Problem 1

The sampling trick is that in the VAE, we generate the samples by

$$z_i = \mu + \Sigma^{\frac{1}{2}} \cdot \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, I) \tag{1}$$

to avoid the error in gradient calculation.

The empirical distribution, i.e. the prior distribution is

$$q_\phi(z|x) \tag{2}$$

The KL divergence between the prior distribution from the real distribution is that

$$\begin{aligned} D_{KL}(q_\phi(z|x) \parallel p(z)) =& \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(x)}\right] \\ =& \frac{1}{2}\left[\text{tr}\left(\Sigma^{*-1}\Sigma_\phi\right) + (\mu^* - \mu_\phi(z))^T\Sigma^{*-1}(\mu^* - \mu_\phi(z)) - d + \log\frac{\det \Sigma^*}{\det \Sigma_\phi}\right] \end{aligned} \tag{3}$$

And for the sample points z_i, the divergence becomes

$$\begin{aligned} D_{KL}(q_\phi(z|x) \parallel p(z|x)) \approx& \frac{1}{K}\sum_{k=1}^{K}\log\frac{q_\phi(z^{(k)}|x)}{p(z^{(k)})} \quad \text{where} \quad z^{(k)} \sim q_\phi(z|x) \\ =& \frac{1}{2K}\sum_{k=1}^{K}\left[\text{tr}\left(\Sigma^{*-1}\Sigma_\phi\right) + (\mu^* - \mu_\theta(z))^T\Sigma^{*-1}(\mu^* - \mu_\phi(z))\right] \\ &+ \frac{1}{2K}\sum_{k=1}^{K}\left[-d + \log\left(\frac{\det \Sigma^*}{\det \Sigma_\phi}\right)\right] \\ =& \frac{1}{2K}\sum_{k=1}^{K}(\mu^* - \mu_\phi(z))^T\Sigma^{*-1}(\mu^* - \mu_\phi(z)) + C \end{aligned} \tag{4}$$

where

$$C = \frac{1}{2K}\sum_{k=1}^{K}\left[\text{tr}\left(\Sigma_\phi^{-1}\Sigma_\theta\right) + -d + \log\left(\frac{\det \Sigma_\phi}{\det \Sigma_\theta}\right)\right] \tag{5}$$

is a constant.

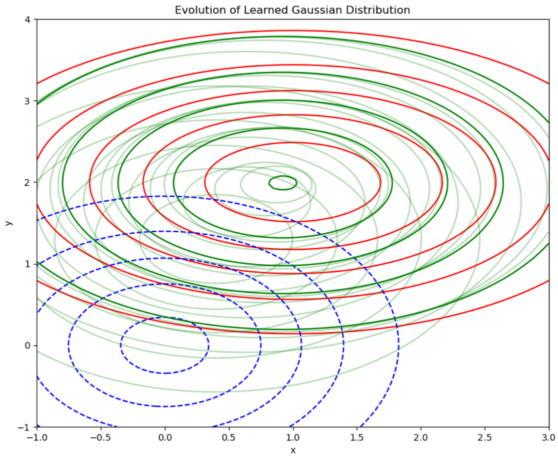
The gradient of KL divergence w.r.t. μ is

$$\begin{aligned} &\frac{\partial}{\partial \mu}D_{KL}(q_\phi(z|x) \parallel p(z)) \\ =& -\Sigma^{*-1}(\mu^* - \mu) \end{aligned} \tag{6}$$

and the gradient w.r.t. Σ is

$$\begin{aligned} &\frac{\partial}{\partial \Sigma}D_{KL}(q_\phi(z|x) \parallel p(z)) \\ =& \frac{1}{2}\frac{\partial}{\partial \Sigma}\left[\text{tr}\left(\Sigma^{*-1}\Sigma\right) - \log\det\Sigma\right] \\ =& \frac{1}{2}(\Sigma^{*-1} - \Sigma^{-1}) \end{aligned} \tag{7}$$

The variation results are shown as below:



Problem 2

An encoder E_θ serves the reconstruction by simply minimizing the reconstruction loss:

$$\mathcal{L}(\theta, \phi) = \|x - D_\phi(E_\theta(x))\|^2 \tag{8}$$

and the training is based on the training set, which might cause "overfitting" (not real overfitting) leading to poorly generated new samples.

The VAE has the object of

$$\begin{aligned} L_{ELBO}(x, \theta, \phi) =& \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)}{q_\phi(z|x)}\right] \\ =& \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) + \log p(z) - \log q_\phi(z|x)\right] \\ =& \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\big(q_\phi(z|x) \parallel p(z)\big) \end{aligned} \tag{9}$$

where the first term is the reconstruction loss, and the second term is the prior regularization. The second term is how VAE solved the problem: by adding the regularization term, VAE is capable of avoiding the overfitting on the training set in the latent space, and was able to give a good result.

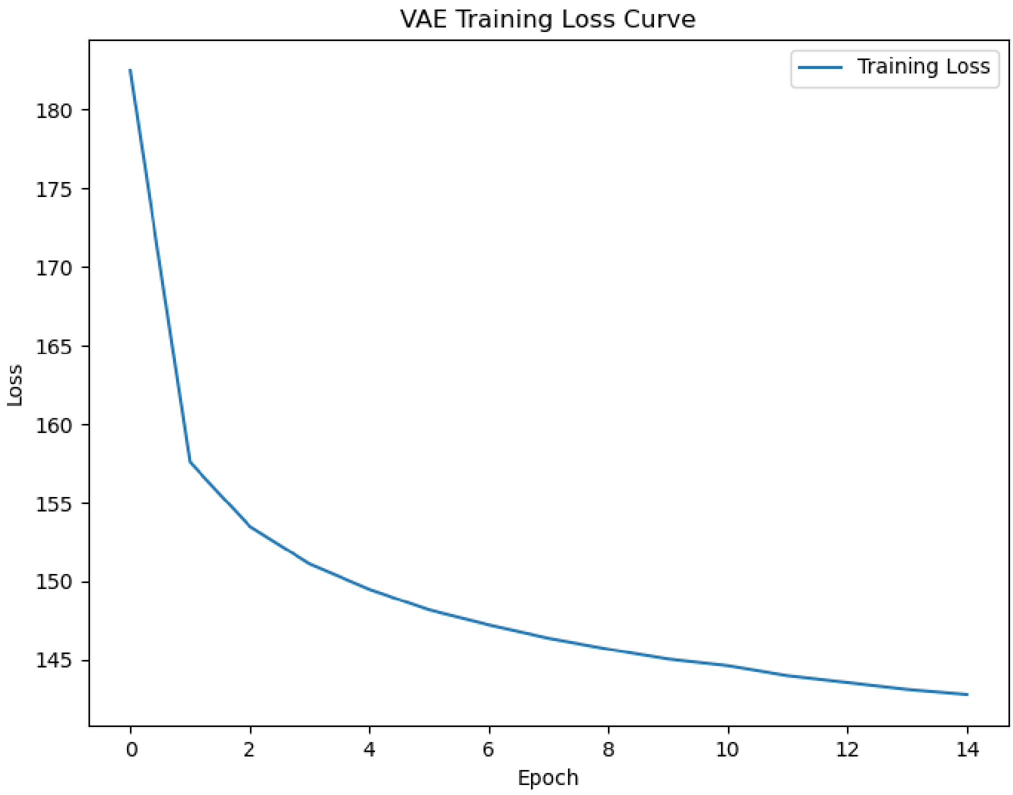
Problem 3

$$\begin{aligned} & \min \mathbb{E}_{x_t\sim p_t}\left[\|s_\theta(x_t, t) - x_{t-1}\|^2\right] \\ =& \min \int p(x_t|x_{t-1})\left[\|s_\theta\|^2 - 2s_\theta x_{t-1} + \|s_{t-1}|^2\right] dx_t \\ \iff & \min \iint p(x_t|x_{t-1})p(x_{t-1}|x_t)\left[\|s_\theta(x_t, t)\|^2 - 2s_\theta x_{t-1} + \|x_{t-1}\|^2\right] dx_t dx_{t-1} \\ =& \min \mathbb{E}_{x_t\sim p_t}[\|s_\theta(x_t, t)\|^2] + \min \iint p(x_{t-1}|x_t)p_t\left[-2s_\theta x_{t-1} + \|x_{t-1}\|^2\right] dx_{t-1}dx_t \\ =& min\mathbb{E}_{x_t\sim p_t}[\|s_\theta(x_t, t)\|^2] + \min \int \left[-2s_\theta\mathbb{E}[x_{t-1}|x_t] + \|\mathbb{E}[x_{t-1}|x_t]\|^2\right] p_t dx_t \\ =& \min \mathbb{E}_{x_t\sim p_t}\left[\|s_\theta(x_t, t) - \mathbb{E}[x_{t-1}|x_t]\|^2\right] \end{aligned} \tag{10}$$

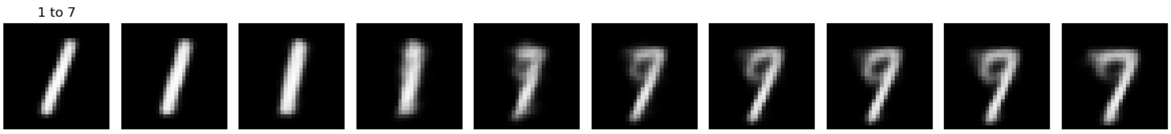
Q.E.D.

Problem 4

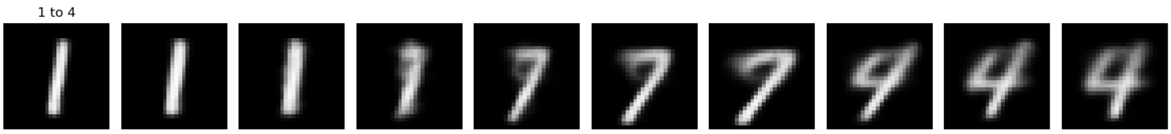
The loss curve (15 epoches)



The interpolation between 1 and 7:



The interpolation between 1 and 4:



We can clearly find that the characters of the generated figures varies continuously, which means that the model has learned the structure of the import data. Note that 7 shows in the interpolation between 1 and 4, indicating the continuous coordinates in the latent space contains the information of the similarity of the handwritten numbers.

The 2-D embedding space:

Reconstructed Images in Latent Space