In the previous class we saw that in the noisy setting we can no longer expect to achieve perfect recovery. Instead we focus on bounding the $\ell_2$ error between a Lasso solution $\hat{\theta}$ and the unknown regression vector $\theta^*$

$$\|\hat{\theta} - \theta^*\|_2 = c \frac{\sqrt{s \log p}}{\sqrt{n}},$$

where $s$ denotes the sparsity, $p$ denotes the ambient dimension and $n$ is the number of samples obtained. Note that this is a huge reduction from the linear rate in $\sqrt{p} \sqrt{n}$ to a logarithmic rate of convergence.

We need to understand if convergence of $\ell_2$ norm is a sufficient metric for recovery of sparse linear models. For example, consider the case where $\theta^* = 0$, $\hat{\theta} = \frac{1}{n} I$. So, $\|\hat{\theta} - \theta^*\|_\infty \leq \frac{1}{n}$, but $S(\theta) = \{1, \cdots, p\} \neq S(\theta^*)$. This example illustrates that just a good convergence rate is not sufficient, the support sets should match too (known as variable selection consistency). We will show in the rest of the lecture that under some assumptions the Lasso solution ensure that the support sets are the same in the limiting case, ie, $S(\hat{\theta}) = S(\theta^*)$.

### 13.1 Lasso Solution

**A1** (Assumption 1): Restricted Eigenvalue Condition states that, if $S = \text{support}(\theta^*)$ then

$$\lambda_{\min}\left(\frac{X^T S X}{n}\right) \geq c_{\min} > 0 \quad (13.1)$$

This assumption is also known as the identifiability condition. We note that the restricted eigenvalue condition alone only assures $\ell_2$ convergence. Another way to state this is

$$\Delta^T \left(\frac{X^T X}{n}\right) \Delta \geq c_{\min} > 0 \quad \forall \Delta : \Delta_{S^c} = 0 \quad (13.2)$$

**A2** (Assumption 2): The irrepresentability assumption states that

$$\|(X^T S X)^{-1} X^T S X_j\|_1 \leq 1 - \alpha < 1 \quad \forall j \in S^c. \quad (13.3)$$

This assumption can be understood as the solution $\hat{w}$ for the ordinary least squares (OLS) problem

$$\hat{w} = \arg \inf_w \|X_j - X_s w\|_2^2.$$ 

This condition implies that the vector $X_j$ is not too related to the support vectors in $X_s$ and hence cannot be represented by them.

Keeping these assumptions in mind, we state one of the main theorems for the Lasso solution.
Theorem 13.1 Consider an $S$-sparse linear regression model for which the design matrix satisfies conditions (A1) and (A2). Then for any choice of regularization parameter such that

$$\lambda_n \geq \frac{2}{\alpha} \| X_S^T \Pi_{S^c}(X)^w \|_\infty$$

(13.4)

the Lagrangian Lasso has the following properties:

1. $\exists$ a unique $\hat{\theta}$ that solves the lasso problem
2. $S(\hat{\theta}) \subseteq S(\theta^*)$.
3. $\| \hat{\theta} - \theta^* \|_\infty \leq \| (X_S^T X_S/n)^{-1} X_S^w/n \|_\infty + \lambda_n \| (X_S^T X_S/n)^{-1} \|_\infty = r_n$
4. No false exclusion of all $j \in S(\theta^*)$ for which $|\theta^*_j| > r_n$

In equation 13.4, $\lambda_n$ gives an upper bound on the maximum noise level in the $(p - s)$ elements in $S^c$. To understand this theorem, we first look at the what the projection matrix $\Pi_{S^c}$ in equation 13.4 means. $\Pi_{S^c}(X) = X_S(X_S^T X_S)^{-1} X_S^T$, where $X_S$ is a $(n \times s)$ matrix and hence $\Pi_{S^c}(X)$ is a $(n \times n)$ matrix. The projection matrix that on pre-multiplying projects a vector $u$ to the $X_S$ space. This can be seen as follows -

$$\hat{\beta} = \arg \inf_{\beta} \| u - X_S \beta \|$$

$$= (X_S^T X_S)^{-1} X_S^T u$$

$$X_S \hat{\beta} = X_S (X_S^T X_S)^{-1} X_S^T u$$

(13.5)

Note that $\Pi_{S^c}(X) = I - \Pi_{S}(X) \neq \Pi_{S^c}(X)$.

Point 4 follows from 3 in that if $\| \hat{\theta} - \theta^* \|_\infty \leq r_n$, then for a false exclusion to occur $|\theta^*_j| > r_n$ would have to be smaller than $r_n$ to go undetected, which directly leads to the conclusion in statement 4.

13.1.1 Side note: Norm Definitions

Before proving the rest of theorem, let's have a quick recap of the definitions of different matrix and vector norms

- Vector $\ell_p$ norm : $\| u \|_p = \left( \sum_{j=1}^p |u_j|^p \right)^{\frac{1}{p}}$
- Matrix operator norm : $\| A \|_p = \sup_{u \neq 0} \frac{\| Au \|_p}{\| u \|_p}$
- Spectral Norm : $\| A \|_2 = \max$ singular value($A$)
- Matrix Infinity norm : $\| A \|_\infty = \max_j \sum_{k=1}^p |A_{jk}|$

13.1.2 Variable Selection Consistency for the Lasso with Gaussian Noise

Theorem 13.1 is a result that applies to any set of linear regression equations. Now, suppose that the noise is Gaussian, i.e., $w_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. There are a few parts to this example:
• Part 1: We will first derive an upper bound for the lower bound of $\lambda_n$ in Theorem 13.1.

• Part 2: Then, looking at Property 3 in Theorem 13.1, we will derive an upper bound for $|| \left( \frac{X^T X_n}{n} \right)^{-1} \frac{X^T w_n}{n} ||_\infty$.

• Put parts 1 and 2 together and make some observations.

**Part 1:** In order to bound $||X^T S \Pi S \perp (X)^w_n ||_\infty$ (which has dimension $(p - s) \times 1$), we consider one element in the vector at a time, and then bound the maximum across all elements. Let $Z_j$ refer to the $j$th entry of the vector:

$$Z_j = X^T j \Pi S \perp (X)^w_n = a^T w$$

for some $j \in S^c$, and where $a = \frac{X^T j \Pi S \perp (X)^w_n}{n}$.

Since $a$ is deterministic and $w_i$ is Gaussian, we know that $Z_j$ is Gaussian with mean 0 and variance $\sigma^2 ||a||_2^2$.

That is,

$$Z_j \sim N(0, \sigma^2 ||a||_2^2)$$

where $||a||_2 = \frac{||X^T j \Pi S \perp (X)||_2}{n}$. If you assume that the columns are normalized, i.e., $\max_{j=1}^p ||X_j||_2 \leq c \sqrt{n}$, then since projection is never expansive (the L2 norm never increases), $||a||_2 = \frac{||X^T j \Pi S \perp (X)||_2}{n} \leq \frac{c}{\sqrt{n}}$.

Then, $Z_j$ is subgaussian with parameter $\frac{c\sigma}{\sqrt{n}}$.

Using the Gaussian tail bound and union bound, we get

$$\mathbb{P}\left( \max_{j \in S^c} |z_j| > t \right) \leq 2(p - s) \exp\left( \frac{-nt^2}{2c^2\sigma^2} \right) \implies \max_{j \in S^c} |z_j| \leq c\sigma \left( \frac{2\log(p - s)}{n} + \delta \right) \text{ w.p. } 1 - 2\exp(\frac{-n\delta^2}{2})$$

Thus, $||X^T S \Pi S \perp (X)^w_n ||_\infty \leq c\sigma \sqrt{\frac{2\log(p - s)}{n}} + \delta$.

**Part 2:** We also have to bound $|| \left( \frac{X^T S X_n}{n} \right)^{-1} \frac{X^T w_n}{n} ||_\infty$ (an $s \times 1$ vector). To do this, we introduce

$$\tilde{z}_j = e^T j \left( \frac{X^T S X_n}{n} \right)^{-1} \frac{X^T w_n}{n} = a^T w$$

for some $j \in \{1, 2, ..., s\}$ where $a^T = e^T j \left( \frac{X^T S X_n}{n} \right)^{-1} \frac{X^T}{n}$. Note that $e_j$ is the $j$th standard basis vector (0’s everywhere and a 1 in the $j$th dimension). As before, we compute $||a||_2$:
\[
\|a\|_2 = a^T a = e_j^T \left(X_s^T X_s \right)^{-1} \left(X_s^T X_s \right)^{-1} e_j \\
= \frac{1}{n} e_j^T \left(X_s^T X_s \right)^{-1} e_j \\
\leq \frac{1}{n} \left\| \left(X_s^T X_s \right)^{-1} \right\|_2 \\
= c_{\text{min}}/n
\]

where \( c_{\text{min}} = \left\| \left(X_s^T X_s \right)^{-1} \right\|_2 \). Thus, using a similar bound on the maximum absolute value of \( \hat{z}_j \)'s, we can bound the \( l_\infty \) norm as follows:

\[
\left\| \left(X_s^T X_s \right)^{-1} X_s^T w \right\|_\infty \leq \frac{1}{\sqrt{c_{\text{min}}}} \left\{ \sqrt{\frac{2\log s}{n} + \sigma} \right\}
\]

**Putting both parts together:** We have bounded with high probability that

\[
\|X_{S^c}^T \Pi_{S^c} (X) w \|_\infty \leq \max_{j \in S^c} |z_j| \leq c\sigma \left( \sqrt{\frac{2\log(p-s)}{n}} + \delta \right)
\]

and that

\[
\left\| \left(X_s^T X_s \right)^{-1} X_s^T w \right\|_\infty \leq \frac{1}{\sqrt{c_{\text{min}}}} \left\{ \sqrt{\frac{2\log s}{n} + \sigma} \right\}
\]

Plugging into Theorem 13.1, we see that we get within \( \sqrt{\log p} \) of the best possible rate.

### 13.1.3 Side Note: Sub-gradients

Let \( f \) be a convex function. When \( f \) is differentiable, the line \( \overline{f}(\theta) = f(\theta_0) + \nabla f(\theta_0)(\theta - \theta_0) \) lies below \( f(\theta) \). However, this requires differentiability.

\( z \) is a sub-gradient of \( f \) at \( \theta_0 \) iff \( f(\theta) \geq \nabla f(\theta_0) + \langle z, \theta - \theta_0 \rangle \) \( \forall \theta \in \Theta \). When \( f \) is not differentiable, we could have several \( z \) values that satisfy this. The sub-differential is the set of all such \( z \) values. A sub-gradient is one such \( z \) value.

If \( f(\theta) \) is the L-1 norm, then when \( \theta \) is non-zero, it is differentiable and its gradient is the sign of \( \theta \). When \( \theta = 0 \), any tangent line with slope between \(-1\) to \( 1 \) lies below \( f(\theta) \) (see Figure 13.1.3).
When solving for a stationary point \((\min_{\theta} f(\theta))\), one normally sets the derivative to 0 \((\nabla f(\theta) = 0)\) and then solves. When \(f\) is not differentiable, can instead require that 0 is in the sub-gradient of \(\theta\).

### 13.1.4 Proof of Lagrangian Lasso Properties (Theorem 13.1)

To solve \(\frac{1}{2n}||y - X\theta||^2_2 + \lambda_n||\theta||_1\), let us use the sub-gradient: \(\frac{1}{n}X^T(X\theta - y) + \lambda_n z = 0\) (for some \(z \in \nabla ||\hat{\theta}||_1\)).

We need to show that for any \(\hat{\theta}\) that satisfies the above, \(\hat{\theta}_{S^c} = 0\) (i.e., no irrelevant coordinates are picked).

Will construct a \(\hat{\theta}, \hat{z}\) pair such that the conditions are already satisfied:

1. \(\hat{\theta}_{s^c} = 0\).
2. The first part of the stationary condition is satisfied.
3. \(\hat{z} \in \nabla ||\hat{\theta}||_1\) with high probability.

We use a constructive procedure, called the **primal-dual witness technique**. This creates a \(\hat{\theta}, \hat{z}\) pair which is primal-dual optimal and satisfies the required conditions. The construction procedure is as follows:

1. \(\hat{\theta}_{s^c} = 0\).
2. Set \(\hat{\theta}_s\) by solving: \(\hat{\theta}_s = \inf_{\theta_s} ||y - X_s\theta_s||^2_2 + \lambda_n ||\theta_s||_1\).
3. Choose \(\hat{z}_s \in \delta||\hat{\theta}_s||_1\) such that \(\frac{1}{n}X^T(X\hat{\theta}_s - y) + \lambda_n \hat{z}_s = 0\).
4. \(\frac{1}{n}X^T(X\hat{\theta} - y) + \lambda_n \hat{z}_s = 0\).

### References