4.3

a) The softmax cost is given by $g\left(\tilde{\mathbf{w}}\right) = \sum_{p=1}^{P} \log\left(1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}\right)$. The partial derivative of the cost with respect to the i th entry in $\tilde{\mathbf{w}}$ can be computed as

$$\frac{\partial}{\partial \tilde{w}_i} g\left(\tilde{\mathbf{w}}\right) = \sum_{p=1}^{P} \frac{\frac{\partial}{\partial \tilde{w}_i}\left(1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}\right)}{1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}}$$

$$= \sum_{p=1}^{P} \frac{\left(e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}\right) \frac{\partial}{\partial \tilde{w}_i}\left(-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}\right)}{1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}} = -\sum_{p=1}^{P} \frac{e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}}{1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}} y_p \tilde{x}_{p,i}.$$

The full gradient therefore can be written as

$$\nabla g = -\sum_{p=1}^{P} \frac{e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}}{1 + e^{-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}} y_p \tilde{\mathbf{x}}_p.$$

Multiplying both the numerator and denominator of each summand by $e^{y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}$ gives

$$\nabla g = -\sum_{p=1}^{P} \frac{1}{1 + e^{y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}}} y_p \tilde{\mathbf{x}}_p = -\sum_{p=1}^{P} \sigma\left(-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}\right) y_p \tilde{\mathbf{x}}_p.$$

b) $\nabla g = -\sum_{p=1}^{P} \sigma\left(-y_p \tilde{\mathbf{x}}_p^T \tilde{\mathbf{w}}\right) y_p \tilde{\mathbf{x}}_p$ can be written as $\tilde{\mathbf{X}}\mathbf{r}$, where

$$\tilde{\mathbf{X}} = \left[\begin{array}{cccc} \tilde{\mathbf{x}}_1 & \tilde{\mathbf{x}}_2 & \cdots & \tilde{\mathbf{x}}_P \end{array}\right],$$

and

$$\mathbf{r} = -\vec{\sigma}\left(-\left(\tilde{\mathbf{X}}^T \tilde{\mathbf{w}}\right) \odot \mathbf{y}\right) \odot \mathbf{y}.$$

In the equation above, $\odot$ denotes the Hadamard (entry-wise) product,

$$\vec{\sigma}\left(\left[\begin{array}{c} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_P \end{array}\right]\right) = \left[\begin{array}{c} \sigma\left(\zeta_1\right) \\ \sigma\left(\zeta_2\right) \\ \vdots \\ \sigma\left(\zeta_P\right) \end{array}\right],$$

and

$$\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_P \end{array}\right].$$

4.5

a) Suppose that the hyperplane $b + \mathbf{x}^T\mathbf{w} = 0$ perfectly separates the two classes of data. Thus for every $\mathbf{x}_p$ we have that $-y_p\left(b + \mathbf{x}_p^T\mathbf{w}\right) < 0$. Multiplying b and $\mathbf{w}$ by a constant $C > 1$ then gives

$$-y_p\left(bC + \mathbf{x}_p^T\mathbf{w}C\right) = C\left(-y_p\left(b + \mathbf{x}_p^T\mathbf{w}\right)\right) < -y_p\left(b + \mathbf{x}_p^T\mathbf{w}\right) < 0.$$

Denoting $-y_p\left(bC + \mathbf{x}_p^T\mathbf{w}C\right)$ by α_p, and $-y_p\left(b + \mathbf{x}_p^T\mathbf{w}\right)$ by β_p, we want to show that $\sum\limits_{p=1}^{P}\log\left(1 + e^{\alpha_p}\right) < \sum\limits_{p=1}^{P}\log\left(1 + e^{\beta_p}\right)$.

Since $\alpha_p < \beta_p$, we have that $e^{\alpha_p} < e^{\beta_p}$. Further, we can write $\log\left(1 + e^{\alpha_p}\right) < \log\left(1 + e^{\beta_p}\right)$. Summing this inequality over all p then gives the desired inequality.

b) According to part a) if we keep multiplying the hyperplane parameters by $C > 1$, the evaluation of the softmax cost will get smaller and smaller. Theoretically speaking, this means that the minimum of the softmax cost (when the data is separable) is at infinity! Practically speaking, this creates a numerical issue since every computer has a limit on the largest number that can be stored in memory.

4.12

We show that the p th summands in $g\left(b, \mathbf{w}\right)$ and $h\left(b, \mathbf{w}\right)$ are identical. Consider the following cases:

Case 1. $y_p = +1$

The p th summand in $g\left(b, \mathbf{w}\right)$ is given by

$$\log\left(1 + e^{-\left(b + \mathbf{x}_p^T\mathbf{w}\right)}\right). \tag{108}$$

With $y_p = +1$, we have $\bar{y}_p = +1$ and the p th summand of $h\left(b, \mathbf{w}\right)$ can be written as $-\bar{y}_p\log\sigma\left(b + \mathbf{x}_p^T\mathbf{w}\right) - (1 - \bar{y}_p)\log\left(1 - \sigma\left(b + \mathbf{x}_p^T\mathbf{w}\right)\right)$, or equivalently

$$-\log\,\sigma\left(b + \mathbf{x}_p^T\mathbf{w}\right). \tag{109}$$

Writing $\sigma\left(b+\mathbf{x}_p^T\mathbf{w}\right)$ as $\frac{1}{1+e^{-\left(b+\mathbf{x}_p^T\mathbf{w}\right)}}$, it is clear that the expressions in (108) and (109) are identical.

Case 2. $y_p = -1$
The p th summand in $g\left(b,\mathbf{w}\right)$ is given by

$$\log\left(1+e^{\left(b+\mathbf{x}_p^T\mathbf{w}\right)}\right). \tag{110}$$

With $y_p = -1$, we have $\bar{y}_p = 0$ and the p th summand of $h\left(b,\mathbf{w}\right)$ can be written as $-\bar{y}_p\log\sigma\left(b+\mathbf{x}_p^T\mathbf{w}\right)-(1-\bar{y}_p)\log\left(1-\sigma\left(b+\mathbf{x}_p^T\mathbf{w}\right)\right)$, or equivalently

$$-\log\ \left(1-\sigma\left(b+\mathbf{x}_p^T\mathbf{w}\right)\right). \tag{111}$$

Now note that $1-\sigma\left(b+\mathbf{x}_p^T\mathbf{w}\right) = \frac{e^{-\left(b+\mathbf{x}_p^T\mathbf{w}\right)}}{1+e^{-\left(b+\mathbf{x}_p^T\mathbf{w}\right)}} = \frac{1}{1+e^{\left(b+\mathbf{x}_p^T\mathbf{w}\right)}}$. Hence, the expressions in (110) and (111) are indeed identical.