Supplement

Representation Learning with Variational Diffusion Models

Egoitz Gonzalez Diego Garcia Cerdas Jacky Chu Maks Kulicki

Project supervised by Samuele Papa for the course Deep Learning II

University of Amsterdam

1 Problem setting

Let's assume we have a true underlying data distribution $p_{\text{data}}(\mathbf{x})$, and a dataset $\mathcal{D}$ of i.i.d. samples from this distribution. Our goal is to learn a likelihood $p_\theta(\mathbf{x})$ from $\mathcal{D}$ that is very similar to $p_{\text{data}}(\mathbf{x})$, so that we can generate new data by sampling from it.

In general, we are also interested in modeling the *factors* $\mathbf{w}$ in data that are crucial for generating an object. Once modeled, these factors can be used as a *representation* of the original data.

Latent variable models introduce the generative process $p_\theta(\mathbf{x}, \mathbf{w}) = p_\theta(\mathbf{x}|\mathbf{w})p(\mathbf{w})$, where

- $p(\mathbf{w})$ is a *prior distribution* over the latent variables, and
- $p_\theta(\mathbf{x}|\mathbf{w})$ is the *conditional distribution*, parameterized by θ.

Since only $\mathbf{x}$ is accessible during training, we marginalize out the latent variables $\mathbf{w}$ to obtain the likelihood of the data:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{w}} p(\mathbf{w})p_\theta(\mathbf{x}|\mathbf{w})d\mathbf{w}.$$

If we want to know how probable a representation $\mathbf{w}$ is for given data $\mathbf{x}$, we are interested in the *posterior distribution*

$$p_\theta(\mathbf{w}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{w})p(\mathbf{w})/p_\theta(\mathbf{x}).$$

We follow a maximum-likelihood-based approach, where the objective is to find the optimal parameters θ according to:

$$\max_\theta \mathbb{E}_{p_{\text{data}}(\mathbf{x})}[\log p_\theta(\mathbf{x})],$$

where the expectation is approximated using a sample average over $\mathcal{D}$.

However, since the integral over $\mathbf{w}$ is very commonly intractable, a common approach is to perform approximate inference.

2 Variational AutoEncoder

We use a VAE [5] to approximate the intractable posterior $p_\theta(\mathbf{w}|\mathbf{x})$ via a tractable *variational posterior* $q_\phi(\mathbf{w}|\mathbf{x})$.

We make the following assumptions:

- The prior over the latent variables is the standard multivariate Gaussian $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- The variational posterior is a Gaussian distribution $q_\phi(\mathbf{w}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})\mathbf{I})$, and $\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi$ are the outputs of a neural network encoder E_ϕ.
- The conditional distribution $p_\theta(\mathbf{x}|\mathbf{w})$ is parameterized by a Variational Diffusion Model D_θ conditioned on the latent variables $\mathbf{w}$.

We are able to jointly optimize network parameters ϕ and θ by maximizing the variational lower bound (VLB) on the (marginal) log-likelihood:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{\mathbf{w}\sim q_\phi(\mathbf{w}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{w})] - D_{\text{KL}}(q_\phi(\mathbf{w}|\mathbf{x})||p(\mathbf{w})) \\ &= -\mathcal{L}(\theta, \phi; \mathbf{x})\end{aligned}$$

For details on the derivation, see Appendix A.

3 Variational Diffusion Model

We employ a VDM [4] for the task of modeling the conditional distribution $p_\theta(\mathbf{x}|\mathbf{w})$. In the following, we repeat the formulation introduced by its authors in order to point out the specific assumptions we make, as well as to outline how we introduce the conditioning on $\mathbf{w}$.

3.1 Forward process

Consider a *variance-preserving* Gaussian diffusion process that defines a sequence of increasingly noise versions of $\mathbf{x}$ represented by the latent variables $\mathbf{z}_t$, with $t \in [0, 1]$, where $t = 0$ is the least noisy version and $t = 1$ is the noisiest version.

The distribution of latent variables $\mathbf{z}_t$ conditioned on $\mathbf{x}$ for $t \in [0, 1]$ is given by:

$$q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\alpha_t\mathbf{x}, \sigma_t^2\mathbf{I}),$$

with $\alpha_t = \sqrt{1 - \sigma_t^2}$ and $q(\mathbf{z}_1|\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We make the simplifying assumption that the noise schedule σ_t^2 for $t \in [0, 1]$ follows a fixed form.

The distributions $q(\mathbf{z}_t|\mathbf{z}_s)$, for $t > s$ are also Gaussian. Given this setting, we can verify through Bayes rule that $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})$ is also Gaussian. The shape of these distributions is given in Appendix B.

3.2 Reverse process

A hierarchical generative model is defined by inverting the diffusion process defined above and conditioning on $\mathbf{w}$. For a finite T, we consider T timesteps of width $\tau = 1/T$.

We define $s(i) = (i-1)/T$ and $t(i) = i/T$, such that our (conditional) hierarchical generative model is given by:

$$p_\theta(\mathbf{x}|\mathbf{w}) = \int_{\mathbf{z}} p(\mathbf{z}_1) p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) \prod_{i=1}^{T} p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w}).$$

With the variance-preserving diffusion process, we have $p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We use a factorized distribution for $p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w})$:

$$p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) = \prod_i p_\theta(x_i|z_{0,i}, \mathbf{w}),$$

where we choose $p_\theta(x_i|z_{0,i}, \mathbf{w}) \propto q(z_{0,i}|x_i)$, which is normalized by summing over all possible discrete values of x_i.
The conditional model distributions are given by:

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

i.e. the same as $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})$, but with the original $\mathbf{x}$ replaced by the output of a (conditional) *denoising model* $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)$.

In practice, we parameterize the conditional denoising model in terms of a conditional *noise prediction model* $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t)$ as follows:

$$\hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t) = (\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t))/\alpha_t,$$

with $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t)$ parameterized by a neural network.

3.3 Variational lower bound

We can optimize the parameters θ by maximizing the variational lower bound of the conditional distribution, given by:

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{w}) &\geq -\mathcal{L}_0(\theta; \mathbf{x}, \mathbf{w}) - \mathcal{L}_1(\mathbf{x}, \mathbf{z}_1) - \mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) \\ &= -\mathcal{L}_{\text{VDM}}(\theta; \mathbf{x}, \mathbf{w}) \end{aligned}$$

where $\mathcal{L}_0(\theta; \mathbf{x}, \mathbf{w})$ is the reconstruction loss

$$\mathcal{L}_0(\theta; \mathbf{x}, \mathbf{w}) = \mathbb{E}_{q(\mathbf{z}_0|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w})],$$

$\mathcal{L}_1(\mathbf{x}, \mathbf{z}_1)$ is the prior loss

$$\mathcal{L}_1(\mathbf{x}, \mathbf{z}_1) = D_{\text{KL}}[q(\mathbf{z}_1|\mathbf{x})||p(\mathbf{z}_1)],$$

which due to our assumptions remains constant and is ignored during training, and $\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w})$ is the diffusion loss

$$\begin{aligned} &\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) = \\ &\sum_{i=1}^{T} \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})} D_{\text{KL}}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})||p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})]. \end{aligned}$$

For details on the derivation, see Appendix C.

The reconstruction loss can be estimated using standard techniques. The diffusion loss is estimated through:

$$\begin{aligned} &\mathcal{L}_t^D(\theta; \mathbf{x}, \mathbf{w}) \\ &= \frac{T}{2} \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), i \sim \mathcal{U}\{1,T\}} \\ &\left[(\text{SNR}(s(i)) - \text{SNR}(t(i)))\, ||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t(i))||_2^2\right], \end{aligned}$$

where $\mathcal{U}\{1, T\}$ is the discrete uniform distribution between 1 and T and $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ is the signal-to-noise ratio at timestep t.

In the case of a noise prediction model $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t)$, the estimator becomes:

$$\begin{aligned} &\mathcal{L}_t^N(\theta; \mathbf{x}, \mathbf{w}) = \\ &\frac{T}{2} \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), i \sim \mathcal{U}\{1,T\}} \\ &[(\text{SNR}(s(i)) - \text{SNR}(t(i)))||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t(i))||_2^2]. \end{aligned}$$

The complete simplification is described in Appendix D.

4 Representation Learning

Putting everything together, we jointly optimize network parameters ϕ and θ by minimizing the following loss function:

$$\begin{aligned} &\mathcal{L}(\theta, \phi; \mathbf{x}) = \\ &\mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \mathcal{L}_{\text{VDM}}(\theta; \mathbf{x}, \mathbf{w}) + D_{\text{KL}}(q_\phi(\mathbf{w}|\mathbf{x})||p(\mathbf{w})). \end{aligned}$$

4.1 Connection to score-based modeling

Consider the marginal distribution of $\mathbf{z}_t$:

$$q(\mathbf{z}_t) = \int q(\mathbf{z}_t|\mathbf{x}) p_{data}(\mathbf{x}) d\mathbf{x}.$$

Explicit score matching (ESM) [2] attempts to learn a score model $\mathbf{s}_\theta(\mathbf{z}_t; t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$, which allows sampling from $q(\mathbf{z}_t)$ using Langevin dynamics.

[6] proves that optimizing the denoising score matching (DSM) objective yields *consistent* score models $\mathbf{s}_\theta(\mathbf{z}_t; t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x})$, in the sense that with infinite data, they are equivalently optimizing the ESM objective.

[3] shows that a certain parameterization of diffusion models reveals an equivalence with DSM over multiple noise levels during training, which implies consistency. Moreover, [4] generalizes the original consistency proof of DSM to show that, with infinite data, the optimal score model $\mathbf{s}_\theta^*(\mathbf{z}_t; t)$ for the derived VLB estimators is such that:

$$\mathbf{s}_\theta^*(\mathbf{z}_t; t) = \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t).$$

It is possible to parameterize the conditional denoising model described in Section 2.3.2 in terms of a conditional score model $\mathbf{s}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t)$ as follows:

$$\hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t) = (\sigma_t^2 \mathbf{s}_\theta(\mathbf{z}_t, \mathbf{w}; t) + \mathbf{z}_t)/\alpha_t.$$

The diffusion loss estimator becomes:

$$\begin{aligned} &\mathcal{L}_t^S(\theta; \mathbf{x}, \mathbf{w}) = \\ &\frac{T}{2} \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), i \sim \mathcal{U}\{1,T\}} \\ &[||\sqrt{c\,(t(i))} \left(\nabla_{\mathbf{z}_{t(i)}} \log q(\mathbf{z}_{t(i)}|\mathbf{x}) - \mathbf{s}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t)\right) ||_2^2], \end{aligned}$$

where $\sqrt{c\,(t(i))}$ is a time-dependent weighting factor.

[1] claims that $\mathcal{L}_t^S(\theta; \mathbf{x}, \mathbf{w})$ is a valid representation learning objective. The score $\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x})$ is a function of

only t, $\mathbf{z}_t$, and $\mathbf{x}$. Thus, when $\mathbf{w}$ contains all information about $\mathbf{x}$, it is possible to learn

$$\mathbf{s}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) = \mathbf{s}^*_\theta(\mathbf{z}_t; t).$$

When $\mathbf{w}$ has no mutual information with $\mathbf{x}$, it is only possible to learn

$$\mathbf{s}_\theta(\mathbf{z}_{t(i)}, \mathbf{w}; t) \approx \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x}),$$

which minimizes the DSM objective only up to a constant

$$||\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t|\mathbf{x}) - \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)||_2^2.$$

Therefore, we expect it to be beneficial for optimization of $\mathcal{L}(\theta, \phi; \mathbf{x})$ to learn a variational posterior $q_\phi(\mathbf{w}|\mathbf{x})$ that maximizes the mutual information between $\mathbf{w}$ and $\mathbf{x}$.

References

[1] Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. 2022.

[2] A. Hyvärinen. Estimation of non-normalized statistical models using score matching, 2005.

[3] Ajay Jain Jonathan Ho and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[4] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 2021.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[6] Pascal Vincent. A connection between score matching and denoising autoencoders, 2011.

Appendices

A VLB on marginal log-likelihood

We are able to jointly optimize network parameters ϕ and θ by maximizing the variational lower bound on the (marginal) log-likelihood:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\
&= \log \int \frac{q_\phi(\mathbf{w}|\mathbf{x})}{q_\phi(\mathbf{w}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{w}) p(\mathbf{w}) dw \\
&= \log \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \left[\frac{p_\theta(\mathbf{x}|\mathbf{w}) p(\mathbf{w})}{q_\phi(\mathbf{w}|\mathbf{x})} \right] \\
&\geq \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \log \left[\frac{p_\theta(\mathbf{x}|\mathbf{w}) p(\mathbf{w})}{q_\phi(\mathbf{w}|\mathbf{x})} \right] \\
&= \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}) - \log q_\phi(\mathbf{w}|\mathbf{x}) \right] \\
&= \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{w}) \right] + \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \left[\log p(\mathbf{w}) - \log q_\phi(\mathbf{w}|\mathbf{x}) \right] \\
&= \mathbb{E}_{\mathbf{w} \sim q_\phi(\mathbf{w}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{w}) \right] - D_{\mathrm{KL}}(q_\phi(\mathbf{w}|\mathbf{x}) || p(\mathbf{w})) \\
&= -\mathcal{L}(\theta, \phi; \mathbf{x}).
\end{aligned}$$

B Forward diffusion process

The following is extracted from [4].

The distribution $q(\mathbf{z}_t|\mathbf{z}_s)$, for any $0 \leq s < t \leq 1$ is given by:

$$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\alpha_{t|s}\mathbf{z}_s, \sigma^2_{t|s}\mathbf{I}),$$

where

$$\alpha_{t|s} = \alpha_t / \alpha_s,$$

and

$$\sigma^2_{t|s} = \sigma^2_t - \alpha^2_{t|s}\sigma^2_s.$$

The distribution $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})$, for any $0 \leq s < t \leq 1$ is given by:

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

where

$$\sigma^2_Q(s, t) = \sigma^2_{t|s}\sigma^2_s / \sigma^2_t$$

and

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

C VLB on conditional log-likelihood

For a finite number of timesteps T, we define $s(i) = (i-1)/T$ and $t(i) = i/T$.

Consider the set of latent variables $\mathbf{z}_{0:1} = [\mathbf{z}_0, ..., \mathbf{z}_1]$, with

$$p_\theta(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{w}) = p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) \left(\prod_{i=1}^{T} p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w}) \right) p(\mathbf{z}_1)$$

and

$$q(\mathbf{z}_{0:1}|\mathbf{x}) = q(\mathbf{z}_0|\mathbf{x}) \prod_{i=1}^{T} q(\mathbf{z}_{t(i)}|\mathbf{z}_{s(i)}).$$

We can optimize the parameters θ by maximizing the variational lower bound of the conditional distribution, given by:

$$
\begin{aligned}
\log p_\theta(\mathbf{x}|\mathbf{w}) &= \log \int p_\theta(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{w}) d\mathbf{z}_{0:1} \\
&= \log \int q(\mathbf{z}_{0:1}|\mathbf{x}) \frac{p_\theta(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{w}) d\mathbf{z}_{0:1}}{q(\mathbf{z}_{0:1}|\mathbf{x})} \\
&= \log \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{w})}{q(\mathbf{z}_{0:1}|\mathbf{x})} \right] \\
&\geq \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{w})}{q(\mathbf{z}_{0:1}|\mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \log \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) \left(\prod_{i=1}^T p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w}) \right) p(\mathbf{z}_1)}{q(\mathbf{z}_0|\mathbf{x}) \prod_{i=1}^T q(\mathbf{z}_{t(i)}|\mathbf{z}_{s(i)})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \log \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) \left(\prod_{i=1}^T p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w}) \right) p(\mathbf{z}_1)}{q(\mathbf{z}_0|\mathbf{x}) \prod_{i=1}^T q(\mathbf{z}_{t(i)}|\mathbf{z}_{s(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) p(\mathbf{z}_1)}{q(\mathbf{z}_0|\mathbf{x})} + \log \prod_{i=1}^T \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{q(\mathbf{z}_{t(i)}|\mathbf{z}_{s(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) p(\mathbf{z}_1)}{q(\mathbf{z}_0|\mathbf{x})} + \log \prod_{i=1}^T \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{\frac{q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x}) q(\mathbf{z}_{t(i)}|\mathbf{x})}{q(\mathbf{z}_{s(i)}|\mathbf{x})}} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) p(\mathbf{z}_1)}{q(\mathbf{z}_0|\mathbf{x})} + \log \frac{q(\mathbf{z}_0|\mathbf{x})}{q(\mathbf{z}_1|\mathbf{x})} + \log \prod_{i=1}^T \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) p(\mathbf{z}_1)}{q(\mathbf{z}_1|\mathbf{x})} + \sum_{i=1}^T \log \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w}) \right] + \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1|\mathbf{x})} \right] + \sum_{i=1}^T \mathbb{E}_{q(\mathbf{z}_{0:1}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_0|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w})] + \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \left[\log \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1|\mathbf{x})} \right] + \sum_{i=1}^T \mathbb{E}_{q(\mathbf{z}_{t(i)}, \mathbf{z}_{s(i)}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})}{q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x})} \right] \\
&= \mathbb{E}_{q(\mathbf{z}_0|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z}_0, \mathbf{w})] - D_{\mathrm{KL}}[q(\mathbf{z}_1|\mathbf{x}) || p(\mathbf{z}_1)] - \sum_{i=1}^T \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})} D_{\mathrm{KL}}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x}) || p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})] \\
&= -\mathcal{L}_0(\theta; \mathbf{x}, \mathbf{w}) - \mathcal{L}_1(\mathbf{x}, \mathbf{z}_1) - \mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) \\
&= -\mathcal{L}_{\mathrm{VDM}}(\theta; \mathbf{x}, \mathbf{w}).
\end{aligned}
$$

D Estimator for diffusion loss

The following derivations are extracted from [4], but we introduce conditioning on the latent representation $\mathbf{w}$. For a finite number of timesteps T, we define $s(i) = (i-1)/T$ and $t(i) = i/T$.

We now derive an estimator for the diffusion loss:

$$\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) = \sum_{i=1}^T \mathbb{E}_{q(\mathbf{z}_{t(i)}|\mathbf{x})} D_{\mathrm{KL}}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x}) || p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{w})].$$

Recall that $p_\theta(\mathbf{z}_s|\mathbf{z}_t, \mathbf{w}) = q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x} = \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t))$ for any $0 \leq s < t \leq 1$. Therefore,

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

and

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

with

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

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

and

$$\sigma_Q^2(s,t) = \sigma_{t|s}^2\sigma_s^2/\sigma_t^2.$$

Using $s = s(i)$ and $t = t(i)$, their KL divergence simplifies as:

$$\begin{aligned}
D_{\text{KL}}[q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})||p_\theta(\mathbf{z}_s|\mathbf{z}_t, \mathbf{w})] &= \frac{1}{2\sigma_Q^2(s,t)}||\boldsymbol{\mu}_Q - \boldsymbol{\mu}_\theta||_2^2 \\
&= \frac{\sigma_t^2}{2\sigma_{t|s}^2\sigma_s^2}\frac{\alpha_s^2\sigma_{t|s}^4}{\sigma_t^4}||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2 \\
&= \frac{1}{2\sigma_s^2}\frac{\alpha_s^2\sigma_{t|s}^2}{\sigma_t^2}||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2 \\
&= \frac{1}{2\sigma_s^2}\frac{\alpha_s^2(\sigma_t^2 - \alpha_{t|s}^2\sigma_s^2)}{\sigma_t^2}||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2 \\
&= \frac{1}{2}\frac{\alpha_s^2\sigma_t^2/\sigma_s^2 - \alpha_t^2}{\sigma_t^2}||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2 \\
&= \frac{1}{2}\left(\frac{\alpha_s^2}{\sigma_s^2} - \frac{\alpha_t^2}{\sigma_t^2}\right)||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2 \\
&= \frac{1}{2}\left(\text{SNR}(s) - \text{SNR}(t)\right)||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2.
\end{aligned}$$

Reparameterizing $\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})$ as $\mathbf{z}_t = \alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the diffusion loss becomes:

$$\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sum_{i=1}^{T}\left(\text{SNR}(s) - \text{SNR}(t)\right)||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2\right].$$

To avoid computing all T terms, we construct an unbiased estimator using:

$$\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) = \frac{T}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), i\sim\mathcal{U}\{1,T\}}\left[\left(\text{SNR}(s) - \text{SNR}(t)\right)||\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2\right],$$

where $\mathcal{U}\{1, T\}$ is the discrete uniform distribution between 1 and T.

In the case of a noise prediction model $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t)$, the estimator becomes:

$$\mathcal{L}_t(\theta; \mathbf{x}, \mathbf{w}) = \frac{T}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), i\sim\mathcal{U}\{1,T\}}\left[\left(\text{SNR}(s) - \text{SNR}(t)\right)||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{w}; t)||_2^2\right].$$