Variational Inference for Diffusion Models

Tomonari MASADA

masada@rikkyo.ac.jp

Contents

Marginal likelihood

We consider a Bayesian generative model whose joint distribution is

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

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 marginal likelihood ($=$ evidence) of $\mathbf{x}_0$ is

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

Contents

ELBO

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

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

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 $\mathcal{X} \equiv \{\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(N)}\}$ in an amortized manner.

Factorization assumption

We assume that $q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)$ factorizes as

$$q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) = q_\psi(\mathbf{x}_1|\mathbf{x}_0)\prod_{t=2}^{T} q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0) \tag{4}$$

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

$$\begin{aligned}
L_{\mathsf{VLB}} &= \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)\log\frac{p_\theta(\mathbf{x}_{0:T})}{q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)}d\mathbf{x}_{1:T} \\
&= \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)\log\frac{p_\theta(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_1|\mathbf{x}_0)\prod_{t=2}^{T} q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)}d\mathbf{x}_{1:T} \\
&= \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)\log p_\theta(\mathbf{x}_T)d\mathbf{x}_{1:T} + \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)\sum_{t=2}^{T}\log\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)}d\mathbf{x}_{1:T} \\
&\quad + \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0)\log\frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q_\psi(\mathbf{x}_1|\mathbf{x}_0)}d\mathbf{x}_{1:T}
\end{aligned} \tag{5}$$

Using Bayes' rule, we have

$$q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0) = \frac{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)q_\psi(\mathbf{x}_t|\mathbf{x}_0)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_0)} \tag{6}$$

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. (6) and rewrite L_{VLB} as follows:

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

(cont.)

$$L_{\mathsf{VLB}} = \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_T)}{q_\psi(\mathbf{x}_T|\mathbf{x}_0)} d\mathbf{x}_{1:T} + \sum_{t=2}^{T} \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T}$$

$$+ \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) d\mathbf{x}_{1:T} \equiv L_T + \sum_{t=2}^{T} L_{t-1} + L_0 \tag{8}$$

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

$$\begin{aligned}
L_{t-1} &\equiv \int q_\psi(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \int q_\psi(\mathbf{x}_1|\mathbf{x}_0) \prod_{t' \neq t} q_\psi(\mathbf{x}_{t'-1}|\mathbf{x}_{t'}, \mathbf{x}_0) \Big(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \Big) d\mathbf{x}_{1:T} \\
&= -\int q_\psi(\mathbf{x}_1|\mathbf{x}_0) \prod_{t' \neq t} q_\psi(\mathbf{x}_{t'-1}|\mathbf{x}_{t'}, \mathbf{x}_0) D_{\mathsf{KL}}(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) d\mathbf{x}_{1:T} \\
&\equiv -\mathbb{E}_{\neg t}\big[D_{\mathsf{KL}}(q_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \big]
\end{aligned} \tag{9}$$

Contents

Tractable variational posterior

We define our variational posterior with parameters $\psi \equiv \{\alpha_t : t = 1, \ldots, T\}$ as follows:

$$q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{10}$$

Therefore, $q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = q_\psi(\mathbf{x}_t|\mathbf{x}_{t-1})$ for $t > 1$.

Based on Eq. (24), we can obtain $q_\psi(\mathbf{x}_t|\mathbf{x}_{t-2})$ as follows:

$$\begin{aligned} q_\psi(\mathbf{x}_t|\mathbf{x}_{t-2}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2}, \big((1-\alpha_t) + \alpha_t(1-\alpha_{t-1})\big)\mathbf{I}) \\ &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2}, (1-\alpha_t\alpha_{t-1})\mathbf{I}) \end{aligned} \tag{11}$$

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

$$q_\psi(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{12}$$

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 from now on.

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

$$
\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \quad \text{(by using Bayes' rule)} \\
&\propto \exp\Bigg(-\frac{1}{2}\Big(\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{1-\alpha_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_{t1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\Big)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\Big(\big(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\big)\mathbf{x}_{t-1}^2 - 2\big(\frac{\sqrt{\alpha_t}}{1-\alpha_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\big)\mathbf{x}_{t-1}\Big)\Bigg)
\end{aligned}
\tag{13}
$$

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,

$$
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t) \tag{14}
$$

Then

$$
\tilde{\beta}_t = 1/\Bigg(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\Bigg) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{\alpha_t - \alpha_t\bar{\alpha}_{t-1} + 1 - \alpha_t} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t) \tag{15}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) &= \left(\frac{\sqrt{\alpha_t}}{1-\alpha_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right) \Big/ \left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) \\
&= \left(\frac{\sqrt{\alpha_t}}{1-\alpha_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right) \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t) \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\mathbf{x}_0
\end{aligned} \tag{16}
$$

Based on Eq. (12), we reparameterize $\mathbf{x}_t$ as $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and rewrite $\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0)$ as follows:

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\Big(\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}\Big) \\
&= \left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{1-\alpha_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\right)\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon} \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\left(\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} + \frac{1-\alpha_t}{1-\bar{\alpha}_t}\right)\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)
\end{aligned} \tag{17}
$$

Contents

Generative modeling of observations

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

$$p_\theta(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \tag{18}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{19}$$

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

Based on Eqs. (14) and (19), it can be said that the KL divergence appearing in L_{t-1} in Eq. (9) is a KL divergence from one Gaussian distribution to another.

Therefore, we can rewrite L_{t-1} as follows:

$$L_{t-1} = -\mathbb{E}_{\neg t}\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + const. \tag{20}$$

By using the reparameterization $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}$ and Eq. (17), we further rewrite L_{t-1} as follows:

$$L_{t-1} = -\mathbb{E}_{\neg t}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t)\right\|^2\right] + const. \tag{21}$$

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

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) \tag{22}$$

where $\boldsymbol{\epsilon}_\theta$ is a function approximator intended to predict $\boldsymbol{\epsilon}$ from $\mathbf{x}_t$.

By using parameterization, we can rewrite L_{t-1} as follows:

$$L_{t-1} = -\mathbb{E}_{\neg t}\left[\frac{(1-\alpha_t)^2}{2\sigma_t^2(1-\bar{\alpha}_t)}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}, t)\|^2\right] + const. \tag{23}$$

Contents

Appendix 1

$$\begin{aligned}
&\int \exp\left(-\frac{(x-ay)^2}{2s^2}-\frac{(y-bz)^2}{2t^2}\right)dy = \int \exp\left(-\frac{t^2(x-ay)^2+s^2(y-bz)^2}{2s^2t^2}\right)dy \\
&= \int \exp\left(-\frac{(s^2+t^2a^2)y^2-2(s^2bz+t^2ax)y+t^2x^2+s^2b^2z^2}{2s^2t^2}\right)dy \\
&= \exp\left(-\frac{t^2x^2+s^2b^2z^2}{2s^2t^2}\right)\int \exp\left(-\frac{s^2+t^2a^2}{2s^2t^2}\left(y^2-\frac{2(s^2bz+t^2ax)}{s^2+t^2a^2}y\right)\right)dy \\
&= \exp\left(-\frac{t^2x^2+s^2b^2z^2}{2s^2t^2}+\frac{(s^2bz+t^2ax)^2}{2s^2t^2(s^2+t^2a^2)}\right)\int \exp\left(-\frac{s^2+t^2a^2}{2s^2t^2}\left(y-\frac{s^2bz+t^2ax}{s^2+t^2a^2}\right)^2\right)dy \\
&\propto \exp\left(-\frac{s^2t^2x^2+s^4b^2z^2+t^4a^2x^2+s^2t^2a^2b^2z^2-t^4a^2x^2-2s^2t^2abzx-s^4b^2z^2}{2s^2t^2(s^2+t^2a^2)}\right) \\
&= \exp\left(-\frac{x^2-2abzx+a^2b^2z^2}{2(s^2+t^2a^2)}\right) = \exp\left(-\frac{(x-abz)^2}{2(s^2+t^2a^2)}\right)
\end{aligned} \tag{24}$$

Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models.
CoRR, abs/2006.11239, 2020.