Recall last week ...

- Introduced the idea of a **statistical model**, a set of assumptions about the data.
- When model has parameters then we have the **Maximum Likelihood** parameter estimate and the **Maximum A Posteriori (MAP)** parameter estimate
- When we have lots of training data then these estimates are much the same, but when only a little data our prior assumptions affect the parameter estimate
- Introduced the linear regression model. Expressed in two equivalent ways:
  - $Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M$, $M \sim N(0,1)$, $\Theta^{(i)} \sim N(0,\lambda)$, λ known
  - $f_{D|\Theta}(d|\vec{\theta}) \propto L(\theta) = \exp(-\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2/2)$, $f_{\Theta^{(i)}}(\theta) \propto \exp(-\theta^2/2\lambda)$
- Noted that $x^{(i)}$ may be a calculated quantity e.g. the square, cube and so on of an observed value when fitting a polynomial.

Overview

- Classification
- Logistic Regression
- Parameter Estimation
- Maximum Likelihood Estimate

Classification

- Suppose we have a collection of objects and each has an unknown **label** associated with it e.g. likes marmite or doesn't
- For a subset of the objects we observe the label plus some other properties e.g. location, nationality (**features**, **explanatory variables**, **independent variables**). This is our **training data**.
- We are willing to make a number of assumptions, our **model**.
- We now want to build a **classifier** that predicts the label of a new object drawn from the collection.

Examples:

- Based on the text within an email, predict whether it is spam or not
- Given the contents of my shopping basket, predict whether I'm vegetarian or not
- Given where I live in Dublin, predict which political party I'll vote for.

Classification: Logistic Regression

- Label Y only takes values 0 or 1. Real-valued vector $\vec{X}$ of m observed features $X^{(1)}, X^{(2)}, \cdots, X^{(m)}$
- In **Logistic regression** our statistical model is that:

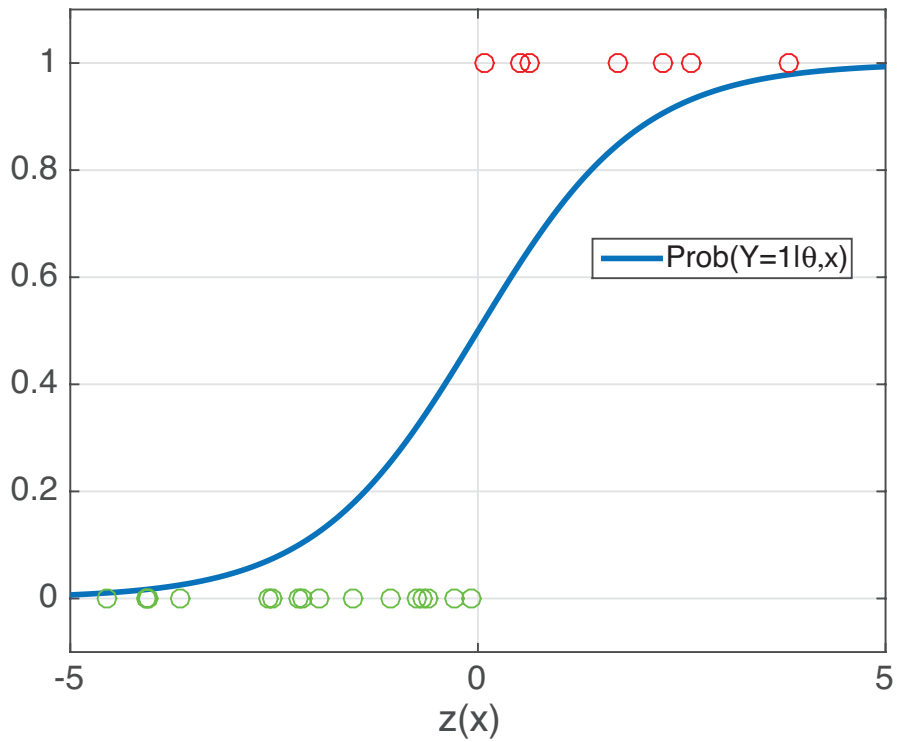
$$P(Y = 1 | \Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-z)} \text{ with } z = \sum_{i=1}^{m} \theta^{(i)} x^{(i)}$$

$$P(Y = 0 | \Theta = \vec{\theta}, \vec{X} = \vec{x}) = 1 - P(Y = 1 | \Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

- Model has m parameters $\theta^{(1)}$, $\theta^{(2)} \cdots, \theta^{(m)}$. We gather these together into a vector $\vec{\theta}$
- Will streamline notation for $P(Y = 1 | \Theta = \vec{\theta}, \vec{X} = \vec{x})$ to $P(Y = 1 | \vec{\theta}, \vec{x})$.

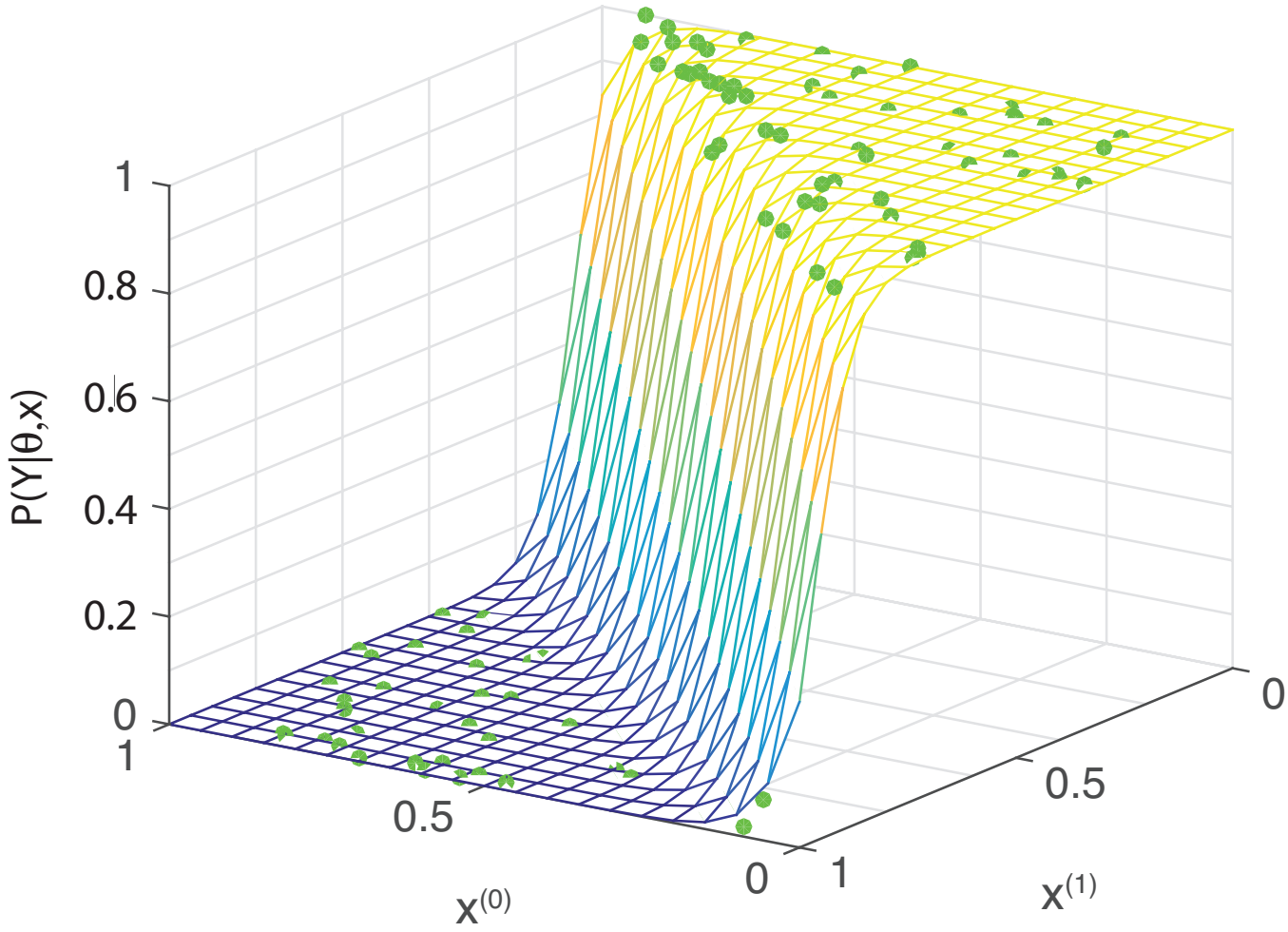
Classification: Logistic Regression

- $P(Y = 1|\vec{\theta}, \vec{x})$ changes smoothly with $\vec{x}$
- Want to try to learn to predict when $Y = 1$ and $Y = 0$ given a value of $\vec{x}$.



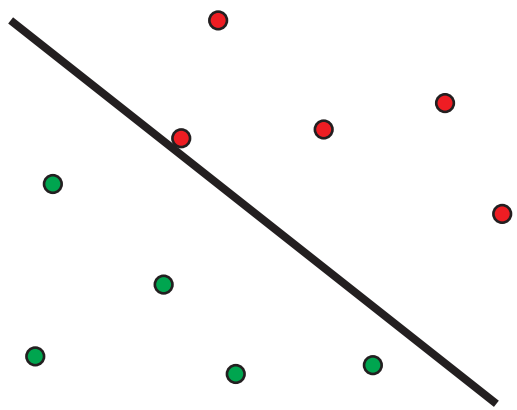
Linear Separability

- Can also plot $P(Y = 1|\vec{\theta}, \vec{x})$ against $\vec{x}$ rather than z.
- Example with vector $\vec{x} = [1, x^{(0)}, x^{(1)}]$



Linear Separability

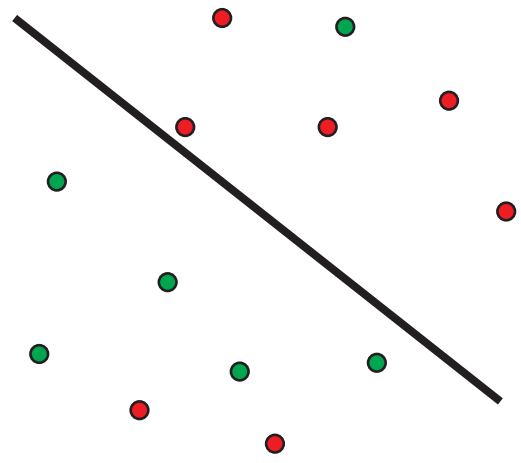
- In general, $\sum_{i=1}^{m} \theta^{(i)} x^{(i)} = 0$ is called a **linear** equation. It defines a plane in m-dimensions.
- Logistic regression thresholds z and predicts $Y = 1$ when $z > 0$ and $Y = 0$ when $z < 0$.
- So we can think of logistic regression as trying to fit a plane that separates the $Y = 1$ data from the $Y = 0$ data.



- We call such data "linearly separable".

Linear Separability

- Not all data is linearly separable e.g.



- Close links between logistic regression and neural networks.

Parameter Estimation

- Training data is RV D. Consists of n observations $d = \{(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)\}$
- Recall Bayes Rule

$$P(\Theta = \vec{\theta}|D = d) = \frac{P(D = d|\Theta = \vec{\theta})P(\Theta = \vec{\theta})}{P(D = d)}$$

posterior *likelihood* *prior*

- Likelihood. Probability of seeing the data d given model with parameter $\Theta = \vec{\theta}$
- Prior. Before seeing any data what is our belief about the model i.e. what is probability of parameter values Θ.
- Posterior. After seeing the data, what is our belief about probability of parameter values Θ now that we have seen the data.
- **Maximum A Posteriori** (MAP) estimate of $\vec{\theta}$ is value that maximises $P(\Theta = \vec{\theta}|D = d)$

Parameter Estimation

- Likelihood is:

$$P(D = d|\Theta = \vec{\theta}) = \prod_{k=1}^{n} P(Y = y_k|\vec{\theta}, \vec{x}_k)$$
$$= \prod_{k=1}^{n} \left(\frac{1}{1+\exp(-z_k)}\right)^{y_k} \left(\frac{\exp(-z_k)}{1+\exp(-z_k)}\right)^{1-y_k}$$

  with $z_k = \sum_{i=1}^{m} \theta^{(i)} x_k^{(i)}$.
- Prior $P(\Theta = \vec{\theta})$. If $\vec{\theta}$ discrete valued then we can use any prior we like. But usually allow $\vec{\theta}$ to be continuous valued in Logistic regression.
- For now let's consider **Maximum Likelihood** estimate of $\vec{\theta}$, the value which maximises $P(D|\vec{\theta})$.

Maximum Likelihood Estimate

Example: have two pairs of observations $(x_1 = 1, y_1 = 1)$ and $(x_2 = -1, y_2 = 0)$, one feature x_k and $z_k = \theta x_k$,

$$P(D = d|\Theta = \vec{\theta}) = p_1^{y_1}(1 - p_1)^{1-y_1} \times p_2^{y_2}(1 - p_2)^{1-y_2}$$

with

$$p_1 = \frac{1}{1 + \exp(-z_1)} = \frac{1}{1 + \exp(-\theta x_1)}, p_2 = \frac{1}{1 + \exp(-z_2)} = \frac{1}{1 + \exp(-\theta x_2)}$$

That is,

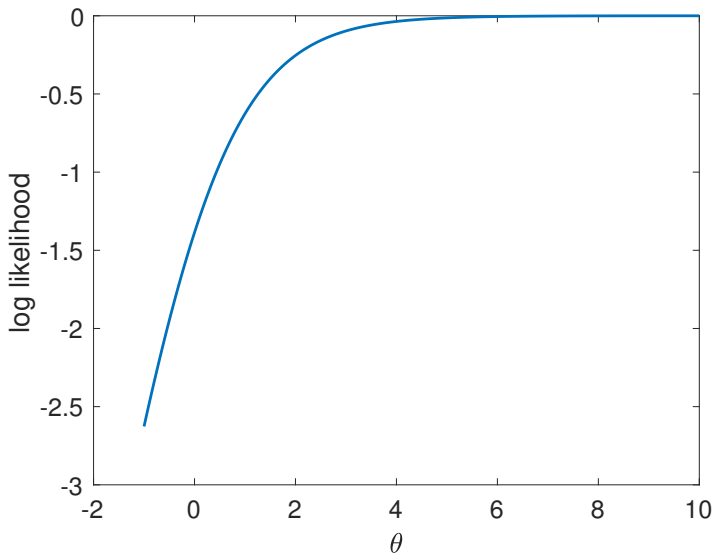
$$P(D = d|\Theta = \vec{\theta}) = p_1 \times (1 - p_2)$$

with

$$p_1 = \frac{1}{1 + \exp(-\theta)}, p_2 = \frac{1}{1 + \exp(+\theta)}$$

Maximum Likelihood Estimate

- Maximum Likelihood estimate is the value of $\vec{\theta}$ which maximises $P(D|\vec{\theta})$
- Maximising $\log P(D = d|\Theta = \vec{\theta})$ is the same as maximising $P(D = d|\Theta = \vec{\theta})$ (why ?)
- $\log P(D = d|\Theta = \vec{\theta})$ is referred to as the **log-likelihood**.
- $\log P(D = d|\Theta = \vec{\theta}) = \log(p_1 \times (1 - p_2)) = \log p_1 + \log(1 - p_2)$ with $p_1 = \frac{1}{1+\exp(-\theta)}$, $p_2 = \frac{1}{1+\exp(+\theta)}$



Maximum Likelihood Estimate

- Log-likelihood maximised by selecting $\theta = +\infty$. What does this mean ?
- $p_1 = \frac{1}{1+\exp(-\theta)} = 1$, $p_2 = \frac{1}{1+\exp(+\theta)} = 0$
- So our prediction is

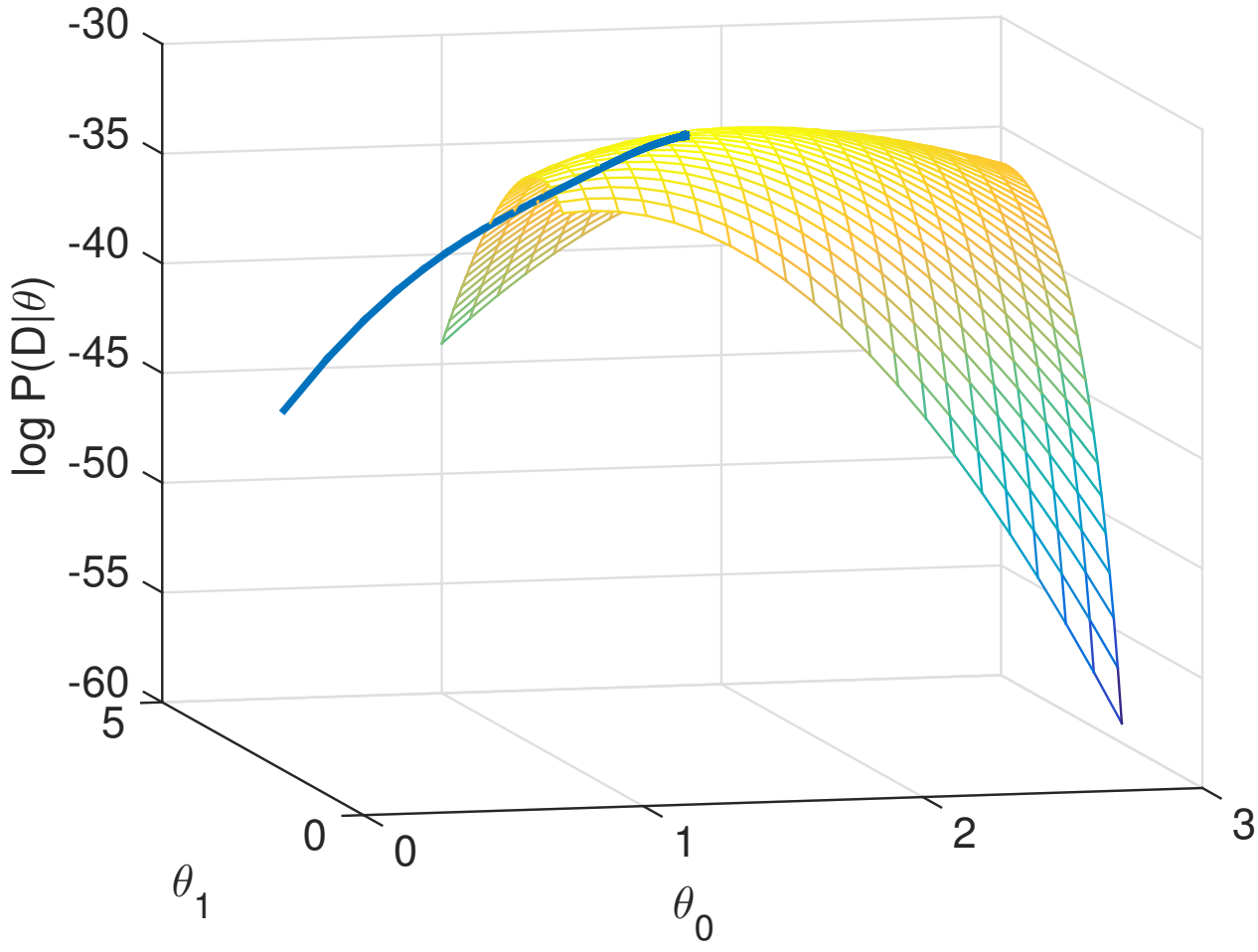
$$P(Y = 1|\Theta = \infty, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-z)}, \; z = \theta^{(1)} x^{(1)} = \begin{cases} 1 & x^{(1)} = -1 \\ 0 & x^{(1)} = 0 \end{cases}$$

$$P(Y = 0|\Theta = \infty, \vec{X} = \vec{x}) = 1 - P(Y = 1|\Theta = \infty, \vec{X}$$

- Recall training data is $(x_1 = 1, y_1 = 1)$ and $(x_2 = -1, y_2 = 0)$

When $\vec{\theta}$ has many elements ...

- log-likelihood is concave, has a single maximum



Alternative Solution Method (Not Examinable)

Alternative approach:

- Compute derivative of log-likelihood with respect to $\theta^{(i)}$:

$$\sum_{k=1}^{n}(y_k - \frac{1}{1+\exp(-z_k)})x_k^{(i)}$$

(Remember $\frac{d\log(x)}{dx} = \frac{1}{x}$, $\frac{d\exp(x)}{dx} = exp(x)$, $\frac{d}{dx}\frac{1}{x} = -\frac{1}{x^2}$ and chain rule $\frac{df(z(x))}{dx} = \frac{df}{dz}\frac{dz}{dx}$)

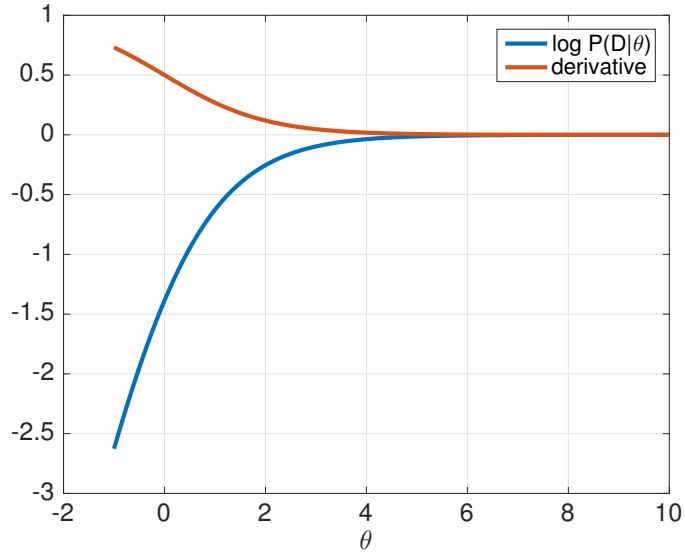
Alternative Solution Method (Not Examinable)

Example: have two pairs of observations $(x_1 = 1, y_1 = 1)$ and $(x_2 = 0, y_2 = 0)$, one feature x_k and $z_k = \theta x_k$,

$$P(D = d|\Theta = \vec{\theta}) = p_1 \times (1 - p_2),\ p_1 = \frac{1}{1 + \exp(-\theta)},\ p_2 = \frac{1}{1 + \exp(+\theta)}$$

Derivative of $\log P(D = d|\Theta = \vec{\theta})$ is

$$(y_1 - p_1)x_1 + (y_2 - p_2)x_2 = (1 - p_1) + (0 - p_2) \times 0 = 1 - p_1$$



Alternative Solution Method (Not Examinable)

- Compute derivative of log-likelihood with respect to $\theta^{(i)}$:

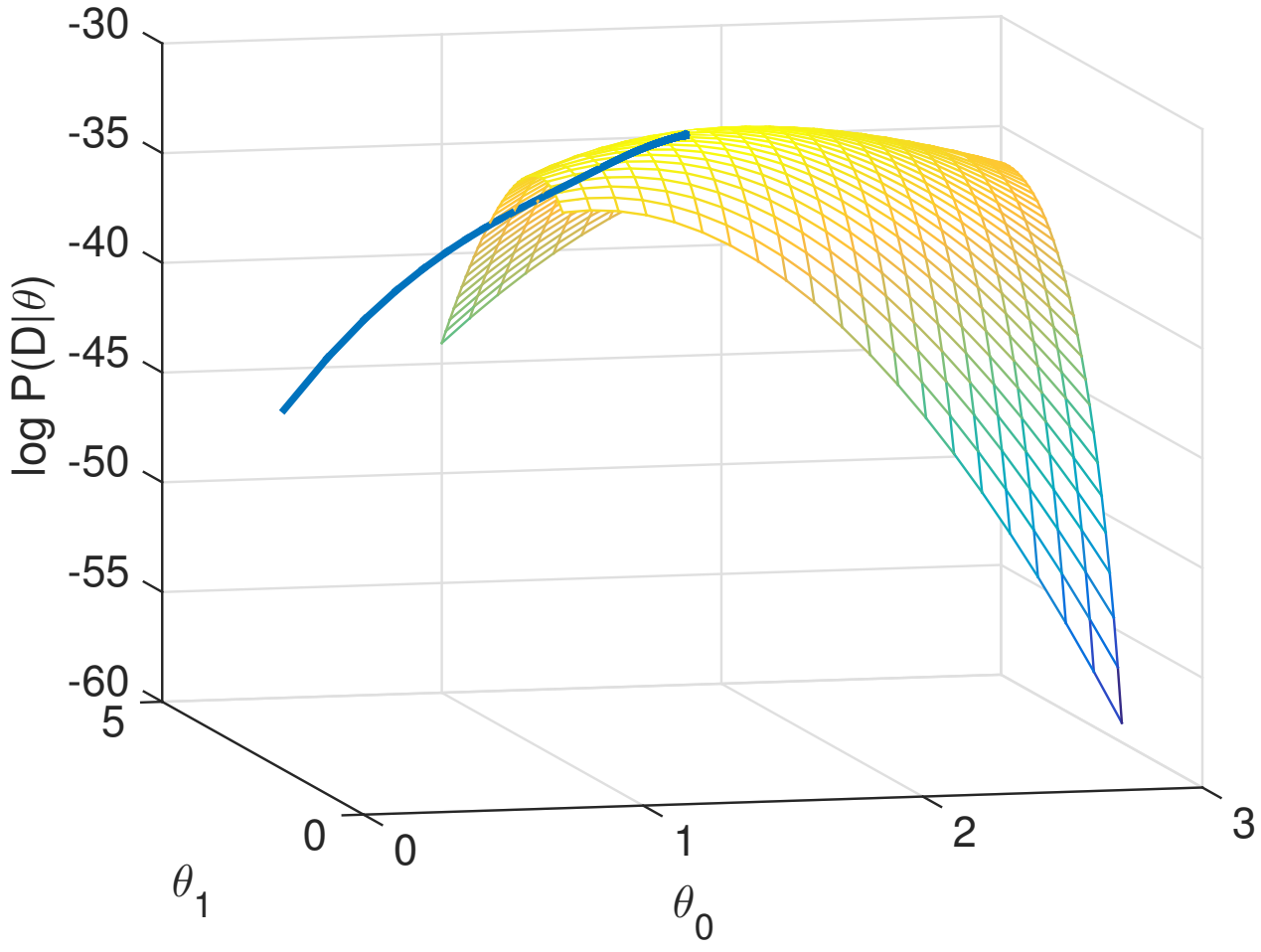
$$\sum_{k=1}^{n}(y_k - \frac{1}{1+\exp(-z_k)})x_k^{(i)}$$

- Would like to set derivative equal to 0 to find ML estimate of $\vec{\theta}$, but hard to do this.
- Instead solve numerically using iteration:

$$\theta_{j+1}^{(i)} = \theta_j^{(i)} + \alpha \sum_{k=1}^{n}(y_k - \frac{1}{1+\exp(-z_{k,j})})X_k^{(i)} \text{ with } z_{k,j} = \sum_{i=1}^{m}\theta_j^{(i)}x_k^{(i)}$$

Alternative Solution Method (Not Examinable)

- log-likelihood is concave, has a single maximum
- iteration climbs uphill until it reaches the maximum

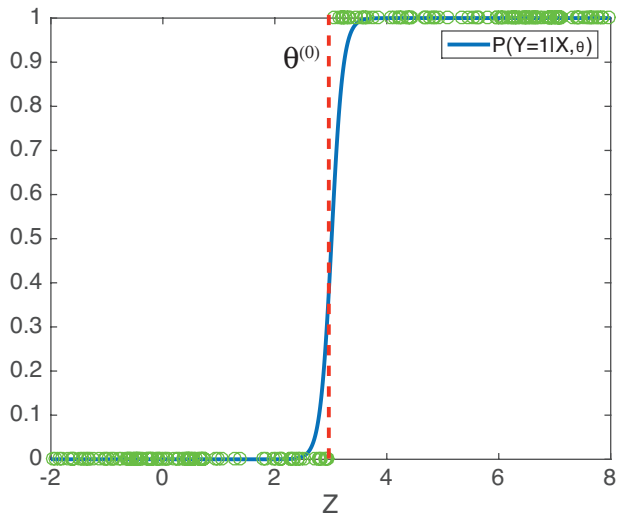


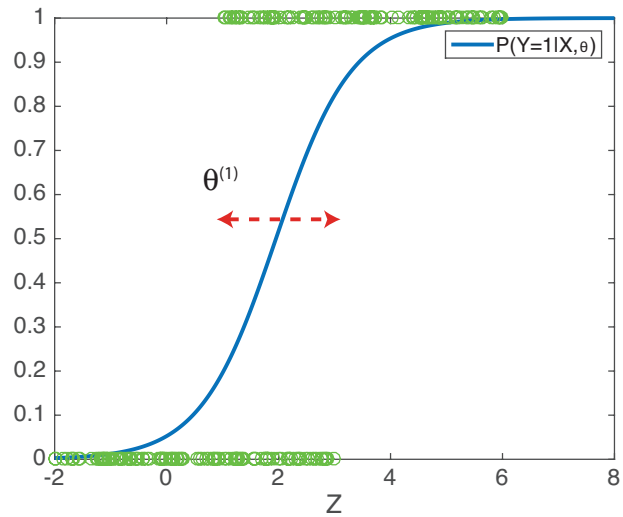
Maximum Likelihood Estimate

```
alpha=0.01; [N,m]=size(X); theta=zeros(1,m);
for l=1:10000,
    grad=zeros(1,m);
    for k=1:N,
        Z=0;
        for i=1:m
            Z=Z+theta(i)*X(k,i);
        end
        for i=1:m
            grad(i) = grad(i) + ...
                (Y(k)-1/(1+exp(-Z)))*X(k,i);
        end
    end
    for i=1:m
        theta(i) = theta(i) + alpha*grad(i);
    end
end
```

Example

- Two inputs $\vec{x} = [1, x]$ so $x = \theta^{(0)} + \theta^{(1)}x$. First input is fixed and means $\theta^{(0)}$ captures offset, $\theta^{(1)}$ captures slope.
- As we increase "noise" on Y then parameter $\theta^{(1)}$ changes to broaden curve reflecting greater uncertainty.





Summary

- Introduced the logistic regression model for classification.

$$P(Y = 1|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-z)} \text{ with } z = \sum_{i=1}^{m} \theta^{(i)} x^{(i)}$$

$$P(Y = 0|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = 1 - P(Y = 1|\Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

- Assumes the data is linearly separable
- Discussed the Maximum Likelihood parameter estimate (there is also a MAP estimate when we include a prior, but we leave that alone for now).