Ex. 4.5

$$\ell(\beta) = \sum_i \left\{ y_i \beta^T x_i - \log(1 + \exp(\beta^T x_i)) \right\}$$

Now $y_i = 1$ if $x_i > x_0$
$y_i = 0$ if $x_i < x_0$ (since classes separated by x_0)

$$\Rightarrow \sum_{x_i < x_0} -\log(1+\exp(\beta^T x_i)) + \sum_{x_i > x_0} \beta^T x_i - \log(1+\exp(\beta^T x_i))$$

Now $\beta^T = (\beta_0, \beta_1)$
and notice $x_i - x_0$ is positive when $x_i > x_0$ $(*)$
and $x_i - x_0$ is negative when $x_i < x_0$

$$\ell(\beta) = \overset{①}{\sum_{x_i < x_0}} -\log(1 + \exp(\beta_0 + \beta_1 x_0 + \beta_1(x_i - x_0)))$$

$$\overset{②}{+ \sum_{x_i > x_0}} \beta_0 + \beta_1 x_0 + \beta_1(x_i - x_0) - \log(1 + \exp(\beta_0 + \beta_1 x_0 + \beta_1(x_i - x_0)))$$

Now ① is maximised at 0 as $\max(-\log) \to 0$.
and ② may approach ∞.

Setting $\beta_0 = -\beta_1 x_0$ we obtain:

$$\ell(\beta) = \sum_{x_i < x_0} -\log(1 + \exp(\beta_1(x_i - x_0))) + \sum_{x_i > x_0} \beta_1(x_i - x_0) - \log(1 + \exp(\beta_1(x_i - x_0)))$$

and using $(*)$ above this is maximised as $\beta_1 \to \infty$ as:

$$-\log(1 + 1/\infty) \to 0$$

and

$$\infty - \log(1 + e^{\infty}) \to \infty$$

Thus MLE's are $\beta_0 = -\beta_1 x_0$ and $\beta_1 = \infty$

(a) Now we consider the case where $x_i \in \mathbb{R}^p$.

If the classes are separable then there exists some hyperplane, α, such that:

$$\alpha x_i > 0 \quad \text{if} \quad y_i = 1$$
$$\alpha x_i < 0 \quad \text{if} \quad y_i = 0$$

Now minimising αx_i when $y_i = 1$ and $-\alpha x_i$ when $y_i = 0$ we find there exists some minimum value $\varepsilon > 0$ such that:

$$\alpha x_i \geq \varepsilon \quad \text{if} \quad y_i = 1$$
$$\alpha x_i \leq -\varepsilon \quad \text{if} \quad y_i = 0 \qquad (*)$$

The problem is reduced to analyzing what happens in the direction of α.

Now h is logistic regressions link function that continuously maps the interval of possible probabilities $(0,1)$ with the real line (e.g. 4.18) and h^{-1} exists so

$$\lim_{x \to \infty} h^{-1}(x) = 1 \qquad \text{and} \qquad \lim_{x \to -\infty} h^{-1}(x) = 0$$

Then the likelihood for coefficient vector β is the product, over all observations, of the chance that a bernoulli variable with parameter $\beta^T x_i = y_i$:

$$L(\beta) = \prod_{i:y_i=1} h^{-1}(\beta^T x_i) \prod_{i:y_i=0} (1 - h^{-1}(\beta^T x_i))$$

Consider a positive real number λ, using $(*)$ and the fact that h^{-1} is increasing, then

$$L(\alpha\lambda) = \prod_{i:y_i=1} h^{-1}(\lambda\alpha x_i) \prod_{i:y_i=0}(1-h^{-1}(\lambda\alpha x_i))$$

$$\geq \prod_{i:y_i=1} h^{-1}(\lambda\varepsilon) \prod_{i:y_i=0} (1-h^{-1}(-\lambda\varepsilon))$$

Now, since $\varepsilon > 0$, as $\lambda \to \infty$

$\varepsilon\lambda \to \infty$

and $-\varepsilon\lambda \to -\infty$

giving:

$$\lim_{\varepsilon\to\infty} h^{-1}(\lambda\varepsilon) = 1 \quad \text{and} \quad \lim_{\varepsilon\to+\infty} 1-h^{-1}(-\lambda\varepsilon) = 1$$

So,

$$\lim_{\varepsilon\to\infty} L(\alpha\lambda) \geq \prod_{i:y_i=1} 1 \prod_{i:y_i=0} 1 = 1$$

and since this is a product of probabilities, this is the global maximum.

Finally we notice that it is possible for components of α to be 0 and when multiplied by large λ would remain 0 (not undefined).

To deal with these cases we notice that we can adjust the non-zero components of α so that all components are greater than some ε^* where $\varepsilon > \varepsilon^* > 0$ and therefore $\lambda\varepsilon^* \to \infty$ as $\lambda \to \infty$ making the above analysis still valid. Now we can find another hyperplane (of the infinite that exist) such that all components are non-zero and still separate the two classes.

Therefore we conclude, when a separating hyperplane exists, the likelihood can be maximised in a manner that causes all of the coefficients of β to diverge.

There may exist separating hyperplanes in which some components of β are finite, but in these cases:

1. any such component must be zero.
2. at least one component will diverge.
3. There can be infinitely many directions α for which the likelihood of $\beta = \lambda\alpha$ is maximised as $\lambda \to \infty$.

(b) In this case we know there exists hyperplanes separating each individual class from all of the others. In this case we may find these $K-1$ hyperplanes sequentially by labelling the current class as $y_i = 1$ and all other classes as $y_i = 0$ and following the exact steps as described in part (a). There are $K-1$ rather than K hyperplanes as once we reach the K^{th} class we can simply use the previous $K-1$ hyperplanes to classify the observations and any remaining observations that do not belong to these $K-1$ classes must belong to this final K^{th} class.