Definitions:

The setting of the problem is the following: we have data in form
\[ \{(x_i, y_i) ; i = 1, \ldots, n\} \], three observations of
\[(X, Y) \sim f\]

- \(X\): feature (input) \(X\)
- \(Y\): response (output) \(Y\)
- \(Y\): or label

GOAL: “Learn from data how to predict \(Y\) from \(X\).”

In some ways the best predictor of \(Y\) given \(X\) is the conditional expectation:
\[ E[Y | X = x] \]
e.g. for it we want a function \(f\) that

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

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

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

(\text{since } \mathbb{E}[Y|X], f(x) \text{ are independent of } f(X))
\[
= 0 \quad \text{since } \mathbb{E}[Y - \mathbb{E}[Y|X] | X] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X]
\]

Thus
\[
\mathbb{E} \left[ (Y - f(x))^2 \right] = 0
\]
\[
\begin{align*}
&= \mathbb{E}\left[ (Y - \mathbb{E}[Y|X])^2 \right] + \mathbb{E}[\mathbb{E}[Y|X] - f(x)]^2 \\
&\geq \mathbb{E}[ (Y - \mathbb{E}[Y|X])^2 ] 
\end{align*}
\]

The function \( \mathbb{E}[Y|X=x] \) is called the regression function of \( Y \) on \( X \).

In general, for any loss/risk we'd be done if we knew the law of \( Y \mid X=x \) for all \( x \).

But we only have access to \( (X_1,Y_1), \ldots, (X_n,Y_n) =: S_n \) data.

So our goal is to use \( S_n \) to construct a function \( \hat{f} \) that approximates
\[ f(x) = \mathbb{E}[y | x=n] \], say \( f_n \)

that is we want \( \hat{f}_n : X \rightarrow Y \)

so that

\[ R(f_n) := \mathbb{E}[(y - \hat{f}(x))^2] \]
\[ = \mathbb{E}[(y - \hat{f}_n(x))^2] \]
\[ = \int (y - \hat{f}_n(x))^2 f(d\mu \times dy) \]

is minimized (in some sense w.r.t. \( S_n \))

\[ \text{Note: In } \int (y - \hat{f}_n(x))^2 f(d\mu \times dy) \]
the integral is w.r.t. the distribution of \((x,y)\)
whereas the dependence on the sample \( S_n \) is still there.
\[ \hat{f} \] if \( \hat{f}_n \) depends non-trivially
on $S_n$, the quantity $R(f_n) = \|Y - \hat{f}_n\|^2$ is **random**.

**Note 2:** We proved earlier that

$$R(f_n) = E[(Y - \hat{f}_n)^2] = E[(Y - f(X))^2] + E[(f - \hat{f}_n)^2(X)]$$

$$= E[Y | X]$$

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

$$R(\hat{f}_n) = R(f) + \|f - \hat{f}_n\|^2$$

**approximation error**

**estimation error**

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

\[ f(x) = \mathbb{E}[Y | X=x] \] is the optimal
when we solve
\[ \min \{ \|Y - f\|_1 \mid Y \text{ is } X\text{-visible} \} \]
+ the optimal in the class of 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
\[ f(\text{unknown}) \] from observations \( S_n \)?

**Notes:** bounds on \( \|f_n - f\|_2 \) will be given in expectation with high prob.

\[ f \] if \( Y \) is labeled then one may want \( \mathbb{P}[Y \neq h(x)] \), 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 very useful is that

\[
\text{EMPIRICAL RISK MINIMISATION} \quad \Rightarrow \quad \text{E R M}
\]

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

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

We don't have access to the full measure \( \mathbb{P} \) generating \((X,Y)\) but we have samples from it, so we can approximate it with the empirical measure

\[
\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, y_i) \quad \text{(d\mathbb{L})}
\]

So we may try \( L_f(x,y) = (y-f(x))^2 \)
\[
\hat{f} = \arg \inf_{f \in \mathcal{F}} \quad h_f(L_f) = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2
\]

\[
\arg \inf_{f \in \mathcal{F}} h_f(L_f) = f.
\]

**Notes**

In certain cases, the function class considered (e.g., neural nets) may be rich enough to include \( \theta \) such that \( y_i = g(x_i) \) \( \forall i = 1, \ldots, n \) for all reasonable sample sets \( n \). *A interpolate*

In principle, this may lead to overfitting.

*\&* memorizing the data w/o extracting any information about \( f \).

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

\[
\inf_{g} \left\{ \text{Rn}(g) + \text{penalty}(s) \right\}
\]
where the penalty may penalize overly complicated answers.

Why? e.g. Suppose

\[ y_i = \sin(x_i) + \epsilon_i \]

the orange curve achieves 0 ERM
\[ \hat{f}(x_i) = y_i \quad \text{for all } i = 1, \ldots, n \]

\textbf{But:} the real generating function is \( \sin(\cdot) \)
so potentially here, if we are thinking of using trigonometric functions
we could perhaps use a penalty
\[ \text{of the form } \sum_{k=0}^{\infty} \left( \epsilon_{k+1} \right)^2, \text{ where} \]
\[ \epsilon_{k+1} \text{ is the coefficient of } \exp(2\pi i k x) \]

to penalize "high energy" solutions.

In fact this penalty is equivalent to
$H'$ (Sobolev) norm so it restricts th norm of the derivatives

$$\| \hat{f}^2 + \hat{f}^2 \| \leq C \sum (\| m \|) \| \hat{f}^2 \|$$

$$\leq C_0 (\| \hat{f}^2 + \hat{f}^2 \|)$$

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 otherwise 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

$$P \gg n$$

- Parameters
- Sample size
**HIGH-DIMENSIONAL REGRESSION**

Observation Model

\[ y = X \theta^* + \varepsilon \]

\[ y \in \mathbb{R}^n \]

\[ X = \begin{bmatrix} x_1 \cdots x_d \\ \vdots \\ x_i \cdots x_d \\ \vdots \\ \vdots \end{bmatrix} \in \mathbb{R}^{n \times d} \]

\[ \varepsilon \in \mathbb{R}^n \]

\[ n = \# \text{st observations} \quad d = \# \text{st degrees of freedom?} \]

*fixed design*: \( X \) is deterministic

*random design*: \( X \) is random

**Warm-up**: let's consider first the classical setting i.e. Olson.
The problem is often written as OLS

\[ \hat{\Theta}^{LS} = \arg\min_{\Theta} \| y - X\Theta \|_2^2 \]

\[ = \arg\min_{\Theta} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{o} x_{ij} \theta_j \right)^2 \]

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

\[ X^TX\hat{\Theta}^{LS} = X^Ty \]

Assume for now that " \( X \) has full rank" i.e. its columns are linearly independent.

Claim: then \( X^TX \) is invertible

Pf of claim: Consider \( X_0 \) for some \( \text{weird} \)
then

\[ X\Sigma = \begin{bmatrix} \Sigma_1 & \cdots & \Sigma_d \\ \vdots & \ddots & \vdots \\ \Sigma_1 & \cdots & \Sigma_d \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_d \end{bmatrix} \]

\[ = \begin{bmatrix} \Sigma_1 v_1 + \Sigma_2 v_2 + \cdots + \Sigma_d v_d \\ \vdots \\ \Sigma_1 v_1 + \Sigma_2 v_2 + \cdots + \Sigma_d v_d \end{bmatrix} \]

\[ = \Sigma v_1 \begin{bmatrix} \Sigma_1 \\ \vdots \\ \Sigma_1 \end{bmatrix} + \Sigma v_2 \begin{bmatrix} \Sigma_2 \\ \vdots \\ \Sigma_2 \end{bmatrix} + \cdots + \Sigma v_d \begin{bmatrix} \Sigma_d \\ \vdots \\ \Sigma_d \end{bmatrix} \]

= linear combination of columns of \( \Sigma \).

By assumption, columns of \( X \) are linearly independent.

So \( X\Sigma = 0 \iff \Sigma = 0 \).

Thus \( v^T X^T X v > 0 \) if \( v \neq 0 \).

This is "strictly" positive definite.

\[ \Rightarrow X^T X \text{ is invertible} \]
Thus if \( \text{rank}(X) = d \)

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

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

**What if \( d > n \):**

we still require \( X^T X \hat{\theta}_{LS} = X^T y \)

But now all we can observe is

\[
X \hat{\theta}_{LS}
\]

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

That is now \( \text{rank}(X) < d \)

so \( \text{null}(X) \neq \emptyset \) if \( \emptyset \neq \text{null}(X) \)

\[
X \Theta = X(\hat{\theta} + \phi).
\]

So the system is underdetermined and we cannot hope to obtain a unique solution.

Instead we get a **subspace of solutions**
At this space we may consider regularizing the problem, that is, perhaps we would consider

\[
\min_{\theta} \| \theta \|_2, \quad \text{s.t.} \quad X^T \theta = X^T y \tag{9.46}
\]

find the minimum norm solution

It is not hard to see that the above always admits a unique solution, think of the subspace of solutions to \( X^T \theta = X^T y \) of its distance from the origin.

In fact, this solution above is denoted by

\[
(X^T X)^+ \quad \text{it is known as the Moore-Penrose inverse (it coincides with normal inverse if invertible)}
\]

So we can write the solution to \( X^T \theta = X^T y \) as

\[
\hat{\theta}_L S = (X^T X)^+ X^T y,
\]

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.

**CASE 1:** we are interested in prediction (also known as the sample prediction).

That is, we want to estimate the value $X_0X$ as well as possible from the noisy observations $Y = X_0X + E$.

**CASE 2:** we are interested in the actual value $X_0X$.

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

$$
\text{MSE}(X \hat{\Theta}^{ls}) = \frac{1}{n} \|X_0X - X\hat{\Theta}^{ls}\|_2^2
$$

$$
= (\hat{\Theta}^{ls} - \Theta^*)^T X^T X (\hat{\Theta}^{ls} - \Theta^*)
$$

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

$$
\|\Theta^* - \hat{\Theta}^{ls}\|_2^2.
$$

Let's consider prediction error for now and more specifically let's focus on the MSE. Notice that since $\hat{\Theta}^{ls}$ solves

$$
\min \|\Theta - X_0\|_2^2
$$

we have by definition that...
\[ \hat{\theta}^{LS} \text{ minimizes the } \|y - x\hat{\theta}^{LS}\|_2^2 \text{ among the } \hat{\theta} \text{'s that minimize } \|y - x\theta\|_2^2. \]

So
\[ \|y - x\hat{\theta}^{LS}\|_2^2 \leq \|y - x\hat{\theta}^x\|_2^2. \]

And since by definition
\[ y = x\hat{\theta}^x + \varepsilon, \]

\[ \|y - x\hat{\theta}^{LS}\|_2^2 \leq \|\varepsilon\|_2^2. \]

Also
\[ \|y - x\hat{\theta}^{LS}\|_2^2 \]
\[ = \|y - x\hat{\theta}^x + x\hat{\theta}^x - x\hat{\theta}^x\|_2^2 \]
\[ = \|y - x\hat{\theta}^x\|_2^2 + \|x(\hat{\theta}^{LS} - \hat{\theta}^x)\|_2^2 - 2 \langle y - x\hat{\theta}^x, x(\hat{\theta}^{LS} - \hat{\theta}^x) \rangle \]
\[ = \|\varepsilon\|_2^2 + \|x(\hat{\theta}^{LS} - \hat{\theta}^x)\|_2^2 - 2 \langle \varepsilon, x(\hat{\theta}^{LS} - \hat{\theta}^x) \rangle. \]

Rearranging (2) we get
\[ \|x(\hat{\theta}^{LS} - \hat{\theta}^x)\|_2^2 \leq \|y - x\hat{\theta}^{LS}\|_2^2 - \|\varepsilon\|_2^2 + 2 \langle \varepsilon, x(\hat{\theta}^{LS} - \hat{\theta}^x) \rangle. \]

\[ \leq 2 \langle \varepsilon, x(\hat{\theta}^{LS} - \hat{\theta}^x) \rangle. \]

Now although we've made some progress
\[ \hat{\theta}^{LS} \text{ depends on } \varepsilon, \text{ and the RHS depends on } \varepsilon. \]
One way out is to consider a worst-case scenario that is taken supremum over all but $\varepsilon$.

$$\| \mathbf{X} (\hat{\theta}^{\text{US}} - \theta^\star) \|_2 \leq 2 \left\langle \varepsilon, \frac{\mathbf{X} (\hat{\theta}^{\text{US}} - \theta^\star)}{\| \mathbf{X} (\hat{\theta}^{\text{US}} - \theta^\star) \|_2} \right\rangle.$$ 

$$\varepsilon \leq \| \mathbf{X} (\hat{\theta}^{\text{US}} - \theta) \|_2 \leq 2 \sup \left\{ \left\langle \varepsilon, \mathbf{v} \right\rangle : \mathbf{v} \in \mathbb{R}^p, \| \mathbf{v} \|_2 \leq 1 \right\}$$

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

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

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

So

$$\mathbb{E} \left[ \text{MSE}(\mathbf{X}\hat{\theta}^{\text{US}}) \right] \leq \frac{2n\sigma^2}{n} = 2\sigma^2.$$ 

So it seems that taking a large sample actually produces no visible benefit, the MSE $\to 0$ as $n \to \infty$.

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

→ It is an inefficiency, what condition can get better results.
Example: Gaussian Sequence:

\[ y_i = \sqrt{n} \sigma_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \]

Thus \( n = d \) \( X = \sqrt{n} n^{-1} \).

So \( \hat{\theta}^T = (X^T X)^{-1} X^T y = \frac{1}{n} \sqrt{n} \frac{y}{n} \).

So \( \text{E}[\text{MSE}(\hat{\theta}^T)] = \frac{1}{n} \text{E} \left[ \| X(\hat{\theta} - \theta) \|^2 \right] \)

\[ = \frac{1}{n} \text{E} \left[ \| X \left( \frac{1}{\sqrt{n}} y - \theta \right) \|^2 \right] \]

\[ = \frac{1}{n} \text{E} \left[ \| y - \sqrt{n} \hat{\theta} \|^2 \right] = \frac{1}{n} \text{E} \left[ \| y \|^2 \right] \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \text{E} \epsilon_i^2 = \frac{n}{n} \sigma^2 = \sigma^2 \]

So Real phenomenon

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

Where could our argument be inefficient (under assumption)
Let's see where our argument may be inefficient.

When taking $\sup$s, we suped over all $\bvec{u}$ in the unit ball in $\bvec{R}^m$.

Q: when could this be wasteful?

Looking back we replaced

$$\bvec{X}(\Theta_{\mathcal{U}} - \Theta_{\mathcal{I}})$$

by the $\sup$ over $\bvec{u} \in \bvec{R}^m$.

But this assumes implicitly that

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

or that $\text{rank}(\bvec{X}) = n$.

What if $\text{rank}(\bvec{X}) = r \ll n$?

Could we hope for improved bounds?

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

1. $\text{Im}(\bvec{X})$ is $r$-dimensional.

   * Let $\{ \bvec{e}_i : i = 1, \ldots, n \}$ be the standard basis.
and \( \{ h_1, h_2, \ldots, h_n \} \) be an orthonormal basis such that \( h_1, \ldots, h_r \) form an orthonormal basis of \( \text{Im}(X) \).

(How to do this? Start with \( h_1, \ldots, h_r \) and complete it with a Hilbert-Schmidt procedure.)

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

\[
\begin{pmatrix}
  h_1^i \\
  \vdots \\
  h_r^i \\
  0 \\
  \vdots \\
  0
\end{pmatrix}
\]

Next express the \( H \) basis in the \( E \) basis

\[
h_j = \sum_{k=1}^{r} h_j^k e_k \quad \text{as columns of vectors} \quad h_j = (h_j^k)_{i=1}^n \quad \text{and the matrix}
\]

Suppose \( v = \sum_{j=1}^{r} \alpha_j h_j \) so \( [v]_H = (\alpha_j) \)

Now we can write

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

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

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

\[\text{This means} \quad [\psi] e = [h_1 \ldots h_n] [\psi]_H\]

\[\text{and since} \quad \langle h_i, h_j \rangle = \delta_{ij} \quad \text{so} \]

\[\begin{bmatrix}
h_i \\
h_j
\end{bmatrix} [\psi] e = [\psi]_H\]

\[
\text{let} \quad 0 = [h_i] \text{, obviously } \sigma 0 = 0 \sigma = I_n.
\]
if $x \in \text{Im}(X)$ then

$$O_x = \begin{pmatrix} x \ell \\ i \\ \frac{\sigma}{\ell} \\ 0 \end{pmatrix}$$

d thus

$$P_x O_x = O_x$$

Now we have

$$(\hat{\Theta} - \Theta^t)^t X \leq 2 \frac{\langle \epsilon, X \epsilon \rangle}{n \times q \| \ell \|} \leq 2 \frac{\langle o \epsilon, o \times q \epsilon \rangle}{n \times q \| \ell \|}$$

$$= 2 \frac{\langle P_r o \epsilon, P_r o \times q \epsilon \rangle}{n \times q \| \ell \|} \equiv \Theta$$

where

$$P_r = \begin{pmatrix} 1_r & 0 \\ 0 & 0 \end{pmatrix}$$

since $O \times q$ will have its last $(n-r)$ entries 0.

$$\Theta \leq 2 \| P_r o \epsilon \|_2$$

Now notice that if

$$\epsilon \sim N(0, \sigma^2 I_n)$$

then

$$\Theta \epsilon = 0 \epsilon \sim N(0, \sigma^2 I_n)$$

$$\Theta \epsilon = P_r \epsilon \sim N(0, \sigma^2 I_{n-r})$$
This finally \( E \left\| P_i \sigma_i \right\|^2 = \sum_{i=0}^{c} E \tilde{E}_i = r \sigma^2 \).

In fact the same is true more generally

if \( \varepsilon \) is a sub-Gaussian vector.

why? Here we need to be a little more careful, let \( q = \tilde{\theta} - \theta^* \)

we have

\[
\|Xq\| \leq 2 \left( \varepsilon, \frac{Xq}{\|Xq\|} \right)
\]

Let as before \( H, \), \( H \) be an orthonormal basis of \( \text{Im}(X) \), as suppose are before

\[
h_j = \xi \theta \psi, \text{ so the mean the vector}
\]

\[
\tilde{\xi} = (\begin{array}{c} \tilde{h}_1 \\ \vdots \\ \tilde{h}_r \end{array}) \quad \text{and} \quad \tilde{\psi} = (\begin{array}{c} \tilde{h}_{r+1} \\ \vdots \\ \tilde{h}_{n} \end{array})
\]

Let \( \Phi = [\tilde{\xi}, \ldots, \tilde{\psi}] \in \mathbb{R}^{n \times r} \).

In particular if \( u \in \text{Im}(X) \) then

\[
u = \sum_{j=1}^{r} \alpha_j h_j = \sum_{j=1}^{r} \alpha_j \sum_{k=1}^{n} \tilde{h}_j \psi_k
\]
\[ \sum_{k=1}^{n} \mathbf{e}_k \left( \sum_{j=1}^{m} \alpha_j \mathbf{h}_j \right) \]

\[ \left[ \Phi \mathbf{v} \right]_e = \Phi \left[ \mathbf{v} \right]_h. \]

Thus, for any \( \mathbf{v} \in \text{span}(\mathbf{h}_j), \forall j \in \mathcal{K} \) and \( \mathbf{v} \neq 0 \),

\[ \mathbf{v} = \Phi \mathbf{v}. \]

Also, the orthonormal property of \( \mathbf{h}_j \):

\[ \langle \mathbf{h}_i, \mathbf{h}_j \rangle = \delta_{ij} \]

Thus:

\[ \frac{\langle \mathbf{E}, \mathbf{X} (\mathbf{\Theta}^{\alpha_i} - \mathbf{\Theta}^{\alpha_j}) \rangle}{\mathbf{X} (\mathbf{\Theta}^{\alpha_i} - \mathbf{\Theta}^{\alpha_j}) \|} = \frac{\langle \mathbf{E}, \Phi \mathbf{v} \rangle}{\| \Phi \mathbf{v} \|}, \quad \frac{\langle \Phi^T \mathbf{E}, \mathbf{v} \rangle}{\| \mathbf{v} \|_2} \]

\[ = \frac{\langle \mathbf{E}, \mathbf{v} \rangle}{\| \mathbf{v} \|_2}, \quad \text{where} \quad \mathbf{E} = \Phi^T \mathbf{E} \in \mathbb{R}^r. \]

\[ \leq \sup_{\mathbf{v} \in \mathcal{B}_2} \langle \mathbf{E}, \mathbf{v} \rangle. \]

Suppose now that \( \mathbf{E} \) is sub-Gaussian:

Let \( \mathbf{v} \in \mathcal{B}^{\alpha_i} \), then

\[ \mathbb{E} \left[ \exp(\lambda \langle \mathbf{u}, \mathbf{v} \rangle) \right] = \mathbb{E} \left[ \exp(\lambda \langle \Phi^T \mathbf{E}, \mathbf{v} \rangle) \right]. \]
\[ = \mathbb{E}\left\{ \exp\left(2\langle \Phi u, \varepsilon \rangle\right) \right\} \]

\underline{Note} that if \( u \in B_{d-1}^{\delta} \), then
\[ ||\Phi u||_2^2 = \langle \Phi u, \Phi u \rangle = \langle u, \Phi^2 \Phi u \rangle = \lambda uu^T. \]
So \( \Phi u \) is a unit vector in \( \mathbb{R}^n \)
\[ = \mathbb{E}\left[ \exp\left(2\langle \Phi u, \varepsilon \rangle\right) \right] = e^{\frac{2^2}{\delta^2}} \]

So \( \varepsilon \) is \( \delta \)-sub-Gaussian (\( \mathbb{R}^n \)).

We are now interested in
\[ \sup_{u \in B_{d-1}^{\delta}} \langle \varepsilon, u \rangle \], \( \varepsilon \) - is \( \delta \)-sub-Gaussian (\( \mathbb{R}^n \)).

We will make use of the following theorem.

**Theorem:** Let \( X \in \mathbb{R}^d \) \( \delta \)-sub-Gaussian vector.

Then for any \( \delta > 0 \), up at least \( 1 - \delta \)
\[ \max_{\theta \in \mathbb{B}_2^d} \mathbb{E}^X \min_{\theta \in \mathbb{B}_2^d} \| \theta^T X \|_1 \leq 4\sigma \sqrt{d} + 2\sigma \sqrt{2 \log(1/\delta)}. \]

**Proof:** The idea is the following.

We will cover \( \mathbb{B}_2^d \) with balls, Strading \( \sqrt{2d} \), and...
at the element $\theta$ of a finite set $\mathcal{N}$.

We will control the sup with the sup over $B^d_1$ + a covariant term

Let $\mathcal{N}$ be a $\frac{1}{2}$-net of $B^d_1$.

if $y \in B^d_1 \cap \mathcal{N}$ s.t. $||y - x|| < \frac{1}{2}$.

Then:

$$\sup_{\theta \in B^d_1} \langle \theta, x \rangle = \sup_{\theta \in \mathcal{N}} \left\{ \langle \theta, x \rangle \right\} \leq \sup_{\theta \in \mathcal{N}} \langle \theta, x \rangle + \sup_{\theta \in \frac{1}{2}B^d_1} \langle \theta, x \rangle$$

(Now)

$\theta \in \frac{1}{2}B^d_1 \implies ||\theta|| \leq \frac{1}{2}$.

so

$$\langle \theta, x \rangle = \frac{1}{2} \langle 2\theta, x \rangle$$ when $2\theta \in B^d_1$

so

$$\sup_{\theta \in \frac{1}{2}B^d_1} \langle \theta, x \rangle = \frac{1}{2} \sup_{2\theta \in B^d_1} \langle \theta, x \rangle$$

so

$$\sup_{\theta \in B^d_1} \langle \theta, x \rangle \leq \sup_{\theta \in \mathcal{N}} \langle \theta, x \rangle + \frac{1}{2} \sup_{\theta \in \frac{1}{2}B^d_1} \langle \theta, x \rangle$$

so

$$\sup_{\theta \in B^d_1} \langle \theta, x \rangle \leq 2 \sup_{\theta \in \mathcal{N}} \langle \theta, x \rangle$$

Thus

$$\mathbb{P} \left( \max_{\theta \in B^d_1} \langle \theta, x \rangle > t \right) \leq \mathbb{P} \left( 2 \sup_{\theta \in \mathcal{N}} \langle \theta, x \rangle > t \right)$$
\[
\mathbb{P}\left[ \exists \omega \text{ s.t. } \langle \theta, x \rangle > t \right] \leq |N| e^{-\frac{t^2}{2b^2}}.
\]

How big is \( |N| \)?

**Lemma:** Let \( \epsilon \in (0,1) \) and \( N \) be an \( \epsilon \)-net of \( B^d \).

**Then:** Start with \( N = \{0\} \) and \( X = B^d \setminus \left( \bigcup_{n \in N} B(x, \epsilon) \right) \).

While \( X \neq \emptyset \), choose a point \( x \in X \) and add it to \( N \).

This will eventually terminate for any \( \epsilon > \frac{1}{d} \). So if we replaced \( \epsilon \) with \( \frac{\epsilon}{\sqrt{d}} \), then the \( \epsilon \)-balls centered at \( N \) will be disjoint.

So \( \bigcup_{n \in N} \left\{ \frac{x+\epsilon B_d}{2} \right\} \subset (1+\frac{\epsilon}{\sqrt{d}})B^d \).

\(d\)-dimensional volumes,

\[
\text{vol}\left( (1+\frac{\epsilon}{\sqrt{d}})B^d \right) \geq \text{vol}\left( \bigcup_{n \in N} \left\{ \frac{x+\epsilon B_d}{2} \right\} \right) = \sum_{n \in N} \text{vol}\left( \frac{2+\epsilon B_d}{2} \right) \]

so \( (1+\frac{\epsilon}{\sqrt{d}})^d \geq |N| \left( \frac{\epsilon}{d} \right)^d \).
\[ \Rightarrow |N| \leq \left( \frac{1 + \frac{3}{\varepsilon}}{\varepsilon/\varepsilon} \right)^d = \left( \frac{2 + 3}{\varepsilon} \right)^d \leq \left( \frac{3}{\varepsilon} \right)^d. \]

Let \( \varepsilon = 1/\delta \) we get that \(|N_c| = 6^d\).

So \( \mathbb{P} \left[ \\sup_{\theta \in \mathcal{E}_c} \langle \theta, X \rangle > t \right] \leq 6^d e^{-\frac{t^2}{80^2}} \leq 5 \)

\[ = \exp \left( -\frac{t^2}{80^2} + d \log 6 \right) \leq 5 \]

\[-\frac{t^2}{80^2} + d \log 6 \leq \log 5 \]

\[ \frac{t^2}{80^2} \geq \log (1/5) + d \log 6. \]

\[ t^2 \geq 80^2 \log (6) d + 80^2 \log (1/5). \]

so that \( t = \sqrt{8 \log (6) \sigma \sqrt{d} + 2 \sigma \sqrt{2 \log (1/5)}}. \)

Since \( \text{MS}E(\hat{x}_{\delta^*}^i) = \frac{1}{n} \| x_{\delta^*}^i - x_{\delta}^i \| \)

we conclude that \( \text{MS}E(\delta_{\delta^*}^i) \leq \frac{\sigma^2 \sqrt{d} + \log (1/\delta)}{n} \).
So we didn't do a lot of work to go from
\[ \frac{d}{n} \quad \text{to} \quad \frac{r}{n}. \]

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

**In fact**

Gaussian sequence model

\[ y_i = \sqrt{n} \theta_i + \varepsilon_i \quad i = 1, \ldots, n \]

\( n = d \)

\[ X = \sqrt{n} \Delta_n \quad \hat{\theta}^* = \mathbf{w}^* y = (X^T X)^{-1} X^T y = \frac{1}{n} \mathbf{w}^* \mathbf{y} = \frac{y}{\sqrt{n}} \]

So
\[ \| X (\hat{\theta}^* - \theta^*) \|_2 \leq \frac{\sum \varepsilon_i^2}{\sqrt{n}} = O(1). \]

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 they can possibly help us we will first consider exact recovery in
the no-relax problem
\[ y = X\theta^* . \]

Suppose we know: \( \theta^* \in K \) \( \left( \begin{array}{c} k\text{-span} \\ \| \theta^* \|_2 \leq m \end{array} \right) \)

We can then consider
\[ \hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \| y - X\theta \|_2^2. \]

Suppose further that \( \theta^* \in K = B_1(1) = \{ x \in \mathbb{R}^p \mid \| x \|_1 \leq 1 \} \)

Continuing from the previous calculation,
\[ \| X(\hat{\theta} - \theta^*) \|_2^2 \leq 2 \langle \varepsilon, X(\hat{\theta} - \theta^*) \rangle \]

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

\[ = 2 \sup_{v \in B_1(1)} \langle \varepsilon, XV \rangle \]

\[ \leq 4 \sup_{v \in B_1(1)} \langle \varepsilon, XV \rangle . \]

(Note) \( \langle \varepsilon, Xv \rangle \) is a linear form in \( v \) of the set \( B_1(1) \) is convex, so it can
Be written as the convex hull of its extreme points.

\[
\sup_{x \in \mathbb{B}(1)} \langle \varepsilon, Xx \rangle = \sup \left\{ \langle \varepsilon, X\xi \rangle \mid \xi \text{ is an extreme point of } \mathbb{B}(1) \right\}
\]

Thinking about \( \mathbb{B}(1) \) we get

\[
= \max \left\{ \langle \varepsilon, X\xi \rangle \mid \xi = \pm \nu_j, \quad j = 1, \ldots, d \right\}
\]

If \( X = [x_1, \ldots, x_n] \) then columns

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

d thus \( \langle \varepsilon, X \nu_j \rangle \) is \( \|X_j\|_2 \sigma^2 \)-sub-Gaussian.

So \( \sup \langle \varepsilon, X\xi \rangle \) is the max of \( 2d \)

\[
\left( \max_{j=1}^d \|X_j\|_2 \right) \sigma^2 \text{-sub-Gaussian r.v.'s.}
\]

d thus

\[
\mathbb{E} \left[ \text{MSE}(X\hat{\theta}_k) \right] \leq \frac{\|X(\hat{\theta}_k - \theta^*)\|_2}{n}
\]
$$\mathbb{E} \left[ \frac{\sum \max_{j=1,d} \left( e_{i,j} - e_{i,j}^{0} \right)}{n} \right]$$

So

$$\leq \frac{C \max_{j=1,d} \|X_{j}l_{2}^{n} \sqrt{\log(n)}}{n}$$

If \( t > 0 \)

$$P \left[ \text{MSE} > t \right] \leq P \left[ \max_{j=1,d} \sum_{i=1}^{n} e_{i,j}^{2} X_{i} > n t / \sqrt{d} \right]$$

$$\leq 2d \exp \left( \frac{-n t^{2}}{160^2 \max_{j=1,d} \|X_{j}l_{2}^{n} \}} \right)$$

If we assume that \( X \) is normalized so that

$$\max_{j=1,d} \|X_{j}l_{2}^{n} \leq \sqrt{n}$$

$$d \text{th} \leq 2d \exp \left( \frac{-n t^{2}}{160^2} \right)$$

A concentration occurs at rate \( n \).

\[ \text{Compare: with previous result in the case where } \text{rank}(X) = \chi_{(d)} \]

So sparsity can get good results even if \( \text{rank}(X) \ll d \).

Similarly if \( h = Bo(k) \)
k-sparse vectors

But to solve over \( B_0(k) \) we need to solve for

\( B_0(\cdot) \), \( B_0(k) \) for each subset you need
to consider \( \mathcal{D}_j = \binom{d}{j} \) subsets as possible supports.
So comp. it becomes quickly infeasible.

Rather than asking a priori that \( \Theta^* \in k = B_0(1) \) say,
we instead penalise solutions with large \( \ell_1 \)-norm.

\[
\hat{\Theta} \in \arg\min \left\{ \frac{1}{2n} \|y - X\Theta\|_2^2 + \alpha \|\Theta\|_1 \right\}
\]

User-specified regularisation parameter

When does this allow us to get close to \( \Theta^* \)?

To understand the effect the regularisation has
w/o the influence of noise we consider noiseless model

\[
y = X\Theta^* \quad \Theta^* \text{ is sparse but we don't know a priori how sparse}
\]
Let $y = 0, \forall j \in S \subset I: i, j \neq i$. Suppose

$$\mathcal{L} = \mathcal{E} \cup \mathcal{G},$$

which can be made locally from the penalty.

Now we try to develop some intuition about $\mathcal{L}$.

**Problem**

To reach $B_0(A) \text{, form } l, l$.

Try to solve with hello, that $x = y$.

Hello.
Q: When does the solution to
\[ \min_{\theta \in \mathbb{R}^d} \| \theta \|_1 \text{ s.t. } X\theta = y \]
give us the solution to
\[ \min_{\theta \in \mathbb{R}^d} \| \theta \|_1 \text{ s.t. } X\theta = y \]

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 \( \theta^* \) is a solution so
\[ S(X,y) = \theta^* + \ker(X) \]
where
\[ \ker(X) = \{ \theta \in \mathbb{R}^d \mid X\theta = 0 \} \]

Now \( (2) \implies \min_{\| \theta \|_1} \text{ s.t. } \theta \in S(X,y) \)
\[ \text{ s.t. } \theta^* \in S(X,y) \text{ the solution to } (2) \]
will be \( \theta^* \) iff
\[ \theta^* \text{ has minimum } \| \cdot \|_1 \text{ norm in } S(X,y) \]
But if \( x \in \ker(X) \) then \( \Theta^* x \in S(\ker(Y)) \) also so \( \text{sol}^u \) to \((2) = \Theta^* \)

iff \[ \| \Theta^* x + y \|_1 \geq \| \Theta^* y \|_1 \quad \forall x \in \ker(X) \]

To visualise the situation consider \( B_{\| \Theta^* \|_1} \)

Note: \( \text{so } \Theta^* \) a sparse it will be an extremum point of \( B_{\| \Theta^* \|_1} \), i.e. a corner \( \Theta^* \ker(X) \)

So \((2)\) will recover the \( \text{sol}^u \) to \((1)\)

iff \( B_{\| \Theta^* \|_1} \text{ inwards } \Theta^* \ker(X) \) only at \( 0^u \).

Definition: set \( (\text{for } S \subseteq [1:d]) \) \( \text{CONE} \)

\[ \text{C} \left( S \right) := \left\{ \mathbf{v} \in \mathbb{R}^d \mid \| v_S \|_1 \leq \| v_S \|_1 \right\} \]

where \( (v_S)_s = v_1 \) if \( s \in S \) \& \( 0 \) otherwise.
Suppose now $\Theta^* = (0, 1)$ and $S = \{e_2\}$.

So $(\Theta^* + \ker X) \cap B_r(\|x\|_1) = \{\Theta^*\}$ if and only if $\ker(X) \cap C(s) = \{0\}$.

**Definition (Restricted Nullspace Property)**

We say $X$ satisfies the (RNP) with respect to $S \subseteq \{1, \ldots, d\}$ if $C(s) \cap \ker(X) = \{0\}$.

**Thm. (i) THDAE**

(a) $\forall \Theta^*$ supported on $S \subseteq \{1, \ldots, d\}$

(b) $X$ satisfies RNP w.r.t. $S$,\[\min \{\|x\|_1 \mid x \Theta^* = y\}\]applied with $y = X \Theta^*$ having solution $\hat{B} = \Theta^*.$
\[ \text{Pf} \quad (b) = 0 \quad (a) \]

\[ \text{We know} \quad y = x\theta^* \quad \text{so} \]

\[ \text{Suffices to show} \quad \theta^* \text{ solution } \rightarrow \| \theta' \|_1 > \| \theta^* \|_1. \]

\[ (a) \Rightarrow \theta^* \in S^{u}, \quad u \in \ker (x) \neq 0 \]

\[ (b) \Rightarrow \| u^*_3 \|_1 > \| u^*_3 \|_1 \]

\[ \theta^* \in S, \quad \text{so} \quad 0 \in \text{supp}(\theta^*) \subseteq S \]

\[ \| \theta' \|_1 = \| \theta^* + u^*_3 \|_1 = \| \theta^* \|_1 + \| u^*_3 \|_1 + \| \theta^* + u^*_3 \|_1. \]

\[ = \| u^*_3 \|_1 + \| \theta^* + u^*_3 \|_1 > \| \theta^* \|_1 - \| u^*_3 \|_1 + \| u^*_3 \|_1 \]

\[ \text{by (6)} \]

\[ \| \theta^* \|_1. \]

\[ (a) \Rightarrow (b) \quad \| \theta^* \|_1 \leq \theta^* \in \ker (x) \quad \theta^* \neq 0 \]

Then \( x\theta^* = 0 \) \text{ and hence rearrange coordinates}:

\[ x \begin{bmatrix} \theta^*_S \\ 0 \end{bmatrix} + x \begin{bmatrix} \theta^*_S \\ \theta^*_S \end{bmatrix} = 0 \]

\[ \Rightarrow \quad x \begin{bmatrix} \theta^*_S \\ 0 \end{bmatrix}^T = x \begin{bmatrix} 0, -\theta^*_S \end{bmatrix}^T \quad \Rightarrow \quad \bar{y} \]

\[ \Rightarrow \quad \begin{bmatrix} -\theta^*_S \\ \theta^*_S \end{bmatrix} \text{ is a solution to} \quad x \theta = \left( x \begin{bmatrix} \theta^*_S \\ 0 \end{bmatrix} \right)
The problem
\[ X \Theta = \tilde{y} \quad (= X \Theta^*_s) \]
has obviously the solution
\[ [\Theta^*_s] \]
but also the solution
\[ [0] \cdot \Theta^*_s \]
by (a) since (2) has a unique solution, so its transversal \( \Theta^*_s \) uniquely it must be that
\[ \| \Theta^*_s X \| > \| \Theta^*_s \| \implies \ker(X) \cap C(s) = \{0\} \].

---

**Verifiable Condition for RNP**

If \( u \in \ker(X) \setminus \{0\} \), then
\[ X u = 0 \implies u \perp X_i \quad \text{and} \quad u \Theta^*_s X d = 0 \]
\[ \implies \{X_i, X d\} \text{ not linearly independent!} \]

So one obvious way to ensure \( X \) satisfies RNP(S) for any \( S \in C(1:d) \)
\( \) is to assume that \( \ker(X) = \{0\} \)
\[ \implies \{X_i, X d\} \text{ lin. independent!} \]
This is already quite restrictive if denoted impossible for $d=n$.

Notice that if indeed the columns are linearly independent, we can transform the model to make them orthonormal, i.e., $\langle \hat{X}_i, \hat{X}_j \rangle = \delta_{ij}$.

\[ \frac{1}{n} \hat{X}^T \hat{X} = \mathbf{1}_d \quad \text{(ORT)} \]

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

**Defn (Pairwise Incoherence Parameter)**

For $n \times d$ matrix $X$ define

\[ \delta_{\text{PW}}(X) := \max_{i,j,k=1,\ldots,d} \left| \frac{1}{n} \langle X_i, X_j \rangle - \delta_{ij} \right| \]

This measures precisely the degree to which ORT fails.

Let $S = \{1:d\}$ with $|S| = s$ if suppose...
Define \( X_S = (X_{i,j})_{ij \in S} \).

Suppose that \( \lambda \) is an eigenvalue of \( \frac{X_S^T X_S}{n} \).

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

\[
\left( \frac{1}{n} X_S^T X_S - I_S \right) w = (\lambda - 1) w
\]

\[
|\lambda - 1| \|w\|_2 \leq \left\| \frac{1}{n} X_S^T X_S - I \right\|_F \|w\|_2
\]

\[
\leq \left( \frac{1}{n} X_S^T X_S - I \right) f \|w\|_2
\]

\[
\|A^{1/2}\|_F = \sum_{i,j} |A_{i,j}|^{1/2}
\]

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

\[
\leq \|s\|_2^2 \|\delta_{pw}(\chi)^2 \leq \|s\|_2^2 \delta_{y}^2
\]
\[
\begin{align*}
\text{Now } & r + \gamma \epsilon (0,1) = 0 \quad 1 - 2 \gamma = \epsilon \\
& 2 \geq 1 - \gamma > 0.
\end{align*}
\]

Thus for any \( \theta \in \mathbb{R}^d \)
\[
\theta^\top X_{s}^\top X_s \theta_s \geq \left( \frac{1}{1-\gamma} \right) \| \theta_s \|_2^2 \\
\text{since } X_s \theta_s = X \theta_s
\]

\[
(\Rightarrow) \quad \| \theta_s \|_2^2 \leq \frac{1}{1-\gamma} \theta_s^\top \frac{X^\top X}{n} \theta_s.
\]

Let now \( \theta \in \ker (X) \) \( (\theta = \theta_s + \theta_{s^c}) \)

\[
\begin{align*}
\text{so } & X \theta_s = -X \theta_{s^c} \\
\| \theta_s \|_2^2 \leq \frac{1}{1-\gamma} \theta_s^\top \frac{X^\top X}{n} \theta_s \\
& = \frac{1}{1-\gamma} \theta_s^\top \frac{X^\top X}{n} \theta_{s^c} \\
& = \frac{1}{1-\gamma} \theta_{s^c}^\top \left( \frac{X^\top X}{n} - X^\top X \right) \theta_{s^c} \\
& \text{since } \theta_s^\top \theta_{s^c} = 0 \\
& \leq \frac{1}{1-\gamma} \| d - \frac{X^\top X}{n} \|_\infty \| \theta_{s^c} \|_1 \| \theta_s \|_1.
\end{align*}
\]
Using $\| - \|_2$ inequality we get

$$\|\theta_{x,1}\|^2 \leq 151 \|\theta_{x,1}\|^2$$

$$\Rightarrow \quad \|\theta_{x,1}\|^2 \leq 7 \|\theta_{x,1}\|^2$$

$$\Rightarrow \quad \|\theta_{x,1}\| \leq \frac{8}{1-\frac{5}{8}} \|\theta_{x,1}\|$$

If $\frac{5}{1-\frac{5}{8}} < 1$ we have for RNP with $S$ with $15.1$.

$$\Rightarrow \quad \frac{8}{1-\frac{5}{8}} = 1$$

$$\Rightarrow \quad \theta_{x,1} < \|\theta_{x,1}\|$$

for $\forall \theta \in \text{ker}(x) \setminus \{0\}$.

This leads to

**Prop.** If for some $s \in S$

$$\delta_{pw}(x) \leq \frac{1}{2s}$$

then

RNP $(S)$ holds $\forall S \subseteq \{1, \ldots, d\}$ with $15.1$. 

The LASSO:

We now have some idea about how $X \Delta \theta \Delta$ may be allowed to interact if we may hope that (2) will actually recover the $\theta_{\Delta \theta \Delta}$ in no useless setting.

**Definition**

$$Y = X\theta + \varepsilon$$

$$\hat{\theta}_L \in \arg \min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}$$

Lasso estimator  Tibshirani 1996

**Structure:**
- First slow rates: minimal assumptions
- Then we will see an example when we can obviously do better than these rates
- Fast rate
- Then we will consider some assumptions that can guarantee us the fast rates.
\[\text{Attempt 1:} \text{ Further will assume that } X \text{ is normalized so that } \max_j \|X_j\|_2 = \frac{1}{\sqrt{n}}.\]

\[\|Y - X \delta^c\|_2^2 = \|X\Theta^c - \delta^c\|_2^2 = \|X(\Theta^c - \delta^c)\|_2^2 + 2 \left\langle \epsilon, X(\Theta^c - \delta^c) \right\rangle + \|\Theta^c\|_2^2 \]

\text{Therefore,}

\[\|X(\Theta^c - \delta^c)\|_2^2 = \|Y - X \delta^c\|_2^2 - 2 \left\langle \epsilon, X(\Theta^c - \delta^c) \right\rangle - \|\Theta^c\|_2^2.\]

\[\text{From the equation } \frac{1}{n} \|Y - X \delta^c\|_2^2 + \zeta \|\Theta^c\|_2^2, \]

\[\text{we have } \frac{1}{n} \|Y - X \delta^c\|_2^2 + 2 \epsilon \|\Theta^c\|_2^2 \leq \frac{1}{n} \|Y - X \delta^c\|_2^2 + 2 \zeta \|\Theta^c\|_2^2.\]

\[\text{So 1 and 2 combined give } \]

\[\|X(\Theta^c - \delta^c)\|_2^2 \leq 2 \left\langle \epsilon, X(\delta^c - \Theta^c) \right\rangle + 2 \epsilon \|\Theta^c\|_2^2 \leq 2 \left\langle X \epsilon, (\delta^c - \Theta^c) \right\rangle + 2 \epsilon \|\Theta^c\|_2^2 \]

\[= 2 \left\langle X \epsilon, (\Theta^c - \delta^c) \right\rangle + 2 \zeta (\|\Theta^c\|_2^2 - \|\delta^c\|_2^2) + 2 \left( \left\langle \epsilon, X \delta^c \right\rangle + \zeta \|\Theta^c\|_2^2 \right) \]

\[\leq 2 \left( \|X \epsilon\|_2^2 - \zeta \right) \|\delta^c\|_2^2 + 2 \left( \|X \epsilon\|_2^2 + \zeta \right) \|\Theta^c\|_2^2.\]

\[\text{Again a union bound gives us,}\]

\[P \left[ \max_j \|X_j\|_2 \geq t \right] = P \left[ \max_j \left| \left\langle \epsilon, X_j \right\rangle \right| \geq t \right] \]

\[\leq \sum_{j=1}^d P \left[ \left| \left\langle \epsilon, X_j \right\rangle \right| \geq t \right] \leq \sum_{j=1}^d 2 \exp \left\{ -\frac{t^2}{2 \epsilon^2 \max_j \|X_j\|_2^2} \right\} \]

\[\leq \sum_{j=1}^d 2 \exp \left\{ -\frac{t^2}{2 \epsilon^2 \max \|X_j\|_2^2} \right\} \leq 2d e^{-\frac{t^2}{2m^2}} \]
Setting \( P \left( \| X^* \| \geq t \right) \leq \delta \)
which can be guaranteed with
\[
t = \sigma \sqrt{2 \log(2d)} + \sigma \sqrt{2 \log(n/\delta)}.
\]
Choose \( \gamma = t/n \) so that:
\[
\|X(\hat{\Theta}^n - \Theta^*)\|_2^2 \leq 2 \left( \|X^*\|_2^2 \gamma^2 + 2 \|X^*\|_{\infty} \| \Theta^* \| \right)
\]
if \( t = 2\gamma \)
then \( \|X^*\|_2 < t = 0 \)
\[
\|X(\hat{\Theta}^n - \Theta^*)\|_2^2 < 2 \left( \|X^*\|_{\infty} \gamma^2 \right) \| \Theta^* \|
\]
so
\[
4 \gamma \| \Theta^* \|
\]
Since \( \|X^*\|_2 < t \) wp \( \gtrsim \delta \) we are done. \( \Box \)

**Theorem:** Suppose \( Y = X \Theta^* + \epsilon \) that \( \Theta^* \) some
argmin \( \{ \frac{1}{n} \| Y - X\Theta \|_2 + \| \Theta \| \} \)
with \( \gamma = \sigma \sqrt{\frac{2}{n}} \left( \sqrt{\log(b)} + \sqrt{\log(1/\delta)} \right) \). Then
\[
wp > 1-\delta
\]
\[
\text{MSE}(X\hat{\Theta}^n) \leq \frac{4 \| \Theta^* \|}{\sqrt{n}} \left( \sigma \sqrt{2 \log(b)} + \sqrt{2 \log(1/\delta)} \right).
\]
This was obtained from assuming simply that
\[
\max \|X_j\|_{\infty} \leq n.
\]
\[
\frac{1}{\sqrt{n}} \quad \text{(Slow rate)}
\]
Gaussian Sequence Revised

\[ n = d, \quad X = \sqrt{n} A, \quad \text{so} \quad Y_i = \sqrt{n} \theta_i + \epsilon_i \]

\[
\arg\min \left[ \frac{1}{2n} \| Y - X \theta \|_2^2 + 2 \| \epsilon \|_2 \right]
\]

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

\[
y_i' = \frac{X^T}{n} y_i = \frac{X^T}{n} \theta_i + \frac{\epsilon_i}{n} = \theta_i + \frac{\epsilon_i}{n}
\]

\( \frac{\epsilon_i}{n} \) independent \( \frac{\epsilon_i}{n} \) - sub-Gaussian.

Notice that

\[
\frac{1}{n} \| Y - X \theta \|_2^2 = \frac{1}{n} \| X^T (Y - X \theta) \|_2^2 = \| y_i' - \theta \|_2^2.
\]

Let \( \Theta \in \mathbb{R} \),

\[
\arg\min \left[ \| y_i' - \theta \|_2^2 + 2 \| \theta \|_2 \right]
\]

\[
= \arg\min_{\theta_i, \theta_d} \sum_{i=1}^{d} (y_i' - \theta_i)^2 + 2 \| \theta \|_2
\]

\( \theta_i > 0 \)

\[
\Rightarrow \quad \theta_i = 2 y_i' \theta_i + y_i'^2 + 2 \epsilon_i = \theta_i^2 + 2 (y_i' \epsilon_i) + y_i'^2
\]

\( \Rightarrow \) minimized \( \hat{\theta}_i = y_i' - 2 \)

\( \theta_i < 0 \)

\[
\Rightarrow \quad \theta_i = -2 (y_i' \epsilon_i) \theta_i + y_i'^2
\]

\( \Rightarrow \) minimized at \( \hat{\theta}_i = y_i' + 2 \epsilon_i \)

Thus if \( y_i' > 2 \)

\( \hat{\theta}_i = y_i' - 2 \)

\( y_i' = 2 \)

\( \hat{\theta}_i = y_i' + 2 \epsilon_1 \)

\& if \( \| y_i' \|_2 < 2 \) no solution so minimum occurs at
boundary = 0 \quad \beta = 0

This can be summarized using the

soft thresholding function

\[ T^S_{\lambda} (x) = \begin{cases} \text{sgn}(x) \left( |x| - \lambda \right) & |x| > \lambda \\ 0 & \text{o.w.} \end{cases} \]

Then \( \hat{x}_i = T_{\lambda} (y_i) \) for \( i = 1, \ldots, d \).

Let \( \Delta = \sigma \sqrt{\frac{2 \log(2d/\delta)}{n}} \)

\[ A = \left\{ \text{max} \left| \beta_i \right| \leq \Delta \right\} \]

Since \( \beta_i \) is \( \frac{\sigma^2}{n} \)-sub Gaussian, we get

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

On the event \( A \) we can estimate
\[ \| \hat{\Theta} - \Theta^* \|_2^2 = \sum_{j=1}^{d} \left( \sum_{i=1}^{n} \left( \mathbb{I}\{y_{ij} \geq \tau \} \left( \delta_{ij}^* + \tau - \Theta_{ij}^* \right) \ight.ight. \\
+ \mathbb{I}\{y_{ij} \leq -\tau \} \left( \delta_{ij}^* + \tau + \Theta_{ij}^* \right) \\
- \mathbb{I}\{y_{ij} \leq \tau \} \left| \delta_{ij}^* \right| \bigg)^2 \\
= \sum_{j=1}^{d} \left( \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \geq \tau \} (\Theta_{ij}^* - \tau) \\
+ \mathbb{I}\{y_{ij} \leq -\tau \} (\Theta_{ij}^* + \tau) \right)^2 \\
= \sum_{j=1}^{d} \left( \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \geq \tau \} \Theta_{ij}^* - \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \leq -\tau \} \Theta_{ij}^* \right)^2 \\
\leq \sum_{j=1}^{d} \left( \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \geq \tau \} \theta_{ij}^* - \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \leq -\tau \} \theta_{ij}^* \right)^2 \\
\leq \sum_{j=1}^{d} \left( 2 \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \geq \tau \} \theta_{ij}^* + 2 \sum_{i=1}^{n} \mathbb{I}\{y_{ij} \leq -\tau \} \theta_{ij}^* \right)^2 \\
\leq \sum_{j=1}^{d} \left( 4 \min \{ \theta_{ij}^*, \frac{\tau}{2} \} \right)^2 \\
\leq \sum_{j=1}^{d} 16 \min \{ \theta_{ij}^*, \frac{\tau}{2} \} \frac{\tau}{2} \\
\leq \frac{16}{2 \cdot 3} \frac{\tau^2}{2} = 4 \| \Theta^* \|_2^2 \| \theta \|_2^2 \frac{\tau^2}{2} \]
\[
\leq \frac{32}{n} \frac{10^4 \theta^2 \sigma^2 \log(2s)}{n}
\]

from def'n of \( \gamma \) (\( \gamma = 2^n \)).

So for the Gaussian reg. model we get the \( \frac{1}{n} \) rule.

In fact all we used was that we can multiply by
\[
\frac{X^T}{n} \rightarrow y' = \frac{X^T y}{n} = \theta^* + \tilde{z}, \quad \tilde{z} = \frac{X^T \varepsilon}{n}
\]

So the same rule holds under CORT.

since we can transform
\[
y = X \theta^* + \varepsilon
\]

\[
\Rightarrow y' = \frac{1}{n} X^T X \theta^* + \tilde{z} \quad \text{iff} \quad \frac{X^T X}{n} = I
\]

\[
= \theta^* + \tilde{z}.
\]

Why does CORT help so much?

It transforms the problem \( y = X \theta^* + \varepsilon \)

when \( \theta^* \) is only observed after it has been corrupted by the action of the operator \( X \)

to a direct prob. \( y' = \theta^* + \tilde{z} \)

when there's no compl.
FAST RATES & RESTRICTED EIGENVALUE CONDITION

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

Under the much weaker normalization condition
\[ \|X_i\|_2 \leq \sqrt{n} \]
we were only able to obtain
\[ \text{the slower \( \sqrt{n} \) rate} \]

We will now attempt to obtain the \( \sqrt{n} \) rate w/o ORT.

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

In the noiseless scenario the
\[ \text{RNP was crucial for recovery}. \]

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

\[ \lambda > 1, \ S \subset \{1:d\} \]
\[ C_\alpha (S) := \{ \theta \in \mathbb{R}^d \mid \| u_S \|_2 \leq \alpha \| \nabla \|_2 \} \]

**Definition (Restricted Eigenvalue Condition)**

\( X \) satisfies \((u,x)\)-REC over \( S \subset \{1:d\} \) if for all \( \psi \in C_\alpha (S) \), for some \( k > 0 \)
\[ \| u_S \|_2^2 \leq \frac{k}{n} \| X u_S \|_2^2. \]
Essentially this attempts to control the minimum eigenvalue of the matrix $\frac{1}{n} X^T X$

But of course when $\ker(x) \neq \emptyset$ this is obviously 0, since $\ker(x) \subseteq \ker(\lambda x)$, $x = 0$.

Instead it tries to control the "minimum eigenvalue" when restricted to the subset $C(x(S))$.

We will see later that the $(k,d)$-REC condition is also useful, and in a certain sense necessary for recovery, i.e., when one wants 0-reconstruction simply prediction.

By that though, let's first see what result we can obtain using the $(k,d)$-REC.

Thus suppose $y = X\theta^* + \epsilon$, with $\text{supp}(\theta^*) = S \subseteq \{1: d\}$, 
\[ \epsilon = (\epsilon_1, \ldots, \epsilon_n) \sim \text{sub-Gaussian}, \] 
Suppose also that
\[ \max \|X_j\|^2 \leq \Delta \] 
and that $X$ satisfies the
\[ (k,d) \text{-REC} \] 
Let $\hat{\theta}$ solve
\[ \arg \min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|^2 + \|\theta\|_1 \|\theta\|_1 \right\} \] 
with $z := \Delta := \sqrt{n \frac{8 \sigma^2 \log(2d/\delta)}{n}}$. Then $\omega_p > 1 - \delta$.
\[ \text{MSE}(\hat{\theta}^c) \leq \frac{24 \|\lambda_i\| \sigma^2 \log(2d/\delta)}{kn} \]

\[ \text{Pf: let } A_i = \left\{ \min_{j \leq n} \frac{1}{n} \langle X_{i,j} e \rangle \leq \frac{\varepsilon}{2} \right\} \]

\[ \Pr(\bar{A}_c) \leq 2d \exp\left( -\frac{n^2 \varepsilon^2}{8 \sigma^2 \min_j \|X_{i,j}\|^2} \right) \leq 2d \exp\left( -\frac{n^2 \varepsilon^2}{8 \sigma^2} \right) \leq \delta \]

Since \( \min_j \|X_{i,j}\|^2 \leq n \).

Let \( L(\theta; z) := L_n(\theta; z) = \frac{1}{2n} \|y - \theta^\top x\|^2 + \frac{1}{n} \langle x\Delta, z \rangle \)

and \( \Delta := \hat{\theta}^c - \theta^* \).

Since \( y - x\hat{\theta}^c = -x\Delta + z \)

\[ L(\hat{\theta}^c; z) = \frac{1}{2n} \|x\Delta\|^2_x + \frac{1}{2n} \langle x\Delta, z \rangle + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x + \frac{1}{n} \langle x\Delta, z \rangle = \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \]

By definition of \( \hat{\theta}^c \) (again)

\[ L(\hat{\theta}^c; z) \leq L(\theta^*; z) = \frac{1}{2n} \|x\Delta\|^2_x + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \]

So \( \frac{1}{2n} \|x\Delta\|^2_x + \frac{1}{2n} \langle x\Delta, z \rangle = \frac{1}{2n} \langle x\Delta, z \rangle + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \leq \frac{1}{2n} \langle x\Delta, z \rangle + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \)

So \( 0 \leq \frac{1}{2n} \|x\Delta\|^2_x + \frac{1}{n} \|x\Delta_{\Delta} \|^2_x + \frac{1}{2n} \langle x\Delta, z \rangle + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \leq \frac{1}{2n} \langle x\Delta, z \rangle + \frac{1}{2n} \|x\Delta_{\Delta} \|^2_x \)

Supp(\( \theta^* \)) = \( S \) by assumption so

\[ \|\theta^* S - \hat{\theta}^c S\|_1 = \|\theta^* S_{\delta} - \hat{\theta}^c S_{\delta} - \|\Delta_{\Delta} S\|_1 \leq \|\theta^* S_{\delta} - \hat{\theta}^c S_{\delta} - (\theta^* + \Delta) S_{\delta}\|_1 \leq \|\Delta S S_{\delta}\|_1 \leq \|\Delta S S_{\delta}\|_1 \]
\( \Delta \geq \frac{1}{\sqrt{n}} \| X \Delta \|_2 \leq \frac{2}{\sqrt{n}} \langle X, \Delta \rangle + 2 \varepsilon \left( \| \Delta_{\text{std}} \|_2 - \| \Delta_{\text{sel}} \|_2 \right) \leq 2 \| \Delta_{\text{std}} \|_2 \max_j \| X_j \|_1 + 2 \varepsilon \left( \| \Delta_{\text{std}} \|_2 - \| \Delta_{\text{sel}} \|_2 \right) \)

and on the event \( A \)

\[
\Delta \leq 2 \sqrt{\frac{\| \Delta_{\text{std}} \|_2 \cdot \varepsilon}{4}} + 2 \varepsilon \left( \| \Delta_{\text{std}} \|_2 - \| \Delta_{\text{sel}} \|_2 \right) \leq 2 \left( 3 \| \Delta_{\text{std}} \|_2 - \| \Delta_{\text{sel}} \|_2 \right)
\]

Recovering: on the event \( A \)

\[
0 \leq \frac{1}{\sqrt{n}} \| X \Delta \|_2 \leq 2 \left( 3 \| \Delta_{\text{std}} \|_2 - \| \Delta_{\text{sel}} \|_2 \right)
\]

and thus \( \Delta \in C_0(S) \).

Continuing from above

\[
\frac{1}{\sqrt{n}} \| X \Delta \|_2 \leq 3 \varepsilon \| \Delta_{\text{std}} \|_2 \leq 3 \varepsilon \frac{S}{\sqrt{n}} \| \Delta \|_2 \leq 3 \varepsilon \frac{S}{\sqrt{n}} \| \Delta \|_2
\]

\[
\Rightarrow \quad \frac{1}{\sqrt{n}} \| X \Delta \|_2 \leq 3 \varepsilon \frac{S}{\sqrt{n}}
\]

Recall that \( \varepsilon = \varepsilon_n = \sqrt{\frac{8 \sigma^2 \log(2d/\delta)}{\kappa n}} \)

\[
\| X \Delta \|_2 \leq 3 \varepsilon \frac{S}{\sqrt{n}} \quad \Rightarrow \quad 8 \sigma^2 \frac{\log(2d/\delta)}{\kappa n}
\]

and

\[
\text{MSE}(\hat{X}^{\Delta'}) = \frac{1}{n} \| X \Delta \|_2 \leq 2 \sqrt{\frac{\log^{10} \sigma^2 \frac{\log(2d/\delta)}{\kappa n}}{\kappa n}}.
\]

So with the \((n, \varepsilon_n)\)-REC we recover the \( \| \Delta \|_2 \) error.
 Bounds on L2-error

It is in this situation where the (un-)REC condition really comes into its ownd can be fully appreciated.

To simplify things, we suppose we are looking at

\[ \hat{\Theta} = \arg \min_{\Theta \in K} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \Theta)^2 \]

where \( L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \Theta)^2 \)

\[ \hat{\Theta} \] will minimize the empirical loss above, whereas

\[ L(\Theta) = \mathbb{E} \left[ (y - x^T \Theta)^2 \right] = \mathbb{E} \left[ (\|x(\Theta - \Theta^*) + \epsilon\|^2) \right] \]

\[ = \mathbb{E} \left[ \|x\|^2 \right] + \mathbb{E} \left[ (\epsilon^T \Theta^* x^T (\Theta - \Theta^*) \epsilon) \right] \]

this is clearly increased at \( \|x^T (\Theta - \Theta^*) \|^2 \)

So we essentially minimize one empirical loss, while the solution (assuming enough large enough) will be close to the minimizer \( \hat{\Theta} \) of \( L(\Theta) \), the true loss.

But is this really what is happening?

When trying to minimize \( L(\Theta) \), the \( \Theta \rightarrow \hat{\Theta} \)

Should guarantee that \( L(\Theta) \approx L(\hat{\Theta}) \)

so that if \( L(\Theta) \) is small, \( L(\hat{\Theta}) \) will also be small.

Think about it differently, if \( L(\Theta) \) is \( L(\hat{\Theta}) \) small, then \( L(\Theta) \) should also be close to its minimum.
\[ \frac{|\ln(\theta^*) - \ln(\hat{\theta})|}{\theta^* - \hat{\theta}} = \text{small} \]

When does this guarantee that \[ |\theta^* - \hat{\theta}| \] is also small?

In the 1-d case the picture is quite simple.

If function \( f \) strongly convex

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\alpha} \| y - x \|^2 \]

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

So \( f(y) - f(x^*) \geq \frac{1}{2\alpha} \| y - x^* \|^2 \), so if \( 5 \alpha \)

\[ |f(y) - f(x^*)| < \varepsilon \quad (\text{ie } y \text{ is a new minimizer}) \]

\[ \frac{1}{2\alpha} \| y - x^* \|^2 < \varepsilon \Rightarrow \| y - x^* \|^2 < 2\alpha \varepsilon \]

Now we know that \( \hat{\theta} \) is a minimizer if that \( \theta^* \) is a new minimizer.

So is our function \( \ln(\theta) \) strongly convex?

Despite "quadratic" it's not.

When does \[ \frac{1}{2\alpha} \| y - x^* \|^2 \] belong to the kernel?

So along the kernel \( S^X \).

The only hope is if we can control the convexity on directions \( \parallel \ker(A) \parallel \) \( \| \theta^* \| \parallel \ker(A) \parallel \).
\[ x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad d = 2, \quad n = 1 \]

\[ y = X \begin{bmatrix} \theta^x \\ \theta^s \end{bmatrix} + \varepsilon = \Theta^t \varepsilon \]

\[ w^* \in \Theta^x \cap [1, 0]^d \] s.t. \[ \sup_{x \in X} x (X, \mathcal{X}, \mathcal{R}) = \text{sec} \]

\[ x_{w_1} = (0, 0, 0, 0, 0, 0, 0, 0) \]

\[ \|x_{w_1}\|^2 = \|x_{w_2}\|^2 = \|x_{w_3}\|^2 = \|x_{w_4}\|^2 = \|x_{w_5}\|^2 = \|x_{w_6}\|^2 = \|x_{w_7}\|^2 = \|x_{w_8}\|^2 \]

\[ w := \min \left\{ (1 - c_1, 2/3) \right\} \]

If \( \hat{\theta} \) is the minimizer

\[ (y - \hat{X}\hat{\theta})^2 \leq 2 \|\theta\|_1 = (1 + 2\hat{\theta})^2 + 2 \|\hat{\theta}\|_1 + 2 \|\hat{\theta}\|_1 \]

minimized at \( \hat{\theta} = [1 + 2, -1/2, 0] \)

prediction \( \hat{\varepsilon} \) error \( (\varepsilon - 2)^2 \)

so for \( \varepsilon < 0 \) and \( a < 1 \) small error is small

\[ \text{given } \Theta^x \cap [0, 1] \text{ so } X \text{ no longer white } \]

\( \text{Chua}(X, \mathcal{X}, \mathcal{R}) \subset \{ x \} \)

for any \( k > 0 \).

Then \( y \approx z, \quad (y - \hat{X}\hat{\theta})^2 + 2 \|\hat{\theta}\|_1 = (1 + \hat{\theta})^2 + 2 \|\hat{\theta}\|_1 + 2 \|\hat{\theta}\|_1 \)

for small \( a \), \( \hat{\theta} \) is minimal at

\[ \hat{\theta} = [0, 1 - 2/2] \]

then prediction error \( \|X(\hat{\theta}^x - \Theta^x)\|_1 = 0 \)

but \( \varepsilon \) error \( \|\hat{\theta}^x - \Theta^x\|_1 = (1 - 2\varepsilon)^2 \) \( \frac{1}{2} \text{ which is } \]

\( \varepsilon \) is small

\( \frac{1}{2} \text{ large} \)
Thus
Suppose \( y = X\theta^* + \varepsilon \), with \( \text{supp}(\theta^*) = S \subseteq \{1:d\} \), \( \varepsilon \) is \( \sigma^2 \)-subGaussian, Suppose also that
\[
\max_{j} \| X_j \|_2^2 \leq \frac{c}{n} \quad \text{and that} \quad X \text{ satisfies the} \quad (k, \delta) \text{-REC with } S.
\]
Let \( \hat{\theta}^* \) solve
\[
\text{argmin}_{\theta} \left\{ \frac{1}{2n} \| y - X\theta \|_2^2 + 2 \| \theta \|_1 \right\}
\]
with \( \omega := \sqrt{2} \cdot \frac{\sqrt{8 \sigma^2 \log(2d/\delta)}}{n} \).
Then \( \omega > 1 - \delta \).
\[
\| \hat{\theta}^* - \theta^* \|_2 \leq \frac{3}{k} \sqrt{8 \sigma^2 \log(2d/\delta) \over n}
\]

**Proof:** essentially the same as the proof. On the event \( \mathcal{A} \)
we showed \( \Delta \in C_{3}(\delta) \), \( (\Delta = \hat{\theta}^* - \theta^*) \).
From the \( (k, \delta) \)-condition
\[
\sqrt{k} \| \Delta \|_2 \leq \frac{1}{\sqrt{n}} \| X \Delta \|_2 \leq \delta \sqrt{\frac{12}{\delta}}
\]
\( \& \text{ done} \)

**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

Then: $X \in \mathbb{R}^{n \times d}$, iid $N(0,1)$ entries. Then there exist universal constants $C < c < C_2$ such that

$$\frac{\|X\theta\|^2}{n} \geq C \|\theta\|^2 - C_1 \frac{\log d}{n} \|\theta\|^2 \quad \text{w.p.} \quad 1 - \frac{e^{-n/32}}{1 - e^{-n/32}} \quad .$$

---

**Proof:** Suffices to look at $\theta \in S^{d-1}$

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

$$E := \left\{ X \in \mathbb{R}^{n \times d} \mid \frac{\|X\theta\|^2}{n} \leq \frac{1}{4} - 2 g(\|\theta\|) \right\}$$

**Exercise:** Prove that on $E^c$ the desired bound holds.

- We need to bound $P[E]$.

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

Split $E$ into smaller pieces according to size of $\|\theta\|$.

For $0 < r < 5$

$$K(r,s) := \left\{ \theta \in S^{d-1} \mid g(\|\theta\|) \in [r,s] \right\}$$
Consider \( A(r,s) := \left\{ X \left| \inf_{g \in \mathcal{K}(r,s)} \frac{\|Xg\|_2}{\sqrt{n}} \leq \frac{1}{2} - s \right. \right\} \).

**Claim:** \( E \subseteq A(0, \frac{M}{4}) \cup \bigcup_{\ell = 1}^{\infty} A\left( \frac{2^{\ell+1}}{q}, \frac{2^\ell}{q} \right) \)

**Proof of Claim:** Let \( \Theta \) attain the infimum in \( E \left( \frac{\|X\Theta\|_2}{n^{1/4}} \right) \).

\( \Theta \) must belong to either \( K(0, \frac{M}{4}) \) or one of the \( K\left( \frac{2^{\ell+1}}{q}, \frac{2^\ell}{q} \right) \).

If \( \Theta \in K(0, \frac{M}{4}) \), then \( \Theta \) is in the family \( A(0, \frac{M}{4}) \).

For \( \Theta \in K\left( \frac{2^{\ell+1}}{q}, \frac{2^\ell}{q} \right) \), then

\[
\frac{\|X\Theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2 \gamma(\|\Theta\|_1) \leq \frac{1}{4} - 2 - \frac{2^\ell}{q} = \frac{1}{4} - \frac{2^\ell}{q}
\]

\( \Rightarrow \Theta \in A\left( \frac{2^{\ell+1}}{q}, \frac{2^\ell}{q} \right) \).

**Union Bound:** \( \mathbb{P}[E] \leq \mathbb{P}[E(0, \frac{M}{4})] + \sum_{\ell=0}^{\infty} \mathbb{P}[E\left( \frac{2^{\ell+1}}{q}, \frac{2^\ell}{q} \right)] \)

**Claim:** \( \mathbb{P}\left[ A(r,s) \right] \leq e^{-\frac{r^2}{2s}} e^{-n\gamma^2/2} \)

**Notice:** we can aim for a lower bound on

\( T(r) = \inf_{g \in \mathcal{K}(r,s)} \frac{\|Xg\|_2}{\sqrt{n}} \).
\[ T(r_i) = \inf_{\theta \in K(r_i)} \sup_{u \in \mathbb{S}^{d-1}} \frac{\langle u, X_\theta \rangle}{\sqrt{n}} \]
\[ = \sup_{\theta \in K(r_i)} \inf_{u \in \mathbb{S}^{d-1}} \frac{\langle u, X_\theta \rangle}{\sqrt{n}} \]

Write \( X_{u, \theta} := \langle u, X_\theta \rangle \)

\( X_{u, \theta} \) is a Gaussian process on \( \mathbb{B}^{m-1} \times \mathbb{S}^{d-1} \)

\( X_{u, \theta} \sim N(0, \frac{1}{n}) \)

We'll use

Then (Gordon '85) let \( \{X_{ij} : i \in [n], j \in [m]\}, \{Y_{ij} : i \in [n], j \in [m]\} \)

be two Gaussian arrays such that

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

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

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

\[ Y_{u, \theta} := \frac{\langle u, \bar{a} \rangle}{\sqrt{n}} + \frac{\langle \theta, \sigma \rangle}{\sqrt{n}}, \quad \bar{a} \sim N(0, 1), \quad \sigma \sim N(0, 1) \]

\[ \mathbb{E}\left[ (X_{u, \theta} - Y_{u, \theta})^2 \right] = \mathbb{E}\left[ \langle u, X(\theta, a) \rangle \right] = \sum_{ij} u_i^2 (\theta_j - \theta_i)^2 = \|\theta - \theta_i\|_2^2 = 1. \]

\[ \mathbb{E}\left[ (Y_{u, \theta} - Y_{u, \theta})^2 \right] = \mathbb{E}\left[ \langle \sigma, \theta \rangle^2 \right] = \|\theta\|_2^2 = 1. \]
if \( u \neq w \)

\[
\mathbb{E} \left[ (y_{w,i} - y_{w,q})^2 \right] = \mathbb{E} \left[ (\langle u, w \rangle + \langle w, q \rangle)^2 \right]
\]

\[= \|u - w\|^2 + \|w - q\|^2 \]

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

\[= \mathbb{E} \left[ \left( \sum_{i,j} X_{ij} (u; q_j - w; q_j) \right)^2 \right]
\]

\[= \sum_{i,j} (u; q_j - w; q_j)^2 = \|u - w\|_{F}^2 \]

Frobenius