Definitions :

The setting of the problem is the following
we have data of the form
$\{(X_i, Y_i)\,;\, i=1,\dots,n\}$, these are observations of

$$(X, Y) \sim \mu$$

feature living in $\mathbb{X}$ (covariates)

response living in $\mathbb{Y}$, or label if Y is discrete e.g. $\{0,1\}$

GOAL : "learn from a data how to predict y from x."

In some ways the best predictor of y given x is the conditional expectation of

$$\mathbb{E}[Y \mid X = x].$$

e.g. for it we want a function f that minimizes

$$\mathbb{E}\left[(Y-f(X))^2\right]$$

then $\mathbb{E}\left[(Y-f(X))^2\right]$

$$= \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f(X)\right)^2\right]$$

$$= \mathbb{E}\left[(Y-\mathbb{E}[Y|X])^2\right] + \mathbb{E}\left\{(\mathbb{E}[Y|X]-f(X))^2\right\}$$

$$+ 2\,\mathbb{E}\left\{(Y-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-f(X))\right\}$$

But $\mathbb{E}\left\{(Y-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-f(X))\right\}$

$$= \mathbb{E}\left\{\mathbb{E}\left[(Y-\mathbb{E}(Y|X)) \times (\mathbb{E}[Y|X]-f(X)) \,\middle|\, X\right]\right\}$$

Tower Law

$$= \mathbb{E}\left\{(\mathbb{E}[Y|X]-f(X)) \times \mathbb{E}\left\{Y - \mathbb{E}[Y|X] \,\middle|\, X\right\}\right\}$$

(Since $\mathbb{E}[Y|X], f(X)$ are m'ble wrt $\sigma(X)$)

$$= 0$$

Since $\mathbb{E}\left[Y - \mathbb{E}[Y|X] \,\middle|\, X\right] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X]$

Thus

$$\mathbb{E}\left[(Y-f(X))^2\right]$$

$$= \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}[Y|X] - f(X)\right)^2\right]$$

$$\geq \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2\right].$$

The function

$\mathbb{E}[Y|X=x]$ is called the

Regression function of Y on X

<u>In general</u>, for any loss/risk

we'd be done if we knew the

Law of $Y | X=x$ for any x.

<u>But</u> we only have access

to $\underbrace{(X_1, Y_1), \ldots, (X_n, Y_n)}_{\text{data}} =: S_n$

So our goal is to use S_n to

construct a function $\hat{f}_n$ that approximates

$$f(x) = \mathbb{E}[Y \mid X = x] \ , \text{ say in } L_2$$

that is we want

$$\hat{f}_n : \mathbb{X} \longrightarrow \mathbb{Y}$$

s. that

$$R(f_n) := \mathbb{E}\left[(Y - \hat{f}_n(X))^2\right]$$
$$= \mathbb{E}\left[(Y - \hat{f}_{S_n}(X))^2\right]$$
$$= \int (y - \hat{f}_{S_n}(x))^2 \mu(dx, dy)$$

is minimized (in some sense wrt S_n)

$\boxed{\text{Note 1:}}$ In $\int \ell(y, \hat{f}_{S_n}(x)) \mu(dx, dy)$

the integral is wrt the distribution of (X, Y) whereas the dependence on the sample S_n is still there.

$\boxed{\text{i.e}}$ if $\hat{f}_{S_n}$ depends non-trivially on S_n, the quantity

$$R(\hat{f}_n) = \|Y - \hat{f}_n\|_2^2$$ is RANDOM

Note 2: We proved earlier that

$$R(\hat{f}_n) = \mathbb{E}\left[(Y - \hat{f}_n)^2\right]$$

$$= \mathbb{E}\left[(Y - f(X))^2\right] + \mathbb{E}\left[(f - \hat{f}_n)^2(X)\right]$$

$$f := \mathbb{E}[Y|X]$$

That is, for any "estimator" $\hat{f}_n$, the risk can be decomposed as

$$R(\hat{f}_n) = \underbrace{R(f)}_{\text{approximation error}} + \underbrace{\|f - \hat{f}_n\|_2^2}_{\text{estimation error}}$$

Approximation error is related to the richness of the function class we're allowed our estimator to come from.

$\boxed{\text{e.g}}$ $f(x) = \mathbb{E}[Y \mid X = x]$ is the optimal

when we solve

$$\inf_f \{ \|Y - f\| : Y \text{ is } X\text{-m'ble } \& L^2 \}$$

& the optimal in this class gives $R(f)$.

This error is not related to the statistical procedure so we will ignore it & focus on the

"estimation error"

which answers the question

Q: how efficiently can we learn f (the optimal) from observations S_n?

$\boxed{\text{Note 3:}}$ bounds on the random g $\|\hat{f}_n - f\|_2^2$ will be given in expectation or with high prob.

$\boxed{\text{e.g}}$ if Y is a label then one may want $\mathbb{P}[Y \neq h(X)]$ i.e. the prob. of misclassification.

We have not so far discussed the question of how to construct the estimator $\hat{f}_n$ from the data.

One principle that is often in many cases is that of

EMPIRICAL RISK MINIMISATION — E R M

That is: the optimal solution of our problem is given as an optimizer of an expectation

$$f(x) = \mathbb{E}[Y|X=x] = \inf_{f \in L^2} \mathbb{E}\left[(Y-f(X))^2\right].$$

We don't have access to the full measure μ generating (X,Y) but we have samples from it, so we can approximate it

with the empirical measure

$$\hat{\mu}_n(dz) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(dz)$$

So we may try $\qquad L_f(x,y) = (y-f(x))^2$

$$\hat{f}_n = \arg\inf_{f\in\mathcal{F}} \mu_n(L_f) = \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

$$\arg\inf_{f\in\mathcal{F}} \mu(L_f) = f.$$

Note: In certain cases the function class considered (e.g. neural nets) may be rich enough to include g s.that $y_i = g(x_i)$ $\forall i=1,\dots,n$ for all reasonable sample sizes n. ← interpolate

In principle, this may lead to overfitting. (i.e) memorizing the data w/o extracting any information about μ.

[deep neural nets tend to interpolate w/o overfitting, but no-one has a clear picture why or how this happens.]

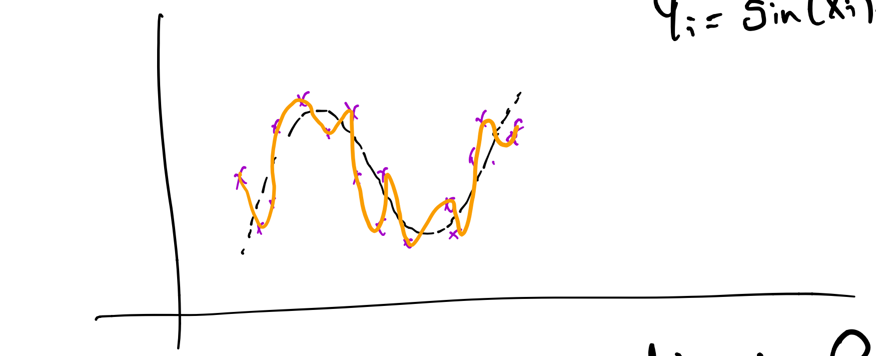
To avoid this we may add a regularizer, e.g. solve instead

$$\inf_g \{ R_n(g) + \text{penalty}(g) \}$$

where the penalty may penalize overly complicated answers.

<u>Why?</u> e.g. suppose

$y_i = \sin(x_i) + \varepsilon_i$



the orange curve achieves 0 ERM

<u>i.e</u> it interpolates $\hat{f}(x_i) = y_i$ for all $i = 1, \dots, n$

<u>BUT</u>: the real generating function is $\sin(\cdot)$

<u>So</u> potentially here, if we are thinking of using trigonometric functions

we could perhaps use a penalty of the form $\boxed{\sum_{k=-\infty}^{\infty} (1+k^2) |\hat{c}_k|^2}$, where

$\hat{c}_k$ is the coefficient of $\exp(2\pi i k x)$ to penalise "high energy" solutions.

In fact this penalty is equivalent to

H^1 (Sobolev) norm so it restricts the norm of the derivatives

i.e $\quad |\hat{f}|^2 + |\hat{f}'|^2 \leq \sum (1+n^2) |\hat{f}_n|^2$

$$\leq C_\epsilon \left(|f|^2 + |f'|^2 \right)$$

So it indeed acts as a regularizer.

BUT : regularization may also play a different role, i.e it may in some way be used to

Choose among many solutions

that would o.w. look identical in terms of our loss function, by imposing restrictions such as sparsity, etc.

This will be crucial in high-dimensional problems where typically there are many more parameters than observations

$$p \gg n$$

params $\quad$ sample size

HIGH-DIMENSIONAL REGRESSION & LASSO

Observation Model

$$y = X\theta^* + \varepsilon$$

$y \in \mathbb{R}^n$ observation vector

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$ design matrix

$\varepsilon \in \mathbb{R}^n$ noise vector.

n = # of observations d = # of degrees of freedom? dimension?

fixed design: X is deterministic

random design: X is random

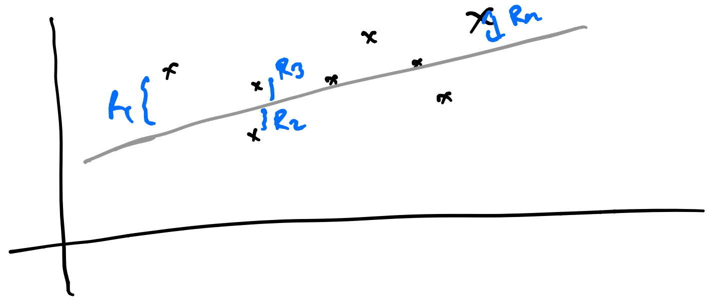
WARM-UP: let's consider first the classical setting i.e $d \ll n$.

The problem is often written as OLS

$$\hat{\theta}^{LS} := \arg\min_{\theta} \|y - X\theta\|_2^2$$

$$= \arg\min_{\theta} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{d} X_{ij}\theta_j\Big)^2$$

$$\sum_{i=1}^{n} R_i^2$$



The quadratic loss function we used is convex so it is not difficult to verify that the first order condition for optimality

$$X^T X \hat{\theta}^{LS} = X^T y$$

Assume for now that

"X has full rank"

i.e its columns are linearly independent.

Claim: then $X^T X$ is invertible

Pf of claim: Consider Xv for some $v \in \mathbb{R}^d$

then $X v = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}$

$$= \begin{pmatrix} x_{11} v_1 + x_{12} v_2 + \cdots + x_{1d} v_d \\ \vdots \\ x_{n1} v_1 + x_{n2} v_2 + \cdots + x_{nd} v_d \end{pmatrix}$$

$$= v_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + v_2 \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + v_d \begin{bmatrix} x_{1d} \\ \vdots \\ x_{nd} \end{bmatrix}$$

= linear combination of column vectors of X.

By assumption columns of X are lin. independent

so $Xv = 0 \iff v = 0_d$.

Thus $v^T X^T X v > 0$ if $v \neq 0 \Rightarrow$

$X^T X$ "strictly" positive definite.

$\Rightarrow$ $X^T X$ invertible $\blacksquare$

Thus if rank$(X) = d$

$$\hat{\theta}^{LS} = (X^T X)^{-1} X^T y.$$

But obviously rank$(X) \leq \min\{d, n\}$ which may only happen if $d < n$.

What if $d \geq n$:

we still require $X^T X \theta^{LS} = X^T y$

But now all we can observe is

$$\underset{(n \times d)}{X} \; \underset{(d \times 1)}{\theta^{LS}}$$

$(n \times 1)$ an n-dimensional vector.

That is now rank $(X) < d$

so null$(X) \neq \phi \Rightarrow$ if $\varphi \in$ null(X)

$$X \theta = X(\theta + \varphi).$$

So the system is underdetermined & we cannot hope to obtain a unique sol^{n}

$\Rightarrow$ instead we get a subspace of solutions

At this space we may consider regularizing the problem, that is perhaps we instead consider

$$\min_{\theta} \|\theta\|_2^2 \quad , \quad \text{s.th. } X^T X \theta = X^T y \qquad (\ast\ast)$$

find the minimum norm solution

It is not hard to see that the above always admits a unique solution, think of the subspace of solu's to $X^T X \theta = X^T y$ & its distance from the origin.

In fact the solution above is denoted by $(X^T X)^+$ & is known as the (it coincides with the usual inverse if invertible) Moore-Penrose Inverse. historical note? $(\ast\ast)$ as.

So we can write the solution to

$$\hat{\theta}^{LS} = (X^T X)^+ X^T y.$$

We would now like to see how well this procedure performs in terms of dimension.

First we need to decide how to measure performance here there is an important decision to be made

& it really depends on the purpose of the analysis, i.e.

vs out of sample

CASE 1: we are interested in prediction that is we want to estimate the value $X\theta^*$ as well as possible from the noisy observations $y = X\theta^* + \varepsilon$. (also known as in sample prediction).

CASE 2: we are interested in the actual value of θ^*.

In case (1) one may consider e.g. the mean squared error, i.e.

$$MSE(X\hat{\theta}^{LS}) = \frac{1}{n}\|X\theta^* - X\hat{\theta}^{LS}\|_2^2$$

$$= (\theta^* - \hat{\theta}^{LS})^T \frac{X^T X}{n} (\theta^* - \hat{\theta}^{LS})$$

whereas in the second case we may consider e.g.

$$\|\theta^* - \hat{\theta}^{LS}\|_2^2.$$

Let's consider prediction error for now & more specifically let's focus on the MSE.

Notice that since $\hat{\theta}^{LS}$ solves $\theta \in \arg\min \|y - X\theta\|_2^2$

min $\|\theta_2\|$ s.t. that

we have by definition that

$\hat{\theta}^{LS}$ minimizes the $\|\cdot\|_2^2$ norm among the θ's that minimize
$$\|y - X\theta\|_2^2.$$

So $$\|y - X\hat{\theta}^{LS}\|_2^2 \leq \|y - X\theta^*\|_2^2$$
& since by definition
$$y = X\theta^* + \varepsilon$$
$$\|y - X\hat{\theta}^{LS}\|_2^2 \leq \|\varepsilon\|_2^2. \quad (1)$$

Also $$\|y - X\hat{\theta}^{LS}\|_2^2$$
$$= \|y - X\hat{\theta}^{LS} + X\theta^* - X\theta^*\|_2^2$$
$$= \|y - X\theta^*\|_2^2 + \|X(\hat{\theta}^{LS} - \theta^*)\|_2^2 - 2\langle y - X\theta^*, X(\hat{\theta}^{LS} - \theta^*)\rangle$$
$$(2) \quad = \|\varepsilon\|_2^2 + \|X(\hat{\theta}^{LS} - \theta^*)\|_2^2 - 2\langle \varepsilon, X(\hat{\theta}^{LS} - \theta^*)\rangle$$

Rearranging (2) we get
$$\|X(\hat{\theta}^{LS} - \theta^*)\|_2^2 \leq \|y - X\hat{\theta}^{LS}\|_2^2 - \|\varepsilon\|_2^2 + 2\langle \varepsilon, X(\hat{\theta}^{LS} - \theta^*)\rangle$$
$$\overset{(1)}{\leq} 2\langle \varepsilon, X(\hat{\theta}^{LS} - \theta^*)\rangle.$$

<u>Now</u> although we've made some progress $\hat{\theta}^{LS}$ depends on ε, & the RHS depends on θ^*.

One way out is to consider a worst case scenario that is take supremum over everything but ε

i.e.

$$\|X(\hat{\theta}^{LS}-\theta^*)\|_2^{\cancel{2}} \leq 2\left\langle \varepsilon, \frac{X(\hat{\theta}^{LS}-\theta^*)}{\|X(\hat{\theta}^{LS}-\theta^*)\|_2}\right\rangle \cdot \cancel{\|X(\hat{\theta}^{LS}-\theta^*)\|_2}$$

$$\Leftarrow \quad \|X(\hat{\theta}^{LS}-\theta)\|_2 \leq 2 \sup\left\{\langle \varepsilon, v\rangle : v \in \mathbb{R}^n, \|v\|_2 \leq 1\right\}$$

$$= 2\|\varepsilon\|_2^2 .$$

Now if $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is iid

$$\mathbb{E}\|\varepsilon\|_2^2 = \sum \sigma_i^2 = n\sigma^2 \quad \text{where } \sigma^2 = Var(\varepsilon)$$

So $$\mathbb{E}[MSE(X\hat{\theta}^{LS})] \leq \frac{2n\sigma^2}{n} = 2\sigma^2 .$$

So it seems that taking a large sample actually produces no visible benefit, the MSE $\not\to 0$ as $n \to \infty$.

Q: Is this a real phenomenon? or an inefficiency of our argument?

$\to$ If it is an inefficiency, what conditions can get us better results.

$\boxed{\text{Example:}}$ Gaussian Sequence:

Let $y_i = \sqrt{n}\,\theta_i^* + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$

Then $n = d$ $\mathbb{X} = \sqrt{n}\,\mathbb{1}_n$.

$X^T = \sqrt{n}\,\mathbb{1}_n$

$X^T X = n\,\mathbb{1}_n$

So $\hat{\theta}^{LS} = (X^T X)^{-1} X^T y =$

$$= \frac{1}{n}\sqrt{n}\,y = \frac{1}{\sqrt{n}}\,y$$

So
$$\mathbb{E}[MSE(\hat{\theta}^{LS})] = \frac{1}{n}\mathbb{E}\left[\|X(\hat{\theta}^{LS} - \theta^*)\|_2^2\right]$$
$$= \frac{1}{n}\mathbb{E}\left[\|\sqrt{n}\cdot(\tfrac{y}{\sqrt{n}} - \theta^*)\|_2^2\right]$$
$$= \frac{1}{n}\mathbb{E}\left[\|y - \sqrt{n}\theta^*\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|\varepsilon\|_2^2\right]$$
$$= \sum_{i=1}^{n} \mathbb{E}\frac{\varepsilon_i^2}{n}$$
$$= \sigma^2$$

So Real phenomenon

Question: Could it be that under additional assumptions we could do better?

Where could our argument be inefficient (under assumptions)

Let's see where our argument may be <u>inefficient</u>.

When taking <u>sups</u>, we sup'ed over all of the ℓ_2-unit ball in $\mathbb{R}^n$.

<u>Q: when could this be wasteful?</u>

looking back we replaced

$$\frac{X(\hat{\theta}^{LS} - \theta^*)}{\| \text{---''---} \|}$$

by the sup over $v \in \mathbb{R}^n$, $\|v\| \leq 1$

<u>But</u> this assumes implicitly that

$$Im(X) = \mathbb{R}^n,$$

or that $\quad Rank(X) = n.$

What if rank $(X) = r << n$?

Could we hope for improved bounds?

We need to figure out a way to exploit this additional information.

(1) $Im(X)$ is r-dimensional

$\Rightarrow$ let $\{ e_i : i = 1, \dots, n \}$ be the standard basis

& $\{\mathbb{h}_1, \mathbb{h}_2, \ldots, \mathbb{h}_n\}$ be ~~an~~ orthonormal basis

such that $\mathbb{h}_1, \ldots, \mathbb{h}_r$ form an orthonormal

basis of $\mathrm{Im}(X)$.

(How to do this? Start with $\mathbb{h}_1, \ldots, \mathbb{h}_r$ & complete it with a Hilbert-Schmidt procedure).

So any elt of $\mathrm{Im}(X)$ expressed as a column vector in the $\mathbb{H}$ basis will take the form

$$\begin{pmatrix} h_1 \\ \vdots \\ h_r \\ \hline 0 \end{pmatrix}$$

n-r elts

Next express the $\mathbb{H}$ basis in the $\mathbb{E}$ basis as

$$\mathbb{h}_j = \sum_{k=1}^{n} h_j^k \, \mathbb{e}_k \quad \rightsquigarrow \text{ define the row vectors}$$

$$h_j = (h_j^i)_{i=1}^{n} \quad \text{\& the matrix}$$

Suppose $v = \sum_{j=1}^{n} \alpha_j \mathbb{h}_j$ so $[v]_{\mathbb{H}} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$

Now we can write $v = \sum_{j=1}^{n} \alpha_j \sum_{k=1}^{n} h_j^k \, \mathbb{e}_k$

$$= \sum_{k=1}^{n} e_k \underbrace{\sum_{j=1}^{n} \alpha_j h_j^k}$$

$\boxed{\text{So}}$

$$\begin{pmatrix} \sum_{j=1}^{n} h_j^1 \alpha_j \\ \sum_{j=1}^{n} h_j^2 \alpha_j \\ \vdots \\ \sum_{j=1}^{n} h_j^n \alpha_j \end{pmatrix} = \begin{bmatrix} h_1^1 & h_2^1 & h_3^1 & \cdots & h_n^1 \\ h_1^2 & h_2^2 & h_3^2 & \cdots & h_n^2 \\ \vdots & & & & \\ h_1^n & h_n^2 & \cdots & & h_n^n \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

$$= \begin{bmatrix} h_1 & h_2 & \cdots & h_n \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

$\boxed{\text{Now}}$ this means.

$$[v]_e = [h_1 \cdots h_n] [v]_H$$

& since $\langle h_i, h_j \rangle = \delta_{ij}$ so

$$\overbrace{\begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}}^{= [h_1 - h_n]^T =} [v]_e = [v]_H$$

let $O = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}$, obviously $O^T O = O O^T = I_n$.

$\boxed{SD}$ if $v \in Im(X)$ then

$$O v = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \\ \hline 0 \end{pmatrix} \quad \& \text{ thus } \quad P_r O v = O v$$

Now we have   write φ for $\hat{\theta}^{LS} - \theta^*$

since O orthogonal

$$\| X(\hat{\theta}^{LS} - \theta^*) \|_2 \leq 2 \frac{\langle \varepsilon, X\varphi \rangle}{\|X\varphi\|} \stackrel{\text{since } O \text{ orthogonal}}{=} 2 \frac{\langle O\varepsilon, OX\varphi \rangle}{\|X\varphi\|}$$

$$= 2 \frac{\langle P_r O\varepsilon, P_r O X\varphi \rangle}{\|X\varphi\|} = (*)$$

where $P_r = \left(\begin{array}{c|c} \mathbb{1}_r & 0 \\ \hline 0 & 0 \end{array} \right)$

Since $OX\varphi$ will have its last $(n-r)$ entries 0.

$$(*) \leq 2 \, \| P_r O \varepsilon \|_2$$

Now notice first that if

$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$ then.

(1) $\tilde{\varepsilon} = O\varepsilon \sim N(0, \sigma^2 \mathbb{1}_n)$

$\tilde{\varepsilon} := P_r O \varepsilon \sim N(0, \sigma^2 \mathbb{1}_r)$ &

thus finally $\mathbb{E}\|P_r O\varepsilon\|_2^2 = \mathbb{E}\sum_{i=1}^{r}\tilde{\varepsilon}_i = r\sigma^2$.

In fact the same is true more generally

if ε is a sub-Gaussian vector.

<u>why?</u> here we need to be

a little more careful, let $\varphi = \hat{\theta}^{LS} - \theta^*$

we have

$$\|X\varphi\| \leq 2\left\langle \varepsilon, \frac{X\varphi}{\|X\varphi\|}\right\rangle$$

let <u>as before</u> $h_1, \dots, h_r$ be an orthonormal basis

of $\mathrm{Im}(X)$, & suppose as before

$h_j = \sum_{k=1}^{n} h_j^k e_k$, so this means the vectors

$j=1,\dots,r$

$$\varphi_1 = \begin{pmatrix} h_1^1 \\ \vdots \\ h_1^n \end{pmatrix}, \dots, \varphi_r = \begin{pmatrix} h_r^1 \\ \vdots \\ h_r^n \end{pmatrix}$$ are an o.n. basis for $\mathrm{Im}(X)$.

let $\Phi = [\varphi_1, \dots, \varphi_r] \in \mathbb{R}^{n\times r}$.

<u>In particular</u> if $v \in \mathrm{Im}(X)$ then

$$v = \sum_{j=1}^{r} \alpha_j h_j = \sum_{j=1}^{r} \alpha_j \sum_{k=1}^{n} h_j^k e_k =$$

$$= \sum_{k=1}^{n} e_k \left(\sum_{j=1}^{r} \alpha_j h_j^k \right)$$

So $[v]_{\mathcal{E}} = \Phi [v]_h$.

Thus for any $v \in \mathrm{Im}(X)$, $\exists \nu \in \mathbb{R}^r$ s.that

$$v = \Phi \nu.$$

Also the o.n property of $\{h_j\}_{j=1}^{n}$

$$\Rightarrow \left(\Phi^T \Phi\right)_{ij} = \langle \varphi_i, \varphi_j \rangle = \langle h_i, h_j \rangle = \delta_{ij}$$

(But $\Phi \Phi^T$ makes no sense!)

Thus:
$$\frac{\langle \varepsilon, X(\hat{\theta}^{LS} - \theta^*) \rangle}{\|X(\hat{\theta}^{LS} - \theta^*)\|} = \frac{\langle \varepsilon, \Phi \nu \rangle}{\|\Phi \nu\|_2} = \frac{\langle \Phi^T \varepsilon, \nu \rangle}{\|\nu\|_2}$$

Since $\Phi^T \Phi = \mathbb{1}_r$

$$= \frac{\langle \tilde{\varepsilon}, \nu \rangle}{\|\nu\|_2}, \text{ where } \tilde{\varepsilon} = \Phi^T \varepsilon \in \mathbb{R}^r.$$

$$\leq \sup_{u \in B_2^r} \langle \tilde{\varepsilon}, u \rangle.$$

Suppose now that $\varepsilon \sim \mathrm{sub}\text{-}G_{\sigma^2}(\mathbb{R}^n)$.

Let $u \in \mathbb{S}^{r-1}$, then

$$\mathbb{E}\left[\exp(\lambda \langle u, \tilde{\varepsilon} \rangle)\right] = \mathbb{E}\left[\exp\left(\lambda \langle u, \Phi^T \varepsilon \rangle\right)\right]$$

$$= \mathbb{E}\left\{ \exp\left(\lambda \langle \Phi u, \varepsilon \rangle\right)\right\}$$

Notice that if $u \in S^{r-1}$, then

$$\|\Phi u\|_2^2 = \langle \Phi u, \Phi u \rangle = \langle u, \Phi^T \Phi u \rangle = \|u\|_2^2.$$

So Φu is a unit vector in $\mathbb{R}^n$

$$\Rightarrow \quad \mathbb{E}\left[\exp\left(\lambda \langle \Phi u, \varepsilon \rangle\right)\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$$

So $\tilde{\varepsilon}$ is σ^2-sub-Gaussian $(\mathbb{R}^r)$.

We are now interested in

$$\sup_{u \in B_2^r} \langle \tilde{\varepsilon}, u \rangle, \quad \tilde{\varepsilon} \text{ - is } \sigma^2\text{-sub-Gaussian} (\mathbb{R}^r).$$

We will make use of the following theorem.

Theorem: Let $X \in \mathbb{R}^d$ σ^2-sub-Gaussian vector.

Then for any $\delta > 0$, wp at least $1-\delta$

$$\max_{\theta \in B_2^d} \theta^T X = \max_{\theta \in B_2^d} |\theta^T X| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

Proof: the idea is the following.

We will cover B_2^d with balls of radius $1/2$, centered at the elements of a finite set $\mathcal{N}$.

We will control the $\sup_{B_2^d}$ with the $\underbrace{\text{sup over } \mathcal{N}}_{\text{Simple union bound.}}$

+ a correction term

Let $\mathcal{N}$ be a $1/2$-net of B_2^d,

ie $\forall x \in B_2^d \ \exists y \in \mathcal{N}$ s.that $\|y-x\| < 1/2$.

Then:

$$\sup_{\theta \in B_2^d} \langle \theta, X \rangle = \sup \left\{ \langle \theta, X \rangle \,\middle|\, \theta = z + x,\ z \in \mathcal{N},\ \|x\| \le 1/2 \right\}$$

$$\overset{(*)}{\le} \sup_{\theta \in \mathcal{N}} \langle \theta, X \rangle + \sup_{\theta \in \frac{1}{2} B_2^d} \langle \theta, X \rangle$$

Now $\theta \in \frac{1}{2} B_2^d \iff \|\theta\|_2 \le 1/2$.

so $\langle \theta, X \rangle = \frac{1}{2} \langle 2\theta, X \rangle$ where $2\theta \in B_2^d$

so $\sup_{\theta \in \frac{1}{2} B_2^d} \langle \theta, X \rangle = \frac{1}{2} \sup_{\theta \in B_2^d} \langle \theta, X \rangle$ so

$$(*)\ \frac{1}{2} \sup_{\theta \in B_2^d} \langle \theta, X \rangle \le \sup_{\theta \in \mathcal{N}} \langle \theta, X \rangle + \cancel{\frac{1}{2} \sup_{\theta \in B_2^d} \langle \theta, X \rangle}$$

So $\sup_{\theta \in B_2^d} \langle \theta, X \rangle \le 2 \sup_{\theta \in \mathcal{N}} \langle \theta, X \rangle$.

Thus $\mathbb{P}\left[\max_{\theta \in B_2^d} \langle \theta, X \rangle > t \right] \le \mathbb{P}\left[2 \sup_{\theta \in \mathcal{N}} \langle \theta, X \rangle > t \right]$

$$= \mathbb{P}\left[2 \max_{\theta \in N} \langle \theta, X \rangle > t \right]$$

this is now a max over $|N|$ r.v.s

So a simple union bound gives ← Lemma 1.3.2 in notes

$$\leq |N| \, e^{-t^2/8\sigma^2}.$$ (*)

how big is $|N|$?

Lemma: let $\varepsilon \in (0,1)$ & N be an ε-net of B_2^d.

$\boxed{\text{Then:}}$ Start with $N = \{0\}$, $X = B_2^d \setminus \left(\bigcup_{z \in N} B_\varepsilon(z) \right)$.

while $X \neq \emptyset$

choose a point in X & add it to N.

This will eventually terminate & $\forall x, y \in N$

$|x-y| > \varepsilon$. So if we replaced ε with $\varepsilon/2$

then the $\frac{\varepsilon}{2}$ balls centered at N will be disjoint.

So $$\bigcup_{z \in N} \left\{ z + \frac{\varepsilon}{2} B_2^d \right\} \subset \left(1 + \frac{\varepsilon}{2}\right) B_2^d$$

& comparing their volumes

$$\mathrm{vol}\left(\left(1 + \frac{\varepsilon}{2}\right) B_2 \right) \geq \mathrm{vol}\left(\bigcup_{z \in N} \left\{ z + \frac{\varepsilon}{2} B_2 \right\} \right) = \sum_{z \in N} \mathrm{vol}\left(z + \frac{\varepsilon}{2} B_2\right)$$

so $$\left(1 + \frac{\varepsilon}{2}\right)^d \geq |N| \left(\frac{\varepsilon}{2}\right)^d$$

$$\Rightarrow \quad |N| \le \left(\frac{1+\frac{\varepsilon}{2}}{\varepsilon/2}\right)^d = \left(\frac{2+\varepsilon}{\varepsilon}\right)^d \le \left(\frac{3}{\varepsilon}\right)^d.$$

---

letting $\varepsilon = 1/2$ we get that $|N_{1/2}| \le 6^d$

so $\mathbb{P}\left[\max_{\theta \in B_2} \langle \theta, X \rangle > t\right] \le 6^d e^{-\frac{t^2}{8\sigma^2}} \le \delta$

$$\Rightarrow \quad \exp\left(-\frac{t^2}{8\sigma^2} + d \log 6\right) \le \delta$$

$$\frac{-t^2}{8\sigma^2} + d\log 6 \le \log(\delta)$$

$$\frac{t^2}{8\sigma^2} \geq \log(1/\delta) + d\log 6.$$

$$t^2 \geq 8\sigma^2 \log(6)\, d + 8\sigma^2 \log(1/\delta).$$

so take $t = \sqrt{8\log(6)}\,\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$.

Since $MSE(X\hat{\theta}^{LS}) = \frac{1}{n}\|X\hat{\theta}^{LS} - X\theta^*\|$

we conclude that

wp at least $1-\delta$

$$MSE(X\hat{\theta}^{LS}) \lesssim \frac{\sigma^2 \overbrace{r}^{\text{rank}(X)} + \log(1/\delta)}{n}.$$

So we did a lot of work to go from

$$\frac{d}{n} \quad \text{to} \quad \frac{r}{n}.$$

If in fact if $\text{rank}(X) = d$ then the bound is the same.

In fact: Gaussian sequence model

$$y_i = \sqrt{n}\,\theta_i^* + \varepsilon_i \qquad i = 1, \dots, n$$

$n = d \qquad X = \sqrt{n}\, \mathbb{1}_n \,, \quad \hat{\theta}^{LS} = n^{-1/2} y = (X^T X)^+ X y.$
$= (n \mathbb{1}_n)^{-1} X y = \frac{1}{n}\sqrt{n}\, y = \frac{y}{\sqrt{n}}$

So $\|X(\hat{\theta}^{LS} - \theta^*)\|_2^2 = \sum_{i=1}^{n} \frac{\varepsilon_i^2}{n} = \Omega(1).$

So it seems

if we are to make any progress we must exploit special structure, e.g. sparsity etc.

Noiseless Recovery: we will now start considering the effect of sparsity.

To build some intuition about how this can possibly help us we will first consider exact recovery in the noiseless problem

$$Y = X\theta^*.$$

Suppose we know : $\theta^* \in K \left(\begin{array}{c} k\text{-sparse} \\ \|\theta^*\|_1 \leq m \\ \text{etc} \end{array} \right)$

we can then consider

$$\hat{\theta} = \underset{\theta \in K}{\operatorname{argmin}} \|y - X\theta\|_2^2.$$

Suppose then that $\boxed{\theta^* \in K = B_1(1) = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq 1\}}$

Continuing from the previous calculation

$$\|X(\hat{\theta}^{LS} - \theta^*)\|_2^2 \leq 2\langle \varepsilon, X(\hat{\theta}^{LS} - \theta^*)\rangle$$

$$\leq 2 \sup_{v,w \in K} \langle \varepsilon, X(v-w)\rangle$$

$$= 2 \sup_{v' \in B_1(2)} \langle \varepsilon, Xv'\rangle$$

$$\leq 4 \sup_{v' \in B_1(1)} \langle \varepsilon, Xv'\rangle.$$

Now $\langle \varepsilon, Xv\rangle$ is a linear form in v &

the set $B_1(1)$ is convex, so it can

be written as the convex hull of its extreme points

so $\sup\limits_{v \in B_1(1)} \langle \varepsilon, Xv \rangle$

$= \sup \left\{ \langle \varepsilon, Xv \rangle \,\middle|\, v \text{ is an extreme point of } B_1(1) \right\}$

thinking about $B_1(1)$ we get

$= \max \left\{ \langle \varepsilon, Xv \rangle : v = \pm e_j,\ j = 1, \ldots, d \right\}$

if $\mathbf{X} = [X_1, \ldots, X_n]$ then

columns

$$X e_j = X \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = [X_j]$$

& thus $\langle \varepsilon, X e_j \rangle$ is $\|X_j\|_2^2 \sigma^2$-sub Gaussian

so $\sup \langle \varepsilon, Xv \rangle$ is the max of $2d$

$\left(\max\limits_{j=1..d} \|X_j\|_2^2 \right) \sigma^2$- sub Gaussian r.v's.

& thus

$$\mathbb{E}\left[\operatorname{MSE}(X \hat{\theta}_K^{LS}) \right] \leq \frac{\|X(\hat{\theta}_K^{LS} - \theta^*)\|_2^2}{n}$$

$$\leq \mathbb{E}\left[\frac{4 \max_{j=1,\dots,d} \langle \varepsilon, \pm X e_j \rangle}{n}\right]$$

so

$$\leq \frac{C \sigma \max \|X_j\|_2^2}{n} \sqrt{2\log(2d)}$$

& if $t > 0$

$$\mathbb{P}[\text{MSE} > t] \leq \mathbb{P}\left[\max_{\substack{v = \pm e_j \\ j=1,\dots,d}} \varepsilon^T X v \geq nt/4\right]$$

$$\leq 2d \exp\left(\frac{-n^2 t^2}{16\sigma^2 \max_j \|X_j\|_2^2}\right)$$

If we assume that X is normalized so that

$\max \|X_j\|_2 \leq \sqrt{n}$

& then $\leq 2d \exp\left(\frac{-n t^2}{16\sigma^2}\right)$

& concentration occurs at rate n.

Compare: with previous result in the case where $\text{rank}(X) = n \ (\ll d)$

So sparsity can get us good results even if $\text{rank}(X) \cancel{\ll} d, n$.

Similarly if $K = B_0(k)$

k-sparse vectors

<u>But</u> to solve over $B_0(k)$ we need to solve first

$B_0(1), \ldots, B_0(k-1)$ & for each of these you need to consider ${}_dC_j = \binom{d}{j}$ subsets as possible supports.

So comp. it becomes quickly infeasible.

$\boxed{\text{Rather}}$ than assuming a priori that $\theta^* \in K = B_p(1)$ say, we instead **penalise** solutions with large $\|\cdot\|_p$ norm.

e.g. $$\hat{\theta} \in \operatorname{argmin} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_0 \right\}$$

<u>user-specified regularisation parameter</u>

When does this allow us to get close to θ^*?

To understand the effect the regularisation has w/o the influence of the noise $\rightsquigarrow$ consider noiseless model

e.g. $Y = X\theta^*$, θ^* is sparse but we don't know a priori how sparse

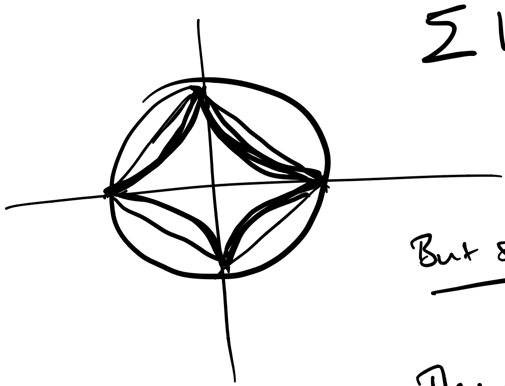
---

Try to solve $\boxed{\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad \text{s.that } X\theta = y}$

Problem non-convex so again we'd have to search $B_0(k)$ for all k.

Idea: replace $\|\cdot\|_0$ by $\|\cdot\|_1$.

We will now try to develop some intuition about geometry.



$\sum |\varepsilon_i|$ vs $\sum |\varepsilon_i|^2$

$\|\varepsilon\|_2 \leq \|\varepsilon\|_1 \leq \sqrt{n}\, \|\varepsilon\|_2$

But suppose ε_i for $i \in I$ are $< \delta$ (small)

The penalty in $L_2 = \sum_{i \in I} \varepsilon_i^2$ which can be much less than the penalty in L_1

Suppose then $y = X\theta^*$, & $\theta_j \neq 0$ $j \in S \subseteq [1:d]$

let $\theta_j = 0$, $j \in S^c \subseteq [1:d]$

support of vector θ.

Q: When does the solution to

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{s.that } X\theta = y \qquad (2)$$

give us the solution to

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \qquad X\theta = y \qquad (1)$$

First let us think about the space of solutions to $X\theta = y$.

$$S(X, y) := \{\theta \in \mathbb{R}^d \mid X\theta = y\}$$

We know θ^* is a solution so

$$S(X, y) = \theta^* + \ker(X)$$

where

$$\ker(X) = \{\theta \in \mathbb{R}^d \mid X\theta = 0\}.$$

Now (2) $\Leftrightarrow \min_{\theta \in S(X,y)} \|\theta\|_1$ so

since $\theta^* \in S(X, y)$ the solution to (2)

will be θ^* iff

θ^* has minimal $\|\cdot\|_1$ norm in $S(X, y)$

But if $x \in \ker(X)$ then $\theta^* + x \in S(X, y)$ also

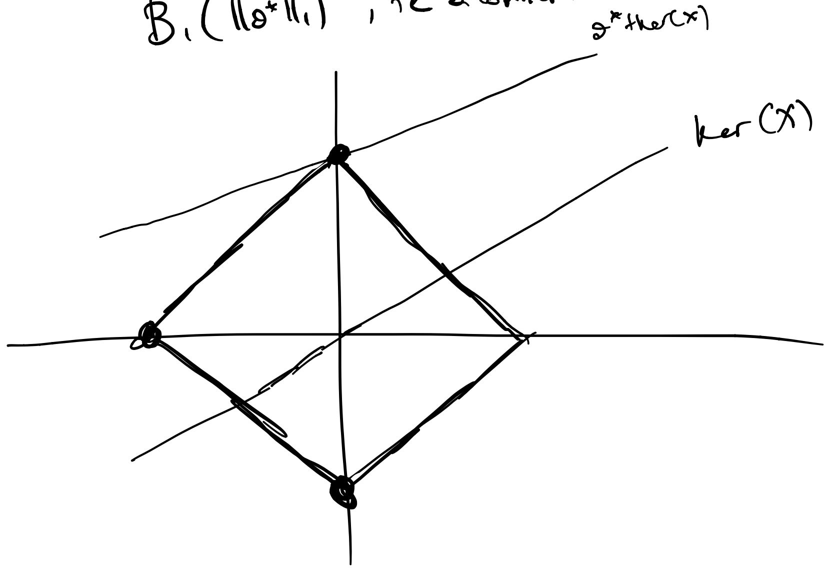
so sol^n to (2) $= \theta^*$

iff $\|\theta^* + x\|_1 \geq \|\theta^*\|_1 \quad \forall x \in \ker(X)$

To visualise the situation consider

$B_1(\|\theta^*\|_1)$

Note: since θ^* is sparse it will be an extreme point of $B_1(\|\theta^*\|_1)$, i.e. a corner.



So (2) will recover the sol^n to (1)

iff $B_1(\|\theta^*\|_1)$ intersects $\theta^* + \ker(X)$ only at θ^*.

Define the set (for $S \subset [1:d]$) CONE

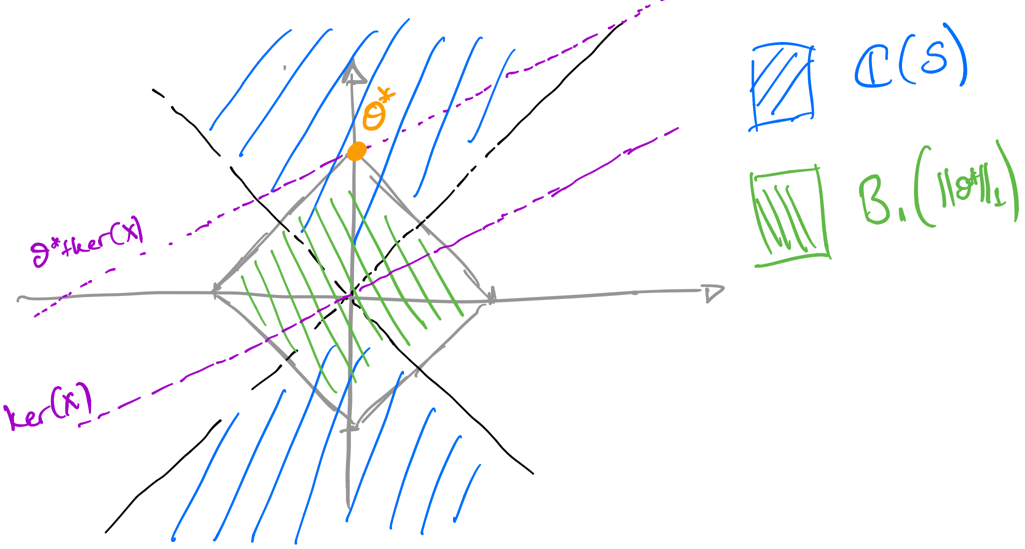
$$\mathbb{C}(S) := \left\{ v \in \mathbb{R}^d \;\middle|\; \|v_{S^c}\|_1 \leq \|v_S\|_1 \right\}$$

where $(v_S)_i = v_i$ if $i \in S$ & 0 o.w.

i.e. $(v_S)_i = v_i \, \mathbb{1}\{i \in S\}$.

Suppose now $\theta^* = (0,1)$ $S = \{2\}$

$\|v_1\| < \|v_2\|$
$\Leftrightarrow$ $v_1 < v_2$



$\boxed{\text{So}}$ $(\theta^* + \ker X) \cap B_1(\|\theta^*\|_1) = \{\theta^*\}$

iff $\ker(X) \cap \mathbb{C}(S) = \{0\}$

Definition: (Restricted Nullspace Property)

We say X satisfies the (RNP) with respect to $S \subseteq [1:d]$ if $\mathbb{C}(S) \cap \ker(X) = \{0\}$

(2) $\min\{\|\theta\|_1 \mid X\theta = y\}$

Thm (8): TFAE

(a) $\forall \theta^*$ supported on $S \subset [1:d]$, (2) applied with $y = X\theta^*$ has unique solution $\hat{\theta} = \theta^*$.

(b) X satisfies RNP wrt S.

Pf: (b) $\Rightarrow$ (a)

We know $Y = X\theta^*$ so

Sufficies to show θ' solution $\Rightarrow \|\theta'\|_1 \geq \|\theta^*\|_1$.

Let $\theta' = \theta^* + v$, $v \in \ker(X) \neq \phi$

(b) $\Rightarrow$ $\|v_{S^c}\|_1 > \|v_S\|_1$

& supp$(\theta^*) = S$, so

$$\|\theta'\|_1 = \|\theta^* + v\|_1 = \|\cancel{\theta^*_{S^c}} + v_{S^c}\|_1 + \|\theta^*_S + v_S\|_1.$$

0 since supp$(\theta^*) = S$

$$= \|v_{S^c}\|_1 + \|\theta^*_S + v_S\|_1 \underset{\Delta\text{ ineq}}{\geq} \|\theta^*_S\|_1 - \|v_S\|_1 + \|v_{S^c}\|$$

$$= \|\theta^*_S\|_1 + \underbrace{(\|v_{S^c}\| - \|v_S\|_1)}_{\geq 0 \text{ by (b)}}$$

$$\geq \|\theta^*_S\|_1.$$

(a) $\Rightarrow$ (b) Let $\theta^* \in \ker(X)$, $\theta^* \neq 0$

Then $X\theta^* = 0$ & thus. rearrange coord's if needed

$$X\begin{bmatrix}\theta^*_S \\ 0\end{bmatrix} + X\begin{bmatrix}0 \\ \theta^*_{S^c}\end{bmatrix} = 0$$

$$\Rightarrow X[\theta^*_S, 0]^T = X[0, -\theta^*_{S^c}]^T =: \tilde{y}$$

$$\Rightarrow \begin{bmatrix}0 \\ -\theta^*_{S^c}\end{bmatrix} \text{ is a sol}^{\underline{n}} \text{ to } X\theta = \left\{X\begin{bmatrix}\theta^*_S \\ 0\end{bmatrix}\right\}$$

i.e. the problem

$$X\theta = \tilde{y} \quad \left(= X\begin{bmatrix}\theta_S^*\\ 0\end{bmatrix}\right)$$

has obviously the solution $\begin{bmatrix}\theta_S^*\\ 0\end{bmatrix}$

but also the solution $\begin{bmatrix}0\\ -\theta_{S^c}^*\end{bmatrix}$.

by (a) since (2) has a unique solution, so it must recover θ_S^* uniquely

it must be that

$$\|\theta_{S^c}^*\| > \|\theta_S^*\|_1 \Rightarrow \ker(X) \cap \mathbb{C}(S) = \{0\}.$$

Verifiable Condition for RNP

If $v \in \ker(X)\setminus\{0\}$, then

$$Xv = 0 \Rightarrow v_1 X_1 + \dots + v_d X_d = 0$$

columns of X

$\Rightarrow \{X_1, \dots, X_d\}$ not linearly independent!

So one obvious way to ensure X satisfies RNP(S) for any $S \subset [1:d]$

is to assume that $\ker(X) = \{0\}$

$\Rightarrow \{X_1, \dots, X_d\}$ lin. independent.

This is already quite restrictive if $d \le n$

& impossible for $d > n$.

Notice that if indeed the colums are lin. ind.

we can transform the model to make them

orthonormal i.e. $\frac{1}{n}\langle \tilde{X}_i, \tilde{X}_j \rangle = \delta_{ij}$

or

$$\boxed{\frac{\tilde{X}^T X}{n} = \mathbb{1}_d \qquad (\text{ORT})}$$

Idea: if we allow ORT to fail in a "controlled" quantifiable way, can we make any progress?

DEF'n (PAIRWISE INCOHERENCE PARAMETER)

For $n \times d$ matrix X define

$$\delta_{PW}(X) := \max_{j,k=1,\dots,d} \left| \frac{1}{n}\langle X_j, X_k \rangle - \delta_{jk} \right|$$

This measures precisely the degree to which ORT fails.

Let $S \subset [1:d]$ $|S| = s$ & suppose

Define $X_S = (X_{i,j})_{i,j \in S}$.

& Suppose that

λ is an eigenvalue of $\frac{X_S^T X_S}{n}$

Then for some $w \in \mathbb{R}^S$

$$\left(\frac{1}{n} X_S^T X_S - \mathbb{1}_S \right) w = (\lambda - 1) w$$

$$|\lambda - 1| \, \|w\|_2 \leq \underbrace{\left\| \frac{1}{n} X_S^T X_S - \mathbb{1} \right\|_2}_{\text{operator norm}} \|w\|_2$$

$$\leq \left\| \frac{1}{n} X_S^T X_S - \mathbb{1} \right\|_F \|w\|_2 \qquad (**)$$

$\|A\|_2 \leq \|A\|_F$ (exercise)

Frobenius norm

$\|A\|_F^2 = \sum_{i,j} |A_{i,j}|^2$

$$(**) \quad |\lambda - 1|^2 \leq \sum_{j,k \in S} \left(\frac{1}{n} \langle X_j, X_k \rangle - \delta_{jk} \right)^2$$

$$\leq |S|^2 \, \delta_{PW}(X)^2 \leq |S|^2 \, \delta^2 / s^2$$

Now if $\gamma \in (0,1) \Rightarrow 1-\lambda \leq \gamma$

$\Rightarrow \lambda \geq 1-\gamma > 0.$

Alternatively

$$\left\| \theta_S^T \left(\frac{X^T X}{n} - 1\right) \right\| \leq \left\| \frac{X^T X}{n} - 1 \right\|_\infty \|\theta_S\|_2^2 = \delta \|\theta_S\|_2^2$$

$$\Rightarrow (1-\delta)\|\theta_S\|_2^2 \leq \theta_S \frac{X^T X}{n} \theta_S$$

Thus for any $\theta \in \mathbb{R}^d$

$$\theta_S^T \frac{X_S^T X_S}{n} \theta_S \geq (1-\gamma)\|\theta_S\|_2^2$$

& since $X_S \theta_S = X \theta_S$

$$\Leftrightarrow \quad \|\theta_S\|_2^2 \leq \frac{1}{1-\gamma} \theta_S^T \frac{X^T X}{n} \theta_S.$$

Let now $\theta \in \ker(X)$ $\quad (\theta = \theta_S + \theta_{S^c})$

so $\boxed{X\theta_S = -X\theta_{S^c}}$

$$\|\theta_S\|_2^2 \leq \frac{1}{1-\gamma} \theta_S^T \frac{X^T X}{n} \theta_S$$

$$= \frac{-1}{1-\gamma} \theta_S^T \frac{X^T X}{n} \theta_{S^c}$$

$$= \frac{1}{1-\gamma} \theta_S^T \left(\mathbb{1} - \frac{X^T X}{n}\right) \theta_{S^c}$$

since $\theta_S^T \theta_S^c = 0$

$$\leq \frac{1}{1-\gamma} \left\| \mathbb{1} - \frac{X^T X}{n} \right\|_\infty \|\theta_{S^c}\|_1 \|\theta_S\|_1. \quad (*)$$

$$\le \frac{1}{1-\gamma}\, \delta_{PW}(X)\, \|\theta_S\|_1\, \|\theta_{S^c}\|_1 .$$

Using ℓ_1-ℓ_2 inequality we get

$$\|\theta_S\|_1^2 \le |S|\, \|\theta_S\|_2^2$$

(*) + $\Rightarrow \quad \|\theta_S\|_1^{\cancel{2}} \le s\, \|\theta_S\|_2^2$

$$\le \frac{s}{1-\gamma}\, \delta_{PW}(X)\, \|\theta_{S^c}\|_1\, \cancel{\|\theta_S\|_1}$$

$$\Rightarrow \quad \|\theta_S\|_1 \le \frac{\cancel{s}}{1-\gamma}\, \frac{\gamma}{\cancel{s}}\, \|\theta_{S^c}\|_1$$

So if $\frac{\gamma}{1-\gamma} < 1$ we have the RNP wrt S with $|S| = s$.

$\Leftrightarrow \quad \gamma < 1-\gamma \Rightarrow$

$\gamma < 1-\gamma$

$2\gamma < 1 \quad \gamma < 1/2 \Rightarrow \quad \|\theta_S\|_1 < \|\theta_{S^c}\|_1$

for any $\theta \in \ker(X) \setminus \{0\}$.

This leads to

Prop: If for some $s \in \mathbb{N}$

$$\delta_{PW}(X) \le \frac{1}{2s} \quad \text{then}$$

RNP(S) holds $\forall\, S \subseteq [1:d]$ with $|S| \le s$.

The LASSO :

We now have some idea about how X & θ^* may be allowed to interact if we may hope that (2) will actually recover the sol^n of (1) in the noiseless setting.

2D Back to $Y = X\theta^* + \varepsilon$

$$(**) \quad \hat{\theta}_L \in \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \tau \|\theta\|_1 \right\}$$

Lasso estimator

Tibshirani 1996

Structure : first slow rates ← minimal assumptions

then we will see an example where we can obviously do better than these rates

⟶ FAST RATE

& then we will consider some assumptions that can guarantee us the fast rates.

Attempt 1: first we will assume that X is normalised so that
$\max_j \|X_j\|_2^2 \le n$.

$$\|Y - X\hat{\theta}^L\|_2^2 = \|X\theta^* + \varepsilon - X\hat{\theta}^L\|_2^2 =$$
$$= \|X(\theta^* - \hat{\theta}^L)\|^2 + 2\langle \varepsilon, X(\theta^* - \hat{\theta}^L)\rangle + \|\varepsilon\|_2^2$$

therefore rearranging $\|X(\theta^* - \hat{\theta}^L)\|_2^2 = \|Y - X\hat{\theta}^L\|_2^2 - 2\langle \varepsilon, X(\theta^* - \hat{\theta}^L)\rangle - \|\varepsilon\|_2^2$ ①

From $\hat{\theta}^L \in \operatorname{argmin}\left\{\frac{1}{2n}\|Y - X\theta\|_2^2 + \tau\|\theta\|_1\right\}$

we have $\frac{1}{n}\|Y - X\theta\|_2^2 + 2\tau\|\hat{\theta}^L\|_1 \le \frac{1}{n}\underbrace{\|Y - X\theta^*\|_2^2}_{\|\varepsilon\|_2^2} + 2\tau\|\theta^*\|_1$ ②

So ① & ② combined give

$$\|X(\theta^* - \hat{\theta}^L)\|_2^2 \le 2\langle \varepsilon, X(\hat{\theta}^L - \theta^*)\rangle + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^L\|_1 + \|\varepsilon\|_2^2 - \|\varepsilon\|_2^2$$
$$= 2\langle X^T\varepsilon, (\hat{\theta}^L - \theta^*)\rangle + 2n\tau(\|\theta^*\|_1 - \|\hat{\theta}^L\|_1)$$
$$= 2\langle X^T\varepsilon, \hat{\theta}^L\rangle - 2n\tau\|\hat{\theta}^L\|_1 + 2(\langle X^T\varepsilon, \theta^*\rangle + n\tau\|\theta^*\|_1)$$
$$\le 2(\|X^T\varepsilon\|_\infty - n\tau)\|\hat{\theta}^L\|_1 + 2(\|X^T\varepsilon\|_\infty + n\tau)\|\theta^*\|_1.$$

Again a union bound gives us

$$\mathbb{P}\left[\|X^T\varepsilon\|_\infty \ge t\right] = \mathbb{P}\left[\max_{j=1,\dots,d} |\langle \varepsilon, X_j\rangle| \ge t\right]$$
$$\le \sum_{j=1}^{d} \mathbb{P}\left[|\langle \varepsilon, X_j\rangle| \ge t\right] \le \sum_{j=1}^{d} 2\exp\left\{-\frac{t^2}{2\sigma^2\|X_j\|_2^2}\right\}$$
$$\le \sum_{j=1}^{d} 2\exp\left\{\frac{-t^2}{2\sigma^2 \max\|X_j\|_2^2}\right\} \le 2d\, e^{-\frac{t}{2n\sigma^2}}$$

since $\boxed{\max\|X_j\|_2^2 \le n.}$

Setting $\mathbb{P}\left[\|X^T\varepsilon\|_\infty \geq t\right] \leq \delta$

which can be guaranteed with

$$t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2\log(1/\delta)n}$$

<u>Choose</u> $\tau = t/n$ so that:

$$\|X(\hat{\theta}^L - \theta^*)\|_2^2 \leq 2\left(\underbrace{\|X^T\varepsilon\|_\infty - n\tau}\right)\|\hat{\theta}^L\| + 2\left(\|X^T\varepsilon\|_\infty + n\tau\right)\|\theta^*\|_1$$

if $t = n\tau$

then $\|X^T\varepsilon\|_\infty < t \Rightarrow$

$$\|X(\hat{\theta}^L - \theta^*)\|_2^2 < 2\left(\|X^T\varepsilon\|_\infty + n\tau\right)\|\theta^*\|_1$$

so $\leq 4n\tau\|\theta^*\|_1$

<u>Since</u> $\|X^T\varepsilon\| < t$ wp $> 1-\delta$ we are done. $\square$

<u>Theorem</u>: Suppose $Y = X\theta^* + \varepsilon$ & let $\hat{\theta}^L$ solve

$$\arg\min\left\{\frac{1}{2n}\|Y - X\theta\|_2^2 + \tau\|\theta\|_1\right\}$$

with $\tau = \sigma\sqrt{\frac{2}{n}}\left(\sqrt{\log(2d)} + \sqrt{\log(1/\delta)}\right)$. Then

wp $> 1-\delta$

$$MSE(X\hat{\theta}^L) \leq \frac{4\|\theta^*\|_1}{\sqrt{n}}\left(\sqrt{2\log(d)} + \sqrt{2\log(1/\delta)}\right).$$

This was obtained from assuming simply that

$$\max\|X_j\|_2^2 \leq n.$$

$\frac{1}{\sqrt{n}}$ (slow rate)

Gaussian Sequence Revisited 6

$y = X\theta + \varepsilon.$

$n = d$, $X = \sqrt{n}\,\mathbb{1}$, so $y_i = \sqrt{n}\,\theta_i^* + \varepsilon_i$

$$\operatorname{argmin}\left[\frac{1}{2n}\|y - X\theta\|_2^2 + \tau\|\theta\|_1\right] \quad (\#).$$

We can transform the original model to obtain (multiply by $\frac{X^T}{n}$)

$$y' := \frac{X^T}{n} y = \frac{X^T X}{n}\theta^* + \xi$$
$$= \theta^* + \xi$$

ξ_i independent $\frac{\sigma^2}{n}$ - sub-Gaussian.

Notice that $\frac{1}{n}\|y - X\theta\|_2^2 = \|\frac{1}{n}X^T(y - X\theta)\|_2^2 = \|y' - \theta\|_2^2$

$$\rightsquigarrow \ (\#) \iff \operatorname{argmin}\left[\|y' - \theta\|_2^2 + 2\tau\|\theta\|_1\right]$$

$$= \operatorname*{argmin}_{\theta_1 \ldots \theta_d} \sum_{i=1}^{d} (y_i' - \theta_i)^2 + 2\tau|\theta_i|$$

$\theta_i > 0 \quad \rightsquigarrow \quad \theta_i^2 - 2y_i'\theta_i + y_i'^2 + 2\tau\theta_i = \theta_i^2 + 2(\tau - y_i')\theta_i + y_i'^2$

$\rightsquigarrow$ minimized at $\hat{\theta}_i = y_i' - \tau$

$\theta_i < 0 \quad \rightsquigarrow \quad \theta_i^2 - 2(\tau + y_i')\theta_i + y_i'^2$

$\rightsquigarrow$ minimized at $\hat{\theta}_i = y_i' + \tau_i$

Thus if $y_i' > \tau \qquad \hat{\theta}_i = y_i' - \tau$

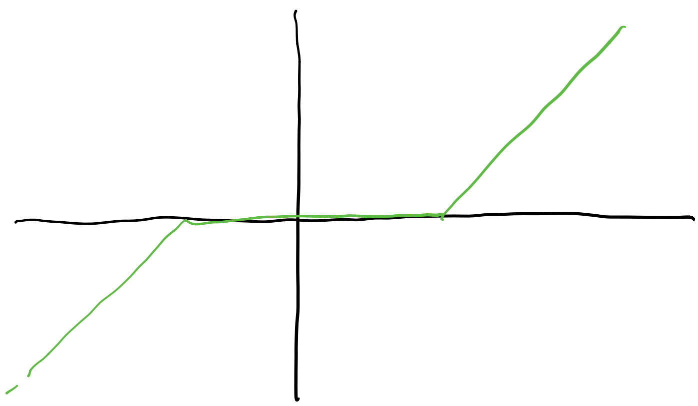
$y_i' < -\tau \qquad \hat{\theta}_i = y_i' + \tau.$

& if $|y_i'| \leq \tau$ no solution so minimum occurs at boundary $\Rightarrow$ $\theta_i = 0$

This can be summarized using the

soft thresholding function

$$T_\tau^{sft}(x) := \begin{cases} \operatorname{sign}(x)\,(|x| - \tau) & |x| \geq \tau. \\ 0 & \text{o.w.} \end{cases}$$



$T_\tau \equiv T_\tau^{sft}$

Then $\hat{\theta}_i = T_\tau(y_i')$ $\quad i = 1, \ldots, d.$,

let $\tau := \tau_n := \sigma \sqrt{\dfrac{2 \log(2d/\delta)}{n}}$

& $A := \left\{ \max_i |\xi_i| \leq \tau/2 \right\}$

Since ξ_i is $\frac{\sigma^2}{n}$ - Sub Gaussian we get

$$\mathbb{P}[A^c] \leq 2d\, e^{-\frac{n\tau^2}{8\sigma^2}} \leq \delta$$

On the event A we can estimate

$$\|\hat{\theta} - \theta^*\|_2^2 = \sum_{j=1}^{d} \left| T_\tau(y_i') - \theta_i^* \right|^2$$

$$= \sum_{j=1}^{d} \Big[\mathbb{1}\{ y_i' \geq \tau \} (\theta_i^* + \xi_i - \tau - \theta_i^*)$$

$$+ \mathbb{1}\{ y_i' \leq -\tau \} (\theta_i^* + \xi_i + \tau - \theta_i^*)$$

$$- \mathbb{1}\{ |y_i| \leq \tau \} |\theta_i^*| \Big]^2$$

$$= \sum_{j=1}^{d} \Big[\mathbb{1}\{ y_i' \geq \tau \} (\xi_i - \tau)$$

$$+ \mathbb{1}\{ y_i' \leq -\tau \} (\xi_i + \tau)$$

$$\cancel{=} \mathbb{1}\{ |y_i'| \leq \tau \} |\theta_i^*| \Big]^2$$

on $\mathcal{A}$ $|\xi|_i \leq \tau/2$ $\forall i$

$$\leq \sum_{j=1}^{d} \Big\{ 2\tau \, \mathbb{1}\{ |y_i'| > \tau \} + \mathbb{1}\{ |y_i'| \leq \tau \} |\theta_i^*| \Big\}^2$$

on $\mathcal{A}$ $\quad |y_i'| \geq \tau \Rightarrow \frac{\tau}{2} \geq |\xi_i| \geq |\xi_i + \theta_i^*| - |\theta_i^*|$

$$\Rightarrow |\theta_j^*| \geq \tau/2$$

& $|y_j'| \leq \tau \Rightarrow |\theta_j^*| \leq \frac{3\tau}{2}$

so $$\|\hat{\theta} - \theta^*\|_2^2 \leq \sum_{j=1}^{d} \Big[2\tau \, \mathbb{1}\{ |\theta_j^*| \geq \tau/2 \} + \mathbb{1}\{ |\theta_j^*| \leq \tfrac{3\tau}{2} \} |\theta_j^*| \Big]^2$$

$$\leq \sum_{j=1}^{d} \Big[4 \min\{ |\theta_j^*|, \tau/2 \} \Big]^2$$

$$\leq \sum_{j=1}^{d} 16 \min\Big\{ |\theta_j^*|, \frac{\tau^2}{4} \Big\} \leq$$

$$\leq \sum_{j \in S} 16 \frac{\tau^2}{4} \leq 4 \|\theta_S^*\|_0 \, \tau_n^2$$

$$\leq \frac{32 \|\theta^*\|_0 \, \sigma^2 \log(2d/\delta)}{n}$$

from def'n of $\tau = \tau_n$.

<u>So</u> for the Gaussian seq^{n} model we get the $\frac{1}{n}$ rate.

In fact all <u>we used</u> was that we can multiply by

$$\frac{X^T}{n} \rightsquigarrow y' = \frac{X^T y}{n} = \theta^* + \xi, \quad \xi = \frac{X^T \varepsilon}{n}$$

<u>So</u> the same rates hold under (ORT)

since we can transform

$$y = X\theta^* + \varepsilon$$

$$\rightsquigarrow y' = \frac{1}{n} X^T X \theta^* + \xi \quad \text{s.t. } \frac{X^T X}{n} = I_n$$

$$= \theta^* + \xi.$$

Why does ORT help so much?

it transforms the problem $y = X\theta^* + \varepsilon$

where θ^* is only observed after

it has been corrupted by the action of the operator X

to a direct problem $y' = \theta^* + \xi$

where there's no corruption

FAST RATES & RESTRICTED EIGENVALUE CONDITION

So far we have seen that we can obtain fast rates ($1/n$) under the ORT condition, at least for the prediction error.

Under the much weaker normalization condition $\max_j \|X_j\|_2 \le \sqrt{n}$ we were only able to obtain the slower $1/\sqrt{n}$ rate.

We will now attempt to obtain the $1/n$ rate w/o ORT.

Here the noiseless scenario will serve as a guide in the following sense:

In the noiseless scenario the RNP was crucial for recovery.

In this setting we will now use a similar albeit stronger assumption

Let $\alpha \geq 1$, $S \in [1:d]$

$$\mathbb{C}_\alpha(S) := \left\{ \theta \in \mathbb{R}^d \,\middle|\, \|v_{S^c}\| \le \alpha \|v_S\|_1 \right\}$$

Definition (Restricted Eigenvalue Condition)

X satisfies (κ, α)-REC over $S \subset [1:d]$ if for all $v \in \mathbb{C}_\alpha(S)$, for some $\kappa > 0$

$$\kappa \|v\|_2^2 \le \frac{1}{n} \|Xv\|_2^2 .$$

Essentially this attempts to control the minimum eigenvalue of the matrix $\frac{1}{n}\cancel{X}^T X$

But of course when $\ker(X) \neq \phi$ this is obviously 0, since $\exists\, v \in \mathbb{R}^d$ $v \neq 0$, $Xv = 0$.

Instead it tries to control the "minimum eigenvalue" when restricted to the subset $\mathbb{C}_\alpha(S)$.

We will see later that the (κ, α)-REC condition is also useful, and in a certain sense necessary, for recovery, i.e. when one wants θ^* rather than simply prediction.

B4 that though, let's first see what result we can obtain using the (κ, α)-REC.

Thm

Suppose $Y = X\theta^* + \varepsilon$, with $\mathrm{supp}(\theta^*) = S \subset [1:d]$, $\varepsilon = (\varepsilon_i)_{i=1}^{d}$ is σ^2-SubGaussian. Suppose also that $\max_j \|X_j\|_2^2 \leq n$ & that X satisfies the $(\kappa, 3)$-REC wrt S. Let $\hat{\theta}^L$ solve

$$\underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \tau \|\theta\|_1 \right\}$$

with $\tau := \tau_n := \sqrt{\dfrac{8\sigma^2 \log(2d/\delta)}{n}}$.

Then wp $> 1 - \delta$

$$\mathrm{MSE}(X\hat{\theta}^L) \le \frac{24\,\|\theta^*\|_0\,\sigma^2 \log(2d/\delta)}{\kappa n}.$$

Pf. Let $\mathcal{A} := \left\{ \max_{j \le n} \frac{1}{n}\langle X_j, \varepsilon\rangle \le \tau/2 \right\}$

Then $\mathbb{P}(\mathcal{A}^c) \le 2d \exp\left(-\frac{n^2\tau^2}{8\sigma^2 \max_j \|X_j\|_2^2}\right) \le 2d \exp\left(\frac{-n\tau^2}{8\sigma^2}\right) \le \delta$

since $\max \|X_j\|_2^2 \le n$.

Let $L(\theta;\tau) := L_n(\theta;\tau) = \frac{1}{2n}\|y - X\theta\|_2^2 + \tau\|\theta\|_1$

& $\Delta := \hat{\theta}^L - \theta^*$.

Since $y - X\hat{\theta}^L = -X\Delta + \varepsilon$

$$L(\hat{\theta}^L;\tau) = \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, \varepsilon\rangle + \frac{\|\varepsilon\|_2^2}{2n} + \tau\|\hat{\theta}^L\|_1$$

By defn of $\hat{\theta}^L$ (argmin)

$$L(\hat{\theta}^L;\tau) \le L(\theta^*;\tau) = \frac{1}{2n}\|\varepsilon\|_2^2 + \tau\|\theta^*\|_1$$

so $\frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, \varepsilon\rangle \cancel{+ \frac{\|\varepsilon\|_2^2}{2n}} + \tau\|\hat{\theta}^L\|_1$

$\le \cancel{\frac{1}{2n}\|\varepsilon\|_2^2} + \tau\|\theta^*\|_1$

so $0 \le \frac{1}{2n}\|X\Delta\|_2^2 \le \frac{1}{n}\langle X\Delta, \varepsilon\rangle + \tau\left(\|\theta^*\|_1 - \|\hat{\theta}^L\|_1\right)$ (*)

$\mathrm{supp}(\theta^*) = S$ by assumption so

$(\hat{\theta}^L)_{S^c} = (\theta^* + \Delta)_{S^c} = \Delta_{S^c}$

$$\begin{aligned}
\|\theta^*\|_1 - \|\hat{\theta}^L\|_1 &= \|\theta^*_S\|_1 - \|\hat{\theta}^L_S\|_1 - \|\hat{\theta}^L_{S^c}\|_1 \\
&= \|\theta^*_S\|_1 - \|\theta^*_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1 \\
&\le \|\theta^*_S\|_1 - \left(\|\theta^*_S\|_1 - \|\Delta_S\|_1\right) - \|\Delta_{S^c}\|_1 = \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1
\end{aligned}$$

$$⊛ \Rightarrow \quad 0 \le \frac{1}{n}\|X\Delta\|_2^2 \le \frac{2}{n}\langle X^T\varepsilon, \Delta\rangle + 2\tau\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right)$$

$$\le 2\|\Delta\|_1 \max_j \frac{|\langle x_j, \varepsilon\rangle|}{n} + 2\tau\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right)$$

and on the event $\mathcal{A}$

$$\le \cancel{2}\|\Delta\|_1 \cdot \frac{\tau}{\cancel{2}} + 2\tau\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right)$$

$$\le \tau\left(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right)$$

<u>Recapping</u>: on the event $\mathcal{A}$

$$0 \le \frac{1}{n}\|X\Delta\|_2^2 \le \tau\left(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right)$$

& thus $\Delta \in \mathbb{C}_3(S)$. $\quad \kappa\|\Delta\|_2^2 \le \frac{1}{n}\|X\Delta\|_2^2$

l_1-l_2 inequality

Continuing from above

$$\frac{1}{n}\|X\Delta\|_2^2 \le 3\tau\|\Delta_S\|_1 \le 3\tau\sqrt{s}\,\|\Delta_S\|_2$$

$$\le 3\tau\sqrt{s}\,\|\Delta\|_2$$

$$\le 3\tau\sqrt{s}\,\frac{\|X\Delta\|_2}{\sqrt{\kappa n}}$$

$$\Rightarrow \quad \frac{1}{\sqrt{n}}\|X\Delta\|_2 \le 3\tau\sqrt{\frac{s}{\kappa}}$$

Recall that $\tau = \tau_n = \sqrt{\dfrac{8\sigma^2\log(2d/\delta)}{\kappa n}}$

$$\|X\Delta\|_2^2 \le 3\cancel{\kappa}\frac{s}{\kappa}\,\frac{8\sigma^2\log(2d/\delta)}{\kappa \cancel{n}}$$

$$\& \quad MSE\left(X\hat{\theta}^L\right) = \frac{1}{n}\|X\Delta\|_2^2 \le \frac{24\,\|\theta^*\|_0\,\sigma^2\log(2d/\delta)}{\kappa\, n}.$$

∎

So with the (κ,α)-REC we recover the $1/n$ error rate.

Bounds on L_2-error

It is in this situation where the (n,α)-REC condition really comes into its own & can be fully appreciated.

To simplify things for now suppose we are looking at

$$\hat{\theta} = \underset{\theta \in K}{\operatorname{argmin}} \frac{1}{2n}\|y - X\theta\|_2^2 = \underset{\theta \in K = B_1(R)}{\operatorname{argmin}} L_n(\theta)$$

$$\underbrace{R = \|\theta^*\|_1}_{\text{truth is feasible}}$$

where $L_n(\theta) := \frac{1}{2n}\|y - X\theta\|_2^2$

$\hat{\theta}$ will minimize the empirical loss above, whereas

$$L(\theta) = \mathbb{E}\left[\|y - X\theta\|_2^2\right] = \mathbb{E}\left[\|X(\theta^* - \theta) + \varepsilon\|_2^2\right]$$

$$= \mathbb{E}\left[\|\varepsilon\|_2^2\right] + \underbrace{\mathbb{E}\left[\langle \varepsilon, X(\theta^* - \theta)\rangle\right]}_{=0} + \|X(\theta^* - \theta)\|_2^2$$

this is clearly minimized at $\theta = \theta^*$.

$\boxed{\text{So}}$ we are essentially minimizing the empirical loss hoping that the solution (when the sample is large enough) will be close to the minimizer θ^* of $L(\theta)$, the true loss.

$\boxed{\text{But:}}$ is this really what is happening?

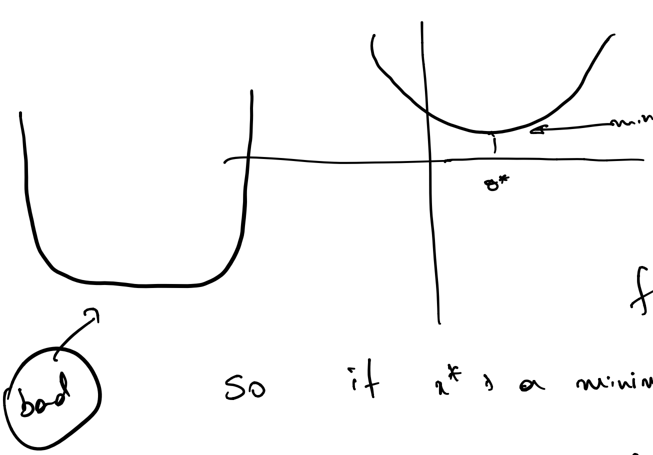
when trying to minimise $L_n(\theta)$, the $n \gg 1$ should guarantee that $L_n(\hat{\theta}) \approx L(\hat{\theta})$

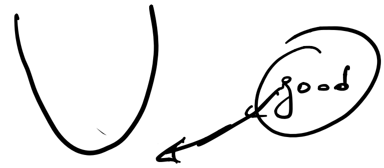
& so that if $L_n(\hat{\theta})$ is small $L(\hat{\theta})$ will also be small.

Think about it differently, if $L_n \approx L$ & $L(\theta^*)$ small then $L_n(\theta^*)$ should also be close to its minimum (i.e.) $|L_n(\theta^*) - L_n(\hat\theta)| \approx$ small

When does this guarantee that $|\theta^* - \hat\theta|$ is also small?

In the 1-d case the picture is quite simple





if function is strongly convex

$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \kappa \|y-x\|_2^2$

So if x^* is a minimiser, $\nabla f(x^*) = 0$

so $f(y) - f(x^*) \geq \kappa \|y - x^*\|_2^2$, so if say

$|f(y) - f(x^*)| < \varepsilon$ (ie y is a near minimizer)

then $\kappa \|y - x^*\|_2^2 \leq \varepsilon \Rightarrow \|y - x^*\|_2^2 \leq \varepsilon/\kappa$.

Now we know that $\hat\theta$ is an minimizer & that θ^* is a near minimizer

so is our function $L_n(\theta)$ strongly convex?

despite "quadratic" it's not

When $d \gg n$ $\frac{1}{2n}\|y - X\theta\|_2^2$ & X has a non-trivial kernel

so along the kernel of X.

The only hope is if we can control the convexity on directions $\perp$ ker(X) & θ^* is $\perp$ ker(X)

Ex: $d=2,\ n=1 \qquad X=[1,0]$

$$y = X[\theta_1^*, \theta_2^*] + \varepsilon = \theta_1^* + \varepsilon$$

let $\theta^* = [1,0]$ & suppose X $(\kappa,3)$-REC wrt SEC
$S=\{1\}$

let $v \in \mathbb{C}_3(S)$, i.e. $3|v_1| \geq |v_2|$, then

$$\|Xv\|_2^2 = v_1^2 = (1-\varepsilon)v_1^2 + \varepsilon v_1^2 \geq (1-\varepsilon)v_1^2 + \frac{\varepsilon}{9}v_2^2 \geq \kappa\|v\|_2^2$$

$$\kappa := \min\{(1-\varepsilon), \varepsilon/9\}$$

If $\hat{\theta}$ is the minimiser

$$(y-X\hat{\theta})^2 + \lambda\|\hat{\theta}\|_1 = (1+\varepsilon-\hat{\theta}_1)^2 + \lambda|\hat{\theta}_1| + \lambda|\hat{\theta}_2|$$

minimised at $\hat{\theta} = [1+\varepsilon-\frac{\lambda}{2}, 0]$

& prediction & ℓ_2 error $(\varepsilon-\lambda)^2$

so for $\varepsilon \ll$ & $\lambda \ll 1$ small — error is small

If now $\theta^* = [0,1]$ so X no longer satisfies $(\kappa,3)$-REC wrt $\{2\}$

for any $\kappa > 0$.

Then $y=\varepsilon$, $(y-X\hat{\theta})^2 + \lambda\|\hat{\theta}\|_1 = (\varepsilon-\hat{\theta}_2)^2 + \lambda|\hat{\theta}_1| + \lambda|\hat{\theta}_2|$

for small λ this is min'ed at

$$\hat{\theta} = [0, \varepsilon-\lambda/2]$$

then prediction error $\|X(\theta^*-\hat{\theta})\|_2^2 = 0$

but ℓ_2 error $\|\theta^*-\hat{\theta}\|_2^2 = (1-\varepsilon+\frac{\lambda}{2})^2$ which is
ε & λ small
is large

Thm

Suppose $Y = X\theta^* + \varepsilon$, with $\mathrm{supp}(\theta^*) = S \subset [1:d]$, $\varepsilon = (\varepsilon_i)_{i=1}^{d}$ is σ^2-SubGaussian. Suppose also that $\max_j \|X_j\|_2^2 \leq n$ & that X satisfies the $(\kappa, 3)$-REC wrt S. Let $\hat{\theta}^L$ solve

$$\underset{\theta}{\mathrm{argmin}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \tau \|\theta\|_1 \right\}$$

with $\tau := \tau_n := \sqrt{\frac{8\sigma^2 \log(2d/\delta)}{n}}$.

Then wp $> 1-\delta$

$$\|\hat{\theta}^L - \theta^*\|_2 \leq \frac{3}{\kappa} \sqrt{\frac{8|S|\sigma^2 \log(2d/\delta)}{n}}$$

Pf: essentially in the last proof, on the event $\mathcal{A}$ we showed $\Delta \in \mathbb{C}_3(S)$, $(\Delta = \hat{\theta}^L - \theta^*)$

From the $(\kappa, 3)$-condition

$$\sqrt{\kappa}\, \|\Delta\|_2 \leq \frac{1}{\sqrt{n}} \|X\Delta\|_2 \leq 3\tau \sqrt{\frac{|S|}{\kappa}}$$

& done $\blacksquare$

RANDOM DESIGN

The idea here is to essentially try and

prove that the conditions on X we have been using so far hold for a random matrix X under some assumptions

Thm: $X \in \mathbb{R}^{n\times d}$, iid $\mathcal{N}(0,1)$ entries. Then $\exists$ universal constants $c_1 < 1 < c_2$ such that

$$(**) \quad \frac{\|X\theta\|_2^2}{n} \geq c_1 \|\theta\|_2^2 - c_2 \frac{\log d}{n} \|\theta\|_1^2 , \quad \theta \in \mathbb{R}^d$$

$$\text{w.p} \geq \quad 1 - \frac{e^{-n/32}}{1-e^{-n/32}} .$$

Pf: Sufficient to look at $\theta \in \mathbb{S}^{d-1}$

Let $g(t) := 2\sqrt{\frac{\log(d)}{n}}\, t$ & define the event

$$\mathcal{E} := \left\{ X \in \mathbb{R}^{nd} \;\middle|\; \inf_{\theta \in \mathbb{S}^{d-1}} \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta\|_1) \right\}$$

$\boxed{\text{exercise:}}$ prove that on $\mathcal{E}^c$ the desired bound $(**)$ holds.

- We need to bound $\mathbb{P}[\mathcal{E}]$.

Although we can now control $\|\theta\|_2^2$, since dimension is high $\|\theta\|_1$ can vary by a factor of $\sqrt{n}$

$\longrightarrow$ Split $\mathcal{E}$ into smaller pieces according to size of $\|\theta\|_1$.

For $0 < r < s$ let

$$K(r,s) := \left\{ \theta \in \mathbb{S}^{d-1} \;\middle|\; g(\|\theta\|_1) \in [r,s] \right\}$$

& consider $\mathcal{A}(r,s) := \left\{ X \,\middle|\, \inf_{\theta \in K(r,s)} \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{2} - 2s \right\}$.

<u>**Claim:**</u> $\mathcal{E} \subset \mathcal{A}(0, 1/4) \cup \bigcup_{l=1}^{\infty} \mathcal{A}\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)$

<u>Pf of Claim:</u> Let θ attain the infimum in $\underline{\mathcal{E}}$ $\left(\frac{\|X\theta\|_2}{\sqrt{n}} = \inf_{\varphi \in \mathbb{S}^{d-1}} \frac{\|X\varphi\|_2}{\sqrt{n}} \right)$

θ must belong to either $K(0, 1/4)$ or one of the $K\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)$

If $\theta \in K(0, 1/4)$ $\quad g(\|\theta\|_1) \leq 1/4$

& for $X \in \mathcal{E}$ $\quad \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2\overset{\geq 0}{g(\|\theta\|_1)} \leq \frac{1}{4} \quad = \frac{1}{2} - \frac{1}{4}$

$\Rightarrow \theta \in \mathcal{A}(0, 1/4)$

<u>Else If</u>: $\theta \in K\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)$, for $l \geq 1$

$$\frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta\|_1) \leq \frac{1}{2} - 2 \times \frac{2^{l-1}}{4} = \frac{1}{2} - \frac{2^l}{4}$$

$\Rightarrow \theta \in \mathcal{A}\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)$.

Union bound $\Rightarrow$ $\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{A}(0, 1/4)] + \sum_{l=1}^{\infty} \mathbb{P}\left[\mathcal{A}\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)\right]$

<u>**Claim:**</u> $\mathbb{P}(\mathcal{A}(r,s)) \leq e^{-n/32}\, e^{-ns^2/2}$

<u>Notice</u> we can aim for a lower bound on

$$T(r,s) = -\inf_{\theta \in K(r,s)} \frac{\|X\theta\|_2}{\sqrt{n}}$$

$$T(r,\lambda) = -\inf_{\theta\in K(r,\lambda)} \sup_{u\in \mathbb{S}^{n-1}} \frac{\langle u, X\theta\rangle}{\sqrt{n}}$$

$$= \sup_{\theta\in K(r,\lambda)} \inf_{u\in \mathbb{S}^{n-1}} \frac{\langle u, X\theta\rangle}{\sqrt{n}}$$

<u>Write</u> $X_{u,\theta} := \langle u, X\theta\rangle$

$X_{u,\theta}$ is a Gaussian process on $\mathbb{S}^{n-1}\times \mathbb{S}^{d-1}$

$X_{u,\theta} \sim \mathcal{N}(0, 1/n)$

We'll use

<u>Thm</u> (Gordon '85) Let $\{X_{ij} : i\in[n], j\in[m]\}$, $\{Y_{ij} : i\in[n], j\in[m]\}$ two Gaussian arrays s.that.

$$\mathbb{E}\left[(X_{ij}-X_{ik})^2\right] \leq \mathbb{E}\left[(Y_{ij}-Y_{ik})^2\right] \quad \forall i,j,k.$$

&

$$\mathbb{E}\left[(X_{ij}-X_{\ell k})^2\right] \geq \mathbb{E}\left[(Y_{ij}-Y_{\ell k})^2\right] \quad i\neq \ell$$

Then $\mathbb{E}\left[\min_i \max_j X_{ij}\right] \leq \mathbb{E}\left[\min_i \max_j Y_{ij}\right]$

Our $X_{u,\theta}$ will be compared with

$$Y_{u,\theta} := \frac{\langle u, \xi\rangle}{\sqrt{n}} + \frac{\langle \theta, J\rangle}{\sqrt{n}}, \quad \xi \sim \mathcal{N}(0, \mathbb{1}_n)$$
$$J \sim \mathcal{N}(0, \mathbb{1}_d)$$

$$\mathbb{E}\left[(X_{u,\theta} - X_{u,\varphi})^2\right] = \mathbb{E}\left[\langle u, X(\theta-\varphi)\rangle\right] = \sum_{i,j} u_i^2 (\theta_j-\varphi_j)^2 = \|u\|_2^2 \, \|\theta-\varphi\|_2^2 = 1.$$

$$\mathbb{E}\left[(Y_{u,\theta} - Y_{u,\varphi})^2\right] = \mathbb{E}\left[\langle J, \theta-\varphi\rangle^2\right] = \|\theta-\varphi\|^2 = 1.$$

if $u \neq w$ $\quad \mathbb{E}\left[\left(Y_{u,\theta} - Y_{w,\varphi}\right)^2\right] = \mathbb{E}\left[\left(\langle u-w, \xi\rangle + \langle \theta-\varphi, \zeta\rangle\right)^2\right]$

$$= \|u-w\|_2^2 + \|\theta-\varphi\|_2^2$$

$$\mathbb{E}\left[\left(X_{u,\theta} - X_{w,\varphi}\right)^2\right] = \mathbb{E}\left[\left(\langle u, X\theta\rangle - \langle w, X\varphi\rangle\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i,j} X_{ij}\left(u_i\theta_j - w_i\varphi_j\right)\right)^2\right]$$

$$= \sum_{i,j}\left(u_i\theta_j - w_i\varphi_j\right)^2 = \|u\theta^T - w\varphi^T\|_F^2$$

Frobenius