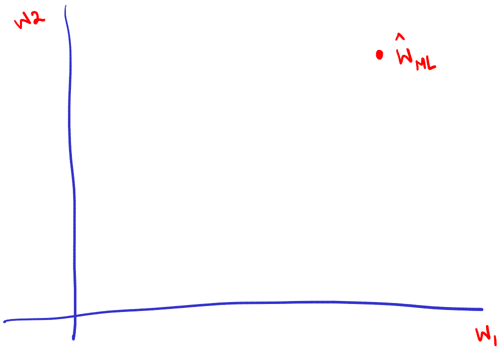
Linear regression / Ridge regression

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \underbrace{\lambda \|w\|^2}_{\text{REGULARIZATION}}$$

$$\lambda \|w\|^2 \rightarrow \sum_{i=1}^{d} w_i^2 = \|w\|^2$$

PARAMETER SPACE

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$



Where is $\hat{w}_R \rightarrow$ solution of the ridge regression problem?

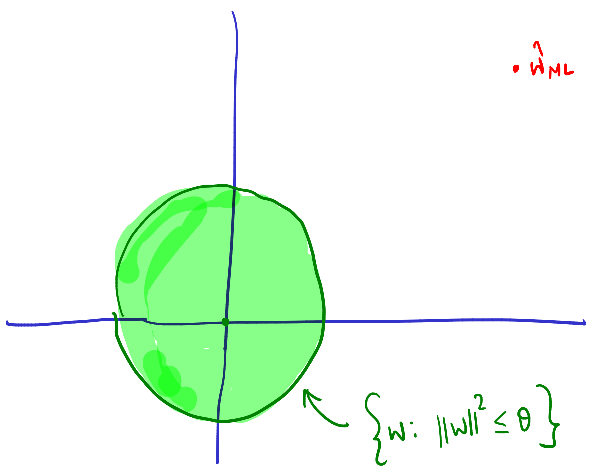
$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2 \quad \leftarrow \text{A}$$

$\equiv$ is equivalent to

$$\boxed{\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \sum_{i=1}^{n} (w^T x_i - y_i)^2 \\ \text{s.t.} \quad & \|w\|^2 \le \theta \end{aligned}} \quad \leftarrow \text{B}$$

$\theta \leftarrow$ depends on λ

For every choice of $\lambda > 0$, $\exists \theta$ s.t. the optimal solutions of problems A and B coincide.



$$\|w\|^2 \le \theta$$

$$\Rightarrow w_1^2 + w_2^2 \le \theta$$

- What is the loss / error / objective function value of linear regression at $\hat{w}_{ML}$

$$\sum_{i=1}^{n} \left(\hat{w}_{ML}^T x_i - y_i\right)^2 = f(\hat{w}_{ML})$$

Consider the set of all w s.t

$$f(w) = f(\hat{w}_{ML}) + c \qquad \underline{\underline{c \geq 0}}$$

$$S_c = \left\{ w : \quad f(w) = f(\hat{w}_{ML}) + c \right\}$$

i.e., every $\underline{w \in S_c}$ satisfies

$$\underbrace{\| X^T w - y \|^2}_{f(w)} = \underbrace{\| X^T \hat{w}_{ML} - y \|^2}_{f(\hat{w}_{ML})} + c$$

On simplification [Please do this],

one gets

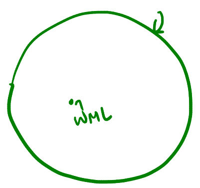
$$\underbrace{(w - \hat{w}_{ML})^T (XX^T)(w - \hat{w}_{ML})} = c'$$

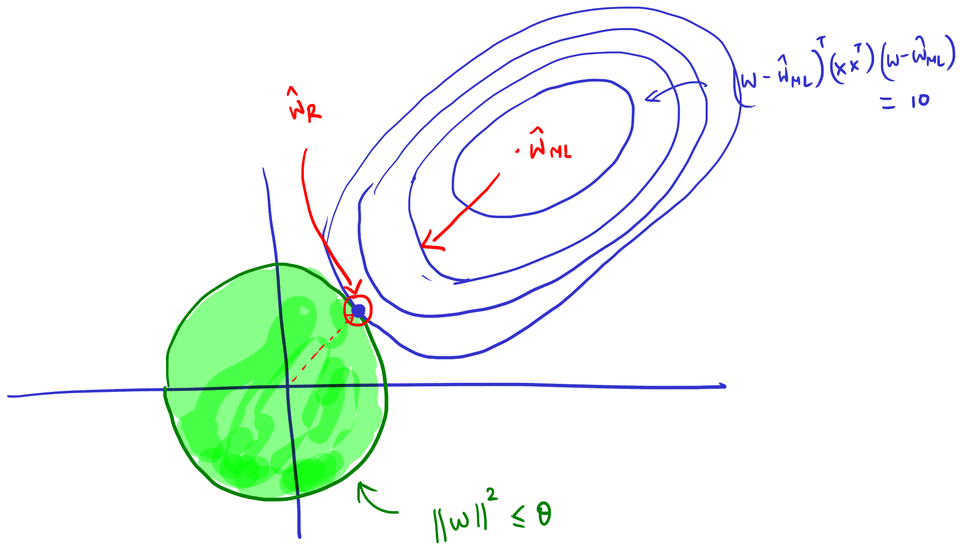
Some constant that depends on c, (XX^T), $\hat{w}_{ML}$ and not on w

If $\boxed{XX^T = I}$,

$$(w - \hat{w}_{ML})^T I (w - \hat{w}_{ML}) = c'$$

$$\| w - \hat{w}_{ML} \|^2 = c'$$





Conclusions:

- Ridge regression pushes feature values towards 0. But does not necessarily make it 0.

- An alternate way to regularize would then be using $\|\cdot\|_1$ norm instead of $\|\cdot\|_2^2$ norm

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|$$

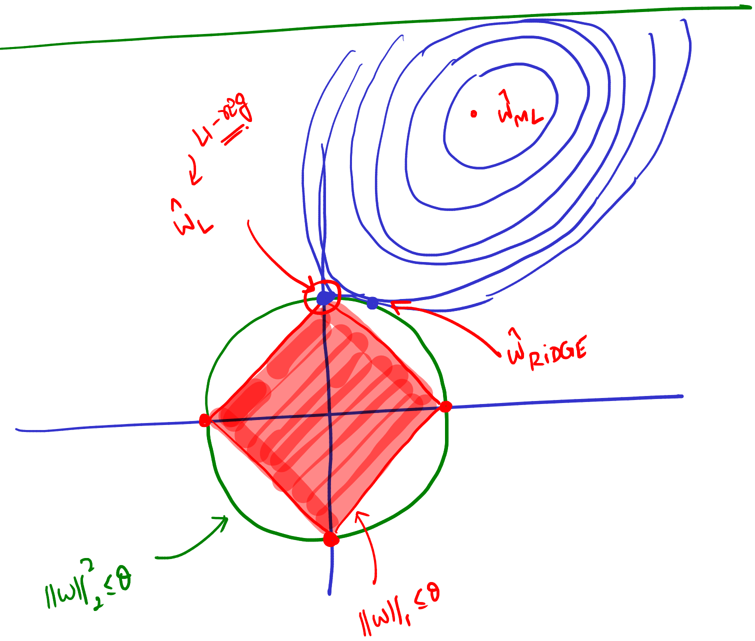
L1 Regularization

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1$$

$$\equiv$$

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\text{s.t. } \|w\|_1 \le \theta$$

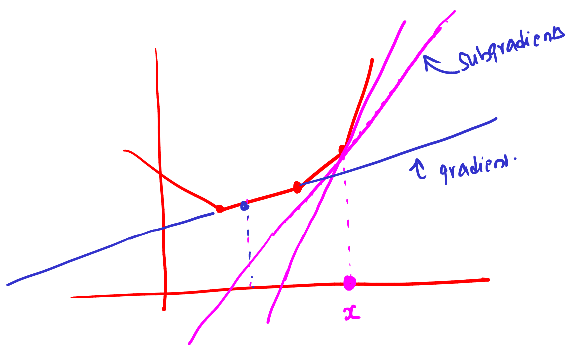


L1 Reg: LASSO - LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR.

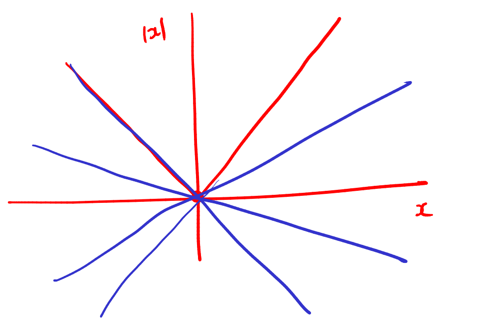
Points

- LASSO does not have a closed form solution

- Sub-gradient methods are usually used to solve LASSO.



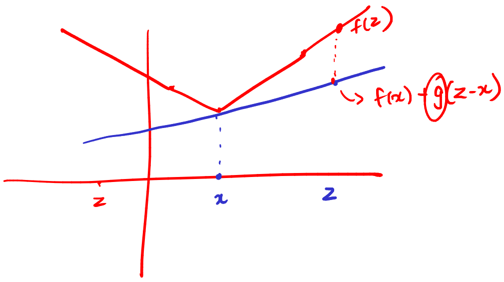
Sub-grad at 0 $= [-1, 1]$



Subgradient

A vector $g \in \mathbb{R}^d$ is a sub-gradient of $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ if

$$\forall z \qquad f(z) \geq f(x) + g^T(z-x)$$



Why Subgradients?

- If function f to minimize is a **convex** function, then sub-gradient descent converges!

- There are other special purpose methods for LASSO → (Ex) IRLS [Iterative Reweighted Least Squares].