

Assignment 5

Name: Twisha Naik, NetID: tn268

Students discussed with:

Problem 1: EM

(1 + 5 + 5 = 11 points)

The variant of bigram language model:

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

1. Maximum likelihood estimate (MLE) of the model parameters

Given: $x_1 \dots x_T \in V$ and $z_1 \dots z_T \in 1 \dots m$

The following can be written as a function of data given. The basic idea is just to count the instances of x_t and z_t from the data at a given time.

$$p(z|x) = \frac{\sum_{t=1}^T \mathbb{1}(x_{t-1} = x, z_t = z)}{\sum_{t=1}^T (\sum_{z'=1}^m \mathbb{1}(x_{t-1} = x, z_t = z'))} \quad \forall z \in \{1 \dots m\}, x \in V \cup \{*\}$$

$$p(x'|z) = \frac{\sum_{t=1}^T \mathbb{1}(x_t = x', z_t = z)}{\sum_{t=1}^T (\sum_{x'' \in V} \mathbb{1}(x_t = x'', z_t = z))} \quad \forall z \in \{1 \dots m\}, x' \in V$$

NOTE: In the above equations, $\mathbb{1}$ 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 $\log \sum_{z_1 \dots z_T \in \{1 \dots m\}} p(x_1 \dots x_T, z_1 \dots z_T)$

The E step calculates the posterior distribution for each data point as a function of the current model parameters $p(z|x)$ and $p(x'|z)$:

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

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 $p(z|x_{t-1}, x_t)$:

$$p^{new}(z|x) = \frac{\sum_{x_t \in V} \text{count}(x, x_t) p(z|x, x_t)}{\sum_{z' \in Z} (\sum_{x_t \in V} \text{count}(x, x_t) p(z'|x, x_t))}$$

$$p^{new}(x'|z) = \frac{\sum_{x_{(t-1)} \in V \cup \{*\}} \text{count}(x_{(t-1)}, x') p(z|x_{(t-1)}, x')}{\sum_{x'' \in V} (\sum_{x_{(t-1)} \in V \cup \{*\}} \text{count}(x_{(t-1)}, x'') p(z|x_{(t-1)}, x''))}$$

Problem 2: VAE

(3 + 3 + 3 = 9 points)

1. VAE objective
- $J^{ELBO}(\theta, \phi)$

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

$$\begin{aligned}
J^{ELBO}(\theta, \phi) &= E_{y \sim q_{Y|X}^\phi(\cdot|x), z \sim q_{Z|X}^\phi(\cdot|x)} [\log(p_{X|YZ}^\theta(x|y, z))] \\
&\quad - D_{KL}[q_{Y|X}(y|x) || p_Y(y)] \\
&\quad - D_{KL}[q_{Z|X}(z|x) || p_Z(z)]
\end{aligned} \tag{1}$$

2. Re-express
- $J^{ELBO}(\theta, \phi)$
- after reparameterization trick

Assuming the latent distributions to be Gaussians:

$$q_{Y|X}^\phi(\cdot|x) = \mathcal{N}(\mu_{\phi,Y}(x), \text{diag}(\sigma_{\phi,Y}^2(x)))$$

$$q_{Z|X}^\phi(\cdot|x) = \mathcal{N}(\mu_{\phi,Z}(x), \text{diag}(\sigma_{\phi,Z}^2(x)))$$

Now,

$$\begin{aligned}
y &\sim \mathcal{N}(y | \mu_{\phi,Y}(x), \text{diag}(\sigma_{\phi,Y}^2(x))) \\
&= \mu_{\phi,Y}(x) + \mathcal{N}(y|0, I_d) \odot \sigma_{\phi,Y}^2(x) \\
y &= \mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma_{\phi,Y}^2(x); \quad \text{where } \epsilon_y \sim \mathcal{N}(\cdot|0, I_d)
\end{aligned}$$

$$\begin{aligned}
z &\sim \mathcal{N}(z | \mu_{\phi,Z}(x), \text{diag}(\sigma_{\phi,Z}^2(x))) \\
&= \mu_{\phi,Z}(x) + \mathcal{N}(z|0, I_d) \odot \sigma_{\phi,Z}^2(x) \\
z &= \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma_{\phi,Z}^2(x); \quad \text{where } \epsilon_z \sim \mathcal{N}(\cdot|0, I_d)
\end{aligned}$$

Rewriting the reconstruction term using this information,

$$\begin{aligned}
&E_{y \sim q_{Y|X}^\phi(\cdot|x), z \sim q_{Z|X}^\phi(\cdot|x)} [\log(p_{X|YZ}^\theta(x|y, z))] \\
&= E_{\epsilon_y \sim \mathcal{N}(\cdot|0, I_d), \epsilon_z \sim \mathcal{N}(\cdot|0, I_d)} [\log(p_{X|YZ}^\theta(x | \mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma_{\phi,Y}^2(x), \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma_{\phi,Z}^2(x)))]
\end{aligned}$$

Rewriting the KL-Divergence terms,

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

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

Putting these values in the equation 1,

$$\begin{aligned}
 J^{ELBO}(\theta, \phi) &= E_{\epsilon_y \sim \mathcal{N}(\cdot|0, I_d), \epsilon_z \sim \mathcal{N}(\cdot|0, I_d)} [\log(p_{X|YZ}^\theta(x|\mu_{\phi,Y}(x) + \epsilon_y \cdot \sigma_{\phi,Y}^2(x), \mu_{\phi,Z}(x) + \epsilon_z \cdot \sigma_{\phi,Z}^2(x)))] \\
 &\quad - \frac{1}{2} \left(\sum_{i=1}^d (\mu_{\phi,Y}^2 + \sigma_{\phi,Y}^2 - 1 - \log(\sigma_{\phi,Y}^2)) \right) \\
 &\quad - \frac{1}{2} \left(\sum_{i=1}^d (\mu_{\phi,Z}^2 + \sigma_{\phi,Z}^2 - 1 - \log(\sigma_{\phi,Z}^2)) \right)
 \end{aligned}$$

3. Computation graph underlying the loss After performing the reparamterization trick, the computational graph formed is as shown in the image below:

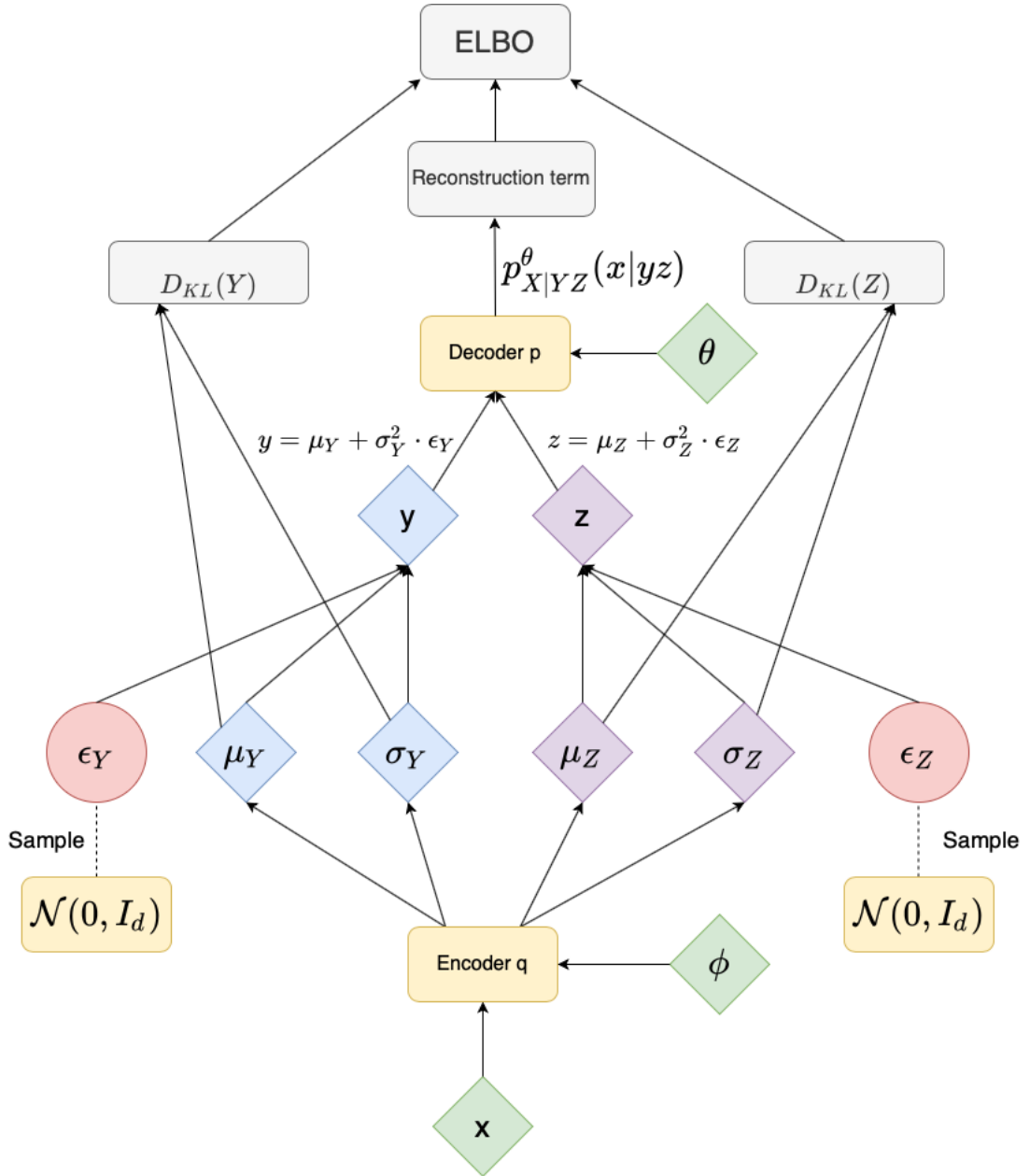


Figure 1: Computational Graph underlying the loss