CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 13: VAE and Diffusion II

Lecturer: Bo Dai *Scribes: Yitong Li, Feng Gao*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

13.1 Recap

- Latent Variable Models: $p(x) = \int p(x,z)\,dz = \int p(x|z)p(z)\,dz$
- This integral is intractable, therefore we introduce a more tractable distribution $q(z)$ to replace $p(x|z)$.
- To better calculate $p(x)$, we introduce Evidence Lower Bound (ELBO), which derives a lower bound on $\log p(x_i)$ using variational inference.
- $\log p(x_i) = \mathbb{E}_{q_i(z)}\left[\log \frac{p(x_i,z)}{q_i(z)}\right] + \mathrm{KL}(q_i(z) \parallel p(z|x_i))$
  KL divergence part is also referred to as $H(q)$.
- Therefore, we have the inequality: $\log p(x_i) \geq \mathbb{E}_{q_i(z)}\left[\log \frac{p(x_i,z)}{q_i(z)}\right]$

13.2 New Content

Given that x_T follows a Gaussian distribution $\mathcal{N}(0, I)$, the update equation for x_{t-1} is given by:

$$x_{t-1} = x_t + \eta_t \nabla f(x_t) + \sqrt{2\eta_t}\epsilon_t, \tag{13.1}$$

where $\epsilon_t \sim N(0, I)$ represents the standard normal noise.

We define the following components:

$$S_\theta(x_t, t) = \eta_t \nabla f(x_t), \quad \text{(deterministic part)} \tag{13.2}$$

$$\Sigma_\theta(x_t, t) = \sqrt{2\eta_t}, \quad \text{(stochastic part)} \tag{13.3}$$

Thus, the conditional probability distribution for x_{t-1} given x_t can be written as:

$$p(x_{t-1}|x_t) = \mathcal{N}(S_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{13.4}$$

Finally, the joint probability distribution for the sequence $x_0, x_1, \ldots x_T$ is:

$$p(x_0, x_1, \ldots x_T) = \prod_{t=1}^{T} p(x_t|x_{t-1})p(x_T). \tag{13.5}$$

Next, we have the marginal distribution $p(x_0)$, which is obtained by integrating over the joint probability $p(x_0, \ldots, x_T)$ with respect to all intermediate variables $x_1, \ldots, x_T$. This can be expressed as:

$$p(x_0) = \int p(x_0, x_1, \ldots, x_T)\, dx_1 \ldots dx_T. \tag{13.6}$$

The log-probability $\log p(x_0)$ can be written as the logarithm of the integral over the joint probability and we now introduce a component $q(x_1, \ldots, x_T|x_0)$, which we divide and multiply inside the integral. The expression for $\log p(x_0)$ becomes:

$$\begin{aligned}
\log p(x_0) &= \log \int p(x_0, x_1, \ldots, x_T) dx_0 dx_1 \ldots dx_T && (13.7)\\
&= \log \int \frac{p(x_0, x_1, \ldots, x_T)}{q(x_1, \ldots, x_T|x_0)} q(x_1, \ldots, x_T|x_0) dx_0 dx_1 \ldots dx_T && (13.8)\\
&\geq \mathbb{E}_q \left[\log \frac{p(x_0, x_1, \ldots, x_T)}{q(x_1, \ldots, x_T|x_0)} \right] && (13.9)\\
&\geq \mathbb{E}_q \left[\log \prod_{i=1}^{T} p(x_{t-1}|x_t) p(x_T) \right] + H(q) && (13.10)
\end{aligned}$$

To optimize the ELBO, we aim to maximize the following objective:

$$\begin{aligned}
&\max_\theta \max_q \mathbb{E}_q \left[\log \prod_{i=1}^{T} p(x_{t-1}|x_t) p(x_T) \right] + H(q) && (13.11)\\
\Rightarrow &\max_q \max_\theta \mathbb{E}_q \left[\log \prod_{i=1}^{T} p(x_{t-1}|x_t) p(x_T) \right] + H(q) && (13.12)\\
\Rightarrow &\max_\theta \mathbb{E}_q \left[\log \prod_{i=1}^{T} p(x_{t-1}|x_t) p(x_T) \right] + H(q) && (13.13)
\end{aligned}$$

Since q depends on θ, we we can change the order of optimization, effectively treating q as arbitrary and leaving the task of optimization to θ.

When introducing q and selecting which q to use, it's important to consider the limitations of the current ELBO equation:

- q is high-dimensional, with T steps, making it computationally expensive to sample the entire trajectory. (using close-form q)
- The current choice of q is complex and performs poorly. (reducing variance in ELBO)

We define $q(x_T, \ldots x_1|x_0) = \prod_{i=1}^{T} q(x_t|x_{t-1})$, where each $q(x_t|x_{t-1})$ is modeled as a Gaussian distribution $\mathcal{N}(\sqrt{1-\beta_{t-1}}x_{t-1}, \beta_{t-1}I)$. This formulation is equivalent to the forward process of adding Gaussian noise in a diffusion model.

Fact 1 (Gaussian Forward Process): Since every timestep follows Gaussian distribution, we have

$$q(x_t|x_0) = \mathcal{N}(\prod_{i=1}^{t} \sqrt{1-\beta_t}x_0, (1-\overline{\alpha_t})I) \tag{13.14}$$

where $\alpha_i = 1 - \beta_i$ and $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$.

We could deduce a close form for $q(x_{t-1}|x_t, x_0)$ from (13.14), which is

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t, x_{t-1}|x_0)}{q(x_t|x_0)} \tag{13.15}$$

$$= \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{13.16}$$

$$\propto \exp\left[-\frac{1}{2}(\frac{1}{\beta_t}||x_t - \sqrt{1-\beta_{t-1}}x_{t-1}||^2 + \frac{1}{1-\overline{\alpha_t}}||x_{t-1} - \overline{\alpha}_{t-1}x_0||^2 - \frac{1}{\bar{\alpha}_t}||x_t - \sqrt{\overline{\alpha}_t}x_0||^2)\right]. \tag{13.17}$$

Fact 2 (Close-form q): From 13.17, we have

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\mu(x_t, x_0), \tilde{\beta}_t) \tag{13.18}$$

$$\mu(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 \tag{13.19}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{13.20}$$

We will now show how **Fact 2** helps reduce variance in the ELBO. Starting from (13.10), we have

$$\mathbb{E}_q\left[\log\prod_{i=1}^{T} p(x_{t-1}|x_t)p(x_T) - \log q(x_1, \ldots, x_T|x_0)\right] \tag{13.21}$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{i=1}^{T}\log\frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] \tag{13.22}$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{i=1}^{T}\log\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}\right] \tag{13.23}$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{i=1}^{T}\log\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{i=1}^{T}\log\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}\right] \tag{13.24}$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{i=1}^{T}\log\frac{p(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log\frac{q(x_T|x_0)}{q(x_1|x_0)}\right] \tag{13.25}$$

$$= \mathbb{E}_q\left[\log p(x_T)\right] - D_{KL}\left[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\right] - D_{KL}\left[q(x_1|x_0)||q(x_T|x_0)\right]. \tag{13.26}$$

Therefore, we get the reduced version as matching $\mathcal{N}(S_\theta(x_t, t), \Sigma_\theta(x_t, t))$ to $\mathcal{N}(\mu(x_t, x_0), \tilde{\beta}_t I)$, which is equivalent to matching score with noise in diffusion model.

Our goal is to maximize the ELBO with respect to $\boldsymbol{\theta}$, and only the second term is relevant to θ. Therefore, we must focus on accurately parameterizing p_θ and optimizing θ to minimize this divergence.