Recap: LS Regression $y \in \mathbb{R}$, $x \in \mathbb{R}^d$

$$p(y \mid x, \theta) = N(w^T x + b, \sigma^2)$$

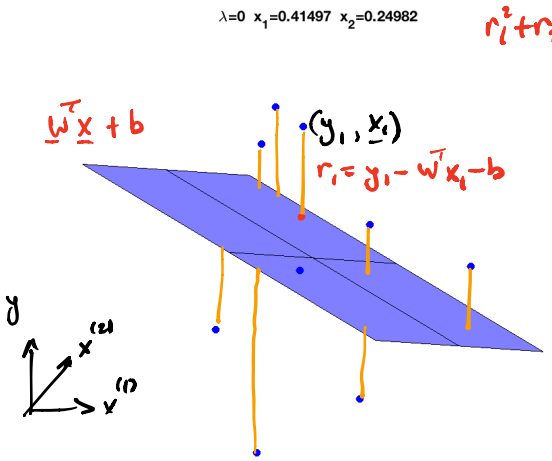
$$\hat{w}_{MAP} = \arg\max_w \sum_{i=1}^{n} \log P(y_i \mid x_i, \theta) + \log P(w)$$

$$P(w) = N(0, t^2 I)$$

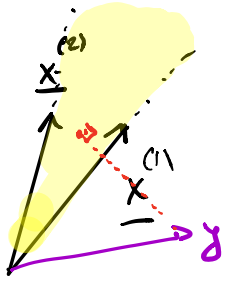
$$= \arg\min_{w,b} \sum (y_i - w^T x_i - b)^2 + \underbrace{\frac{\sigma^2}{t^2}}_{=\lambda} \cdot \|w\|_2^2$$

$\lambda = 0$ Linear Least Squares

① Row-wise interpretation



$$\underset{n \times d}{\underline{\underline{X}}} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$



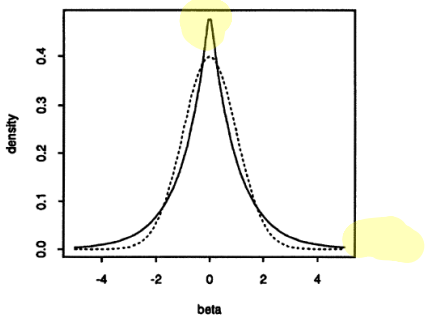
② Column-wise

$$\underline{\underline{X}} = \begin{bmatrix} \underline{X}^{(1)} & \cdots & \underline{X}^{(d)} \end{bmatrix}$$

L1 Regularization (Lasso)

$$\arg\min \sum_i (y_i - \underline{w}^T \underline{x}_i - b)^2 + \lambda \|w\|_1$$

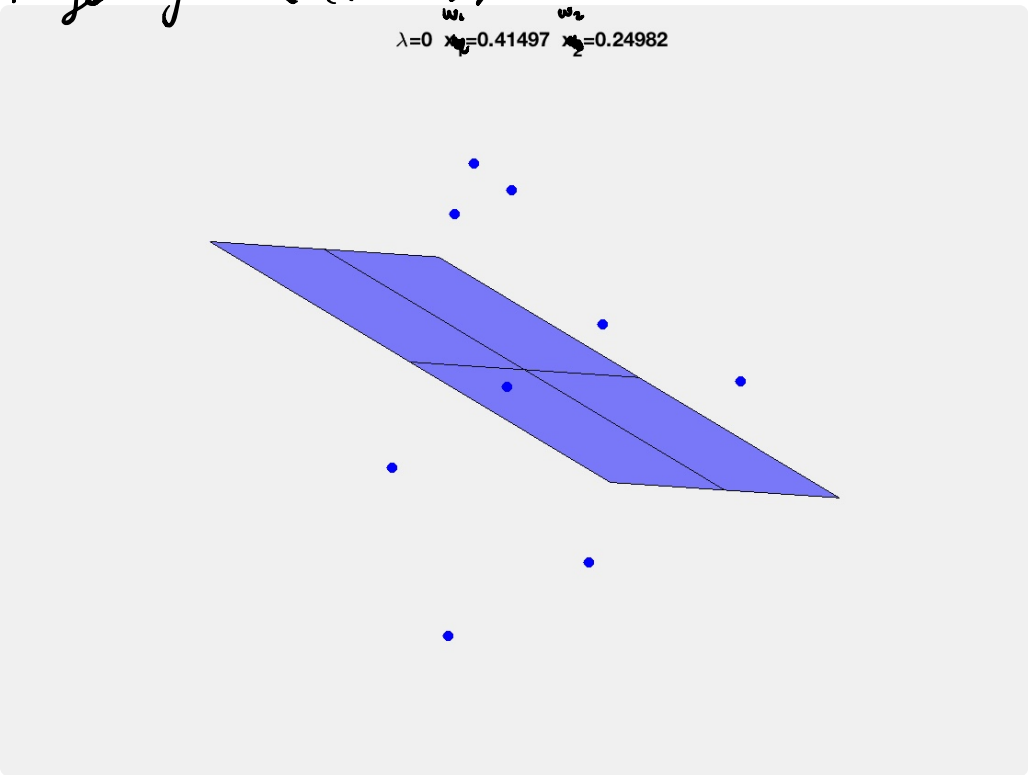
Laplace prior $\propto e^{\frac{-|w_1|}{t}} \cdot e^{\frac{-|w_2|}{t}} \cdots e^{\frac{-|w_d|}{t}} = e^{-\frac{1}{t}\|w\|_1}$



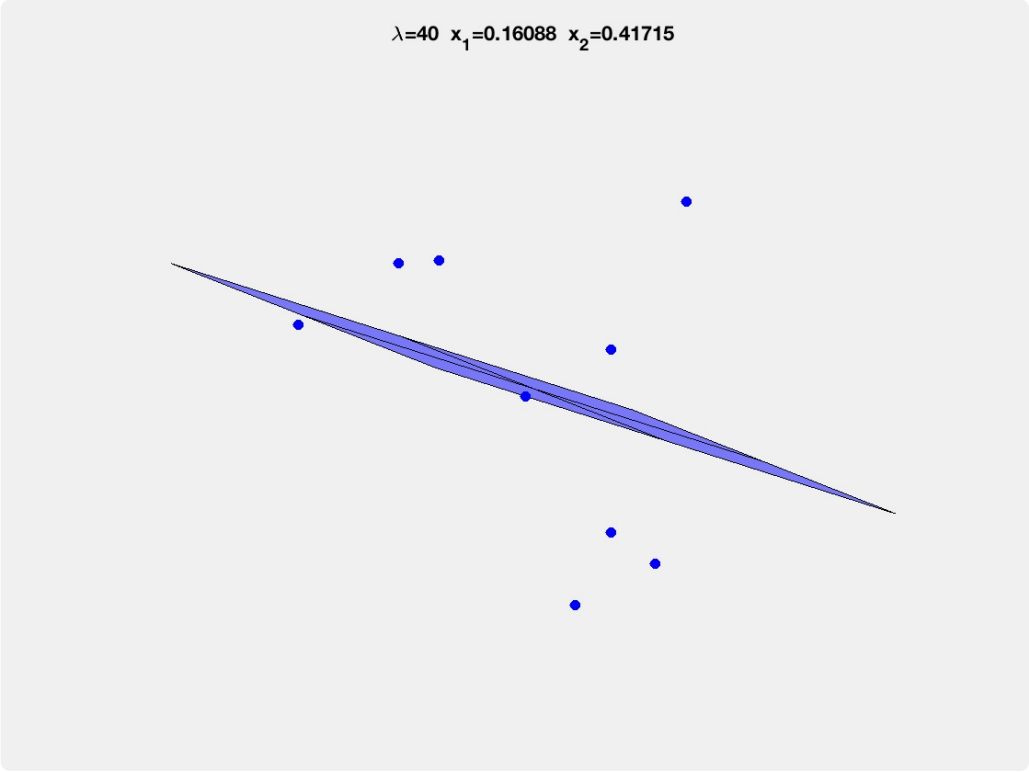
(Laplacian)

Fig. 7. Double-exponential density (———) and normal density (- - - - -): the former is the implicit prior used by the lasso; the latter by ridge regression

Ridge Regression ($\lambda \|w\|_2^2$)



Lasso ($\lambda \|w\|_1$)



Distribution of $\hat{w}_{LS}$

$$y_i = w^T x_i + e_i \qquad e_i \sim N(0, \sigma^2)$$

$$\underline{y} = \underline{\underline{X}}\,\underline{w} + \underline{e}$$

$(X^T X)^{-1} X^T$ → Pseudo-inverse of X

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y$$
$$= (X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T e$$
$$= w + (X^T X)^{-1} X^T e$$

$$E\,\hat{w}_{LS} = w$$

$$E(\hat{w}_{LS} - w)(\hat{w}_{LS} - w)^T = E\,(X^T X)^{-1} X^T e e^T X (X^T X)^{-1}$$
$$= (X^T X)^{-1} X^T \cdot \sigma^2 I \cdot X (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1}$$

$$\hat{w}_{LS} \sim N(w, \sigma^2 (X^T X)^{-1})$$

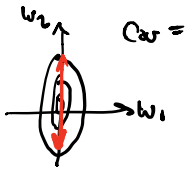
$$X = U \Sigma V^T \qquad X^T X = V \Sigma \underbrace{U^T U}_{I} \Sigma V^T = V \Sigma^2 V^T$$

$$(X^T X)^{-1} = V \Sigma^{-2} V^T$$

Eg.

$$(X^T X) = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$$



$$Cov = \begin{matrix} 1 & 0 \\ 0 & 10 \end{matrix}$$

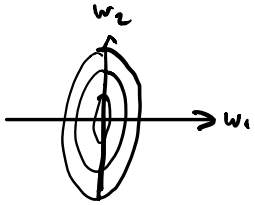
$$\|Xw - y\|_2^2$$

Hessian: $\min_w L(w)$ $\qquad \nabla^2 L(w) = X^T X$

eg $X^T X = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$

$$\begin{matrix} 1 & 0 \\ 0 & 0.1 \end{matrix}$$

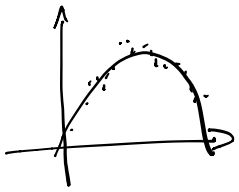
$$w_1^2 + 0.1 w_2^2$$



$$\hat{w}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$
$$= V (\Sigma^2 + \lambda I)^{-1} V^T V \Sigma U^T y$$

$$= V(\Sigma + \lambda I)^{-1} U^T y = \sum_i v_i u_i^T y \cdot \frac{w_i}{\sigma_i^2 + \lambda}$$

Nonlinear features



$$f(x, \theta) = b + \sum_{j=0}^{m-1} w_j \phi_j(x)$$

$\phi_j(x)$ are basis functions

$$\theta = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$\min \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$$

Examples: $x \in \mathbb{R}$ $\quad \phi_j(x) = x^j$ $\quad f(x)$ is a degree $m-1$ polynom

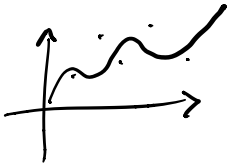
$\underline{x} \in \mathbb{R}^2$ $\quad \underline{\phi}(\underline{x}) = \left(1, x^{(1)}, x^{(2)}, (x^{(1)})^2, (x^{(2)})^2, x^{(1)} \cdot x^{(2)}\right)$

Gaussian basis $\phi_j(x) = e^{-\frac{\|\underline{x} - \mu_j\|_2^2}{2\sigma_j^2}}$ $\quad x \in \mathbb{R}^d$

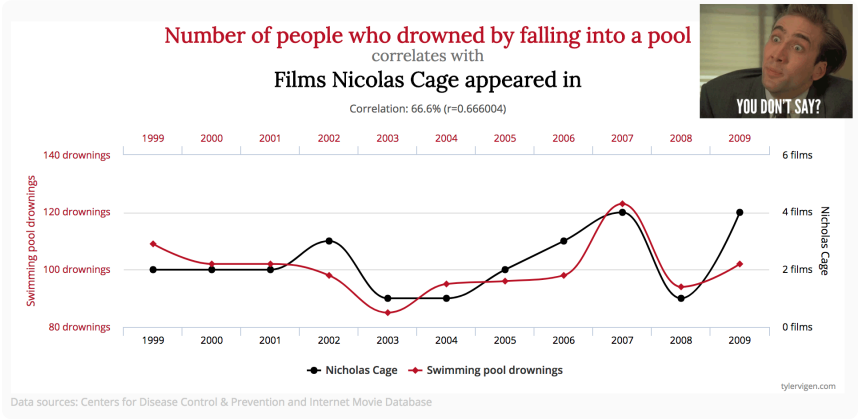
$$\min \|y - \underline{\underline{\Phi}} w\|_2^2$$

$$\underline{\underline{\Phi}} = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{m-1}(x_1) \\ \vdots & & \\ \phi_0(x_n) & \cdots & \phi_{m-1}(x_n) \end{bmatrix}$$

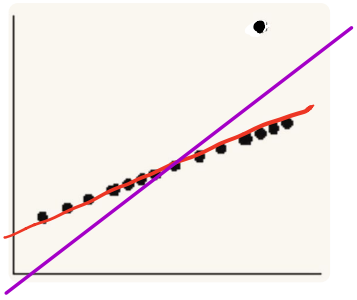
→ overfit



→ spurious correlation



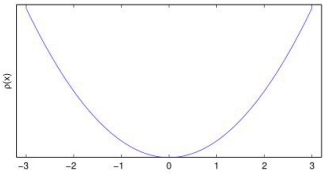
Robust Regression



LS is not robust to outliers

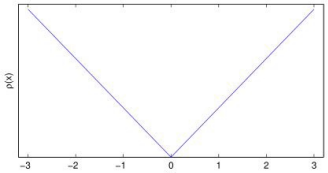
$$\min \sum_{i=1}^{n} \rho(y_i - w^T x_i - b)$$

Where ρ is a robust loss function



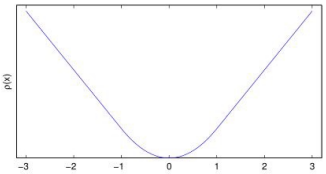
squared loss

(a)



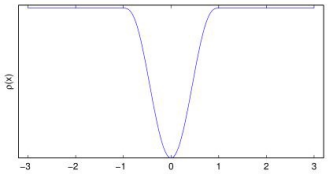
$\rho(x) = |x|$ absolute error

(b)



$\rho(x) = \begin{cases} x^2 & |x| \le 1 \\ 2|x| - 1 & |x| > 1 \end{cases}$ Huber

(c)



Biweight

(d)

Kernels

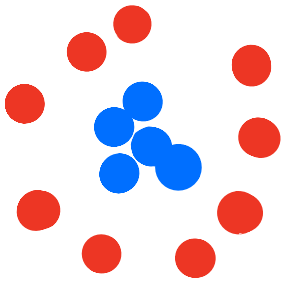
Non linear feature maps: transform feature vectors via

$$\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}^m$$

and apply a linear method (e.g. least-squares, SVM) to the transformed data $\Phi(x_1), \ldots, \Phi(x_n)$.

Classification: $u \longrightarrow \text{sign}(w^T \Phi(x) + b)$

Example $\underline{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \in \mathbb{R}^2$ Consider $\Phi(x) = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(1)} \cdot x^{(2)} \\ (x^{(1)})^2 \\ (x^{(2)})^2 \end{bmatrix}$



Can be separated by a circular classifier

$$y \rightarrow \text{sign}\left((x^{(1)} - c_1)^2 + (x^{(2)} - c_2)^2 - r^2 \right)$$

This is a linear classifier in the transformed space $\Phi(x)$