Denoising Diffusion Probabilistic Models

Pramook Khungurn

November 7, 2022

This note is written to summarize papers related the **denoising diffusion probabilistic models** (DDPMs), which are a new type of generative models that have become popular since 2020.

1 Notations

- In general, we shall denote a data item with the symbol $\mathbf{x}$, which is generally a vector in $\mathbb{R}^d$.
- When several data items are sampled from the same distribution, we differentiate between them by subscripts; for examples, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and so on.
- We will also deal with sequence of data items. In particular, we will deal with Markov chains in which the next element in the sequence is generated from the previous one. Different elements of the sequence are differentiated by superscripts which are always inside a pair of parentheses; for example $(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)})$.
- Writing $(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)})$ is cumbersome. We can abbreviate it with $\mathbf{x}^{(0:100)}$.
- When referring to a subsequence such as $(\mathbf{x}^{(2)}, \mathbf{x}^{(5)}, \mathbf{x}^{(11)}, \mathbf{x}^{(29)})$, we can abbreviate it with $\mathbf{x}^{(2,5,11,29)}$.
- Sometimes, we need to refer to a subsequence whose elements have indices in a set $\mathtt{I} = \{i_1, i_2, \ldots, i_k\}$. In such a case, we can abbreviate $(\mathbf{x}^{(i_1)}, \mathbf{x}^{(i_2)}, \ldots, \mathbf{x}^{(i_k)})$ with just $\mathbf{x}^{(\mathtt{I})}$.
  - In this way, $\mathbf{x}^{(2,5,11,29)}$ is an abbreviation of $\mathbf{x}^{(\{2,5,11,29\})}$.
- In fact, we shall let $\{0:100\}$ denote the set $\{0, 1, 2, \ldots, 100\}$. In this way, $\mathbf{x}^{(0:100)}$ is just an abbreviation for $\mathbf{x}^{(\{0:100\})}$.
- Using a set as a superscript allows us to do $\mathbf{x}^{(\{0:100\}-\{2,5,11,29\})}$, which denotes the subsequence of $\mathbf{x}^{(0:100)}$ with $\mathbf{x}^{(2)}$, $\mathbf{x}^{(5)}$, $\mathbf{x}^{(11)}$, and $\mathbf{x}^{(29)}$ removed.
- We will also deal with probabilities of data items and sequences. Let $\mathtt{I} = \{i_1, i_2, \ldots, i_k\}$. Then, we may denote the joint probability $p(\mathbf{x}^{(i_1)}, \mathbf{x}^{(i_2)}, \ldots, \mathbf{x}^{(i_k)})$ with just $p(\mathbf{x}^{(\mathtt{I})})$.
- We will also deal with conditional probabilities. Let $\mathtt{J} = \{j_1, j_2, \ldots, j_\ell\}$. Then, the conditional probability $p(\mathbf{x}^{(j_1)}, \mathbf{x}^{(j_2)}, \ldots, \mathbf{x}^{(j_\ell)} | \mathbf{x}^{(i_1)}, \mathbf{x}^{(i_2)}, \ldots, \mathbf{x}^{(i_k)})$ can be abbreviated with $p(\mathbf{x}^{(\mathtt{J})} | \mathbf{x}^{(\mathtt{I})})$.
- In some cases, we might put the superscripts on the probability function itself. That is, we may write

  $$p^{(100|0)}(\mathbf{x}^{(100)} | \mathbf{x}^{(0)})$$

  to mean the same thing as $p(\mathbf{x}^{(100)} | \mathbf{x}^{(0)})$. The great thing about the newly introduced notation is that we can write

  $$p^{(100|0)}(\mathbf{x} | \mathbf{x}'),$$

  to signify that we are treating the conditional probability as a function of two data items, and we are evaluating it at $(\mathbf{x}, \mathbf{x}')$.

2 The Generative Modeling Problem

- We are given n data items $\mathbf{x}_1^{(0)}$, $\mathbf{x}_2^{(0)}$, …, $\mathbf{x}_N^{(0)}$ that are sampled i.i.d. from a probability distribution $q(\mathbf{x}^{(0)})$, which is unknown to us.
  - As usual, each data item is a d-dimensional vector. In other words, $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ for all i.
- We are interested in modeling $q(\mathbf{x})$ by finding a model $p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})$ with parameters $\boldsymbol{\theta}$ that best approximates it.
  - To reduce levels of subscription, we will sometimes write $p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})$ as $p(\mathbf{x}^{(0)}; \boldsymbol{\theta})$.
- We would like to know (1) how to estimate the parameters $\boldsymbol{\theta}$ and (2) how to sample from the model given the parameters.

3 Denoising Diffusion Probabilistic Models (DDPMs)

- Sohl-Dickstein et al. first described the idea of such a model in 2015 [SWMG15], and then Ho et al. popularized it in 2020 [HJA20].
- The main ideas are as follows:
  - There is a Markov chain $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, …, $\mathbf{x}^{(T)}$ called the **forward process**, with the following properties:
    1. $\mathbf{x}^{(t)}$ is obtained by scaling $\mathbf{x}^{(t-1)}$ down and adding a small amount of noise.
    2. $\mathbf{x}^{(T)}$ is very close to the isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$.
  - Because each step of the forward process is simple, we can revert it to get $\mathbf{x}^{(t-1)}$ from $\mathbf{x}^{(t)}$.
  - If we can do so, we can sample $\mathbf{x}^{(0)}$ according to $q(\mathbf{x}^{(0)})$ as follows:
    * Sample $\mathbf{x}^{(T)}$ according to $\mathcal{N}(\mathbf{0}, I)$.
    * Use the above reversal process to compute $\mathbf{x}^{(T-1)}$, $\mathbf{x}^{(T-2)}$, …, $\mathbf{x}^{(1)}$, and finally $\mathbf{x}^{(0)}$.

3.1 The Forward Process

- Ho et al. uses the following Markov chain:

$$\begin{aligned}\mathbf{x}^{(0)} &\sim q(\mathbf{x}^{(0)}),\\ \mathbf{x}^{(t)} &\sim \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t I)\end{aligned}$$

  for all $t = 1, 2, \ldots, T$. Here, β_1, β_2, …, β_T are small positive constants collectively called the **variance schedule**.
- They fix $T = 1000$.
- The picked a linear progression where $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$ as the variance schedule.
- Note that we can write $\mathbf{x}^{(t)}$ in terms for $\mathbf{x}^{(t-1)}$ as follows:

$$\mathbf{x}^{(t)} = \sqrt{1-\beta_t}\mathbf{x}^{(t-1)} + \sqrt{\beta_t}\boldsymbol{\xi}$$

  where $\boldsymbol{\xi}$ is a random variable distributed according to $\mathcal{N}(\mathbf{0}, I)$. The conditional probability of $\mathbf{x}^{(t)}$ given $\mathbf{x}^{(t-1)}$ (i.e., the transition kernel) is given by:

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t I) = \frac{1}{(2\pi\beta_t)^{d/2}} \exp\left(\frac{-\|\mathbf{x}^{(t)} - \sqrt{1-\beta_t}\mathbf{x}^{(t-1)}\|_2^2}{2\beta_t}\right).$$

- With the conditional probablity above, the probability of sampling $\mathbf{x}^{(0:T)} = (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$ is given by

$$q(\mathbf{x}^{(0:T)}) = q(\mathbf{x}^{(0)}) \prod_{t=1}^{T} q(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}).$$

- After carrying out the t transitions, here's what the forward process does to the original data.

  Proposition 1. *Let $\alpha_t = 1 - \beta_t$. Let $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. For any $1 \leq t \leq T$, we have that*

$$q(\mathbf{x}^{(t)} | \mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1 - \overline{\alpha}_t)I).$$

- The particular choice of the variance schedule in the Ho et al. paper yields $\overline{\alpha}_T \approx 10^{-4.385}$.
- So, if $\mathbf{x}^{(0)}$'s components are always much smaller than $10^{-4.385/2} \approx 10^{-2.192}$, then $\mathbf{x}^{(T)}$'s distribution would be very close to $\mathcal{N}(\mathbf{0}, I)$.
  - This would always happen if $q(\mathbf{x}^{(0)})$ is a distribution of RGB images because each component of $\mathbf{x}^{(0)}$ would be in the range $[0, 1]$.
- We can also show that

  Proposition 2. *If $t \geq 2$, then*

$$q(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}, \mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(t)}, \mathbf{x}^{(0)}), \tilde{\beta}_t I)$$

  where

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) = \frac{\beta_t \sqrt{\overline{\alpha}_{t-1}}}{1 - \overline{\alpha}_t}\mathbf{x}^{(0)} + \frac{\sqrt{\alpha_t}(1 - \overline{\alpha}_{t-1})}{1 - \overline{\alpha}_t}\mathbf{x}^{(t)},$$
$$\tilde{\beta}_t = \beta_t \frac{1 - \overline{\alpha}_{t-1}}{1 - \overline{\alpha}_t}.$$

3.2 The Backward Process

- The backward process is sometimes called the **generative process**.
- It is denoted by the probability function $p_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}$ denotes trainable parameters.
- It goes as follows:
  1. Sample $\mathbf{x}^{(T)} \sim p_{\boldsymbol{\theta}}^{(T)} = \mathcal{N}(\mathbf{0}, I)$.
  2. Sample $\mathbf{x}^{(t-1)} \sim p_{\boldsymbol{\theta}}^{(t-1|t)}(\cdot | \mathbf{x}^{(t)})$ for all $t = T, T-1, \ldots, 1$.

  We will specify the exact form of $p_{\boldsymbol{\theta}}^{(t-1|t)}$ later. (Spoiler: it is a Gaussian distribution whose mean is computed by a neural network.)
- With the above process, the probability of sampling $\mathbf{x}^{(0:T)}$ is

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

3.2.1 The Loss Function

- To find the optimal parameters $\boldsymbol{\theta}$, the standard approach would be to optimize the log-likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})$, which can be rewritten as

$$p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}) = E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\left[p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\right].$$

  (See Claim 10 in Appendix B.) So, the loss function would be

$$E_{\mathbf{x}^{(0)}\sim q}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})] = E_{\mathbf{x}^{(0)}\sim q}\left[-\log\left(E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\left[p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\right]\right)\right]. \quad (1)$$

  However, we are at a deadend because of the logartihm in the expectation.

- The solution would be to apply a trick commonly used in variation inferrence: use Jensen's inequality. Because log is a concave function, we have that

$$E[\log f(X)] \leq \log E[f(X)]$$

  for any random variable X. In other words,

$$-\log E[f(X)] \leq -E[\log f(X)].$$

  So, instead of minimizing $-\log E[f(X)]$ like in (1), we would be minimizing its upper bound $-E[\log f(X)]$.

- Using Jensen's inequality, we can show that

$$E_{\mathbf{x}^{(0)}\sim q}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})] \leq E_{\mathbf{x}^{(0:T)}\sim q}\Big[\log q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\Big] \quad (2)$$

  Again, see the proof in Claim 11 of Appendix B. Our goal, then, becomes minimizing the RHS of (2).

- To faciliate further discussion, let us call the RHS of (2) "L."

$$L := E_{\mathbf{x}^{(0:T)}\sim q}\Big[\log q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\Big]. \quad (3)$$

  We will also refer to it as the "variational lower bound" or the "evidence lower bound" of the log-likelihood (because L is the upperbound of the negative log-likelihood).

- We can write L as a sum of three terms:

$$\begin{aligned} L = &\underbrace{E_{\mathbf{x}^{(0)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\Big]}_{L_T} \\ &+ \underbrace{E_{\mathbf{x}^{(0,1)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\Big]}_{L_0} \\ &+ \sum_{t=2}^{T}\underbrace{E_{\mathbf{x}^{(0,t)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big)\Big]}_{L_{t-1}}. \end{aligned}$$

  See the proof in Claim 12 of Appendix B.

- We will now consider how to minimize each term in turn.

3.2.2 The L_T Term

- First, consider the L_T term:

$$L_T = E_{\mathbf{x}^{(0)} \sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)}) \,\big\|\, p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\Big].$$

  We can see that it does not depend on $\boldsymbol{\theta}$ at all. This is because $p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})$ is just $\mathcal{N}(\mathbf{0}, I)$, which does not depend on $\boldsymbol{\theta}$. Moreover, according to Proposition 1, $q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})$ is a fixed Gaussian distribution that is determined by the noise schedule. Because the noise schedule is a hyperparameter, it follows that $q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})$ also does not depend on $\boldsymbol{\theta}$. As a result, we can ignore this term during the optimization process.

3.2.3 The L_1, L_2, ..., L_{T-1} Terms

- To optimize these terms, we have to specify what $p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ is.
- The Ho et al. paper chooses it to be

$$p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) := \mathcal{N}\big(\mathbf{x}^{(t-1)}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t), \beta_t I\big)$$

  where $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\cdot, \cdot)$ is a neural network. The variance of the Gaussian was chosen empirically.

- Now, consider

$$D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big).$$

  We know from Proposition 2 that $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})$ is a Gaussian distribution of $\mathbf{x}^{(t-1)}$. Moreover, $p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ is a Gaussian distribution of $\mathbf{x}^{(t-1)}$ by definition. So, the KL-divergence can be computed using Corollary 9. Applying it, we have that

$$D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) = \frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\|^2}{2\beta_t} + C \tag{4}$$

  where C is a constant that does not depend on $\boldsymbol{\theta}$ or $\mathbf{x}^{(0)}$ or $\mathbf{x}^{(t-1)}$. (More specifically, C can be written in terms of d and the β_t's in the noise schedule.)

- Now,

$$L_{t-1} = E_{\mathbf{x}^{(0,t)} \sim q}\left[\frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\|^2}{2\beta_t}\right] + C,$$

  which we can rewrite as

$$L_{t-1} - C = E_{\mathbf{x}^{(0)} \sim q, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)}\left[\frac{1}{2\beta_t}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}\right) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right\|^2\right].$$

  See the derivation in the proof of Claim 14.

- We see from the above equation that we want $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ to predict $\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}\right)$ from $\mathbf{x}^{(t)}$ and t. This is equivalent to predicting $\boldsymbol{\xi}$ from $\mathbf{x}^{(t)}$. The Ho et al. paper utilizes this structure training a network $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, \cdot)$ so that $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t) = \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)$ would return $\boldsymbol{\xi}$ instead.

- We can write the relationship between $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ as follows. Because we want $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)$ to predict $\boldsymbol{\xi}$, it follows that we want the following equation to hold:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right).$$

  This gives

$$\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t) = \frac{\sqrt{\alpha_t(1-\overline{\alpha}_t)}}{\beta_t}\left(\frac{\mathbf{x}^{(t)}}{\sqrt{\alpha_t}} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right).$$

- Using $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ instead of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, we can rewrite $L_{t-1} - C$ as

$$L_{t-1} - C = E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)\right\|^2\right].$$

  See the derivation in the proof of Claim 15.

3.2.4 The L_0 Term

- The Ho et al. paper derives an expression for L_0 that involves taking into the fact that $\mathbf{x}^{(0)}$ have discrete values when each data item is an image. However, this leads to a complicated expression that they end up not using. So, we will derive a simpler expression.

- Recall that

$$p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) = \mathcal{N}(\mathbf{x}^{(0)}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, 1), \beta_1 I).$$

  So,

$$-\log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) = \frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, 1)\right\|^2 + C'$$

  where C' is a constant that does not depend on $\boldsymbol{\theta}$ or $\mathbf{x}^{(0)}$ or $\mathbf{x}^{(1)}$.

- The expression can be simplified further to

$$-\log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) = \frac{1}{2\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, 1)\right\|^2 + C'$$

  using the derivation in the proof of Claim 16. As a result,

$$L_0 - C' = E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_1}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_1}\boldsymbol{\xi}, 1)\right\|^2\right].$$

3.2.5 The Overall Loss Function

- Taking stock, we want to train the network $\boldsymbol{\xi}_\theta$ by finding

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} L = \arg\min_{\boldsymbol{\theta}}\left(L_T + L_0 + \sum_{t=2}^{T} L_{t-1}\right) = \arg\min_{\boldsymbol{\theta}}\left(L_0 + \sum_{t=2}^{T} L_{t-1}\right) \\ &= \arg\min_{\boldsymbol{\theta}}\left(E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_1}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_1}\boldsymbol{\xi}, 1)\right\|^2\right]\right. \\ &\qquad \left. + \sum_{t=2}^{T} E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)\right\|^2\right]\right)\end{aligned}$$

  Note that the terms are almost the same, except for the coefficients of the norm squared $\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot,\cdot)\|^2$.

- Ho et al. simplifies the above loss function further, dropping all the norm squared's coefficients. So, the optimization problem becomes

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left(\sum_{t=1}^{T} E_{\mathbf{x}^{(0)} \sim q, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)} \left[\left\| \boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t) \right\|^2 \right] \right) \tag{5}$$

$$= \arg\min_{\boldsymbol{\theta}} \left(E_{t \sim \{1,2,\ldots,T\}, \mathbf{x}^{(0)} \sim q, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)} \left[\left\| \boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t) \right\|^2 \right] \right). \tag{6}$$

- The process to train $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ is as follows.

  Initialize $\boldsymbol{\theta}$.

  while not satisfied **do**

    Sample t uniformly from the set $\{1, 2, \ldots, T\}$.
    Sample $\mathbf{x}^{(0)} \sim q$.
    Sample $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)$.
    Compute $\mathbf{x}^{(t)} \leftarrow \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}$.
    Compute the loss $\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\|^2$.
    Use the gradient of the loss to update $\boldsymbol{\theta}$.

  end while

3.2.6 Pseudocode for the Backward Process

- While we have specified how to use the backward process to generate a sample before, it was not concrete enough as $p_{\boldsymbol{\theta}}$ was not explicitly specified.

- The pseudocode that use the trained network is as follows.

  Sample $\mathbf{x}^{(T)} \sim \mathcal{N}(\mathbf{0}, I)$.
  for $t = T, T-1, T-2, \ldots, 1$ **do**
    Sample $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)$.
    Compute $\mathbf{x}^{(t-1)} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}_{\boldsymbol{\theta}^*}(\mathbf{x}^{(t)}, t)\right) + \sqrt{\beta_1}\boldsymbol{\xi}$.
  end for
  return $\mathbf{x}^{(0)}$

3.3 Alternative Derivation Through Score Matching

- The derivation we have been going through in the last section is long and tedious. However, I have a shorter but less rigorous derivation.

- The derivation relies on the following theorem.

  Theorem 3 (Tweedie's Formula). *Let $\mathbf{x}$ be a random variable. Let $\boldsymbol{\xi}$ be a random variable whose distribution is $\mathcal{N}(\mathbf{0}, I)$. Let $\mathbf{y} = \mathbf{x} + \sigma I$ where $\sigma > 0$. (In other words, $\mathbf{y} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)$.) Then,*

$$E[\mathbf{x}|\mathbf{y}] = \mathbf{y} + \sigma^2 \nabla \log p(\mathbf{y}).$$

- The gradient of the logarithm of a probability function $\nabla \log p(\mathbf{y})$ is called the **score** of $\mathbf{y}$.

- For the forward process, we have that

$$\mathbf{x}^{(t)} = \sqrt{1-\beta_t}\mathbf{x}^{(t-1)} + \sqrt{\beta_t}\boldsymbol{\xi}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)$. Applying Tweedie's formula, we have that

$$E[\sqrt{1-\beta_t}\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}] = \mathbf{x}^{(t)} + \beta_t \nabla \log q(\mathbf{x}^{(t)})$$
$$E[\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}] = \frac{\mathbf{x}^{(t)} + \beta_t \nabla \log q(\mathbf{x}^{(t)})}{\sqrt{1-\beta_t}}.$$

This means that, when we want to sample $\mathbf{x}^{(t-1)}$ given $\mathbf{x}^{(t)}$, the mean of the distribute had better be the RHS of the above equation. In other words, we should sample $\mathbf{x}^{(t-1)}$ according to:

$$\mathbf{x}^{(t-1)} \sim \mathcal{N}\left(\frac{\mathbf{x}^{(t)} + \beta_t \nabla \log q(\mathbf{x}^{(t)})}{\sqrt{1-\beta_t}}, \beta_t I\right).$$

- What is left to do is to estimate the score $\nabla \log q(\mathbf{x}^{(t)})$. We do this by training a score network $\mathbf{s}_{\boldsymbol{\theta}}(\cdot, \cdot)$ so that $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)$ would give a good estimate of $\nabla \log q(\mathbf{x}^{(t)})$.
- Training such a network relies on a result that Pascal Vincent [Vin11].

Theorem 4. *Let p_{data} be an arbitrary probability density, $\mathbf{x} \sim p_{\text{data}}$, $\mathbf{y} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)$, and $\mathbf{f}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ be a neural network with parameters $\boldsymbol{\theta}$. Moreover, let*

$$J^{\sigma}(\boldsymbol{\theta}) = E_{\mathbf{x}\sim p_{\text{data}},\ \boldsymbol{\xi}\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{\xi} + \sigma \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x} + \sigma\boldsymbol{\xi})\|_2^2\right].$$

Then,

$$\arg\min_{\boldsymbol{\theta}} J^{\sigma}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \left(E_{\mathbf{x}\sim p_{\text{data}},\ \mathbf{y}\sim\mathcal{N}(\mathbf{x},\sigma^2)}\left[\left\|f_{\boldsymbol{\theta}}(\mathbf{y}) - \nabla \log p(\mathbf{y})\right\|_2^2\right]\right).$$

- We can apply the above result to our problem. We have that

$$\mathbf{x}^{(t)} \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I).$$

Let q_t denote the probability distribution of $\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}$. Now, we can apply Theorem 4 by making the following substituion.

  - Substitute $\mathbf{x}$ with $\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}$.
  - Substitute p_{data} with q_t.
  - Substitute $\mathbf{y}$ with $\mathbf{x}^{(t-1)}$.
  - Substitute σ^2 with $1-\overline{\alpha}_t$.
  - Substitute $\mathbf{f}_{\boldsymbol{\theta}}(\cdot)$ with $\mathbf{s}_{\boldsymbol{\theta}}(\cdot, t)$.

So, by solving the optimization problem

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left(E_{\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}\sim q_t,\ \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\left\|\boldsymbol{\xi} + \sqrt{1-\overline{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)\right\|_2^2\right]\right), \tag{7}$$

we will end up with a network $\mathbf{s}_{\boldsymbol{\theta}^*}$ such that $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}^{(t)}, t)$ would give a good approximation of $\nabla \log q(\mathbf{x}^{(t)})$.

- However, there are still two problems with the optimization problem in (7). The first is that we can simplify it further. Sampling $\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}$ from q_t is the same as sampling $\mathbf{x}^{(0)}$ from q and then multiplying it with $\sqrt{\overline{\alpha}_t}$. As a result, we can apply the law of the unconcious statistician (LOTUS) to rewrite the problem as:

$$\arg\min_{\boldsymbol{\theta}} \left(E_{\mathbf{x}^{(0)}\sim q,\ \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)} \left[\left\| \boldsymbol{\xi} + \sqrt{1-\overline{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t) \right\|_2^2 \right] \right).$$

- The second problem is that the loss function only applies for a single value of t. However, we need our network to work for all $t = 1, 2, \ldots, T$. So, let us add loss terms for all of them to arrive at

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} \left(\sum_{t=1}^{T} E_{\mathbf{x}^{(0)}\sim q,\ \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)} \left[\left\| \boldsymbol{\xi} + \sqrt{1-\overline{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t) \right\|_2^2 \right] \right) \\ &= \arg\min_{\boldsymbol{\theta}} \left(E_{t\sim\{1,2,\ldots,T\},\mathbf{x}^{(0)}\sim q,\ \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)} \left[\left\| \boldsymbol{\xi} + \sqrt{1-\overline{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t) \right\|_2^2 \right] \right) \end{aligned} \tag{8}$$

  Now, if we find the solution θ^* to this optimziation problem, we will have that $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}^{(t)}, t)$ would be a good approximation for $\nabla \log q(\mathbf{x}^{(t)})$ for all $t = 1, 2, \ldots, T$.

- We can see that the optimization problem in (8) is similar to (6). This gives us a relationship between $\mathbf{s}_{\boldsymbol{\theta}}$ and $\xi_{\boldsymbol{\theta}}$ when the parameters are optimal:

$$\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t) \equiv -\sqrt{1-\overline{\alpha}_t}\mathbf{s}_{\theta}(\mathbf{x}^{(t)}, t). \tag{9}$$

4 Denoising Diffusion Implicit Models (DDIMs)

- The **denoising diffusion implicity model** (DDIM) was described by Song et al. in 2020 [SME20].
- The goal of the paper is to speed up the forward process.

4.1 A Non-Markovian Forward Processes

- Recall that the forward process of the DDPM is given by

$$\begin{aligned} \mathbf{x}^{(0)} &\sim q(\mathbf{x}^{(0)}), \\ \mathbf{x}^{(t)} &\sim \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t I). \end{aligned}$$

  In other words, we first sample $\mathbf{x}^{(0)}$, then we generative $\mathbf{x}^{(1)}$, then $\mathbf{x}^{(2)}$, and so on until we get $\mathbf{x}^{(1000)}$. We have that the following equality is true:

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I).$$

- The DDIM paper seeks to create another stochastic process.
  - It is defined by the probability function $q_{\boldsymbol{\sigma}}$, where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_T) \in [0, \infty)^T$.
  - First, sample $\mathbf{x}^{(0)}$ according to $q_{\boldsymbol{\sigma}}(\mathbf{x}^{(0)})$, which is defined to be the data distribution.
  - Then, sample $\mathbf{x}^{(T)}$, the last element, according to

$$q_{\boldsymbol{\sigma}}(\mathbf{x}^{(T)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(T)}; \sqrt{\overline{\alpha}_T}\mathbf{x}^{(0)}, (1-\overline{\alpha}_T)I).$$

  - Given that $\mathbf{x}^{(t)}$ for some $t > 1$ has been sampled, sample $\mathbf{x}^{(t-1)}$ according to

$$q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) = \mathcal{N}\left(\mathbf{x}^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_{t-1}-\sigma_t^2}\frac{\mathbf{x}^{(t)} - \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}}{\sqrt{1-\overline{\alpha}_t}}, \sigma_t^2 I\right)$$

- So, this new process samples $\mathbf{x}^{(0)}$ and then $\mathbf{x}^{(T)}$. Then, it works backward from $\mathbf{x}^{(T)}$ to $\mathbf{x}^{(1)}$.

- It can be shown that the new forward process has the same "marginal" probability as the old one. In other words,

$$q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I).$$

  See Appendix C for a proof.

- The process above is not a "forward" process per se because it goes backward in time.

- However, a forward process can be defined by using Bayes' rule:

$$q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)}) = \frac{q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}{q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}.$$

  It can be shown that $q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)})$ is a Gaussian, but this fact is not used in the paper. This forward process is not a Markov chain because $\mathbf{x}^{(t)}$ depends on both $\mathbf{x}^{(t-1)}$ and $\mathbf{x}^{(0)}$.

- The parameters $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_T)$ controls how stochastic the process is. When $\boldsymbol{\sigma} = \mathbf{0}$, the process becomes deterministic because $\mathbf{x}^{(t-1)}$ is determined deterministicly from $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(0)}$.

  - When the model becomes deterministic, it becomes an **implicit model**, a deterministic procedure for turning noise into a data sample [ML16].
  - This is why the model is called a denoising diffusion "implicit" model (DDIM).

4.2 The Backward Process

- Now, we shall derive a backward process that reverts the above non-Markovian forward process. The probability function of the backward process is denoted by $p_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ denote trainable parameters.

- The task of the backward process is to sample $\mathbf{x}^{(t-1)}$ when given $\mathbf{x}^{(t)}$. In other words, we would like to find the form of $p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$.

- This time, what is different from DDPM is that, we now have $q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})$ that we can leverage. This is done by first making a prediction of $\mathbf{x}^{(0)}$ from $\mathbf{x}^{(t)}$. Then, we can use $q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})$ to sample $\mathbf{x}^{(t-1)}$.

- Here's how we predict $\mathbf{x}^{(0)}$ from $\mathbf{x}^{(t)}$. Recall that $\mathbf{x}^{(t)} \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I)$. By Tweedie's formula,

$$\begin{aligned} E[\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}|\mathbf{x}^{(t)}] &= \mathbf{x}^{(t)} + (1-\overline{\alpha}_t)\nabla \log q^{(t)}_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}) \\ E[\mathbf{x}^{(0)}|\mathbf{x}^{(t)}] &= \frac{\mathbf{x}^{(t)}}{\sqrt{\overline{\alpha}_t}} + \frac{(1-\overline{\alpha}_t)}{\sqrt{\overline{\alpha}_t}}\nabla \log q^{(t)}_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}) \end{aligned}$$

  So, the RHS of the above equation is the best prediction of $\mathbf{x}^{(0)}$ from $\mathbf{x}^{(t)}$.

- There are two ways to compute the RHS above.

  1. The first way is the direct one. We shall train a **score network** $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ that would estimate $\nabla \log q^{(t)}_{\boldsymbol{\sigma}}(\mathbf{x})$. Then, the prediction function would be

$$\mathbf{f}^{(t)}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\overline{\alpha}_t}} + \frac{(1-\overline{\alpha}_t)}{\sqrt{\overline{\alpha}_t}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t).$$

2. The second way uses the relationship between the score network and the noise network (9)

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) \equiv -\frac{\xi_{\boldsymbol{\theta}}(\mathbf{x}, t)}{\sqrt{1-\overline{\alpha}_t}}.$$

So, we train a **noise network** $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ to predict the noise $\boldsymbol{\xi}$. Our prediction would then be

$$\mathbf{f}_{\boldsymbol{\theta}}^{(t)}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\overline{\alpha}_t}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}, t).$$

We shall use the noise network approach.

- With the prediction function defined, we are ready to define the backward process. It is given by

$$p^{(t-1|t)}(\mathbf{x}|\mathbf{x}') = \begin{cases} \mathcal{N}(\mathbf{x}; \mathbf{f}_{\boldsymbol{\theta}}^{(1)}(\mathbf{x}'), \sigma_1^2 I), & t = 1 \\ q_{\boldsymbol{\sigma}}^{(t-1|t,0)}(\mathbf{x}|\mathbf{x}', \mathbf{f}_{\boldsymbol{\theta}}^{(t)}(\mathbf{x}')), & 1 < t \leq T \end{cases}$$

- The above distribution tells us to sample $\mathbf{x}^{(t-1)}$ from $\mathbf{x}^{(t)}$ by evaluating

$$\begin{aligned} \mathbf{x}^{(t-1)} &= \sqrt{\overline{\alpha}_{t-1}} \cdot \mathbf{f}_{\boldsymbol{\theta}}^{(t)}(\mathbf{x}^{(t)}) + \sqrt{1-\overline{\alpha}_{t-1}-\sigma_t^2}\frac{\mathbf{x}^{(t)} - \sqrt{\overline{\alpha}_t} \cdot \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)})}{\sqrt{1-\overline{\alpha}_t}} + \sigma_t \boldsymbol{\xi} \\ &= \sqrt{\overline{\alpha}_{t-1}} \underbrace{\left(\frac{\mathbf{x}^{(t)} - \sqrt{1-\overline{\alpha}_t} \cdot \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)}{\sqrt{\overline{\alpha}_t}}\right)}_{\text{predicted } \mathbf{x}^{(0)}} + \underbrace{\sqrt{1-\overline{\alpha}_{t-1}-\sigma_t^2} \cdot \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)}_{\text{diretion pointing to } \mathbf{x}^{(t)}} + \underbrace{\sigma_t \boldsymbol{\xi}}_{\text{random noise}}. \end{aligned}$$

Here, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)$ as usual.

- Again, we note that, for the sampling process above, different choices of $\boldsymbol{\sigma}$ result in different generative process.
  - When $\boldsymbol{\sigma} = \mathbf{0}$, the generative process becomes deterministic.
  - When $\sigma_t = \sqrt{(1-\overline{\alpha}_{t-1})/(1-\overline{\alpha}_t)}\sqrt{\beta_t}$ for all t, the generative process becomes the same as that of the DDPM.

4.3 Training

- The loss function that we will be minimizing is (3).

$$J_{\boldsymbol{\sigma}}(\boldsymbol{\theta}) = E_{\mathbf{x}^{(0:T)} \sim q_{\boldsymbol{\sigma}}}[\log q_{\boldsymbol{\sigma}}(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})].$$

- Recall from (5) that the (weighted) version of the loss for the vanilla DDPM is given by

$$L_{\mathbf{w}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t E_{\mathbf{x}^{(0)} \sim q, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)}\left[\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)\right\|_2^2\right]$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_T) \in [0, \infty)^T$.

- The paper shows that the two losses are equivalent.

Theorem 5. *For each $\boldsymbol{\sigma} \in [0, \infty)^T$, there exists $\mathbf{w} \in [0, \infty)^T$ and $C \in \mathbb{R}$ such that $J_{\boldsymbol{\sigma}}(\boldsymbol{\theta}) = L_{\mathbf{w}}(\boldsymbol{\theta}) + C$.*

- Consider the loss $L_{\mathbf{w}}(\boldsymbol{\theta})$.

  - If the parameters $\boldsymbol{\theta}$ for $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, t)$ are not shared between different t's, then the parameters do not depend on the weights $\mathbf{w}$.
  - The above assertion justifies the use of $L_{\mathbf{1}}$ as a surrogate for (3).
- Now, consider the loss $J_{\boldsymbol{\sigma}}$.
  - By Theorem 5, $J_{\boldsymbol{\sigma}}$ is equivalent to some $L_{\mathbf{w}}$.
  - So, if parameters are not shared across different t's, it must be the case that the optimal solution to $J_{\boldsymbol{\sigma}}$ must the the same as that of $L_{\mathbf{1}}$.
  - As a result, $L_{\mathbf{1}}$ can be used as surrogate for $J_{\boldsymbol{\sigma}}$ as well.
- The above reasoning tells us to train a DDIM in the same way we train a DDPM in the Ho et al. paper. In other words, a DDPM can be turned into a DDIM by just changing the sampling method.
  - Of course, the model we train in practice will share parameters across all t's. The conclusion we just derived above would not be true, but we still use $L_{\mathbf{1}}$ any way.

4.4 Advantanges of DDIM

- There are two main advantages.
  1. Sampling can be done much faster.
  2. The sampling process can be made deterministic by setting $\boldsymbol{\sigma} = \mathbf{0}$.

4.4.1 Faster Sampling

- Instead of generating a sample using all time steps from the sequence $(T, T-1, \ldots, 1)$, Song et al. experimented with using a subsequence $\boldsymbol{\tau} = (\tau_S, \tau_{S-1}, \ldots, \tau_1)$.
- They found that:
  - The FID score of DDIMs for CIFAR10 and CelebA are worse than those of DDPMs when the full set of steps are used.
  - However, when subsequences of the full sequence of time steps are used, DDIMs performed much better across the board.
  - In their experiments, the full sequence has 1000 time steps. The subsequences have 10, 20, 50, and 100 time steps. FID score deteriorates as fewer time steps are used. However, for subsequences with 50 and 100 steps, it seems that the FID scores did not become much worse.

4.4.2 Determinism

- Because the backward process can be made deterministic, we can treat $\mathbf{x}^{(T)}$ as a latent code of a data sample.
- Song et al. experimented with interpolating between two latent codes. It seems to work just like GAN.
- It is also possible to find a latent code ($\mathbf{x}^{(T)}$) corresponding a data item $\mathbf{x}^{(0)}$. A process to do so will be discussed in the next section.

4.5 Connection to Neural ODE

- With $\boldsymbol{\sigma} = \mathbf{0}$, the update equation becomes

$$\mathbf{x}^{(t-1)} = \sqrt{\overline{\alpha}_{t-1}}\left(\frac{\mathbf{x}^{(t)} - \sqrt{1-\overline{\alpha}_t}\cdot\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)}{\sqrt{\overline{\alpha}_t}}\right) + \sqrt{1-\overline{\alpha}_{t-1}}\cdot\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t).$$

  In other words,

$$\frac{\mathbf{x}^{(t-1)}}{\sqrt{\overline{\alpha}_{t-1}}} - \frac{\mathbf{x}^{(t)}}{\sqrt{\overline{\alpha}_t}} = \left(\frac{\sqrt{1-\overline{\alpha}_{t-1}}}{\sqrt{\overline{\alpha}_{t-1}}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}}\right)\boldsymbol{\xi}(\mathbf{x}^{(t)}, t). \tag{10}$$

- We can turn the above difference equation. First, we think of t as a continuous variable and treat $\mathbf{x}^{(t)}$, $\overline{\alpha}_t$ as functions of t. In other words, we make the following transformations

$$\mathbf{x}^{(t)} \to \mathbf{x}(t),$$
$$\overline{\alpha}_t \to \overline{\alpha}(t).$$

  Second, we define

$$\sigma(t) := \sqrt{\frac{1-\overline{\alpha}(t)}{\overline{\alpha}(t)}} = \sqrt{\frac{1}{\overline{\alpha}(t)} - 1},$$
$$\overline{\mathbf{x}}(t) := \frac{\mathbf{x}(t)}{\sqrt{\overline{\alpha}(t)}} = \sqrt{1+\sigma^2(t)}\cdot\mathbf{x}(t).$$

  Writing $t-1$ in (10) as $t - \Delta t$, we have that

$$\overline{\mathbf{x}}(t-\Delta t) - \overline{\mathbf{x}}(t) = \big(\sigma(t-\Delta t) - \sigma(t)\big)\boldsymbol{\xi}(\mathbf{x}(t), t) = \big(\sigma(t-\Delta t) - \sigma(t)\big)\boldsymbol{\xi}_{\boldsymbol{\theta}}\left(\frac{\overline{\mathbf{x}}(t)}{\sqrt{1+\sigma^2(t)}}, t\right).$$

  Dividing both sides by $-\Delta t$, we have that

$$\frac{\overline{\mathbf{x}}(t) - \overline{\mathbf{x}}(t-\Delta t)}{\Delta t} = \frac{\sigma(t) - \sigma(t-\Delta t)}{\Delta t}\boldsymbol{\xi}_{\boldsymbol{\theta}}\left(\frac{\overline{\mathbf{x}}(t)}{\sqrt{1+\sigma^2(t)}}, t\right).$$

  Taking the limit as $\Delta t \to 0$, we have

$$\frac{\mathrm{d}\overline{\mathbf{x}}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\sigma(t)}{\mathrm{d}t}\boldsymbol{\xi}_{\boldsymbol{\theta}}\left(\frac{\overline{\mathbf{x}}(t)}{\sqrt{1+\sigma^2(t)}}, t\right)$$
$$\mathrm{d}\overline{\mathbf{x}}(t) = \boldsymbol{\xi}_{\boldsymbol{\theta}}\left(\frac{\overline{\mathbf{x}}(t)}{\sqrt{1+\sigma^2(t)}}, t\right)\mathrm{d}\sigma(t). \tag{11}$$

- The above ODE can be used to sample a data item by solving the initial value problem where the initial condition is $\overline{\mathbf{x}}(T) \sim \mathcal{N}(\mathbf{0}, 1+\sigma^2(T))$.
  - Here, note that, because we want $\overline{\alpha}(t)$ to be small, $1+\sigma^2(t)$ would be very large.
- The ODE can also be used to find the corresponding $\mathbf{x}^{(T)}$ for each $\mathbf{x}^{(0)}$ by running it forward:

$$\overline{\mathbf{x}}(t+1) - \overline{\mathbf{x}}(t) = (\sigma(t+1) - \sigma(t))\boldsymbol{\xi}_{\boldsymbol{\theta}}\left(\frac{\overline{\mathbf{x}}(t)}{\sqrt{1+\sigma^2(t)}}, t\right).$$

  In other words,

$$\frac{\mathbf{x}^{(t+1)}}{\sqrt{\overline{\alpha}_{t+1}}} - \frac{\mathbf{x}^{(t)}}{\sqrt{\overline{\alpha}_t}} = \left(\frac{\sqrt{1-\overline{\alpha}_{t+1}}}{\sqrt{\overline{\alpha}_{t+1}}} - \frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}}\right)\boldsymbol{\xi}(\mathbf{x}^{(t)}, t).$$

- The paper also notes that the ODE (11) is equivalent to the "variance-exploding SDE" in the score-based SDE paper by Yang Song et al., which is also published in the same year [SSDK+20]. However, we will not go into the details of the derivation.

A Gaussian Identities

- **Proposition 6.**

$$\mathcal{N}(ax+b;\mu,\sigma^2) = \frac{1}{|a|}\mathcal{N}\left(x;\frac{\mu-b}{a},\frac{\sigma^2}{a^2}\right)$$

Proof.

$$\begin{aligned}
\mathcal{N}(ax+b;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2\sigma^2}(ax+b-\mu)^2\Big)\\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{a^2}{2\sigma^2}\left(x-\frac{\mu-b}{a}\right)^2\right)\\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2(\sigma^2/a^2)}\left(x-\frac{\mu-b}{a}\right)^2\right)\\
&= \frac{1}{|a|}\frac{1}{\sqrt{2\pi}(\sigma/|a|)}\exp\left(-\frac{1}{2(\sigma^2/a^2)}\left(x-\frac{\mu-b}{a}\right)^2\right)\\
&= \frac{1}{|a|}\mathcal{N}\left(x;\frac{\mu-b}{a},\frac{\sigma^2}{a^2}\right)
\end{aligned}$$

as required. □

- **Proposition 7.**

$$\mathcal{N}(x;\mu_1,\sigma_1^2) * \mathcal{N}(x;\mu_2,\sigma_2^2) = \int \mathcal{N}(t;\mu_1,\sigma_1^2)\mathcal{N}(x-t;\mu_1,\sigma_1^2)\,\mathrm{d}t = \mathcal{N}(x;\mu_1+\mu_2,\sigma_1^2+\sigma_2^2).$$

Proof. Let X_1 be a random variable distributed according to (μ_1,σ_1^2), and let X_2 be a random variable distributed according to (μ_2,σ_2^2). Let X_1 and X_2 be indepdendnet of each other. Define $X = X_1 + X_2$. Then, we have that the probability distribution function of X is given by

$$p_X(x) = \mathcal{N}(x;\mu_1,\sigma_1^2) * \mathcal{N}(x;\mu_2,\sigma_2^2).$$

Observe that

$$E[X] = E[X_1] + E[X_2] = \mu_1 + \mu_2.$$

Moreover,

$$\mathrm{Var}(X_1+X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \sigma_1^2 + \sigma_2^2.$$

So, the proposition would be true if we can show that $p_X(x)$ is a Gaussian. In other words, if we can show that there are constants A, B, and C such that

$$p_X(x) = A\exp\big(-B(x-C)^2\big)$$

with $A, B > 0$, then we would be done.

So, let us that that

$$\begin{aligned}
\mathcal{N}(t;\mu_1,\sigma^2) &= A_1\exp\big(-B_1(t-C_1)^2\big),\\
\mathcal{N}(x-t;\mu_2,\sigma^2) &= A_2\exp\big(-B_2(x-t-C_2)^2\big)
\end{aligned}$$

with $A_1, B_1, A_2, B_2 > 0$. It follows that

$$\begin{aligned} p_X(x) &= \int A_1 \exp\big(-B_1(t-C_1)^2\big) A_2 \exp\big(-B_2(x-t-C_2)^2\big)\,\mathrm{d}t \\ &= A_1 A_2 \int \exp\big(-B_1(t-C_1)^2 - B_2(x-t-C_2)^2\big)\,\mathrm{d}t. \end{aligned}$$

Now, we have that

$$\begin{aligned} &-B_1(t-C_1)^2 - B_2(x-t-C_2)^2 \\ &= -B_2x^2 + 2B_2C_2x - (B_1+B_2)t^2 + 2(B_1C_1 + B_2x - B_2C_2)t - B_1C_1^2 - B_2C_2^2. \end{aligned}$$

Now, let's simplify the expression by given the constants new names.

$$-B_1(t-C_1)^2 - B_2(x-t-C_2)^2 = -B_2x^2 + 2D_2x - (B_1+B_2)t^2 + 2(B_2x+D_4)t + D_5.$$

We know that $D_3 = B_1 + B_2 > 0$. Let's complete the square of the polynomial that involves t. We have that

$$\begin{aligned} &-(B_1+B_2)t^2 + 2(B_2x+D_4)t \\ &= -(B_1+B_2)\left(t^2 - 2\frac{B_2x+D_4}{B_1+B_2}t\right) \\ &= -(B_1+B_2)\left(t^2 - 2\frac{B_2x+D_4}{B_1+B_2}t + \frac{(B_2x+D_4)^2}{(B_1+B_2)^2} - \frac{(B_2x+D_4)^2}{(B_1+B_2)^2}\right) \\ &= -(B_1+B_2)\left(t^2 - 2\frac{B_2x+D_4}{B_1+B_2}t + \frac{(B_2x+D_4)^2}{(B_1+B_2)^2}\right) + \frac{(B_2x+D_4)^2}{B_1+B_2} \\ &= -(B_1+B_2)\left(t - \frac{B_2x+D_4}{B_1+B_2}\right)^2 + \frac{(B_2x+D_4)^2}{B_1+B_2} \\ &= -(B_1+B_2)(t - D_6x - D_7)^2 + \frac{B_2^2}{B_1+B_2}x^2 + 2D_8x + D_9. \end{aligned}$$

So,

$$\begin{aligned} &-B_2x^2 + 2D_2x - (B_1+B_2)t^2 + 2(B_2x+D_4)t + D_5 \\ &= -B_2x^2 + 2D_2x - (B_1+B_2)(t - D_6x - D_7)^2 + \frac{B_2^2}{B_1+B_2}x^2 + 2D_8x + D_9 + D_5 \\ &= -\left(B_2 - \frac{B_2^2}{B_1+B_2}\right)x^2 + 2(D_2+D_8)x - (B_1+B_2)(t - D_6x - D_7)^2 + D_{10} \\ &= -\left(\frac{B_1B_2}{B_1+B_2}\right)x^2 + 2(D_2+D_8)x - (B_1+B_2)(t - D_6x - D_7)^2 + D_{10}. \end{aligned}$$

Note that, because $B_1, B_2 > 0$, we have that $B_1B_2/(B_1+B_2) > 0$. Now, we can complete the square of the terms that involve x:

$$-\left(\frac{B_1B_2}{B_1+B_2}\right)x^2 + 2(D_2+D_8)x = -\left(\frac{B_1B_2}{B_1+B_2}\right)(x - D_{11})^2 + D_{12}.$$

This gives

$$\begin{aligned} &-\left(\frac{B_1B_2}{B_1+B_2}\right)x^2 + 2(D_2+D_8)x - (B_1+B_2)(t - D_6x - D_7)^2 + D_{10} \\ &= -\left(\frac{B_1B_2}{B_1+B_2}\right)(x - D_{11})^2 - (B_1+B_2)(t - D_6x - D_7)^2 + D_{10} + D_{12} \\ &= -\left(\frac{B_1B_2}{B_1+B_2}\right)(x - D_{11})^2 - (B_1+B_2)(t - D_6x - D_7)^2 + D_{13}. \end{aligned}$$

In other words,

$$-B_1(t-C_1)^2 - B_2(x-t-C_2)^2 = -D_{14}(x-D_{11})^2 - D_{15}(t-D_6x-D_7)^2 + D_{13}$$

for some $D_{14}, D_{15} > 0$.

Hence,

$$\begin{aligned} p_X(x) &= A_1A_2 \int \exp\big(-D_{14}(x-D_{11})^2 - D_{15}(t-D_6x-D_7)^2 + D_{13}\big)\,\mathrm{d}t \\ &= A_1A_2e^{D_{13}}\bigg(\int \exp(-D_{15}(t-D_6x-D_7)^2)\,\mathrm{d}t\bigg)\exp(-D_{14}(x-D_{11})^2). \end{aligned}$$

The integral on the last line is a positive constant because it is an integral of a Gaussian. So, $p_X(x)$ is a Gaussian. We are done. □

- **Proposition 8.** *Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be positive definite matrices. We have that*

$$D_{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1) \,\|\, \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)) = \frac{1}{2}\bigg(\log\frac{\det\Sigma_2}{\det\Sigma_1} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\Sigma_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - d\bigg).$$

Proof. See other sources. We will not prove this. □

- **Corollary 9.**

$$D_{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \sigma_1^2 I) \,\|\, \mathcal{N}(\boldsymbol{\mu}_2, \sigma_2^2 I)) = \frac{1}{2}\bigg(\frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2}{\sigma_2^2}2d(\log|\sigma_2| - \log|\sigma_1|) + d\frac{\sigma_1^2}{\sigma_2^2} - d\bigg).$$

Proof. Applying Proposition 8, we have that

$$\begin{aligned} &D_{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \sigma_1^2 I) \,\|\, \mathcal{N}(\boldsymbol{\mu}_2, \sigma_2^2 I)) \\ &= \frac{1}{2}\bigg(\log\frac{\det(\sigma_2^2 I)}{\det(\sigma_1^2 I)} + \mathrm{tr}((\sigma_2^2 I)^{-1}\sigma_1^2 I) + (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T(\sigma_2^2 I)^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - d\bigg) \\ &= \frac{1}{2}\bigg(\log\frac{\sigma_2^{2d}}{\sigma_1^{2d}} + \mathrm{tr}\Big(\frac{\sigma_1^2}{\sigma_2^2}I\Big) + \frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2}{\sigma_2^2} - d\bigg) \\ &= \frac{1}{2}\bigg(2d(\log\sigma_2 - \log\sigma_1) + d\frac{\sigma_1^2}{\sigma_2^2} + \frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2}{\sigma_2^2} - d\bigg) \\ &= \frac{1}{2}\bigg(\frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2}{\sigma_2^2}2d(\log|\sigma_2| - \log|\sigma_1|) + d\frac{\sigma_1^2}{\sigma_2^2} - d\bigg) \end{aligned}$$

as required. □

B Proofs for the DDPM paper (Ho et al. 2020)

- **Claim 10.**

$$p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}) = E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\bigg[p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\bigg].$$

Proof.

$$\begin{aligned}
p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}) &= \int p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\,\mathrm{d}\mathbf{x}^{(1:T)} \\
&= \int p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\frac{q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\,\mathrm{d}\mathbf{x}^{(1:T)} \\
&= \int q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})}{q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\,\mathrm{d}\mathbf{x}^{(1:T)} \\
&= \int q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\frac{\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{\prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\,\mathrm{d}\mathbf{x}^{(1:T)} \\
&= \int q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\,\mathrm{d}\mathbf{x}^{(1:T)} \\
&= E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\left[p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\right]
\end{aligned}$$

as required. □

- **Claim 11.**

$$E_{\mathbf{x}^{(0)}\sim q}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})] \leq E_{\mathbf{x}^{(0:T)}\sim q}\Big[\log q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\Big].$$

Proof.

$$\begin{aligned}
E_{\mathbf{x}^{(0)}\sim q}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)})] &= E_{\mathbf{x}^{(0)}\sim q}\left[-\log\left(E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\left[p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\right]\right)\right] \\
&\leq E_{\mathbf{x}^{(0)}\sim q}\left[-E_{\mathbf{x}^{(1:T)}\sim q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)})}\left[\log\left(p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\prod_{t=1}^{T}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\right)\right]\right] \\
&= -E_{\mathbf{x}^{(0:T)}\sim q}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) + \sum_{t=1}^{T}\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) - \sum_{t=1}^{T}\log q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})\right] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\left[\sum_{t=1}^{T}\log q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \sum_{t=1}^{T}\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\right] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[\log q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\Big]
\end{aligned}$$

as required. □

- **Claim 12.** *Let L be the RHS of* (3)*. Then, it can be rewritten as*

$$\begin{aligned}
L = &\underbrace{E_{\mathbf{x}^{(0)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\,\big\|\,p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\Big]}_{L_T} \\
&+ \underbrace{E_{\mathbf{x}^{(0,1)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\Big]}_{L_0} \\
&+ \sum_{t=2}^{T}\underbrace{E_{\mathbf{x}^{(0,t)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big)\Big]}_{L_{t-1}}.
\end{aligned}$$

Proof. We can rewrite L as follows.

$$\begin{aligned}
L &= E_{\mathbf{x}^{(0:T)}\sim q}\Big[\log q(\mathbf{x}^{(1:T)}|\mathbf{x}^{(0)}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0:T)})\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \sum_{t=1}^{T}\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} - \sum_{t=2}^{T}\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\Big]
\end{aligned}$$

Applying Bayes' rule, we have that

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)},\mathbf{x}^{(0)}) = \frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}.$$

However, because q is a Markov chain, we have that $\mathbf{x}^{(t)}$ depends only on $\mathbf{x}^{(t-1)}$ and not on $\mathbf{x}^{(0)}$. So, $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)},\mathbf{x}^{(0)}) = q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$, and we have that

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}.$$

So,

$$\begin{aligned}
L &= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} - \sum_{t=2}^{T}\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} - \sum_{t=2}^{T}\log\left(\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})}\frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}\right)\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)}) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} - \sum_{t=2}^{T}\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})} - \log\frac{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})} - \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) - \sum_{t=2}^{T}\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})}\Big] \\
&= E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\Big] + E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\Big] + \sum_{t=2}^{T}E_{\mathbf{x}^{(0:T)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})}\Big] \\
&= E_{\mathbf{x}^{(0,T)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\Big] + E_{\mathbf{x}^{(0,1)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\Big] + \sum_{t=2}^{T}E_{\mathbf{x}^{(0,t-1,t)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})}\Big].
\end{aligned}$$

The last line comes from using Claim 13 to eliminate unused indices in the sampling process. We will

now simplify each term, one by one. For the first term from the left, we have that

$$\begin{aligned}
E_{\mathbf{x}^{(0,T)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\Big] &= E_{\mathbf{x}^{(0,T)}\sim q}\Big[\log\frac{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}\Big] \\
&= \iint q(\mathbf{x}^{(0)},\mathbf{x}^{(T)})\log\frac{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}\,\mathrm{d}\mathbf{x}^{(T)}\mathrm{d}\mathbf{x}^{(0)} \\
&= \iint q(\mathbf{x}^{(0)})q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\log\frac{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}\,\mathrm{d}\mathbf{x}^{(T)}\mathrm{d}\mathbf{x}^{(0)} \\
&= \int q(\mathbf{x}^{(0)})\bigg(\int q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\log\frac{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})}\,\mathrm{d}\mathbf{x}^{(T)}\bigg)\mathrm{d}\mathbf{x}^{(0)} \\
&= \int q(\mathbf{x}^{(0)})D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\,\big\|\,p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\,\mathrm{d}\mathbf{x}^{(0)} \\
&= E_{\mathbf{x}^{(0)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\,\big\|\,p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\Big].
\end{aligned}$$

We cannot simplify the second term further. However, for the last term, we have that

$$\begin{aligned}
&E_{\mathbf{x}^{(0,t-1,t)}\sim q}\Big[-\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})}\Big] \\
&= \int q(\mathbf{x}^{(0)},\mathbf{x}^{(t-1)},\mathbf{x}^{(t)})\log\frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}\,\mathrm{d}\mathbf{x}^{(0,t-1,t)} \\
&= \iint q(\mathbf{x}^{(0)},\mathbf{x}^{(t)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\log\frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}\,\mathrm{d}\mathbf{x}^{(t-1)}\mathrm{d}\mathbf{x}^{(0,t)} \\
&= \int q(\mathbf{x}^{(0)},\mathbf{x}^{(t)})\bigg(\int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\log\frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}\,\mathrm{d}\mathbf{x}^{(t-1)}\bigg)\mathrm{d}\mathbf{x}^{(0,t)} \\
&= \int q(\mathbf{x}^{(0)},\mathbf{x}^{(t)})D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big)\,\mathrm{d}\mathbf{x}^{(0,t)} \\
&= E_{\mathbf{x}^{(0,t)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big)\Big]
\end{aligned}$$

As a result, we can write L as a sum of three terms

$$\begin{aligned}
L = &\underbrace{E_{\mathbf{x}^{(0)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})\,\big\|\,p_{\boldsymbol{\theta}}(\mathbf{x}^{(T)})\big)\Big]}_{L_T} \\
&+\underbrace{E_{\mathbf{x}^{(0,1)}\sim q}\Big[-\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\Big]}_{L_0} \\
&+\sum_{t=2}^{T}\underbrace{E_{\mathbf{x}^{(0,t)}\sim q}\Big[D_{KL}\big(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)},\mathbf{x}^{(t)})\big\|p_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\big)\Big]}_{L_{t-1}}
\end{aligned}$$

as required. □

- **Claim 13.** *Let f be a function of $\mathbf{x}^{(t_1)}$, $\mathbf{x}^{(t_2)}$, …, and $\mathbf{x}^{(t_k)}$. Then,*

$$E_{\mathbf{x}^{(0:T)}\sim q}\big[f(\mathbf{x}^{(t_1)},\mathbf{x}^{(t_2)},\ldots,\mathbf{x}^{(t_k)})\big] = E_{\mathbf{x}^{(t_1,t_2,\ldots,t_k)}\sim q}\big[f(\mathbf{x}^{(t_1)},\mathbf{x}^{(t_2)},\ldots,\mathbf{x}^{(t_k)})\big].$$

In other words, if $\mathtt{T} = \{t_1, t_2, \ldots, t_k\}$, then

$$E_{\mathbf{x}^{(0:T)}\sim q}\big[f(\mathbf{x}^{(\mathtt{T})})\big] = E_{\mathbf{x}^{(\mathtt{T})}\sim q}\big[f(\mathbf{x}^{(\mathtt{T})})\big].$$

Proof. We have that

$$\begin{aligned}
&E_{\mathbf{x}^{(0:T)}\sim q}\big[f(\mathbf{x}^{(\mathtt{T})})\big] \\
&= \int q(\mathbf{x}^{(0:T)}) f(\mathbf{x}^{(\mathtt{T})})\,\mathrm{d}\mathbf{x}^{(0:T)} \\
&= \int\int q(\mathbf{x}^{(\mathtt{T})}) q(\mathbf{x}^{(\{0:T\}-\mathtt{T})}|\mathbf{x}^{(\mathtt{T})}) f(\mathbf{x}^{(\mathtt{T})})\,\mathrm{d}\mathbf{x}^{(\{0:T\}-\mathtt{T})}\mathrm{d}\mathbf{x}^{(\mathtt{T})} \\
&= \int \bigg(\int q(\mathbf{x}^{(\{0:T\}-\mathtt{T})}|\mathbf{x}^{(\mathtt{T})})\,\mathrm{d}\mathbf{x}^{(\{0:T\}-\mathtt{T})}\bigg) q(\mathbf{x}^{(\mathtt{T})}) f(\mathbf{x}^{(\mathtt{T})})\,\mathrm{d}\mathbf{x}^{(\mathtt{T})} \\
&= \int q(\mathbf{x}^{(\mathtt{T})}) f(\mathbf{x}^{(\mathtt{T})})\,\mathrm{d}\mathbf{x}^{(\mathtt{T})} \\
&= E_{\mathbf{x}^{(\mathtt{T})}\sim q}\big[f(\mathbf{x}^{(\mathtt{T})})\big].
\end{aligned}$$

as required. □

- **Claim 14.** *It holds that*

$$L_{t-1} - C = E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\beta_t}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}\right) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right\|^2\right]$$

where C is the value defined in (4)*.*

Proof. First, we have that

$$\begin{aligned}
L_{t-1} &= E_{\mathbf{x}^{(0,t)}\sim q}\left[\frac{\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\|^2}{2\beta_t}\right] + C \\
L_{t-1} - C &= E_{\mathbf{x}^{(0,t)}\sim q}\left[\frac{1}{2\beta_t}\left\|\frac{\beta_t\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\mathbf{x}^{(0)} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}^{(t)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right\|^2\right].
\end{aligned}$$

From Proposition 1, we know that $\mathbf{x}^{(t)} = \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}$ where $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}, I)$. As a result,

$$\mathbf{x}^{(0)} = \frac{\mathbf{x}^{(t)} - \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}}{\sqrt{\overline{\alpha}_t}}.$$

Using the above fact, we can show that

$$\frac{\beta_t\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\mathbf{x}^{(0)} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{x}^{(t)} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}\right).$$

(The derivation is straightforward but tedious.) So, we may rewrite $L_{t-1} - C$ as

$$L_{t-1} - C = E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\beta_t}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}\right) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)}, t)\right\|^2\right]$$

as required. □

- **Claim 15.**

$$L_{t-1} - C = E_{\mathbf{x}^{(0)}\sim q, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}, t)\right\|^2\right].$$

Proof.

$$\begin{aligned}
L_{t-1} - C &= E_{\mathbf{x}^{(0)}\sim q,\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\beta_t}\left\|\frac{1}{\sqrt{\alpha_t}}\frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi} - \frac{1}{\sqrt{\alpha_t}}\frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)},t)\right\|^2\right] \\
&= E_{\mathbf{x}^{(0)}\sim q,\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{1}{2\beta_t}\frac{\beta_t^2}{\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)},t)\right\|^2\right] \\
&= E_{\mathbf{x}^{(0)}\sim q,\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(t)},t)\right\|^2\right] \\
&= E_{\mathbf{x}^{(0)}\sim q,\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\left[\frac{\beta_t}{2\alpha_t(1-\overline{\alpha}_t)}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi},t)\right\|^2\right]
\end{aligned}$$

as required. □

- **Claim 16.** *It holds that*

$$\frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right\|^2 = \frac{1}{2\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right\|^2$$

where $\boldsymbol{\xi}$ is a random variable distributed according to $\mathcal{N}(\mathbf{0},I)$ such that $\mathbf{x}^{(1)} = \sqrt{\overline{\alpha}_1}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\xi}$.

Proof.

$$\begin{aligned}
\frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right\|^2 &= \frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \frac{1}{\sqrt{\alpha_1}}\left(\mathbf{x}^{(1)} - \frac{\beta_1}{\sqrt{1-\overline{\alpha}_1}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right)\right\|^2 \\
&= \frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \frac{1}{\sqrt{\alpha_1}}\left(\sqrt{\overline{\alpha}_1}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_1}\boldsymbol{\xi} - \frac{\beta_1}{\sqrt{1-\overline{\alpha}_1}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right)\right\|^2 \\
&= \frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \frac{1}{\sqrt{\alpha_1}}\left(\sqrt{\alpha_1}\mathbf{x}^{(0)} + \sqrt{\beta_1}\boldsymbol{\xi} - \frac{\beta_1}{\sqrt{\beta_1}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right)\right\|^2 \\
&= \frac{1}{2\beta_1}\left\|\mathbf{x}^{(0)} - \mathbf{x}^{(0)} + \frac{1}{\sqrt{\alpha}_1}\left(\sqrt{\beta_1}\boldsymbol{\xi} - \sqrt{\beta_1}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right)\right\|^2 \\
&= \frac{1}{2\beta_1}\frac{\beta_1}{\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right\|^2 \\
&= \frac{1}{2\alpha_1}\left\|\boldsymbol{\xi} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}^{(1)},1)\right\|^2
\end{aligned}$$

as required. □

C Proofs for the DDIM Paper (Song et al. 2020)

- **Proposition 17.** *For all $1 \le t \le T$, we have that*

$$q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I).$$

Proof. We will prove the proposition by induction from T down to 1. The base case is $t = T$, which is already true by definition.

Suppose by way of induction that the proposition is true for some $t \le T$. We have that

$$\begin{aligned}
q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}) &= \int q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)})q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})\,\mathrm{d}\mathbf{x}^{(t)} \\
&= \int \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, (1-\overline{\alpha}_t)I) \\
&\quad \mathcal{N}\left(\mathbf{x}^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}\mathbf{x}^{(0)} + \sqrt{1-\overline{\alpha}_{t-1}-\sigma_t^2}\frac{\mathbf{x}^{(t)} - \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}}{\sqrt{1-\overline{\alpha}_t}}, \sigma_t^2 I\right)\mathrm{d}\mathbf{x}^{(t)}.
\end{aligned}$$

Now, the above equation is quite handful to write, so let us introduce some shorthands. Let

$$\gamma_t = 1 - \overline{\alpha}_t$$
$$\delta_t = 1 - \overline{\alpha}_{t-1} - \sigma_t^2.$$

With them, the equation becomes,

$$\begin{aligned} q_{\boldsymbol{\sigma}}(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}) &= \int \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, \gamma_t I)\mathcal{N}\left(\mathbf{x}^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}\mathbf{x}^{(0)} + \sqrt{\delta_t}\frac{\mathbf{x}^{(t)} - \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}}{\sqrt{\gamma_t}}, \sigma_t^2 I\right) \mathrm{d}\mathbf{x}^{(t)} \\ &= \int \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\overline{\alpha}_t}\mathbf{x}^{(0)}, \gamma_t I)\mathcal{N}\left(\mathbf{x}^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}\mathbf{x}^{(0)} + \frac{\sqrt{\delta_t}\mathbf{x}^{(t)} - \sqrt{\delta_t\overline{\alpha}_t}\mathbf{x}^{(0)}}{\sqrt{\gamma_t}}, \sigma_t^2 I\right) \mathrm{d}\mathbf{x}^{(t)}. \end{aligned}$$

Because the dimensions of the random variables are independent, it suffices to prove the proposition for the 1D case. In this case, we have that

$$q_{\boldsymbol{\sigma}}(x^{(t-1)}|x^{(0)}) = \int \mathcal{N}(x^{(t)}; \sqrt{\overline{\alpha}_t}x^{(0)}, \gamma_t)\mathcal{N}\left(x^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}x^{(0)} + \frac{\sqrt{\delta_t}x^{(t)} - \sqrt{\delta_t\overline{\alpha}_t}x^{(0)}}{\sqrt{\gamma_t}}, \sigma_t^2\right) \mathrm{d}x^{(t)}.$$

Note that,

$$\begin{aligned} &\mathcal{N}\left(x^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}x^{(0)} + \frac{\sqrt{\delta_t}x^{(t)} - \sqrt{\delta_t\overline{\alpha}_t}x^{(0)}}{\sqrt{\gamma_t}}, \sigma_t^2\right) \\ &= \mathcal{N}\left(x^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}}x^{(0)} + \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)} - \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}x^{(0)}, \sigma_t^2\right) \\ &= \mathcal{N}\left(x^{(t-1)} - \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \left(\sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}\right)x^{(0)}, \sigma_t^2\right). \end{aligned}$$

Now, by applying Proposition 6,

$$\begin{aligned} \mathcal{N}(x^{(t)}; \sqrt{\overline{\alpha}_t}x^{(0)}, \gamma_t) &= \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathcal{N}\left(\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\sqrt{\overline{\alpha}_t}x^{(0)}, \frac{\delta_t}{\gamma_t}\gamma_t\right) \\ &= \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathcal{N}\left(\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\sqrt{\overline{\alpha}_t}x^{(0)}, \delta_t\right) \\ &= \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathcal{N}\left(\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}x^{(0)}, \delta_t\right) \end{aligned}$$

So,

$$\begin{aligned} &q_{\boldsymbol{\sigma}}(x^{(t-1)}|x^{(0)}) \\ &= \int \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathcal{N}\left(\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}x^{(0)}, \delta_t\right) \mathrm{d}x^{(t)}\mathcal{N}\left(x^{(t-1)} - \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \left(\sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}\right)x^{(0)}, \sigma_t^2\right) \mathrm{d}x^{(t)} \\ &= \int \mathcal{N}\left(\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}x^{(0)}, \delta_t\right) \mathrm{d}x^{(t)}\mathcal{N}\left(x^{(t-1)} - \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}; \left(\sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{\delta_t\overline{\alpha}_t}}{\sqrt{\gamma_t}}\right)x^{(0)}, \sigma_t^2\right) \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathrm{d}x^{(t)} \end{aligned}$$

Let

$$u = \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}.$$

It follows that

$$\mathrm{d}u = \frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}\mathrm{d}x^{(t)}.$$

Making the substitution from $\frac{\sqrt{\delta_t}}{\sqrt{\gamma_t}}x^{(t)}$ to u, we have that

$$\begin{aligned}
&q_{\boldsymbol{\sigma}}(x^{(t-1)}|x^{(0)}) \\
&= \int \mathcal{N}\left(u; \frac{\sqrt{\delta_t \overline{\alpha}_t}}{\sqrt{\gamma_t}} x^{(0)}, \delta_t\right) \mathrm{d}x^{(t)} \mathcal{N}\left(x^{(t-1)} - u; \left(\sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{\delta_t \overline{\alpha}_t}}{\sqrt{\gamma_t}}\right) x^{(0)}, \sigma_t^2\right) \mathrm{d}u.
\end{aligned}$$

Applying Proposition 7, we have that

$$\begin{aligned}
q_{\boldsymbol{\sigma}}(x^{(t-1)}|x^{(0)}) &= \mathcal{N}\left(x^{(t-1)}; \frac{\sqrt{\delta_t \overline{\alpha}_t}}{\sqrt{\gamma_t}} x^{(0)} + \left(\sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{\delta_t \overline{\alpha}_t}}{\sqrt{\gamma_t}}\right) x^{(0)}, \delta_t + \sigma_t^2\right) \\
&= \mathcal{N}\left(x^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}} x^{(0)}, 1 - \overline{\alpha}_{t-1} - \sigma_t^2 + \sigma_t^2\right) \\
&= \mathcal{N}\left(x^{(t-1)}; \sqrt{\overline{\alpha}_{t-1}} x^{(0)}, 1 - \overline{\alpha}_{t-1}\right).
\end{aligned}$$

As a result, the proposition is true for $t-1$ as well. By induction, we are done. □

References

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[ML16] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models, 2016.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020.

[SWMG15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, jul 2011.