Last Time:
- Polynomial Regression
- Overfitting vs. underfitting

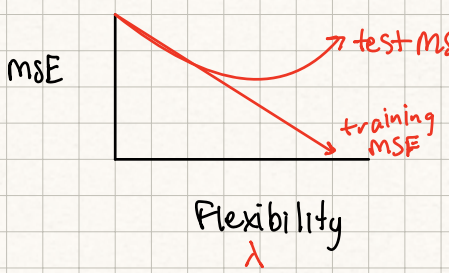
Ref
ISL: 2.2.1, 2.2.2
ESL: 7.2, 7.3
Bishop: 3.2

- Machine Learning finds patterns to data that human intuition cannot find

- Test data is not used in training and is used to measure fit

  - Measuring quality of fit

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



Demo Notes:
Overfitting the data results in an unintuitive fit that doesn't accurately model data

Extremely large weights can also indicate overfitting

Regularization: Avoiding Overfitting

Ref: Bishop 3.1.4

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

LS method: $\min_w \sum (y_i - f(x_i; w))^2$ to find w given (x_i, y_i), $i = 1 \dots n$

$$\min_w \left[\sum (y_i - f(x_i; w))^2 + \lambda \sum_{i=1} w_i^2\right] = \min \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

ridge regression

positive regularization constant

regularization term

Also known as hyper parameter

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

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

not used numerically

Lasso: $\min_w \left[\sum (\quad)^2 + \lambda |w_i|\right]$

finding w is much more complicated
- coordinate descent

the weights are constrained to be small and also sparse → (many zero weights)

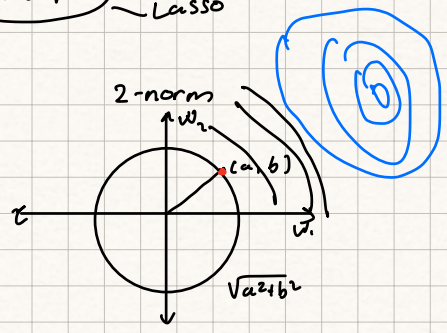
Demo Notes:

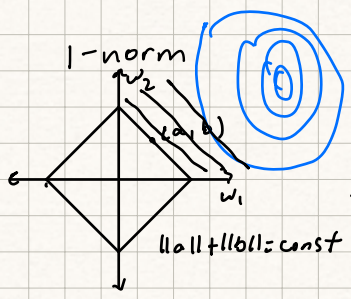
weights are reasonable at high order, more intuitive fit

- Ridge Reg.
  - Plotting with λ as ind. var.
  - As λ inc. MSE inc.
  - As λ dec., MSE dec.

Why is solution sparse?

$\lambda \|w\|$ — Lasso





many points of non-intersection