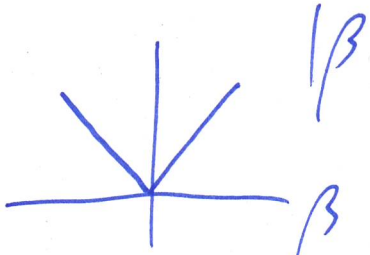
Lasso = soft-threshold estimator if $X^TX/n = I$

Proof:

$$\|Y - X\beta\|_2^2/n + \lambda\|\beta\|_1$$



$$= \|Y\|_2^2/n + \sum_{j=1}^{p} \underbrace{\{-2\underbrace{(X^TY)_j/n}_{z_j}\cdot\beta_j + \beta_j^2 + \lambda|\beta_j|\}}_{\text{minimize w.r.t. } \beta_j \;\; \forall j}$$

decoupled problem!

$$-2 z_j \beta_j + \beta_j^2 + \lambda |\beta_j| \overset{!}{=} \min_{\beta_j}$$

if $0 \leq z_j \leq \lambda/2 \Rightarrow \hat{\beta}_j = 0$

[because $\lambda |\beta_j| - 2 z_j \beta_j \geq 0 \quad \forall \beta_j$

$= 0 \quad \beta_j = 0$

$\beta_j^2 \geq 0 \quad \forall \beta_j \quad = 0 \quad \beta_j = 0$]

likewise: $-\lambda/2 \leq z_j \leq 0 \Rightarrow \hat{\beta}_j = 0$

for $z_j > \lambda/2 \rightsquigarrow \hat{\beta}_j > 0$

and hence objective fct. is differentiable for > 0

$$\frac{\partial}{\partial \beta_j} : -2z_j + 2\beta_j + \lambda \stackrel{!}{=} 0$$

$$\rightsquigarrow \hat{\beta}_j = z_j - \lambda/2$$

likewise for $z_j \leq -\lambda/2$. ∎

in general: sub-differential calculus

II.4. Prediction with the Lasso

goal: estimation of regression function

$$f(x) = \mathbb{E}[Y|X=x] = \sum_{j=1}^{p} \beta_j x_j = \beta^T x = x^T \beta$$

II.4.1 Practical aspects

use $\hat{f}(x) = \hat{\beta}(\lambda)^T x$

and choose λ via cross-validation

Why prediction?

training data $(X_1, Y_1), \ldots (X_n, Y_n)$

new data point: X_{new}, Y_{new}

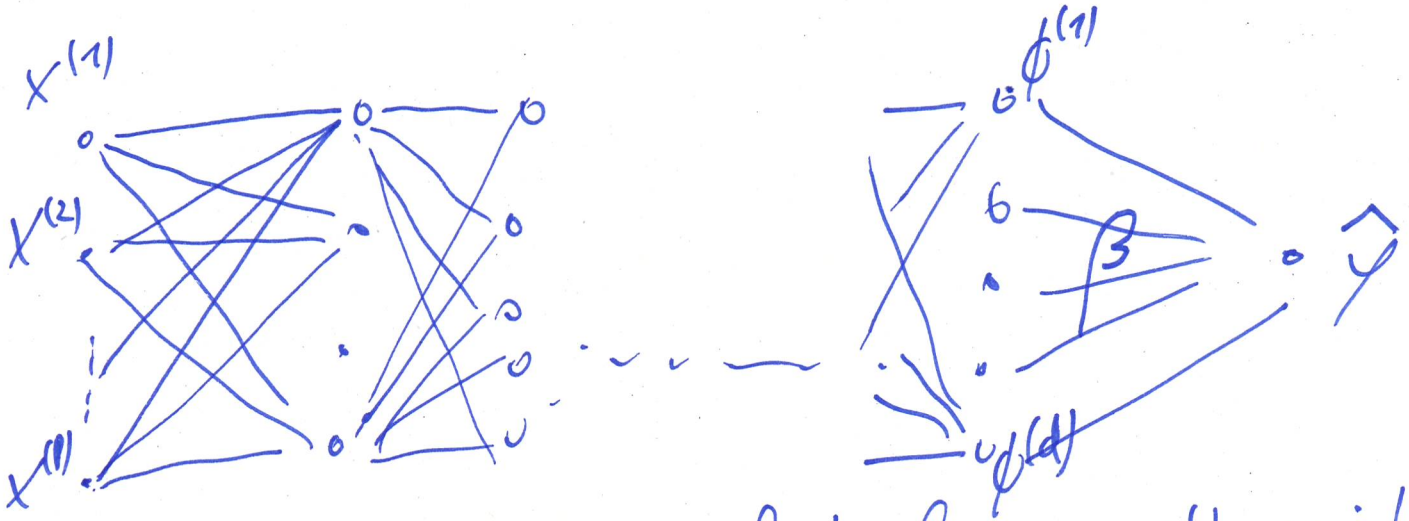
X_{new}: observe — Y_{new}: to be predicted

$$\mathbb{E}\left[(Y_{new} - \hat{\beta}^T X_{new})^2\right] = \sigma^2_{new} + \underbrace{\mathbb{E}\left[(\hat{\beta}^T X_{new} - \beta^T X_{new})^2\right]}$$

$Y_{new} = \beta^T X_{new} + \varepsilon_{new}$, ε_{new} indep. of training data, $\mathbb{E}[\varepsilon^2_{new}]$

mean squared error for estimating regression function

Comment: deep neural network



last layer with variables/features

$\phi^{(1)}(x)$

$\phi^{(2)}(x)$ $\quad d \gg n$

$\vdots$

$\phi^{(d)}(x)$

learned from data

last step:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\|y - \underline{\Phi}\beta\|_2^2/n + \lambda \|\beta\|_1 \right)$$

or $\lambda \|\beta\|_2^2$

$$\underset{n \times d}{\underline{\Phi}} = \begin{bmatrix} \phi^{(1)}(x_1) & \cdots & \phi^{(d)}(x_1) \\ \phi^{(1)}(x_2) & & \vdots \\ \vdots & & \\ \phi^{(1)}(x_n) & \cdots & \phi^{(d)}(x_n) \end{bmatrix} \qquad d \gg n$$

Choice of λ: cross-validation $\hat{\lambda}_{CV}$

(typically 10-fold CV; default in glmnet R-package)

good prediction and increased sparsity: "1se rule" from glmnet

largest λ which is at most 1se away from minimum

II.4.2 Some results from asymptotic theory

goal: what are "some" properties of the Lasso in terms of mathematical guarantees

Consider asymptotics:

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta^0_{n;j} X^{(j)}_{n;i} + \varepsilon_{n;i} \qquad i = 1, \dots, n$$

$$p = p_n$$

$$n = 1, 2, 3, \dots$$

allow for $\underline{p_n \gg n}$ $\left(\frac{p_n}{n} \to \infty \ (n \to \infty) \right)$

What happens (to the Lasso) as $n \to \infty$ (and $p_n \to \infty$)?

notationally: just write $\beta^0_{p_n \times 1}$

etc...

Consider (here) fixed design X

Theorem

Consider the Lasso $\hat{\beta}(\lambda)$.

Assumptions:

(1) $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$

(not crucial, but simplifies the proof)

(2) $\frac{1}{n} \sum_{i=1}^{n} \left(X_i^{(j)}\right)^2 \equiv 1 \quad \forall j = 1, \dots, p_n$

Scaled covariates

(3) $\|\beta^0\|_1 = o\left(\sqrt{\frac{n}{\log(p_n)}}\right) \quad (n \to \infty)$

(makes only sense if $\log(p_n) \ll n$) e.g. $p_n = n^\alpha$ works for any $\alpha < \infty$

choose $\lambda = \lambda_n = 4\,\hat{\sigma}_n \sqrt{\frac{t_n^2 + 2\log(p_n)}{n}}$ with

- $t_n^2 \to \infty$, $t_n^2 = o(\log(p_n))$ $\quad (t_n^2 = \log(p_n))$
- $\mathbb{P}[C > \hat{\sigma}_n \geq \sigma] \longrightarrow 1$

  ($\hat{\sigma}_n$ should overestimate σ)

Then: $\|X(\hat{\beta}(\lambda_n) - \beta^0)\|_2^2 / n \xrightarrow{P} 0 \quad (n \to \infty)$

convergence in probability

main ass: $\|\beta^0\|_1$ is suffic. sparse

$\lambda = \lambda_n \asymp \sqrt{\log(p_n)/n}$ $\qquad$ $\log(p_n) \ll n$