CS 236 Autumn 2019/2020 Homework 2

SUNet ID: 06009508

Name: Brandon McKinzie

Collaborators: N/A

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

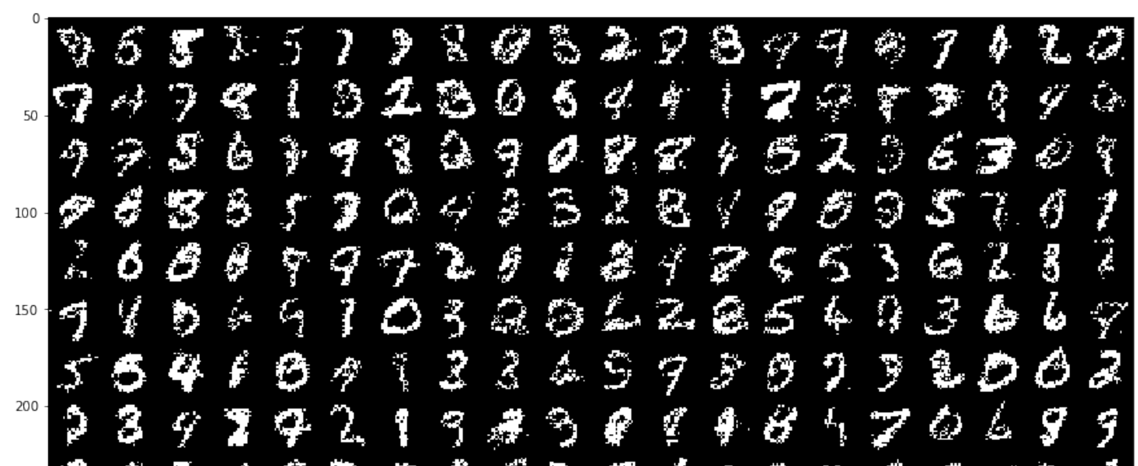
Problem 1: Implementing the Variational Autoencoder (VAE)

3. *Report the three numbers you obtain as part of the write-up.*

My numbers for the log-likelihood lower bounds on the test subset are reported below.

- **NELBO**: 100.8358154296875
- **KL**: 19.305727005004883
- **Rec**: 81.53005981445312

5. *Visualize 200 digits.*

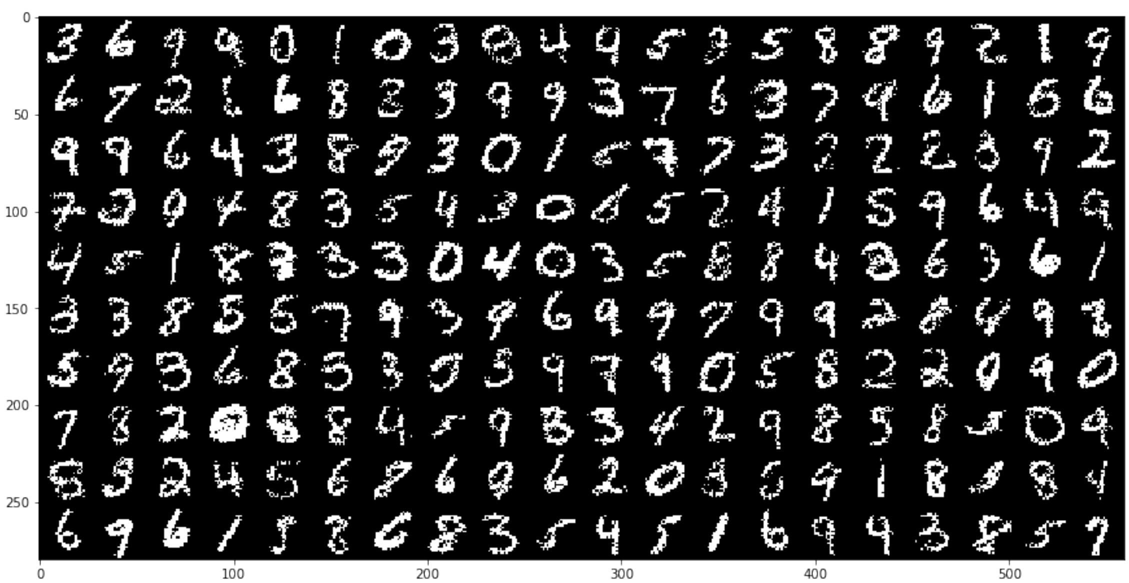


Problem 2: Implementing the Mixture of Gaussians VAE (GMVAE)

2. My numbers for the log-likelihood lower bounds on the test subset are reported below.

- **NELBO**: 97.71849060058594
- **KL**: KL: 17.689722061157227
- **Rec**: 80.02876281738281

3. *Visualize 200 digits.*



Problem 3: IWAE

1. *Prove that IWAE is a valid lower bound of the log-likelihood, and that the ELBO lower bounds IWAE*

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}_m(\boldsymbol{x}) \geq \mathcal{L}_1(\boldsymbol{x}) \tag{1}$$

for any $m \geq 1$.

The IWAE bound is defined as

$$\mathcal{L}_m(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}^{(i)})}{q_\phi(\boldsymbol{z}^{(i)} \mid \boldsymbol{x})} \right] \tag{2}$$

which for $m = 1$ reduces to the standard ELBo:

$$\mathcal{L}_1(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{\boldsymbol{z}^{(1)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}^{(1)})}{q_\phi(\boldsymbol{z}^{(1)} \mid \boldsymbol{x})} \right] \tag{3}$$

Jensen's inequality tells us that

$$\log \left(\mathbb{E} \left[x \right] \right) \geq \mathbb{E} \left[\log x \right] \tag{4}$$

and more generally, that the logarithm of any *convex combination* of x is greater than or equal to that convex combination over the logarithm of x. Any simple average, such as the average over the m unnormalized densities above, is a convex combination.

First, note that an expectation, taken over m i.i.d. samples, of an average over those samples, is equal to the expectation taken over single samples of the quantity being averaged over. Formally,

$$\mathbb{E}_{x^1, x^2, \ldots, x_m \sim p(x)} \left[\frac{1}{m} \sum_{i=1}^{m} f(x^i) \right] = \mathbb{E}_{x \sim p(x)} \left[f(x) \right] \tag{5}$$

which is really just another way of stating the fact that the Monte-Carlo average is an unbiased estimator.

Next, we use Jensen's inequality to show that $\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}_m(\boldsymbol{x})$ for $m \geq 1$:

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) &= \log \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] && (6)\\
&= \log \mathbb{E}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(i)})}{q_\phi(\boldsymbol{z}^{(i)}\mid\boldsymbol{x})}\right] && (7)\\
&\geq \mathbb{E}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{1}{m}\sum_{i=1}^{m}\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(i)})}{q_\phi(\boldsymbol{z}^{(i)}\mid\boldsymbol{x})}\right] && (8)\\
&= \mathcal{L}_m(\boldsymbol{x};\theta,\phi) && (9)
\end{aligned}$$

which proves that $\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}_m(\boldsymbol{x})$ for $m \geq 1$.

Next, we need to show that $\mathcal{L}_m(\boldsymbol{x}) \geq \mathcal{L}_1(\boldsymbol{x})$ for $m \geq 1$. To do this, note that, by definition of the uniform distribution over integers $1 \leq j \leq m$,

$$\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}^{(i)}) = \mathbb{E}_{i\sim U(1..m)}\left[f(\boldsymbol{z}^{(i)})\right] \tag{10}$$

We can use this, combined with Jensen's inequality, to show

$$\begin{aligned}
\mathcal{L}_m(\boldsymbol{x};\theta,\phi) &= \mathbb{E}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{1}{m}\sum_{i=1}^{m}\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(i)})}{q_\phi(\boldsymbol{z}^{(i)}\mid\boldsymbol{x})}\right] && (11)\\
&= \mathbb{E}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \mathbb{E}_{j\sim U(1..m)}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(j)})}{q_\phi(\boldsymbol{z}^{(j)}\mid\boldsymbol{x})}\right]\right] && (12)\\
&\geq \mathbb{E}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\mathbb{E}_{j\sim U(1..m)}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(j)})}{q_\phi(\boldsymbol{z}^{(j)}\mid\boldsymbol{x})}\right]\right] && (13)\\
&= \mathbb{E}_{\boldsymbol{z}^{(1)}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}^{(1)})}{q_\phi(\boldsymbol{z}^{(1)}\mid\boldsymbol{x})}\right] && (14)\\
&= \mathcal{L}_1(\boldsymbol{x};\theta,\phi) && (15)
\end{aligned}$$

which proves that $\mathcal{L}_m(\boldsymbol{x}) \geq \mathcal{L}_1(\boldsymbol{x})$ for $m \geq 1$

3. My numbers for the log-likelihood lower bounds on the test subset are reported below.
   - Negative IWAE-1: 100.11393737792969
   - Negative IWAE-10: 78.5806655883789
   - Negative IWAE-100: 46.388160705566406
   - Negative IWAE-1000: 45.54800796508789

4. My numbers for the log-likelihood lower bounds on the test subset are reported below.
   - Negative IWAE-1: 97.7275619506836
   - Negative IWAE-10: 77.10411071777344
   - Negative IWAE-100: 43.673885345458984
   - Negative IWAE-1000: 43.121673583984375

The IWAE bounds for GMVAE have the same trend as VAE: increasing the number of importance samples m decreases the NIWAE. The numbers above also confirm that, for $m = 1$, the NIWAE-1 values match the associated NELBo from the previous question.

Problem 4: SSVAE

1. My classification accuracy on the test set: 0.7531999945640564

3. My classification accuracy on the test set: 0.9271000027656555

Problem 5: SVHN

Since fully-supervised VAE (FSVAE) always conditions on an observed y in order to generate the sample x, it is a special case of the conditional variational autoencoder. Derive the Evidence Lower Bound $\mathcal{L}(\boldsymbol{x};\theta,\phi,y)$ of the conditional log probability $\log p_\theta(x|y)$. You are allowed to introduce the amortized inference model $q_\phi(z|x,y)$.

The model defines the distribution[1]

$$p_\theta(\boldsymbol{x} \mid y) = \int p_\theta(\boldsymbol{x},\boldsymbol{z} \mid y)\mathrm{d}\boldsymbol{z} \tag{16}$$

$$= \int p(\boldsymbol{z} \mid y)p_\theta(\boldsymbol{x} \mid y,\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{17}$$

$$= \int p(\boldsymbol{z})p_\theta(\boldsymbol{x} \mid y,\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{18}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[p_\theta(\boldsymbol{x} \mid y,\boldsymbol{z})\right] \tag{19}$$

Note that $p(\boldsymbol{z} \mid y) = p(\boldsymbol{z})$ due to the independence assumptions defined by the graphical model. As in the original ELBo derivation, we proceed by acknowledging Jensen's inequality:

$$\log p_\theta(\boldsymbol{x} \mid y) = \log \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[p_\theta(\boldsymbol{x} \mid y,\boldsymbol{z})\right] \tag{20}$$

$$\geq \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log p_\theta(\boldsymbol{x} \mid y,\boldsymbol{z})\right] \tag{21}$$

$$= \mathcal{L}(\boldsymbol{x};\theta,\phi,y) \tag{22}$$

Although technically we have "derived" the ELBo as the question has asked, I'm going to assume the instructors actually want us to derive a form reminiscent of a VAE. There are many ways we can write $\mathcal{L}(\boldsymbol{x};\theta,\phi,y)$, but the form most associated with VAEs can be derived by first acknowledging that, for any valid probability distribution $q(\boldsymbol{z})$ over $\boldsymbol{z}$

$$p_\theta(\boldsymbol{x} \mid y) = \int p_\theta(\boldsymbol{x},\boldsymbol{z} \mid y)\mathrm{d}\boldsymbol{z} \tag{23}$$

$$= \int \frac{p_\theta(\boldsymbol{x},\boldsymbol{z} \mid y)}{q(\boldsymbol{z})}q(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{24}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x},\boldsymbol{z} \mid y)}{q(\boldsymbol{z})}\right] \tag{25}$$

Therefore, we can apply the exact same earlier derivation of $\mathcal{L}(\boldsymbol{x};\theta,\phi,y)$ with Jensen's

[1] I've written the prior $p(z)$ without dependence on θ because part 2 of the question defines it as such.

inequality to obtain an equivalent definition in a different form[2]:

$$\log p_\theta(\boldsymbol{x} \mid y) = \log \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{p_\theta(\boldsymbol{x}, \boldsymbol{z} \mid y)}{q(\boldsymbol{z})}\right] \tag{26}$$

$$\geq \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z} \mid y)}{q(\boldsymbol{z})}\right] \tag{27}$$

$$= \mathcal{L}(\boldsymbol{x}; \theta, \phi, y) \tag{28}$$

From basic definitions of probability[3], logarithms[4], expectation, and fractions, we know that the above is maximized when $q(\boldsymbol{z}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x}, y)$. Of course, this distribution is (potentially) different depending on $\boldsymbol{x}$ and y. Therefore, we use amortized inference as the question suggests and instead learn a parameterized $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, y)$ with parameters ϕ shared (i.e. not depending on) for all $\boldsymbol{x}, y$. Now we have the form

$$\mathcal{L}(\boldsymbol{x}; \theta, \phi, y) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z} \mid y)}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, y)}\right] \tag{29}$$

$$\tag{30}$$

Using the exact same derivations from the lectures, we know that we can also write this in the form

$$\mathcal{L}(\boldsymbol{x}; \theta, \phi, y) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, y)\right] - D_{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, y) || p(\boldsymbol{z})\right) \tag{31}$$

$$\tag{32}$$

which can interpreted from the VAE perspective with $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, y)$ representing a decoder and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, y)$ representing an encoder.

[2] Again, there are many ways of writing $\mathcal{L}(\boldsymbol{x}; \theta, \phi, y)$, but I'm providing a couple because I was docked severely on the last homework for "insufficient analysis."

[3] I'm assuming Bayes rule is obvious.

[4] I'm assuming that monotonicity of log and the fact that $\log(1) =$ is obvious.