Sparse linear models in high dimensions

Han Zhang & Yu Zhang

October 24, 2022

Table of Contents

Table of Contents

Suppose that we observe a vector $y \in \mathbb{R}^n$ and a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ that are linked via the standard linear model

$$y = \mathbf{X}\theta^* + w, \tag{7.1}$$

where $w \in \mathbb{R}^n$ is a vector of noise variables.
The focus of this chapter is settings in which n is smaller than d ,it is necessary to impose additional structure on the unknown regression vector $\theta^* \in \mathbb{R}^d$.

Different sparsity models

- hard sparsity ,meaning that

$$\theta^* \in S(\theta^*) := \left\{ j \in \{1, 2, \ldots, d\} \mid \theta_j^* \neq 0 \right\}, \tag{7.2}$$

  which is known as the support set of θ^*

- weak sparsity .Roughly speaking ,a vectorθ^* is *weakly sparse* if it can be closely approximated by a sparse vector.
  One way via the l_q-"norm".For a parameter $q \in [0, 1]$and radius $R_q > 0$,consider the set

$$\mathbb{B}_q(R_q) = \left\{ \theta \in \mathbb{R}^d | \sum_{j=1}^{d} |\theta_q|^q \leq R_q \right\}. \tag{7.3}$$

  It is known as the l_q-ball of radiusR_q.

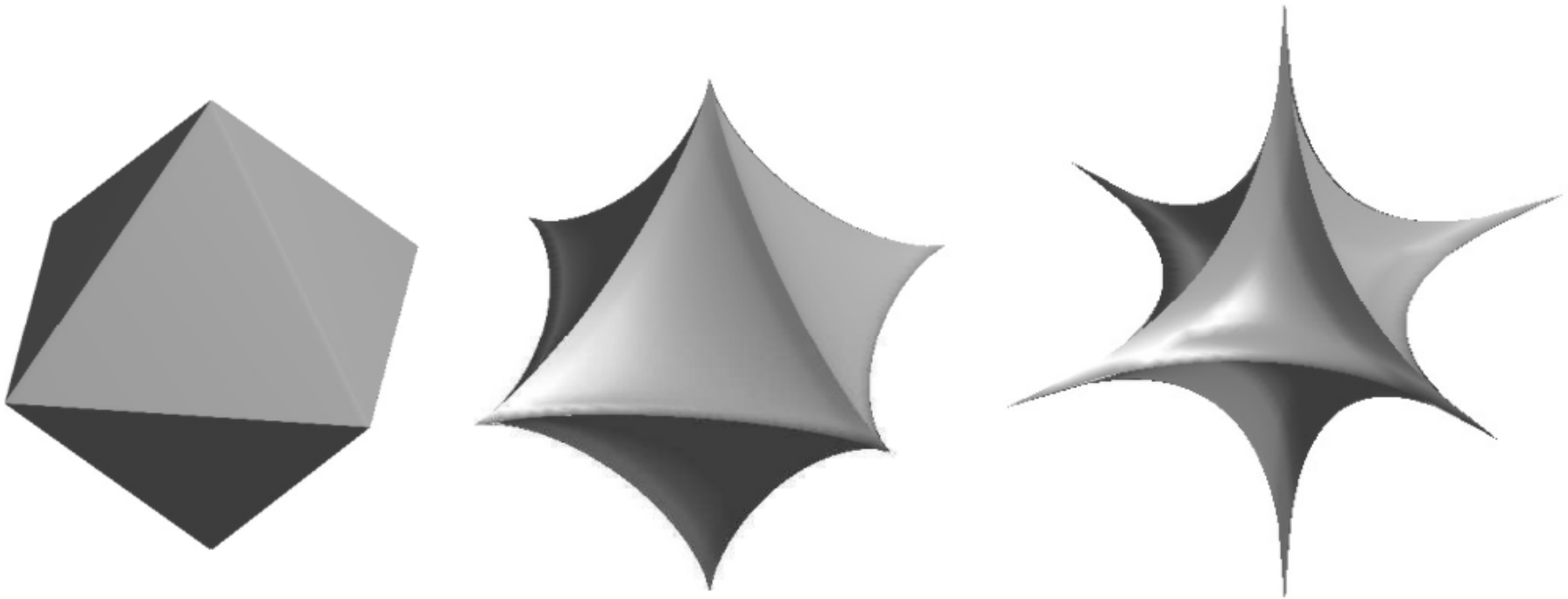

Figure 7.1 Illustrations of the ℓ_q-"balls" for different choices of the parameter $q \in (0, 1]$. (a) For $q = 1$, the set $\mathbb{B}_1(R_q)$ corresponds to the usual ℓ_1-ball shown here. (b) For $q = 0.75$, the ball is a non-convex set obtained by collapsing the faces of the ℓ_1-ball towards the origin. (c) For $q = 0.5$, the set becomes more "spiky", and it collapses into the hard sparsity constraint as $q \to 0^+$. As shown in Exercise 7.2(a), for all $q \in (0, 1]$, the set $\mathbb{B}_q(1)$ is star-shaped around the origin.

Table of Contents

l_q-based relaxation

Let us define $\|\theta\|_0 := \sum_{j=1}^{d} \mathbb{I}[\theta_j \neq 0]$.
We consider the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad \text{such that} \mathbf{X}\theta = y. \tag{7.8}$$

- the cost function is non-differentiable and non-convex.
- by traversing,but it is not practical.

A strategy is to replace with the l_0-norm.We get the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that} \mathbf{X}\theta = y. \tag{7.9}$$

Unlike the l_0-version,this is now a convex program,we refer to it as the *basis pursuit linear program*.

Exact recovery and restricted nullspace

Q:when it is equivalent?
There is a vector $\theta^* \in \mathbb{R}^d$ such that $y = \mathbf{X}\theta^*$,the vectorθ^* has support $S \subset \{1, 2, \ldots, d\}$,meaing that $\theta_j^* = 0$ for all $j \in S^c$. The nullspace of $\mathbf{X}$is givenby $null\,(\mathbf{X}) := \left\{\Delta \in \mathbb{R}^d | \mathbf{X}\Delta = 0\right\}$.The *tangent cone* of the l_1-ball at θ^* is given by

$$\mathbb{T}(\theta^*) = \left\{\Delta \in \mathbb{R}^d | \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{for some} t > 0\right\}. \qquad (7.10)$$

If $\theta^*$is the unique optimal solution of the basis pursuit (LP),then it must be the case that the intersection of the nullspace $null(\mathbf{X})$with this tangent cone contains only the zero vector.

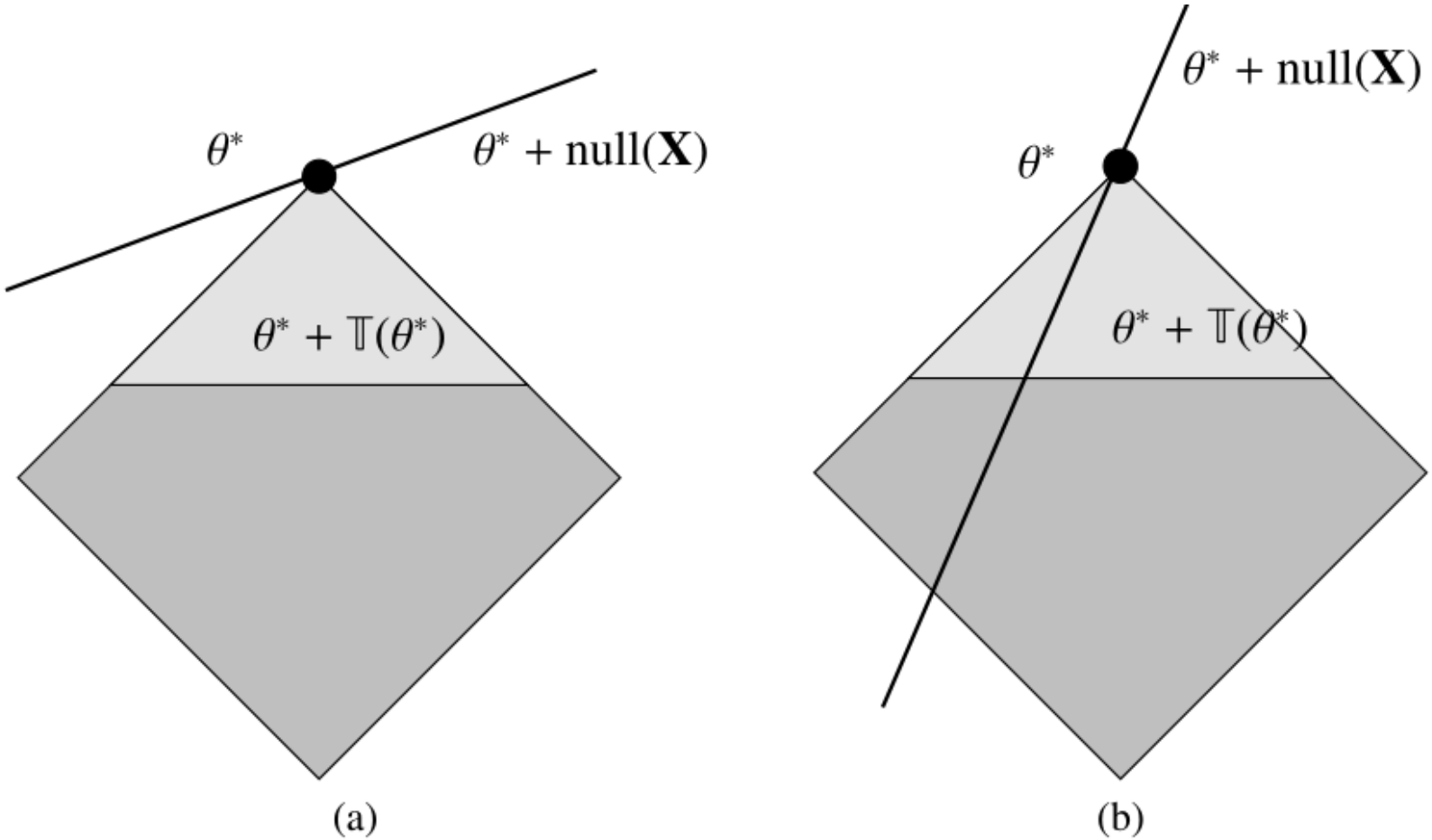

Figure 7.2 Geometry of the tangent cone and restricted nullspace property in $d = 2$ dimensions. (a) The favorable case in which the set $\theta^* + \text{null}(\mathbf{X})$ intersects the tangent cone only at θ^*. (b) The unfavorable setting in which the set $\theta^* + \text{null}(\mathbf{X})$ passes directly through the tangent cone.

Let us define the subset

$$\mathbb{C}(S) = \left\{ \Delta \in \mathbb{R}^d \middle| \|\Delta_{S^c}\|_1 \leq \|\Delta_{S\|_1} \right\},$$

Definition (7.7)

The matrix $\mathbb{X}$ satisfies the *restricted nullspace property* with respect to Si if $\mathbb{C}(S) \cap null(\mathbb{X}) = \{0\}$.

Theorem (7.8)

The following two properties are equivalent:
(a)For any vector $\theta^* \in \mathbb{R}^d$ with support S,the basis pursuit program(7.9)applied with $y = \mathbf{X}\theta^*$ has unique solution $\widehat{\theta} = \theta^*$.
(b)The matrix $\mathbf{X}$ satisfies the restricted nullspace property with respect to S.

Proof.

- $(b) \Rightarrow (a)$ Since both $\widehat{\theta}$ and θ^* are feasible for the basis pursuit program, and since$\widehat{\theta}$ is optimal, we have $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1$.Defining the error vector$\widehat{\Delta} := \widehat{\theta} - \theta^*$,we have

$$\begin{aligned}\|\theta^*_S\|_1 = \|\theta^*\|_1 &\geq \|\theta^* + \widehat{\Delta}\|_1 \\ &= \|\theta^*_S + \widehat{\Delta_S}\|_1 + \|\widehat{\Delta_{S^c}}\|_1 \\ &\geq \|\theta^*_S\|_1 - \|\widehat{\Delta_S}\|_1 + \|\widehat{\Delta_{S^c}}\|_1,\end{aligned}$$

where we have used the fact that $\theta^*_{S^c} = 0$ We conclude that the error $\widehat{\Delta} \in \mathbb{C}(S)$. However ,by construction $\mathbf{X}\widehat{\Delta} = 0, \widehat{\Delta} \in \mathit{null}(\mathbf{X})$.So $\widehat{\Delta} = 0$

or equivalently that $\widehat{\theta} = \theta^*$.

- $(a) \Rightarrow (b)$ it suffices to show that, if the l_1-relaxation succeeds for all S-sparse vectors, then the set $null(\mathbf{X})\backslash\{0\}$ has no intersection with $\mathbb{C}(S)$.For a given vector$\theta^* \in null(\mathbf{X})\backslash\{0\}$, consider the basis pursuit problem

$$min_{\beta \in \mathbb{R}^d} \|\beta\|_1 \qquad \text{such that } \mathbf{X}\beta = \mathbf{X} \begin{bmatrix} \theta^*_S \\ 0 \end{bmatrix}. \tag{7.11}$$

  By assumption, the unique optimal solution will be $\widehat{\beta} = [\theta^*_S \quad 0]^T$. Since $\mathbf{X}\theta^* = 0$by assumption, the vector $[0 \quad -\theta^*_{S^c}]^T$is also feasible for the problem, and, by uniqueness, we must have $\|\theta^*_S\|_1 < \|\theta^*_{S^c}\|_1$ implying that$\theta^* \notin \mathbb{C}(S)$as claimed.

Sufficient conditions for restricted nullspace

The earliest sufficient conditions were based on the incoherence parameter of the$\mathbf{X}$,namely the quantity

$$\delta_{PW}(\mathbf{X}) := max_{j,k=1,\ldots,d} \left| \frac{|\langle X_j, X_k \rangle|}{n} - \mathbb{I}[j = k] \right|, \tag{7.12}$$

Proposition (7.9)

If the pairwise incoherence satisfies the bound

$$\delta_{PW}(\mathbf{X}) \leq \frac{1}{3s}, \tag{7.13}$$

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

The proof is involved in Exercise7.3.

Definition (7.10 Restricted isometry property)

For a given integer $s \in \{1, \ldots, d\}$,we say that $\mathbf{X} \in \mathbb{R}^{n \times d}$satisfies a restricted isometry property of order s with constant $\delta_s(\mathbf{X}) > 0$ if

$$\left\| \frac{\mathbf{X}_S^T \mathbf{X}_S}{n} - \mathbf{I}_s \right\|_2 \leq \delta_s(\mathbf{X}) \qquad \text{for all subsets S of size at most s.} \quad (7.14)$$

For s=2,the RIP constant δ_2 is very closely related to the pairwise incoherence parameter $\delta_{PW}(\mathbf{X})$,in the case when the matrix$\frac{\mathbf{X}}{\sqrt{n}}$ has unit-norm columns,we have

Although RIP imposes constraints on much lager submatrices than pairwise incoherence,it is milder.

Proposition (7.11)

If the RIP constant of order 2s is bounded as $\delta_{2s}(\mathbf{X}) < \frac{1}{3}$, then the uniform restricted nullspace property holds for any subset S of cardinality $|S| \leq s$.

Proof.

Let $\theta \in null(\mathbf{X})$ be an arbitrary non-zero member of the nullspace. For any subset A, we let $\theta_A \in \mathbb{R}^{|A|}$ denote the subvector of elements indexed by A, and we define the vector$\tilde{\theta_A} \in \mathbb{R}^d$ with elements

$$\tilde{\theta_j} = \begin{cases} \theta & \text{if } j \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Let S be the subset of $\{1, 2, \ldots, d\}$corresponding to the s entries of θ that are largest in absolute value. It suffices to show that$\|\theta_{S^c}\|_1 > \|\theta_S\|_1$or this subset. Let us write $S^c = \cup_{j \geq 1} S_j$, where $S_1$is the subset of indices given by the s largest values of $\tilde{\theta_{S^c}}$; the subset S_2 is the largest s in the subset $S^c \backslash S_1$, and the final subset may contain fewer than s entries. Using this notation, we have the decomposition $\theta = \tilde{\theta_S} + \sum_{k \geq 1} \tilde{\theta_{S^k}}$.

The RIP property guarantees that $\|\tilde{\theta}_S\|_2^2 \leq \frac{1}{1-\delta_{2s}} \|\frac{1}{\sqrt{n}}\mathbf{X}\tilde{\theta}_S\|_2^2$. Moreover, since$\theta \in null(\mathbf{X})$,we have $\mathbf{X}\tilde{\theta}_S = -\sum_{j\geq 1}\mathbf{X}\tilde{\theta}_{S_j}$, and hence

$$\|\tilde{\theta}_{S_0}\|_2^2 \leq \frac{1}{1-\delta_{2s}}\left|\sum_{j\geq 1}\frac{\langle \mathbf{X}\tilde{\theta}_{S_0}, \mathbf{X}\tilde{\theta}_{S_j}\rangle}{n}\right| = \frac{1}{1-\delta_{2s}}\left|\sum_{j\geq 1}\tilde{\theta}_{S_0}\left[\frac{\mathbf{X}^T\mathbf{X}}{n} - \mathbf{I}^d\right]\tilde{\theta}_{S_j}\right|,$$

By the RIP property,$\|n^{-1}\mathbf{X}_{S_0\cup S_j}^T\mathbf{X}_{S_0\cup S_j}^T - \mathbf{I}_2\|_2 \leq \delta_{2s}$,and hence we have

$$\|\tilde{\theta}_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\sum_{j\geq 1}\|\tilde{\theta}_{S_j}\|_2, \tag{7.16}$$

Finally,by construction of the sets S_j,for each $j \geq 1$,we have $\|\tilde{\theta}_{S_j}\|_{infty} \leq \frac{1}{s}\|\tilde{\theta}_{S_{j-1}}\|_1$,which implis that $\|\tilde{\theta}_{S_j}\|_2 \leq \frac{1}{\sqrt{s}}\|\tilde{\theta}_{S_{j-1}}\|_1$.Applying these upper bounds to the inequality(7.16),we obtain

$$\|\tilde{\theta}_{S_0}\|_1 \leq \sqrt{s}\|\tilde{\theta}_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\left\{\|\tilde{\theta}_{S_0}\|_1 + \sum_{j\geq 1}\|\tilde{\theta}_{S_j}\|_1\right\}.$$

or equivalently $\|\tilde{\theta}_{S_0}\|_1 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\left\{\|\tilde{\theta}_{S_0}\|_1 + \|\tilde{\theta}_{S^c}\|_1\right\}$.Some simple algebra verifies that this inequality implies that $\|\tilde{\theta}_{S_0}\|_1 < \|\tilde{\theta}_{S^c}\|_1$ as long as $\delta_{2s} < \frac{1}{3}$.

A major advantage of the RIP approach is,for many classes of random design matrices,to guarantee exactness of basis pursuit ,it uses much smaller sample compared with pairwise incoherence.
Unlike the restricted nullspace property ,the pairwise incoherence condition and the RIP condition are not necessary conditions for basis pursuit problem.

Table of Contents

A natural extension of the basis pursuit program is based on minimizing a weighted combination of the term $\|y - \mathbf{X}\theta\|_2^2$with the l_1-norm penalty,say of the form

$$\widehat{\theta} \in argmin_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}. \tag{7.18}$$

We refer to it as the Lasso program. We can consider different forms of the Lasso,that is either

$$min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \right\} \qquad \text{such that} \|\theta\|_1 \leq R \tag{7.19}$$

for some radius R¿0,or

$$min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \qquad \text{such that} \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \leq b^2 \tag{7.20}$$

We refer to it as the *relaxed basis pursuit*.By Lagrangian duality ,all three families of convex programs are equivalent.

Restricted eigenvalue condition

Let us define the set

$$\mathbb{C}_\alpha(S) = \left\{\Delta \in \mathbb{R}^d \middle| \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_{S\|_1}\right\}, \tag{7.21}$$

Definition (7.12)

The matrix $\mathbf{X}$ satisfies the restricted eigenvalue(RE)conditionover S with parameters(k, α) if

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq k\|\Delta\|_2^2 \qquad \text{for all}\Delta \in \mathbb{C}_\alpha(S). \tag{7.22}$$

If the RE condition holds with parameters(k,1)for any k¿0,the the restricted nullspace property holds **how?**
About why the RE condition is useful ,we will have a intuitive illustration.

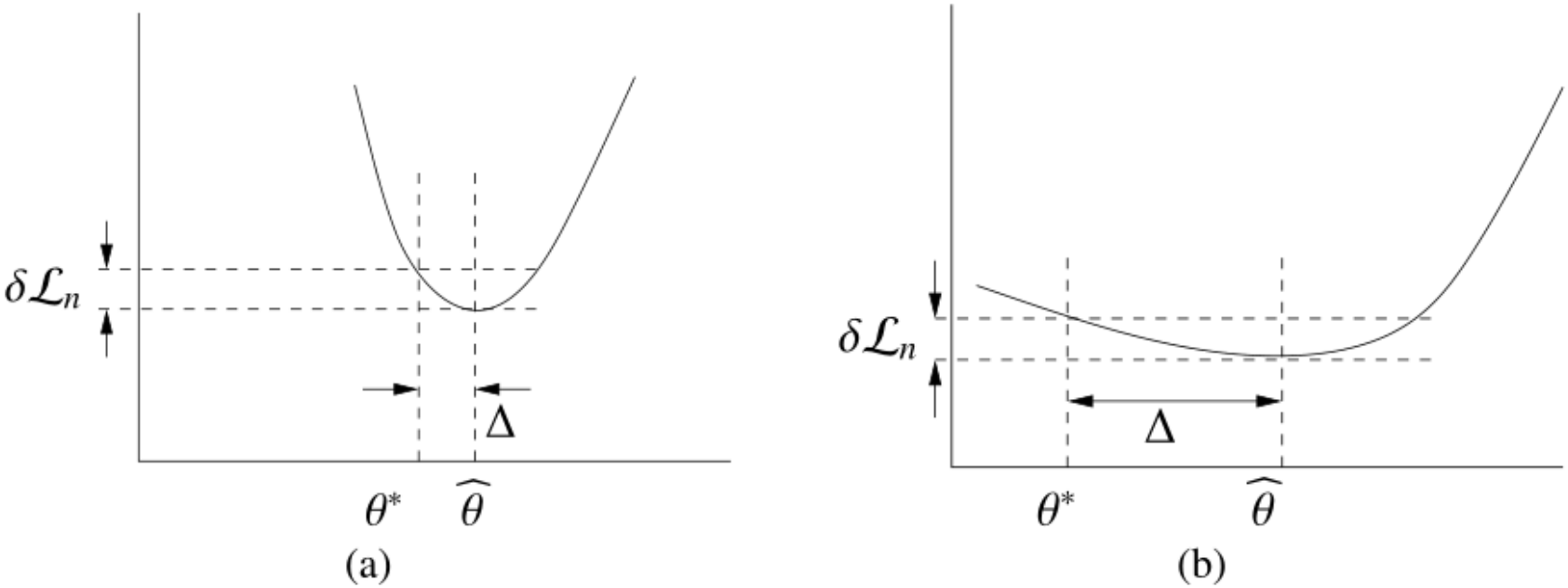

Figure 7.5 Illustration of the connection between curvature (strong convexity) of the cost function, and estimation error. (a) In a favorable setting, the cost function is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) -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

The curvature of a cost function ($\mathcal{L}_n(\widehat{\theta}) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$) is captured by the structure of its Hessian matrix $\nabla^2\mathcal{L}_n(\theta)$,which here is $\frac{1}{n}\mathbf{X}^T\mathbf{X}$.
So with the RE condition holds ,we will see Δ has curvature in all directions ,for all $\Delta \in \mathbb{R}^d \backslash \{0\}$
Let us impose the following conditions for next subsection :
(A1) The vectorθ^* is supported on a subset $S \subseteq \{1, 2, \ldots, d\}$ with $|S| = s$. (A2)The design matrix satisfies the restricted eigenvalue condition(7.22)over S with parameters(k,3).

Bounds on l_2-error for hard sparse models

Theorem (7.13)

Under assumptions(A1)and(A2):
(a)Any solution of the Lagrangian Lasso(7.18)with regularization parameter lower bounded as $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$ *satisfies the bound*

$$\|\theta - \theta^*\|_2 \leq \frac{3}{k}\sqrt{s}\lambda_n. \tag{7.25a}$$

(b)Any solution of the constrained Lasso (7.19)with $R = \|\theta^*\|_1$ *satisfies the bound*

$$\|\theta - \theta^*\|_2 \leq \frac{4}{k}\sqrt{s}\|\frac{\mathbf{X}^T w}{n}\|_\infty. \tag{7.25b}$$

(c)Any solution of the relaxed basis pursuit program(7.20)with $b^2 \geq \frac{\|w\|_2^2}{2n}$ *satisfies the bound*

$$\|\theta - \theta^*\|_2 \leq \frac{4}{k}\sqrt{s}\|\frac{\mathbf{X}^T w}{n}\|_\infty + \frac{2}{\sqrt{k}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}. \tag{7.25c}$$

In adition,all three solutions satisfy the
l_1-bound$\|\theta - \theta^*\|_1 \leq 4\sqrt{s}\|\theta - \theta^*\|_2$.

Proof.
(b)Given the choice R $= \|\theta^*\|_1$ the target vectorθ^* is feasible. Since $\widehat{\theta}$ is optimal, we have the inequality $\frac{1}{2n}\|y - \mathbf{X}\widehat{\theta}\|_2^2 \leq \frac{1}{2n}\|y - \mathbf{X}\theta^*\|_2^2$. Defining the error vector$\widehat{\Delta} := \widehat{\theta} - \theta^*$and performing some algebra yields the basic inequality

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leq \frac{2w^T\mathbf{X}\widehat{\Delta}}{n}. \tag{7.28}$$

Applying Holder inequality to the right-hand side yields $\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leq 2\left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty \|\widehat{\Delta}\|_1$ As shown in the proof of Theorem 7.8, whenever for an S -sparse vector, the error $\widehat{\Delta}$ belongs to the cone$\mathbb{C}_1(S)$,whence

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

Since $\mathbb{C}_1(S)$ is a subset of$\mathbb{C}_3(S)$, we may apply the restricted eigenvalue condition (7.22) to the left-hand side of the inequality (7.28), thereby obtaining $\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \geq k\|\widehat{\Delta}\|_2^2$. Putting together the pieces yields the claimed bound.

(c) Note that $\frac{1}{2n}\|y - \mathbf{X}\theta^*\|_2^2 = \frac{\|w\|_2^2}{2n} \leq b^2$, where the inequality follows by our assumed choice of b. Thus, the target vectorθ^* is feasible, and since $\widehat{\theta}$ is optimal, we have$\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1$. As previously reasoned, the error vector $\widehat{\Delta} := \widehat{\theta} - \theta^*$ must then belong to the cone $\mathbb{C}_1(S)$. Now by the feasibility of$\widehat{\theta}$, we have

$$\frac{1}{2n}\|y - \mathbf{X}\widehat{\theta}\|_2^2 \leq b^2 = \frac{1}{2n}\|y - \mathbf{X}\theta^*\|_2^2 + \left(b^2 - \frac{\|w\|_2^2}{2n}\right).$$

Rearranging yields the modified basic inequality

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leq \frac{2w^T\mathbf{X}\widehat{\Delta}}{n} + 2\left(b^2 - \frac{\|w\|_2^2}{2n}\right).$$

Applying the same argument as in part (b)—namely, the RE condition to the left-hand side and the cone inequality to the right-hand side—we obtain

$$k\|\widehat{\Delta}\|_2^2 \leq 4\sqrt{s}\|\widehat{\Delta}\|_2 \left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty + 2\left(b^2 - \frac{\|w\|_2^2}{2n}\right),$$

which implies that $\|\widehat{\Delta}\|_2 \leq \frac{8}{k}\sqrt{s}\left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty + \frac{2}{\sqrt{k}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}$,as claimed.

(a)Our first step is to show that, under the condition $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$, the error vector$\widehat{\Delta}$ belongs to$\mathbb{C}_3(S)$. To establish this intermediate claim, let us define the Lagrangian $L(\theta; \lambda_n) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + \lambda_n\|\theta\|_1$Since $\widehat{\theta}$ is optimal, we have

$$L(\widehat{\theta}; \lambda_n) \leq L(\theta^*; \lambda_n) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + \lambda_n\|\theta^*\|_1.$$

Rearranging yields the lagrangian basic inequality

$$0 \leq \frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq \frac{w^T\mathbf{X}\widehat{\Delta}}{n} + \lambda_n\left\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right\}. \tag{7.29}$$

Now since θ^* is S-sparse,we can write

$$\|\theta^*\|_1 - \|\widehat{\theta}\|_1 = \|\theta^*_S\|_1 - \|\theta^*_S + \widehat{\Delta}_S\|1 - \|\widehat{\Delta}_{S^c}\|_1.$$

Substituting into the basic inequality(7.29)yields

$$\begin{aligned} 0 \leq \frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 &\leq 2\frac{w^T\mathbf{X}\widehat{\Delta}}{n} + 2\lambda_n\|\theta^*_S\|_1 - \|\theta^*_S + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \\ &\leq 2\left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty \|\widehat{\Delta}\|_1 + 2\lambda_n\left\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right\} \\ &\leq \lambda_n\left\{3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\right\} \end{aligned} \tag{7.30}$$

where step (i) follows from a combination of Holder inequality and the triangle inequality, whereas step (ii) follows from the choice ofλ_n. Inequality (7.30) shows that $\widehat{\Delta} \in \mathbb{C}_3(S)$,so that the RE condition may be applied. Doing so, we obtain $k\|\widehat{\Delta}\|_2^2 \leq 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2$, which implies the claim (7.25a).

Example(7.14)(classical linear Guassian model)The noise vector $w \in \mathbb{R}^n$has i.i.d.$\mathcal{N}(0.\sigma^2)$entries,the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is fixed .Suppose that $\mathbf{X}$ satisfies the RE condition(7.22)and that it is C-column normalized ,meaning that $max_{j=1,\ldots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$,where $X_j \in \mathbb{R}^n$ denotes the jth column of $\mathbf{X}$.

So the random variables$\|\frac{\mathbf{X}^T w}{n}\|_\infty$ corresponds to the absolute maximum of d zero-mean Guassian variables,each with variance at most $\frac{C^2\sigma^2}{n}$.From Exercise2.12 we have

$$\mathbb{P}\left[\|\frac{\mathbf{X}^T w}{n}\|_\infty \geq C\sigma\left(\sqrt{\frac{2logd}{n}}+\delta\right)\right] \leq 2e^{-\frac{n\delta^2}{2}} \qquad \text{for all}\delta > 0.$$

Consequently, if we set $\lambda_n = 2C\sigma\left(\sqrt{\frac{2logd}{n}}+\delta\right)$,then Theorem 7.13(a)implies

$$\|\theta - \theta^*\|_2 \leq \frac{4C\sigma}{k}\sqrt{s}\left\{\sqrt{\frac{2logd}{n}}+\delta\right\}. \tag{7.26}$$

with probabllity at least $1-2e^{-\frac{n\delta^2}{2}}$.
Similarly,we can get other two equality.

Restricted nullspace and eigenvalues for random designs

In practice ,it is difficult to verify that a given design matrix $\mathbf{X}$ satisfies theRE condition(7.22).However,it is possible to give high-probability results in the case of random design matrices.
We here define the maximum diagonal entry $\rho^2(\Sigma)$ of a covariance matrix Σ.

Theorem (7.16)

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

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\theta\|_2^2 - c_2 \rho^2(\Sigma) \frac{log d}{n} \|\theta\|_1^2 \qquad for\ all \theta \in \mathbb{R}^d \qquad (7.31)$$

with probabllity at least $1 - \frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{32}}}$.

- The proof is provided in the Appendix.
- Theorem 7.16 can be used to establish restricted nullspace and eigenvalue conditions for various matrix ensembles that do not satisfy incoherence or RIP conditions.
- When a bound of the form (7.31) holds, it is also possible to prove a more general result on the Lasso error, known as an *oracle inequality.*

We let $\bar{\kappa} := \gamma_{min}(\Sigma)$.

Theorem (7.19)

(Lasso oracle inequality)Under the condition(7.31),consider the Lagrangian Lasso(7.18) with regularization parameter $\lambda_n \geq 2\|\frac{\mathbf{X}w}{n}\|_\infty$*.For any* $\theta^* \in \mathbb{R}^d$*,and optimal solution*$\widehat{\theta}$*satisfies the bound*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{144}{c_1^2}\frac{\lambda_n}{\bar{\kappa}^2}|S| + \frac{16}{c_1}\frac{\lambda_n}{\bar{\kappa}}\|\theta^*_{S^c}\|_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma)}{\bar{\kappa}}\frac{logd}{n}\|\theta^*_{S^c}\|_1^2, \quad (7.32)$$

valid for any subset S with cardinality $|S| \leq \frac{c_1}{64c_2}\frac{\bar{\kappa}}{\rho^2(\Sigma)}\frac{n}{logd}$.

- This result holds without any assumptions whatsoever on the underlying regression vector $\theta^* \in \mathbb{R}^d$and it actually yields a family of upper bounds with a tunable parameter to be optimized.
- An optimal bound is obtained by choosing S to balance these two terms.

Proof:
$\rho^2 := \rho^2(\Sigma)$. Recall the argument leading to the bound (7.30). For a general vector $\theta^* \in \mathbb{R}^d$, the same argument applies with any subset S. Doing so yields that

$$0 \leq \frac{1}{2n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\left\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\right\} \tag{7.33}$$

This inequality implies that the error vector $\hat{\Delta}$satisfies the constraint

$$\|\hat{\Delta}\|_1^2 \leq \left(4\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1\right)^2 \leq 32|S|\|\hat{\Delta}\|_2^2 + 8\|\theta^*_{S^c}\|_1^2. \tag{7.34}$$

Combined with the bound (7.31), we find that

$$\begin{aligned}\frac{\|\mathbf{X}\hat{\Delta}\|_2^2}{n} &\geq c_1\|\sqrt{\Sigma}\hat{\Delta}\|_2^2 - c_2\rho^2(\Sigma)\frac{logd}{n}\|\hat{\Delta}\|_1^2 \\ &\geq \left\{c_1\bar{\kappa} - 32c_2\rho^2|S|\frac{logd}{n}\right\}\|\hat{\Delta}\|_2^2 - 8c_2\rho^2\frac{logd}{n}\|\theta^*_{S^c}\|_1^2 \\ &\geq c_1\frac{\bar{\kappa}}{2}\|\hat{\Delta}\|_2^2 - 8c_2\rho^2\frac{logd}{n}\|\theta^*_{S^c}\|_1^2 \qquad (7.35)\end{aligned}$$

where the second inequality uses $\|\sqrt{\Sigma}\hat{\Delta}\|_2 \geq \bar{\kappa}\|\hat{\Delta}\|_2$ and the final inequality uses the condition $32c_2\rho^2|S|\frac{logd}{n} \leq c_1\frac{\bar{\kappa}}{2}$ by $|S| \leq \frac{c_1}{64c_2}\frac{\bar{\kappa}}{\rho^2(\Sigma)}\frac{n}{logd}$. We split the remainder of the analysis into two cases.

Case1: Suppose $c_1 \frac{\bar{\kappa}}{4}\|\hat{\Delta}\|_2^2 \geq 8c_2\rho^2 \frac{logd}{n}\|\theta^*_{S^c}\|_1^2$.
Combining the bounds (7.35) and (7.33) yields

$$c_1 \frac{\bar{\kappa}}{8}\|\hat{\Delta}\|_2^2 \leq c_1 \frac{\bar{\kappa}}{4}\|\hat{\Delta}\|_2^2 - 4c_2\rho^2 \frac{logd}{n}\|\theta^*_{S^c}\|_1^2 \leq \frac{\|\mathbf{X}\hat{\Delta}\|_2^2}{2n}$$

$$\frac{1}{2n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\left\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\right\} \leq \frac{\lambda_n}{2}\left\{3\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1\right\}$$

$$c_1 \frac{\bar{\kappa}}{8}\|\widehat{\Delta}\|_2^2 \leq \frac{\lambda}{2}\left\{3\sqrt{|S|}\|\widehat{\Delta}\|_2 + 2\|\theta^*_{S^c}\|_1^2\right\} \tag{7.36}$$

This bound involves a quadratic form in $\|\widehat{\Delta}\|_2$; computing the zeros of this quadratic form,we find that

$$\|\widehat{\Delta}\|_2^2 \leq \frac{144}{c_1^2}\frac{\lambda_n}{\bar{\kappa}^2}|S| + \frac{16}{c_1}\frac{\lambda_n^2}{\bar{\kappa}}\|\theta^*_{S^c}\|_1$$

Case2: Support $c_1 \frac{\bar{\kappa}}{4} \|\hat{\Delta}\|_2^2 \leq 8 c_2 \rho^2 \frac{log d}{n} \|\theta^*_{S^c}\|_1^2$.
Then we get

$$\|\widehat{\Delta}\|_2^2 \leq \frac{32 c_2}{c_1} \frac{\rho^2(\Sigma)}{\bar{\kappa}} \frac{log d}{n} \|\theta^*_{S^c}\|_1^2$$

Taking into account both cases, we combine this bound with the earlier inequality (7.36),thereby obtaining the claim (7.32).

Table of Contents

In other applications, the actual value of the regression vector θ^* may not be of primary interst. We might be interested in finding a good predictor, meaning a vector$\widehat{\theta} \in \mathbb{R}^d$ such that the *mean-squared prediction error*.

$$\frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, \widehat{\theta} - \theta^* \rangle)^2 \tag{7.37}$$

is small.

- Suppose that we estimate $\widehat{\theta}$ on the basis of the response vector $y = \mathbf{X}\theta^* + w$. Then receive a "fresh" vector of responses, say $\tilde{y} = \mathbf{X}\theta^* + \widetilde{w}$, where $\tilde{w} \in \mathbb{R}^n$ is a noise vector, with i.i.d. zero-mean entries with variance σ^2.
- We can then measure the quality of our vector $\widehat{\theta}$ by how well it predicts the vector $\tilde{y}$ in terms of squared error, taking averages over instantiations of the noise vector $\tilde{w}$. With the design matrix held fixed,we find that

$$\frac{1}{n}\mathbb{E}\left[\|\tilde{y} - \hat{y}\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|\tilde{y} - \mathbf{X}\widehat{\theta}\|_2^2\right] = \frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2 + \sigma^2,$$

- It is important to note that, at least in general, the problem of finding a good predictor should be easier than estimating θ^*well in l_2-norm.

Theorem (7.20)

(Prediction error bounds) Consider the Lagrangian Lasso (7.18) with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$.
(a)Any optimal solution $\widehat{\theta}$ *satisfies the bound*

$$\frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \leq 12\|\theta^*\|_1 \lambda_n. \tag{7.38}$$

(b)If θ^* *is supported on a subset of cardinality s, and the design matrix satisfies the* $(\kappa; 3)-$*RE condition over S ,then any optimal solution satisfies the bound*

$$\frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{9}{\kappa} s\lambda_n^2. \tag{7.39}$$

Proof: $\widehat{\Delta} := \widehat{\theta} - \theta^*$

(a) We first show that $\|\widehat{\Delta}\|_1 \leq 4\|\theta^*\|_1$under the stated conditions. From the Lagrangian basic inequality (7.29), we have

$$0 \leq \frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq \frac{w^T\mathbf{X}\widehat{\Delta}}{n} + \lambda_n\left\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right\}. \tag{7.42}$$

By Holder inequality and the choice of $\lambda_n(\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty)$, we have

$$\left|\frac{w^T\mathbf{X}\widehat{\Delta}}{n}\right| \leq \left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty \left\|\widehat{\Delta}\right\|_1 \leq \frac{\lambda_n}{2}\left\|\widehat{\Delta}\right\|_1 \leq \frac{\lambda_n}{2}\left\{\|\theta^*\|_1 + \|\widehat{\theta}\|_1\right\},$$

where the final step also uses the triangle inequality. Putting together the pieces yield

$$0 \leq \frac{\lambda_n}{2}\left\{\|\theta^*\|_1 + \|\widehat{\theta}\|_1\right\} + \lambda_n\left\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right\},$$

which (for $\lambda_n > 0$) implies that $\|\widehat{\theta}\|_1 \leq 3\|\theta^*\|_1$.

Consequently, $\|\widehat{\Delta}\|_1 \leq \|\theta^*\|_1 + \|\widehat{\theta}\|_1 \leq 4\|\theta^*\|_1$.

Returning to our earlier inequality (7.42),

$$0 \leq \frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq \frac{w^T\mathbf{X}\widehat{\Delta}}{n} + \lambda_n\left\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right\}.$$

By Holder inequality and the choice of $\lambda_n(\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty)$, we have

$$\left|\frac{w^T\mathbf{X}\widehat{\Delta}}{n}\right| \leq \left\|\frac{\mathbf{X}^T w}{n}\right\|_\infty \left\|\widehat{\Delta}\right\|_1 \leq \frac{\lambda_n}{2}\left\|\widehat{\Delta}\right\|_1,$$

Then we have

$$\begin{aligned}
0 \leq \frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 &\leq \frac{w^T\mathbf{X}\widehat{\Delta}}{n} + \lambda_n\left\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\right\} \\
&\leq \frac{\lambda_n}{2}\left\|\widehat{\Delta}\right\|_1 + \lambda_n\left\{\|\theta^*\|_1 - \|\theta^* + \widehat{\Delta}\|_1\right\} \\
&\leq \frac{3\lambda_n}{2}\left\|\widehat{\Delta}\right\|_1
\end{aligned}$$

where the final equation is based on the triangle inequality bound $\|\theta^* + \widehat{\Delta}\|_1 \geq \|\theta^*\|_1 - \|\widehat{\Delta}\|_1$.

Combined with the upper bound $\|\widehat{\Delta}\|_1 \leq 4\|\theta^*\|_1$,we have

$$0 \leq \frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq 3\lambda_n \left\|\widehat{\Delta}\right\|_1 \leq 12\lambda_n\|\theta^*\|_1$$

(b) In this case, the same argument as in the proof of Theorem 7.13(a) leads to the basic inequality

$$\frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq 3\lambda_n\|\widehat{\Delta_S}\|_1 \leq 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2,$$

Similarly, the proof of Theorem 7.13(a) shows that the error vector $\widehat{\Delta}$ belongs to $\mathbb{C}_3(S)$, whence the $(\kappa;3)$-RE condition can be applied:

$$\kappa\|\widehat{\Delta}\|_2^2 \leq \frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \Rightarrow \|\widehat{\Delta}\|_2 \leq \frac{\|\mathbf{X}\widehat{\Delta}\|_2}{\sqrt{n\kappa}}$$

Combined the upper inequations, we have

$$\frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2 \leq 3\lambda_n\sqrt{s}\frac{\|\mathbf{X}\widehat{\Delta}\|_2}{\sqrt{n\kappa}},$$

Simplify,

$$\frac{1}{\sqrt{n}}\|\mathbf{X}\widehat{\Delta}\|_2 \leq \frac{3\lambda_n\sqrt{s}}{\sqrt{\kappa}},$$

Then $\frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \leq \frac{9}{\kappa}s\lambda_n^2$

Table of Contents

- In other settings, we are interested in somewhat more refined question, namely whether or not a Lasso estimate $\widehat{\theta}$ has non-zero entries in the same positions as the true regression vectorθ^* : *variable selection consistency*.
- Note that it is possible for the l_2- error$\|\widehat{\theta} - \theta^*\|_2$ to be quite small even if $\widehat{\theta}$ and θ^* have different supports.
- On the other hand, given an estimate $\widehat{\theta}$ that correctly recovers the support of θ^*, we can estimate θ^* very well simply by performing an ordinary least-squares regression restricted to this subset.

Variable selection consistency for the Lasso

We introduce the following two conditions:
(A3)*Lower eigenvalue:* The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:

$$\gamma_{min}\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right) \geq c_{min} > 0. \tag{7.43a}$$

(A4) *Mutual incoherence:* There exists some $\alpha \in [0, 1)$ such that

$$max_{j\in S^c}\|(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T X_j\|_1 \leq \alpha. \tag{7.43b}$$

- Assumption (A3) is very mild: in fact, it would be required in order to ensure that the model is identifiable, even if the support set S were known a priori. In particular, the submatrix $\mathbf{X}_S \in \mathbb{R}^{n \times s}$ corresponds to the subset of covariates that are in the support set, so that if assumption (A3) were violated, then the submatrix $\mathbf{X}_S$ would have a non-trivial nullspace, leading to a non-identifiable model.
- Assumption (A4) is a more subtle condition. In order to gain intuition, suppose that we tried to predict the column vector X_j using a linear combination of the columns of $\mathbf{X}_S$. The best weight vector $\widehat{\omega} \in \mathbb{R}^{|S|}$ is given by

$$\widehat{\omega} = \arg \min_{\omega \in \mathbb{R}^{(i)}} \|X_j - \mathbf{X}_S \omega\|_2^2 = \left(\mathbf{X}_S^{\mathrm{T}} \mathbf{X}_S\right)^{-1} \mathbf{X}_S^{\mathrm{T}} X_j,$$

and the mutual incoherence condition is a bound on $\|\omega\|_1$. In the ideal case, if the column space of $\mathbf{X}_S$ were orthogonal to X_j, then the optimal weight vector $\widehat{\omega}$ would be identically zero. In general, we cannot expect this orthogonality to hold, but the mutual incoherence condition (A4) imposes a type of approximate orthogonality.

Theorem (7.21)

Consider an S-sparse linear regression model for which the design matrix satisfies conditions (A3) and (A4). Then for any choice of regularization parameter such that

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| \mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X}) \frac{w}{n} \right\|_\infty, \tag{7.44}$$

the Lagrangian Lasso (7.18) has the following properties:
(a)Uniqueness: There is a unique optimal solution $\widehat{\theta}$.
(b)No false inclusion:This solution has its support set $\widehat{S}$ contained within the true support set S.

(c)$l_\infty-$bounds: The error $\widehat{\theta} - \theta^*$ satisfies

$$\|\widehat{\theta_S} - \theta^*_S\|_\infty \leq \left\| \left(\frac{\mathbf{X}_S^T \mathbf{X}_S}{n} \right)^{-1} \mathbf{X}_S^T \frac{w}{n} \right\|_\infty + \left\| \left| \left(\frac{\mathbf{X}_S^T \mathbf{X}_S}{n} \right)^{-1} \right| \right\|_\infty \lambda_n, \quad (7.45)$$

where $|||A|||_\infty = max_{i=1,\ldots,s} \sum_j |A_{ij}|$ is the matrix $l_\infty-$norm.
(d)*No false exclusion:* The Lasso includes all indices $i \in S$ such that$|\theta_i^*| > B(\lambda_n; \mathbf{X})$, and hence is *variable selection consistent* if $min_{i \in S} |\theta_i^*| > B(\lambda_n; \mathbf{X})$.

Theorem 7.21 is a deterministic result that applies to any set of linear regression equations. It implies more concrete results when we make specific assumptions about the noise vector w.

Corollary (7.22)

Consider the S-sparse linear model based on a noise vector w with zero-mean i.i.d.σ-sub-Gaussian entries, and a deterministic design matrix $\mathbf{X}$ *that satisfies assumptions (A3) and (A4), as well as the C-column normalization condition(*$\frac{max_{j=1,\ldots,d}\|\mathbf{X}_j\|_2}{\sqrt{n}} \leq C$*). Suppose that we solve the Lagrangian Lasso (7.18) with regularization parameter*

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2log(d-s)}{n}} + \delta\right\} \tag{7.46}$$

for some $\delta > 0$*. Then the optimal solution* $\widehat{\theta}$ *is unique with its support contained within S,and satisfies the* $l_\infty-$*error bound*

$$\|\widehat{\theta_S} - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{2logs}{n}} + \delta\right\} + \left|\left|\left|\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right)^{-1}\right|\right|\right|_\infty \lambda_n, \tag{7.47}$$

all with probability at least $1 - 4e^{-\frac{n\delta^2}{2}}$.

Proof:
We first verify that the given choice (7.46) of regularization parameter satisfies the bound (7.44) with high probability.

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| \mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X}) \frac{w}{n} \right\|_\infty, \tag{7.44}$$

For $\left\| \mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X}) \frac{w}{n} \right\|_\infty$:
It suffices to bound the maximum absolute value of the random variables

$$Z_j := X_j^T [\mathbf{I_n} - \mathbf{X_S}(\mathbf{X_S^T X_S})^{-1} \mathbf{X_S^T}](\frac{w}{n}) \qquad \text{for} j \in S^c.$$

Since $\Pi_{S^\perp}(X)$is an orthogonal projection matrix, we have

$$\|\Pi_{S^\perp}(X) X_j\|_2 \leq \|X_j\|_2 \leq C\sqrt{n},$$

where inequality (i) follows from the column normalization assumption. Therefore, each variable Z_j is sub-Gaussian with parameter at most $\frac{C^2\sigma^2}{n}$.

From standard sub-Gaussian tail bounds (Chapter 2), we have

$$\mathbb{P}\left[max_{j\in S^c}|Z_j| \geq t\right] \leq 2(d-s)e^{-\frac{t^2}{2}\frac{n}{C^2\sigma^2}},$$

$$\Leftrightarrow \mathbb{P}\left[\left\|\mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X})\frac{w}{n}\right\|_\infty\right] \leq 2(d-s)e^{-\frac{t^2}{2}\frac{n}{C^2\sigma^2}},$$

Then,

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2log(d-s)}{n}} + \delta\right\} \geq \frac{2}{1-\alpha}\left\|\mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X})\frac{w}{n}\right\|_\infty$$

$$\Leftrightarrow \left\|\mathbf{X}_{S^c}^T \Pi_{S^\perp}(\mathbf{X})\frac{w}{n}\right\|_\infty \leq C\sigma\left\{\sqrt{\frac{2log(d-s)}{n}} + \delta\right\} := t$$

Consequently,the given choice (7.46) of regularization parameter satisfies the bound (7.44) with probability at least $1-4e^{-\frac{n\delta^2}{2}}$.

The only remaining step is to simplify the $l_\infty-$bound (7.45). The second term in this bound is a deterministic quantity, so we focus on bounding the first term $\left\| \left(\frac{\mathbf{X}_S^T \mathbf{X}_S}{n} \right)^{-1} \mathbf{X}_S^T \frac{w}{n} \right\|_\infty$.

For each $i = 1, \ldots, s$, consider the random variable $\tilde{Z}_i := \frac{e_i^T (\frac{1}{n} \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_S^T w}{n}$. Since the elements of the vector w are i.i.d. $\sigma-$sub-Gaussian, the variable $\tilde{Z}_i$is zero-mean and sub-Gaussian with parameter at most

$$\frac{\sigma^2}{n} \left| \left| \left| (\frac{1}{n} X_S^T X_S)^{-1} \right| \right| \right|_2 \leq \frac{\sigma^2}{c_{min} n}$$

,where we have used the eigenvalue condition (7.43a).
Consequently, for any $\delta > 0$, we have

$$\mathbb{P} \left[max_{i=1,\ldots,s} |\tilde{Z}_j| > \frac{\sigma}{\sqrt{c_{min}}} \left\{ \sqrt{\frac{2 log s}{n}} + \delta \right\} \right] \leq 2 e^{-\frac{n\delta^2}{2}}$$

- Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$, we say that $z \in \mathbb{R}^d$ is a subgradient of f at θ, denoted by $z \in \partial f(\theta)$, if we have

$$f(\theta + \Delta) \geq f(\theta) + \langle z, \Delta \rangle \quad \text{for all } \Delta \in \mathbb{R}^d.$$

- When $f(\theta) = \|\theta\|_1$, it can be seen that $z \in \partial\|\theta\|_1$ if and only if $z_j = \text{sign}(\theta_j)$ for all $j = 1, 2, \ldots, d$. Here we allow sign(0) to be any number in the interval $[-1, 1]$.
- In application to the Lagrangian Lasso program (7.18), we say that a pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^d \times \mathbb{R}^d$ is primal-dual optimal if $\widehat{\theta}$ is a minimizer and $\widehat{z} \in \partial\|\widehat{\theta}\|_1$. Any such pair must satisfy the zero-subgradient condition

$$\frac{1}{n}\mathbf{X}^{\mathrm{T}}(\widehat{\mathbf{X}}\widehat{\theta} - y) + \lambda_n \widehat{z} = 0, \tag{7.48}$$

Primal-dual witness (PDW) construction:

- 1 Set $\widehat{\theta}_{S^c} = 0$.
- 2 Determine $\left(\widehat{\theta}_s, \widehat{z}_s\right) \in \mathbb{R}^s \times \mathbb{R}^s$ by solving the oracle subproblem

$$\widehat{\theta}_S \in \arg\min_{\theta_s \in \mathbb{R}^2} \{ \underbrace{\frac{1}{2n} \|y - \mathbf{X}_s \theta_S\|_2^2}_{=:f(\theta_s)} + \lambda_n \|\theta_S\|_1 \}, \tag{7.49}$$

  and then choosing $\widehat{z}_S \in \partial \|\widehat{\theta_S}\|_1$ sunch that $\nabla f(\theta_S)|_{\theta_S = \widehat{\theta_S}} + \lambda_n \widehat{z}_S = 0$.
- 3 Solve for $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ via the zero-subgradient equation (7.48), and check whether or not the strict dual feasibility condition $\|\widehat{Z_{S^c}}\|_\infty < 1$ holds.

We say that the PDW construction succeeds if the vector $\widehat{z}_{S^c}$ constructed in step 3 satisfies the strict dual feasibility condition. The following result shows that this success acts as a witness for the Lasso:

Lemma (7.23)

If the lower eigenvalue condition (A3) holds, then success of the PDW construction implies that the vector $\left(\widehat{\theta}_S, 0\right) \in \mathbb{R}^d$ *is the unique optimal solution of the Lasso.*

Proof: When the PDW construction succeeds, then $\widehat{\theta} = \left(\widehat{\theta}_S, 0\right)$ is an optimal solution with associated subgradient vector $\widehat{z} \in \mathbb{R}^d$ satisfying $\|\widehat{z_{S^c}}\|_\infty < 1$, and $\langle \widehat{z}, \widehat{\theta} \rangle = \|\widehat{\theta}\|_1$. Now let $\widetilde{\theta}$ be any other optimal solution.
$F(\theta) := \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$
Then we are guaranteed that

$$F(\widehat{\theta}) + \lambda_n \langle \widehat{z}, \widehat{\theta} \rangle = F(\widetilde{\theta}) + \lambda_n \|\widetilde{\theta}\|_1$$
$$F(\widehat{\theta}) - \lambda_n \langle \widehat{z}, \widetilde{\theta} - \widehat{\theta} \rangle = F(\widetilde{\theta}) + \lambda_n \left(\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \widetilde{\theta} \rangle \right)$$

By the zero-subgradient conditions (7.48), we have $\lambda_n \widehat{z} = -\nabla F(\widehat{\theta})$, which implies that

$$F(\widehat{\theta}) + \langle \nabla F(\widehat{\theta}), \widetilde{\theta} - \widehat{\theta} \rangle - F(\widetilde{\theta}) = \lambda_n \left(\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \widetilde{\theta} \rangle \right).$$

By convexity of F, the left-hand side is negative, which implies that $\|\widetilde{\theta}\|_1 \leq \langle \widehat{z}, \widetilde{\theta} \rangle$.

But since we also have $\langle \widehat{z}, \widetilde{\theta} \rangle \leq \|\widehat{z}\|_\infty \|\widetilde{\theta}\|_1$, we must have $\|\widetilde{\theta}\|_1 = \langle \widehat{z}, \widetilde{\theta} \rangle$. Since $\|\widehat{z_{S^c}}\|_\infty < 1$, this equality can only occur if $\widetilde{\theta}_j = 0$ for all $j \in S^c$.

Thus, all optimal solutions are supported only on S, and hence can be obtained by solving the oracle subproblem (7.49). Given the lower eigenvalue condition (A3), this subproblem is strictly convex, and so has a unique minimizer.

Proof of Theorem 7.21:

In order to prove Theorem 7.21(a) and (b), it suffices to show that the vector $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ constructed in step 3 satisfies the strict dual feasibility condition.

By construction, the subvectors $\widehat{\theta}_S, \widehat{z}_S$ and $\widehat{z}_S^c$ satisfy the zero-subgradient condition (7.48). By using the fact that $\widehat{\theta}_{S^c} = \theta^*_{S^c} = 0$ and writing out this condition in block matrix form, we obtain

$$\frac{1}{n}\mathbf{X}^T(\mathbf{X}\widehat{\theta} - y) + \lambda_n \widehat{z} = 0, \quad y = \mathbf{X}\theta^* + w$$

$$\frac{1}{n}\mathbf{X}^T(\mathbf{X}(\widehat{\theta} - \theta^*) - w) + \lambda_n \widehat{z} = 0$$

$$\frac{1}{n}\begin{bmatrix} \mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S & \mathbf{X}_S^{\mathrm{T}}\mathbf{X}_{S^c} \\ \mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_S & \mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_{S^c} \end{bmatrix} \begin{bmatrix} \widehat{\theta}_S - \theta^*_S \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathbf{X}_S^{\mathrm{T}} w \\ \mathbf{X}_{S^c}^{\mathrm{T}} w \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z_{S^c}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{7.50}$$

Using the zero-subgradient conditions (7.50), we can obtain

$$\widehat{z}_S^{c\,*} = -\frac{1}{\lambda_n n}\mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_S\left(\widehat{\theta}_S - \theta_S^*\right) + \mathbf{X}_{S^c}^{\mathrm{T}}\left(\frac{w}{\lambda_n n}\right). \tag{7.51}$$

$$\widehat{\theta}_S - \theta_S^* = \left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\mathbf{X}_S^{\mathrm{T}}w - \lambda_n n\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\widehat{z}_s. \tag{7.52}$$

Substituting this expression back into equation (7.51) and simplifying yields

$$\widehat{z}_{S^c} = \underbrace{\mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_S\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\widehat{z}_S}_{\mu} + \underbrace{\mathbf{X}_{S^c}^{\mathrm{T}}\left[I - \mathbf{X}_S\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\mathbf{X}_S^{\mathrm{T}}\right]\left(\frac{w}{\lambda_n n}\right)}_{V_{S^c}}. \tag{7.53}$$

By the triangle inequality, we have $\|\widehat{z}_{S^c}\|_\infty \leq \|\mu\|_\infty + \|V_{S^c}\|_\infty$. By the mutual incoherence condition (7.43b), we have $\|\mu\|_\infty \leq \alpha$. By our choice (7.44) of regularization parameter, we have $\|V_{S^c}\|_\infty \leq \frac{1}{2}(1-\alpha)$. Putting together the pieces, we conclude that $\|\widehat{z}_{S^c}\|_\infty \leq \frac{1}{2}(1+\alpha) < 1$, which establishes the strict dual feasibility condition.

It remains to establish a bound on the ℓ_∞-norm of the error $\widehat{\theta}_S - \theta^*_S$. From equation (7.52) and the triangle inequality, we have

$$\left\|\widehat{\theta}_S - \theta^*_S\right\|_\infty \leq \left\|\left(\frac{\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S}{n}\right)^{-1}\mathbf{X}_S^{\mathrm{T}}\frac{w}{n}\right\|_\infty + \left|\left|\left|\left(\frac{\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S}{n}\right)^{-1}\right|\right|\right|_\infty \lambda_n,$$

which completes the proof.

Thanks for listening