1(c) Means of metrics from 7 runs of the training are as following.

NELBO : 101.687

KL Divergence : 19.261

Reconstruction loss : 82.425

Std. Dev. for the same experiment are:

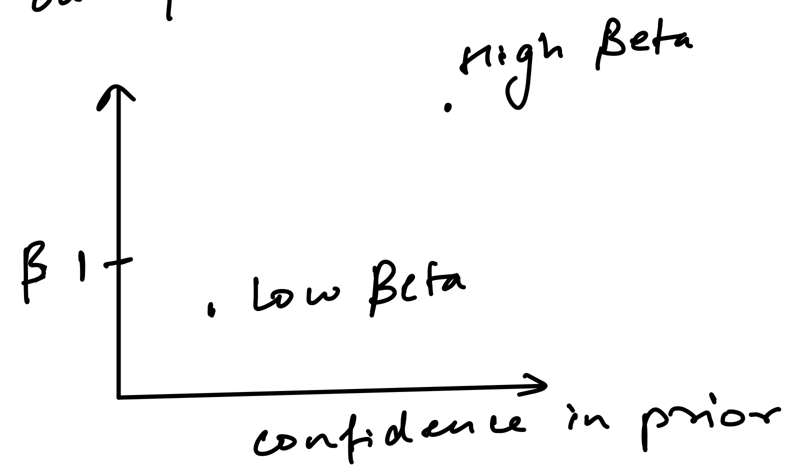
NELBO : 0.674

KL Divergence : 0.143

Reconstruction Loss : 0.781

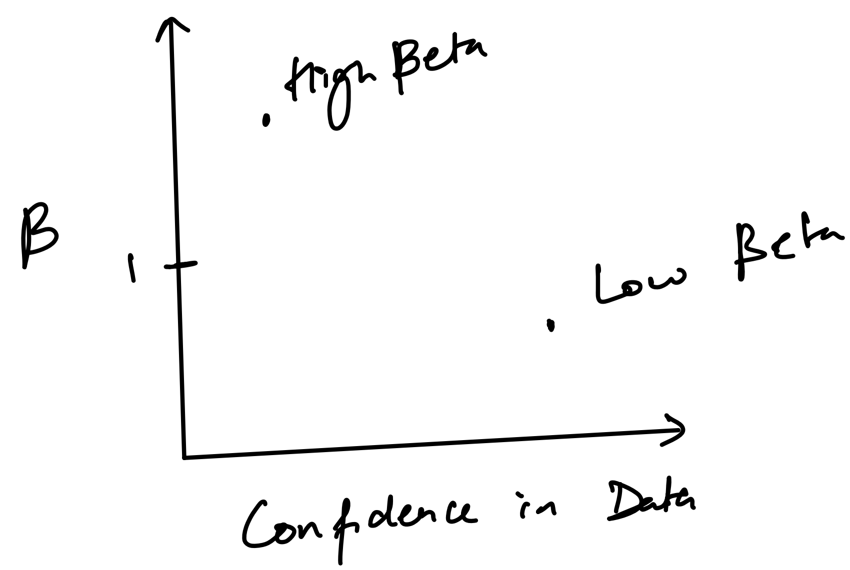
1(d) 'β' in β-VAE seems to be acting like a weightage to KL Divergence. By tuning this, we are trying to convey how much we care for posterior $q_\phi(z|x)$ to be close to prior $p(z)$ - which is gaussian in our case.

- When we set $\beta = 1$, then the ELBO remains same as VAE.
- When we set $\beta < 1$, we are giving less weightage to KL Divergence as compared to Reconstruction Loss. This may make intuitive sense to have since we really don't have any knowledge about the prior and it's just a guess.
- When we set $\beta > 1$, we are giving more weightage to KL Divergence. We would probably want to do only when we want our posterior $p(z|x)$ to follow a certain shape which we would only want to do if we have some knowledge about our prior ($p(z)$).



The above perspective is from the perspective of prior. We can see this from the perspective of data also i.e. how much we want to learn from data. If we $\beta < 1$, what we are saying is that $p(z|x)$ is allowed to diverge from our prior $p(z)$ i.e. we want to learn more from data. This may result in overfitting.

For $\beta > 1$, we want to stick closer to prior. Essentially, we want to learn less from data and we are putting our bias in terms of selection of prior. This may lead to underfitting.



3 a) In this problem, we need to prove

$$\log p_\theta(x) \geq \mathcal{L}_m(x) \geq \mathcal{L}_1(x)$$

where $\mathcal{L}_m(x) = E_{z^{(1)} \dots z^{(m)} \sim q_\phi(z|x)} \left(\log \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)} \right)$

$$\mathcal{L}_1(x) = E_{z \sim q_\phi(z|x)} \left(\log \frac{p_\theta(x,z)}{q_\phi(z|x)} \right)$$

First, we'll talk about

$$\log p_\theta(x) \geq \mathcal{L}_m(x)$$

$$p_\theta(x) = \sum_z p_\theta(x,z)$$

$$= \sum_{z \in Z} \frac{q(z)}{q(z)} p_\theta(x,z)$$

$$= E_{z \sim q(z)} \left[\frac{p_\theta(x,z)}{q(z)} \right]$$

Using Monte Carlo which samples 'k' elements at a time,

$$p_\theta(x) \approx E_{z^{(j)} \sim q(z)} \left[\frac{1}{k} \sum_{j=1}^{k} \frac{p_\theta(x, z^{(j)})}{q(z^{(j)})} \right]$$

Taking log on both si ,

$$\log(p_\theta(x)) \approx \log\left(E_{z^{(j)} \sim q(z)}\left[\frac{1}{k}\sum_{j=1}^{k}\frac{p_\theta(x, z^{(j)})}{q(z^{(j)})}\right]\right)$$

Using Jensen's Inequality, we get

$$\geq E_{z^{(j)} \sim q(z)}\left[\log\left(\frac{1}{k}\sum_{j=1}^{k}\frac{p_\theta(x, z^{(j)})}{q(z^{(j)})}\right)\right]$$

For better estimates of $z^{(j)}$ we want $z^{(j)}$ to be picked from $q_\phi(z|x)$. So the above equation becomes.

$$\geq E_{z^{(j)} \sim q_\phi(z|x)}\left[\log\left(\frac{1}{k}\sum_{j=1}^{k}\frac{p_\theta(x, z^{(j)})}{q_\phi(z^{(j)}|x)}\right)\right]$$

which is RHS. Hence,

$$\boxed{\log p_\theta(x) \geq \mathcal{L}_m(x)}$$

Now coming to the second part,

$$\mathcal{L}_m(x) \geq \mathcal{L}_1(x),$$

$$\log \frac{1}{m} \sum_{i=1}^{m} \left(\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)} \right)$$

$$= \log \mathbb{E}_{z \sim q_\phi(z^{(i)}|x)} \left[\frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)}|x)} \right]$$

Jensen's Inequality states that,

$$\log \mathbb{E}[x] \geq \mathbb{E}[\log(x)]$$

$$\therefore \geq \mathbb{E}_{z \sim \phi(z^{(i)}/x)} \left[\log \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(1)}/x)} \right]$$

$$\geq L_1(x)$$

Hence, $\boxed{L_m(x) \geq \mathcal{L}_1(x)}$

There are two intuitions as well for this:

1. The more samples we draw from $q_\phi(z|x)$, the better estimator it is for the actual $q_\phi(z|x)$ and hence while training we can bring it closer to the actual $q_\phi(z)$.

2. If we could draw all the samples from z i.e $m \to \infty$, we could compute the actual $p_\theta(x)$ and hence the $\mathcal{L}_m(x) \to 0$. So higher the value of m, lesser is $\mathcal{L}_m(x)$.