Sparse classifers / models give insight

(x_i, d_i), $i = 1 \cdots N$ features, labels $x_i^T w \approx d_i$

$$Aw = [a_1 \; a_2 \; \cdots \; a_M] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} = \sum_{i=1}^{M} w_i a_i$$

a_ℓ: ℓ^{th} feature component

Suppose $w_\ell \approx 0 \Rightarrow a_\ell$ is unimportant

If a small number of w_ℓ are nonzero, then only these few features matter! w is sparse.

$\|w_0\| = \sum_{i=1}^{M} \mathbb{1}_{\{w_i \neq 0\}}$ (number of non-zero elements)

ℓ_0-norm

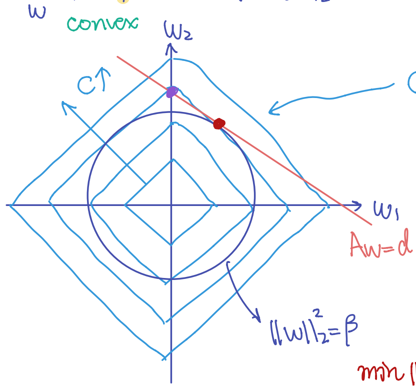
$\|aw\|_0 \neq a\|w\|_0$

Consider $\min_w \|w\|_0$ s.t. $\|Aw - d\|_2^2 < \varepsilon$ non-convex, intractable

Convex relaxation gives tractable problem

$\min_w \|w\|_1$ s.t. $\|Aw - d\|_2^2 < \varepsilon$ LASSO: Least Absolute Selection and Shrinkage Operator

convex



$C = \|w\|_1 = \sum_{i=1}^{M} |w_i| : |w_1| + |w_2| = C$, 1^{st} quad: $w_1 + w_2 = C$

$\min \|w\|_1$ s.t. $Aw = d$

corners on $\|w\|_1 \Rightarrow$ sparse solution

$\min \|w\|_2^2$ s.t. $Aw = d$ circular $\|w\|_2^2 \Rightarrow$ non sparse solutions

LASSO is a regularized least-squares problem

$\min_w \|w\|_1$ s.t. $\|Aw - d\|_2^2 < \varepsilon$ is equivalent to $\min_w \|Aw - d\|_2^2 + \lambda \|w\|_1$ for some λ, ε

Note: $\min_w \|w\|_1 + \frac{1}{\lambda} \|Aw - d\|_2^2$

LASSO	Ridge Regression
$w_L = \arg\min_w \|Aw - d\|_2^2 + \lambda \|w\|_1$	$w_R = \arg\min_w \|Aw - d\|_2^2 + \lambda \|w\|_2^2$
Sparse w_L	non-sparse w_R
Can have small model error $w_{op} - w_L$	great prediction error $\|Aw_{op} - Aw_R\|_2^2$
iterative solution	closed form solution

LASSO maybe used for model / feature selection

$w_L = \arg\min_w \|Aw - d\|_2^2 + \lambda \|w\|_1$ $S_L = \{i : [w_L]_i \neq 0\}$ selected features

$$Aw_L = \sum_{i=1}^{M} a_i [w_L]_i = \sum_{i \in S_L} a_i [w_L]_i$$

Debiasing $A_L = \{a_i : i \in S_L\}$

$\hat{w}_L = \underset{w}{\operatorname{argmin}} \|A_L w - d\|_2^2 = (A_L^T A_L)^{-1} A_L^T d$ avoids shrinkage due to $\|w\|_1$