Exercise: Gradient and Hessian of log-likelihood for multinomial logistic regression

1. Let $\mu_{ik} = \boldsymbol{\sigma}(\boldsymbol{\eta}_i)_k$, where $\boldsymbol{\eta}_i = \mathbf{w}^T \mathbf{x}_i$. Prove that the Jacobian of the softmax is

$$\frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \mu_{ik}(\delta_{kj} - \mu_{ij}) \tag{1}$$

   where $\delta_{kj} = I(k = j)$.

2. Hence show that the gradient of the NLL is given by

$$\nabla_{\mathbf{w}_c} \ell \quad = \quad \sum_i (y_{ic} - \mu_{ic}) \mathbf{x}_i \tag{2}$$

   Hint: use the chain rule and the fact that $\sum_c y_{ic} = 1$.

3. Show that the block submatrix of the Hessian for classes c and c' is given by

$$\mathbf{H}_{c,c'} \quad = \quad -\sum_i \mu_{ic}(\delta_{c,c'} - \mu_{i,c'}) \mathbf{x}_i \mathbf{x}_i^T \tag{3}$$

Hence show that the Hessian of the NLL is positive definite.