[27/10/2020] Department of Mathematics and Statistics
Lancaster University

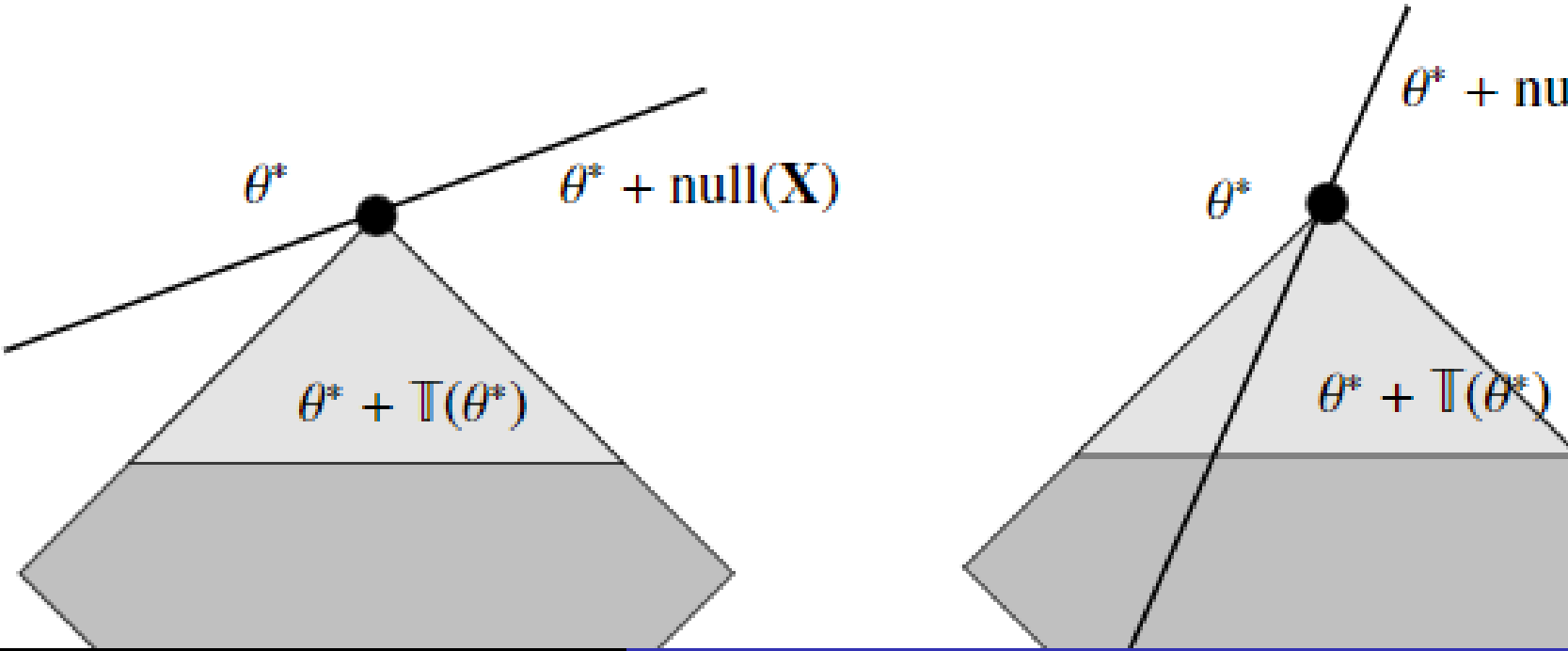
now let us consider the tangent cone of the l_1-ball at θ^*, given by

$$\mathbb{T}(\theta^*) = \Delta \in \mathbb{R}^d \mid \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \quad \text{for some} \quad t > 0$$

The set $\mathbb{T}(\theta^*)$ captures the set of all directions relative to θ^*along which the l_1-norm remains constant decreases.

Proof

Figure:



$$\begin{aligned}
\|\theta^*\|_1 &\geq \|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\
\|\theta^*\|_1 - \|\theta^*_S\|_1 &\geq - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\
0 &\geq - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\
\|\hat{\Delta}_S\|_1 &\geq \|\hat{\Delta}_{S^c}\|_1
\end{aligned}$$

which proves that $\hat{\Delta} \in \mathbb{C}(S)$. However, by construction, $X\hat{\Delta} = 0$, which means that $\hat{\Delta} \in \text{null}(X)$ too. By the assumption imposed earlier, this implies that $\hat{\Delta} = 0$ or that $\hat{\theta} = \theta^*$.

Sufficient conditions for restricted nullspace

In order to ensure that for any vector $\mathbb{R}^d$with support S, the basis pursuit programme applied with $y = X\theta^*$has unique solution $\hat{\theta} = \theta^*$, the matrix X has to satisfy the restricted nullspace property. The earliest sufficient conditions were based on the incoherence parameter of the design matrix, which is the quantity

$$\delta_{pw}(X) = \max_{j,k=1,\cdots,d} |\frac{\langle X_j, X_k \rangle}{n} - 1j = k|$$

where $X_j$and $X_k$are the kth and jth columns of the matrix X respectively and 1. denotes an indicator function. If X is rescaled by dividing by $\sqrt{n}$, then $X_j' X_j = 1$, which makes is more readily interpretable. The parameter $\delta_{pw}(X)$ essentially defines the maximum absolute value of cross-correlations between the columns of X.

In what follows, through Exercise 7.3 of Wainright (2019), it will be shown that a small mutual (pairwise) incoherence is sufficient to guarantee a uniform version of the restricted nullspace property.4

Proposition:

If the pairwise incoherence satisfies the bound

$$\delta_{pw}(X) \leq \frac{1}{3s}$$

then the restricted nullspace property holds for all subsets S of cardinality at most s.

Proof:

Choose a vector θ such that $X\theta = 0$. For some set S subject to $|S| \leq s$, we have $\theta = \theta_S + \theta_{S^c}$,and $X(\theta_S + \theta_{S^c}) = 0$. Thus, $X\theta_S = -X\theta_{S^c}$. Let us lower bound the $l_2$norm of the left hand side of the former equation

$$\begin{aligned}\|\frac{X\theta_S}{\sqrt{n}}\|_2^2 =& \frac{(X\theta_S)'X\theta_S}{n} = \frac{\theta_S'X'X\theta_S}{n} \\ =& \frac{\theta_S'X'X\theta_S}{n} - \theta_S'\theta_S + \theta_S'\theta_S \\ =& \theta_S'\left(\frac{X'X}{n} - I\right)\theta_S + \|\theta_S\|_2^2\end{aligned}$$

since we have the inequality

$$u'Mv \leq \|M\|_2\|u\|_1\|v\|_1$$

The term $\theta_S'\left(\frac{X'X}{n} - I\right)\theta_S$can be expressed as follows

$$\theta_S'\left(\frac{X'X}{n} - I\right)\theta_S \leq |||\frac{X'X}{n} - I|||_2\|\theta_S\|_1\|\theta_S\|_1$$

Thus,

$$\begin{aligned}\|\frac{X\theta_S}{\sqrt{n}}\|_2^2 =& \theta_S' \left(\frac{X'X}{n} - I\right)\theta_S + \|\theta_S\|_2^2 \\ \geq& -|||\frac{X'X}{n} - I|||_2 \|\theta_S\|_1^2 + \|\theta_S\|_2^2\end{aligned}$$

since the mutual incoherence parameter is the smallest constant $\delta_{pw}(X)$ such that

$$|||\frac{X'X}{n} - I|||_2 \leq \delta_{pw}(X)$$

we would thus have

$$\begin{aligned}\|\frac{X\theta_S}{\sqrt{n}}\|_2^2 =& \theta_S' \left(\frac{X'X}{n} - I\right)\theta_S + \|\theta_S\|_2^2 \\ \geq& -|||\frac{X'X}{n} - I|||_2 \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\ \geq& -\delta\|\theta_S\|_1^2 + \|\theta_S\|_2^2\end{aligned}$$

Moreover, we have the inequality

$$\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2$$

which leads to

$$\begin{aligned}
\|\frac{X\theta_S}{\sqrt{n}}\|_2^2 =& \theta_S' \left(\frac{X'X}{n} - I\right)\theta_S + \|\theta_S\|_2^2 && (1)\\
\geq& - |||\frac{X'X}{n} - I|||_2 \|\theta_S\|_1^2 + \|\theta_S\|_2^2 && (2)\\
\geq& - \delta\|\theta_S\|_1^2 + \|\theta_S\|_2^2 && (3)\\
\geq& - \delta s\|\theta_S\|_2^2 + \|\theta_S\|_2^2 = (1 - \delta s)\|\theta_S\|_2^2 && (4)
\end{aligned}$$

Since $X\theta_S = -X\theta_{S^c}$ we would also have

$$\begin{aligned}
\|\frac{X\theta_S}{\sqrt{n}}\|_2^2 =& |\theta_S' \left(\frac{X'X}{n} - I\right)\theta_S + \underbrace{\theta_S'\theta_{S^c}}_{=0}| && (5)\\
\leq& \delta\|\theta_S\|_1\|\theta_{S^c}\|_1 && (6)\\
\leq& \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1 && (7)
\end{aligned}$$

Relating equations (1) and (5), we have

$$(1-\delta s)\|\theta_S\|_2^2 \leq \|\frac{X\theta_S}{\sqrt{n}}\|_2^2 \leq \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

Hence, we may write the above as

$$(1-\delta s)\|\theta_S\|_2^2 \leq \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1 \tag{8}$$

$$\|\theta_S\|_2^2 \leq \frac{\delta\sqrt{s}}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1 \tag{9}$$

Recall the inequality $\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2$. Thus, multiplying equation (9) by $\sqrt{s}$, we will have

$$\sqrt{s}\|\theta_S\|_2^2 \leq \frac{s\delta}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

$$\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2 \leq \frac{s\delta}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

Exercise 7.3: Given a matrix $X \in \mathbb{R}^{n\times d}$, suppose that it has pairwise incoherence upper bounded as

$$\delta_{pw}(X) < \frac{\gamma}{s},$$

1. Let $S \subset \{1, \cdots, d\}$be any subset of size s. Show that there is a function $\gamma \to c(\gamma)$ such that $\gamma_{\min}(\frac{X_S' X_S}{n}) \geq c(\gamma) > 0$, as long as γ is sufficiently small.
2. Prove that X satisfies the restricted nullspace property wrt S as long as $\gamma < \frac{1}{3}$.

A more related but sophisticated sufficient condition is the Restricted Isometry Property (RIP). This can be understood as a generalisation of the pairwise incoherence condition, based on looking at conditioning of larger subsets of columns.

Definion (Restricted isometry property):
For a given integer $s \in \{1, \cdots, d\}$, we say that $X \in \mathbb{R}^{n \times d}$ satisfies the RIP of order s with constant $\delta_s(X) > 0$, if

$$|||\frac{X_S' X_S}{n} - I_s|||_2 \leq \delta_s(X) \quad \text{for all subests } S \text{ of size at most } s$$

For $s = 1$,we would have

$$|||\frac{X_j' X_j}{n} - 1|||_2 \leq \delta_1$$
$$|\frac{\|X_j\|_2^2}{n} - 1| \leq \delta_1$$

which implies

$$1 - \delta_1 \leq \frac{\|X_j\|_2^2}{n} \leq 1 + \delta_1$$

for all $j = 1, \cdots, d$. Now consider $s = 2$, and suppose the matrix $X/\sqrt{n}$ has unit-norm columns. Then we would have

$$\frac{X'_{\{j,k\}} X_{\{j,k\}}}{n} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{\|X_j\|_2^2}{n} - 1 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & \frac{\|X_k\|_2^2}{n} - 1 \end{bmatrix} = \begin{bmatrix} 0 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & 0 \end{bmatrix}$$

Now let us consider the l_2-matrix norm, which is the maximum singular value - i.e.,

$$|||\frac{X'_{\{j,k\}} X_{\{j,k\}}}{n} - I_2|||_2 = \max_{j \neq k} |\frac{\langle X_j, X_k \rangle}{n}| = \delta_{pw}(X)$$

Definition (Sandwich relation)

For any matrix X any sparsity level $s \in 2, \cdots, d$, we have the sandwich relation

$$\delta_{pw}(X) \leq \delta_s(X) \leq s\delta_{pw}(X)$$

and neither bound can be improved in general.

Although RIP imposes constraints on much larger submatrices than pairwise incoherence, the magnitude of the constraints required to guarantee uniform RNS property can be milder. Suitable control on the RIP constants implies that the RNS property holds:
If the RIP constant of order $2s$ is bounded as $\delta_{2s}(X) < 1/2$, then the uniform RNS holds for any subset S of cardinality $|S| \leq s$.
Like pairwise incoherence constant, control on the RIP constants is sufficient condition for the BPLP to succeed. A major advantage of the RIP is that for various classes of random design matrices, it can be used to guarantee exactness of basis pursuit using a sample size n that is much smaller than that guaranteed by pairwise incoherence. The RIP approach overcomes the "quadratic barrier" - i.e., the requirement that the sample

size n scales quadratically in the sparsity s, as in the pairwise incoherence approach.

We now consider the noisy setting, in which we observe $(y, X) \in \mathbb{R}^d \times \mathbb{R}^{n \times d}$, which are linked by the model

$$y = X\theta^* + \underbrace{\varepsilon}_{\text{noise}}$$

where $\varepsilon \in \mathbb{R}^n$. Thus, a natural extension of the basis pursuit programme introduced in the noiseless setting is the Lasso programme, which is based on minimizing a weighted combination of the term $\|y - X\theta\|_2^2$ with the l_1-norm penalty, say of the form

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}$$

where $\lambda_n > 0$ is a regularisation parameter, which is chosen by the user.

Different constrained forms of Lasso are as follows

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \quad \text{s.t. } \|\theta\|_1 \leq R$$

for some $R > 0$, or

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{s.t. } \|y - X\theta\|_2^2 \leq b^2$$

for some noise tolerance $b > 0$. The latter is referred to as relaxed basis pursuit by **?**. By Lagrangian duality theory, all three families of convex programmes are equivalent.

subsectionRestricted eigenvalue condition
Achieving perfect recovery is no longer feasible in noisy setting. Therefore, we focus on bounding the l_2-error $\|\hat{\theta} - \theta^*\|_2$, between a Lasso solution $\hat{\theta}$ and the unknown regression vector θ^*. In noisy setting, the required condition is one that is slightly stronger than the RNS property - namely that the restricted eigenvalues of the matrix $\frac{X'X}{n}$ are lower bounded over a cone. In particular, for a constant $\alpha \geq 1$, let us define the set

$$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}$$

Notice that the above set is a special case of our definition in the RNS property, which corresponds to the special case of $\alpha = 1$.

Definition (Restricted eigenvalue condition)

The matrix X satisfies the RE condition over S with parameters (κ, α), if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2, \quad \forall \Delta \in \mathbb{C}_\alpha(S)$$

The RE condition, strengthens the RNS property. In particular, if the RE condition holds with parameters $(\kappa, 1)$ for any $\kappa > 0$, then the RNS property holds. In what follows, we will prove that the error $\|\hat{\theta} - \theta^*\|_2$in the Lasso condition is well controlled.

Intuition for the RE condition:

In the optimisation problem

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n}\|y - X\theta\|_2^2 \right\} \quad \text{s.t. } \|\theta\|_1 \leq R$$

let us consider the radius $R = \|\theta^*\|_1$. With this setting, the true parameter vector $\theta^*$is feasible for the problem. By definition, the Lasso estimate $\hat{\theta}$minimised the quadratic cost function

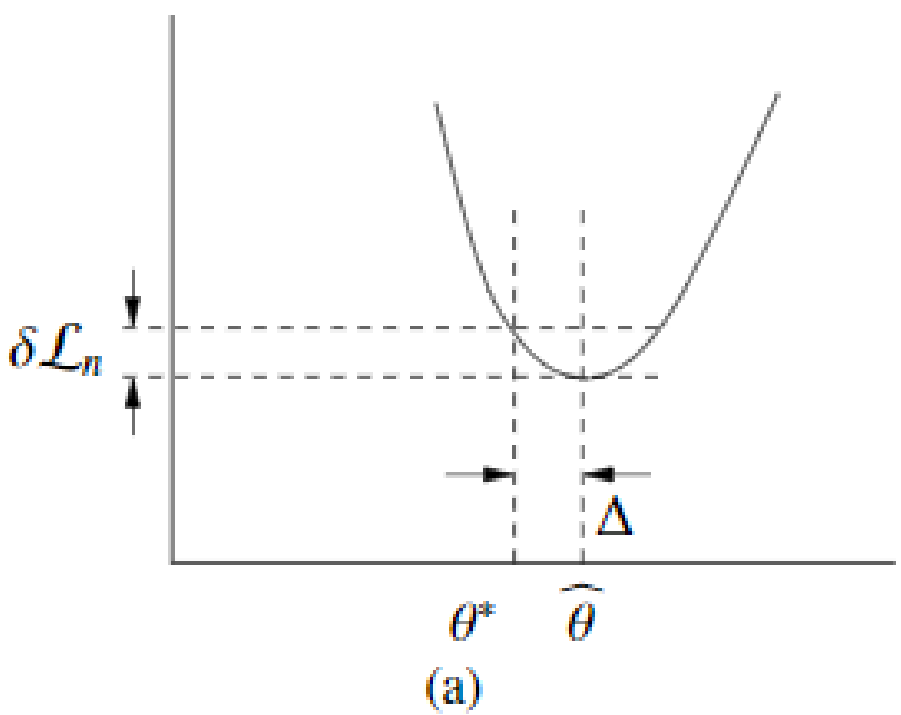
$$L_n(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$$

over the l_1-ball of radius R. As $n \to \infty$, it is expected that θ^*becomes a near-minimiser of the same cost function, so that

$$L_n(\hat{\theta}) \approx L_n(\theta).$$

But when does closeness in cost imply that the error vector $\Delta := \hat{\theta} - \theta^*$is small? Let us first take a look at the following image

Figure:



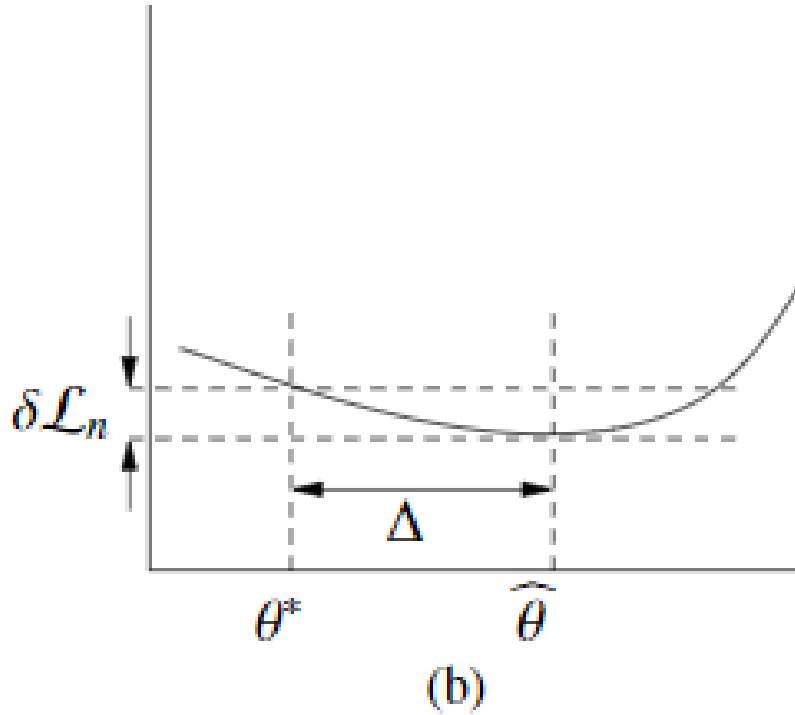


Figure 7.5 Illustration of the connection between curvature (strong convexity) the cost function, and estimation error. (a) In a favorable setting, the cost fu tion is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\delta\mathcal{L}_n$ $\mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ in the cost implies that the error vector $\Delta = \widehat{\theta} - \theta^*$ is not too la (b) In an unfavorable setting, the cost is very flat, so that a small cost difference need not imply small error.

As it is evident from the figure above there is a link between the cost difference of the one-dimensional function $\delta L_n := L_n(\theta^*) - L_n(\hat{\theta})$ and the error $\Delta = \hat{\theta} - \theta^*$ is controlled by the curvature of the cost function. Which of the figures above are favorable and why? For a function in d dimensions, the curvature of a cost function is captured by the structure of its Hessian matrix $\nabla^2 L_n(\theta)$, which is symmetric positive semidefinite matrix.
Notice that in the special case of the quadratic cost function that underlies the Lasso

$$\nabla L_n(\theta) = \frac{1}{n}(\theta X'X - X'y)$$

and the Hessian is calculated as

$$\nabla^2 L_n(\theta) = \frac{1}{n} X'X$$

If we could guarantee that the eigenvalues of this matrix were uniformly bounded away from zero, say that

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 > 0, \quad \forall \Delta \in \mathbb{R}^d \backslash \{0\}$$

then we would be assured of having curvature in all directions.

In the high-dimensional setting $d > n$ and the Hessian is a $d \times d$ with rank at most n, so it is impossible to guarantee that it has positive curvature in all directions. The quadratic cost function always has the below form

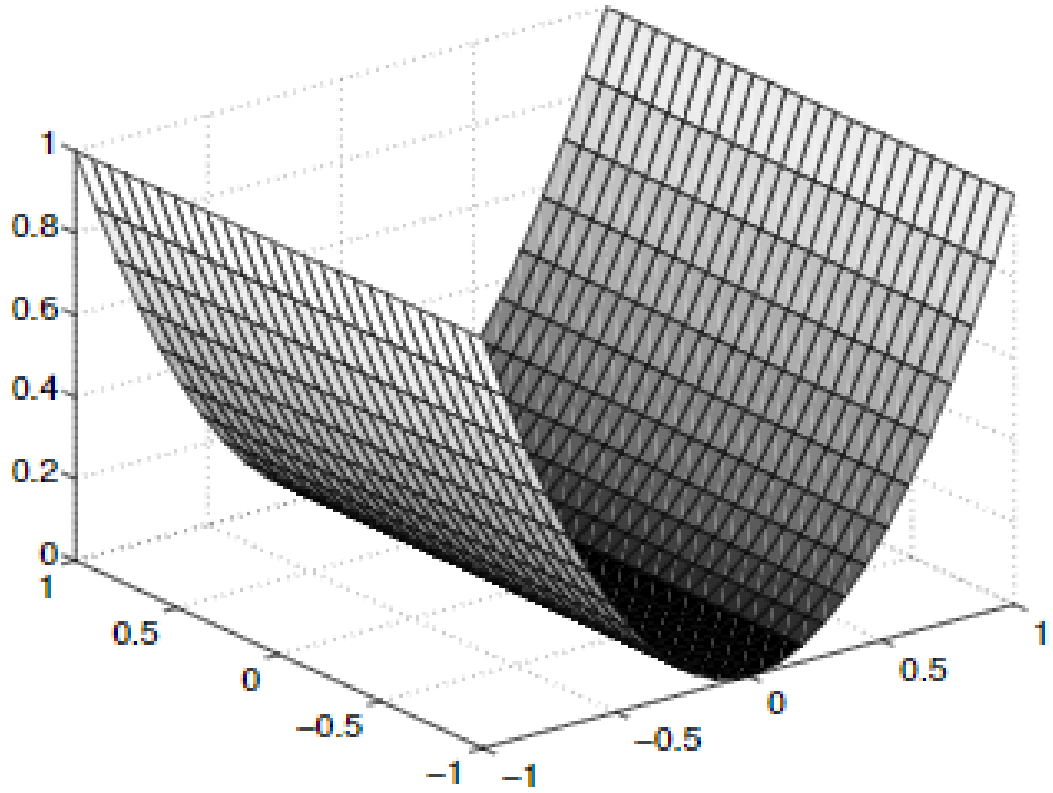


Figure:

Although, it is curved in some directions, there is always $(d-n)$-dimensional

subspace of directions in which it is completely flat, and consequently the uniform lower bound above is never satisfied. Thus, it is necessary to relax the stringency of the uniform curvature condition and require that it holds only for a subset $\mathbb{C}_\alpha(S)$ of vectors. If we can be assured that the subset $\mathbb{C}_\alpha(S)$ is well aligned with the curved direction of the Hessian, then a small difference in the cost function will translate into bounds on the difference between $\hat{\theta}$ and θ^*.
Let us now provide the bound on the error $\|\hat{\theta} - \theta^*\|_2$in the case of a "hard sparse" vector θ^*. In particular, let us impose the following conditions:

1. The vector $\theta^*$is supported on the subset $S \subseteq \{1, 2, \cdots, d\}$ with $|S| = s$.
2. The design matrix X satisfies the restricted eigenvalue condition

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2, \quad \forall \Delta \in \mathbb{C}_\alpha(S)$$

   with parameter $(\kappa, 3)$.

The following results provides bounds on the l_2-error between any Lasso solution $\hat{\theta}$ and the true vector θ^*.

Theorem

Under the above assumptions:

1. *Any solutions of the* **Lagrangian Lasso**

$$\hat{\theta} \in \min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}$$

with

$$\lambda_n \geq 2 \| \frac{X'\varepsilon}{n} \|_\infty$$

satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa} \sqrt{s} \lambda_n.$$

2. *Any solution of the* **constrained Lasso**

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\}, \quad \textit{such that } \|\theta\|_1 \leq \|\theta^*\|_1,$$

satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2 < \frac{4}{-} \sqrt{s} \| \frac{X'\varepsilon}{-} \|_\infty$$

The above results are deterministic and apply to any set of linear regression equations. However, the results involve unknown quantities stated in terms of ε and/or θ^* . Obtaining results for specific statistical models - as determined by the noise vector ε and the random design matrix X - involves bounding or approximating these quantities.

Intuition:

- Based on our earlier discussion of the role of convexity, naturally all three upper bounds are inversely proportional to the restricted eigenvalue constant $\kappa \geq 0$.
- Their scaling with $\sqrt{s}$ is also natural, since we are trying to estimate the unknown regression vector with s unknown entries.
- The remaining terms involve the unknown noise vector.

Before, we carry on with an example for classical linear Gaussian models, let us first preface the theory, with Exercise 2.12 of the second chapter.

Exercise 2.12 (Upper bounds for sub-Gaussian maxima): Let $\{X_i\}_{i=1}^n$ be a sequence of zero mean random variables, each sub-Gaussian with parameter σ. (No independence assumption are needed). Prove that

$$\mathbb{E}[\max_{i=1,\cdots,n} X_i] \leq \sqrt{2\sigma^2 \log n}$$

For any $\lambda > 0$, we can use the convexity of the exponential function and Jensen's inequality to obtain

$$\exp\{\lambda \mathbb{E}[\max_{i=1,\cdots,n} X_i]\} \leq \mathbb{E}[\exp \lambda \max_{i=1,\cdots,n} X_i]$$

using the monotonicity of exponential function

$$\mathbb{E}[\exp \lambda \max_{i=1,\cdots,n} X_i] = \mathbb{E}[\max_{i=1,\cdots,n} \exp \lambda X_i] \leq \sum_{i=1}^{n} \mathbb{E}[\exp \lambda X_i] \leq n \exp \frac{\lambda^2 \sigma^2}{2}$$

hence, $\mathbb{E}[\max_{i=1,\cdots,n} X_i] \leq \frac{\log n}{\lambda} + \lambda \frac{\sigma^2}{2}$ and optimising over $\lambda > 0$ yields $\lambda = \frac{\sqrt{2 \log n}}{\sigma}$; substituting

$$\mathbb{E}[\max_{i=1,\cdots,n} X_i] \leq \frac{\sigma}{\sqrt{2}} \sqrt{\log n} + \frac{\sigma}{\sqrt{2}} \sqrt{\log n} = \sqrt{2\sigma^2 \log n}$$

Example (Classical linear Gaussian model):
Consider a classical linear Gaussian model with noise vector $\varepsilon \in \mathbb{R}^n$with i.i.d. $N(0, \sigma^2)$ entries. Let us consider a fixed design matrix $X \in \mathbb{R}^{n \times d}$. Suppose that matrix X satisfies the RE condition over S with parameters (κ, α), meaning that

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2, \quad \forall\Delta \in \mathbb{C}_\alpha(S)$$

and that is C column normalised, meaning that $\max_{j=1,\cdots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$. With this setup, the r.v. $\|\frac{X'\varepsilon}{n}\|_\infty$ corresponds to the absolute maximum of d zero-mean Gaussian variables, each with variance at most $\frac{C^2\sigma^2}{n}$. From exercise 2.12, it is evident that

$$P\left[\|\frac{X'\varepsilon}{n}\|_\infty \geq C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right)\right] \leq 2\exp\left(-\frac{n\delta^2}{2}\right), \quad \forall\delta > 0$$

Now let us set $\lambda_n = 2C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right)$, which would turn

$$\begin{aligned}\|\frac{X'\varepsilon}{n}\|_\infty \geq C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right) &= 2\|\frac{X'\varepsilon}{n}\|_\infty \geq 2C\sigma\left(\sqrt{\frac{2\log d}{n}} + \delta\right) \\ &= 2\|\frac{X'\varepsilon}{n}\|_\infty \geq \lambda_n\end{aligned}$$

From Theorem 7.13, we have that the optimal solution to of the Lagrangian Lasso satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{6C\sigma}{\kappa}\sqrt{s}\left\{\sqrt{\frac{2\log d}{n}} + \delta\right\}$$

with probability at least $1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$. The same Theorem can be used to show the bounds for the constrained Lasso.

The most significant difference between the constrained Lasso and the Lagrangian Lasso is that the Constrained Lasso assumed exact knowledge of the l_1-norm $\|\theta^*\|_1$, whereas the Lagrangian Lasso only requires knowledge of the noise variance σ^2. As you may suspect, in practice it is straightforward to estimate the noise variance, whereas the l_1-norm is a more delicate object.

Example (Compressed sensing):
In the domain of compressed sensing,the design matrix X can be chosen by the user, and one standard choice is the standard Gaussian matrix with i.i.d. $N(0,1)$ entries. Suppose that the noise vector $\varepsilon \in \mathbb{R}^n$is deterministic, say with bounded entries ($\|\varepsilon\|_\infty \leq \sigma$). Under these assumptions, each variable

$\frac{X_j'\varepsilon}{\sqrt{n}}$ is a zero mean Gaussian with variance at most σ^2. Thus, by following the same argument as in the preceding example, we conclude that Lasso will again satisfy the bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{6\sigma}{\kappa}\sqrt{s}\left\{\sqrt{\frac{2\log d}{n}} + \delta\right\}$$

with $C = 1$.

In what follows, we prove the bounds on the l_2- error between any Lasso solution $\hat{\theta}$and the true vector θ^*. Let us first consider the constrained Lasso

$$\min_{\theta \in \mathbb{R}^d}\left\{\frac{1}{2n}\|y - X\theta\|_2^2\right\}, \quad \text{such that } \|\theta\|_1 \leq R,$$

Given the choice $R = \|\theta^*\|_1$, the tangent vector θ^* is feasible. Since $\hat{\theta}$ is optimal,we have the inequality

$$\frac{1}{2n}\|y - X\hat{\theta}\|_2^2 \leq \frac{1}{2n}\|y - X\theta^*\|_2^2$$

Let us define the error vector $\hat{\Delta} = \hat{\theta} - \theta^*$. Substituting y with $X\theta^* + \varepsilon$

and expanding the above inequality we get

$$\begin{aligned}
\frac{1}{2n}\|y - X\hat{\theta}\|_2^2 &\leq \frac{1}{2n}\|y - X\theta^*\|_2^2 \\
\|X\theta^* + \varepsilon - X\hat{\theta}\|_2^2 &\leq \|X\theta^* + \varepsilon - X\theta^*\|_2^2 \\
\|X(\theta^* - \hat{\theta}) + \varepsilon\|_2^2 &\leq \|\varepsilon\|_2^2 \\
\|X(\theta^* - \hat{\theta})\|_2^2 + \|\varepsilon\|_2^2 + 2\varepsilon' X(\theta^* - \hat{\theta}) &\leq \|\varepsilon\|_2^2 \\
\|X(\theta^* - \hat{\theta})\|_2^2 + 2\varepsilon' X(\theta^* - \hat{\theta}) &\leq 0 \\
\|X\hat{\Delta}\|_2^2 &\leq 2\varepsilon' X\hat{\Delta} \\
\frac{\|X\hat{\Delta}\|_2^2}{n} &\leq \frac{2\varepsilon' X\hat{\Delta}}{n}
\end{aligned}$$

[

Holder's inequality]
Recall that for any $p \geq 1$,

$$\mathbb{E}|XY| \leq \|X\|_p \|Y\|_q$$

where $q = p/(p-1)$ if $p > 1$, and $q = \infty$ if $p = 1$.

Given the above definition, we can apply Holder inequality to the right hand-side of the penultimate inequality to obtain

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq 2\|\frac{X'\varepsilon}{n}\|_\infty \|\hat{\Delta}\|_1 \tag{10}$$

As it was shown in an earlier Theorem, whenever $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$for an S-sparse vector, the error$\hat{\Delta}$belongs to the cone $\mathbb{C}_1(S)$, whence

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \leq 2\|\hat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\hat{\Delta}\|_2$$

Since $\mathbb{C}_1(S)$ is a subset of $\mathbb{C}_3(S)$, we may apply the restricted eigenvalue

condition to the left hand side of inequality (10), thereby obtaining

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \geq \kappa \|\hat{\Delta}\|_2^2$$

Putting all this together yields the claimed bound.
The earlier Theorem assumed that the design matrix X satisfies the restricted eigenvalue condition (RE). In practice, it is difficult to verify that a given design matrix X satisfies this condition. However, it is possible to give high-probability results in the case of random design matrices. As discussed previously, pairwise incoherence and RIP conditions are one way in which to certify the restricted nullspace and eigenvalue properties, and are well suited to isotropic designs (in which the population covariance matrix of the rows X_i is the identity). Many other random design matrices encountered in practice do not have such an isotropic structure, so that it is desirable to have alternative direct verifications of the restricted nullspace property.

Theorem

Consider a random matrix $X \in \mathbb{R}^{n \times d}$,in which each row $x_i \in \mathbb{R}^d$ is drawn i.i.d from $N(0, \Sigma)$ distribution. Then there are universal positive constants $c_1 < 1 < c_2$ such that

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

with probability at least $1 - \frac{\exp(-n/32)}{1-\exp(-n/32)}$, where $\rho^2(\Sigma)$ is the maximum diagonal entry of the covariance matrix Σ.

The above Theorem can be used to establish restricted nullspace and eigenvalue conditions for various matrix ensembles that do not satisfy incoherence or RIP conditions.

Example (Geometric decay):

Consider a covariance matrix with Toeplitz structure $\Sigma_{ij} = \nu^{|i-j|}$for some parameter $\nu \in [0, 1)$. This type of geometrically decaying covariance struc-

ture arises naturally from autoregressive processes, e.g.,

$$\begin{bmatrix} 1 & \nu & \nu^2 & \cdots & \nu^{n-2} & \nu^{n-1} \\ \nu & 1 & \nu & \nu^2 & \vdots & \nu^{n-2} \\ \nu^2 & \nu & 1 & \nu & \ddots & \vdots \\ \vdots & \nu^2 & \ddots & \ddots & \ddots & \nu^2 \\ \nu^{n-2} & \vdots & \ddots & \nu & 1 & \nu \\ \nu^{n-1} & \nu^{n-2} & \cdots & \nu^2 & \nu & 1 \end{bmatrix}$$

where ν allows for the tuning of the memory process. By classical results, we have $\gamma_{\min}(\Sigma) \geq (1-\nu)^2 > 0$ and $\rho^2(\Sigma) = 1$, independently of dimension d. Consequently, the earlier Theorem implies that, with high probability, the sample covariance matrix $\hat{\Sigma} = \frac{X'X}{n}$ obtained by sampling from this distribution will satisfy the RE condition for all subsets S of cardinality at most $|S| \leq \frac{c_1}{32c_2}(1-\nu^2)\frac{n}{\log d}$