

# Regularization and Ridge Regression for Supervised Learning

# Objectives

- Extend supervised learning to rank deficient  $\underline{A}$

- Introduce ridge regression and regularization to control bias/variance tradeoff

- Solve ridge regression problem

# Supervised Learning Solves Linear Equations<sup>2</sup>

Classification/Modeling:  $\underline{A} \underline{w} \approx \underline{d}$        $\underline{x}_i^T \underline{w} \approx d_i$   
 ( $N \times P$ ) ( $P \times 1$ )    ( $N \times 1$ )      feature model      label

Least Squares Solution       $\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2$

Overdetermined system:  $N \geq P$ ,  $\text{rank}(\underline{A}) = P$

$$\underline{w} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{d} \quad \underline{A}^T \underline{A} \text{ invertible}$$

Underdetermined system:  $N < P$  or  $\text{rank}(\underline{A}) < P$

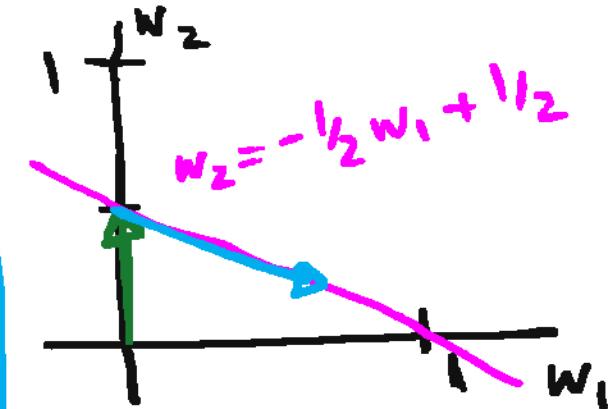
No unique solution  $\underline{w}$ !

Example:  $N=1, P=2$      $\underline{A} = [1 \ 2]$ ,  $d=1$

$$\underline{A} \underline{w} = \underline{d} \Rightarrow w_1 + 2w_2 = 1 \Rightarrow w_2 = -\frac{1}{2}w_1 + \frac{1}{2}$$

Infinite # solns zero error

$$\underline{w} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + d \begin{bmatrix} 1 \\ -0.5 \end{bmatrix}$$



How do we choose a unique solution? 3

if  $\text{rank}(\underline{A}) < P$  and  $\underline{A}\underline{w}_0 = \underline{d}$ , then  $\underline{A}\underline{w} = \underline{d}$  for  $\underline{w} = \underline{w}_0 + \underline{v}$   
where  $\text{cols}(\underline{v})$  span space  $\perp \text{span}\{\underline{A}\}$  ( $\underline{A}\underline{v} = \underline{0}$ )

Squared error is not sufficient to give unique solution

Consider robustness to errors/noise  $\underline{\Sigma}$

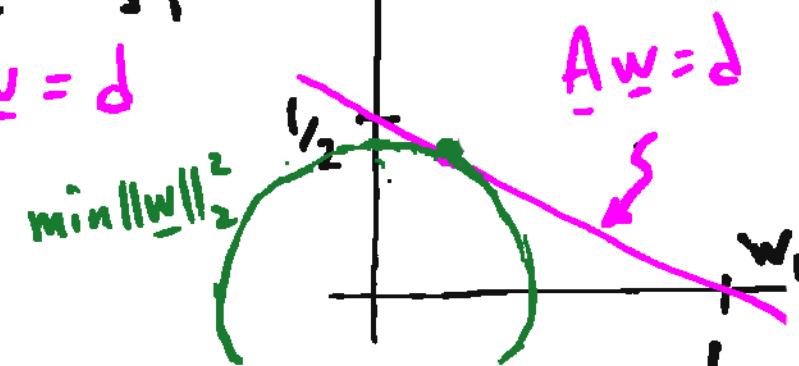
$$\underline{x}_i^T \underline{w} = d_i \Rightarrow (\underline{x}_i + \underline{\Sigma})^T \underline{w} = d_i + \underline{\Sigma}^T \underline{w} \quad \underline{\Sigma}^T \underline{w}: \text{"variance" term}$$

$$|\underline{\Sigma}^T \underline{w}|^2 = \|\underline{w}\|_2^2 \|\underline{\Sigma}\|_2^2 \cos^2 \theta_{\underline{\Sigma}, \underline{w}}$$
 no info on  $\underline{\Sigma} \Rightarrow \min \|\underline{w}\|_2^2$

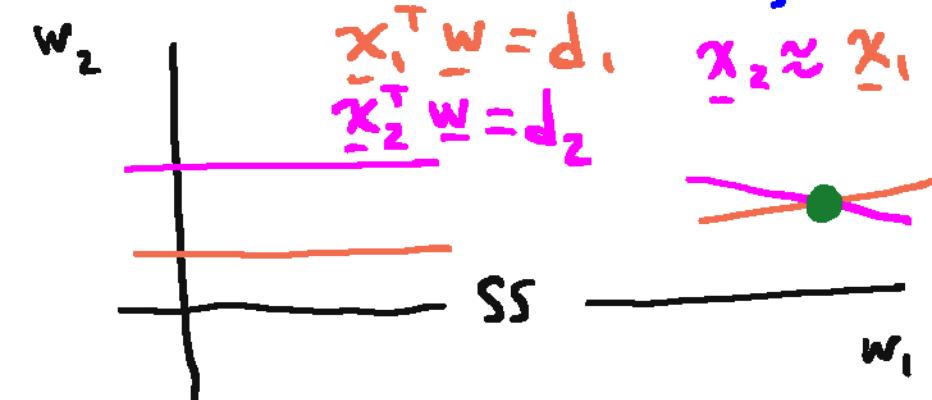
Example:  $\underline{A} = [1 \ 2]$ ,  $d = 1$

$$\min_{\underline{w}} \|\underline{w}\|_2^2 \text{ s.t. } \underline{A}\underline{w} = \underline{d}$$

$$\begin{aligned} \min_{w_2} & 5w_2^2 - 4w_2 + 1 \\ \Rightarrow \underline{w} = & \begin{bmatrix} .2 \\ .4 \end{bmatrix} \end{aligned}$$



Similar features  $\Rightarrow$  large  $\|\underline{w}\|_2^2$



Choose  $\underline{w}$  to balance bias and variance

4

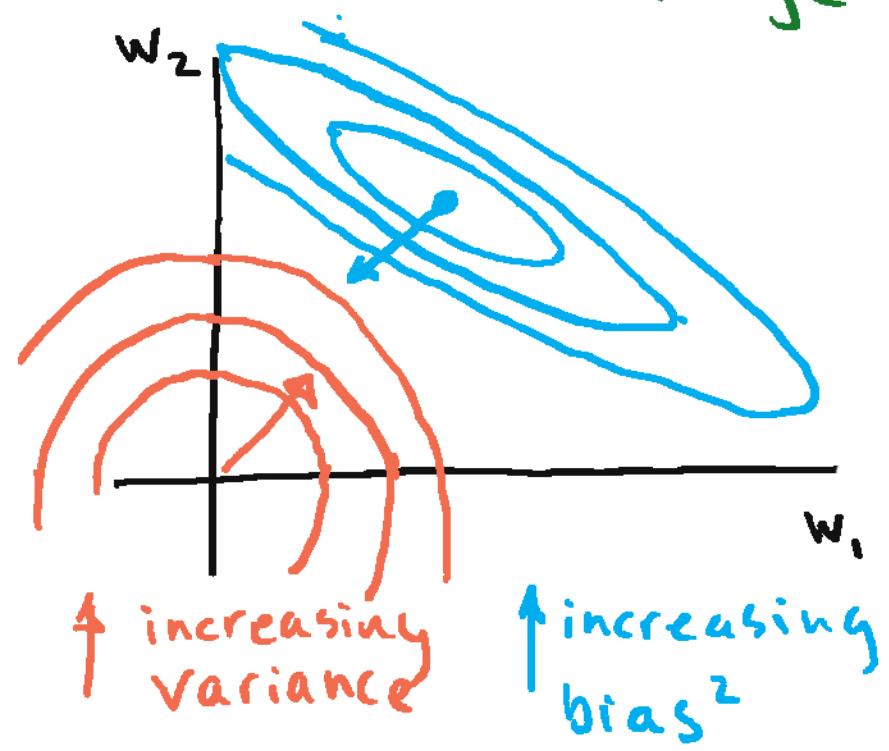
Consider  $(\underline{x}_i + \underline{\epsilon})^T \underline{w} - d_i = \underbrace{\underline{x}_i^T \underline{w} - d_i}_{\text{bias}} + \underbrace{\underline{\epsilon}^T \underline{w}}_{\text{variance term}}$

$$\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda \|\underline{w}\|_2^2$$

bias<sup>2</sup>
variance

ridge regression

bias<sup>2</sup>       $\|\underline{w}\|_2^2$  - regularizer  
 $\lambda$  - regularization parameter  
 (Tikhonov regularized)



**Solution:**  $\|\underline{a}\|_2^2 + \|\underline{b}\|_2^2 = \underline{a}^\top \underline{a} + \underline{b}^\top \underline{b} = \|\begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix}\|_2^2$

$$\min_w \left\| \begin{bmatrix} A_w - d \\ \lambda^T w \end{bmatrix} \right\|_2^2 \rightarrow \min_w \left\| \begin{bmatrix} A \\ \lambda^T I \end{bmatrix} \begin{bmatrix} w \\ 1 \end{bmatrix} - \begin{bmatrix} d \\ 0 \end{bmatrix} \right\|_2^2$$

$$\underline{W} = (\tilde{\underline{A}}^T \tilde{\underline{A}})^{-1} \tilde{\underline{A}}^T \tilde{\underline{d}} = ([\underline{A}^T \quad \lambda^2 \underline{I}] \begin{bmatrix} \underline{A} \\ \lambda^2 \underline{I} \end{bmatrix})^{-1} \underline{A}^T \underline{d}$$

$\underline{A}^T \underline{A} + \lambda \underline{I} > 0$  ( $\lambda > 0$ )  $\rightarrow$  inverse exists

Ridge regression results in a robust solution 5

Suppose  $\underline{d} = \underline{A}\underline{w} + \underline{n}$   $\underline{n}$ : model error or label error

Least Squares:  $\underline{w} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{d} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{A} \underline{w} + (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{n}$

$$= \underline{w} + (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{n} \leftarrow \text{variance}$$

Ridge:  $\underline{w} = (\underline{A}^T \underline{A} + \lambda \underline{I})^{-1} \underline{A}^T \underline{d} = (\underline{A}^T \underline{A} + \lambda \underline{I})^{-1} \underline{A}^T \underline{A} \underline{w} + (\underline{A}^T \underline{A} + \lambda \underline{I})^{-1} \underline{A}^T \underline{n}$

bias variance

LS is problematic (large variance term) when  $\underline{A}$  is nearly low rank

Example:

$$\underline{A} = \frac{1}{2} \begin{bmatrix} 1 & 1 & .001 \\ 1 & -1 & .001 \\ 1 & 1 & -.001 \\ 1 & -1 & -.001 \end{bmatrix}$$

$$\underline{A}^T \underline{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-6} \end{bmatrix} \rightarrow (\underline{A}^T \underline{A})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^6 \end{bmatrix} \rightarrow \text{large variance!}$$

$$(\underline{A}^T \underline{A} + \lambda \underline{I})^{-1} = \begin{bmatrix} \frac{1}{1+\lambda} & 0 & 0 \\ 0 & \frac{1}{1+\lambda} & 0 \\ 0 & 0 & \frac{1}{10^{-6}+\lambda} \end{bmatrix} \approx \begin{bmatrix} \frac{1}{1+\lambda} & 0 & 0 \\ 0 & \frac{1}{1+\lambda} & 0 \\ 0 & 0 & \frac{1}{\lambda} \end{bmatrix} \rightarrow \begin{array}{l} \lambda \text{ controls} \\ \text{Variance} \end{array}$$

or orthogonal cols,  $\text{rank}(\underline{A}) = 3$

# Use Cross validation to choose $\lambda$

6

- 1) Choose  $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$  (often logarithmic spacing)
- 2) Split data: training ( $t$ ) and validation ( $v$ ) sets

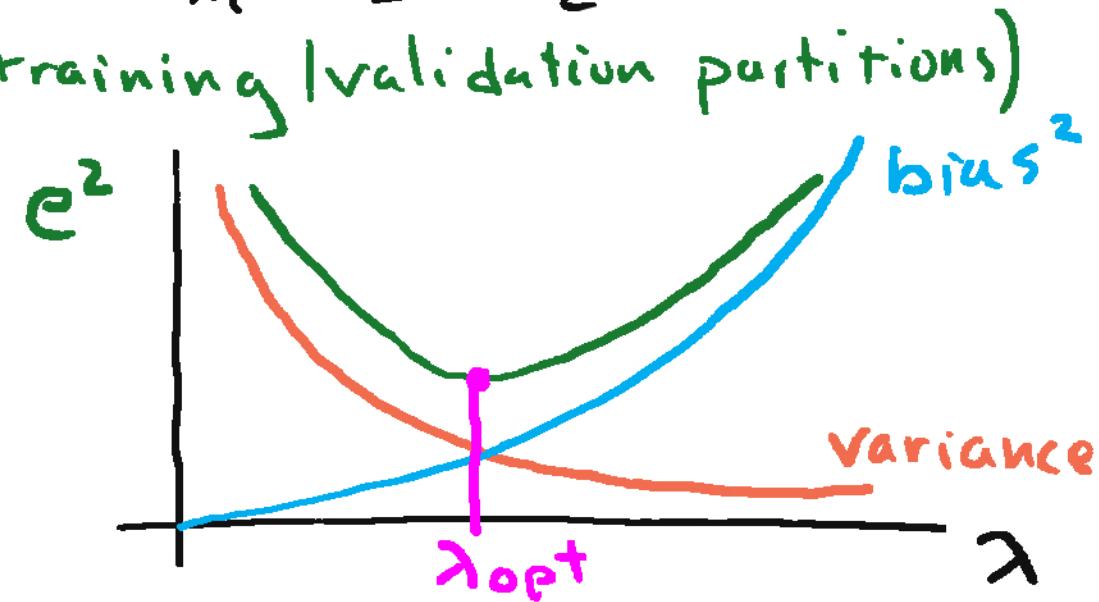
$$\underline{A}^{(t)}, \underline{d}^{(t)}, \underline{A}^{(v)}, \underline{d}^{(v)}$$

$$3) \text{ Compute } \underline{w}_{\lambda_i} = (\underline{A}^{(t)T} \underline{A}^{(t)} + \lambda_i \underline{I})^{-1} \underline{A}^{(t)T} \underline{d}^{(t)}$$

$$4) \text{ Compute } e^2(\lambda_i) = \|\underline{A}^{(v)} \underline{w}_{\lambda_i} - \underline{d}^{(v)}\|_2^2$$

(average over different training/validation partitions)

- 5) Choose  $\lambda$ : that minimizes  $e^2(\lambda_i)$



**Copyright 2019  
Barry Van Veen**