

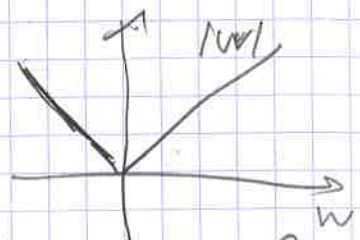
Lasso regul^o

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N (w x_n - t_n)^2 + \lambda \|w\|_1$$

$$\begin{aligned} \nabla \mathcal{L} &= \frac{2}{N} X^T (X^T w - T) + \lambda \vec{\nabla}_w \sum_{d=1}^D |w_d| \\ &= \text{---} + \lambda \sum_{d=1}^D \text{sign}(w_d) \end{aligned}$$

Here we defined a sub-gradient:

$$\nabla_w |w| = \begin{cases} +1 & \text{if } w_d > 0 \\ 0 & \text{if } w_d = 0 \\ -1 & \text{if } w_d < 0 \end{cases}$$



We may define $\text{sign}(\vec{w})$, but it won't factor with \vec{w} .

⇒ There is no simple mathematical form
(to frame it as an algebra problem)

• The GD update step is:

$$\begin{aligned} \vec{w} &\rightarrow \vec{w} - \eta \vec{\nabla}_w \mathcal{L} = \vec{w} - \eta \lambda \sum_{d=1}^D \text{sign}(w_d) \\ &\quad + (\text{usual MSE term}) \end{aligned}$$

$$= \vec{w} - \eta \lambda \vec{\text{sign}}(\vec{w}) + \vec{\nabla}(\text{MSE})$$

For a given dim^o, say, where $w_d > 0$, we have

$$w_d \mapsto w_d - \eta \lambda (+1) + \vec{\nabla}(\text{MSE})$$

shrinks w_d , independently of its magnitude

→ go code it!