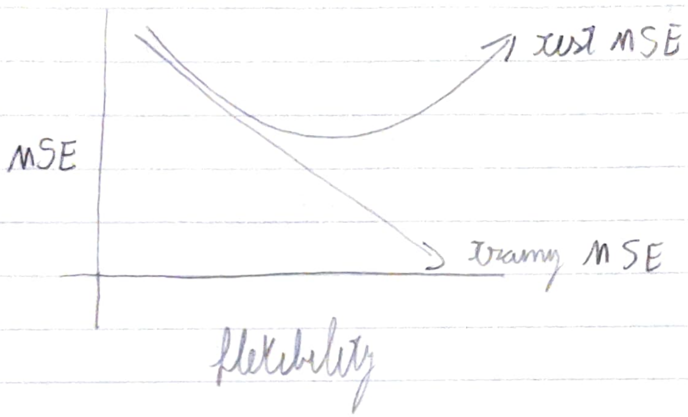
01/31 Lecture

Daniel Vidal

Measuring quality of fit

ref: ISL 2.2.1, 2.2.2
ESL 7.2, 7.3
Bishop 3.2

$$MSE = \frac{1}{n}\sum (y_i - f(x_i; w))^2$$



* see matlab example

- want a function that is smooth and flexible
- large weights suggest overfitting.

Avoiding overfitting regularization   ref Bishop 3.1.4

Least Squares Method: $\min \sum (y_i - f(x_i, w))^2$ to find w
(w is a vector of weights)

given w/ data (x_i, y_i) $i = 1, \dots n$

Regularized LS: $\min \left[\sum (y_i - f(x_i; w))^2 + \underbrace{\lambda \sum_{i=1}^{m} w_i^2}_{\text{regularization term}} \right]$

regularization constant
(hyper parameter)

in case of lin or poly regression

$$= \min \underbrace{\|y - Xw\|_2^2}_{w = (X^T X)^{-1} X^T y} + \lambda \|w\|_2^2$$

note:
$\|a\|_2^2 + \|b\|_2^2 = \left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2$
$\sum a^2 + \sum b^2$

$$= \min_w \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} w \right\|_2^2$$

$$w = \left(\begin{bmatrix} X^T & -\sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X \\ -\sqrt{\lambda} I \end{bmatrix} \right)^{-1} \begin{bmatrix} X^T & -\sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} = (X^T X + \lambda I)^{-1} X^T y$$

when regularization is quadratic ($\lambda \sum_{i=1}^{m} w_i^2$), ridge regression

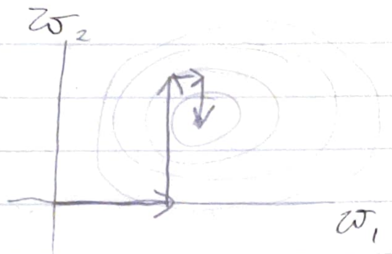
lasso $\quad \min_w \left[\sum (\quad)^2 + \lambda \sum |w_i| \right]$ $\qquad \lambda \|w\|_1 \rightarrow$ lasso

the weights are restrained to be small and sparse

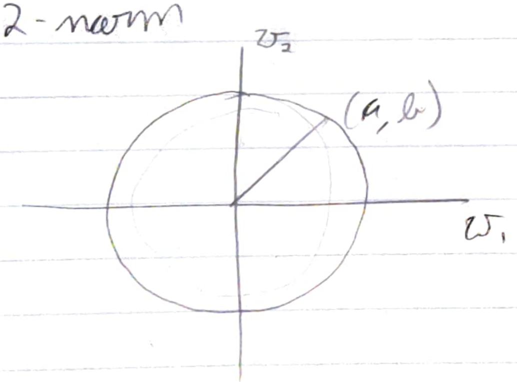
many zero weights

finding w is much more complex

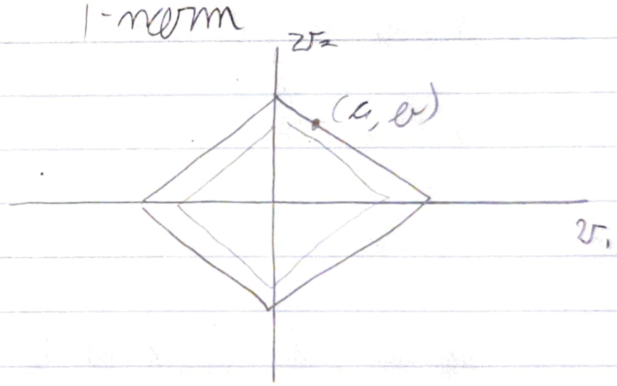
— coordinate descent



2-norm



1-norm



read ahead: bishop 3.1