Lesson 22 (Maximimum Likelihood Estimation) Consider the simple bias network shown below and applied to the MNIST classification problem. (Recall that the MNIST classification problem consists of classifying images of the digits 0 through 9.)

$$\boxed{\mathbf{b}} \rightarrow \hat{\mathbf{P}}$$

Note that

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{p}_0 & \hat{p}_1 & \cdots & \hat{p}_9 \\ \hat{p}_0 & \hat{p}_1 & \cdots & \hat{p}_9 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{p}_0 & \hat{p}_1 & \cdots & \hat{p}_9 \end{pmatrix}_{n\times 10} .$$

In particular, $\mathbf{b} = \hat{\mathbf{p}} = (\hat{p}_0, \hat{p}_1, \ldots, \hat{p}_9)$ where $\hat{p}_0$ through $\hat{p}_9$ are the respective network output probabilities of the 10 digits 0 through 9. Let $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}$ be the targets of a training dataset of size n.

(a) Show that the cross-entropy function corresponding to the simple bias network for the training dataset targets $\mathbf{y}$ is given by

$$\mathrm{CE}(\hat{\mathbf{P}}, \mathbf{P}) = -\left[q_0 \ln(\hat{p}_0) + q_1 \ln(\hat{p}_1) + \cdots + q_9 \ln(\hat{p}_9)\right]$$

where q_0 equals the number of 0's, q_1 represents the number of 1's, etc. contained in the training set. Hint: The $\mathbf{P}$ matrix results from one-hot-encoded targets and consists only of zeros and ones.

(b) What does $q_0 + q_1 + \cdots + q_9$ equal? Explain.

(c) We have shown that minimizing the cross-entropy $CE(\hat{\mathbf{P}}, \mathbf{P})$ is equivalent to minimizing the negative log likelihood function (or equivalently maximizing the likelihood function) subject to the constraint

$$\hat{p}_0 + \hat{p}_1 + \cdots + \hat{p}_9 = 1.$$

Use Lagrange multipliers to show that the optimal values for the bias $\mathbf{b}$ is given by

$$\mathbf{b} = \hat{\mathbf{p}} = \begin{pmatrix} \hat{p}_0 & \hat{p}_1 & \cdots & \hat{p}_9 \end{pmatrix} = \begin{pmatrix} \frac{q_0}{n} & \frac{q_1}{n} & \cdots & \frac{q_9}{n} \end{pmatrix}.$$