Variational Diffusion Models

Pramook Khungurn

November 12, 2022

This note is written as I read the paper "Variational Diffusion Models" by Kingma et al.. [KSPH21].

1 Introduction

- The paper lists two contributions.
- First, it proposes a new family of diffusion-based generative models.
  - It incorporate Fourier features.
  - It can joinly optimize the noise schedule together with the rest of the model.
  - It can be easily casted to continuous time settings.
- Second, it contributes new theoretical understanding of diffusion-based generative models.
  - Derive a simple expression of the variational lower bound (VLB) in terms of the signal-to-noise ratio.
  - Prove a new invariance in the continuous time setting.
  - Show that various diffusion models in literature are equivalent up to a trivial time-dependent rescaling of the data.
- The end result is that the authors got a model that achieved SOTA log likelihood at the time.
  - However, FID score was not the best when compared to other models, so the method might not lead to the best looking images.
  - Their model is also large, deep, and kind of impossible to train if you don't have enough resource.

2 Model

- A data item is represented by $\mathbf{x} \in \mathbb{R}^d$.
- The data distribution is denoted by $p(\mathbf{x})$, which we want to model.

2.1 Forward Time Diffusion Process

- We start with a data item $\mathbf{x}$ sampled according to $p(\mathbf{x})$.
- We define a sequence increasingly noisy versions of $\mathbf{x}$, which we call the **latent variables $\mathbf{z}_t$**.
  - Here, t runs from $t = 0$ (least noisy) to $t = 1$ (most noisy).

- The distribution of the latent variable $\mathbf{z}_t$ conditioned on $\mathbf{x}$ is given by
$$q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \alpha_t \mathbf{x}, \sigma_t^2 I) \tag{1}$$
where α_t and σ_t^2 are strictly positive scalar-valued functions of t.
- We also assume that α_t and σ_t are smooth.
  - In other words, they have continuous first derivatives with respect to t, and the derivatives are finite.
- Define the **signal-to-noise radio (SNR)** to be
$$\mathrm{SNR}(t) = \alpha_t^2/\sigma_t^2.$$
- The SNR should be monotonically decreating in time.
  - In other words, $t > s$ implies $\mathrm{SNR}(t) < \mathrm{SNR}(s)$.
  - This formalizes the notion that, as t increases, the latent variable should become noisier.
- In the original DDPM paper [HJA20], we have that $\alpha_t = \sqrt{1 - \sigma_t^2}$.
  - So, $\alpha_t^2 + \sigma_t^2 = 1$ for all t.
  - As a result, we call such a model **variance preserving**.
- In the paper by Song et al. on the SDE formulation of score-based models [SSDK$^+$20], we have a model where $\alpha_t = 1$ for all t.
  - As $t \to 1$, σ_t^2 must increase in order for the SNR to decrease.
  - This means that $\alpha_t^2 + \sigma_t^2 = 1 + \sigma_t^2$, which increase as t increase.
  - As a result, we call such a model **variance exploding**.
  - In fact, the SDE for such a model is calle the variance-exploding SDE (VE-SDE).
- We also require that the forward time process also satisfies the following properties.
  1. For any $0 \leq s < t \leq 1$, we have that
  $$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\mathbf{z}_t; \alpha_{t|s}\mathbf{z}_s; \sigma_{t|s}^2 I) \tag{2}$$
  where $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2$.
  2. The joint distribution $(\mathbf{z}_s, \mathbf{z}_t, \mathbf{z}^u)$ for any $0 \leq s < t < u \leq 1$ is Markov. In other words,
  $$q(\mathbf{z}_u|\mathbf{z}_t, \mathbf{z}_s) = q(\mathbf{z}_u|\mathbf{z}_t).$$
- We want the model to be consistent. In other words, it should be the case that
  - (1) should be consistent with (2), and
  - (2) should be consistent with itself.

  This is indeed the case, and the proofs can be found in Appendix B.
- It can be shown that, for any $0 \leq s < t \leq 1$, we have that
$$q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) = \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_Q(\mathbf{z}_t, \mathbf{x}; s, t), \sigma_Q^2(s,t) I)$$
where
$$\sigma_Q^2(s,t) = \sigma_{t|s}^2 \sigma_s^2 / \sigma_t^2,$$
$$\boldsymbol{\mu}_Q(\mathbf{z}_t, \mathbf{x}; s, t) = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s \sigma_{t|s}^2}{\sigma_t^2}\mathbf{x}.$$
See a proof also in Appendix B.

2.2 Noise Schedule

- In works such as [HJA20], the nosie schedule has a fixed form.
- The paper proposes learning the noise schedule through the parameterization

$$\sigma_t^2 = \mathrm{sigmoid}(\gamma_{\boldsymbol{\eta}}(t)) = \frac{1}{1+\exp(-\gamma_{\boldsymbol{\eta}}(t))}$$

  where $\gamma_{\boldsymbol{\eta}}(t)$ is a monotonic neural network with parameter $\boldsymbol{\eta}$.
- The specification of $\gamma_{\boldsymbol{\eta}}(t)$.
  - It has 3 linear layers with weights that are restricted to be positive.
  - Let us call the layers l_1, l_2, and l_3. Then,

$$\gamma_{\boldsymbol{\eta}}(t) := l_1(t) + l_3(\phi(l_2(l_1(t))))$$

    where ϕ is the sigmoid function.
  - l_2 has 1024 outputs while other layers have only a single output.
- The paper fixes $\alpha_t = \sqrt{1-\sigma_t^2}$, subscribing to the variance-perserving camp.
  - However, we will show later that variance-preserving models and variance-exploding models are equivalent.
- We now have that

$$\begin{aligned}
\alpha_t^2 &= 1-\sigma_t^2 = 1 - \frac{1}{1+\exp(-\gamma_{\boldsymbol{\eta}}(t))} = \frac{\exp(-\gamma_{\boldsymbol{\eta}}(t))}{1+\exp(-\gamma_{\boldsymbol{\eta}}(t))} = \frac{1}{1+\exp(\gamma_{\boldsymbol{\eta}}(t))} \\
&= \mathrm{sigmoid}(-\gamma_{\boldsymbol{\eta}}(t)), \\
\mathrm{SNR}(t) &= \frac{\alpha_t^2}{\sigma_t^2} = \frac{\exp(-\gamma_{\boldsymbol{\eta}}(t))}{1+\exp(-\gamma_{\boldsymbol{\eta}}(t))}(1+\exp(-\gamma_{\boldsymbol{\eta}}(t))) \\
&= \exp(-\gamma_{\boldsymbol{\eta}}(t)).
\end{aligned}$$

2.3 Reverse Time Generative Process

- The generative model is defined by inverting the forward time process.
- It samples a sequence of latent variables $\mathbf{z}_t$ with time running backward from $t=1$ to $t=0$.
- The model can be defined in the discrete time and continuous time setting. We will discuss the discrete time setting first.
- Definitions for the discrete time settings.
  - Let T be a positive integer.
  - We split the time interval $[0,1]$ into T segments, each with width $\tau = 1/T$.
  - Define $s(i) = (i-1)/T$ and $t(i) = i/T$.
  - The generative model for data item $\mathbf{x}$ is given by:

$$p(\mathbf{x}) = \int_z p(\mathbf{z}_1)p(\mathbf{x}|\mathbf{z}_0)\prod_{i=1}^{T} p(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)})\,\mathrm{d}\mathbf{z}.$$

    Here, $\mathbf{z}$ denotes $(\mathbf{z}_0, \mathbf{z}_{1/T}, \mathbf{z}_{2/T}, \dots, \mathbf{z}_1)$.

- With the variance preserving setting and sufficiently small SNR(1), we have that $q(\mathbf{z}_1|\mathbf{x}) \approx \mathcal{N}(\mathbf{z}_1; \mathbf{0}, I)$. So, we can model the marginal distribution of $\mathbf{z}_1$ with $\mathcal{N}(\mathbf{0}, I)$. In other words,

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; \mathbf{0}, I).$$

- For $p(\mathbf{x}|\mathbf{z}_0)$, the paper factors the terms into independent components. Let the ith component of $\mathbf{x}$ and $\mathbf{z}_0$ be denoted by x_i and $z_{0,i}$, respectively. We set

$$p(\mathbf{x}|\mathbf{z}_0) = \prod_{i=1}^{d} p(x_i|z_{0,i})$$

and

$$p(x_i|z_{0,i}) = \frac{q(z_{0,i}|x_i)}{\sum_{x=0}^{255} q(z_{0,i}|x)}$$

taking into account that each x_i is an 8-bit pixel value. The last equation is just applying Bayes' rule assuming that each pixel value is equally likely.

- Lastly, we choose

$$p(\mathbf{z}_s|\mathbf{z}_t) = q(\mathbf{z}_t|\mathbf{z}_t, \mathbf{x} = \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)).$$

This is the same as $q(\mathbf{z}_s; \mathbf{z}_t, \mathbf{x})$ we discussed in the last section but the sampled data $\mathbf{x}$ is replaced by a **denoising model** $\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$ that predicts $\mathbf{x}$ from $\mathbf{z}_t$.

- To be more concrete, we can also rewrite $p(\mathbf{z}_s|\mathbf{z}_t)$ as

$$p(\mathbf{z}_s|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t), \sigma_Q^2(s,t)I)$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \boldsymbol{\mu}_Q(\mathbf{z}_t, \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t); s, t) = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s \sigma_{t|s}^2}{\sigma_t^2}\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t).$$

- The mean of the backward step $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t)$ can also be written as

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{1}{\alpha_{t|s}}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) = \frac{1}{\alpha_{t|s}} + \frac{\sigma_{t|s}^2}{\alpha_{t|s}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)$$

where

$$\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) = \frac{\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\sigma_t}$$

is the **noise prediction model** that predicts that Gaussian noise $\boldsymbol{\xi} \sim \mathcal{N}(0, I)$ that is used to make $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\xi}$, and

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) = -\frac{\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\sigma_t}$$

is the **score model** that predicts the score $\nabla \log q(\mathbf{z}_t)$ from $\mathbf{z}_t$.

- Moreover, we can simplify $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t)$ and $\sigma_Q^2(s,t)$ further:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{\mathbf{z}_t + \sigma_t \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\alpha_{t|s}}$$
$$\sigma_Q^2(s,t) = -\sigma_s^2 \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))$$

where $\mathrm{expm1}(u) = e^u - 1$.[1] See the proof at Proposition 10 in Appendix B.

[1] In numerical software packages such as NumPy, Torch, and JAX, expm1 is available as a function because the straightforward computation is not very accurate and numerically stable.

2.4 Noise Prediction Model

- Following Ho et al. [HJA20], the paper trains the noise prediction model $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\cdot;\cdot)$.
- The relationship between $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\cdot;\cdot)$ and $\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\cdot;\cdot)$ is as follows:

$$\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t) = \frac{\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)}{\alpha_t}.$$

- The paper uses an architecture similar to that of Ho et al.with a number of modifications. The modification that they want to highlight the most is the use of Fourier features [TSM+20].
  - The paper optimizes the network for likelihood, which is sensitive to the exact pixel values. So, it needs to capture all the fine details in the data.
  - To do so, the authors propose adding a set of Fourier features to the input of the noise prediction model.
    * Let $\mathbf{x}$ be the original data, scaled to the range $[-1,1]$, and let $\mathbf{z}_t$ be a latent code.
    * They concatenate to $\mathbf{z}_t$ channels $\sin(2^n \pi \mathbf{z}_t)$ and $\cos(2^n \pi \mathbf{z}_t)$ where n runs over a range of integers from n_{min} to n_{max}, and then they feed the concatenated tensor to the noise prediction model.
  - Including the features led to large improvements in log-likelihood, especially when combined with learned noise schedule. In particular, it allows the network to learn with much higher value of $\mathrm{SNR}_{\mathrm{max}}$ (i.e., much lower value for σ_0^2) than without.
  - The authors got the best results with $n_{\mathrm{min}} = 7$ and $n_{\mathrm{max}} = 8$.
    * This is quite surprising because it's just only 4 more channels.
    * The author says lower frequencies can be learned from $\mathbf{z}$ itself, and high frequencies are simply not present or irrelevant for likelihood.
- Other modifications include:
  - The paper's network does not perform nay downsampling or upsampling. The tensors remain at the original input resolution.
  - The network is deeper than ones used by Ho et al. in [HJA20].
    * For the CIFAR10 and the 32×32 ImageNet datasets, the authors uses U-Nets with depth of 32 in the downsampling and upsampling (which are not actually performed).
    * For the 64×64 ImageNet dataset, they double the depth!
  - Intead of taking time t as input to the noise prediction model, they feed a scaled version of $\gamma_{\boldsymbol{\eta}}(t)$ as input to the network. The scaling is done in such a way that the value is in the range $[0,1]$.
  - Apart from the attention block that connects the upward and downard branches of the U-Net in [HJA20], the authors remove all attention blocks from the model.
  - The model use dropout of rate 0.1.
  - The authors optimized the model with the AdamW algorithm [LH17]. The settings are as follows.
    * Learning rate of 2×10^{-4}.
    * $\beta_1 = 0.9, \beta_2 = 0.99$.
    * Weight decay coefficient of 0.01.
  - The model weights are accumulated with exponential moving average with decay rate of 0.9999.

2.5 Variational Lower Bound

- We train the model by trying to minimize the variational lower bound of the log likelihood. This is given by

$$-\log p(\mathbf{x}) \leq \mathrm{VLB}(\mathbf{x}) = \underbrace{D_{KL}(q(\mathbf{x}_1|x)\|p(\mathbf{z_1}))}_{\text{prior loss}} + \underbrace{E_{\mathbf{z}_0\sim q(\mathbf{z}_0|x)}[-\log p(\mathbf{x}|\mathbf{z}_0)]}_{\text{reconstruction loss}} + \underbrace{\mathcal{L}_T(\mathbf{x})}_{\text{diffusion loss}}.$$

  You can find how to derive the above expression in another note of mine [Khu22].

- The prior loss and the reconstruction loss can be estimated using standard techniques.
- The diffusion loss depends on the number of time steps T, and we will discuss it in the next sections.

3 Discrete-Time Model

- In case of finite T, the diffusion loss is

$$\mathcal{L}_T(\mathbf{x}) = \sum_{i=1}^{T} E_{\mathbf{z}_{t(i)}\sim q(\mathbf{z}_{t(i)}|\mathbf{x})}\big[D_{KL}(q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)},\mathbf{x})\|p(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}))\big].$$

- We can simplify the diffusion loss to

$$\mathcal{L}_T(\mathbf{x}) = \frac{T}{2} E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I), i\sim\mathcal{U}\{1:T\}}\Big[\big(\mathrm{SNR}(s(i)) - \mathrm{SNR}(t(i))\big)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_{t(i)}; t(i))\|^2\Big].$$

  where $\mathbf{z}_{t(i)} = \alpha_{t(i)}\mathbf{x} + \sigma_{t(i)}\boldsymbol{\xi}$. See the proof in Proposition 11 of Appendix B.

- Another expression is for the loss is given by

$$\mathcal{L}_T(\mathbf{x}) = \frac{T}{2} E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I), i\sim\mathcal{U}\{1:T\}}\Big[\mathrm{expm1}\big(\gamma_{\boldsymbol{\eta}}(t(i)) - \gamma_{\boldsymbol{\eta}}(s(i))\big)\|\xi - \hat{\xi}_{\boldsymbol{\theta}}(\mathbf{z}_{t(i)}; t(i))\|^2\Big].$$

  See Proposition 12 in Appendix B for the proof.

- Note that the rewritten loss contains explicit dependencies on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. So, if we optimize it, we optimize both the noise prediction model and the noise schedule.
  - This is different from the simplified loss in [HJA20], which can only be used to optimize the noise prediction model.
  - It is also much simpler than the loss in Nichol and Dhariwal [ND21], which treats the loss for the noise prediction model and the loss for the noise schedule differently.
- The paper also observes that more timesteps are always better in terms of minimizing the loss value.
  - Imagine you graph $\mathrm{SNR}(t)$ versus $\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2$ where $\mathrm{SNR}(t)$ goes from 0 (much noise) to 1 (no noise).
  - You would have that, when $\hat{\mathbf{x}}_{\boldsymbol{\theta}}$ is good enough, the graph would be descreasing as you go from $\mathrm{SNR}(t) = 0$ to $\mathrm{SNR}(t) = 1$. This is simply because it is easier to denoise an image when there is less noise in the image.
  - Now, we can interpret $\mathrm{SNR}(s) - \mathrm{SNR}(t)$ as the with of an interval, and $\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2$ as the height of the graph at the beginning of the interval in the graph above.
  - So, the discrete time diffusion loss is an upper Riemann sum approximation of an integral of a strictly decreasing function.
  - This implies that more time steps leads to a more accurate upper bound, which is lower.
  - See Figure 2 in the paper for an illustation.

4 Continuous-Time Model

- We now take $T \to \infty$. The limit of $\mathcal{L}_T(\mathbf{x})$ is given by

$$\begin{aligned}
\mathcal{L}_\infty(\mathbf{x}) &= -\frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\bigg[\int_0^1 \mathrm{SNR}'(t)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2\,\mathrm{d}t\bigg] \\
&= -\frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)}\big[\mathrm{SNR}'(t)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2\big] \\
&= \frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)}\big[\gamma'_{\boldsymbol{\eta}}(t)\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2\big]
\end{aligned}$$

  where $\mathrm{SNR}'(t) = \mathrm{dSNR}(t)/\mathrm{d}t$ and $\gamma'_{\boldsymbol{\eta}}(t) = \mathrm{d}\gamma_{\boldsymbol{\theta}}(t)/\mathrm{d}t$. Note that the last equality is not trivial, and its proof can be found in Appendix B.

- The signal-to-noise function $\mathrm{SNR}(t)$ is invertible because it is monotonically decreasing. So, we can perform a change of variable with $v = \mathrm{SNR}(t)$. This gives

$$\begin{aligned}
\mathcal{L}_\infty(\mathbf{x}) &= -\frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\int_0^1 \mathrm{SNR}'(t)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2\,\mathrm{d}t \\
&= -\frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\bigg[\int_{\mathrm{SNR}_{\max}}^{\mathrm{SNR}_{\min}} \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_{\mathrm{SNR}^{-1}(v)};\mathrm{SNR}^{-1}(v))\|^2\,\mathrm{d}v\bigg] \\
&= \frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\bigg[\int_{\mathrm{SNR}_{\min}}^{\mathrm{SNR}_{\max}} \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_{\mathrm{SNR}^{-1}(v)};\mathrm{SNR}^{-1}(v))\|^2\,\mathrm{d}v\bigg]
\end{aligned}$$

  where $\mathrm{SNR}_{\min} = \mathrm{SNR}(1)$ and $\mathrm{SNR}_{\max} = \mathrm{SNR}(0)$.

- The above equation shows us that the only effect the function $\alpha(t)$ and $\sigma(t)$ have on the diffusion loss is the value of $\mathrm{SNR}(t)$ at endpoints $t = 0$ and $t = 1$. The loss value is invarient to the shape of the function $\mathrm{SNR}(t)$ between $t = 0$ and $t = 1$.

- Moreover, the distribution $p(\mathbf{x})$ defined by the generative model is also invariant to the specification of the diffusion model.
  - Let $p^A(\mathbf{x})$ denote the distribution defined by the combination of α_t^A, $\hat{\sigma}_t^A$, and $\mathbf{x}_{\boldsymbol{\theta}}^A$.
    Let $p^B(\mathbf{x})$ be defined similarly for α_t^B, σ_t^B, and $\hat{\mathbf{x}}_{\boldsymbol{\theta}}^B$.
    We require that both distributions have the same values of $\mathrm{SNR}_{\min}$ and $\mathrm{SNR}_{\max}$.
  - Then, we can show that $p^A(\mathbf{x}) = p^B(\mathbf{x})$ if $\hat{\mathbf{x}}_{\boldsymbol{\theta}}^A(\mathbf{z}_t,t) = \hat{\mathbf{x}}_{\boldsymbol{\theta}}^A((\alpha_t^A/\alpha_t^B)\mathbf{z}_t,t)$. Moreover, the distribution of all latents $\mathbf{z}_t$ is the same up to scaling.
  - Hence, all models that satisfies the following mild conditions are equivalent (up to scaling).
    * α_t and σ_t are positive scalar value functions.
    * $\mathrm{SNR}(t) = \alpha_t^2/\sigma_t^2$ is monotonically decreasing in t.
    * $q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t;\alpha_t\mathbf{x},\sigma_t^2 I)$.
    * For all $0 \le s < t \le 1$, it is true that $q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\mathbf{z}_t;\alpha_{t|s}\mathbf{z}_s,\sigma_{t|s}^2 I)$.
    * The forward process is Markov. That is, for any $0 \le s < t < u \le 1$, it follows that $q(\mathbf{z}_u|\mathbf{z}_t,\mathbf{z}_s) = q(\mathbf{z}_u|\mathbf{z}_t)$.
    * $\mathrm{SNR}(0)$ and $\mathrm{SNR}(1)$ are fixed constants that agree with other models.
  - This means that the models based on the variance-exploding SDE and variance-preserving SDE in [SSDK+20] are equivalent in continuous time up to time-dependent scaling factors.

- The equivalence between diffusion models continues to hold even if the loss is weighted and of the form:

$$\mathcal{L}_\infty(\mathbf{x},w) = \frac{1}{2}E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I)}\bigg[\int_{\mathrm{SNR}_{\min}}^{\mathrm{SNR}_{\max}} w(v)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_{\mathrm{SNR}^{-1}(v)};\mathrm{SNR}^{-1}(v))\|^2\,\mathrm{d}v\bigg]$$

- Optimization of $\mathcal{L}_\infty$ requires a lot of care. The paper has the details on how to compute the gradient of the loss in its appendix.
  - We will not cover it now because I've become tired of reading.

5 Summary

- The paper gives a new formulation of the DDPM that deals with the noise schedule in a systematic way.
  - It yields a loss function that can be used to optimize both the noise prediction model and the noise schedule in one go.
  - It also shows that diffusion models that can be formulated in the paper's framework are equivalent up to scaling if the $\mathrm{SNR}_{\min}$ and $\mathrm{SNR}_{\max}$ match.
- While the theoretical component of the paper is certainly valuable, I double whether the proposed new model architecture and losses are practical.
  - The paper's model is very deep and hard to train.
  - The loss is still quite complicated and require a lot of care, especially in the continuous-time setting.
  - In the end, the architecture and the loss are designed to get better likelihood, not image quality as measured by FID scores.

A Gaussian Identities

- Many of these identities come from a lecture note by Marc Toussaint [Tou11].
- A multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ, denoted by $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is the distribution:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(\det 2\pi\Sigma)^{1/2}} \exp\bigg(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \bigg).$$

  It is defined only if the covariance matrix is positive definite.
- **Proposition 1.** *For any invertible matrix A and any vector $\mathbf{b}$, we have that*

$$\mathcal{N}(A\mathbf{x} + \mathbf{b}; \boldsymbol{\mu}, \Sigma) = \frac{1}{|\det A|} \mathcal{N}(\mathbf{x}, A^{-1}(\boldsymbol{\mu} - \mathbf{b}), A^{-1}\Sigma A^{-T}).$$

Proof.

$$\begin{aligned}
\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma) &= \frac{1}{(\det 2\pi\Sigma)^{1/2}} \exp\bigg(-\frac{1}{2}(A\mathbf{x}+\mathbf{b}-\boldsymbol{\mu})^T\Sigma^{-1}(A\mathbf{x}+\mathbf{b}-\boldsymbol{\mu})\bigg) \\
&= \frac{1}{(\det 2\pi\Sigma)^{1/2}} \exp\bigg(-\frac{1}{2}(A\mathbf{x}+\mathbf{b}-\boldsymbol{\mu})^T A^{-T}A^T\Sigma^{-1}AA^{-1}(A\mathbf{x}+\mathbf{b}-\boldsymbol{\mu})\bigg) \\
&= \frac{1}{(\det 2\pi\Sigma)^{1/2}} \exp\bigg(-\frac{1}{2}(\mathbf{x}-A^{-1}(\boldsymbol{\mu}-\mathbf{b}))^T (A^{-1}\Sigma A^{-T})^{-1}(\mathbf{x}+A^{-1}(\boldsymbol{\mu}-\mathbf{b}))\bigg) \\
&= \frac{1}{(\det AA^T)^{1/2}}\frac{1}{(\det A^{-1}A^{-T})^{1/2}(\det 2\pi\Sigma)^{1/2}} \\
&\qquad \exp\bigg(-\frac{1}{2}(\mathbf{x}-A^{-1}(\boldsymbol{\mu}-\mathbf{b}))^T (A^{-1}\Sigma A^{-T})^{-1}(\mathbf{x}+A^{-1}(\boldsymbol{\mu}-\mathbf{b}))\bigg) \\
&= \frac{1}{|\det A|}\frac{1}{(\det 2\pi A^{-1}\Sigma A^{-T})^{1/2}} \\
&\qquad \exp\bigg(-\frac{1}{2}(\mathbf{x}-A^{-1}(\boldsymbol{\mu}-\mathbf{b}))^T (A^{-1}\Sigma A^{-T})^{-1}(\mathbf{x}+A^{-1}(\boldsymbol{\mu}-\mathbf{b}))\bigg) \\
&= \frac{1}{|\det A|}\mathcal{N}(\mathbf{x}, A^{-1}(\boldsymbol{\mu}-\mathbf{b}), A^{-1}\Sigma A^{-T})
\end{aligned}$$

as required. □

- **Corollary 2.** *if $a \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^d$ is a vector, then*

$$\mathcal{N}(a\mathbf{x}+\mathbf{b};\boldsymbol{\mu},\Sigma) = \frac{1}{|a|^d}\mathcal{N}\bigg(\mathbf{x};\frac{\boldsymbol{\mu}-\mathbf{b}}{a},\frac{\Sigma}{a^2}\bigg).$$

- **Proposition 3.**

$$\mathcal{N}(\mathbf{x};\mu_1,\Sigma_1)\mathcal{N}(\mathbf{x};\mu_2,\Sigma_2) = \mathcal{N}(\mu_1;\mu_2,\Sigma_1+\Sigma_2)\mathcal{N}(\mathbf{x};\mu_3,\Sigma_3)$$

where

$$\begin{aligned}
\mu_3 &= \Sigma_2(\Sigma_1+\Sigma_2)^{-1}\mu_1 + \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\mu_2, \\
\Sigma_3 &= \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\Sigma_2.
\end{aligned}$$

We will not prove this proposition. It looks painful.

- **Proposition 4.** *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 5.**

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

Proof. Applying Proposition 4, we have that

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

as required. □

B Proofs of Model Properties

- **Proposition 6.** *The property in Equation* (1) *is consistent with the property in Equation* (2)*. In other words, for any* $0 \leq s < t \leq 1$*, it holds that*

$$q(\mathbf{z}_t|\mathbf{x}) = \int q(\mathbf{z}_t|\mathbf{z}_s)q(\mathbf{z}_s|\mathbf{x})\,\mathrm{d}\mathbf{z}_s.$$

Proof.

$$\begin{aligned}
&\int q(\mathbf{z}_t|\mathbf{z}_s)q(\mathbf{z}_s|\mathbf{x})\,\mathrm{d}\mathbf{z}_s \\
&= \int \mathcal{N}(\mathbf{z}_t; \alpha_{t|s}\mathbf{z}_s, \sigma_{t|s}^2 I)\mathcal{N}(\mathbf{z}_s; \alpha_s\mathbf{x}, \sigma_s^2 I)\,\mathrm{d}\mathbf{z}_s \\
&= \int \mathcal{N}(\alpha_{t|s}\mathbf{z}_s; \mathbf{z}_t, \sigma_{t|s}^2 I)\mathcal{N}(\mathbf{z}_s; \alpha_s\mathbf{x}, \sigma_s^2 I)\,\mathrm{d}\mathbf{z}_s \\
&= \int \frac{1}{\alpha_{t|s}^d}\mathcal{N}\left(\mathbf{z}_s; \frac{\mathbf{z}_t}{\alpha_{t|s}}, \frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} I\right)\mathcal{N}(\mathbf{z}_s; \alpha_s\mathbf{x}, \sigma_s^2 I)\,\mathrm{d}\mathbf{z}_s && \text{(Corollary 2)} \\
&= \int \frac{1}{\alpha_{t|s}^d}\mathcal{N}\left(\frac{\mathbf{z}_t}{\alpha_{t|s}}; \alpha_s\mathbf{x}, \left(\frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} + \sigma_s^2\right) I\right)\mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_3, \Sigma_3)\,\mathrm{d}\mathbf{z}_s && \text{(Proposition 3)} \\
&= \frac{1}{\alpha_{t|s}^d}\mathcal{N}\left(\frac{\mathbf{z}_t}{\alpha_{t|s}}; \alpha_s\mathbf{x}, \left(\frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} + \sigma_s^2\right) I\right)\int \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_3, \Sigma_3)\,\mathrm{d}\mathbf{z}_s \\
&= \frac{1}{\alpha_{t|s}^d}\mathcal{N}\left(\frac{\mathbf{z}_t}{\alpha_{t|s}}; \alpha_s\mathbf{x}, \left(\frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} + \sigma_s^2\right) I\right) \\
&= \mathcal{N}\left(\mathbf{z}_t; \alpha_{t|s}\alpha_s\mathbf{x}, (\sigma_{t|s}^2 + \alpha_{t|s}^2\sigma_s^2) I\right) && \text{(Corollary 2)} \\
&= \mathcal{N}\left(\mathbf{z}_t; \frac{\alpha_t}{\alpha_s}\alpha_s\mathbf{x}, (\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2 + \alpha_{t|s}^2\sigma_s^2) I\right) \\
&= \mathcal{N}\left(\mathbf{z}_t; \alpha_t\mathbf{x}, \sigma_t^2 I\right) \\
&= q(\mathbf{z}_t|\mathbf{x})
\end{aligned}$$

as required. □

- **Proposition 7.** *The property in Equation 2 is consistent with itself. In other words, for any $0 \leq s < t < u \leq 1$, it holds that*

$$q(\mathbf{z}_u|\mathbf{z}_s) = \int q(\mathbf{z}_u|\mathbf{z}_t)q(\mathbf{z}_t|\mathbf{z}_s)\,\mathrm{d}\mathbf{z}_t.$$

Proof.

$$\begin{aligned}
&\int q(\mathbf{z}_u|\mathbf{z}_t)q(\mathbf{z}_t|\mathbf{z}_s)\,\mathrm{d}\mathbf{z}_t \\
&= \int q(\mathbf{z}_u;\alpha_{u|t}\mathbf{z}_t,\sigma^2_{u|t}I)q(\mathbf{z}_t;\alpha_{t|s}\mathbf{z}_s,\sigma^2_{t|s}I)\,\mathrm{d}\mathbf{z}_t \\
&= \int q(\alpha_{u|t}\mathbf{z}_t;\mathbf{z}_u,\sigma^2_{u|t}I)q(\mathbf{z}_t;\alpha_{t|s}\mathbf{z}_s,\sigma^2_{t|s}I)\,\mathrm{d}\mathbf{z}_t \\
&= \int \frac{1}{\alpha^2_{u|t}} q\bigg(\mathbf{z}_t;\frac{\mathbf{z}_u}{\alpha_{u|t}},\frac{\sigma^2_{u|t}}{\alpha^2_{u|t}}I\bigg)q(\mathbf{z}_t;\alpha_{t|s}\mathbf{z}_s,\sigma^2_{t|s}I)\,\mathrm{d}\mathbf{z}_t && \text{(Corollary 2)} \\
&= \frac{1}{\alpha^2_{u|t}} q\bigg(\frac{\mathbf{z}_u}{\alpha_{u|t}};\alpha_{t|s}\mathbf{z}_s,\Big(\frac{\sigma^2_{u|t}}{\alpha^2_{u|t}}+\sigma^2_{t|s}\Big)I\bigg) && \text{(same reasoning as Proposition 6)} \\
&= q\big(\mathbf{z}_u;\alpha_{u|t}\alpha_{t|s}\mathbf{z}_s,(\sigma^2_{u|t}+\alpha^2_{u|t}\sigma^2_{t|s})I\big) \\
&= q\big(\mathbf{z}_u;\alpha_{u|s}\mathbf{z}_s,\big((\sigma^2_u-\alpha^2_{u|t}\sigma^2_t)+\alpha^2_{u|t}(\sigma^2_t-\alpha^2_{t|s}\sigma^2_s)\big)I\big) \\
&= q\big(\mathbf{z}_u;\alpha_{u|s}\mathbf{z}_s,(\sigma^2_u-\alpha^2_{u|t}\sigma^2_t+\alpha^2_{u|t}\sigma^2_t-\alpha^2_{u|t}\alpha^2_{t|s}\sigma^2_s)I\big) \\
&= q\big(\mathbf{z}_u;\alpha_{u|s}\mathbf{z}_s,(\sigma^2_u-\alpha^2_{u|s}\sigma^2_s)I\big) \\
&= q\big(\mathbf{z}_u;\alpha_{u|s}\mathbf{z}_s,\sigma^2_{u|s}I\big) \\
&= q(\mathbf{z}_u|\mathbf{z}_s)
\end{aligned}$$

as required. ☐

- **Proposition 8.** *For any $0 \leq s < t \leq 1$, we have that*

$$q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x}) = \mathcal{N}(\mathbf{z}_s;\boldsymbol{\mu}_Q(\mathbf{z}_t,\mathbf{x};s,t),\sigma^2_Q(s,t)I)$$

where

$$\sigma^2_Q(s,t) = \sigma^2_{t|s}\sigma^2_s/\sigma^2_t,$$
$$\boldsymbol{\mu}_Q(\mathbf{z}_t,\mathbf{x};s,t) = \frac{\alpha_{t|s}\sigma^2_s}{\sigma^2_t}\mathbf{z}_t + \frac{\alpha_s\sigma^2_{t|s}}{\sigma^2_t}\mathbf{x}.$$

Proof. By Baye's rule,

$$q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x}) = \frac{q(\mathbf{z}_t|\mathbf{z}_s,\mathbf{x})q(\mathbf{z}_s|\mathbf{x})}{q(\mathbf{z}_t|\mathbf{x})} = \frac{q(\mathbf{z}_t|\mathbf{z}_s)q(\mathbf{z}_s|\mathbf{x})}{q(\mathbf{z}_t|\mathbf{x})}.$$

The last equality follows from the fact that we require q to be Markov: $q(\mathbf{z}_t|\mathbf{z}_s,\mathbf{x}) = q(\mathbf{z}_t|\mathbf{z}_s)$. Now, we apply Proposition 3 to get

$$\begin{aligned}
q(\mathbf{z}_t|\mathbf{z}_s)q(\mathbf{z}_s|\mathbf{x}) &= \mathcal{N}(\mathbf{z}_t;\alpha_{t|s}\mathbf{z}_s,\sigma^2_{t|s}I)\mathcal{N}(\mathbf{z}_s;\alpha_s\mathbf{x},\sigma^2_sI) \\
&= \mathcal{N}(\alpha_{t|s}\mathbf{z}_s;\mathbf{z}_t,\sigma^2_{t|s}I)\mathcal{N}(\mathbf{z}_s;\alpha_s\mathbf{x},\sigma^2_sI) \\
&= \frac{1}{\alpha^d_{t|s}}\mathcal{N}\bigg(\mathbf{z}_s;\frac{\mathbf{z}_t}{\alpha_{t|s}},\frac{\sigma^2_{t|s}}{\alpha^2_{t|s}}I\bigg)\mathcal{N}(\mathbf{z}_s;\alpha_s\mathbf{x},\sigma^2_sI) \\
&= q(\mathbf{z}_t|\mathbf{x})\mathcal{N}(\mathbf{z}_s;\boldsymbol{\mu}_3,\Sigma_3).
\end{aligned}$$

where $\boldsymbol{\mu}_3$ and Σ_3 are as described in the statement of Proposition 3. The $q(\mathbf{z}_t|\mathbf{x})$ comes from the reasoning we used in the proof of Proposition 6. So, it turns out that

$$q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) = \frac{q(\mathbf{z}_t|\mathbf{x})\mathcal{N}(\mathbf{x}_s; \boldsymbol{\mu}_3, \Sigma_3)}{q(\mathbf{z}_t|\mathbf{x})} = \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_3, \Sigma_3).$$

So, what is left for us to do is to compute $\boldsymbol{\mu}_3$ and Σ_3 and see if the results agree with $\boldsymbol{\mu}_Q(\mathbf{z}_t, \mathbf{x}; s, t)$ and $\sigma_Q^2(s,t)I$.

We have that $\boldsymbol{\mu}_1 = \mathbf{z}_t/\alpha_{t|s}$, $\Sigma_1 = (\sigma_{t|s}^2/\alpha_{t|s}^2)I$, $\boldsymbol{\mu}_2 = \alpha_s\mathbf{x}$, and $\Sigma_2 = \sigma_s^2 I$, so

$$\begin{aligned}
\boldsymbol{\mu_3} &= \Sigma_2(\Sigma_1+\Sigma_2)^{-1}\boldsymbol{\mu}_1 + \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\boldsymbol{\mu}_2 \\
&= \frac{\sigma_s^2}{\sigma_{t|s}^2/\alpha_{t|s}^2 + \sigma_s^2}\frac{\mathbf{z}_t}{\alpha_{t|s}} + \frac{\sigma_{t|s}^2/\alpha_{t|s}^2}{\sigma_{t|s}^2/\alpha_{t|s}^2 + \sigma_s^2}\alpha_s\mathbf{x} \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_{t|s}^2/\alpha_{t|s} + \alpha_{t|s}\sigma_s^2}\frac{\mathbf{z}_t}{\alpha_{t|s}} + \frac{\sigma_{t|s}^2}{\sigma_{t|s}^2 + \alpha_{t|s}^2\sigma_s^2}\alpha_s\mathbf{x} \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_{t|s}^2 + \alpha_{t|s}^2\sigma_s^2}\mathbf{z}_t + \frac{\sigma_{t|s}^2}{\alpha_s\sigma_{t|s}^2 + \alpha_{t|s}^2\sigma_s^2}\mathbf{x} \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2 + \alpha_{t|s}^2\sigma_s^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2 + \alpha_{t|s}^2\sigma_s^2}\mathbf{x} \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\mathbf{x} \\
&= \boldsymbol{\mu}_Q(\mathbf{z}_t, \mathbf{x}; s, t).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\Sigma_3 = \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\Sigma_2 &= \frac{\sigma_{t|s}^2}{\alpha_{t|s}^2}\left(\frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} + \sigma_s^2\right)^{-1}\sigma_s^2 I = \frac{\sigma_{t|s}^2\sigma_s^2}{\alpha_{t|s}^2(\sigma_{t|s}^2/\alpha_{t|s}^2 + \sigma_s^2)}I = \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_{t|s}^2 + \alpha_{t|s}^2\sigma_s^2}I = \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}I \\
&= \sigma_Q^2(s,t)I
\end{aligned}$$

as required. $\square$

- **Proposition 9.**

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{1}{\alpha_{t|s}}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)$$

Proof. We have that

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) &= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\hat{\mathbf{x}}(\mathbf{z}_t; t) \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\left(\frac{\mathbf{z}_t - \sigma_t\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\alpha_t}\right) \\
&= \left(\frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} + \frac{\alpha_s\sigma_{t|s}^2}{\alpha_t\sigma_t^2}\right)\mathbf{z}_t - \frac{\alpha_s\sigma_t\sigma_{t|s}^2}{\alpha_t\sigma_t^2}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\sigma_t^2}\left(\frac{\alpha_t\sigma_s^2}{\alpha_s} + \frac{\alpha_s\sigma_{t|s}^2}{\alpha_t}\right)\mathbf{z}_t - \frac{\alpha_s\sigma_{t|s}^2}{\alpha_t\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\sigma_t^2}\left(\frac{\alpha_t^2\sigma_s^2 + \alpha_s^2\sigma_{t|s}^2}{\alpha_s\alpha_t}\right)\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\sigma_t^2}\left(\frac{\alpha_t^2\sigma_s^2 + \alpha_s^2(\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2)}{\alpha_s\alpha_t}\right)\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\sigma_t^2}\left(\frac{\alpha_t^2\sigma_s^2 + \alpha_s^2\sigma_t^2 - \alpha_t^2\sigma_s^2}{\alpha_s\alpha_t}\right)\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\sigma_t^2}\frac{\alpha_s^2\sigma_t^2}{\alpha_s\alpha_t}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{\alpha_s}{\alpha_t}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \\
&= \frac{1}{\alpha_{t|s}}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)
\end{aligned}$$

as required. □

- **Proposition 10.**

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{\mathbf{z}_t + \sigma_t \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\alpha_{t|s}}$$
$$\sigma_Q^2(s, t) = -\sigma_s^2 \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))$$

where $\mathrm{expm1}(u) = e^u - 1$.

Proof. First, we have that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{1}{\alpha_{t|s}}\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) = \frac{1}{\alpha_{t|s}}\left(\mathbf{z}_t - \frac{\sigma_{t|s}^2}{\sigma_t}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right).$$

Now,

$$\begin{aligned}
\frac{\sigma_{t|s}^2}{\sigma_t} &= \frac{\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2}{\sigma_t} = \sigma_t - \frac{\alpha_t^2\sigma_s^2}{\alpha_s^2\sigma_t} = \sigma_t\left(1 - \frac{\alpha_t^2\sigma_s^2}{\alpha_s^2\sigma_t^2}\right) = \sigma_t\left(1 - \frac{(1-\sigma_t^2)\sigma_s^2}{(1-\sigma_s^2)\sigma_t^2}\right) = \sigma_t\left(1 - \frac{(1-\sigma_t^2)\sigma_s^2}{(1-\sigma_s^2)\sigma_t^2}\right) \\
&= \sigma_t\left(1 - \frac{\sigma_t^{-2} - 1}{\sigma_s^{-2} - 1}\right) = \sigma_t\left(1 - \frac{1 + \exp(-\gamma_{\boldsymbol{\eta}}(t)) - 1}{1 + \exp(-\gamma_{\boldsymbol{\eta}}(s)) - 1}\right) = \sigma_t\big(1 - \exp(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\big) \\
&= -\sigma_t \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t)).
\end{aligned}$$

So,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{\mathbf{z}_t + \sigma_t \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)}{\alpha_{t|s}},$$

and we are done with $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t)$.

For $\sigma_Q^2(s,t)$, we have that

$$\begin{aligned}\sigma_Q^2(s,t) &= \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2} = \frac{\sigma_s^2}{\sigma_t}\frac{\sigma_{t|s}^2}{\sigma_t} = \frac{\sigma_s^2}{\sigma_t}\big(-\sigma_t \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\big)\\ &= -\sigma_s^2 \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t))\end{aligned}$$

as required. □

- **Proposition 11.** *In the case of finite T, the diffusion loss $\mathcal{L}_T(\mathbf{x})$ can be expressed as*

$$\mathcal{L}_T(\mathbf{x}) = \frac{T}{2} E_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},I), i\sim\mathcal{U}\{1:T\}}\Big[\big(\mathrm{SNR}(s(i)) - \mathrm{SNR}(t(i))\big)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_{t(i)}; t(i))\|^2\Big].$$

Proof. Let us use s and t as shorthands for $s(i)$ and $t(i)$. We have that

$$\begin{aligned}q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x}) &= \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_Q(\mathbf{z}_t,\mathbf{x}; s,t), \sigma_Q^2(s,t)),\\ p(\mathbf{z}_s|\mathbf{z}_t) &= \mathcal{N}(\mathbf{z}_s, \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s,t), \sigma_Q^2(s,t)),\\ \boldsymbol{\mu}_Q(\mathbf{z}_t,\mathbf{x}; s,t) &= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\mathbf{x}\\ \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s,t) &= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; s,t),\\ \sigma_Q^2 &= \sigma_{t|s}^2\sigma_s^2/\sigma_t^2.\end{aligned}$$

Applying Proposition 5, we have that

$$D_{KL}(q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x})\|p(\mathbf{z}_s|\mathbf{z}_t)) = \frac{1}{2\sigma_Q^2(s,t)}\|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_{\boldsymbol{\theta}}\|^2 = \frac{\sigma_t^2}{2\sigma_{t|s}^2\sigma_s^2}\frac{\alpha_s^2\sigma_{t|s}^4}{\sigma_t^4}\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2.$$

Now,

$$\begin{aligned}\frac{\sigma_t^2}{2\sigma_{t|s}^2\sigma_s^2}\frac{\alpha_s^2\sigma_{t|s}^4}{\sigma_t^4} &= \frac{1}{2\sigma_s^2}\frac{\alpha_s^2\sigma_{t|s}^2}{\sigma_t^2} = \frac{1}{2\sigma_s^2}\frac{\alpha_s^2(\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2)}{\sigma_t^2} = \frac{1}{2}\frac{\alpha_s^2\sigma_t^2 - \alpha_t^2\sigma_s^2}{\sigma_s^2\sigma_t^2} = \frac{1}{2}\left(\frac{\alpha_s^2}{\sigma_s^2} - \frac{\alpha_t^2}{\sigma_t^2}\right)\\ &= \frac{1}{2}\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big).\end{aligned}$$

As a result,

$$D_{KL}(q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x})\|p(\mathbf{z}_s|\mathbf{z}_t)) = \frac{1}{2}\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big)\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2,$$

and

$$\begin{aligned}
\mathcal{L}_T(\mathbf{x}) &= \sum_{i=1}^{T} E_{\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})}[D_{KL}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \| p(\mathbf{z}_s|\mathbf{z}_t))] \\
&= \frac{1}{2} \sum_{i=1}^{T} E_{\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})} \Big[\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big) \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2 \Big] \\
&= \frac{1}{2} \sum_{i=1}^{T} E_{\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I)} \Big[\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big) \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2 \Big] \\
&= \frac{T}{2} \sum_{i=1}^{T} E_{\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I), i \sim \mathcal{U}\{1:T\}} \Big[\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big) \|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|^2 \Big]
\end{aligned}$$

as required. □

- **Proposition 12.** *In the case of finite T, the diffusion loss $\mathcal{L}_T(\mathbf{x})$ can be expressed as*

$$\mathcal{L}_T(\mathbf{x}) = \frac{T}{2} E_{\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I), i \sim \mathcal{U}\{1:T\}} \Big[\mathrm{expm1}\big(\gamma_{\boldsymbol{\eta}}(t(i)) - \gamma_{\boldsymbol{\eta}}(s(i))\big) \|\xi - \hat{\xi}_{\boldsymbol{\theta}}(\mathbf{z}_{t(i)}; t(i))\|^2 \Big].$$

Proof. The reasoning of this proposition is similar to the last one. We start with

$$D_{KL}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \| p(\mathbf{z}_s|\mathbf{z}_t)) = \frac{1}{2\sigma_Q^2(s,t)} \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_{\boldsymbol{\theta}}\|^2 = \frac{\sigma_t^2}{2\sigma_{t|s}^2 \sigma_s^2} \frac{\sigma_{t|s}^4}{\alpha_{t|s}^2 \sigma_t^2} \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|.$$

We have that

$$\frac{\sigma_t^2}{2\sigma_{t|s}^2 \sigma_s^2} \frac{\sigma_{t|s}^4}{\alpha_{t|s}^2 \sigma_t^2} = \frac{1}{2\sigma_s^2} \frac{\sigma_{t|s}^2}{\alpha_{t|s}^2} = \frac{\alpha_s^2(\sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2)}{2\sigma_s^2 \alpha_t^2} = \frac{\alpha_s^2 \sigma_t^2 - \alpha_t^2 \sigma_s^2}{2\alpha_t^2 \sigma_s^2} = \frac{1}{2}\bigg(\frac{\alpha_s^2 \sigma_t^2}{\alpha_t^2 \sigma_s^2} - 1 \bigg).$$

In the proof of Proposition 10, we showed that

$$\frac{\alpha_t^2 \sigma_s^2}{\alpha_s^2 \sigma_t^2} = \exp(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t)).$$

As a result,

$$\frac{\alpha_s^2 \sigma_t^2}{\alpha_t^2 \sigma_s^2} = \exp(\gamma_{\boldsymbol{\eta}}(t) - \gamma_{\boldsymbol{\eta}}(s)).$$

Hence,

$$\frac{\sigma_t^2}{2\sigma_{t|s}^2 \sigma_s^2} \frac{\sigma_{t|s}^4}{\alpha_{t|s}^2 \sigma_t^2} = \frac{1}{2}\big(\exp(\gamma_{\boldsymbol{\eta}}(t) - \gamma_{\boldsymbol{\eta}}(s)) - 1 \big) = \mathrm{expm1}(\gamma_{\boldsymbol{\eta}}(t) - \gamma_{\boldsymbol{\eta}}(s)).$$

We are done. □

- **Proposition 13.**

$$\mathcal{L}_\infty(\mathbf{x}) = \frac{1}{2} E_{\xi \sim \mathcal{N}(\mathbf{0}, I), t \sim \mathcal{U}(0,1)} \big[\gamma_\eta'(t) \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}(\mathbf{z}_t; t)\|^2 \big]$$

Proof. First, we have that

$$\mathcal{L}_\infty(\mathbf{x}) = -\frac{1}{2} E_{\xi\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)} \big[\mathrm{SNR}'(t) \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z}_t;t)\|^2 \big].$$

Because $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\xi}$, we have that $\mathbf{x} = (\mathbf{z}_t - \sigma_t \boldsymbol{\xi})/\alpha_t$. It follows that

$$\|\mathbf{x} - \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2 = \bigg\| \frac{\mathbf{z}_t - \sigma_t \boldsymbol{\xi}}{\alpha_t} - \frac{\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)}{\alpha_t} \bigg\|^2 = \frac{\sigma_t^2}{\alpha_t^2} \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2 = \frac{1}{\mathrm{SNR}(t)} \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|^2.$$

Now,

$$\frac{\mathrm{SNR}'(t)}{\mathrm{SNR}(t)} = \frac{\{\exp(-\gamma_{\boldsymbol{\eta}}(t))\}'}{\exp(-\gamma_{\boldsymbol{\eta}}(t))} = -\frac{\exp(-\gamma_{\boldsymbol{\eta}}(t))}{\exp(-\gamma_{\boldsymbol{\eta}}(t))} \gamma'_{\boldsymbol{\eta}}(t) = -\gamma'_{\boldsymbol{\eta}}(t).$$

As a result,

$$\begin{aligned}
\mathcal{L}_\infty(\mathbf{x}) &= -\frac{1}{2} E_{\xi\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)} \big[\mathrm{SNR}'(t) \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z}_t;t)\|^2 \big] \\
&= -\frac{1}{2} E_{\xi\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)} \bigg[\frac{\mathrm{SNR}'(t)}{\mathrm{SNR}(t)} \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}(\mathbf{z}_t;t)\|^2 \bigg] \\
&= \frac{1}{2} E_{\xi\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(0,1)} \big[\gamma'_{\boldsymbol{\eta}}(t) \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}(\mathbf{z}_t;t)\|^2 \big]
\end{aligned}$$

as required. □

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.

[Khu22] Pramook Khungurn. Denoising diffusion probabilistic models. `https://pkhungurn.github.io/notes/notes/ml/ddpm/ddpm.pdf`, 2022. Accessed: 2022-11-11.

[KSPH21] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2021.

[LH17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[ND21] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[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.

[Tou11] Marc Toussaint. Lecture notes: Gaussian identities. `https://www.user.tu-berlin.de/mtoussai/notes/gaussians.pdf`, 2011. Accessed: 2022-11-08.

[TSM+20] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *CoRR*, abs/2006.10739, 2020.