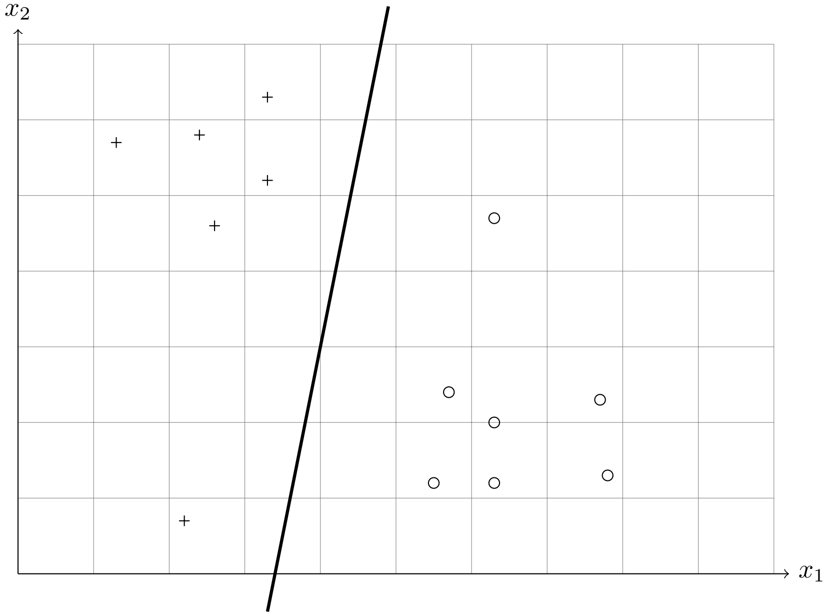
# 1 Q2 solutions

## 1.1

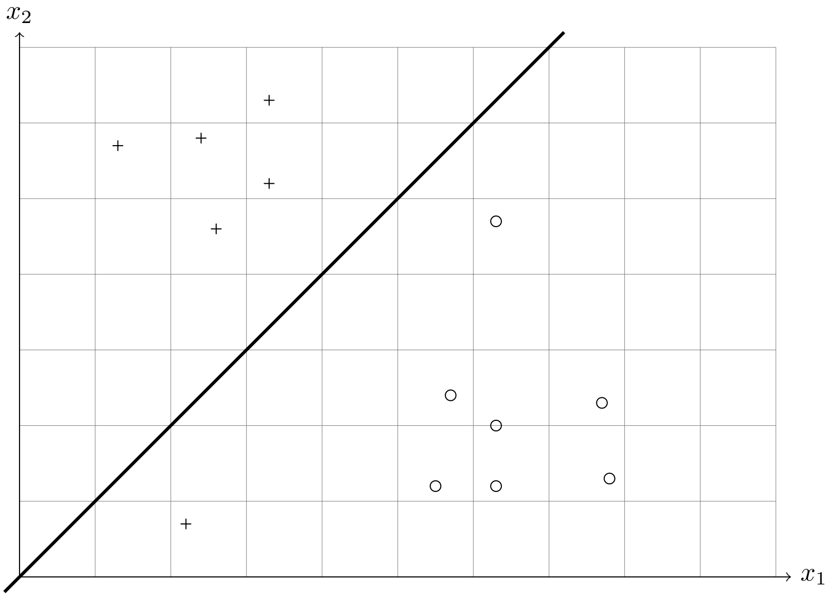
No regularization.



No training errors. The decision boundary is not unique.

## 1.2

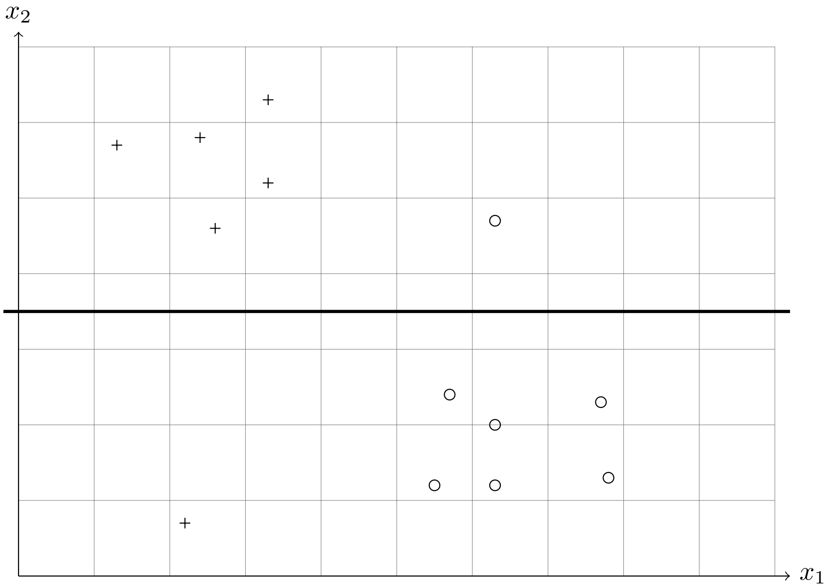
$w_0 = 0$.



One training error. The decision boundary is unique.

## 1.3

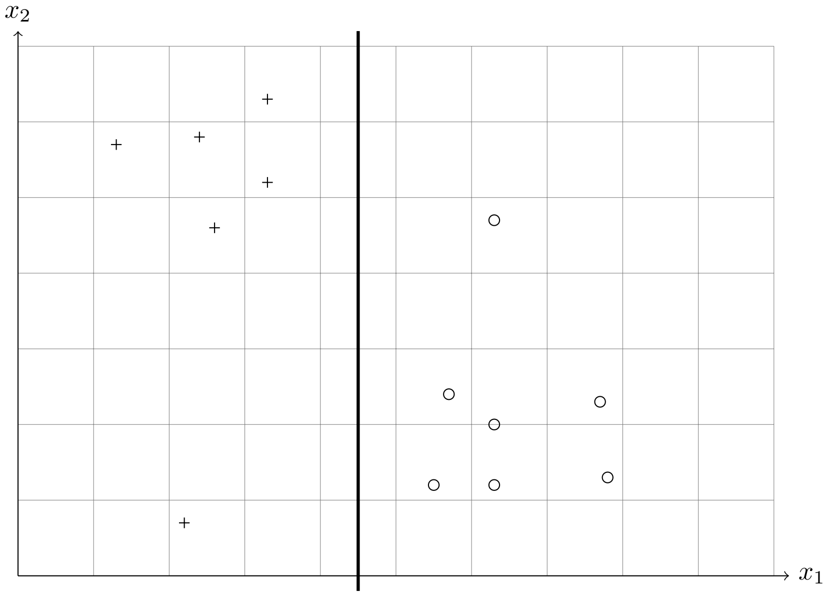
$w_1 = 0$.



Two classification errors. The decision boundary is unique.

### 1.4

$w_2 = 0$.



No classification errors. The decision boundary is not unique.

## 2 Uniqueness of MLE estimates in logistic regression

Here we show how maximum likelihood estimation for logistic regression can break down when training on linearly separable data. In this case, a step function will fit the training data perfectly. This means that maximum likelihood estimation will select parameter values of infinite magnitude, and will allow for many different possible parameter values.

### 2.1 Scaling w

The decision boundary is comprised of all the $\mathbf{x}$ for which we say $p(y = 1|\mathbf{x}, \mathbf{w}) = 0.5$, that is, all the points where we think the class probabilities are equal. On one side of this decision boundary we think $y$ is more likely to be 1, and on the other side we think $y$ is more likely to be 0.

We can easily show that the points where $p(y = 1|\mathbf{x}, \mathbf{w}) = 0.5$ are the points where $\mathbf{w}^T\mathbf{x} = 0$. The $\mathbf{x}$ values which satisfy this equation form a hyperplane.

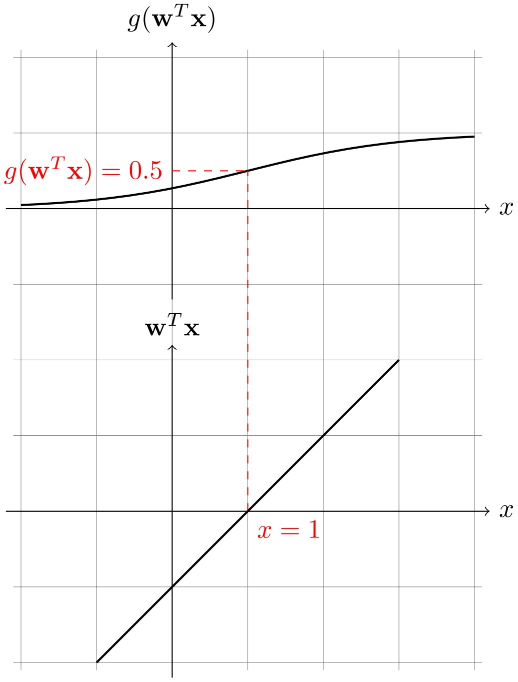
Below we consider the case where $\mathbf{x}$ is of a single dimension. Here we define

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

so $\mathbf{w}^T\mathbf{x} = w_0 + w_1 x$. It is easy to see that in this one-dimensional example, out decision boundary is simply the point where $x = -\frac{w_0}{w_1}$.

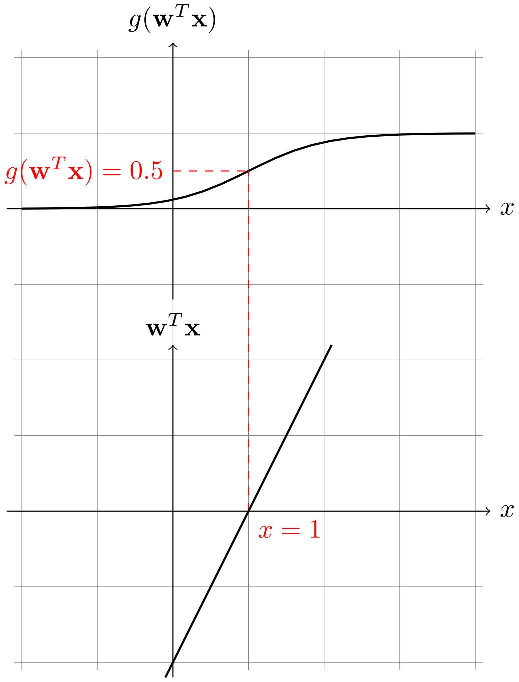
Here we illustrate the relation between the line formed by $\mathbf{w}^T\mathbf{x}$, and the logistic function $g(\mathbf{w}^T\mathbf{x})$. We can think of the logistic function as "squashing" the line, while $g(\mathbf{w}^T\mathbf{x}) = 0.5$ where $\mathbf{x}^T\mathbf{w} = 0$.

In this example, $w_0 = -1, w_1 = 1$.

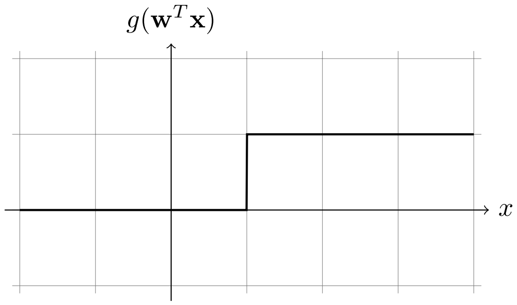


Notice now that since the decision boundary is a ratio of the weights, we can multiply the weights by any constant and still preserve the same decision boundary. I.e. $x = -\frac{w_0}{w_1} = -\frac{cw_0}{cw_1}$.

What changes if we increase the weights? In the diagram below, we set $c = 2$. This increases the slope of the line $\mathbf{w}^T\mathbf{x}$, and makes the function $g(\mathbf{w}^T\mathbf{x})$ increase more sharply.



We can imagine that if we set $c$ above to be very large, as $c \to \infty$ then $g(\mathbf{w}^T\mathbf{x})$ approaches the step function:

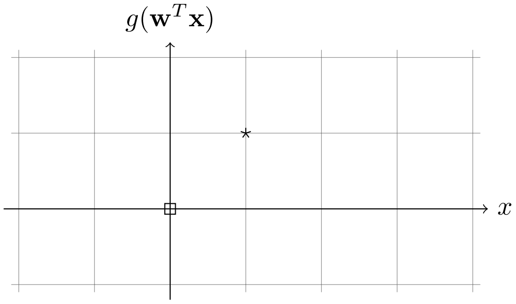


## 2.2 Linearly separable data

Imagine a very simple linear regression example where we have two data points:

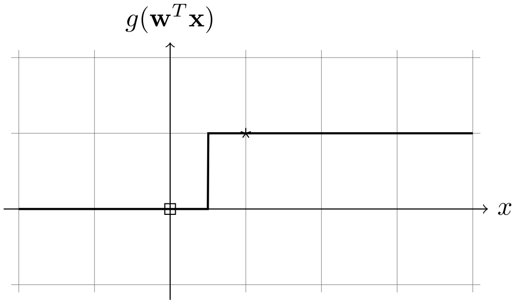
$$(\mathbf{x}_1, y_1) = (0, 0), (\mathbf{x}_2, y_2) = (1, 1)$$

We can visualize this data as in the following plot. (This is similar to the plot of the SAT scores we saw in lecture.)



The MLE estimate of $\mathbf{w}$ attempts to maximize $\prod_i p(y_i|\mathbf{x}_i, \mathbf{w})$. In the above diagram, this is interpreted as fitting the logistic function to pass through, or as close as possible to, the training data points.

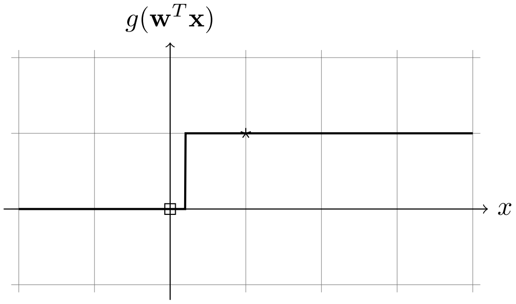
Here we note that the maximum likelihood would be achieved by the step function, which could pass through both of the data points, as illustrated here:



This gives a likelihood of $\prod_{i=1}^{2} p(y_i|\mathbf{x}_i, \mathbf{w}) = 1$. Any logistic function which decreases less sharply would give a lower likelihood.

Now it should be easy to see that any step function which passes through the data points would give the same likelihood.

Specifically, we can shift the step function to the left or right, and still have it pass through both data points, as illustrated here:



This leads to a problem when trying to estimate $\mathbf{w}$. Not only are the ideal $\mathbf{w}$ parameters of infinite magnitude, but any $\mathbf{w}$ with a decision boundary which lies between the two separable classes gives the same likelihood. Since we are trying to optimize the likelihood, this means there are many solutions!

### 2.3 Negative semi-definiteness of $\nabla^2 L(\mathbf{w})$

Note that we only proved that the Hessian of the log-likelihood is negative *semi*-definite, i.e. $\mathbf{v}^T \nabla^2 L(\mathbf{w})\mathbf{v} \leq 0, \forall \mathbf{v}$. (Compare this to the definition of negative definite, which would require a strict inequality: $\mathbf{v}^T \nabla^2 L(\mathbf{w})\mathbf{v} < 0$.)

Note that negative semi-definiteness does not imply a unique optimum. Take for example, a constant function, which is negative semi-definite, and certainly does not have a unique optimum!

We can relate the geometric intepretation given above to the Hessian by examining the gradient and Hessian matrices which were definted in the assignment as follows:

$$\nabla L(\mathbf{w}) = \Phi^T(\mathbf{y} - \mu)$$

$$\nabla^2 L(\mathbf{w}) = -\Phi^T U \Phi$$

where

$$[U]_{i,i} = g(\mathbf{w}^T \mathbf{x}_i)(1 - g(\mathbf{w}^T \mathbf{x}_i))$$

Note that when our training data is predicted perfectly, for the data with label 1, $g(\mathbf{w}^T \mathbf{x}_i) = 1$ and for the data with label 0, $g(\mathbf{w}^T \mathbf{x}_i) = 0$. This means our gradient and Hessian will both tend to zero for these optimal values of $\mathbf{w}$.

### 2.4 Training errors and priors

Note that when we do not have linearly separable data, the above problems will not occur. The maximum likelihood estimate will attempt to accomodate all of the data points, and will never allow a data point to have a zero likelihood. Because of this, the MLE will not tend to a step function as above.

If we do have linearly separable data, we can always avoid the above problem by adding a prior which penalizes large magnitudes of the weights.