to go from one time step to the next by adding noise. It takes the form of a Markov Chain in the form of Gaussian transitions.

Forward Process Posteriors - Describe the probability distribution over possible states at each time step that is conditioned on the original data. Expressed in terms of the posteriors parameters which is mean and variance.

More Notes - The equation provides a procedural generation path, and the posteriors provide a statistical description of where you might end up along that path.