1 Regression with High Dimensional Data

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

\[
\hat{w} \in \arg \min_{w \in \mathbb{R}^d} \|y - Xw\|_2^2
\]

Here \(y = (y_1, y_2, ..., y_N)^T\), and \(X\) is the measurement matrix whose \(i\)th row is \(x_i^T\). Many 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 \(\arg \min_{w \in \mathbb{R}^d} \|y - Xw\|_2^2\) and thus recovery of \(w\) is generally impossible.

However, we can hope to recover \(w\) if \(w\) has some low-dimensional structures. In this lecture, we assume \(w\) is \(k\)-sparse: \(|S| = |\text{supp}(w)| \leq k\). There are some additional motivations for focusing on sparse \(w\): sparsity in \(w\) 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. $y = Xw$. We formulate the regression problem as an optimization problem

$$w^* = \arg\min_{w : Xw = y} \|w\|_0 \quad (*)$$

Minimizing $\ell_0$-norm is in general NP-hard. Instead, consider the relaxation to $\ell_1$-norm:

$$\hat{w} = \arg\min_{w : Xw = y} \|w\|_1 \quad (***)$$

(***) is a convex optimization problem and we can hope to solve it. A natural question is, when does this relaxation work (i.e. $\hat{w}$ is close to $w^*$ 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 $X$ can be large. But as long as the nullspace does not contain directions in which the $\ell_1$-norm decreases, we can still hope to recover $w^*$ by minimizing $\ell_1$-norm in the nullspace. Let us denote $\nu = \hat{w} - w^*$, this intuition is more precisely described by the restricted nullspace condition below:

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

The set of directions in which $\ell_1$-norm decreases is referred to as the cone of descent directions: $\mathcal{C} \triangleq \{\nu : \|\hat{w}\|_1 \leq \|w^*\|_1\}$

Moreover, as illustrated in Figure 2, the nullspace is likely to intersect with the $\ell_1$-norm ball at the axis (thus results in sparse solutions). In contrast, intersections with $\ell_2$-norm balls are likely to be non-sparse.

2.2 Curvature

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

As $N \to \infty$, we do expect $|L_N(\hat{w}) - L_N(w^*)| \to 0$. The question is, what additional conditions are needed to ensure that the $\ell_2$-norm also vanishes (i.e. $\|\hat{w} - w^*\|_2 \to 0$)? 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 $X$ intersecting $\ell_1$-norm ball and $\ell_2$-norm ball

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

Specifically, assume $L_N(\cdot)$ id differentiable, it is $\gamma$-strongly convex if $\forall w, w'$, the following equation holds:

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

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

However, notice $\nabla^2 L_N(w) = X^T X / N \in \mathbb{R}^{d \times d}$ and with high dimensional data ($d \gg N$), $\text{rank}(X^T X) \leq N < d$. Thus $\lambda_{\min}(\nabla^2 L_N(w)) = 0$! 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.
\[
\frac{\|X\nu\|_2^2}{N\|\nu\|_2^2} \geq \gamma, \quad \forall \nu \in \mathcal{C}
\]

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 \( w \), we want to solve
\[
\hat{w} = \arg \min_w \|w\|_0 \leq R \|Xw - y\|^2
\]
\( (\dagger) \)

This is in general NP-hard, and we consider its relaxation to \( \ell_1 \)-norm instead:
\[
\hat{w} = \arg \min_{w: \|w\|_1 \leq R} \|Xw - y\|^2
\]
\( (\dagger\dagger) \)

(\(\dagger\dagger\)) 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 \( \lambda \)):
\[
\hat{w} = \arg \min_w \|Xw - y\|^2 + \lambda\|w\|_1
\]
\( (\dagger\dagger\dagger) \)

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

1. \( \ell_2 \) error: \( \|\hat{w} - w^*\|_2 \);
2. prediction error: \( \|X\hat{w} - Xw^*\|_2 \);
3. support recovery: whether or not \( \text{supp}(\hat{w}) = \text{supp}(w^*) \).
3.2 Bounding $\ell_2$-Error and Prediction Error

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

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

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

2. $\|w^*\|_1 = R$.

Then we have

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

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

$$\|\epsilon\|_2^2 \geq \|y - X\hat{w}\|_2^2 = \|Xw^* + \epsilon - X(w^* + \nu)\|_2^2 = \|X\nu - \epsilon\|_2^2$$

$$= \|\epsilon\|_2^2 - 2\epsilon^T X\nu + \|X\nu\|_2^2$$

$$\therefore \|X\nu\|_2^2 \leq \frac{2\epsilon^T X\nu}{N} = \frac{2(X^T \epsilon)^T \nu}{N} \leq \frac{2}{N} \|X^T \epsilon\|_\infty \|\nu\|_1$$

The last inequality above follows from Hölder’s inequality.

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

$$\|w^*\|_1 = \|w^*_S\|_1 + \|w^* + \nu\|_1 = \|w^*_S + \nu_S\|_1 + \|\nu_S\|_1 \geq \|w^*_S\|_1 - \|\nu_S\|_1 + \|\nu_S\|_1$$

Thus $\forall \nu \in \mathcal{C}, \|\nu_S\|_1 \geq \|\nu_S\|_1$

$$\therefore \|\nu\|_1 = \|\nu_S\|_1 + \|\nu_S\|_1 \leq 2\|\nu_S\|_1 \leq 2\sqrt{k}\|\nu\|_2$$

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

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

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

Following similar arguments, one can derive bound for $\|\hat{w} - w^*\|_2$ in the regularized version (optimization problem (†††)) or bound for the prediction error $\|X\hat{w} - Xw^*\|_2$. 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 $\epsilon \in \mathbb{R}^N$ is a vector with i.i.d Gaussian entries, i.e. $\epsilon_j \sim N(0, \sigma^2)$, $\forall j \in \{1, 2, \ldots, N\}$. We will view the measurement matrix $X$ as fixed and normalized ($\|\tilde{x}_j\|_2 / \sqrt{N} = 1$, $\forall j$, $\tilde{x}_j$ here denotes the $j$th column of $X$). Then $x_j^T \epsilon / N$ is also a Gaussian random variable with mean 0 and variance $\sigma^2 \|x_j\|^2 / N = \sigma^2 / N$. From the Gaussian tail bound, we have

$$P\left( \frac{|x_j^T \epsilon|}{N} \geq t \right) \leq 2 \exp \left( -\frac{Nt^2}{2\sigma^2} \right)$$

Apply union bound,

$$P\left( \left\| X^T \epsilon \right\|_\infty / N \geq t \right) = P\left( \max_j \left| \tilde{x}_j^T \epsilon \right| / N \geq t \right) \leq 2d \exp \left( -\frac{Nt^2}{2\sigma^2} \right)$$

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

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

Plug into the bound in Theorem 1, we obtain

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

3.4 Recovery of Support

Thus far we have discussed bounds on the $\ell_2$-error ($\| \hat{w} - w^* \|_2$) or the prediction error ($\|X \hat{w} - X w^* \|_2$). In this subsection, we consider a somewhat more refined question: how well does $\hat{w}$ recover the support of $w^*$?

**Theorem 2.** If the following assumptions hold:

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

$$\max_{j \in Sc} \left\| (X_S^T X_S)^{-1} X_S^T x_j \right\|_1 \leq 1 - \gamma$$

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

---

$^1$This assumption essentially requires that the columns in $X_{Sc}$ cannot be well represented as linear combinations of columns in $X_S$
(3) \((X_S \text{‘well-behaved’ and invertible}) \lambda_{\text{min}}(X^T_S X_S/N) \geq C, \text{ where } C \text{ is a positive constant.}\)

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

\[
\hat{w} \in \arg \min_w \left\{ \frac{1}{2N} \|y - Xw\|_2^2 + \lambda \|w\|_1 \right\}
\]

Then with probability \(1 - c_1 e^{-c_2 N \lambda^2}\):

(a) (Uniqueness) solution \(\hat{w}\) is unique;

(b) (No false inclusion) \(\text{supp}(\hat{w}) \subseteq \text{supp}(w*)\)

(c) (Bounds) \(\|\hat{w} - w*\|_{\infty} \leq \lambda \left[ \frac{4\sigma}{\sqrt{C}} + \|((X^T_S X_S)^{-1})_{\infty}\right] \triangleq B(\lambda, \sigma; X)\)

(d) (No false exclusion) if \(\forall j, |w^*_j| > B(\lambda, \sigma; X)\), then \(\text{supp}(\hat{w}) \supseteq \text{supp}(w*)\)

Before proving Theorem 2, let us interpret the theorem:

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

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 \((\hat{w}, \hat{z}) \in \mathbb{R}^d \times \mathbb{R}^d\) that are primal-dual optimal for

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

As before, we will denote the support of \(w^*\) as \(S\). Let us describe the construction procedure as follows:

(i) Set \(\hat{w}_{SC} = 0\);

(ii) Set \(\hat{w}_S = \arg \min_{w_S} \left\{ \frac{1}{2N} \|X_S w_S - y\|_2^2 + \lambda \|w_S\|_1 \right\}\)

Thus \(\hat{z}_S\) is an element of subdifferential \(\partial\|\hat{w}_S\|_1\) satisfying

\[
\lambda \hat{z}_S - \frac{1}{N} X^T_S (y - X_S \hat{w}_S) = 0 \quad \text{and} \quad \hat{z}_S = \text{sign}(w^*_S)
\]

(iii) Solve for \(\hat{z}_{SC}\) using \(\lambda \hat{z} - \frac{1}{N} X^T (y - Xw) = 0\). Check if the strict dual feasibility condition (i.e. \(\|\hat{z}_{SC}\|_{\infty} < 1\)) 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 \( w^* \).

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

This claim is Lemma 11.2 in [1] and its proof can be found there. Next we prove \( \|\hat{z}_{SC}\|_\infty \leq 1 \) with high probability (3° – 5°).

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

\[
-\frac{1}{N} \begin{pmatrix} X_T^{\top}X_S & X_T^{\top}X_{SC} \\ X_S^{\top}X_S & X_S^{\top}X_{SC} \end{pmatrix} \begin{pmatrix} \hat{w}_S - w_S^* \\ 0 \end{pmatrix} + \frac{1}{N} \begin{pmatrix} X_S^{\top}\epsilon \\ X_{SC}^{\top}\epsilon \end{pmatrix} - \lambda \begin{pmatrix} \hat{z}_S \\ \hat{z}_{SC} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\]

\[
\Rightarrow \hat{z}_{SC} = \frac{1}{\lambda} \left( \frac{1}{N}X_{SC}^{\top}\epsilon - \frac{1}{N}X_S^{\top}X_S(\hat{w}_S - w_S^*) \right) \text{ and}
\]

\[
\hat{w}_S - w_S^* = -\lambda \left( \frac{1}{N}X_S^{\top}X_S \right)^{-1} \hat{z}_S + \left( \frac{1}{N}X_S^{\top}X_S \right)^{-1} \frac{1}{N}X_S^{\top}\epsilon
\]

\[
\therefore \hat{z}_{SC} = \frac{1}{N}X_{SC}^{\top}X_S \left( \frac{1}{N}X_S^{\top}X_S \right)^{-1} \text{sign}(w_S^*) + X_{SC}^{\top}(I - X_S(X_S^{\top}X_S)^{-1}X_S^{\top}) \frac{\epsilon}{\lambda N}
\]

\[
\triangleq \mu + V_{SC}
\]

\[
\therefore \|\hat{z}_{SC}\|_\infty \leq \|\mu\|_\infty + \|V_{SC}\|_\infty
\]

4°. Let us bound the term \( \|\mu\|_\infty \). Notice \( \mu \) is deterministic. According to assumption 1 (mutual incoherence), \((X_S^{\top}X_S)^{-1}X_S^{\top}X_{SC} \) is a matrix whose columns all have \( \ell_1 \)-norm upper bounded by \( 1 - \gamma \). Thus \( \frac{1}{N}X_{SC}^{\top}X_S \left( \frac{1}{N}X_S^{\top}X_S \right)^{-1} = \left( (X_S^{\top}X_S)^{-1}X_S^{\top}X_{SC} \right)^\top \) is a matrix whose rows all have \( \ell_1 \)-norm upper bounded by \( 1 - \gamma \).

\[
\therefore \|\mu\|_\infty = \|\frac{1}{N}X_{SC}^{\top}X_S \left( \frac{1}{N}X_S^{\top}X_S \right)^{-1} \text{sign}(w_S^*)\|_\infty \leq 1 - \gamma
\]

5°. Let us now bound \( \|V_{SC}\|_\infty \). Let \( V_j \) be the \( j^{\text{th}} \) element of \( V_{SC} \), let \( \tilde{x}_j \) denote the \( j^{\text{th}} \) column of \( X_{SC} \), then

\[
V_j = \tilde{x}_j^{\top}(I - X_S(X_S^{\top}X_S)^{-1}X_S^{\top}) \frac{\epsilon}{\lambda N}
\]

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

\[
\mathbb{P}(\|V_{SC}\|_\infty \geq \gamma) \leq 2(d - k) \exp\left(-\frac{\gamma^2 N\lambda^2}{2\sigma^2 K^2}\right)
\]
Combining $4^\circ - 5^\circ$, we have shown
\[ \|\hat{z}_S^c\|_\infty < 1 - \gamma + \gamma = 1 \quad \text{with probability} \ 1 - 2(d - k) \exp(-\frac{\gamma^2 N \lambda^2}{2\sigma^2 K^2}) \]

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

$6^\circ$. From previous discussion, we have
\[
\hat{w}_S - w_S^* = -\lambda \left( \frac{1}{N} X_S^T X_S \right)^{-1} \text{sign}(w_S^*) + \left( \frac{1}{N} X_S^T X_S \right)^{-1} \frac{1}{N} X_S^T \epsilon
\]

\[ \therefore \|\hat{w}_S - w_S^*\|_\infty \leq \lambda \left\| \left( \frac{1}{N} X_S^T X_S \right)^{-1} \text{sign}(w_S^*) \right\|_\infty + \left\| \left( \frac{1}{N} X_S^T X_S \right)^{-1} \frac{1}{N} X_S^T \epsilon \right\|_\infty \]
\[ \leq \lambda \left( \left\| \left( \frac{1}{N} X_S^T X_S \right)^{-1} \right\|_\infty + \frac{4\sigma}{\sqrt{C}} \right) \quad \text{with probability} \ 1 - 2 \exp(-c_2 \lambda^2 N) \]

The last inequality follows from similar arguments as those in $5^\circ$ and uses assumption 3 ($X_S$ ‘well-behaved’)$^2$. So we have proved (c) and it is not difficult to see that result (d) is a direct consequence of (c).

\[ \square \]

4 Other Sparsity Patterns

So far we have considered $w$ being $k$-sparse, i.e. $\|w\|_0 \leq k$. 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].