CS 533: Natural Language Processing **(Due: 04/21/20)**

Assignment 5

Name: Keya Desai*, NetID:* kd706 *Students discussed with:* tn268, phj15

Problem 1: EM (1 + 5 + 5 = 11 points)

The variant of bigram language model is given as follows:

$$p(x_1...x_T, z_1...z_T) = \prod_{t=1}^{T} p(z_t|x_{t-1})p(x_t|z_t)$$

1. **Maximum likelihood estimate (MLE) of the model parameters**
   Given: $x_1...x_T \in V$ and $z_1...z_T \in 1...m$
   The following can be written as a function of data given. The idea is to count the instances of x_t and z_t from the data at a given time.

$$\begin{aligned} p(z|x) &= \frac{\sum_{t=1}^{T} \mathbb{1}(x_{t-1} = x, z_t = z)}{\sum_{t=1}^{T}(\sum_{z'=1}^{m} \mathbb{1}(x_{t-1} = x, z_t = z'))} \qquad \forall z \in \{1...m\}, x \in V \cup \{*\} \\ p(x'|z) &= \frac{\sum_{t=1}^{T} \mathbb{1}(x_t = x', z_t = z)}{\sum_{t=1}^{T}(\sum_{x'' \in V} \mathbb{1}(x_t = x'', z_t = z))} \qquad \forall z \in \{1...m\}, x' \in V \end{aligned} \tag{1}$$

   In the equation (1), $\mathbb{1}$ is an indicator function which is 1 only when the condition in the bracket is satisfied, 0 otherwise.

2. **E step**
   The conditional distributions $p(z|x)$ and $p(x'|z)$ that maximize $log \sum_{z_1...z_T \in \{1...m\}} p(x_1...x_T, z_1...z_T)$
   The E step calculates the posterior distribution for each data point as a function of the current model parameters $p(z|x)$ and $p(x'|z)$:

$$p(z|x_{t-1}, x_t) = \frac{p(z|x = x_{t-1})p(x' = x_t|z)}{\sum_{z' \in Z} p(z'|x = x_{t-1})p(x' = x_t|z')} \tag{2}$$

3. **M step**
   M step computes new parameter values $p^{new}(z|x)$ and $p^{new}(x'|z)$ as a function of the perdatum posterior probabilities $p(z|x_{t-1}, x_t)$:

$$\begin{aligned} p^{new}(z|x) &= \frac{\sum_{x_t \in V} count(x, x_t)p(z|x, x_t)}{\sum_{z' \in Z}(\sum_{x_t \in V} count(x, x_t)p(z'|x, x_t))} \\ p^{new}(x'|z) &= \frac{\sum_{x_{(t-1)} \in V \cup \{*\}} count(x_{(t-1)}, x')p(z|x_{(t-1)}, x')}{\sum_{x'' \in V}(\sum_{x_{(t-1)} \in V \cup \{*\}} count(x_{(t-1)}, x'')p(z|x_{(t-1)}, x''))} \end{aligned} \tag{3}$$

Problem 2: VAE (3 + 3 + 3 = 9 points)

1. VAE objective $J^{ELBO}(\theta, \phi)$ is given as follows:

   $J^{ELBO}(\theta, \phi)$ = (Reconstruction term) - (KL Divergence for the parameters y and z)

$$\begin{aligned} J^{ELBO}(\theta, \phi) = & E_{y \sim q^{\phi}_{Y|X}(.|x), z \sim q^{\phi}_{Z|X}(.|x)}[log(p^{\theta}_{X|YZ}(x|y,z))] \\ & - D_{KL}[q_{Y|X}(y|x)||p_Y(y)] \\ & - D_{KL}[q_{Z|X}(z|x)||p_Z(z)] \end{aligned} \tag{4}$$

2. Re-expressing $J^{ELBO}(\theta, \phi)$ after reparameterization trick as follows:
   Assuming the latent distributions to be Gaussians:

$$q^{\phi}_{Y|X}(.|x) = \mathcal{N}(\mu_{\phi,Y}(x), diag(\sigma^2_{\phi,Y}(x))$$

$$q^{\phi}_{Z|X}(.|x) = \mathcal{N}(\mu_{\phi,Z}(x), diag(\sigma^2_{\phi,Z}(x))$$

   Now,

$$\begin{aligned} & y \sim \mathcal{N}(y|\mu_{\phi,Y}(x), diag(\sigma^2_{\phi,Y}(x)) \\ & \quad = \mu_{\phi,Y}(x) + \mathcal{N}(y|0, I_d) \odot \sigma^2_{\phi,Y}(x) \\ & y = \mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma^2_{\phi,Y}(x); \qquad \text{where } \epsilon_y \sim \mathcal{N}(.|0, I_d) z \quad \sim \mathcal{N}(z|\mu_{\phi,Z}(x), diag(\sigma^2_{\phi,Z}(x)) \\ & \quad = \mu_{\phi,Z}(x) + \mathcal{N}(z|0, I_d) \odot \sigma^2_{\phi,Z}(x) \\ & z = \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma^2_{\phi,Z}(x); \qquad \text{where } \epsilon_z \sim \mathcal{N}(.|0, I_d) \end{aligned} \tag{5}$$

   Rewriting the reconstruction term using this information,

$$\begin{aligned} & E_{y \sim q^{\phi}_{Y|X}(.|x), z \sim q^{\phi}_{Z|X}(.|x)}[log(p^{\theta}_{X|YZ}(x|y,z))] \\ & = E_{\epsilon_y \sim \mathcal{N}(.|0,I_d), \epsilon_z \sim \mathcal{N}(.|0,I_d)}[log(p^{\theta}_{X|YZ}(x|\mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma^2_{\phi,Y}(x), \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma^2_{\phi,Z}(x)))] \end{aligned} \tag{6}$$

   Rewriting the KL-Divergence terms,

$$\begin{aligned} & D_{KL}[q_{Y|X}(y|x)||p_Y(y)] \\ & = D_{KL}[\mathcal{N}(\mu_{\phi,Y}(x), diag(\sigma^2_{\phi,Y}(x))||\mathcal{N}(y|0, I_d)] \\ & = \frac{1}{2}(\sum_{i=1}^{d}(\mu^2_{\phi,Y} + \sigma^2_{\phi,Y} - 1 - log(\sigma^2_{\phi,Y})) \end{aligned} \tag{7}$$

$$\begin{aligned} & D_{KL}[q_{Z|X}(z|x)||p_Z(z)] \\ & = D_{KL}[\mathcal{N}(\mu_{\phi,Z}(x), diag(\sigma^2_{\phi,Z}(x))||\mathcal{N}(z|0, I_d)] \\ & = \frac{1}{2}(\sum_{i=1}^{d}(\mu^2_{\phi,Z} + \sigma^2_{\phi,Z} - 1 - log(\sigma^2_{\phi,Z})) \end{aligned} \tag{8}$$

   Putting these values in the equation 4,

$$\begin{aligned} J^{ELBO}(\theta, \phi) = & E_{\epsilon_y \sim \mathcal{N}(.|0,I_d), \epsilon_z \sim \mathcal{N}(.|0,I_d)}[log(p^{\theta}_{X|YZ}(x|\mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma^2_{\phi,Y}(x), \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma^2_{\phi,Z}(x)))] \\ & - \frac{1}{2}(\sum_{i=1}^{d}(\mu^2_{\phi,Y} + \sigma^2_{\phi,Y} - 1 - log(\sigma^2_{\phi,Y})) \\ & - \frac{1}{2}(\sum_{i=1}^{d}(\mu^2_{\phi,Z} + \sigma^2_{\phi,Z} - 1 - log(\sigma^2_{\phi,Z})) \end{aligned} \tag{9}$$

3. Performing the reparamterization trick, the computational graph is as given in Figure 1.:

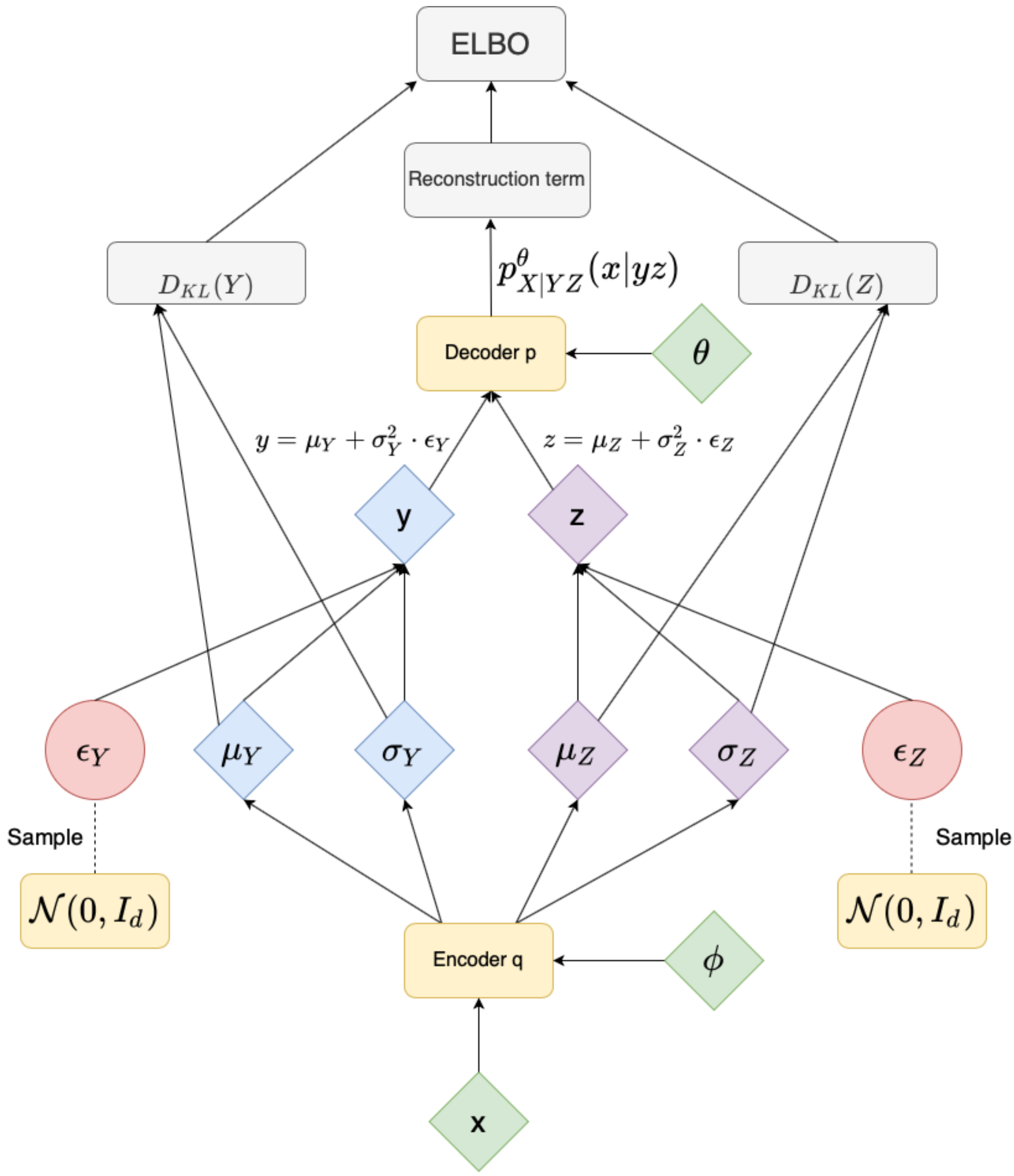

Figure 1: Computational Graph underlying the loss