Lecture 10: Tikhinov Regularisation / Intro to SVD

Recall: $X \in \mathbb{R}^{n \times p}$ (n training samples, p features)

$y \in \mathbb{R}^n$ (n labels)

Model: $y \approx Xw$, $y_i \approx x_i^T w$ for some $w \in \mathbb{R}^p$

Least Squares: $\hat{w}_{LS} = \underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T w)^2$

- if X is full rank (cols are linearly indep), then $\hat{w}_{LS}$ is unique and $\hat{w}_{LS} = (X^T X)^{-1} X^T y$
- if X not full rank, $X^T X$ is not invertible

  $\hat{w}_{LS}$ is unique.

Alternative: Tikhinov Regularization / Ridge Regressions

$$\hat{w} = \arg\min_{w} \underbrace{\|y - Xw\|_2^2}_{\text{measures fit to data}} + \lambda \underbrace{\|w\|_2^2}_{\text{"regularizer" measures "energy" in } w}$$

$\lambda > 0$ "regularization parameter" or "tuning parameter"

- $\hat{w}$ unique even when no least squares soln exists.
- even when X is full ~~rank~~ rank, it can be "badly behaved", and regularization adjusts for this.

e.g. $X^TX = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-12} \end{bmatrix}$ $(X^TX)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{12} \end{bmatrix}$

Let $f(\vec{w}) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2$

$= y^Ty - 2w^TX^Ty + w^TX^TXw + \lambda w^Tw$

$= y^Ty - 2w^TX^Ty + w^T(X^TX + \lambda I)w$

$\nabla_w f = 0 - 2X^Ty + 2(X^TX + \lambda I)w = 0$

if $(X^TX + \lambda I)$ is invertible, then $\hat{w} = (X^TX + \lambda I)^{-1}X^Ty$

always true

note: $X^TX + \lambda I$ is always invertible

recall Q is invertible is it is positive definite

i.e. $a^TQa > 0$ for all $\boxed{a \neq 0}$

$$a^T(X^TX + \lambda I)a = a^TX^TXa + \lambda a^Ta$$

$$= \underbrace{\|Xa\|_2^2}_{\geq 0} + \underbrace{\lambda \|a\|_2^2}_{>0} > 0$$

alternative derivation of $\hat{w}$:

1st note $\|a\|_2^2 + \|b\|_2^2 = \left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2$

$= \sum a_i^2 + \sum b_i^2$ $\quad \sum a_i^2 + b_i^2$

so $\|y - Xw\|_2^2 + \lambda \|w\|_2^2 = \|y - Xw\|_2^2 + \|\sqrt{\lambda}\, w\|_2^2$

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

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

$$\hat{w} = \underset{w}{\operatorname{argmin}} \|\tilde{y} - \tilde{X} w\|_2^2 \Rightarrow \hat{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y}$$

$$\tilde{X}^T \tilde{X} = \begin{bmatrix} X^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} = X^T X + \lambda I$$

$$\tilde{X}^T \tilde{y} = \begin{bmatrix} X^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} = X^T y$$

$$\Rightarrow \hat{w} - (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y = (X^T X + \lambda I)^{-1} X^T y$$

example: $y = Xw + \varepsilon \iff y_i = x_i^T w + \varepsilon_i$

ε: noise or errors in y or our model

~~gen~~

$$X = \begin{bmatrix} 1 & 1 & .01 \\ 1 & -1 & .01 \\ 1 & 1 & -.01 \\ 1 & -1 & -.01 \end{bmatrix}$$

cols of X are lin. indep. (orthogonal)

$\Rightarrow X$ is full rank, X^TX is invertible.

$$X^TX = 4 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-4} \end{bmatrix} \Rightarrow (X^TX)^{-1} = \frac{1}{4} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^4 \end{bmatrix}$$

If least squares:

$$\hat{w} = (X^T X)^{-1} X^T y$$

$$= (X^T X)^{-1} X^T (Xw + \varepsilon)$$

$$= (X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T \varepsilon$$

$$= w + (X^T X)^{-1} X^T \varepsilon$$

good

$$\begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 10^4 \end{bmatrix}$$

LS soln exists (unique)
but small noise in ε gets amplified by large values in $(X^T X)^{-1}$

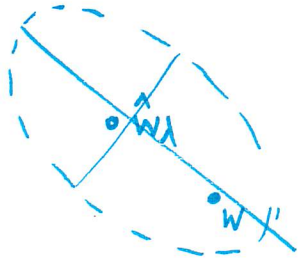
$$(X^T X + \lambda I)$$

$$= \begin{bmatrix} 4+\lambda & 0 & \\ & 4+\lambda & \\ 0 & & 4\cdot 10^{-4} + \lambda \end{bmatrix}$$

$$(X^T X + \lambda I)^{-1} = \begin{bmatrix} 1/4+\lambda & 0 & \\ & 1/4+\lambda & \\ 0 & & \approx 1/\lambda \end{bmatrix}$$

↳ avoids amplifying noise

Singular Value Decomposition



Finding λ

- split data into training and validation test set. $(X_i^{(t)}, y_i^{(t)}, X_i^{(v)}, y_i^{(v)})$

- consider a set of possible λ's : $\Lambda = \{\lambda_1, \lambda_2, .., \lambda_M\}$

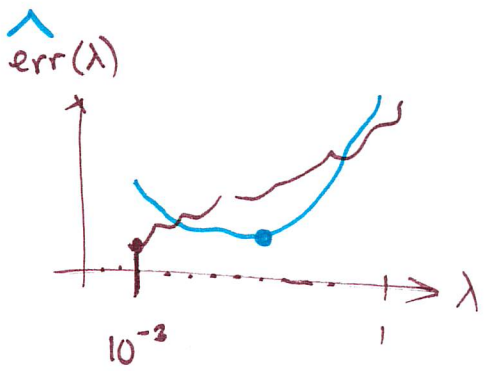
- for each $\lambda \in \Lambda$, compute $\hat{w}_\lambda = (X^{(t)T} X^{(t)} + \lambda I)^{-1} X^{(t)T} y$

measure error of $\hat{w}_\lambda$ on validation set

$$\hat{err}(\lambda) = \| y^{(v)} - X^{(v)} \hat{w}_\lambda \|_2^2$$

choose $\lambda_{opt} = \underset{\lambda}{\text{argmin}}\ err(\lambda)$

$$\hat{w} = (X^T X + \lambda_{opt} I)^{-1} X^T y$$



$$err(\lambda) = E\,(y - x^T \hat{w}_\lambda)^2$$

$$\hat{w} = \arg\min \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

Sing. Val. Decomp. (SVD)

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix}}_{X} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ 2 \end{bmatrix}}_{y} \Rightarrow X^{-1} y = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix}}_{X} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ 2.01 \end{bmatrix}}_{y} \Rightarrow X^{T} y = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

system is very sensitive to small changes in y.

SVD will help us quantify how well-behaved X is in terms of it's inverse or rank

~~another eko~~