10-708: Probabilistic Graphical Models 10-708, Spring 2017

Homework3

Lecturer: Eric P. Xing *Name: Yuan Liu(yuanl4)*

1 Variational Autoencoders

1.1 Derive the evidence lower bound

$$
\begin{aligned}
D_{KL}(q_\phi(z|x)||p_\theta(z|x)) &= \int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x)}dz = \int q_\phi(z|x)\log\frac{q_\phi(z|x)p_\theta(x)}{p_\theta(z|x)p_\theta(x)}dz \\
&= E_{q_\phi(z|x)}\left[\log\frac{q_\phi(z|x)p_\theta(x)}{p_\theta(x,z)}\right] = E_{q_\phi(z|x)}\left[\log\frac{q_\phi(z|x)}{p_\theta(x,z)}\right] + \log p_\theta(x)
\end{aligned}
$$

Because we know $D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \geq 0$. Then we can get:

$$\log p_\theta(x) \geq E_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

$$logp_\theta(D) = \sum_{i=1}^{n}\log p_\theta(x^{(i)}) \geq \sum_{i=1}^{n} E_{q_\phi(z|x^{(i)})}\left[\log\frac{p_\theta(x^{(i)},z)}{q_\phi(z|x^{(i)})}\right]$$

1.2 Wake-sleep algorithm

- **Wake phase:** Maximize the bound with respect to p_θ, which means

$$\theta := arg\max_\theta \sum_{i=1}^{n} E_{q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right]$$

  - We first generate samples $\{z^{(i,j)}\}_{j=1}^m$ from $q(z|x^{(i)})$ for each i.
  - Then we maximize

$$\theta := arg\max_\theta \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\log p_\theta(x^{(i)}|z^{(i,j)})$$

- **Sleep phase:** Maximize $F'(\theta,\psi;x) = -\log p_\theta(x) + D_{KL}(p_\theta(z|x)||q_\phi(z|x))$ with respect to $q_\phi(z|x)$, which means:

$$\phi := arg\max_\phi E_{p_\theta(z,x)}\left[\log q_\phi(z|x)\right]$$

  - We first generate samples $\{(x^{(j)}, z^{(j)})\}_{j=1}^m$ from $p_\theta(z,x)$ through top-down pass.
  - Then we maximize

$$\phi := arg\max_\phi \frac{1}{m}\sum_{j=1}^{m}\log q_\phi(z^{(j)}|x^{(j)})$$

- **Advantage:** It is generally applicable to a wide range of generative models by training a separate inference network.

- **Disadvantages:**(1) KL distance is not symmetric. (2) Doesn't optimize a well-defined objective function. (3) Not guaranteed to converge.

1.3 Autoencoding variational Bayes approach

According to ELBO we know:

$$\begin{aligned}\log p_\theta(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] &= E_{q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right] \\ &= E_{q_\phi(z|x)}\left[\log p_\theta(x|z)p(z) - \log q_\phi(z|x)\right] \\ &= E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}(q_\phi(z|x)||p(z))\end{aligned}$$

Let

$$L(\theta,\phi,x) := E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}(q_\phi(z|x)||p(z))$$

- Maximize $L(\theta,\phi,x)$ with respect to $p_\theta(x|z)$. This step is the same as wake phase.
  - We first generate samples $\{z^{(i,j)}\}_{j=1}^m$ from $q(z|x^{(i)})$ for each i.
  - Then we maximize

$$\theta := arg\max_\theta \frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m \log p_\theta(x^{(i)}|z^{(i,j)})$$

$$\nabla_\theta L(\theta,\phi,x) \approx \frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m \nabla_\theta \log p_\theta(x^{(i)}|z^{(i,j)})$$

- Maximize $L(\theta,\phi,x)$ with respect to $q_\phi(z|x)$.

$$\nabla_\phi E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] = E_{\epsilon\sim N(0,I)}\left[\nabla_\phi \log p_\theta(x|z_\phi(\epsilon))\right]$$

$$\nabla_\phi D_{KL}(q_\phi(z|x)||p(z)) = E_{\epsilon\sim N(0,I)}\left[\nabla_\phi\left(\log q_\phi(z_\phi(\epsilon)|x) - \log p(z_\phi(\epsilon))\right)\right]$$

  - We first generate samples $\{\epsilon^{(i)}\}_{i=1}^m$ from normal distribution.
  - Then me get

$$\nabla_\phi L(\theta,\phi,x) \approx \frac{1}{m}\sum_{i=1}^m \nabla_\phi[\log p_\theta(x|z_\phi(\epsilon^{(i)})) - \log q_\phi(z_\phi(\epsilon^{(i)})|x) + \log p(z_\phi(\epsilon^{(i)})))]$$

- **Advantage** (1)Enjoy similar applicability with wake-sleep algorithm. (2)Reduce variance through reparameterization of the recognition distribution

- **Disadvantage** (1)Not applicable to discrete latent variables. (2)Usually use a fixed standard normal distribution as prior, leading to limited flexibility

1.4 Tighter bound

$$L_K = E[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}] \leq E[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}] = \log p(x)$$

Let $I \in 1, .., k+1$ with $|I| = k$ be a uniformly distributed subset of distinct indices from $1, ..., k+1$. We will use the following simple observation: $E_{I=\{i_1,..i_k\}}[\frac{a_{i_1}+...+a_{i_k}}{k}] = \frac{a_1+...+a_{k+1}}{k+1}$ for any sequence of numbers $a_{i_1} + ... + a_{i_k}$.

$$\begin{aligned}
L_{k+1} &= E[\log \frac{1}{k+1}\sum_{i=1}^{k+1}\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}] \\
&= E\left[\log E_{I=\{i_1,..i_k\}}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(x, z^{(i_k)})}{q_\phi(z^{(i_k)}|x)}\right]\right] \\
&\geq E\left[E_{I=\{i_1,..i_k\}}\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(x, z^{(i_k)})}{q_\phi(z^{(i_k)}|x)}\right]\right] \\
&= E\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)}\right] \\
&= L_k
\end{aligned}$$

Then we can get

$$\log p(x) \geq L_{k+1}(x) \geq L_k(x)$$

1.5 Counterexample

We know if $\frac{p_\theta(x,z)}{q_\phi(z|x)}$ is bounded, from the strong law of large numbers, $\lim_{n\to\infty} L_n \to \log p(x)$. Then we can get: if we want to find counter example, we need to make $\frac{p_\theta(x,z)}{q_\phi(z|x)}$ unbounded. For example, let $p_\theta(x, z)$ is supported on $[0, 1] \times [0, 1]$, and make $q_\phi(z|x) \equiv 0$ on $[0, 1] \times [0, 1]$.

2 Implementation and Experience

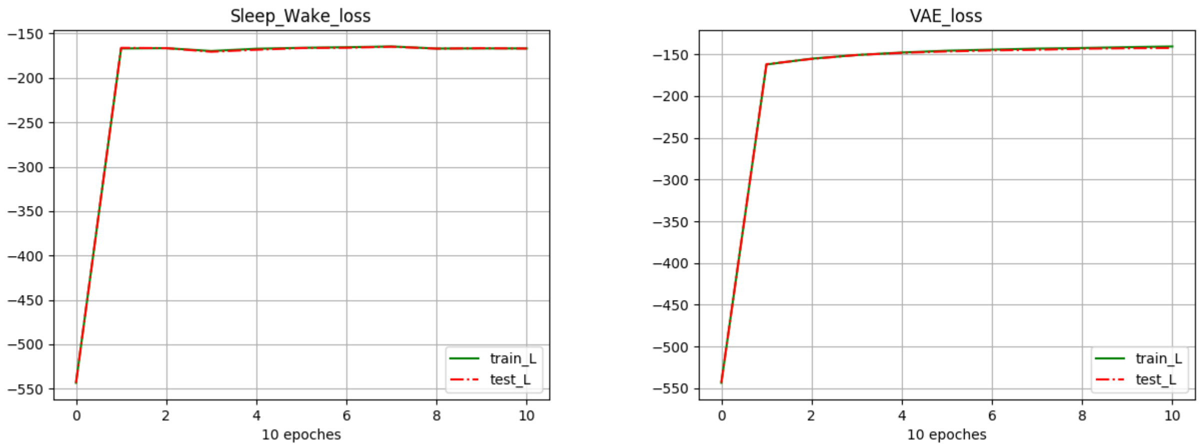

Figure 1: L_{1000} of both algorithm. The left one is about sleep wake algorithm, and the right part is about AEVB algorithm

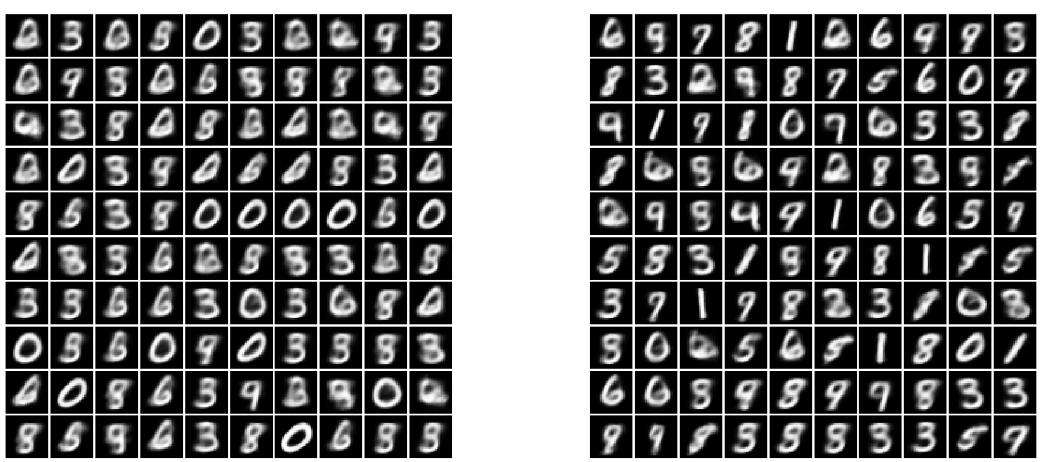

Figure 2: Visualize of both algorithm. The left one is about sleep wake algorithm, and the right part is about AEVB algorithm

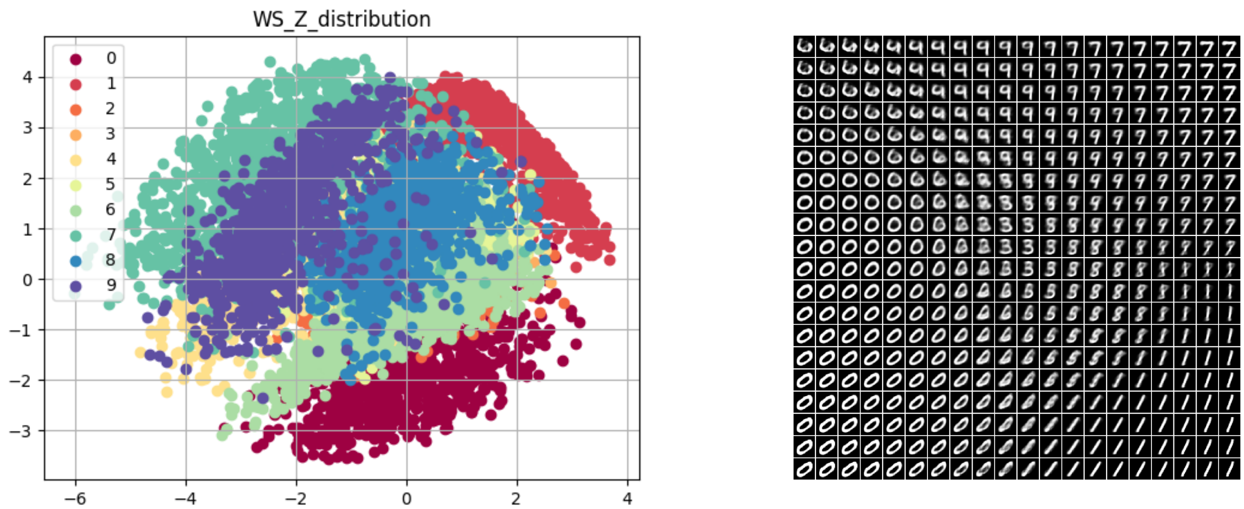

Figure 3: The distribution of wake_sleep algorithm

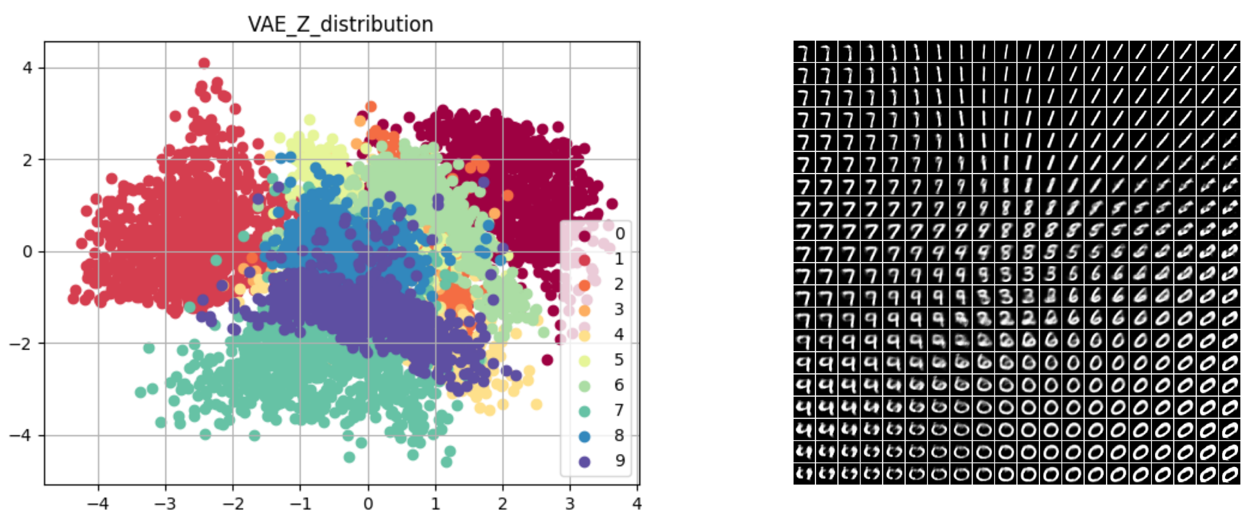

Figure 4: The distribution of AEVB algorithm