6.883 Learning with Combinatorial Structure
Note for Lecture 11
Instructor: Prof. Stefanie Jegelka
Scribe: Xuhong Zhang

1 Regression with High Dimensional Data

Consider the following regression problem: given data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ generated from the model $y = \mathbf{x}^T\mathbf{w} + \epsilon$, where $\mathbf{x}_i \in \mathbb{R}^d$, $\forall i$, and ϵ denotes the noise. Our goal is to recover the unknown signal/function $\mathbf{w}$:

$$\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Here $\mathbf{y} = (y_1, y_2, ..., y_N)^T$, and $\mathbf{X}$ is the measurement matrix whose i$^{\text{th}}$ row is $\mathbf{x}_i^T$. Many

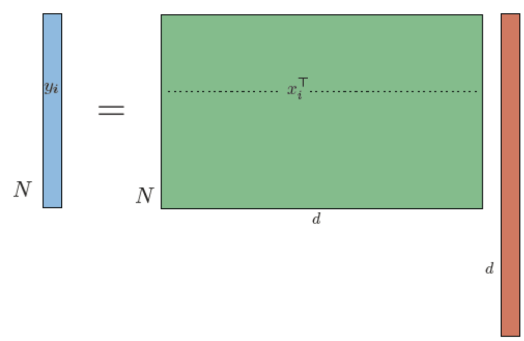

Figure 1: Schematic of regression with high dimensional data

applications of interest deal with high-dimensional data, i.e. $d \gg N$. For example, if the input is an image, d can be the number of pixels in the image. In such cases, the problem is underdetermined: there are many solutions to $\arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ and thus recovery of $\mathbf{w}$ is generally impossible.

However, we can hope to recover $\mathbf{w}$ if $\mathbf{w}$ has some low-dimensional structures. In this lecture, we assume $\mathbf{w}$ is k-sparse: $|S| = |\text{supp}(\mathbf{w})| \le k$. There are some additional motivations for focusing on sparse $\mathbf{w}$: sparsity in $\mathbf{w}$ helps to improve computational efficiency, as well as making the solution more interpretable.

2 Intuitive Arguments

Let us consider the noiseless case first, i.e. $\mathbf{y} = \mathbf{X}\mathbf{w}$. We formulate the regression problem as an optimization problem

$$\mathbf{w}^* = \arg\min_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_0 \qquad (*)$$

Minimizing ℓ_0-norm is in general NP-hard. Instead, consider the relaxation to ℓ_1-norm:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_1 \qquad (**)$$

$(**)$ is a convex optimization problem and we can hope to solve it. A natural question is, when does this relaxation work (i.e. $\hat{\mathbf{w}}$ is close to $\mathbf{w}^*$ in some sense)? We start with some intuitive arguments and provide a more formal analysis in Section 3.

2.1 Restricted Nullspace Condition

Nullspace of $\mathbf{X}$ can be large. But as long as the nullspace does not contain directions in which the ℓ_1-norm decreases, we can still hope to recover $\mathbf{w}^*$ by minimizing ℓ_1-norm in the nullspace. Let us denote $\boldsymbol{\nu} = \hat{\mathbf{w}} - \mathbf{w}^*$, this intuition is more precisely described by the *restricted nullspace condition* below:

$$\{\boldsymbol{\nu} \in \mathbb{R}^d : \mathbf{X}\boldsymbol{\nu} = \mathbf{0}\} \cap \{\boldsymbol{\nu} : \|\hat{\mathbf{w}}\|_1 \le \|\mathbf{w}^*\|_1\} = \{\mathbf{0}\}$$

The set of directions in which ℓ_1-norm decreases is referred to as the *cone of descent directions*: $\mathcal{C} \triangleq \{\boldsymbol{\nu} : \|\hat{\mathbf{w}}\|_1 \le \|\mathbf{w}^*\|_1\}$

Moreover, as illustrated in Figure 2, the nullspace is likely to intersect with the ℓ_1-norm ball at the axis (thus results in sparse solutions). In contrast, intersections with ℓ_2-norm balls are likely to be non-sparse.

2.2 Curvature

Now consider the general case where noise is present. Suppose that $\mathbf{w}^*$ is the true optimal solution and we estimate $\mathbf{w}$ by minimizing a data-dependent objective function $L_N(\mathbf{w}) = \frac{1}{2N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$ over some constrained set D, i.e. $\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in D} L_N(\mathbf{w})$.

As $N \to \infty$, we do expect $|L_N(\hat{\mathbf{w}}) - L_N(\mathbf{w}^*)| \to 0$. The question is, what additional conditions are needed to ensure that the ℓ_2-norm also vanishes (i.e. $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \to 0$)? As illustrated in Figure 3, it is important to have a sufficiently large curvature. A natural way to specify that a function is suitably "curved" is via the notion of *strong convexity*. More

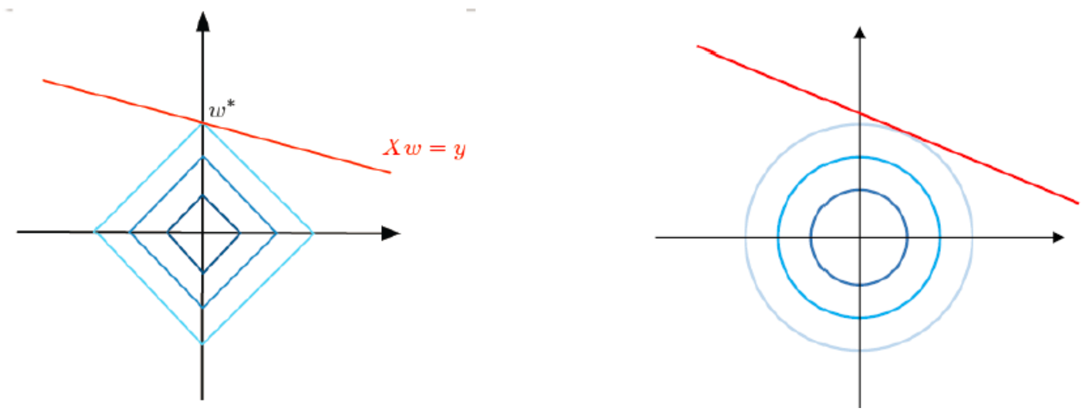

Figure 2: Nullspace of $\mathbf{X}$ intersecting ℓ_1-norm ball and ℓ_2-norm ball

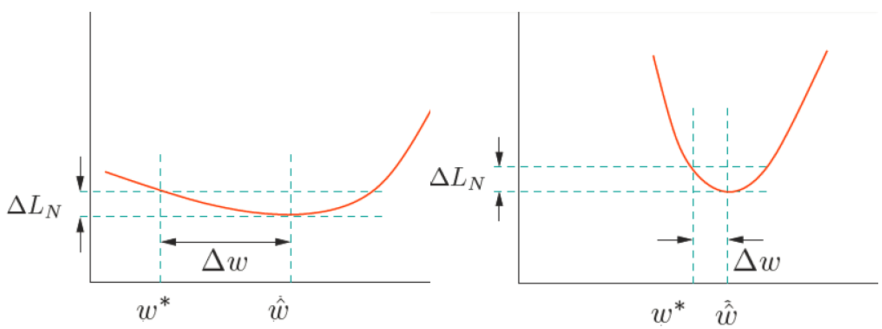

Figure 3: Difference in objective function vs difference in parameter values

specifically, assume $L_N(\cdot)$ id differentiable, it is γ-strongly convex if $\forall \mathbf{w}, \mathbf{w}'$, the following equation holds:

$$L_N(\mathbf{w}') - L_N(\mathbf{w}) \geq \nabla L_N(\mathbf{w})^T(\mathbf{w}' - \mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2$$

If $L_N(\cdot)$ is twice differentiable, this is equivalent to $\lambda_{\min}(\nabla^2 L_N(\mathbf{w})) \geq \gamma$, where $\lambda_{\min}(\nabla^2 L_N(\mathbf{w}))$ denotes the smallest eigenvalue of $\nabla^2 L_N(\mathbf{w})$.

However, notice $\nabla^2 L_N(\mathbf{w}) = \mathbf{X}^T\mathbf{X}/N \in \mathbb{R}^{d\times d}$ and with high dimensional data ($d \gg N$), rank($\mathbf{X}^T\mathbf{X}$) $\leq N < d$. Thus $\lambda_{\min}(\nabla^2 L_N(\mathbf{w})) = 0$! In other words, $L_N(\cdot)$ is not strongly convex. It turns out that we can relax the notion of strong convexity: we only need $L_N(\cdot)$ to have sufficient curvature in the cone of descent directions, as demonstrated in Figure

4. In particular, we require $\exists \gamma > 0$, s.t.

$$\frac{\|\mathbf{X}\boldsymbol{\nu}\|_2^2}{N\|\boldsymbol{\nu}\|_2^2} \geq \gamma, \quad \forall \boldsymbol{\nu} \in \mathcal{C}$$

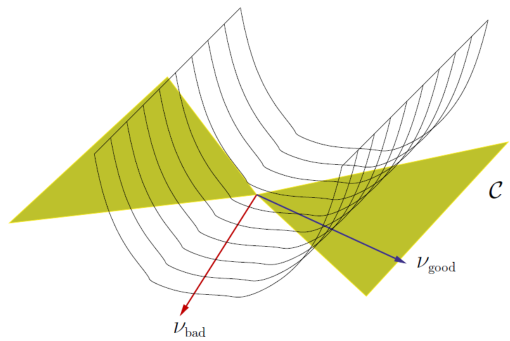

Figure 4: Difference in objective function vs difference in parameter values

3 Formal Analysis

3.1 Formulation

To solve the regression problem with sparsity constraint on $\mathbf{w}$, we want to solve

$$\mathbf{w}^* = \arg\min_{\mathbf{w}:\|\mathbf{w}\|_0 \leq R} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \qquad (\dagger)$$

This is in general NP-hard, and we consider its relaxation to ℓ_1-norm instead:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}:\|\mathbf{w}\|_1 \leq R} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \qquad (\dagger\dagger)$$

($\dagger\dagger$) is referred to as the constrained form of the Lasso problem. It can be equivalently written as its regularized version (by appropriate choice of parameters R and λ):

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1 \qquad (\dagger\dagger\dagger)$$

As discussed before, we want to establish that $\hat{\mathbf{w}}$ is close to $\mathbf{w}^*$ in some sense. In particular, we will look at the following measures of 'error':

(1) ℓ_2 error: $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$;

(2) prediction error: $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{w}^*\|_2$;

(3) support recovery: whether or not supp($\hat{\mathbf{w}}$) = supp($\mathbf{w}^*$).

3.2 Bounding ℓ_2-Error and Prediction Error

Analysis in this subsection assumes problem (††) since it is easier to analyse.

Theorem 1. *Let $S = supp(\boldsymbol{w}^*)$ and $|S| = k$. Assume*

(1) $\boldsymbol{X}$ satisfies restricted eigenvalue bound: $\forall \boldsymbol{\nu} \in \mathcal{C}$, $\frac{\|\boldsymbol{X\nu}\|_2^2}{N\|\boldsymbol{\nu}\|_2^2} \geq \gamma > 0$

(2) $\|\boldsymbol{w}^*\|_1 = R$.

Then we have

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 \leq \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{\boldsymbol{X}^T\boldsymbol{\epsilon}}{\sqrt{N}}\right\|_\infty$$

Proof. Due to the optimality of $\hat{\mathbf{w}}$, we have $\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 \leq \|\mathbf{y} - \mathbf{X}\mathbf{w}^*\|_2^2 = \|\boldsymbol{\epsilon}\|_2^2$. Thus

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_2^2 \geq \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 &= \|\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{w}^* + \boldsymbol{\nu})\|_2^2 = \|\mathbf{X}\boldsymbol{\nu} - \boldsymbol{\epsilon}\|_2^2 \\
&= \|\boldsymbol{\epsilon}\|_2^2 - 2\boldsymbol{\epsilon}^T\mathbf{X}\boldsymbol{\nu} + \|\mathbf{X}\boldsymbol{\nu}\|_2^2 \\
\therefore \frac{\|\mathbf{X}\boldsymbol{\nu}\|_2^2}{N} \leq \frac{2\boldsymbol{\epsilon}^T\mathbf{X}\boldsymbol{\nu}}{N} &= \frac{2(\mathbf{X}^T\boldsymbol{\epsilon})^T\boldsymbol{\nu}}{N} \leq \frac{2}{N}\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\nu}\|_1
\end{aligned}$$

The last inequality above follows from Hölder's inequality.

Let $S = \text{supp}(\mathbf{w}^*)$ (i.e. $\mathbf{w}_S^* \neq \mathbf{0}$ and $\mathbf{w}_{S^C}^* = \mathbf{0}$). Then $\forall \boldsymbol{\nu} \in \mathcal{C}$,

$$\|\mathbf{w}^*\|_1 = \|\mathbf{w}_S^*\|_1 \geq \|\mathbf{w}^* + \boldsymbol{\nu}\|_1 = \|\mathbf{w}_S^* + \boldsymbol{\nu}_S\|_1 + \|\boldsymbol{\nu}_{S^C}\|_1 \geq \|\mathbf{w}_S^*\|_1 - \|\boldsymbol{\nu}_S\|_1 + \|\boldsymbol{\nu}_{S^C}\|_1$$

Thus $\forall \boldsymbol{\nu} \in \mathcal{C}$, $\|\boldsymbol{\nu}_S\|_1 \geq \|\boldsymbol{\nu}_{S^C}\|_1$

$\therefore \|\boldsymbol{\nu}\|_1 = \|\boldsymbol{\nu}_S\|_1 + \|\boldsymbol{\nu}_{S^C}\|_1 \leq 2\|\boldsymbol{\nu}_S\|_1 \leq 2\sqrt{k}\|\boldsymbol{\nu}\|_2$

$\therefore \frac{\|\mathbf{X}\boldsymbol{\nu}\|_2^2}{N} \leq \frac{2}{N}\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\nu}\|_1 \leq \frac{4\sqrt{k}}{N}\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\nu}\|_2$

From assumption (1), we have $\|\mathbf{X}\boldsymbol{\nu}\|_2^2/N \geq \gamma\|\boldsymbol{\nu}\|_2^2$.

$$\therefore \|\boldsymbol{\nu}\|_2 = \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq \frac{4\sqrt{k}}{\gamma N}\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty = \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{\mathbf{X}^T\boldsymbol{\epsilon}}{\sqrt{N}}\right\|_\infty$$

□

Following similar arguments, one can derive bound for $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$ in the regularized version (optimization problem (†††)) or bound for the prediction error $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{w}^*\|_2$. For more details, please refer to Theorem 11.2 in [1].

3.3 Example: Classical Linear Gaussian Model

In classical linear Gaussian model, the observation noise $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is a vector with i.i.d Gaussian entries, i.e. $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, $\forall j \in \{1, 2, ...N\}$. We will view the measurement matrix $\mathbf{X}$ as fixed and normalized ($\|\tilde{\mathbf{x}_j}\|_2/\sqrt{N} = 1$, $\forall j$, $\tilde{\mathbf{x}_j}$ here denotes the j^{th} column of $\mathbf{X}$). Then $\tilde{\mathbf{x}_j}^T \boldsymbol{\epsilon}/N$ is also a Gaussian random variable with mean 0 and variance $\frac{\sigma^2}{N}\frac{\|\tilde{\mathbf{x}_j}\|_2^2}{N} = \sigma^2/N$. From the Gaussian tail bound, we have

$$\mathbb{P}\Big(\frac{|\tilde{\mathbf{x}_j}^T \boldsymbol{\epsilon}|}{N} \geq t\Big) \leq 2\exp\big(-\frac{Nt^2}{2\sigma^2}\big)$$

Apply union bound,

$$\mathbb{P}\Big(\frac{\|\mathbf{X}^T \boldsymbol{\epsilon}\|_\infty}{N} \geq t\Big) = \mathbb{P}\Big(\max_j \frac{|\tilde{\mathbf{x}_j}^T \boldsymbol{\epsilon}|}{N} \geq t\Big) \leq 2d\exp\big(-\frac{Nt^2}{2\sigma^2}\big)$$

If we set $t = \sigma\sqrt{\frac{\tau \log d}{N}}$ for some constant $\tau > 2$, we have

$$\frac{\|\mathbf{X}^T \boldsymbol{\epsilon}\|_\infty}{N} \leq \sigma\sqrt{\frac{\tau \log d}{N}} \quad \text{with probability } 1 - 2\exp(-\frac{1}{2}(\tau - 2)\log d)$$

Plug into the bound in Theorem 1, we obtain

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq \frac{4\sigma}{\gamma}\sqrt{\frac{\tau k \log d}{N}} \quad \text{with probability } 1 - 2\exp(-\frac{1}{2}(\tau - 2)\log d)$$

3.4 Recovery of Support

Thus far we have discussed bounds on the ℓ_2-error ($\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$) or the prediction error ($\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{w}^*\|_2$). In this subsetion, we consider a somewhat more refined question: how well does $\hat{\mathbf{w}}$ recover the support of $\mathbf{w}^*$?

Theorem 2. *If the following assumptions hold:*

(1) *(Mutual incoherence[1]) Let S be the support set of $\boldsymbol{w}^*$ and $\mathbf{X}_S \in \mathbb{R}^{N \times k}$ be the columns of $\boldsymbol{X}$ corresponding to S. $\tilde{\boldsymbol{x}}_j$ is j^{th} column in $\mathbf{X}_{S^C}$. There exists some $\gamma > 0$ s.t.*

$$\max_{j \in S^c} \|(\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^T \boldsymbol{x}_j\|_1 \leq 1 - \gamma$$

(2) *(Bounded columns) $\forall j \in \{1, 2, ..., d\}$, $\|\boldsymbol{x}_j\|_2/\sqrt{N} \leq K$, where K is a constant;*

[1] This assumption essentially requires that the columns in $\mathbf{X}_{S^C}$ cannot be well represented as linear combinations of columns in $\mathbf{X}_S$

(3) ($\boldsymbol{X}_S$ 'well-behaved' and invertible) $\lambda_{\min}(\boldsymbol{X}_S^T\boldsymbol{X}_S/N) \geq C$, where C is a positive constant.

Under the three assumptions above, with iid Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\lambda \geq \frac{8K\sigma}{\gamma}\sqrt{\frac{\log(d)}{N}}$ and

$$\hat{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w}}\{\frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1\}$$

Then with probability $1 - c_1 e^{-c2N\lambda^2}$:

- *(a) (Uniqueness) solution $\hat{\boldsymbol{w}}$ is unique;*
- *(b) (No false inclusion) supp($\hat{\boldsymbol{w}}$) $\subseteq$ supp($\boldsymbol{w}^*$)*
- *(c) (Bounds) $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_\infty \leq \lambda[\frac{4\sigma}{\sqrt{C}} + \|(\frac{\boldsymbol{X}_S^T\boldsymbol{X}_S}{N})^{-1}\|_\infty] \triangleq B(\lambda, \sigma; \boldsymbol{X})$*
- *(d) (No false exclusion) if $\forall j$, $|\boldsymbol{w}_j^*| > B(\lambda, \sigma; \boldsymbol{X})$, then supp($\hat{\boldsymbol{w}}$) $\supseteq$ supp($\boldsymbol{w}^*$)*

Before proving Theorem 2, let us interpret the theorem:

- The uniqueness result in (a) allows us to talk about supp($\hat{\mathbf{w}}$) unambiguously;
- (b) guarantees that $\hat{\mathbf{w}}$ does not include non-zero entries outside the support of $\mathbf{w}^*$;
- Result in (c) guarantees that $\hat{\mathbf{w}}$ is uniformly close to $\mathbf{w}^*$ in the ℓ_∞-norm;
- (d) states that as long as the non-zero entries of $\mathbf{w}^*$ are reasonably far away from 0, the support of $\hat{\mathbf{w}}$ actually agrees with the support of $\mathbf{w}^*$.

Proof. 1°. The proof of Theorem 2 is based on a constructive procedure, known as a *primal-dual witness method* (PDW). We construct a pair $(\hat{\mathbf{w}}, \hat{\mathbf{z}}) \in \mathbb{R}^d \times \mathbb{R}^d$ that are primal-dual optimal for

$$\min_{\mathbf{w}}\left\{\frac{1}{2N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1\right\} = \min_{\mathbf{w}}\max_{\mathbf{z}\in[-1,1]^d}\left\{\frac{1}{2N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\mathbf{z}^T\mathbf{w}\right\}$$

As before, we will denote the support of $\mathbf{w}^*$ as S. Let us describe the construction procedure as follows:

(i) Set $\hat{\mathbf{w}}_{S^C} = \mathbf{0}$;

(ii) Set $\hat{\mathbf{w}}_S = \arg\min_{\mathbf{w}_S}\left\{\frac{1}{2N}\|\mathbf{X}_S\mathbf{w}_S - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}_S\|_1\right\}$
Thus $\hat{\mathbf{z}}_S$ is an element of subdifferential $\partial\|\hat{\mathbf{w}}_S\|_1$ satisfying

$$\lambda\hat{\mathbf{z}}_S - \frac{1}{N}\mathbf{X}_S^T(\mathbf{y} - \mathbf{X}_S\hat{\mathbf{w}}_S) = \mathbf{0} \quad \text{and} \quad \hat{\mathbf{z}}_S = \text{sign}(\mathbf{w}_S^*)$$

(iii) Solve for $\hat{\mathbf{z}}_{S^C}$ using $\lambda\hat{\mathbf{z}} - \frac{1}{N}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = \mathbf{0}$. Check if the strict dual feasibility condition (i.e. $\|\hat{\mathbf{z}}_{S^C}\|_\infty < 1$) holds. If so, the constructive procedure succeeds.

It should be noticed that the above procedure is a proof technique and cannot be carried out in practice because we generally do not know the support of $\mathbf{w}^*$.

2°. **Claim:** if the constructive procedure succeeds and assumption (3) holds, then $(\hat{\mathbf{w}}_S, \mathbf{0})$ is the unique optimal solution of $\min_{\mathbf{w}} \left\{ \frac{1}{2N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1 \right\}$.

This claim is Lemma 11.2 in [1] and its proof can be found there. Next we prove $\|\hat{\mathbf{z}}_{S^C}\|_\infty \leq 1$ with high probability ($3^\circ - 5^\circ$).

3°. Notice $\hat{\mathbf{w}}_{S^C} = \mathbf{w}^*_{S^C} = \mathbf{0}$. Thus $\lambda\hat{\mathbf{z}} - \frac{1}{N}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = \lambda\hat{\mathbf{z}} - \frac{1}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon} - \mathbf{X}\hat{\mathbf{w}}) = \mathbf{0}$ can be written as

$$-\frac{1}{N}\begin{pmatrix} \mathbf{X}_S^T\mathbf{X}_S & \mathbf{X}_S^T\mathbf{X}_{S^C} \\ \mathbf{X}_{S^C}^T\mathbf{X}_S & \mathbf{X}_{S^C}^T\mathbf{X}_{S^C} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{w}}_S - \mathbf{w}_S^* \\ \mathbf{0} \end{pmatrix} + \frac{1}{N}\begin{pmatrix} \mathbf{X}_S^T\boldsymbol{\epsilon} \\ \mathbf{X}_{S^C}^T\boldsymbol{\epsilon} \end{pmatrix} - \lambda\begin{pmatrix} \hat{\mathbf{z}}_S \\ \hat{\mathbf{z}}_{S^C} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

$$\begin{aligned}
\Rightarrow \hat{\mathbf{z}}_{S^C} &= \frac{1}{\lambda}(\frac{1}{N}\mathbf{X}_{S^C}^T\boldsymbol{\epsilon} - \frac{1}{N}\mathbf{X}_{S^C}^T\mathbf{X}_S(\hat{\mathbf{w}}_S - \mathbf{w}_S^*)) \text{ and} \\
\hat{\mathbf{w}}_S - \mathbf{w}_S^* &= -\lambda(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\hat{\mathbf{z}}_S + (\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\frac{1}{N}\mathbf{X}_S^T\boldsymbol{\epsilon} \\
\therefore \hat{\mathbf{z}}_{S^C} &= \frac{1}{N}\mathbf{X}_{S^C}^T\mathbf{X}_S(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*) + \mathbf{X}_{S^C}^T(I - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T)\frac{\boldsymbol{\epsilon}}{\lambda N} \\
&\triangleq \boldsymbol{\mu} + V_{S^C} \\
\therefore \|\hat{\mathbf{z}}_{S^C}\|_\infty &\leq \|\boldsymbol{\mu}\|_\infty + \|V_{S^C}\|_\infty
\end{aligned}$$

4°. Let us bound the term $\|\boldsymbol{\mu}\|_\infty$. Notice $\boldsymbol{\mu}$ is deterministic. According to assumption 1 (mutual incoherence), $(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{X}_{S^C}$ is a matrix whose columns all have ℓ_1-norm upper bounded by $1-\gamma$. Thus $\frac{1}{N}\mathbf{X}_{S^C}^T\mathbf{X}_S(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1} = ((\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{X}_{S^C})^T$ is a matrix whose rows all have ℓ_1-norm upper bounded by $1-\gamma$.

$$\therefore \|\boldsymbol{\mu}\|_\infty = \|\frac{1}{N}\mathbf{X}_{S^C}^T\mathbf{X}_S(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*)\|_\infty \leq 1-\gamma$$

5°. Let us now bound $\|V_{S^C}\|_\infty$. Let V_j be the j^{th} element of V_{S^C}, let $\tilde{\mathbf{x}}_j$ denote the j^{th} column of X_{S^C}, then

$$V_j = \tilde{\mathbf{x}}_j^T(I - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T)\frac{\boldsymbol{\epsilon}}{\lambda N}$$

Notice that $I - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$ is an orthogonal projection matrix, and from assumption (2) (bounded columns), $\|\tilde{\mathbf{x}}_j\|_2 \leq K\sqrt{N}$. Therefore V_j is a Gaussian random variable with zero mean and variance upper bounded by $\sigma^2K^2/(N\lambda^2)$. From Gaussian tail bound and union bound, we obtain:

$$\mathbb{P}(\|V_{S^C}\|_\infty \geq \gamma) \leq 2(d-k)\exp(-\frac{\gamma^2N\lambda^2}{2\sigma^2K^2})$$

Combining $4° - 5°$, we have shown

$$\|\hat{\mathbf{z}}_{S^C}\|_\infty < 1 - \gamma + \gamma = 1 \quad \text{with probability } 1 - 2(d-k)\exp(-\frac{\gamma^2 N\lambda^2}{2\sigma^2 K^2})$$

Apply the claim in 2°, we know that $(\hat{\mathbf{w}}_S, \mathbf{0})$ is the unique optimal solution with high probability (result (a) proved). Result (b) follows directly from the construction of $\hat{\mathbf{w}}$. Next let us establish (c) and (d), i.e. bound the ℓ_∞-norm of $\hat{\mathbf{w}}_S - \mathbf{w}_S^*$.

6°. From previous discussion, we have

$$\begin{aligned}
\hat{\mathbf{w}}_S - \mathbf{w}_S^* &= -\lambda(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*) + (\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\frac{1}{N}\mathbf{X}_S^T\boldsymbol{\epsilon} \\
\therefore \|\hat{\mathbf{w}}_S - \mathbf{w}_S^*\|_\infty &\le \lambda\|(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*)\|_\infty + \|(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\frac{1}{N}\mathbf{X}_S^T\boldsymbol{\epsilon}\|_\infty \\
&\le \lambda(\|(\frac{1}{N}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\|_\infty + \frac{4\sigma}{\sqrt{C}}) \quad \text{with probability } 1 - 2\exp(-c_2\lambda^2 N)
\end{aligned}$$

The last inequality follows from similar arguments as those in $5°$ and uses assumption 3 ($\mathbf{X}_S$ 'well-behaved')[2]. So we have proved (c) and it is not difficult to see that result (d) is a direct consequence of (c).

□

4 Other Sparsity Patterns

So far we have considered $\mathbf{w}$ being $k-$sparse, i.e. $\|\mathbf{w}\|_0 \le k$. Other types of sparsity patterns/low dimensional structures may be useful too. Examples include

(1) we may prefer solutions that are not only sparse, but also have subsequent nonzeros (i.e. non-zero entries are grouped together);

(2) the signals of interest may correspond to a tree graph and we prefer solutions where the nonzero entries form sub-trees.

Next lecture, we will see how to generalize the formulation we have in today's lecture to accommodate more general sparsity patterns.

REFERENCES

1 . "*Theoretical Results for the Lasso*" by M. Wainwright

2 . "*Learning with Submodular Functions - A Convex Optimization Perspective*" by F.Bach

[2]For more details on the constant c_2, please refer to [1].