Deep Learning Final HW

Liam Carpenter

December 2021

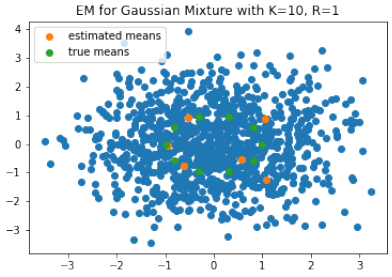
Section 1

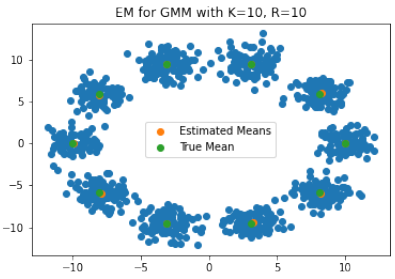
Problem 1

The density of the the GMM is

$$p(x) = p(x) = \sum_{k=1}^{10} \frac{1}{10} \mathcal{N}(x|10(cos(2\pi k/K), sin(2\pi k/K))), I)$$

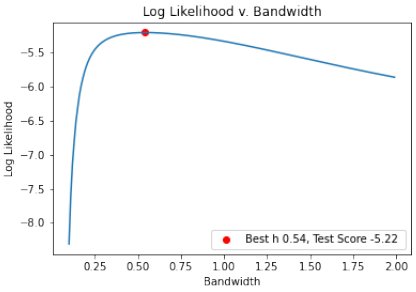
Problem 2





We can see that the EM algorithm performs better as the clusters become more differentiated from each other.

Problem 3



Problem 4

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

$$\implies \nabla_\theta \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \int \frac{\exp(E_\theta(x_i))}{\int \exp(E_\theta(x_i))dx} \nabla_\theta E(x)dx$$

$$= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(x)]$$

Problem 5

$$\mathrm{F} = \mathrm{E}_{x\sim p_\theta}[f(x)] = \int p(x)f(x)dx = \int \exp(E_\theta(x)) \exp(-\log(\int \exp(E_\theta(x)))dx) f(x)dx$$

$$= \int \frac{\exp(E_\theta(x))}{\int \exp(E_\theta(x))dx} f(x)dx = \frac{\int \exp(E_\theta(x))f(x)dx}{\int \exp(E_\theta(x))dx} = \frac{\int \frac{\exp(E_\theta(x))f(x)q(x)}{q(x)}dx}{\int \frac{\exp(E_\theta(x))q(x)}{q(x)}dx} = \frac{\mathbb{E}_q[\exp(E_\theta(x))f(x)/q(x)]}{\mathbb{E}_q[\exp(E_\theta(x))/q(x)]}$$

Problem 7

The test likelihood of our ebm is -7.7469. Noticably higher than our kernel density estimator but close.

Section 2

1

First we note,

$$\log p(x|\theta) = \mathbb{E}_q[\log p(x|\theta)]$$

From bayes rule we have that,

$$p(x) = \frac{p(x,z)}{p(z|x)}$$

so,

$$= \mathbb{E}_q[\log \frac{p(x,z|\theta)}{p(z|x,\theta)}]$$

and multiplying by $1 = \frac{q(z)}{q(z)}$

$$= \mathbb{E}_q[\log(\frac{p(x,z|\theta)}{q(z)} \frac{q(z)}{p(z|x,\theta)})]$$

$$= \mathbb{E}_q[\log \frac{p(x,z \mid \theta)}{q(z)}] + \mathbb{E}_q[\log(\frac{q(z)}{p(z,\theta)})]$$

$$\implies \log p(x \mid \theta) = \mathbb{E}_q[\log \frac{p(x,z \mid \theta)}{q(z)}] + \mathrm{KL}(q(z) \mid\mid p(z,\theta))$$

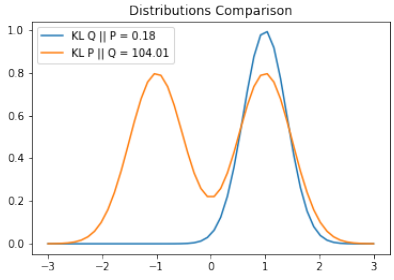
The likelihood $\mathcal{L}(q,\theta)$ is a lower bound on the log marginal likelihood of the data becaus the KL divergence is always greater than or equal to 0. That is

$$\log p(x \mid \theta) = \mathbb{E}_q[\log \frac{p(x,z \mid \theta)}{q(z)}] + \mathrm{KL}(q(z) \mid\mid p(z,\theta)) \geq \mathbb{E}_q[\log \frac{p(x,z \mid \theta)}{q(z)}]$$

The lower bound is actually achieved only when the Kullback Leibler divergence is zero, meaning our approximate distribution over Z is exactly equal to the true distribution $p(z|x,\theta)$

From this we can see the tradeoff between maximizing the variational lower bound and minimizing the KL divergence. As the variational lower bound approaches the true likelihood, the approximate variational distribution approaches the true distribution of $p(z|x,\theta)$ and the KL divergence tends to 0.

Problem 2



From this we can see that in regions where the density is very different, for a tighter bound, we want the probability attributed to that region to be low, as with the case where we are taking the KL divergence of Q with respect to P. In the bimodal case we can see that there is a high probability being attributed to this same region of low similarity results in a much higher KL divergence.

Problem 3

1. Assuming $q(z) = Unif[1, \ldots, K]$,

$$\text{L(q, } \theta) = \mathbb{E}_q[\log(\tfrac{p(x,z|\theta)}{q(z)})] = \textstyle\sum_{k=1}^{K} 1/K \log(\tfrac{p(x|z,\theta)p(z|\theta)}{\frac{1}{K}})$$
$$= \textstyle\sum_{k=1}^{K} \frac{1}{K} \log(\mathcal{N}(\mu_k, \Sigma_k)) = \sum_{k=1}^{K} \frac{1}{2K}(-\|x - \mu_k\|^2 - \log((2\pi)^2)))$$

2. Likewise assuming $q = p(z, x|\theta)$

$$\mathcal{L}(q, \theta) = \mathbb{E}_q(\log \frac{p(x, z|\theta)}{p(z|x, \theta)}) = \mathbb{E}_q(\log \frac{p(x, z|\theta)}{\frac{p(x,z|\theta)}{\sum_z p(x,z|\theta)}})$$
$$= \mathbb{E}_q(\log \frac{p(x, z|\theta)}{p(x, z|\theta)} p(x|\theta)) = \mathbb{E}_q[\log p(x|\theta)] = \log p(x|\theta)$$

As we can see when we use the true distribution for $p(z|x, \theta)$ we end up with exactly the log likelihood. Looking at the variational distribution decomposition this makes sense since the KL divergence would be zero.

Problem 4

For the second case. It is obvious that as K goes to infinity we will still end up with the log likelihood $\log p(x|\theta)$ as the dependence on the latent variables is removed. In the limit the sum becomes the integral. Integration over this would give us the lower bound.