Machine Learning

4771

Instructor: Itsik Pe’er

Reminder: Parameter Estimation

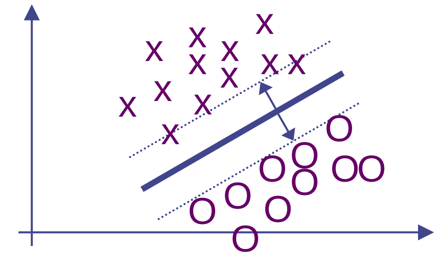
A-Posteriori → Regularization

$$R_{regularized}(\theta) = \frac{1}{N}\sum_{i=1}^{N} Loss(y_i, f(x_i;\theta)) + \frac{\lambda}{2}\|\theta\|^2$$

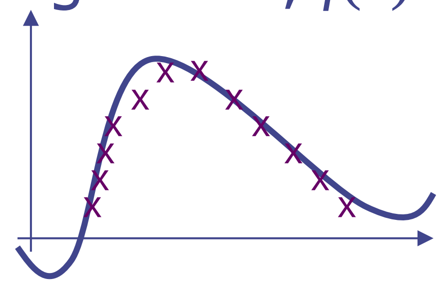
$$\theta^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda N I)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Regression

Classification

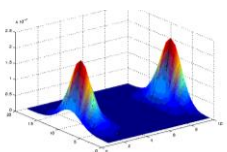


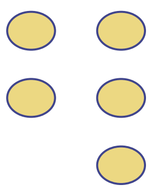
Regression, $f(x)=y$

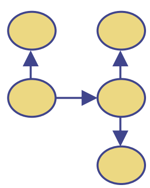


Supervised

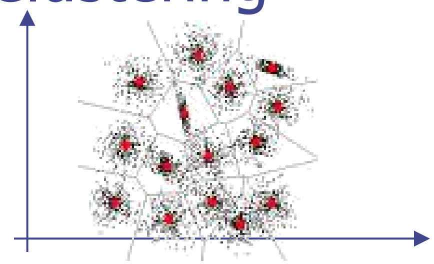
Density/Structure Estimation



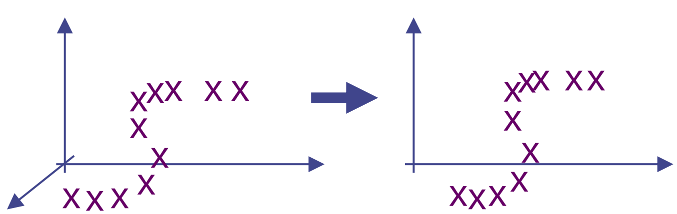




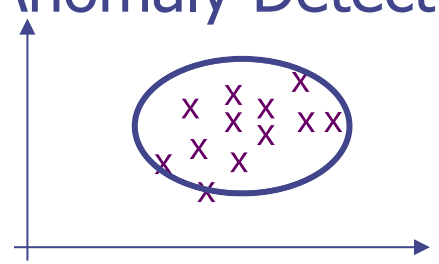
Clustering



Feature Selection



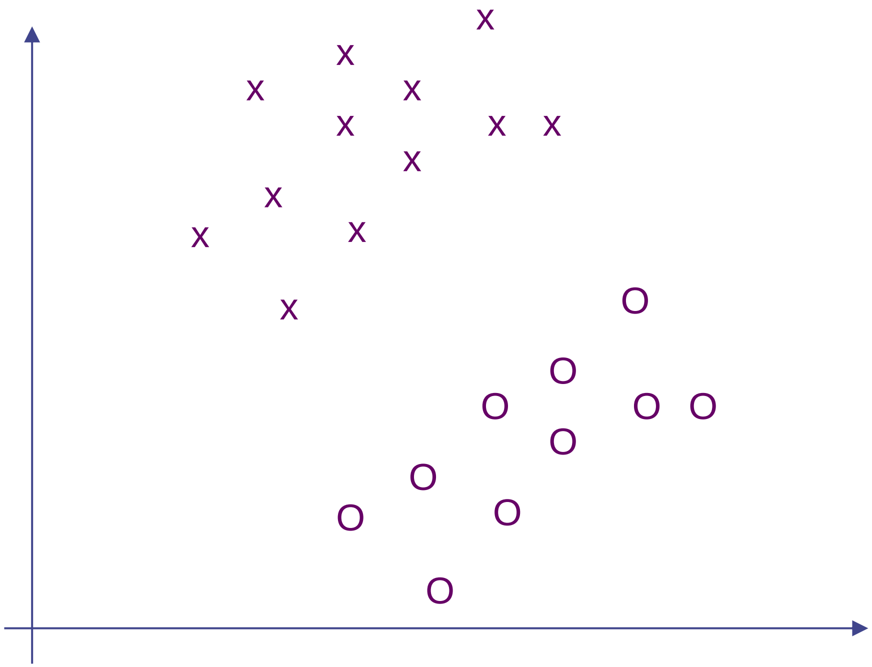
Anomaly Detection



Unsupervised

Class 6

- **Classification**
- Logistic Regression
- Gradient Descent



Classification Problems

- Determine student admission to Columbia based on GPA, prev. school rank, tests

Classification Problems

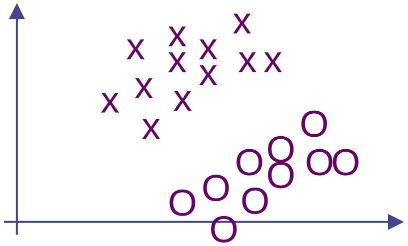
- Determine student admission to Columbia based on GPA, prev. school rank, tests
- Decide malignant or benign tumors based on size, density, speed of growth

Kim et al. '07

Formalizing Classification

- Classification is another important learning problem

Classification: $X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, y_i \in \{0,1\}$

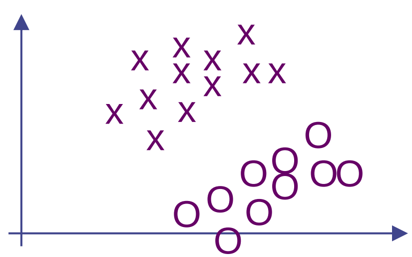


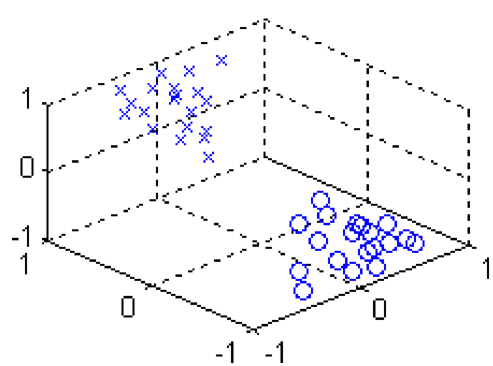
Classification is like Regression

- Classification is another important learning problem

Classification: $X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, y_i \in \{0,1\}$

Regression: $X = \{(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots, (\boldsymbol{x}_N, t_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, t_i \in \boldsymbol{R}$





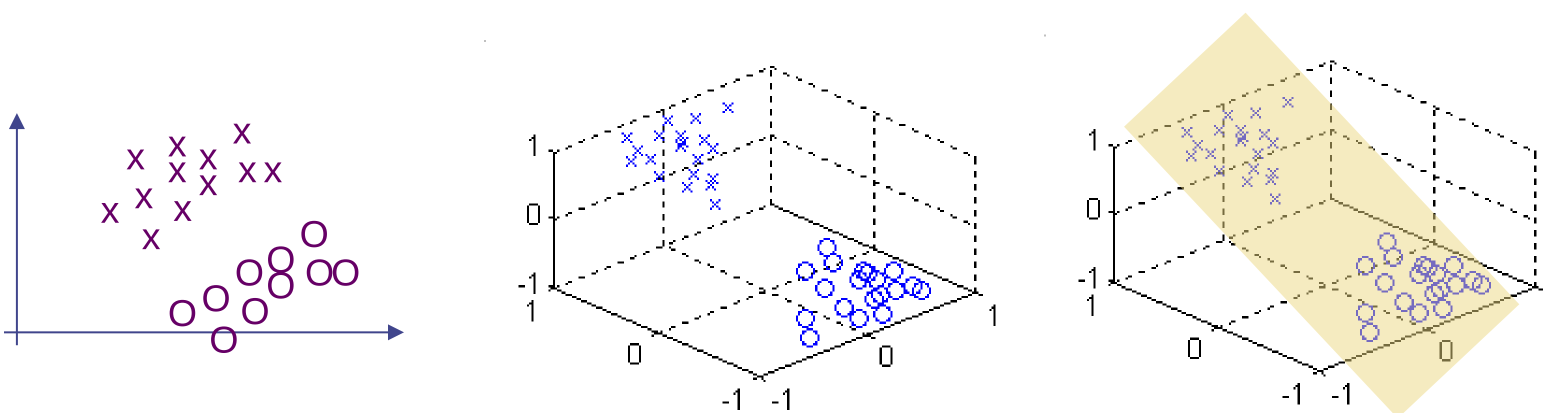
Classification is like Regression

- Classification is another important learning problem

Classification: $X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, y_i \in \{0,1\}$

Regression: $X = \{(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots, (\boldsymbol{x}_N, t_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, t_i \in \boldsymbol{R}$

- Should we solve this as a least squares regression problem?



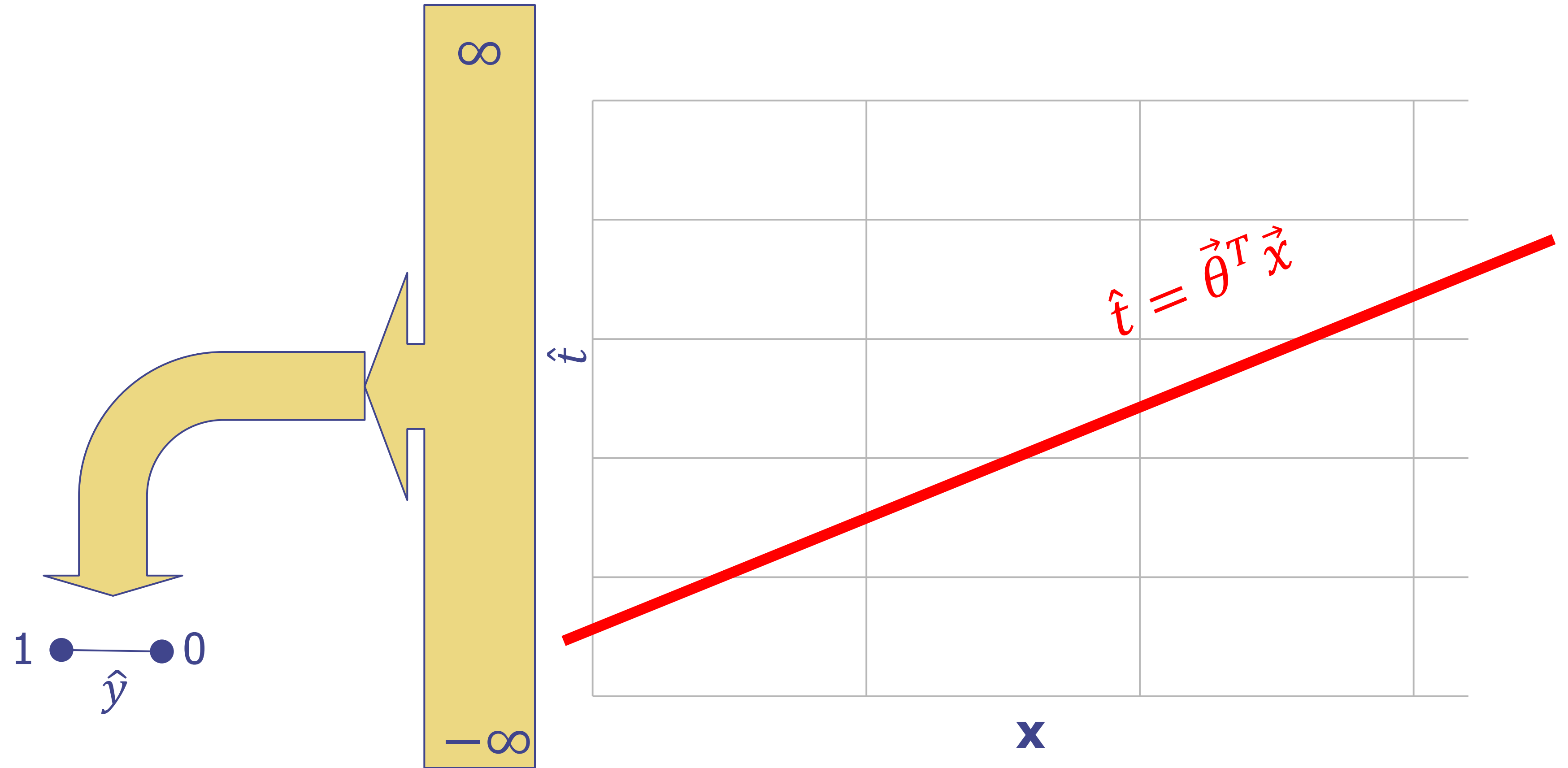
Short hand for Linear Functions

- Hiding the intercept by notation

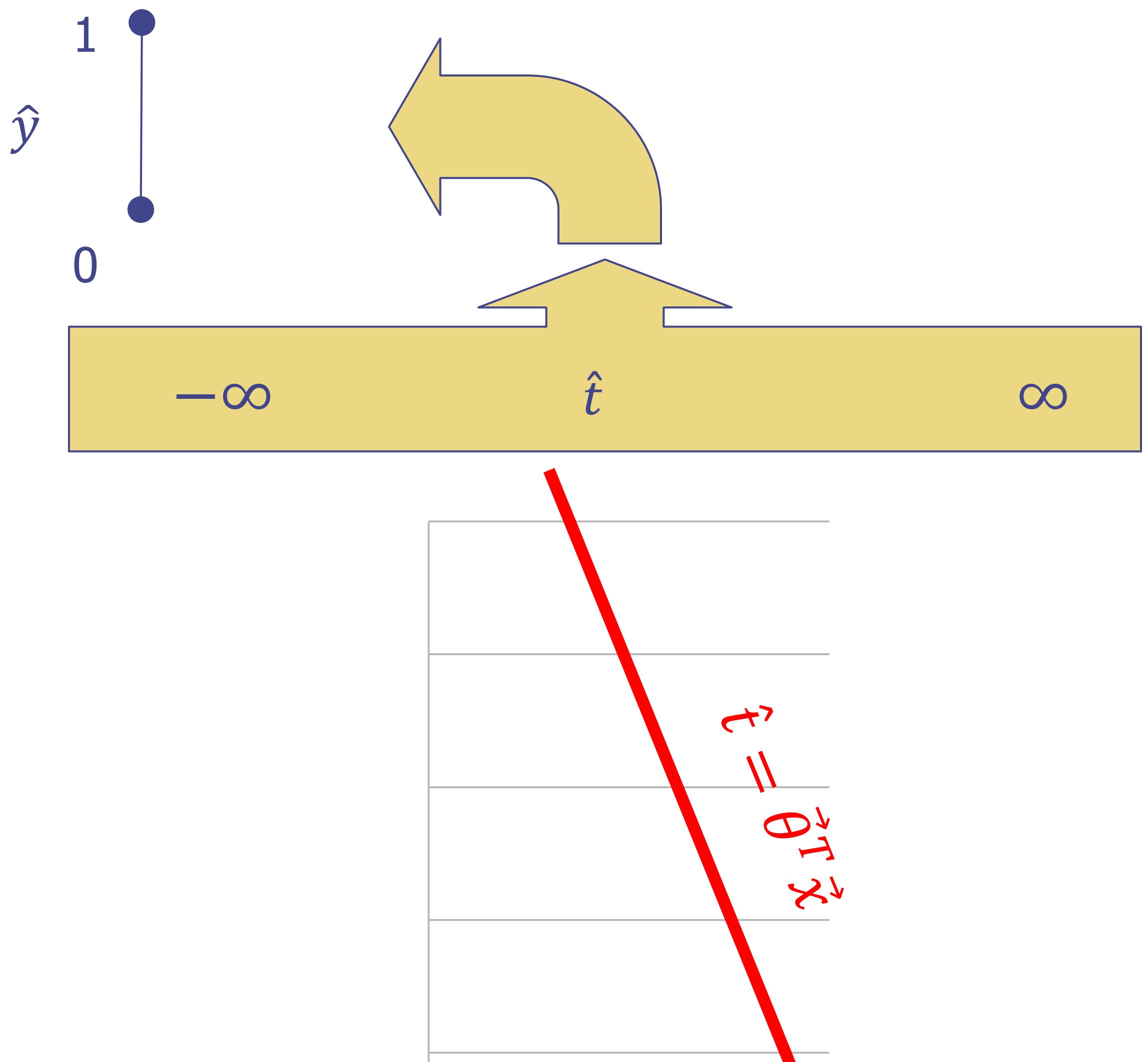
$$f(\boldsymbol{x};\theta) = \theta^T\boldsymbol{x} + \theta_0$$

$$= \begin{bmatrix} \theta(1) \\ \theta(2) \\ \vdots \\ \theta(D) \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(1) \\ \boldsymbol{x}(2) \\ \vdots \\ \boldsymbol{x}(D) \end{bmatrix} + \theta_0 = \begin{bmatrix} \theta_0 \\ \theta(1) \\ \theta(2) \\ \vdots \\ \theta(D) \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{x}(1) \\ \boldsymbol{x}(2) \\ \vdots \\ \boldsymbol{x}(D) \end{bmatrix} = \vec{\theta}^T\vec{\boldsymbol{x}}$$

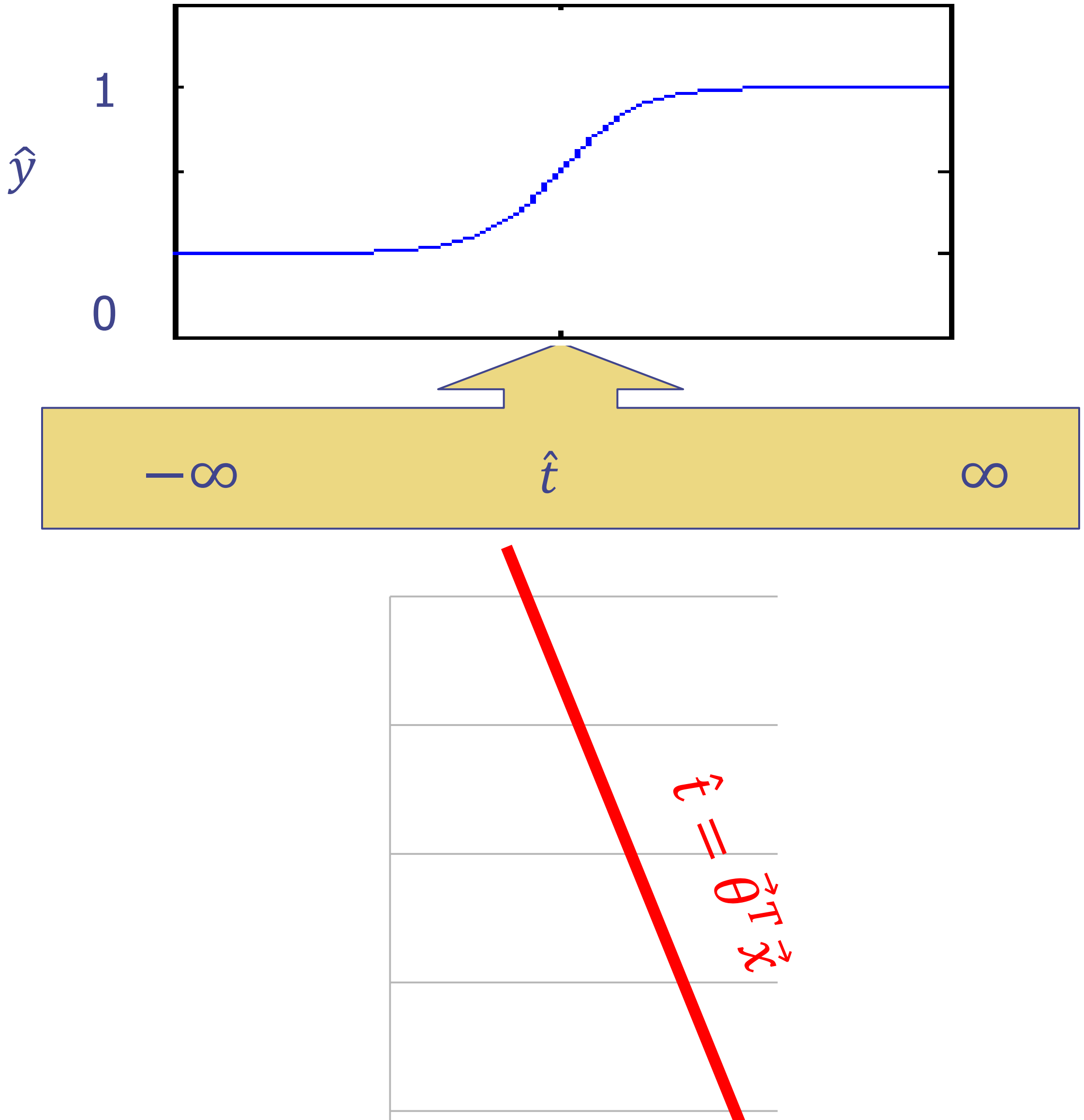
Recalibrating Range



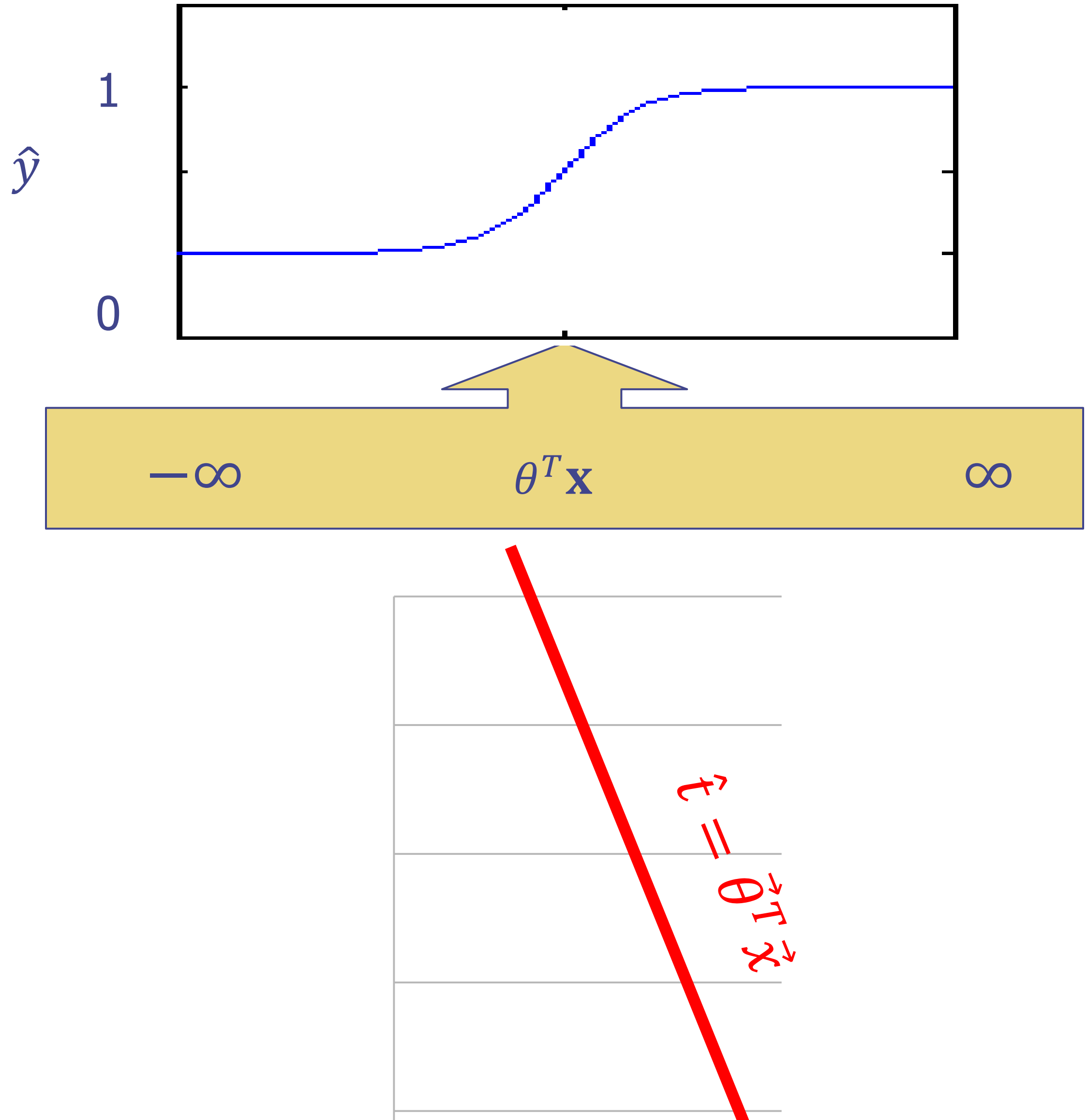
Recalibrating Range



Recalibrating Range



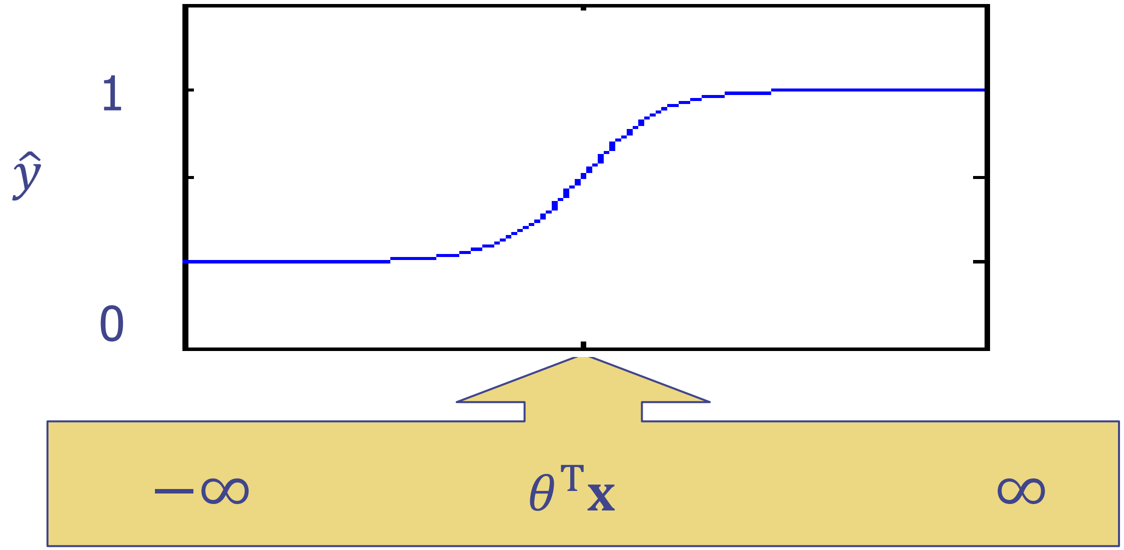
Recalibrating Range



$$\lim_{\theta^T x \to \infty} \hat{y} = 1$$

$$\lim_{\theta^T x \to -\infty} \hat{y} = 0$$

Recalibrating Range



$$\lim_{\theta^T \boldsymbol{x}\to\infty} \hat{y} = 1$$

$$\hat{Y} \underset{\theta^T \boldsymbol{x} \to \infty}{\cong} 1 - \exp(-\theta^T \boldsymbol{x})$$

$$\lim_{\theta^T \boldsymbol{x}\to-\infty} \hat{y} = 0$$

$$\hat{Y} \underset{\theta^T \boldsymbol{x} \to -\infty}{\cong} \exp(\theta^T \boldsymbol{x})$$

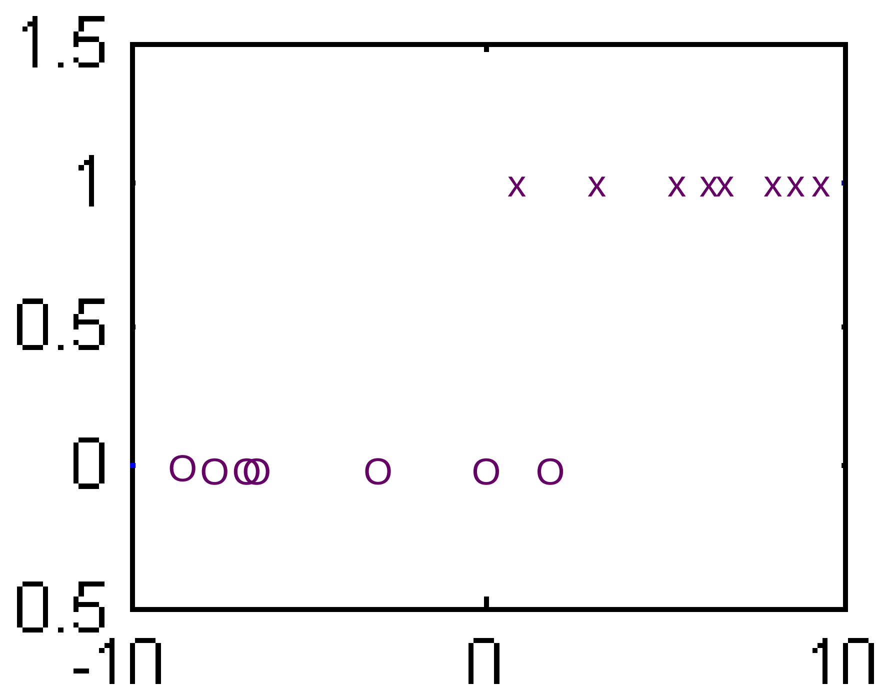
Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x}_i \in \boldsymbol{R}^D, y_i \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x};\theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$



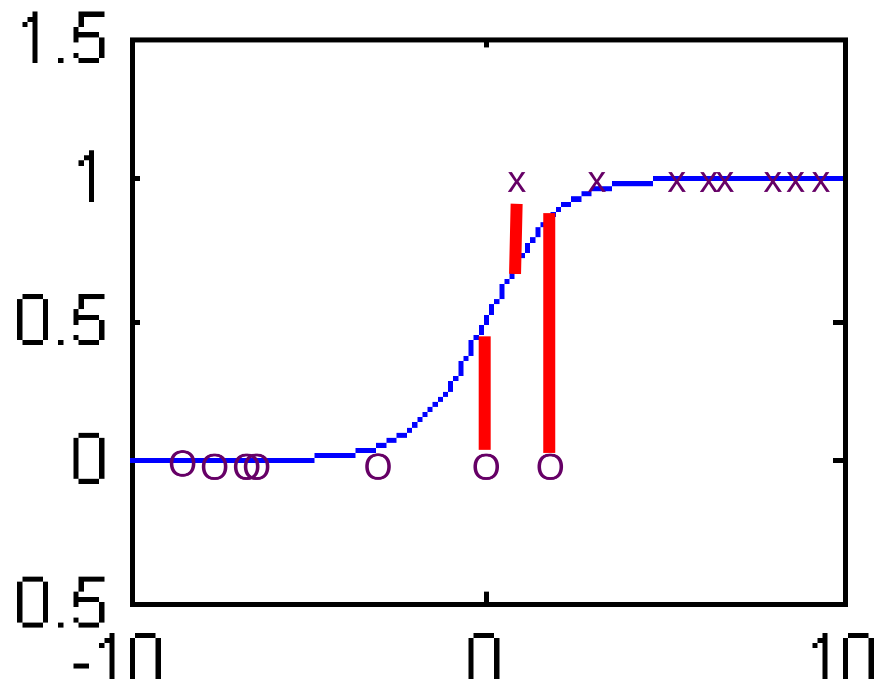
Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x}; \theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$



Logistic Regression

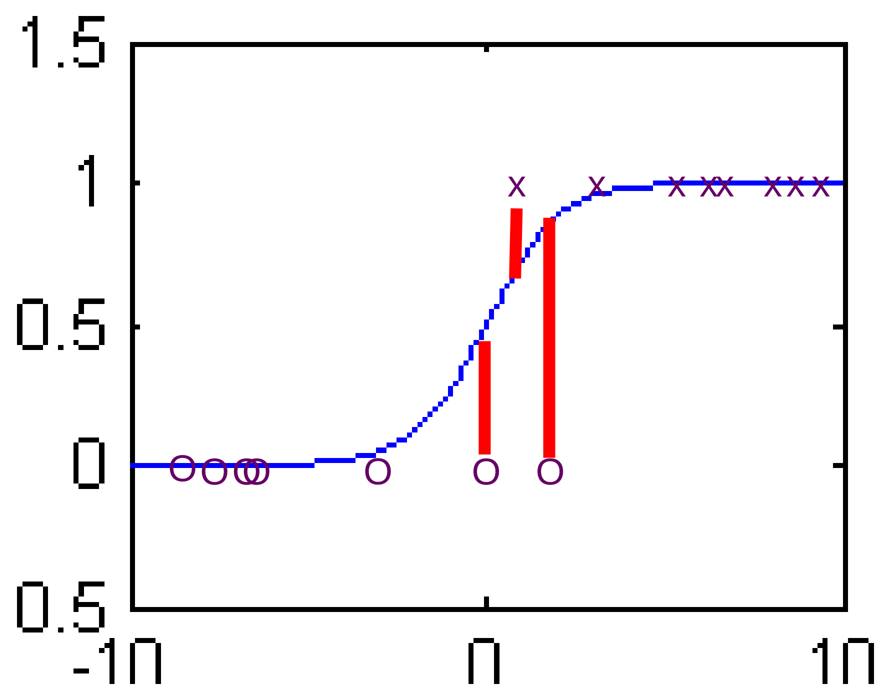
- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x};\theta) = \frac{1}{1+exp(-\theta^T\boldsymbol{x})}$$

- Assume $\Pr(y=1) = f(\boldsymbol{x};\theta)$



Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

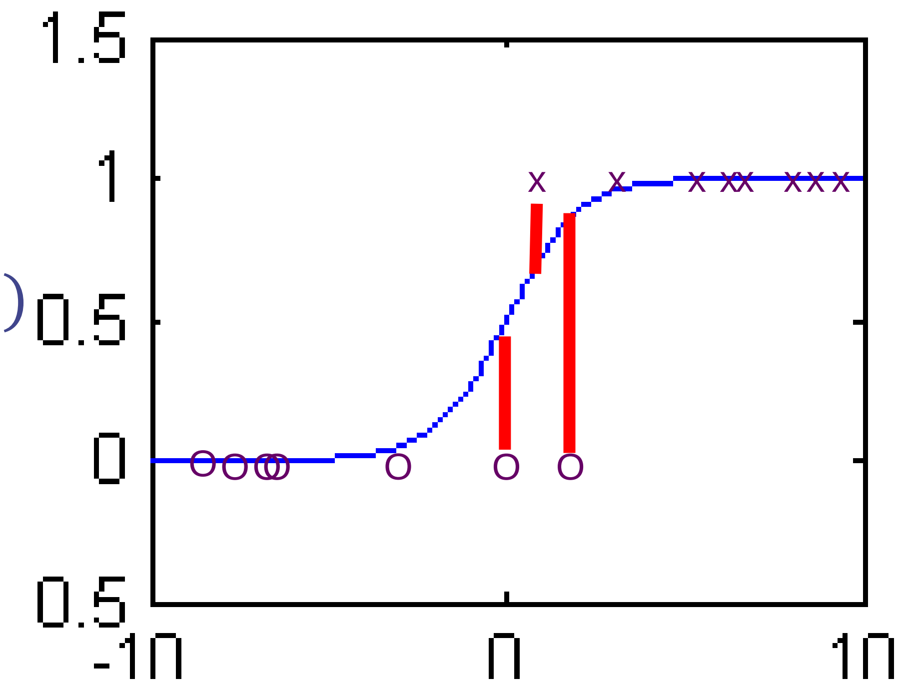
- Use this function and output 1 if $f(\boldsymbol{x}) > 0.5$ and 0 otherwise

$$f(\boldsymbol{x}; \theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$

- Assume $\Pr(y_i = 1) = f(\boldsymbol{x}_i; \theta)$

- $\Pr(y | f(\boldsymbol{x}; \theta)) =$

$$= \prod_{y_i=0} \left(1 - f(\boldsymbol{x}_i; \theta)\right) \prod_{y_i=1} f(\boldsymbol{x}_i; \theta)$$



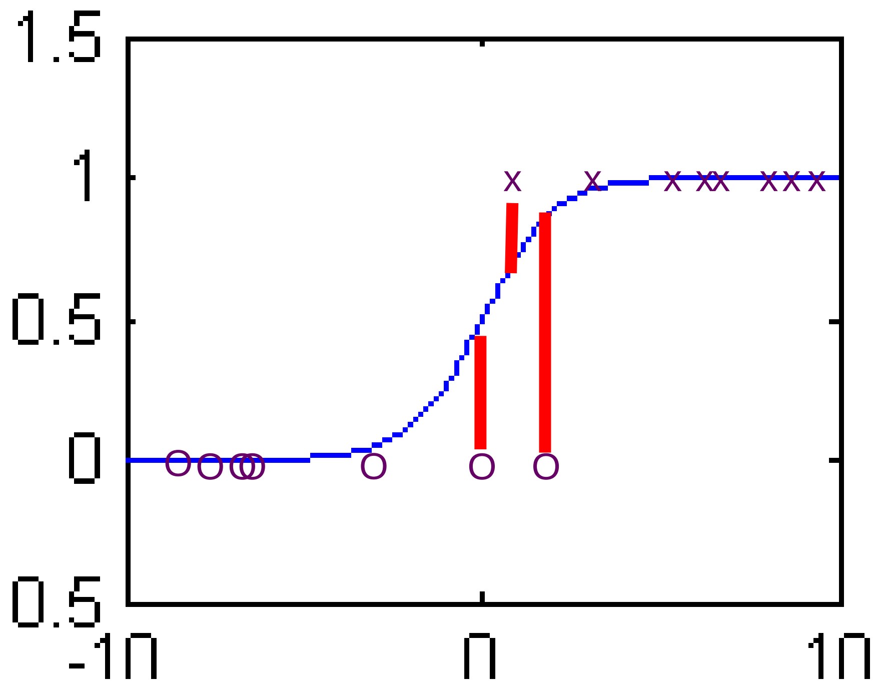
Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x};\theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$



- Assume $\Pr(y_i = 1) = f(\boldsymbol{x}_i;\theta)$
- $\Pr(y|f(\boldsymbol{x};\theta)) =$
$=\prod_{y_i=0}\left(1 - f(\boldsymbol{x}_i;\theta)\right)\prod_{y_i=1} f(\boldsymbol{x}_i;\theta)$
- $\log \Pr(y|f(\boldsymbol{x};\theta)) =$
$=\sum_i\left[y_i \log f(\boldsymbol{x}_i;\theta) + (1 - y_i)\log\left(1 - f(\boldsymbol{x}_i;\theta)\right)\right]$

Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x}; \theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$

- Instead of squared loss, use Logistic Loss (i.e. negative binomial likelihood)

$$Loss_{log}\big(y, f(\boldsymbol{x}; \theta)\big) = (y - 1)\log\big(1 - f(\boldsymbol{x}; \theta)\big) - y\log(f(\boldsymbol{x}; \theta))$$

- The resulting method is called Logistic Regression.
- Empirical Risk:

Logistic Regression

- Given a classification problem with binary outputs

$$X = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \boldsymbol{x} \in \boldsymbol{R}^D, y \in \{0,1\}$$

- Use this function and output 1 if $f(\boldsymbol{x})>0.5$ and 0 otherwise

$$f(\boldsymbol{x};\theta) = \frac{1}{1 + exp(-\theta^T \boldsymbol{x})}$$

- Instead of squared loss, use Logistic Loss (i.e. negative binomial likelihood)

$$Loss_{log}\big(y, f(\boldsymbol{x};\theta)\big) = (y-1)\log\big(1 - f(\boldsymbol{x};\theta)\big) - y\log(f(\boldsymbol{x};\theta))$$

- The resulting method is called Logistic Regression.
- Empirical Risk:

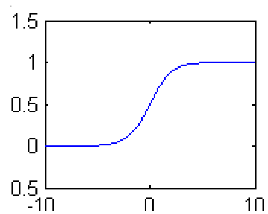
$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[(y_i - 1)\log\big(1 - f(\boldsymbol{x}_i;\theta)\big) - y_i\log(f(\boldsymbol{x}_i;\theta))\right]$$

Logistic Regression

- With empirical logistic risk has no closed form solution:

$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)\log\big(1 - f(\boldsymbol{x_i};\theta)\big) - y_i\log(f(\boldsymbol{x_i};\theta))$$

$$f(\boldsymbol{x};\theta) = \frac{1}{1 + exp(-\theta^T\boldsymbol{x})}$$



Logistic Regression

- With empirical logistic risk has no closed form solution:

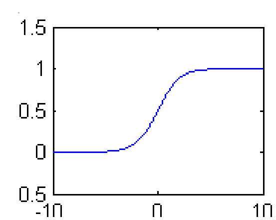
$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)\log\big(1 - f(\boldsymbol{x}_i;\theta)\big) - y_i\log(f(\boldsymbol{x}_i;\theta))$$

$$\nabla_\theta R = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1 - y_i}{1 - f(\boldsymbol{x}_i;\theta)} - \frac{y_i}{f(\boldsymbol{x}_i;\theta)}\right)f'(\boldsymbol{x}_i;\theta) = 0 \quad ??????$$

Find best θ numerically!

where

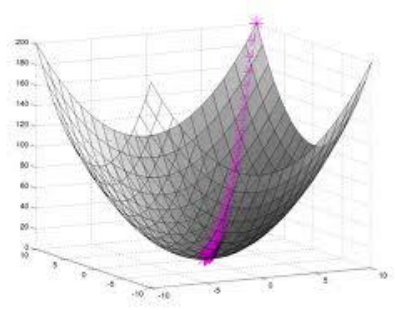
$$f(\boldsymbol{x};\theta) = \frac{1}{1 + exp(-\theta^T\boldsymbol{x})} = g(\theta^T\boldsymbol{x})$$



$$g(z) = \frac{1}{1 + exp(-z)} \qquad g'(z) = g(z)(1 - g(z))$$

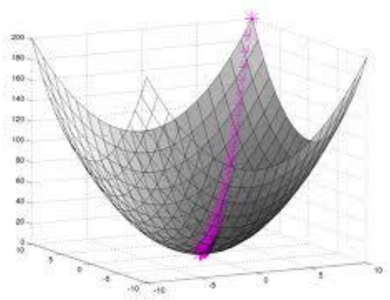
Gradient Descent

- Useful when we can't get minimum solution in closed form
- Gradient points in direction of fastest increase
- Take step in the opposite direction!



Gradient Descent

- Useful when we can't get minimum solution in closed form
- Gradient points in direction of fastest increase
- Take step in the opposite direction!



- Gradient Descent Algorithm

choose scalar step size η, *& tolerance* ε
initialize θ^0 *= small random vector*

$$\theta^1 \leftarrow \theta^0 - \eta_0 \nabla_\theta R_{emp}|_{\theta^0} \;\; ; t \leftarrow 1$$

while $\|\theta^t - \theta^{t-1}\| \geq \epsilon$ $\{$

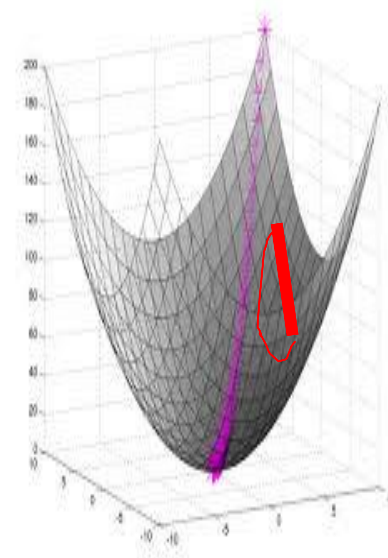
$$\theta^{t+1} \leftarrow \theta^t - \eta_t \nabla_\theta R_{emp}|_{\theta^t} \;\; ; t \leftarrow t+1 \;\}$$

- For appropriate $\{\eta_t\}$, this will converge to local minimum

Gradient Descent Convergence

$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)\log\big(1 - f(\boldsymbol{x_i};\theta)\big) - y_i\log(f(\boldsymbol{x_i};\theta))$$

is a convex function, so local minimum is global



Gradient Descent Convergence

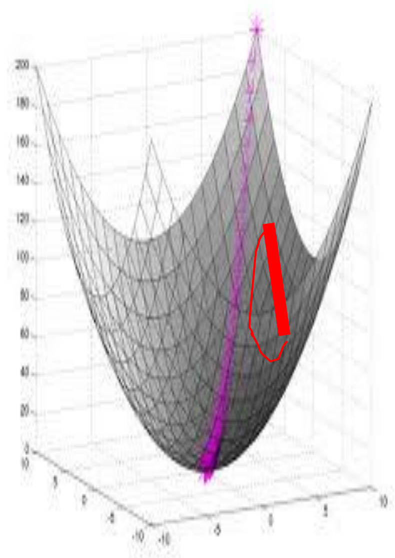
$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)\log\big(1 - f(\boldsymbol{x_i};\theta)\big) - y_i\log(f(\boldsymbol{x_i};\theta))$$

is a convex function, so local minimum is global

Proof:

Convex combination of convex functions

$-\log\big(1 - f(\boldsymbol{x_i};\theta)\big)$ and $-\log(f(\boldsymbol{x_i};\theta))$



Convexity of Regression Loss

$-\log(f(\boldsymbol{x_i};\theta))$ is a convex function

Convexity of Regression Loss

$-\log(f(\boldsymbol{x}_i;\theta))$ is a convex function

Proof:

$$\nabla_\theta\left[-\log\left(f(\boldsymbol{x}_i;\theta)\right)\right] =$$

$$= \nabla_\theta\left[-\log\left(\frac{1}{1+\exp(-\theta^T x_i)}\right)\right]$$

$$= \nabla_\theta\left[\log(1+\exp(-\theta^T x_i))\right]$$

$$= \frac{-\exp(-\theta^T x_i)\,x_i}{1+\exp(-\theta^T x_i)}$$

$$= \left(\frac{1}{1+\exp(-\theta^T x_i)} - 1\right) x_i$$

Convexity of Regression Loss

$-\log(f(\boldsymbol{x}_i;\theta))$ is a convex function

Proof:

$$\nabla_\theta^2\left[-\log\left(f(\boldsymbol{x}_i;\theta)\right)\right]$$

$$= \nabla_\theta\left[\left(\frac{1}{1+\exp(-\theta^T x_i)} - 1\right) x_i\right] =$$

Convexity of Regression Loss

$-\log(f(\boldsymbol{x}_i;\theta))$ is a convex function

Proof:

$$\nabla_\theta^2\left[-\log\left(f(\boldsymbol{x}_i;\theta)\right)\right]$$

$$= \nabla_\theta\left[\left(\frac{1}{1+\exp(-\theta^T x_i)} - 1\right)x_i\right]$$

$$= \frac{1}{1+\exp(-\theta^T x_i)}\frac{\exp(-\theta^T x_i)}{1+\exp(-\theta^T x_i)}x_i x_i^T$$

Convexity of Regression Loss

$-\log(f(\boldsymbol{x}_i;\theta))$ is a convex function

Proof:

$$z^T\ \nabla_\theta^2\left[-\log\left(f(\boldsymbol{x}_i;\theta)\right)\right]z =$$

$$z^T\ \frac{1}{1+\exp(-\theta^T x_i)}\frac{\exp(-\theta^T x_i)}{1+\exp(-\theta^T x_i)} x_i x_i^T z$$

$$= \frac{1}{1+\exp(-\theta^T x_i)}\frac{\exp(-\theta^T x_i)}{1+\exp(-\theta^T x_i)}\left(x_i^T z\right)^2$$

Convexity of Regression Loss

$-\log(1 - f(\boldsymbol{x_i}; \theta))$ is a convex function

Convexity of Regression Loss

$-\log(1 - f(\boldsymbol{x}_i; \theta))$ is a convex function

Proof:

$$\nabla_\theta\left[-\log\left(1 - f(\boldsymbol{x}_i; \theta)\right)\right] =$$

$$= \nabla_\theta\left[-\log\left(\frac{\exp(-\theta^T x_i)}{1 + \exp(-\theta^T x_i)}\right)\right]$$

$$= \nabla_\theta[\theta^T x_i + \log(1 + \exp(-\theta^T x_i))]$$

$$= x_i + \nabla_\theta[\log(1 + \exp(-\theta^T x_i))]$$

Convexity of Regression Loss

$-\log(1 - f(\boldsymbol{x_i}; \theta))$ is a convex function

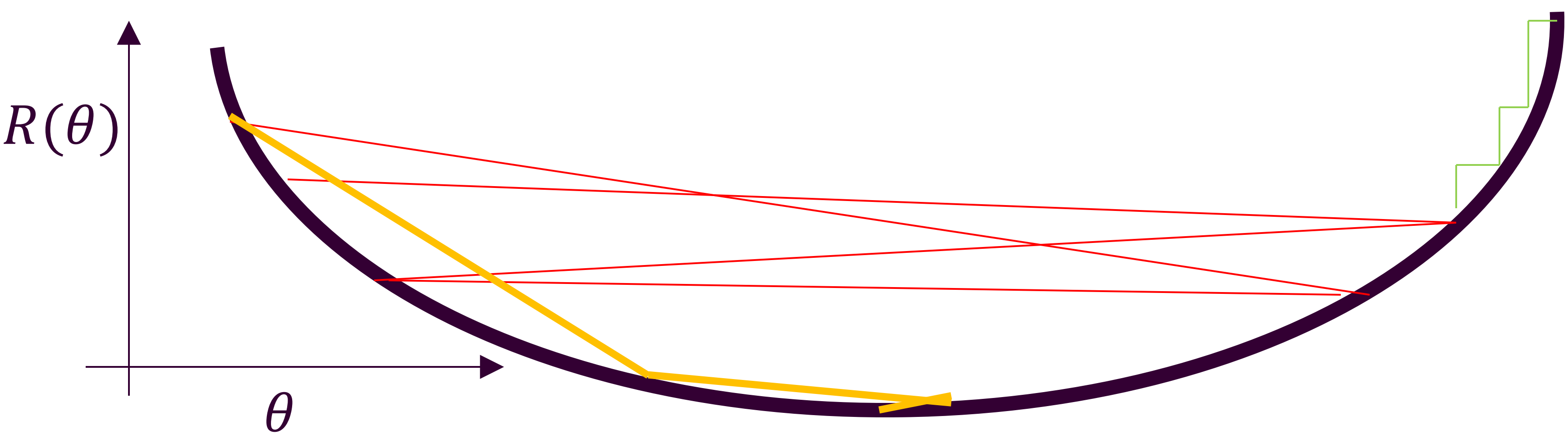
Proof:

$$\nabla_\theta^2\left[-\log\left(1 - f(\boldsymbol{x_i}; \theta)\right)\right] =$$
$$= \nabla_\theta^2[\log(1 + \exp(-\theta^T x_i))]$$

How fast to descend?

How fast to descend?

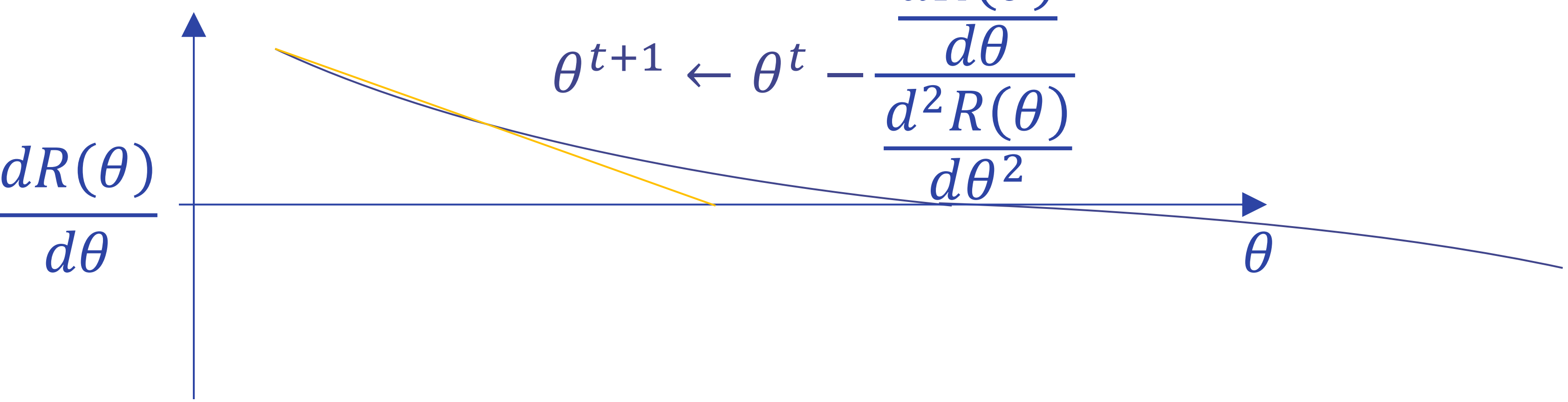
- Not too fast, not too slow, just right

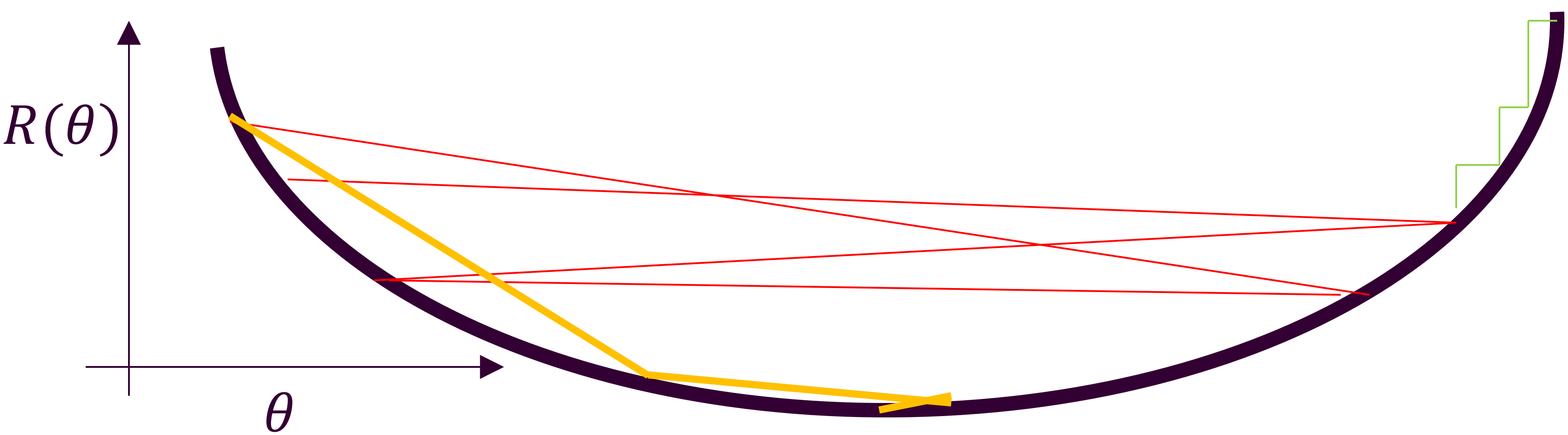


How fast to descend?

- Newton's method for the derivative

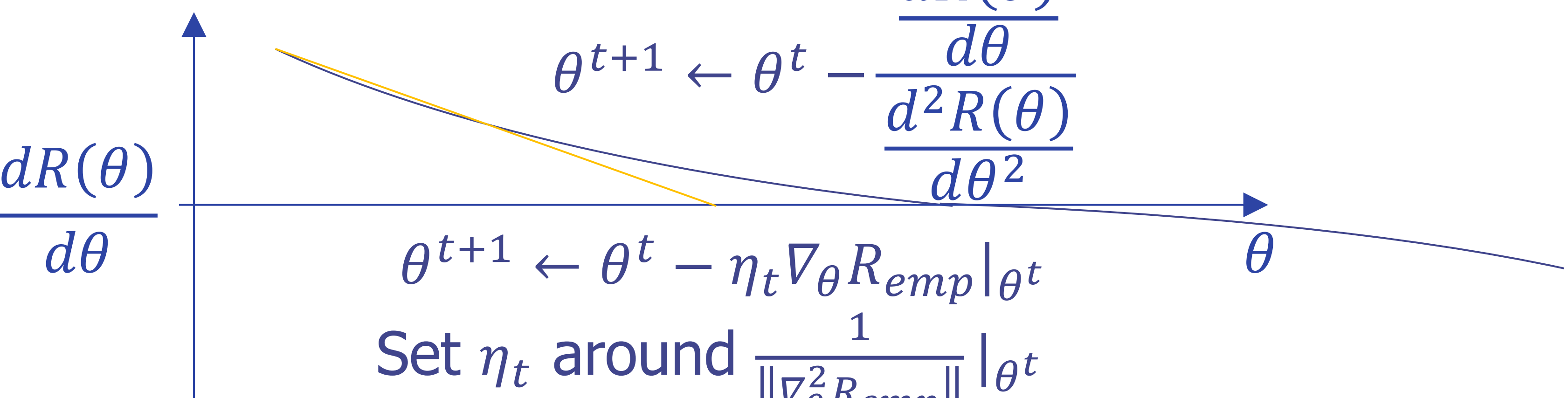
$$\theta^{t+1} \leftarrow \theta^{t} - \frac{\frac{dR(\theta)}{d\theta}}{\frac{d^2R(\theta)}{d\theta^2}}$$





How fast to descend?

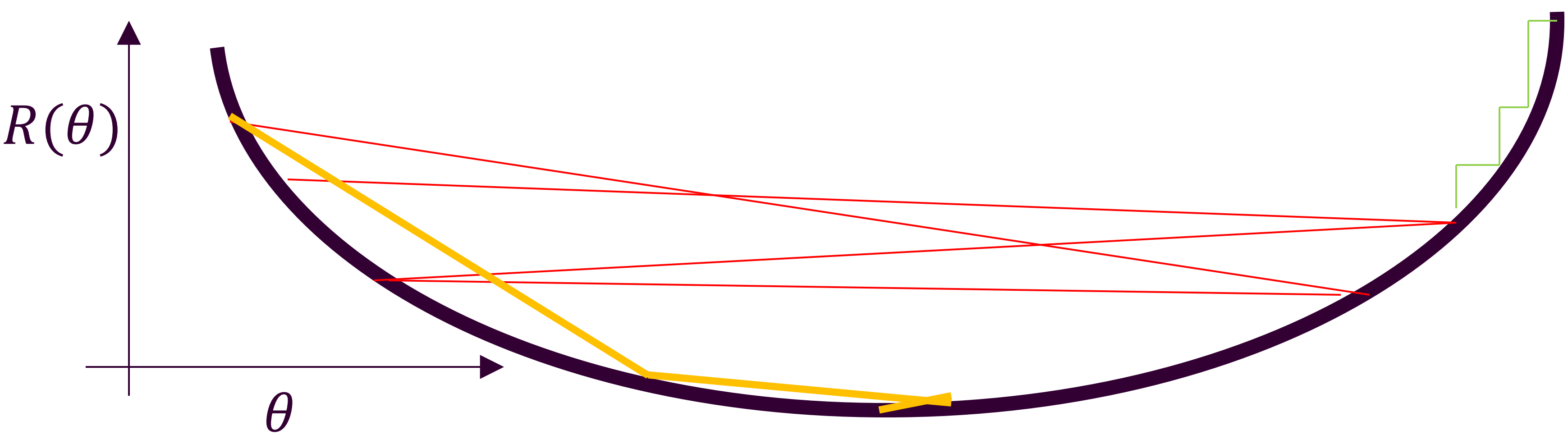
- Newton's method for the derivative



$$\theta^{t+1} \leftarrow \theta^{t} - \frac{\frac{dR(\theta)}{d\theta}}{\frac{d^2R(\theta)}{d\theta^2}}$$

$$\theta^{t+1} \leftarrow \theta^{t} - \eta_t \nabla_\theta R_{emp}|_{\theta^t}$$

Set η_t around $\frac{1}{\|\nabla_\theta^2 R_{emp}\|}|_{\theta^t}$



Summary

- Classification
- Logistic Regression
- Gradient Descent