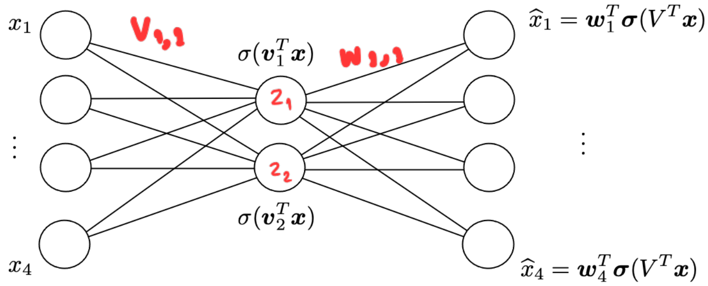
Autoencoders, Multivariate Normal KL

Submit a PDF of your answers to Canvas

1. Consider the undercomplete autoencoder shown in the figure below. In this problem, there is no activation function on the output layer (the first output is given as $\widehat{x}_1 = \boldsymbol{w}_1^T \boldsymbol{\sigma}(V^T \boldsymbol{x})$).



→ V and W V → 2×4, W → 4×2 → 16 parameters in total

a) How many parameters do you need to learn to train your autoencoder?

b) Let $w_{1,1}$ denote the weight associated with the connection between the first (top) hidden node and the first output node, and $w_{2,1}$ denote the weight associated with the bottom hidden node and the first output. Compute the partial derivative of the output $\widehat{x}_1$ with respect to weight $w_{1,1}$ when $\sigma(\cdot)$ is the logistic activation function. $\sigma(v_1^T x)$

c) Let $v_{1,1}$ denote the weight associated with the connection between the first (top) input note and the first hidden note. Compute the partial derivative of the output $\widehat{x}_1$ with respect to weight $v_{1,1}$ when $\sigma(\cdot)$ is the logistic activation function.

$\frac{d}{dv_{1,1}} \hat{x}_1 = \frac{d}{dz_1} \hat{x} \frac{d}{dv_{1,1}} z_1 = w_{1,1} (\sigma(z_1)(1-\sigma(z_1)) x_1$

d) You decide to use the squared error loss function, i.e, $\ell(\widehat{\boldsymbol{x}}, \boldsymbol{x}) = ||\widehat{\boldsymbol{x}} - \boldsymbol{x}||^2$. Compute the partial derivative of $\ell(\widehat{\boldsymbol{x}}, \boldsymbol{x})$ first with respect to $w_{1,1}$, and then with respect to $v_{1,1}$.

$\frac{\partial}{\partial v_{1,1}} \ell(\hat{x}, x) = \sum_{i=1}^{4} \left(\frac{\partial}{\partial v_{1,1}} \hat{x}_i \right) 2(x_i - \hat{x}_i) = 2 \sum_{i=1}^{4} (x_i - \hat{x}_i) w_{1,i} (\sigma(z_1)(1-\sigma(z_1)) x_i$

2. In this problem we will compute the Kullback–Leibler divergence between two multivariate normal distributions (a quantity we will use in a later lecture). For reference, recall the general form of a multivariate normal distribution:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

and the definition of KL divergence:

$$D(p(\boldsymbol{x}) || q(\boldsymbol{x})) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}.$$

Let $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $q(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Describe and justify each step in the derivation of the KL divergence between the two multivariate normal distributions:

$$D(p(\boldsymbol{x})||q(\boldsymbol{x}))$$

$$\overset{(1)}{=} \int p(\boldsymbol{x}) \left(\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \right) d\boldsymbol{x}$$

$$\overset{(2)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} E_p \left[(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right] - \frac{1}{2} E_p \left[(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \right]$$

de fn. of expectation
$E[g(x)] = \int g(x) p(x) dx$

$$\overset{(3)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} E_p \left[\mathrm{tr} \left((\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right) \right] - \frac{1}{2} E_p \left[\mathrm{tr} \left((\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \right) \right]$$

$\to$ is a scalar,
trace is the identity
transformation!

$$\overset{(4)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_2^{-1} E_p \left[(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \right] \right) - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_1^{-1} E_p \left[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \right] \right)$$

$\to$ $tr(AB) = tr(BA)$
we may remove Σ_1^{-1} out of expectation
bc. $A = (X-\mu)^T$, $B = \Sigma^{-1}(X-\mu)$

$$\overset{(5)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_2^{-1} E_p \left[(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \right] \right) - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_1 \right)$$

definition of covariance
$\Sigma_1 = E_p[(X_1 - \mu_1)^T (X_1 - \mu_1)]$

$$\overset{(6)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_2^{-1} E_p \left[\boldsymbol{x}\boldsymbol{x}^T - 2\boldsymbol{x}\boldsymbol{\mu}_2^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T \right] \right) - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_1 \right)$$

expansion

$$\overset{(7)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - 2\boldsymbol{\mu}_1\boldsymbol{\mu}_2^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T \right) \right) - \frac{1}{2} n$$

$E_p[X] = \mu_1$, $E_p[X^T X] = \Sigma_1$

$$\overset{(8)}{=} \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1 \right) + \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} n$$

arrangement of terms