Regularization In Regression

(x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Linear Regression: $w_0 + w_1 x(1) + w_2 x(2) + \dots + w_d x(d)$ – prediction for a new x

(Univariate) Polynomial Fitting

$x \in \mathbb{R}$, $d = 1$, $x \mapsto \phi(x)$ → of dimension d'

$\mathbb{R}^d \mapsto \mathbb{R}^{d'}$

$$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \\ \vdots \\ x^M \end{bmatrix} = z \qquad M = d'$$

Polynomial fit at x

equivalent to

Linear Regression for $\phi(x) = z$: $w_0 + w_1 z(1) + w_2 z(2) + \dots + w_M z(M)$

$= w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$

$x \in \mathbb{R}^d$

$\mathbb{R}^d \to \mathbb{R}^{d'}$, $d' >> d$

$x \mapsto \phi(x)$ → do prediction using $\phi(x)$

Least Squares Regression

$$\min_w \|X^T w - y\|_2^2$$

Solution: $(XX^T) w^* = Xy$

$\Rightarrow w^* = (XX^T)^{-1} Xy$

Prediction on training data is

$X^T w^* = X^T (XX^T)^{-1} X y$

Ridge Regression

$$\boxed{\min_w \|X^T w - y\|_2^2 + \lambda \|w\|_2^2}$$

Solution: $(XX^T + \lambda I) w^* = Xy$

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

Prediction on training data is

$X^T w^* = X^T (XX^T + \lambda I)^{-1} X y$

$$X^T = U\Sigma V^T \quad \text{- SVD of } X^T$$

$$XX^T = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$$

$$(XX^T)^{-1} = V\Sigma^{-2}V^T$$

$$X^T w^* = X^T (XX^T)^{-1} X y$$

$$= (U\Sigma \underbrace{V^T)(V}_{I}\Sigma^{-2}\underbrace{V^T)(V}_{I}\Sigma U^T) y$$

$$= U\Sigma\Sigma^{-2}\Sigma U^T y$$

$$= UU^T y$$

$$\boxed{\sum_{i=1}^{d+1} u_i u_i^T y}$$

Orthogonal projection onto range space of X^T

$$X^T = U\Sigma V^T$$

$$XX^T = V\Sigma^2 V^T$$

$$XX^T + \lambda I = V(\Sigma^2 + \lambda I)V^T$$

$$X^T w^* = X^T (XX^T + \lambda I)^{-1} X y$$

$$(XX^T + \lambda I)^{-1} = \left(V(\Sigma^2 + \lambda I)V^T\right)^{-1}$$

$$= V(\Sigma^2 + \lambda I)^{-1} V^T$$

$$X^T w^* = U\Sigma \underbrace{V^T V}_{I} (\Sigma^2 + \lambda I)^{-1} \underbrace{V^T V}_{I} \Sigma U^T y$$

$$\boxed{X^T w^* = U \underbrace{\Sigma(\Sigma^2 + \lambda I)^{-1}\Sigma}_{} U^T y}$$

$$\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & 0 \\ & \frac{\sigma_2^2}{\sigma_2^2+\lambda} & & \\ 0 & & \ddots & \\ & & & \frac{\sigma_{d+1}^2}{\sigma_{d+1}^2+\lambda} \end{bmatrix}$$

$$= \sum_{i=1}^{d+1} \underbrace{\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)}_{=1,\ \lambda=0} u_i u_i^T y$$

$$\le 1$$

$$\sigma_i^2 \gg \lambda, \quad \frac{\sigma_i^2}{\sigma_i^2+\lambda} \simeq \frac{\sigma_i^2}{\sigma_i^2} = 1$$

$$\sigma_i^2 \ll \lambda, \quad \frac{\sigma_i^2}{\sigma_i^2+\lambda} \simeq \frac{\sigma_i^2}{\lambda} \simeq 0$$

Ridge Regression "shrinks" the small singular values to 0.

Hard "shrinkage" : σ_i^2 as is for $i = 1, 2, \ldots, k$

$\sigma_i^2 \to 0$ for $i = k+1, \ldots, d+1$

k-truncated SVD

Ridge Regression : $\boxed{\min_w \|X^T w - y\|_2^2 + \lambda \|w\|_2^2}$

Equivalent to the following constrained optimization problem

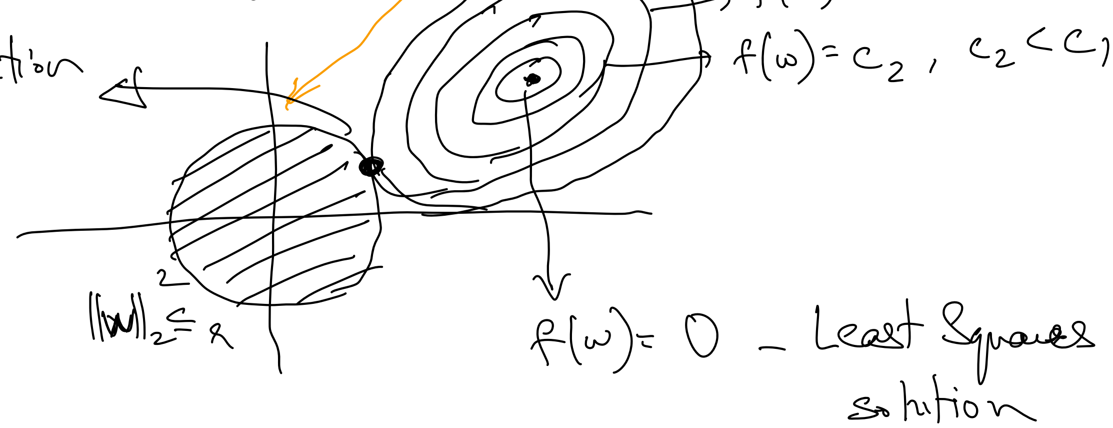
$$\boxed{\min_w \|X^T w - y\|_2^2 \text{ such that } \|w\|_2^2 \le r \text{ for some } r}$$

Geometric Interpretation of constrained optimization problem

Level sets of $f(w) = \|X^T w - y\|_2^2$

$f(w) = c$

Ridge Regularization Solution



$\boxed{\text{Lasso}}$: $\underbrace{\boxed{\min_w \|X^T w - y\|_2^2 + \lambda \|w\|_1}}_{\text{leads to a solution}}$ → 1-norm of w instead of squared 2-norm

with sparse w, i.e., w with many zeros.

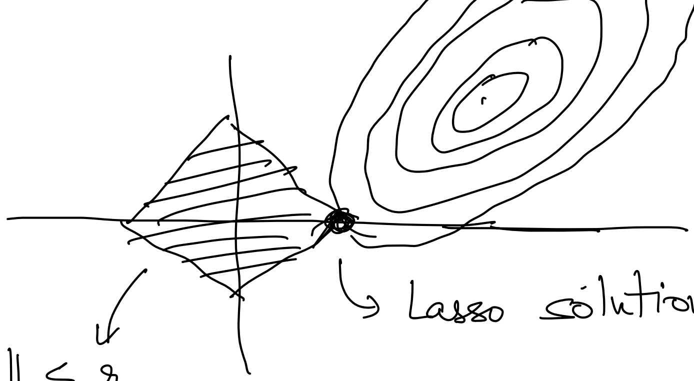
Parsimonious model

Lasso is equivalent to the following constrained optimization problem :

$$\min_w \|X^T w - y\|^2$$
$$\text{such that } \|w\|_1 \le r$$

Level sets of $f(w)$

Contour map



$\|w\|_1 \le r$

Lasso solution is usually sparse