netid: lw2350
name: Lizhong Wang

Homework 4

1. ELBO

1.

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x}) d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \left(\frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z} \mid \mathbf{x})} \cdot \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right) d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \left(\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} - \log p_\theta(\mathbf{z} \mid \mathbf{x}) + \log q_\phi(\mathbf{z} \mid \mathbf{x}) \right) d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} - \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] + \left(- \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} - \mathbb{H}\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \right] \right) \\
&= \mathrm{ELBO}(\theta, \phi; \mathbf{x}) + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right]
\end{aligned}
$$

2. By definition, KL divergence is non-negative. Therefore, $\log p_\theta(\mathbf{x})$ is the sum of ELBO and a non-negative term. Therefore, $\log p_\theta(\mathbf{x}) \geq \mathrm{ELBO}(\theta, \phi; \mathbf{x})$.

   $\log p_\theta(\mathbf{x}) = \mathrm{ELBO}(\theta, \phi; \mathbf{x})$ when $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] = 0$, which implies that $q_\phi(\mathbf{z} \mid \mathbf{x})$ is equivalent to $p_\theta(\mathbf{z} \mid \mathbf{x})$.
   Therefore, $\log p_\theta(\mathbf{x}) = \mathrm{ELBO}(\theta, \phi; \mathbf{x})$ when $q_\phi(\mathbf{z} \mid \mathbf{x})$ is equivalent to $p_\theta(\mathbf{z} \mid \mathbf{x})$.

2. ELBO surgery

1.

$$\begin{aligned}
\mathrm{ELBO}(\theta,\phi;\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\right] \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log \frac{p_\theta(\mathbf{x}\mid\mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log p_\theta(\mathbf{x}\mid\mathbf{z}) d\mathbf{z} + \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log p_\theta(\mathbf{z}) d\mathbf{z} - \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log q_\phi(\mathbf{z}\mid\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] - \left(- \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log p_\theta(\mathbf{z}) d\mathbf{z} - \mathbb{H}\left[q_\phi(\mathbf{z}\mid\mathbf{x})\right]\right) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z}\mid\mathbf{x}) \| p_\theta(\mathbf{z})\right]
\end{aligned}$$

2.

- In practice, we minimize $-\mathrm{ELBO}(\theta,\phi;\mathbf{x})$ instead of maximize $\mathrm{ELBO}(\theta,\phi;\mathbf{x})$. That is we will minimize:

$$-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] + \mathrm{KL}\left[q_\phi(\mathbf{z}\mid\mathbf{x}) \| p_\theta(\mathbf{z})\right].$$

- The $-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right]$ term is the reconstruction loss. It measures how different the reconstructed $\hat{x}$ and the original x are.

  Geometrically, this term pushes the center of z "bubbles" away from each other. So that for any value of the latent variable z, there's only one x that is highly likely been generated by this z.

- The $\mathrm{KL}\left[q_\phi(\mathbf{z}\mid\mathbf{x}) \| p_\theta(\mathbf{z})\right]$ term acts as a regularization term. It prevents the z generation distribution, $q_\phi(\mathbf{z}\mid\mathbf{x})$, goes too far from $p_\theta(\mathbf{z})$ in order to overfit x.

  Geometrically, the $\mathrm{KL}\left[q_\phi(\mathbf{z}\mid\mathbf{x}) \| p_\theta(\mathbf{z})\right]$ term keeps the z "bubbles" together, and prevents the z "bubbles" from going to infinitely far away from each other.

3.

$$\begin{aligned}
\mathrm{ELBO}(\theta,\phi;\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\right] \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log p_\theta(\mathbf{x},\mathbf{z}) d\mathbf{z} - \int q_\phi(\mathbf{z}\mid\mathbf{x}) \log q_\phi(\mathbf{z}\mid\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[p_\theta(\mathbf{x},\mathbf{z})\right] + \mathbb{H}\left[q_\phi(\mathbf{z}\mid\mathbf{x})\right]
\end{aligned}$$

3. Reconstruction loss

1.

$$
\begin{aligned}
-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) &= -\log \prod_{d=1}^{D} \mathrm{Bernoulli}(x_d\ ; \hat{x}_d) \\
&= -\log \prod_{d=1}^{D} \hat{x}_d^{x_d}\ (1-\hat{x}_d)^{1-x_d} \\
&= -\sum_{d=1}^{D} \log \left[\hat{x}_d^{x_d}\ (1-\hat{x}_d)^{1-x_d}\right] \\
&= -\sum_{d=1}^{D} \left[x_d \log \hat{x}_d + (1-x_d)\log(1-\hat{x}_d)\right] \\
&= \text{binary cross-entropy loss summed over dimensions 1...D}
\end{aligned}
$$

2.

$$
\begin{aligned}
-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) &= -\log \prod_{d=1}^{D} \mathcal{N}(x_d\ ; \hat{x}_d, \sigma^2) \\
&= -\log \prod_{d=1}^{D} \frac{e^{-\frac{1}{2}\left(\frac{x_d-\hat{x}_d}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} \\
&= -\sum_{d=1}^{D} \log \frac{e^{-\frac{1}{2}\left(\frac{x_d-\hat{x}_d}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} \\
&= -\sum_{d=1}^{D} \left[-\frac{1}{2}\left(\frac{x_d-\hat{x}_d}{\sigma}\right)^2 - \log(\sigma\sqrt{2\pi})\right] \\
&= \sum_{d=1}^{D} \left[\frac{1}{2}\left(\frac{x_d-\hat{x}_d}{\sigma}\right)^2\right] + D\log(\sigma\sqrt{2\pi}) \\
&= \left[\frac{1}{2\sigma^2}\sum_{d=1}^{D}(x_d-\hat{x}_d)^2\right] + \left[D\log(\sigma\sqrt{2\pi})\right] \\
&\text{,where } D\log(\sigma\sqrt{2\pi}) \text{ is a constant}
\end{aligned}
$$

Therefore, under these assumptions, $-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}})$ equals the MSE loss summed over dimensions 1...D up to a constant.

4. Short answer

1. **Reparameterization**
   Backward propagation in PyTorch won't work if we directly sample z. By using reparameterization, all randomness goes to ϵ, which allows us to calculate the gradient with respect to μ and Σ using PyTorch.

2. **Overlapping latents**
   When $p_\theta(x^{(1)}) \approx p_\theta(x^{(2)})$, meaning $x^{(1)}$ and $x^{(2)}$ are roughly equally likely to occur, we get $p_\theta(x^{(1)} \mid z) \approx p_\theta(x^{(2)} \mid z)$ according to Bayes' theorem.
   This is problematic because we are not certain whether we should reconstruct back to $x^{(1)}$ or $x^{(2)}$ for the given z.

3. **Missing labels**
   We can use denoising autoencoder and restricted Boltzmann machine to leverage all data.

4. **Discrete latent variables**
   The reconstruction term will become problematic because we don't know how to reparameterize z anymore. Reinforce estimator can be used for a VAE with a discrete, categorical latent variable.