Setup

fixed-design matrix $\in \mathbb{R}^{n \times d}$

$n \ll d$

$\in \mathbb{R}^d$ "true" parameter

Model $\quad y = Xw^* + \xi$

$\xi \in \mathbb{R}^n \quad N(0, \sigma^2 I_{n\times n})$

(A1) For each $i = 1, \dots, d$, $\quad \frac{1}{\sqrt{n}} \|X_i\|_2 \leq 1$

(A2) w^* is k-sparse.

<u>Performance metric</u> for $\hat{w} = \hat{w}(X, y)$

$$\mathcal{E}(\hat{w}) = \mathbb{E}_\xi \left[\frac{1}{n} \| X\hat{w} - Xw^* \|_2^2 \right]$$

<u>Benchmark</u>

$$\hat{w}_{\ell_0} \in \operatorname*{argmin}_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \frac{1}{2n} \|Xw - y\|_2^2$$

$\binom{d}{k}$

$$\mathcal{E}(\hat{w}_{\ell_0}) \lesssim \frac{k \cdot \sigma^2 \log\left(\frac{d}{k}\right)}{n}$$

minimax-optimal

$$\inf_{\hat{w}} \sup_{\substack{w^* \\ \|w^*\|_0 = k}} \mathbb{E}\left[\frac{1}{n} \|X\hat{w} - Xw^*\|_2^2 \right] \asymp \frac{k\sigma^2 \log\left(\frac{d}{k}\right)}{n}$$

Lasso

$$\hat{w}_\lambda \boxed{\in} \operatorname*{argmin}_{w \in \mathbb{R}^d} \Big\{ \overbrace{\frac{1}{2n}\|Xw-y\|_2^2 + \lambda\|w\|_1}^{f(w)} \Big\}$$

Basic properties

Fact 1 All solutions to the above provide the same predictions.

Proof By contradiction.

Assume that $\hat{w}_1, \hat{w}_2$ are two predictions such that

$$X\hat{w}_1 \neq X\hat{w}_2.$$

$$f(\hat{w}_1) = f(\hat{w}_2)$$

$u \in (0,1)$.

Consider $\hat{w}_u = u\,\hat{w}_1 + (1-u)\,\hat{w}_2$

$$f(\hat{w}_1) \leq f(\hat{w}_u)$$

(by strict convexity of $\alpha \mapsto \|\alpha - y\|_2^2$)

$$< u\, f(\hat{w}_1) + (1-u)\, f(\hat{w}_2)$$

$$= u\, f(\hat{w}_1) + (1-u)\, f(\hat{w}_1)$$

$$= f(\hat{w}_1)$$

$f(\hat{w}_1) < f(\hat{w}_1)$ which is impossible. ∎

$$\frac{1}{2n}\|Xw-y\|_2^2 + \lambda\|w\|_1$$

$\hat{w}$ is a solution to the above

if $\boxed{\nabla_w \frac{1}{2n}\|Xw-y\|_2^2 + \lambda s = 0}$

$s \in \partial\|\hat{w}\|_1$

$$s_i = \begin{cases} +1 & \text{if } \hat{w}_i > 0 \\ -1 & \text{if } \hat{w}_i < 0 \\ \in[-1,1] & \text{if } \hat{w}_i = 0 \end{cases}$$

$$\boxed{\frac{1}{n}X^T(y-X\hat{w}) = \lambda s}$$ ← satisfies this.

What is the smallest $\lambda > 0$ s.t.

$\hat{w} = 0$ is a solution?

Suppose: $\boxed{\lambda \geq \frac{\|X^Ty\|_\infty}{n}}$

Take $\hat{w} = 0$, $s = \frac{X^Ty}{n} \cdot \frac{1}{\lambda}$

$\|s\|_\infty \leq 1$

What λ should we choose?

Suppose $\frac{X^TX}{n} = I$ (hypothetical)

$$\frac{1}{n}X^T(y-X\hat{w}) = \lambda s$$

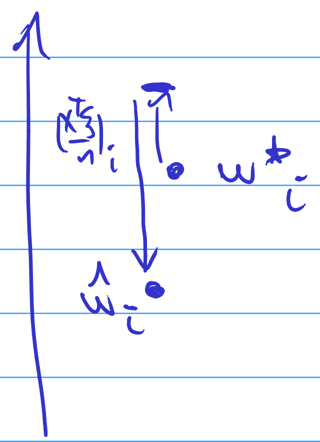
$$\frac{1}{n}X^T(Xw^* + \xi - X\hat{w}) = \lambda s$$

$$\boxed{w^* - \hat{w} + \frac{X^T\xi}{n} = \lambda s}$$

Suppose $\lambda = \left\|\frac{X^T\xi}{n}\right\|_\infty$

if $w^*_i = 0$, then $\hat{w}_i = 0$

$$\hat{w}_i = w^*_i + \left(\frac{X^T\xi}{n}\right)_i - \left\|\frac{X^T\xi}{n}\right\|_\infty s_i$$



If $|w^*_i| > 0$, then

$$|\hat{w}_i - w^*_i| \lesssim \left\|\frac{X^T\xi}{n}\right\|_\infty$$

$$\|\hat{w} - w^*\|_2^2 = \frac{1}{n}\|X\hat{w} - Xw^*\|_2^2$$

$$\leq k \cdot \left\|\frac{X^T\xi}{n}\right\|_\infty^2$$

$$\left\|\frac{X^T\xi}{\sqrt{n}}\right\|_\infty \lesssim \sigma\sqrt{\log d}$$

$$\left(\frac{1}{\sqrt{n}}\begin{bmatrix}\langle x_1, \xi\rangle/\sqrt{n} \\ \langle x_2, \xi\rangle/\sqrt{n} \\ \vdots \\ \langle x_d, \xi\rangle/\sqrt{n}\end{bmatrix}\right)^2$$

$$\leq k\sigma^2 \frac{\log d}{n}$$

$$\lambda \asymp \left\|\frac{X^T\xi}{n}\right\|_\infty$$

What can we prove for the lasso?

- Example. $\lambda \geq 2\|\frac{X^T\xi}{n}\|_\infty$

Let $\hat{w}$ be a lasso solution.

$$\underbrace{\frac{1}{2n}\|X\hat{w}-y\|_2^2+\lambda\|\hat{w}\|_1 \leq \frac{1}{2n}\|Xw^*-y\|+\lambda\|w^*\|_1}_{\text{Basic Inequality (Rearranging)}}$$

$$0 \leq \frac{1}{n}\|X\hat{w}-Xw^*\|_2^2 \leq 2\langle \frac{X^t\xi}{n}, \hat{w}-w^*\rangle + 2\lambda(\|w^*\|_1-\|\hat{w}\|_1)$$

$$\leq \underbrace{2\|\frac{X^T\xi}{n}\|_\infty}_{\leq \lambda} \overset{\leq 4\|w^*\|_1}{\underbrace{\|\hat{w}-w^*\|_1}_{\leq \|\hat{w}\|_1+\|w^*\|_1}} + 2\lambda(\|w^*\|_1-\|\hat{w}\|_1)$$

$$\leq 3\lambda(\|w^*\|_1) - \lambda\|\hat{w}\|_1.$$

$$\Rightarrow \|\hat{w}\|_1 \leq 3\|w^*\|_1$$

$$\frac{1}{n}\|X\hat{w}-Xw^*\| \leq \lambda \cdot 4\|w^*\|_1 + 2\lambda\|w^*\|_1$$

$$\leq 6\lambda\|w^*\|_1$$

$$\lambda \simeq 2\|\frac{X^t\xi}{n}\|_\infty \simeq \frac{\sigma\sqrt{\log d}}{\sqrt{n}}$$

Zhang, Jordan, Wainwright 2017

$$f_p^\lambda(w) = \frac{1}{2n}\|Xw - y\|_2^2 + \lambda\rho(w)$$

↑ more general penalty.

ρ needs to satisfy:

1) $\rho(w) = \sum_{i=1}^{d} \rho_i(w_i)$ | $\|w\|_1 = \sum |w_i|$

2) $\rho(0) = 0$

3) $\rho(-w) = \rho(w)$

4) ρ non-decreasing on $[0, \infty)$.

$$W_\rho^\lambda = \left\{ w \in \mathbb{R}^d : w \longmapsto \frac{1}{2n}\|Xw - y\|_2^2 + \lambda\rho(w) \text{ is a local min of the above} \right\}$$

<u>Thm</u>

There exists a "bad" design matrix X such that for any ρ there exists a 2-sparse vector w^* such that

$$\mathbb{E}_\xi \inf_{\lambda \geq 0} \sup_{w \in W_\rho^\lambda} \frac{1}{n}\|Xw - Xw^*\|_2^2 \geq \frac{\sigma\sqrt{\log n}}{\sqrt{n}}$$

Intuition on 2x2 matrices

Let $A = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}$ $\qquad w^* = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$

$\|A_1\|_2 = 1$

$\|A_2\|_2 = 1$

$$\boxed{A^T (y - A\hat{w}) = \lambda s}$$

$$A^T (A w^* + \xi - A\hat{w}) = \lambda s$$

$\hat{w} = 0$

$$\big(\underbrace{A^T A w^*}_{\text{signal}} + \underbrace{A^T \xi}_{\text{noise}}\big) = \lambda s$$

$$A^T A = \begin{bmatrix} 1 & 2\varepsilon - 1 \\ 2\varepsilon - 1 & 1 \end{bmatrix} \qquad A^T A w = 0$$

$$\|A \cdot 0 - A w^*\|_2^2 = \|A w^*\| = w^{*T} \underbrace{A^T A w^*}_{0} = 0$$

$$A^T A w^* = \begin{bmatrix} \varepsilon \\ \varepsilon \end{bmatrix}$$

If $\hat{w} = 0$, then $\|A\hat{w} - A w^*\|_2^2 = \varepsilon$

choose $\varepsilon = \Omega\left(\frac{1}{\sqrt{n}}\right)$

$$A^T A = \begin{bmatrix} 1 & -1+2\varepsilon \\ -1+2\varepsilon & 1 \end{bmatrix} \qquad w^* = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$X = \Bigg[\underbrace{\begin{matrix} \boxed{\sqrt{n} A} & & & \\ & \boxed{\sqrt{n} A} & & \\ & & \boxed{\sqrt{n} A} \cdots & \\ & & & \boxed{\sqrt{n} A} \end{matrix}}_{n \times n} \quad \underset{n\times(d-n)}{0} \Bigg] \qquad w^* = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \hat{w}_{2i-1} \\ \hat{w}_{2i} \end{bmatrix} \in \underset{w \in \mathbb{R}^2}{\operatorname{argmin}} \frac{1}{2n} \left\| \sqrt{n} A - \begin{bmatrix} y_{2i-1} \\ y_{2i} \end{bmatrix} \right\|_2^2 + \lambda |\hat{w}_{2i-1}| + \lambda |\hat{w}_{2i}|$$

Everything is in 2 dimensions and in the first block.

optimality conditions for $\hat{w}$

$$\frac{1}{n} \sqrt{n} A^T (y + \cancel{\sqrt{n} A \hat{w}}) = \lambda s$$

$\hat{w} = 0$

$\hookrightarrow \sqrt{n} A w^* + \xi$

$$\to A^T A w^* + \frac{A^T \xi}{\sqrt{n}}$$

$$= \begin{bmatrix} \varepsilon \\ \varepsilon \end{bmatrix} + \boxed{\frac{A^T \xi}{\sqrt{n}}} = \lambda s$$

$$\frac{1}{\sqrt{n}} \left\| \begin{bmatrix} \langle A_1, \xi \rangle \\ \langle A_2, \xi \rangle \end{bmatrix} \right\|_\infty \le \frac{1}{\sqrt{n}} \cdot \sqrt{\xi_1^2 + \xi_2^2}$$

$\|A_1\|_2 = 1 \qquad \|A_2\|_2 = 1$

$$A^TA = \begin{bmatrix} 1 & 1-2\varepsilon \\ 1-2\varepsilon & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} \cos(\alpha) & -\cos(\alpha) \\ \sin(\alpha) & \sin(\alpha) \end{bmatrix}$$

$$\|A_1\|_2 = \sqrt{\cos^2(\alpha) + \sin^2(\alpha)} = 1$$

$$A^TA = \begin{bmatrix} 1 & -\cos^2(\alpha) + \sin^2(\alpha) \\ & 1 \end{bmatrix}$$

$$-1 + \underbrace{2\sin^2(\alpha)}_{=\varepsilon}$$

$$\begin{bmatrix} \varepsilon \\ \varepsilon \end{bmatrix} + \begin{bmatrix} \langle A_1, \xi\rangle / \sqrt{n} \\ \langle A_2, \xi\rangle / \sqrt{n} \end{bmatrix} = \lambda s$$

$$\|\cdot\|_\infty \le N \frac{\sqrt{\xi_1^2 + \xi_2^2}}{\sqrt{n}} \ll \frac{\sqrt{\log n}}{\sqrt{n}}$$

$\sigma = 1$

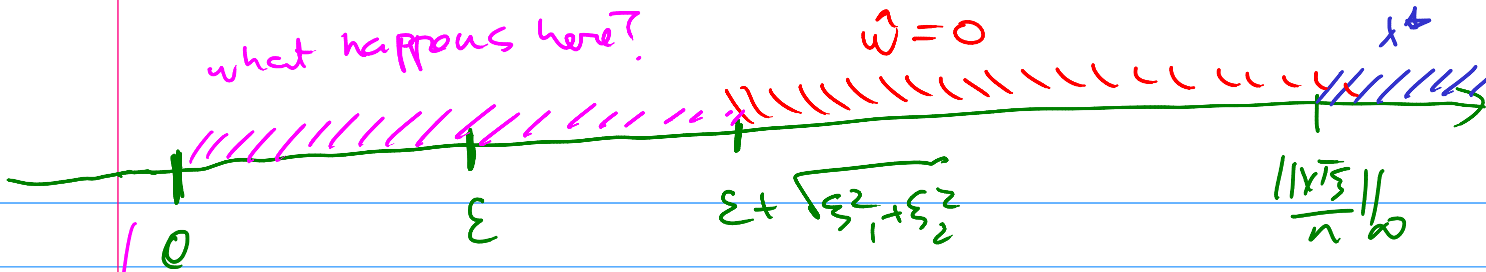
for the ε

$$\lambda \asymp \left\|\frac{X^T\xi}{n}\right\|_\infty = \frac{\sqrt{\log n}}{\sqrt{n}}$$

choose ε

$$c_1 \cdot \left\|\frac{X^T\xi}{n}\right\|_\infty$$

some very small abs const.



$\varepsilon \quad \frac{1}{n}\|\frac{X^T\xi}{n}\| \gtrsim \frac{\sqrt{\log n}}{\sqrt{n}}$ slow rate !

Intuition

Take some block such that

$$\frac{\langle A_1, \xi \rangle}{\sqrt{n}} \gtrsim \frac{\sqrt{\log n}}{\sqrt{n}} \qquad \boxed{n^{3/4}}$$

Take a "bad" block above holds.

Let $\hat{w} \in \mathbb{R}^2$ be some minimizer of the lasso objective on this block.

$$\tilde{w} = \begin{bmatrix} \langle A_1, \xi \rangle / \sqrt{n} \\ 0 \end{bmatrix}$$

want this to be large

$$E_{rr} \quad \|A\hat{w} - Aw^*\|_2^2 = \|A\hat{w}\|_2^2$$

$$\|A\hat{w} - \xi/\sqrt{n}\|_2^2$$

$$\leq \|A\tilde{w} - \xi/\sqrt{n}\|_2^2 + \frac{\langle A_1, \xi \rangle}{\sqrt{n}} \cdot \lambda$$

$$= \left\| A_1^T A_1 \frac{\xi}{\sqrt{n}} - \frac{\xi}{\sqrt{n}} \right\|_2^2 + \frac{\langle A_1, \xi \rangle}{\sqrt{n}} \cdot \lambda$$

$$= \frac{\|\xi\|_2^2}{n} - \frac{\langle A_1, \xi \rangle^2}{\sqrt{n}} + \frac{\langle A_1, \xi \rangle}{\sqrt{n}} \cdot \lambda$$

$$\|A\hat{w} - \xi/\sqrt{n}\|_2^2 \geq \frac{\|\xi\|_2^2}{n} - \|A\hat{w}\|_2^2$$

$$\Rightarrow \|A\hat{w}\|_2^2 \geq \left(\frac{\langle A_1, \xi \rangle}{\sqrt{n}}\right)^2 - \frac{\langle A_1, \xi \rangle}{\sqrt{n}} \lambda$$

$$\gtrsim \left\| \frac{X^T \xi}{n} \right\|_\infty^2 \gtrsim \frac{\log n}{n}$$

$\#$ bad blocks $\asymp n^{3/4}$.

Total Error from bad blocks

$$\gtrsim \frac{\log n}{n^{1/4}} \gtrsim \frac{\sqrt{\log n}}{\sqrt{n}}.$$

This concludes the proof for the whole regularization path.