1. [30 points] We want to estimate the standard deviation from dataset \( \mathcal{D} = \{x_n, n = 1, \ldots, N\} \). Suppose that each observation from independent and identically distributed Normal distribution, i.e. \( x_n \sim \mathcal{N}(\mu, \sigma^2) \). Derive formula for the following:

(a) What is the maximum likelihood estimator (MLE) for \( \mu \) and \( \sigma \)?

(b) Are they unbiased estimators? Hint: given parameter \( \theta \), an unbiased estimator \( \hat{\theta} \) should yield \( \theta = \mathbb{E}_\mathcal{D}[\hat{\theta}] \).

2. [40 points] Consider Gaussian Mixture Model \( p(x|\theta) = \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \), and the log likelihood \( \ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \). Define the posterior responsibility that cluster \( k \) has for datapoint \( x_n \) as:

\[
r_{nk} = p(z_n = k|x_n, \theta) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k'}^{K} \pi_{k'} \mathcal{N}(x_n|\mu_{k'}, \Sigma_{k'})}
\]

(a) Derive formula for the gradient of the log likelihood w.r.t. \( \mu_k \): \( \frac{\partial \ell(\theta)}{\partial \mu_k} = \ldots \)

(b) Do the same for \( \frac{\partial \ell(\theta)}{\partial \pi_k} = \ldots \)

(c) We need to fix the above formula since we need to impose the constraint \( \sum_k \pi_k = 1 \). This can be done by re-parameterizing with the softmax function \( \pi_k = \exp(w_k) / \sum_{k'} \exp(w_{k'}) \). Now, derive the formula for the gradient \( \frac{\partial \ell(\theta)}{\partial w_k} = \ldots \)

(d) Assume diagonal matrices for \( \Sigma_k \), and derive the formula for \( \frac{\partial \ell(\theta)}{\partial \Sigma_k} = \ldots \). You should be careful in making sure that \( \Sigma_k \) are positive semi-definite.

3. [30 points] Consider a two-layer network with one hidden layer, where all the activation functions are given by logistic sigmoid functions \( \sigma(a) = 1/(1+\exp(-a)) \). Show that there exists an equivalent network, which produces exactly the same network output, but with hidden unit activation functions given by \( \tanh(a) = (\exp(a) - \exp(-a))/(\exp(a) + \exp(-a)) \). Hint: the parameters of the two networks differ by linear transformations.