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 μ and σ?

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

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

$$r_{nk} = p(z_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

   (a) Derive formula for the gradient of the log likelihood w.r.t. $\boldsymbol{\mu}_k$: $d\ell(\boldsymbol{\theta})/d\boldsymbol{\mu}_k = ...$

   (b) Do the same for $d\ell(\boldsymbol{\theta})/d\pi_k = ...$

   (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 $d\ell(\boldsymbol{\theta})/dw_k = ...$

   (d) Assume diagonal matrices for Σ_k, and derive the formula for $d\ell(\boldsymbol{\theta})/d\Sigma_k = ...$. You should be careful in making sure that Σ_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.