Deep Generative models

HW 1

Yosef Shimon Sella 318731205

1.

a) Given vector $\mathrm{x} \in \mathbb{R}^n$ with covariance matrix Σ, and vector $\mathrm{w} \in \mathbb{R}^n$ such that $y = \mathrm{w}^T\mathrm{x} = \sum_i^n \mathrm{w}_i\mathrm{x}_\mathrm{i}$.

The variance of y is defined by:

$$var(y) = \mathrm{E}[(\mathrm{y}-\mathrm{E}[\mathrm{y}])(\mathrm{y}-\mathrm{E}[\mathrm{y}])^\mathrm{T}] =$$

$$= \mathrm{E}[(\mathrm{w}^T\mathrm{x}-\mathrm{E}[\mathrm{w}^T\mathrm{x}])(\mathrm{w}^T\mathrm{x}-\mathrm{E}[\mathrm{w}^T\mathrm{x}])^\mathrm{T}] = \mathrm{E}[(\mathrm{w}^T\mathrm{x}-\mathrm{w}^T\mathrm{E}[\mathrm{x}])(\mathrm{w}^T\mathrm{x}-\mathrm{w}^T\mathrm{E}[\mathrm{x}])^\mathrm{T}]$$

$$= \mathrm{E}[\mathrm{w}^T(\mathrm{x}-\mathrm{E}[\mathrm{x}])(\mathrm{x}-\mathrm{E}[\mathrm{x}])^\mathrm{T}\mathrm{w}] = \mathrm{w}^T\mathrm{E}[(\mathrm{x}-\mathrm{E}[\mathrm{x}])(\mathrm{x}-\mathrm{E}[\mathrm{x}])^\mathrm{T}]\mathrm{w} =$$

$$\mathrm{w}^T\Sigma\mathrm{w} \blacksquare$$

The passes are made from linear algebra calculations and definitions of random and scalar vectors.

b) By definition a matrix is positive semi definite if for any vector $w \neq 0$ the matrix $w^TAw \geq 0$. In part a) we've shown that for vector w that satisfies the condtion stated, the value $\mathrm{w}^T\Sigma\mathrm{w}$ is greater or equals 0, because the variance by definition is always greater or equals 0. Therefore the matrix Σ is positive semi definite.

2.

a) Lets see first $KL(p \parallel q) \geq 0$:
Jensen's Inequality states:

$$E[f(X)] \geq f(E[X])$$

For a convex function f.

Since the function $\log(x)$ is concave, the function $-\log(x)$ is convex for which Jensen's inequality holds.

Therefore we can write the following:

$$E_{x\sim p}\left[log\left(\frac{p(x)}{q(x)}\right)\right] = E_{x\sim p}\left[-log\left(\frac{q(x)}{p(x)}\right)\right] \geq -\log\left(E_{x\sim p}\left[\frac{q(x)}{p(x)}\right]\right)$$
$$= -\log\left(\sum p(x)\frac{q(x)}{p(x)}\right) = -\log\left(\sum q(x)\right)$$

since $\sum q(x) = 1$ we get

$$-\log\left(\sum q(x)\right) = 0$$

and overall:

$$KL(p \parallel q) \geq 0 \blacksquare$$

Now lets focus on the private case, $KL(p \parallel q) = 0$ iff $p = q$:

$$KL(p \parallel q) = \sum p(x)\log(\frac{p(x)}{q(x)})$$

We can observe that the only way for the expression to be equal to 0 is iff $p = q$, since $p(x), q(x) > 0$.

$$\log\left(\frac{p(x)}{q(x)}\right)|_{p=q} = \log(1) = 0 \blacksquare$$

b) We can write the KL divergence as:

$$KL(p \parallel q) = \sum p(x)\log\left(\frac{p(x)}{q(x)}\right) = \sum p(x)\log\big(p(x)\big) - \sum p(x)\log\big(q(x)\big)$$

Therefore:

$$\underset{\theta}{\operatorname{argmin}} KL(p \parallel q_\theta) = \underset{\theta}{\operatorname{argmin}}(\sum p(x)\log\big(p(x)\big) - \sum p(x)\log\big(q_\theta(x)\big))$$
$$= \underset{\theta}{\operatorname{argmin}}(-\sum p(x)\log\big(q_\theta(x)\big)) = \underset{\theta}{\operatorname{argmin}}(-\sum \log\big(q_\theta(x)\big))$$

Which is the same as minimizing the negative log likelihood:

$$\underset{\theta}{\operatorname{argmin}}\left(\frac{1}{N}\sum -\log\big(q_\theta(x)\big)\right) = \underset{\theta}{\operatorname{argmin}}\left(-\sum \log\big(q_\theta(x)\big)\right) \blacksquare$$

c) As defined previously the KL divergence is

$$KL(p \parallel q) = \sum_{x_i} p(x_i)\log(\frac{p(x_i)}{q(x_i)})$$

In our case we will calculate the KL divergence when

$$p(x_1, \dots, x_n) = \prod_i p_i(x_i) \;, q(x_1, \dots, x_n) = \prod_i q_i(x_i)$$

The KL divergence in this case is:

$$KL\big(p(x_1, \dots, x_n) \parallel q(x_1, \dots, x_n)\big) = \sum_{x_i} p(x_1, \dots, x_n)\log\left(\frac{p(x_1, \dots, x_n)}{q(x_1, \dots, x_n)}\right)$$
$$= \sum_{x_i}(\prod_i p_i(x_i)\log\left(\frac{\prod_i p_i(x_i)}{\prod_i q_i(x_i)}\right)) = \sum_{x_i}(\prod_i p_i(x_i)\log\left(\prod_i \frac{p_i(x_i)}{q_i(x_i)}\right))$$

Using logarithm rules we get:

$$= \sum_{x_i}(\prod_i p_i(x_i)\sum_i \log\left(\frac{p_i(x_i)}{q_i(x_i)}\right))$$

We can see that for each i we fix x_i and therefore only $p_i(x_i)$ contributes to that expression for every iteration and we get:

$$= \sum_i \sum_{x_i} p_i(x_i)\log\left(\frac{p_i(x_i)}{q_i(x_i)}\right) = \sum_i KL(p_i \parallel q_i) \blacksquare$$

d) The KL divergence in the continuous case is:

$$KL(p \parallel q) = \int p(x)\log(\frac{p(x)}{q(x)})dx$$

And the probability density function of a gaussian is:

$$p(x) \sim \mathcal{N}(\mu, \sigma^2)\ ;\ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We can plug the probability density function into the KL divergence for two Gaussians

$$p(x) \sim \mathcal{N}\left(\mu_p, \sigma_p^2\right)\ ;\ q(x) \sim \mathcal{N}\left(\mu_q\ , \sigma_q^2\right)$$

And obtain:

$$\int \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{\left(x-\mu_p\right)^2}{2\sigma_p^2}} \log(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{\left(x-\mu_p\right)^2}{2\sigma_p^2}}}{\frac{1}{\sqrt{2\pi\sigma_q^2}} e^{-\frac{\left(x-\mu_q\right)^2}{2\sigma_q^2}}})dx$$

Lets simplify the term in the log:

$$\log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{\left(x-\mu_p\right)^2}{2\sigma_p^2}}}{\frac{1}{\sqrt{2\pi\sigma_q^2}} e^{-\frac{\left(x-\mu_q\right)^2}{2\sigma_q^2}}}\right) = \frac{\left(x-\mu_q\right)^2}{2\sigma_q^2} - \frac{\left(x-\mu_p\right)^2}{2\sigma_p^2} + \log(\frac{\sigma_q}{\sigma_p})$$

Now lets revisit the KL divergence using the simplified log term:

$$KL(p \parallel q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx =$$

$$\int p(x) \left[\frac{\left(x-\mu_q\right)^2}{2\sigma_q^2} - \frac{\left(x-\mu_p\right)^2}{2\sigma_p^2} + \log\left(\frac{\sigma_q}{\sigma_p}\right)\right] dx$$

Notice that the integral is the same as the expression for the expected value over $p(x)$.

$$= \mathbb{E}_p\left[\frac{\left(x-\mu_q\right)^2}{2\sigma_q^2} - \frac{\left(x-\mu_p\right)^2}{2\sigma_p^2} + \log\left(\frac{\sigma_q}{\sigma_p}\right)\right]$$

$$= \frac{1}{2\sigma_q^2}\mathbb{E}_p\left[\left(x-\mu_q\right)^2\right] - \frac{1}{2\sigma_p^2}\mathbb{E}_p\left[\left(x-\mu_p\right)^2\right] + \log\left(\frac{\sigma_q}{\sigma_p}\right)$$

By definition:

$$\mathbb{E}_p\left[\left(x-\mu_p\right)^2\right] = \sigma_p^2$$

Using mathematical manipulation we get:

$$
\begin{aligned}
\mathbb{E}_p\left[\left(x-\mu_q\right)^2\right] &= \mathbb{E}_p\left[\left(x-\mu_p+\mu_p-\mu_q\right)^2\right] \\
&= \mathbb{E}_p\left[\left(x-\mu_p\right)^2+2\left(x-\mu_p\right)\left(\mu_p-\mu_q\right)+\left(\mu_p-\mu_q\right)^2\right] \\
&= \sigma_p^2+2\cdot\underbrace{\left(\mu_p-\mu_p\right)}_{0}\cdot\left(\mu_p-\mu_q\right)+\left(\mu_p-\mu_q\right)^2
\end{aligned}
$$

$$
=\sigma_p^2+\left(\mu_p-\mu_q\right)^2
$$

Overall, we get:

$$
=\tfrac{1}{2\sigma_q^2}\left(\sigma_p^2+\left(\mu_p-\mu_q\right)^2\right)-\tfrac{1}{2\sigma_p^2}\sigma_p^2+\log\left(\tfrac{\sigma_q}{\sigma_p}\right)=
$$

$$
KL(p\parallel q)=\frac{\sigma_p^2+\left(\mu_p-\mu_q\right)^2}{2\sigma_q^2}+\log\left(\frac{\sigma_q}{\sigma_p}\right)-\frac{1}{2}\blacksquare
$$