1) a) There are $10 \times 784 = 7840$ parameters for this model.

(b) (10pts) Write down the log-likelihood and convert it into a minimization problem over the cross-entropy loss E. Derive the gradient of E with respect to each $\mathbf{w}_k$, i.e., $\nabla_{\mathbf{w}_k} E(\mathbf{w})$.

$$p(t_k = 1|\mathbf{x}, \mathbf{w}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{i=0}^{9} \exp(\mathbf{w}_i^T \mathbf{x})}$$

We can write down the likelihood function

$$p(T|x,w) = \prod_{n=1}^{N} \prod_{k=1}^{k} p(t_k = 1 | x, w)^{t_{nk}}$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{k} \left(\frac{\exp(w_k^T x_n)}{\sum_{i=0}^{9} \exp(w_i^T x_n)} \right)^{t_{nk}}$$

Then log likelihood is

$$E(w) = -\ln P(T|x,w) = -\sum_{n=1}^{N} \sum_{k=1}^{k} t_{nk} \ln(y_{nk})$$

Where $\ln(y_{nk}) = \left[w_k^T x_n - \ln\left(\sum_{i=0}^{9} \exp(w_i^T x_n) \right) \right]$

Then $\frac{\partial}{\partial w_i} \ln(y_{nk}) = x_n \cdot 1\{i=k\} - \frac{x_n \exp(w_i^T x_n)}{\sum_{j=1}^{k} \exp(w_j^T x_n)}$

$$= x_n \left[\{i=k\} - y_{ni} \right]$$

So, deriving WRT W_i:

$$\frac{\partial}{\partial w_i} -\ln P(T|x,w) = -\sum_{n=1}^{N} \sum_{k=1}^{k} t_{nk} x_n \left[\{i=k\} - y_{ni} \right]$$

$$= -\sum_{n=1}^{N} \sum_{k=1}^{k} t_{nk} x_n \cdot 1\{i=k\} + \sum_{n=1}^{N} \sum_{k=1}^{k} t_{nk} x_n y_{ni} \cdot 1\{i=k\}$$

$$= -\sum_{n=1}^{N} t_{ni} x_n + \sum_{n=1}^{N} x_n y_{ni} \qquad \text{because } \sum^{k} t_{nk} = 1$$

$$\nabla_{w_i} E(w) = \sum_{n=1}^{N} (y_{ni} - t_{ni}) x_n$$

c) Training accuracy : 89.43%

Test accuracy : 88%

Samples used : 60,000