

Variational Inference for Diffusion Models

Tomonari MASADA

masada@rikkyo.ac.jp

Contents

Marginal likelihood

Variational lower bound

Variational posterior

Generative modeling of observations

Appendix

Marginal likelihood

We consider a Bayesian generative model whose joint distribution is

$$p_{\theta}(\mathbf{x}_{0:T}) = p_{\theta}(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) \quad (1)$$

where \mathbf{x}_0 is an observation and $\{\mathbf{x}_t : t = 1, \dots, T\}$ are latent random variables. The distribution of \mathbf{x}_t is only conditioned on \mathbf{x}_{t+1} . The log of the marginal likelihood (= the evidence) of \mathbf{x}_0 is

$$\log p_{\theta}(\mathbf{x}_0) = \log \int p_{\theta}(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \quad (2)$$

Contents

Marginal likelihood

Variational lower bound

Variational posterior

Generative modeling of observations

Appendix

ELBO

Jensen's inequality gives a lower bound of $\log p_{\theta}(\mathbf{x}_0)$ as follows:

$$\begin{aligned}\log p_{\theta}(\mathbf{x}_0) &= \log \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_{\theta}(\mathbf{x}_{0:T})}{q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\ &\geq \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\ &= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \equiv L_{\text{VLB}}\end{aligned}\quad (3)$$

where $q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is a variational posterior, which is conditioned on the observation \mathbf{x}_0 as in VAE. We train our model over many observations $\mathcal{X} \equiv \{\mathbf{x}_0^{(1)}, \dots, \mathbf{x}_0^{(N)}\}$ in an amortized manner.

Markov assumption

The definition of the conditional distribution gives the following equations:

$$\begin{aligned} q_\psi(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_0)q_\psi(\mathbf{x}_1|\mathbf{x}_0) &= \frac{q_\psi(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_1, \mathbf{x}_0)} \frac{q_\psi(\mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_0)} = \frac{q_\psi(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_0)} \\ &= q_\psi(\mathbf{x}_2, \mathbf{x}_1|\mathbf{x}_0) \end{aligned} \quad (4)$$

$$\begin{aligned} q_\psi(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)q_\psi(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_0)q_\psi(\mathbf{x}_1|\mathbf{x}_0) &= \frac{q_\psi(\mathbf{x}_3, \mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)} \frac{q_\psi(\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_1, \mathbf{x}_0)} \frac{q_\psi(\mathbf{x}_1, \mathbf{x}_0)}{q_\psi(\mathbf{x}_0)} \\ &= q_\psi(\mathbf{x}_3, \mathbf{x}_2, \mathbf{x}_1|\mathbf{x}_0) \end{aligned} \quad (5)$$

...

Therefore, by assuming that the equation $q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \dots, \mathbf{x}_1, \mathbf{x}_0) = q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$ holds for $t = 2, \dots, T$, we obtain the following factorization of $q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)$:

$$q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) = q_\psi(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^T q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \quad (6)$$

Then the variational lower bound L_{VLB} can be rewritten as follows:

$$\begin{aligned}
L_{\text{VLB}} &= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^T q_{\psi}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_{\theta}(\mathbf{x}_T) d\mathbf{x}_{1:T} + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \sum_{t=2}^T \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_{\psi}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&\quad + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)}{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0)} d\mathbf{x}_{1:T} \tag{7}
\end{aligned}$$

Using Bayes' rule, we have

$$q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q_\psi(\mathbf{x}_t|\mathbf{x}_0)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_0)} \quad (8)$$

We will discuss later how we make the above $q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$ tractable.

We replace $q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$ appearing in L_{VLB} with $\frac{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q_\psi(\mathbf{x}_t|\mathbf{x}_0)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_0)}$ based on Eq. (8) and rewrite L_{VLB} as follows:

$$\begin{aligned} L_{\text{VLB}} = & \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_\theta(\mathbf{x}_T) d\mathbf{x}_{1:T} + \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \sum_{t=2}^T \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\ & + \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \sum_{t=2}^T \log \frac{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q_\psi(\mathbf{x}_t|\mathbf{x}_0)} d\mathbf{x}_{1:T} + \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\psi(\mathbf{x}_1|\mathbf{x}_0)} d\mathbf{x}_{1:T} \quad (9) \end{aligned}$$

(continued on the next page)

$$\begin{aligned}
L_{\text{VLB}} &= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_{\theta}(\mathbf{x}_T) d\mathbf{x}_{1:T} + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \sum_{t=2}^T \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_{\psi}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&\quad + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0) \cancel{q_{\psi}(\mathbf{x}_2|\mathbf{x}_0)} \cdots \cancel{q_{\psi}(\mathbf{x}_{T-1}|\mathbf{x}_0)}}{\cancel{q_{\psi}(\mathbf{x}_2|\mathbf{x}_0)} \cancel{q_{\psi}(\mathbf{x}_2|\mathbf{x}_0)} \cdots q_{\psi}(\mathbf{x}_T|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&\quad + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)}{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_{\theta}(\mathbf{x}_T) d\mathbf{x}_{1:T} + \sum_{t=2}^T \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_{\psi}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&\quad + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{\cancel{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0)}}{q_{\psi}(\mathbf{x}_T|\mathbf{x}_0)} d\mathbf{x}_{1:T} + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)}{\cancel{q_{\psi}(\mathbf{x}_1|\mathbf{x}_0)}} d\mathbf{x}_{1:T} \\
&= \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_T)}{q_{\psi}(\mathbf{x}_T|\mathbf{x}_0)} d\mathbf{x}_{1:T} + \sum_{t=2}^T \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_{\psi}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&\quad + \int q_{\psi}(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_{\theta}(\mathbf{x}_0|\mathbf{x}_1) d\mathbf{x}_{1:T} \equiv L_T + \sum_{t=2}^T L_{t-1} + L_0
\end{aligned} \tag{10}$$

We can rewrite L_{t-1} as follows:

$$\begin{aligned}
L_{t-1} &\equiv \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \int \left(q_\psi(\mathbf{x}_1|\mathbf{x}_0) \prod_{t' \neq t} q_\psi(\mathbf{x}_{t'-1}|\mathbf{x}_{t'}, \mathbf{x}_0) \right) \left(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right) d\mathbf{x}_{1:T} \\
&= - \int \left(q_\psi(\mathbf{x}_1|\mathbf{x}_0) \prod_{t' \neq t} q_\psi(\mathbf{x}_{t'-1}|\mathbf{x}_{t'}, \mathbf{x}_0) \right) D_{\text{KL}}(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) d\mathbf{x}_{1:T} \\
&\equiv -\mathbb{E}_{\neg t} [D_{\text{KL}}(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))] \tag{11}
\end{aligned}$$

It can be said that, by minimizing this expectation of the KL-divergence

$D_{\text{KL}}(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$ for each t , we can maximize the ELBO in Eq. (7).

Contents

Marginal likelihood

Variational lower bound

Variational posterior

Generative modeling of observations

Appendix

Parameterization of variational posterior

We parameterize our variational posterior with the parameters $\psi \equiv \{\alpha_t : t = 1, \dots, T\}$ as

$$q_\psi(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \quad (12)$$

Therefore, $q_\psi(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) = q_\psi(\mathbf{x}_t | \mathbf{x}_{t-1})$ for $t = 2, \dots, T$. This parameterization makes $q_\psi(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0)$ in Eq. (8) tractable.

Based on Eq. (28) in Appendix, we obtain $q_\psi(\mathbf{x}_t | \mathbf{x}_{t-2})$ as follows:

$$\begin{aligned} q_\psi(\mathbf{x}_t | \mathbf{x}_{t-2}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2}, ((1 - \alpha_t) + \alpha_t(1 - \alpha_{t-1})) \mathbf{I}) \\ &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2}, (1 - \alpha_t \alpha_{t-1}) \mathbf{I}) \end{aligned} \quad (13)$$

By repeating the same argument, we obtain $q_\psi(\mathbf{x}_t | \mathbf{x}_0)$ as follows:

$$q_\psi(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (14)$$

where $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. It is easy to sample from $q_\psi(\mathbf{x}_t | \mathbf{x}_0)$.

We regard ψ as free parameters and drop ψ from our notations for the rest of this presentation.

We rewrite $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ appearing in L_{t-1} of Eq. (11) as follows:

$$\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \quad (\text{based on Eq. (8)}) \\
&\propto \exp \left(-\frac{1}{2} \left(\frac{(\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_{t-1})^2}{1 - \alpha_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{1 - \bar{\alpha}_t} \right) \right) \\
&\propto \exp \left(-\frac{1}{2} \left(\left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t}}{1 - \alpha_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \mathbf{x}_{t-1} \right) \right)
\end{aligned} \tag{15}$$

We denote the element-wise mean and variance of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ by $\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0)$ and $\tilde{\beta}_t$ respectively. That is,

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \equiv \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t) \tag{16}$$

Then we obtain

$$\tilde{\beta}_t = 1 / \left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{\alpha_t - \alpha_t \bar{\alpha}_{t-1} + 1 - \alpha_t} = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} (1 - \alpha_t) \tag{17}$$

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) &= \left(\frac{\sqrt{\alpha_t}}{1 - \alpha_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) / \left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \\
&= \left(\frac{\sqrt{\alpha_t}}{1 - \alpha_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} (1 - \alpha_t) \\
&= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \mathbf{x}_0
\end{aligned} \tag{18}$$

Based on Eq. (14), we reparameterize \mathbf{x}_t as

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + (1 - \bar{\alpha}_t) \boldsymbol{\epsilon} \quad \text{for } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{19}$$

and rewrite $\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0)$ as follows:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \left(\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \right) \\
&= \frac{1}{\sqrt{\alpha_t}} \left(\left(\frac{\alpha_t(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} + \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} \right) \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon} \right) \\
&= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon} \right)
\end{aligned} \tag{20}$$

Contents

Marginal likelihood

Variational lower bound

Variational posterior

Generative modeling of observations

Appendix

Generative modeling of observations

Here we specify the details of our Bayesian generative model firstly in this presentation.

$$p_{\theta}(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \quad (21)$$

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t)) \quad (22)$$

We assume that $\boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$ as discussed in [1].

Eqs. (16) and (22) show that the KL divergence appearing in L_{t-1} of Eq. (11) is from one Gaussian distribution to another. Therefore, we can rewrite L_{t-1} as follows¹:

$$L_{t-1} = -\mathbb{E}_{\neg t} \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t)\|^2 \right] + \text{const.} \quad (23)$$

¹https://scoste.fr/posts/dkl_gaussian/

By using the reparameterization $\mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1 - \bar{\alpha}_t)\epsilon$ in Eq. (19) and the result in Eq. (20), we further rewrite L_{t-1} as

$$L_{t-1} = -\mathbb{E}_{\neg t} \left[\frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \epsilon), t) \right\|^2 \right] + \text{const.} \quad (24)$$

We may parameterize $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ as follows [1]:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) \quad (25)$$

where $\boldsymbol{\epsilon}_\theta$ is a function approximator intended to predict ϵ from \mathbf{x}_t . Then we can rewrite L_{t-1} as follows:

$$L_{t-1} = -\mathbb{E}_{\neg t} \left[\frac{(1 - \alpha_t)^2}{2\sigma_t^2(1 - \bar{\alpha}_t)} \left\| \epsilon - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1 - \bar{\alpha}_t)\epsilon, t) \right\|^2 \right] + \text{const.} \quad (26)$$

We consider L_T in Eq. (10):

$$L_T \equiv \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_T)}{q_\psi(\mathbf{x}_T|\mathbf{x}_0)} d\mathbf{x}_{1:T} \quad (27)$$

Both of the noise distribution $p_\theta(\mathbf{x}_T)$ and the approximate posterior $q_\psi(\mathbf{x}_T|\mathbf{x}_0)$ have no trainable parameters. Therefore, L_T can be regarded as a constant.

Next, we consider L_0 in Eq. (10). How we maximize L_0 depends on how we specify the distribution $p_\theta(\mathbf{x}_0|\mathbf{x}_1)$, which directly models the observation. For example, see Sec. 3.3 of [1].

Contents

Marginal likelihood

Variational lower bound

Variational posterior

Generative modeling of observations

Appendix

Appendix

$$\begin{aligned} & \int \exp \left(-\frac{(x-ay)^2}{2s^2} - \frac{(y-bz)^2}{2t^2} \right) dy = \int \exp \left(-\frac{t^2(x-ay)^2 + s^2(y-bz)^2}{2s^2t^2} \right) dy \\ &= \int \exp \left(-\frac{(s^2 + t^2a^2)y^2 - 2(s^2bz + t^2ax)y + t^2x^2 + s^2b^2z^2}{2s^2t^2} \right) dy \\ &= \exp \left(-\frac{t^2x^2 + s^2b^2z^2}{2s^2t^2} \right) \int \exp \left(-\frac{s^2 + t^2a^2}{2s^2t^2} \left(y^2 - \frac{2(s^2bz + t^2ax)}{s^2 + t^2a^2} y \right) \right) dy \\ &= \exp \left(-\frac{t^2x^2 + s^2b^2z^2}{2s^2t^2} + \frac{(s^2bz + t^2ax)^2}{2s^2t^2(s^2 + t^2a^2)} \right) \int \exp \left(-\frac{s^2 + t^2a^2}{2s^2t^2} \left(y - \frac{s^2bz + t^2ax}{s^2 + t^2a^2} \right)^2 \right) dy \\ &\propto \exp \left(-\frac{s^2t^2x^2 + s^4b^2z^2 + t^4a^2x^2 + s^2t^2a^2b^2z^2 - t^4a^2x^2 - 2s^2t^2abzx - s^4b^2z^2}{2s^2t^2(s^2 + t^2a^2)} \right) \\ &= \exp \left(-\frac{x^2 - 2abzx + a^2b^2z^2}{2(s^2 + t^2a^2)} \right) = \exp \left(-\frac{(x-abz)^2}{2(s^2 + t^2a^2)} \right) \end{aligned} \tag{28}$$



Jonathan Ho, Ajay Jain, and Pieter Abbeel.

Denoising diffusion probabilistic models.

CoRR, [abs/2006.11239](https://arxiv.org/abs/2006.11239), 2020.