# Final Homework

Phillip Wang pw1287

December 15, 2021

## 1 Generative Model and EBMs

### 1.

The probability density function is

$$
\begin{aligned}
p(x|\lambda) &= \sum_{k}^{K} \pi_k g(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \sum_{k}^{K} \pi_k \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k}))}{(2\pi)^{D/2}|\boldsymbol{\Sigma_k}|^{1/2}} \\
&= \frac{1}{10}\sum_{k}^{K} \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^\top (\boldsymbol{x}-\boldsymbol{\mu_k}))}{2\pi} \\
&= \frac{1}{10}\sum_{k}^{K} \frac{\exp(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu_k}\|^2))}{2\pi}
\end{aligned}
$$

### 2.

**R=10**

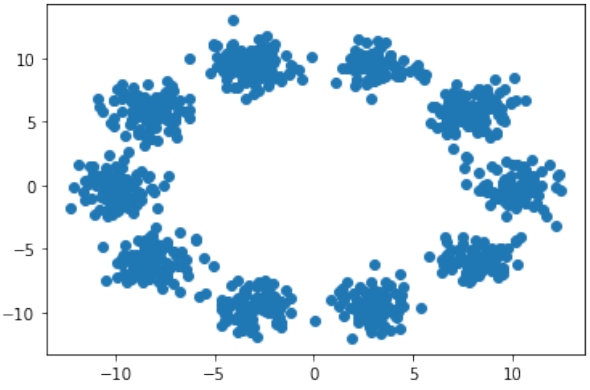
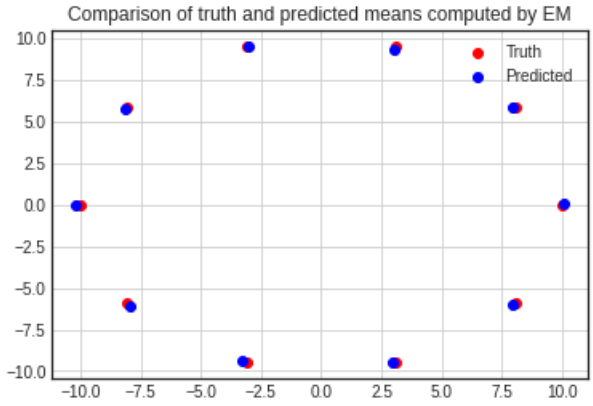
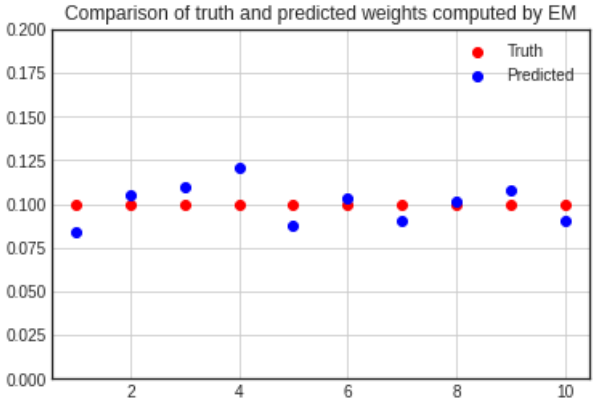



Figure 1: Sampled from Gaussian Distribution with R=10




**R=1**

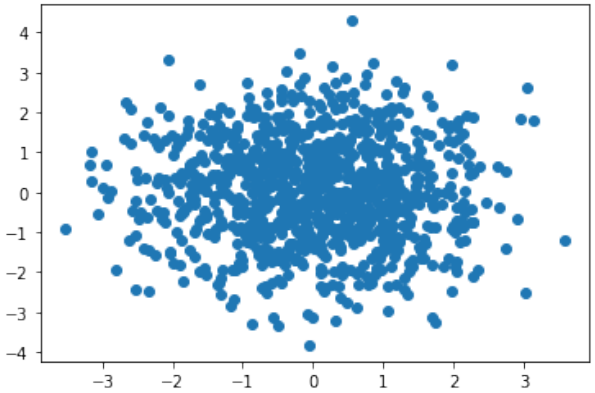
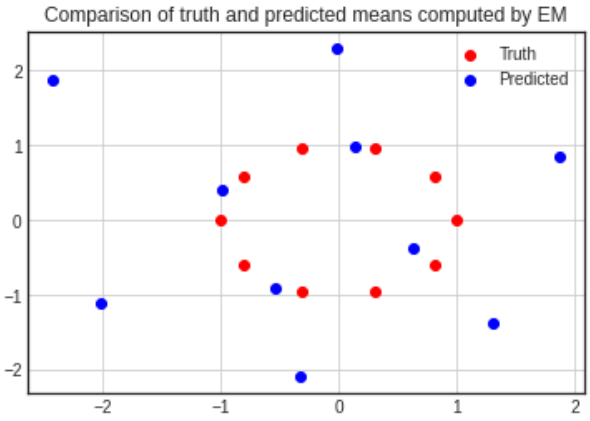
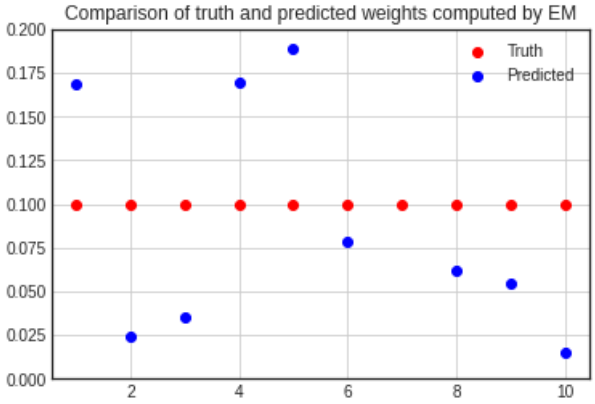


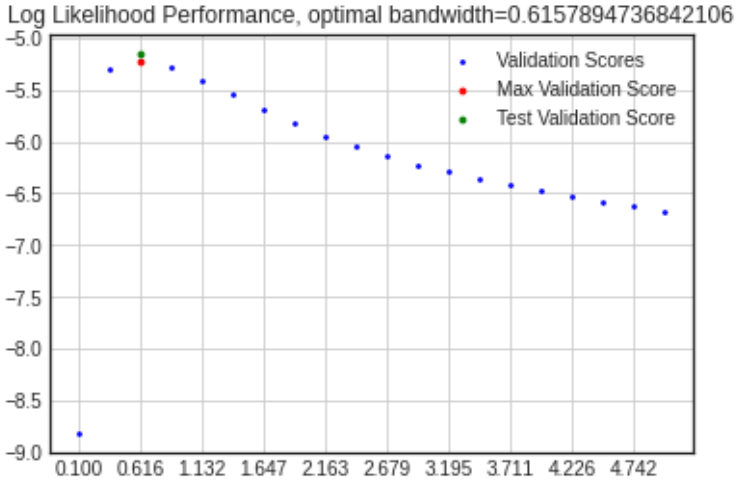

Figure 2: Sampled from Gaussian Distribution with R=1




If we compared the distributions, the deviation between the actual means/weights and the predicted means/weights computed by EM, we could see that it is hard to locate the mean and the weights in the GMM when R=1. This is because when R is large, the given clusters were able to separate from each other very well. When R is small, different components of Gaussian are mixed together and it is hard to tell each other apart. With the interference from other clusters, it is also difficult to compute the right mean.

**3.**



The above figure is the log likelihoods of the models with bandwidth($\sigma$) varying from 0.1 to 5 with 20 steps. We then evaluated the models using a validation set and found out that the optimal $\sigma$ is around 0.616. Then, the optimal estimator is picked and evaluated on the test set.

**4.**

The MLE is to maximize the log-likelihood function of

$$\log(\prod_{i=1}^{n} p_\theta(x_i)) = \sum_{i=1}^{n} \log(p_\theta(x_i)) = \sum_{i=1}^{n} E_\theta(x_i) - A(\theta)$$

This is equivalent of optimizing the loss

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} E_\theta(x_i) - A(\theta)$$

For the gradient of $\mathcal{L}(\theta)$, we have

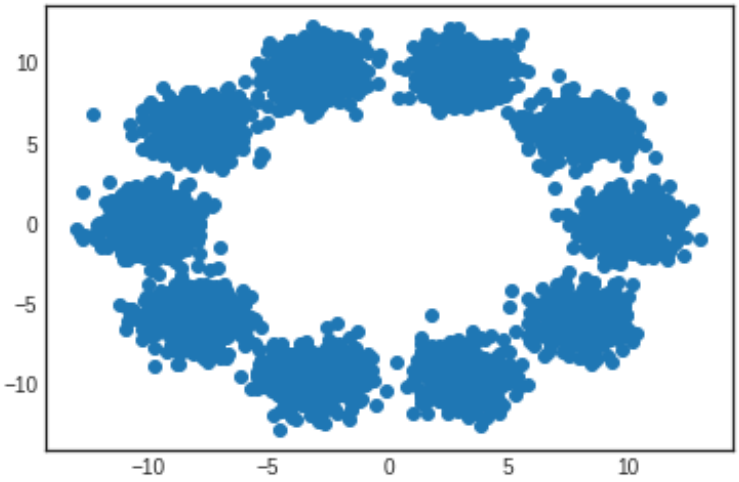
$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \nabla_\theta A(\theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \nabla_\theta \log(\int \exp(E_\theta(x)) dx) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \mathbb{E}_{x \sim p_\theta} \nabla_\theta E_\theta(x)
\end{aligned}$$

**5.**

$$\begin{aligned}
F &= \mathbb{E}_{x \sim p_\theta} f(x) \\
&= \int p_\theta(x) dx \cdot f(x) \\
&= \int q(x) dx \frac{p_\theta(x) f(x)}{q(x)} \\
&= \mathbb{E}_{x \sim q}[\frac{p_\theta(x) f(x)}{q(x)}] \\
&= \mathbb{E}_{x \sim q}[\frac{\exp(E_\theta(x) - A(\theta)) f(x)}{q(x)}] \\
&= \mathbb{E}_{x \sim q}[\frac{\exp(E_\theta(x)) f(x)}{\exp(A(\theta)) q(x)}] \\
&= \mathbb{E}_{x \sim q}[\frac{\exp(E_\theta(x)) f(x)/q(x)}{\exp(\log(\int \exp(E_\theta(x)) dx))}] \\
&= \mathbb{E}_{x \sim q}[\frac{\exp(E_\theta(x)) f(x)/q(x)}{\int \frac{\exp(E_\theta(x)) q(x)}{q(x)} dx}] \\
&= \mathbb{E}_{x \sim q}[\frac{f(x) \exp(E_\theta(x))/q(x)}{\exp(E_\theta(x))/q(x)}] \\
&= \frac{\mathbb{E}_{x \sim q}[f(x) \exp(E_\theta(x))/q(x)]}{\mathbb{E}_{x \sim q}[\exp(E_\theta(x))/q(x)]}
\end{aligned}$$

### 6.

Since computing $A(\theta)$ is intractable in this case, we use important sampling $\hat{F}$ to estimate its gradients. We used a Energy Based Model consisting of three Linear() layers and two ReLU() layers and we iterate through every parameter to accumulate our estimation of $\hat{F}$ and we manually update the model parameters using the gradients. The distribution of $X_m$ that is used for important sampling is as follows




### 7.

We compute the sum of log probabilities from the Energy-based model and the kernel density, then sample from the same Gaussian Mixture Model another 1000 points as a testing set. After training, we evaluated the EBM and Kernel Density model on the testing set. The energies output by EBM is converted to log likelihoods using Soft(arg)Max() and averaged on all samples. The total likelihoods from KD is also divided by the number of samples as well.

EBM achieved -7.95 score while KD achieved -5.22 score. In this case, Kernel Density performs better than EBM in maximizing the log likelihoods of the points from the target distribution. This might because that our EBM needs more epochs to train, now it takes the model a long time to converge because we go through the whole dataset to accumulate the estimation of $A(\theta)$ by accessing each parameter individually. We could improve the performance by doing this in batch. Also, the EBM does not pushes the energy high enough for either the center or the area outside of the Gaussian ring.

## 2 Variational Inference

### 1.

$$\begin{aligned}\mathcal{L}(q,\theta) &= \int_{\mathcal{Z}} \log(\frac{p(x,z|\theta)}{q(z)})q(z)dz \\ &= \int_{\mathcal{Z}} (\log(p(x,z|\theta)) - \log(q(z)))q(z)dz \\ &= \int_{\mathcal{Z}} \log(p(x,z|\theta))q(z)dz - \int_{\mathcal{Z}} \log(q(z))q(z)dz \end{aligned} \tag{1}$$

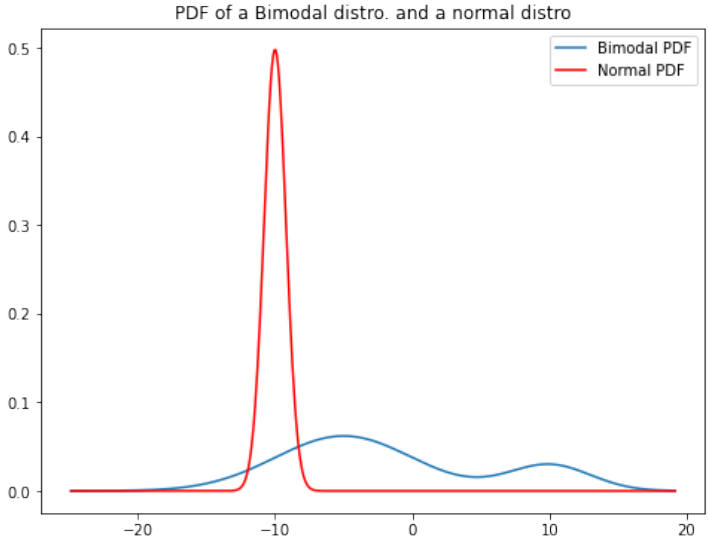
$$\begin{aligned}D_{KL}(q||p(z|x,\theta)) &= \int_{\mathcal{Z}} q(z)\log(\frac{q(z)}{p(z|x,\theta)})dz \\ &= \int_{\mathcal{Z}} q(z)\log(q(z))dz - \int_{\mathcal{Z}} q(z)\log(p(z|x,\theta))dz \end{aligned} \tag{2}$$

From equation 1 and 2, we get

$$\begin{aligned}
\mathcal{L}(q,\theta) + D_{KL}(q||p(z|x,\theta)) &= \int_{\mathcal{Z}} \log(p(x,z|\theta))q(z)dz - \int_{\mathcal{Z}} \log(q(z))q(z)dz + \int_{\mathcal{Z}} q(z)\log(q(z))dz - \int_{\mathcal{Z}} q(z)\log(p(z|x,\theta))dz \\
&= \int_{\mathcal{Z}} \log(p(x,z|\theta))q(z)dz - \int_{\mathcal{Z}} q(z)\log(p(z|x,\theta))dz \\
&= \int_{\mathcal{Z}} \log(p(x,z|\theta))q(z)dz - \int_{\mathcal{Z}} q(z)\log(\frac{p(z,x|\theta)}{p(x|\theta)})dz \\
&= \int_{\mathcal{Z}} \log(p(x,z|\theta))q(z)dz - \int_{\mathcal{Z}} q(z)\log(p(z,x|\theta))dz + \int_{\mathcal{Z}} q(z)log(p(x|\theta))dz \\
&= log(p(x|\theta)) \int_{\mathcal{Z}} q(z)dz = log(p(x|\theta)) \\
&\implies \mathcal{L}(q,\theta) = log(p(x|\theta)) - D_{KL}(q||p(z|x,\theta)) \qquad (3)
\end{aligned}$$

As we see in eq. 3, maximizing $\mathcal{L}(q,\theta)$ is equivalent of minimizing $D_{KL}(q||p(z|x,\theta))$.

## 2.




The blue bimodal distribution $\mathcal{D}_1$ is constructed by using two normal distributions, $\mathcal{D}_1' \sim \mathcal{N}(\mu = -5, \sigma^2 = 5)$ and $\tilde{\mathcal{D}}_1 \sim \mathcal{N}(\mu = 10, \sigma^2 = 3)$. The red normal distribution is just $\mathcal{D}_2 \sim \mathcal{N}(\mu = -10, \sigma^2 = 0.8)$. Such large divergence is obtained by having $\mathcal{D}_2$ reaching a high probability density on locations where $\mathcal{D}_1$ has low probability density. If we shift the red normal distribution toward 0, the ratio between two KL Divergence shrinks.

If $p = PDF(\mathcal{D}_1), q = PDF(\mathcal{D}_2)$, we have $D_{KL}(p||q) = 97.77 >> D_{KL}(q||p) = 2.10$. In order to achieve a tighter bound, our variational distribution should be resemble the prior distribution so that the Kullback-Leibler divergence is minimized.

## 3.

### Choice 1

We can first establish that $p(x,z|\theta)p(\theta) = p(x|z,\theta)p(z,\theta) = p(x,z,\theta) \implies \frac{p(x,z|\theta)}{p(x|z,\theta)} = \frac{p(z,\theta)}{p(\theta)} = p(z|\theta)$, then we have

$$\begin{aligned}
\mathcal{L}(q,\theta) &= \mathbb{E}_q \log(\frac{p(x,z|\theta)}{q(z)}) \\
&= \mathbb{E}_q \log(\frac{p(x|z,\theta)p(z|\theta)}{q(z)}) \\
\text{Following our assumption in GMM, } p(z=k|\theta) &= \pi_k = \frac{1}{K} \\
\text{Also we have } q \sim Unif[1,...,K] \text{ so we have } q(z) &= \frac{1}{K} \\
\implies \mathcal{L}(q,\theta) &= \mathbb{E}_q \log(\frac{p(x|z,\theta)\frac{1}{K}}{\frac{1}{K}}) \\
&= \mathbb{E}_q \log(p(x|z,\theta) = \log(p(x|z,\theta)) \\
\text{Since } p(x|z=k) \sim \mathcal{N}(\mu_k, \Sigma_k) \text{ and in GMM we have } \Sigma_k &= \boldsymbol{I} \\
\implies \log(p(\boldsymbol{x}|z,\boldsymbol{\theta})) &= \log(\frac{1}{\sqrt{2\pi}} \exp(-\frac{||\boldsymbol{x}-\boldsymbol{\mu_k}||^2}{2})) \\
&= \log(\frac{1}{\sqrt{2\pi}}) - \frac{||\boldsymbol{x}-\boldsymbol{\mu_k}||^2}{2} \qquad (4)
\end{aligned}$$

Hence, eq. 4 estimated $\log(p(x|\theta))$ to be $-\frac{||x-\mu_k||^2}{2} + \log(\frac{1}{\sqrt{2\pi}})$. This means that when the given data point is situated to match every component of a cluster mean in the Gaussian Mixture Model, the probability that it belong to that cluster is $\frac{1}{\sqrt{2\pi}}$. As it moves further from that mean, the probability decreases.

**Choice 2**

If we have $q = p(z|x,\theta)$, then we have $D_{KL}(q||p(z|x,\theta)) = 0$. Therefore, as we have seen in the previous choice, our estimation of $\log p(x) = \mathcal{L}(q,\theta) = \log(p(x|z,\theta)$.

The variational lower bound is an estimator of $\log(p(x|\theta)$ in both cases, but in choice 2 the estimator is more(most) accurate.

## 4.

In Choice 1, as $K \to \infty$, the components means $\mu_k = (R\cos(\frac{2\pi k}{K}), R\sin(\frac{2\pi k}{K}))$ forms a circle with radius $R$ centered around (0,0). In this case, the variational lower bound is the log likelihood of picking that point for a specific Gaussian distribution where its mean is located on the circle.

In Choice 2, the variational lower bound is rid of $K$, so it does not change anything.