Shubham Shrivastava email: shubhams@stanford.edu

Problem 1: Implementing the Variational Autoencoder (VAE)

3. Metrics on Test Samples

NELBO: 102.23046112060547
Reconstruction Loss: 83.24808502197266
KL Divergence: 18.982376098632812

4. Visualization of 200 digits

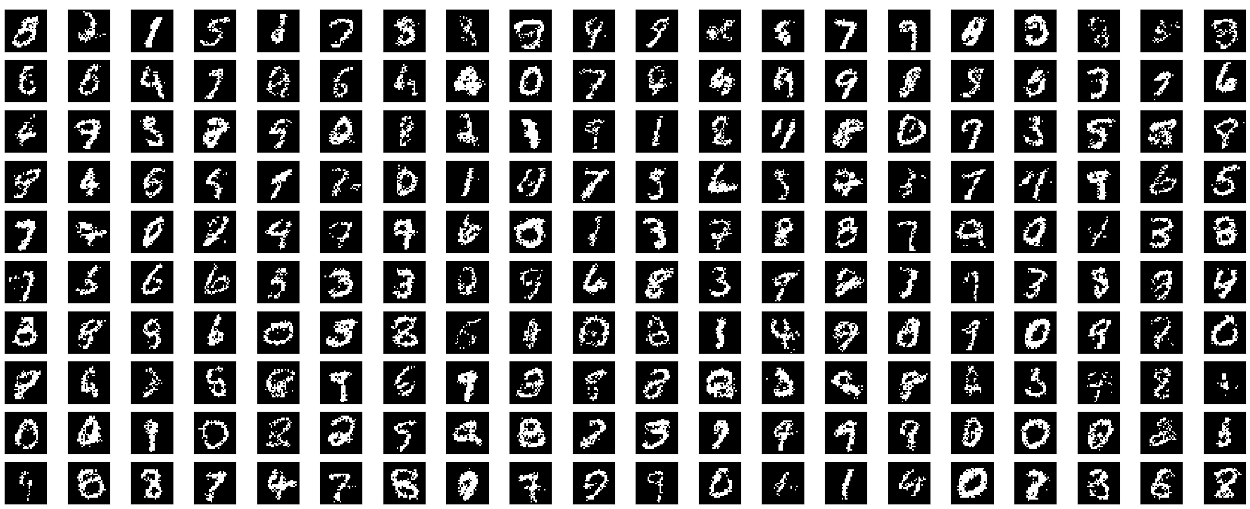

Figure 1: Samples generated from VAE trained with MNIST dataset

5. β-VAE

For any positive value of β, the objective function will dictate how important it is for the variational approximation, $q_\phi(z|x)$, to very closely match the assumed latent gaussian distribution $p(z)$. For $\beta = 1$, this takes the exact form of a normal VAE, whereas for a value of β greater than 1, the encoder output $q_\phi(z|x)$, starts looking more and more like the prior, $p(z)$. $\beta > 1$ introduces higher bottleneck constraint over the latent representation by restricting its distribution, which in effect encourages a more efficient encoding and helps the representation to be disentangled.

Problem 2: Implementing the Mixture of Gaussians VAE (GM-VAE)

1. Implement the (1) log normal and (2) log normal mixture

$\boxed{P2}$

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

log normal:

$$\log \mathcal{N}(x|\mu,\sigma^2) = \log\left(\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)\right) - \log\left(\sqrt{2\pi\sigma^2}\right)$$

$$= \frac{-(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

$$= \frac{-(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\left[\log 2\pi + 2\log\sigma\right]$$

$$= \frac{-(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log\sigma$$

log normal <u>mixture</u>:

$$\log \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(x|\mu_i,\sigma_i^2) = \log \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(x|\mu_i,\sigma_i^2)$$

$$= \log \frac{1}{K} \sum_{i=1}^{K} \exp\left(\log \mathcal{N}(x|\mu_i,\sigma_i^2)\right)$$

$$= \log\left[\sum_{i=1}^{K} \exp\left(\log \mathcal{N}(x|\mu_i,\sigma_i^2)\right)\right] - \log K$$

$$= \log\left[\sum_{i=1}^{K} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log\sigma\right)\right] - \log K$$

2. Metrics on Test Samples

NELBO: 98.30365753173828
Reconstruction Loss: 80.51519775390625
KL Divergence: 17.788480758666992

3. Visualization of 200 digits

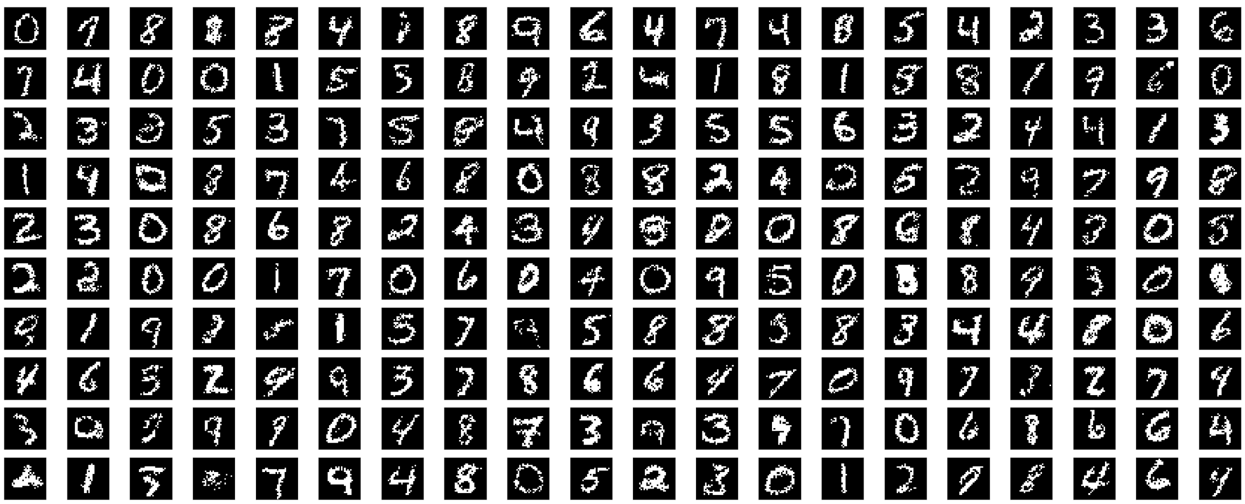

Figure 2: Samples generated from GMVAE trained with MNIST dataset

Problem 3: Implementing the Importance Weighted Autoencoder (IWAE)

1. Proof of IWAE valid lower bound of the log-likelihood

$\boxed{P3}$ In importance sampling the likelihood function $p_\theta(x)$ can be given as:

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

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

This can be approximated as:

$$p_\theta(x) = \mathbb{E}_{z \sim q(z)} \left[\frac{p_\theta(x, z)}{q(z)} \right] \approx \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(x, z^{(i)})}{q(z^{(i)})}$$

For IWAE, the log-likelihood can be estimated as:

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

$$= \log\left(\mathbb{E}_{z^{(1)}, z^{(2)}, \ldots, z^{(m)} \overset{iid}{\sim} q_\phi(z|x)} \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(x, z^{(i)})}{q(z^{(i)}|x)} \right)$$

i

Using Jenson's inequality, we can write:

$$\log(p_\theta(x)) \geq \underbrace{E_{z^{(1)}, z^{(2)}, \ldots, z^{(m)} \overset{iid}{\sim} q_\phi(z|x)} \log\left(\frac{1}{m}\sum_{i=1}^{m} \frac{p_\theta(x, z^{(i)})}{q(z^{(i)}|x)}\right)}_{\mathcal{L}_m(x;\theta,\phi)}$$

$$\Rightarrow \log(p_\theta(x)) \geq \mathcal{L}_m(x;\theta,\phi) \quad \text{ii}$$

We can further use Jenson's inequality to get:

$$\mathcal{L}_m(x;\theta,\phi) \geq \underbrace{E_{z^{(1)}, \ldots, z^{(m)} \overset{iid}{\sim} q_\phi(z|x)} \left(\frac{1}{m}\sum_{i=1}^{m} \log \frac{p_\theta(x, z^{(i)})}{q(z^{(i)}|x)}\right)}_{E_{z \sim q_\phi(z|x)}\left(\log \frac{p_\theta(x,z)}{q(z|x)}\right)}$$

$$= \mathcal{L}_1(x)$$

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

3. IWAE bounds for VAE [m = {1, 10, 100, 1000}]

NELBO: 101.51044464111328. KL: 19.448644638061523. Rec: 82.06179809570312
Negative IWAE-1: 101.43938446044922
Negative IWAE-10: 98.5534439086914
Negative IWAE-100: 97.41525268554688
Negative IWAE-1000: 96.80988311767578

4. IWAE bounds for GMVAE [m = {1, 10, 100, 1000}]

NELBO: 98.30223083496094. KL: 17.77159881591797. Rec: 80.5306167602539
Negative IWAE-1: 98.31681060791016
Negative IWAE-10: 95.99824523925781
Negative IWAE-100: 95.24256896972656
Negative IWAE-1000: 94.83466339111328

Comparing IWAE for VAE and GMVAE, we find that the IWAE in general for every value of m is lower for GMVAE. This means that with GMVAE, (1) the estimated lower bound (ELBO) in general is higher than that of VAE for all values of m, (2) as we increase the number of samples for importance sampling (m), the IWAE bound increases, which means that the variational posterior, $q_\phi(z|x)$ gets closer to the true posterior, $p_\theta(z|x)$, with increasing number of samples, m.

Problem 4: Implementing the Semi-Supervised VAE (SSVAE)

1. Test classification accuracy in supervised setting

Test set classification accuracy: 0.7339000105857849

3. Test classification accuracy in semi-supervised setting

Test set classification accuracy: 0.937999963760376

Bonus: Style and Content Disentanglement in SVHN

1. Derivation of the Evidence Lower Bound

$\boxed{\text{Bonus}}$

1

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

$$P_\theta(x) = \sum_z p(z)\, p_\theta(x|z)$$

If x is conditioned on y:

$$P_\theta(x|y) = \sum_z p(z)\, p_\theta(x|z, y) \quad \text{— (i)}$$

The expression in equation (i) is intractable, hence we introduce an amortized inference model $q_\phi(z|x, y)$ to sample z.

$$\Rightarrow P_\theta(x|y) = \sum_{z \sim q_\phi(z|x,y)} q_\phi(z|x, y) \cdot \frac{p(z) \cdot p_\theta(x|z, y)}{q_\phi(z|x, y)}$$

$$= \mathbb{E}_{q_\phi(z|x,y)}\left(\frac{p(z) \cdot p_\theta(x|z, y)}{q_\phi(z|x, y)} \right) \quad \text{— (ii)}$$

Log-likelihood of the conditional can be given as:

$$\log P_\theta(x|y) = \log\left(E_{q_\phi(z|x,y)}\left(\frac{p(z)\cdot p_\theta(x|z,y)}{q_\phi(z|x,y)}\right)\right) \quad \text{(iii)}$$

Using Jenson's inequality:

$$\log P_\theta(x|y) \geq E_{q_\phi(z|x,y)}\left(\log \frac{p(z)\cdot p_\theta(x|z,y)}{q_\phi(z|x,y)}\right) \quad \text{(iv)}$$

RHS of equation (iv) serves as the evidence lower bound and can be further expanded as:

$$E_{q_\phi(z|x,y)}\left(\log \frac{p(z)\cdot p_\theta(x|z,y)}{q_\phi(z|x,y)}\right) =$$

$$E_{q_\phi(z|x,y)}\left(\log p_\theta(x|z,y)\right) + E_{q_\phi(z|x,y)} \log \frac{p(z)}{q_\phi(z|x,y)}$$

$$= \underbrace{\underbrace{E_{q_\phi(z|x,y)}\left(\log p_\theta(x|z,y)\right)}_{\text{Reconstruction objective}} - E_{q_\phi(z|x,y)} \log \frac{q_\phi(z|x,y)}{p(z)}}_{}$$

$$= \underbrace{\underbrace{E_{q_\phi(z|x,y)}\left(\log p_\theta(x|z,y)\right)}_{\text{Reconstruction objective}} - D_{KL}\left(q_\phi(z|x,y) \,||\, p(z)\right)}_{\text{ELBO}}$$

2. Visualization of 200 SVHN digits

Figure 3: Samples (Style and Content Disentangled) generated from FSVAE trained with SVHN dataset