CHAPTER 7

Sparse linear models in high dimensions

The linear model is one of the most widely used in statistics, and has a history dating back to the work of Gauss on least-squares prediction. In its low-dimensional instantiation, in which the number of predictors d is substantially less than the sample size n, the associated theory is classical. In contrast, our aim in this chapter is to develop theory that is applicable to the high-dimensional regime, meaning that it allows for scalings such that $d \asymp n$, or even $d \gg n$. As one might intuitively expect, if the model lacks any additional structure, then there is no hope of obtaining consistent estimators when the ratio d/n stays bounded away from zero.[1] For this reason, when working in settings in which $d > n$, it is necessary to impose additional structure on the unknown regression vector $\theta^* \in \mathbb{R}^d$, and this chapter focuses on different types of sparse models.

7.1 Problem formulation and applications

Let $\theta^* \in \mathbb{R}^d$ be an unknown vector, referred to as the regression vector. 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. This model can also be written in a scalarized form: for each index $i = 1, 2, \ldots, n$, we have $y_i = \langle x_i, \theta^* \rangle + w_i$, where $x_i^T \in \mathbb{R}^d$ is the i^{th} row of $\mathbf{X}$, and y_i and w_i are (respectively) the i^{th} entries of the vectors y and w. The quantity $\langle x_i, \theta^* \rangle := \sum_{j=1}^d x_{ij}\theta_j^*$ denotes the usual Euclidean inner product between the vector $x_i \in \mathbb{R}^d$ of predictors (or covariates), and the regression vector $\theta^* \in \mathbb{R}^d$. Thus, each response y_i is a noisy version of a linear combination of d covariates.

The focus of this chapter is settings in which the sample size n is smaller than the number of predictors d. In this case, it can also be of interest in certain applications to

[1] Indeed, this intuition will be formalized as a theorem in Chapter 15 using information-theoretic methods.

consider a *noiseless linear model*, meaning the special case of equation (7.1) with $w = 0$. When $d > n$, this noiseless model is also interesting, since the observations $y = \mathbf{X}\theta^*$ correspond to an under-determined linear system.

7.1.1 Different sparsity models

At the same time, when $d > n$, it is impossible to obtain any meaningful estimates of θ^* unless the model is equipped with some form of low-dimensional structure. One of the simplest kinds of structure in a linear model is a *hard sparsity* assumption, meaning that the set

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

known as the *support set* of θ^*, has cardinality $s = |S|$ substantially smaller than d. Assuming that the model is exactly supported on s coefficients may be overly restrictive, in which case it is also useful to consider various relaxations of hard sparsity, which we refer to as *weakly sparse* models. Roughly speaking, a vector θ^* is weakly sparse if it can be closely approximated by a sparse vector. There are different ways in which to formalize such an idea, one way being via the ℓ_q-"norms".

For a parameter $q \in [0, 1]$ and radius $R_q > 0$, consider the set

$$\mathbb{B}_q(R_q) = \big\{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_i|^q \leq R_q \big\}, \tag{7.3}$$

which is known as the ℓ_q-ball of radius R_q.

As illustrated in Figure 7-1, for $q \in [0, 1)$, it is not a ball in the strict sense of the word, since it is a non-convex set. In the special case $q = 0$, any vector $\theta^* \in \mathbb{B}_0(R_0)$ can have at most $s = R_0$ non-zero entries. In the more general setting $q \in (0, 1]$, membership in $\mathbb{B}_q(R_q)$ has different interpretations. One of them involves how quickly the ordered coefficients

$$\underbrace{|\theta^*_{(1)}|}_{\max_{j=1,2,\ldots,d} |\theta^*_j|} \geq |\theta^*_{(2)}| \geq \cdots \geq |\theta^*_{(d-1)}| \geq \underbrace{|\theta^*_{(d)}|}_{\min_{j=1,2,\ldots,d} |\theta^*_j|} \tag{7.4}$$

decay. More precisely, as we explore in Exercise 7.2, if these ordered coefficients satisfy the bound $|\theta^*_{(j)}| \leq C\, j^{-\alpha}$ for a suitable exponent α, then θ^* belongs to $\mathbb{B}_q(R_q)$ for a radius R_q depending on (C, α).

7.1.2 Applications of sparse linear models

Although quite simple in appearance, the high-dimensional linear model is fairly rich. We illustrate here with some examples and applications.

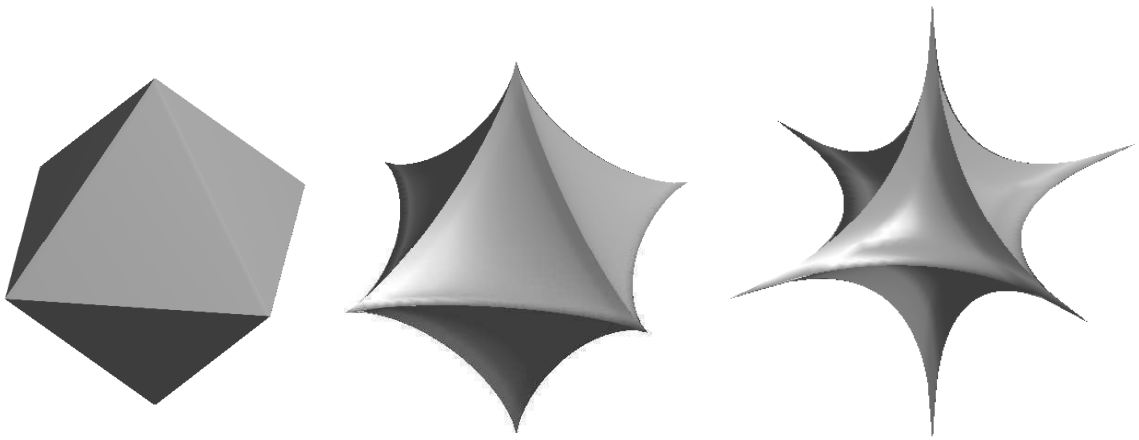


Figure 7-1. Illustrations of the ℓ_q "balls" for different choices of the parameter $q \in (0, 1]$. (a) For $q = 1$, the set $\mathtt{B}_1(R_q)$ corresponds to the usual ℓ_1 ball shown here. (b) For $q = 0.75$, the ball is a non-convex set obtained by collapsing the faces of the ℓ_1-ball towards to origin. (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 $\mathtt{B}_q(1)$ is star-shaped around the origin.

Example 7.1 (Gaussian sequence model)**.** In a finite-dimensional version of the Gaussian sequence model, we make observations of the form

$$y_i = \sqrt{n}\theta_i^* + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{7.5}$$

where $w_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. noise variables. This model is a special case of the general linear regression model (7.1) with $n = d$, and a design matrix $\mathbf{X} = \sqrt{n}\mathbf{I}_n$. It is a truly high-dimensional model, since the sample size n is equal to the number of parameters d. Although it appears simple on the surface, it is a suprisingly rich model: indeed, many problems in non-parametric estimation, among them regression and density estimation, can be reduced to an "equivalent" instance of the Gaussian sequence model, in the sense that the optimal rates for estimation are the same under both models. For non-parametric regression, when the function f belongs a certain type of function class (known as a Besov space), then the vector of its wavelet coefficients belongs to a certain type of ℓ_q-ball with $q \in (0, 1)$, so that the estimation problem corresponds to a version of the Gaussian sequence problem with an ℓ_q-sparsity constraint. Various methods for

estimation, such as wavelet thresholding, exploit this type of approximate sparsity. See the bibliographic section for additional references on this connection. ♣

Example 7.2 (Signal denoising in orthonormal bases)**.** Sparsity plays an important role in signal processing, both for compression and denoising of signals. In abstract terms, a signal can be represented as a vector $\beta^* \in \mathbb{R}^d$. Depending on the application, the signal length d could represent the number of pixels in an image, or the number of discrete samples of a time series. In a denoising problem, one observes a corrupted set of samples of the form $\widetilde{y}_i = \beta_i^* + \widetilde{w}_i$, where $\widetilde{w}_i$, $i = 1, 2, \ldots, d$ are some kind of additive noise. Based on the observation vector $\widetilde{y} \in \mathbb{R}^n$, the goal is to "denoise" the signal, meaning to reconstruct β^* as accurately as possible. In a compression problem, the goal is to produce a representation of β^*, either exact or approximate, that is "smaller" than its original representation.

Many classes of signals exhibit sparsity when transformed into an appropriate basis, such as a wavelet basis. This sparsity can be exploited both for compression and denoising. In abstract terms, any such transform can be represented as an orthogonal matrix $\Psi \in \mathbb{R}^{d \times d}$, so that $\theta^* := \Psi^T \beta^* \in \mathbb{R}^d$ corresponds to the vector of transform coefficients. If the vector θ^* is known to be sparse, then it can be compressed by retaining only some number $s < d$ of its coefficients, say the largest s in absolute value. Of course, if θ^* were exactly sparse, then this representation would be exact. It is more realistic to assume that θ^* satisfies some form of approximate sparsity, and as we explore in Exercise 7.2, such conditions can be used to provide guarantees on the accuracy of the reconstruction.

Returning to the denoising problem, in the transformed space, the observation model takes the form $y = \theta^* + w$, where $y := \Psi^T \widetilde{y}$ and $w := \Psi^T \widetilde{w}$ are the transformed observation and noise vector, respectively. When the observation noise is assumed to be i.i.d. Gaussian (and hence invariant under orthogonal transformation), then both the original and the transformed observations are instances of the Gaussian sequence model with $n = d$ from the preceding example. If θ^* is known to be sparse, then it is natural to consider estimators based on thresholding. In particular, for a threshold $\lambda > 0$ to be chosen, the hard-thresholded estimate of θ^* is defined as

$$[H_\lambda(y)]_i = \begin{cases} y_i & \text{if } |y_i| \geq \lambda, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{7.6}$$

A closely related technique is the soft-thresholded estimate given by

$$[T_\lambda(y)]_i = \begin{cases} \text{sign}(y_i)\left(|y_i| - \lambda\right) & \text{if } |y_i| \geq \lambda, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{7.7}$$

As we explore in Exercise 7.1, each of these estimators have interpretations as mini-

mizing the quadratic loss $\|y - \theta\|_2^2$ subject to ℓ_0 and ℓ_1 constraints, respectively. ♣

Example 7.3 (Lifting and non-linear functions)**.** Despite its superficial appearance as representing purely linear functions, augmenting the set of predictors allows for non-linear models to be represented by the standard equation (7.1). As an example, let us consider polynomial functions in a scalar variable $t \in \mathbb{R}$ of degree k, say of the form

$$f_\theta(t) = \theta_1 + \theta_2 t + \ldots + \theta_{k+1} t^k.$$

Suppose that we observe n samples of the form (y_i, t_i), where $y_i = f_\theta(t_i) + w_i$, for $i = 1, 2, \ldots, n$. This problem can be converted into an instance of the linear regression model by using the sample points $(t_1, \ldots, t_n)$ to define the $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^k \end{bmatrix}.$$

When expressed in this lifted space, the polynomial functions are linear in θ, and so we can write the observations (y_i, t_i) in the standard vector form $y = \mathbf{X}\theta + w$.

This lifting procedure is not limited to polynomial functions. The more general setting is to consider functions that are linear combinations of some set of basis functions—say of the form

$$f_\theta(t) = \sum_{j=1}^{b} \theta_j \phi_j(t),$$

where $\{\phi_1, \ldots, \phi_b\}$ are some known functions. Given n observation pairs (y_i, t_i), this model can also be reduced to the form $y = \mathbf{X}\theta + w$, where the design matrix $\mathbf{X}$ has entries $X_{ij} = \phi_j(t_i)$.

Although the preceding discussion has focused on univariate functions, the same ideas apply to multivariate functions, say in D dimensions. Returning to case of polynomial functions, we note that there are $\binom{D}{k}$ possible multinomials of degree k in dimension D. This leads to the model dimension growing exponentially as D^k, so that sparsity assumptions become essential in order to produce manageable classes of models. ♣

Example 7.4 (Signal compression in overcomplete bases)**.** We now return to an extension of the signal processing problem introduced in Example 7.2. As we observed previously, many classes of signals exhibit sparsity when represented in an appropriate basis, such as a wavelet basis, and this sparsity can be exploited for both compression and denoising purposes. Given a signal $y \in \mathbb{R}^n$, classical approaches to signal denoising and compression are based on orthogonal transformations, where the basis functions

are represented by the columns of an orthonormal matrix $\Psi \in \mathbb{R}^{n \times n}$. However, it can be useful to consider an *overcomplete* set of basis functions, represented by the columns of a matrix $X \in \mathbb{R}^{n \times d}$ with $d > n$. Within this framework, signal compression can be performed by finding a vector $\theta \in \mathbb{R}^d$ such that $y = \mathbf{X}\theta$. Since $\mathbf{X}$ has rank n, we can always find a solution with at most n non-zero co-ordinates, but the hope is to find a solution $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_0 = s \ll n$ non-zeros.

Problems involving ℓ_0-constraints are computationally intractable, so that it is natural to consider relaxations. As we will discuss at more length later in the chapter, the ℓ_1-relaxation has proven very successful. In particular, one seeks a sparse solution by solving the basis pursuit program

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \underbrace{\sum_{j=1}^{d} |\theta_j|}_{\|\theta\|_1} \qquad \text{such that } y = \mathbf{X}\theta.$$

Later sections of the chapter will provide theory under which the solution to this ℓ_1-relaxation is equivalent to the original ℓ_0-problem. ♣

Example 7.5 (Compressed sensing)**.** Compressed sensing is based on the combination of ℓ_1-relaxation with the random projection method (see Example 2.6 from Chapter 2). It is motivated by the inherent wastefulness of the classical approach to exploiting sparsity for signal compression. As previously described in Example 7.2, given a signal $\beta^* \in \mathbb{R}^d$, the standard approach is to first compute the full vector $\theta^* = \Psi^T \beta^* \in \mathbb{R}^d$ of transform coefficients, and then to *discard* all but the top s coefficients. Is there a more direct way of estimating β^*, without pre-computing the full vector θ^* of its transform coefficients?

The compressed sensing approach is to take $n \ll d$ random projections of the original signal $\beta^* \in \mathbb{R}^d$, each of the form $y_i = \langle x_i, \beta^* \rangle := \sum_{j=1}^d x_{ij} \beta_j^*$, where $x_i \in \mathbb{R}^d$ is a random vector. Various choices are possible, including the standard Gaussian ensemble ($x_{ij} \sim \mathcal{N}(0,1)$, i.i.d.), or the Rademacher ensemble ($x_{ij} \in \{-1, +1\}$, i.i.d.). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a measurement matrix with x_i as its i^{th} row and $y \in \mathbb{R}^n$ be the concatenated set of random projections. In matrix-vector notation, the problem of exact reconstruction amounts to finding a solution of the under-determined linear system $y = \mathbf{X}\beta^*$ such that $\Psi^T \beta$ is as sparse as possible. The standard ℓ_1-relaxation of this problem takes the form $\min_{\beta \in \mathbb{R}^d} \|\Psi^T \beta\|_1$ such that $y = \mathbf{X}\beta$, or equivalently, in the transform domain,

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \qquad \text{such that } y = \widetilde{\mathbf{X}}\theta, \tag{7.8}$$

where $\widetilde{\mathbf{X}} := \mathbf{X}\Psi$. In asserting this equivalence, we have used the relation $\beta^* = \Psi^T \theta^*$. This is another instance of the basis pursuit LP with a random design matrix $\widetilde{\mathbf{X}}$.

Compressed sensing has become a very popular approach to recovering sparse sig-

nals, with a variety of applications. Later in the chapter, we will develop theory that guarantees the success of ℓ_1-relaxation for the random design matrices that arise from taking random projections.

♣

Example 7.6 (Selection of Gaussian graphical models)**.** Any zero-mean Gaussian random vector $(Z_1, \ldots, Z_d)$ has a density of the form

$$p_{\Theta^*}(z_1, \ldots, z_d) = \frac{1}{\sqrt{(2\pi)^d \det((\Theta^*)^{-1})}} \exp\big(-\frac{1}{2} z^T \Theta^* z\big),$$

where $\Theta^* \in \mathbb{R}^{d \times d}$ is the inverse covariance matrix, also known as the precision matrix. For many interesting models, the precision matrix is sparse, with relatively few non-zero entries. The problem of Gaussian graphical model selection, as discussed at more length in Chapter 11, is to infer the non-zero entries in the matrix Θ^*.

This problem can be reduced to an instance of sparse linear regression as follows. For a given index $s \in V := \{1, 2, \ldots, d\}$, suppose that we are interested in recovering its neighborhood, meaning the subset $\mathcal{N}(s) := \{t \in V \mid \Theta^*_{st} \neq 0\}$. In order to do so, imagine performing a linear regression of the variable Z_s on the $(d-1)$-dimensional vector $Z_{\backslash\{s\}} := \big\{z_t, t \in V \backslash \{s\}\big\}$. As we explore in Exercise 11.3 in Chapter 11, we can write

$$\underbrace{Z_s}_{\text{Response } y} = \langle \underbrace{Z_{\backslash\{s\}}}_{\text{Predictors}}, \theta^* \rangle + w_s,$$

where w_s is a zero-mean Gaussian variable, independent of the vector $Z_{\backslash\{s\}}$. Moreover, the vector $\theta^* \in \mathbb{R}^{d-1}$ has the same sparsity pattern as the s^{th} off-diagonal row $\big(\Theta^*_{st}, t \in V \backslash \{s\}\big)$ of the precision matrix. ♣

■ 7.2 Recovery in the noiseless setting

In order to build intuition, we begin by focusing on the simplest case in which the observations are perfect or noiseless. More concretely, we are given the linear equation $y = \mathbf{X}\theta^*$, where $y \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ are known, and $\theta^* \in \mathbb{R}^d$ is unknown. When $d > n$, this is an *underdetermined* set of linear equations, so that there is a whole subspace of solutions. But what if we are told that there is a sparse solution, meaning that $\theta^* \in \mathbb{R}^d$ has at most $s \ll d$ non-zero entries? In this setting, our goal is to find the sparsest solution to an underdetermined set of linear equations. This noiseless problem has applications in signal representation and compression, as discussed in Examples 7.4 and 7.5.

7.2.1 ℓ_1-based relaxation

This problem can be cast as a (nonconvex) optimization problem involving the ℓ_0-"norm". Let us define

$$\|\theta\|_0 := \sum_{j=1}^{d} \mathbb{I}[\theta_j \neq 0],$$

where the function $t \mapsto \mathbb{I}[t \neq 0]$ is equal to one if $t \neq 0$, and zero otherwise. Strictly speaking, this is not a norm, but it serves to counts the number of non-zero entries in the vector $\theta \in \mathbb{R}^d$. We now consider the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \qquad \text{such that } \mathbf{X}\theta = y. \tag{7.9}$$

If we could solve this problem, then we would obtain a solution to the linear equations that has the fewest number of non-zero entries.

But how to solve the problem (7.9)? Although the constraint set is simply a subspace, the cost function is non-differentiable and non-convex. The most direct approach would be to search exhaustively over subsets of the columns of $\mathbf{X}$. In particular, for each subset $S \subset \{1, \ldots, d\}$, we could form the matrix $\mathbf{X}_S \in \mathbb{R}^{n \times |S|}$ consisting of the columns of $\mathbf{X}$ indexed by S, and then examine the linear system $y = \mathbf{X}_S\, \theta$ to see whether or not it had a solution $\theta \in \mathbb{R}^{|S|}$. If we iterated over subsets in increasing cardinality, then the first solution found would be the sparsest solution. Let's now consider the associated computational cost. If the sparsest solution contained s non-zero entries, then we would have to search over at least $\sum_{j=1}^{s-1} \binom{d}{j}$ subsets before finding it. But the number of such subsets grows exponentially in s, so the procedure would not be computationally feasible for anything except toy problems.

Given the computational difficulties associated with ℓ_0-minimization, a natural strategy is to replace the troublesome ℓ_0-objective by the nearest convex member of the ℓ_q-family, namely the ℓ_1-norm. This is an instance of a *convex relaxation*, in which a non-convex optimization problem is approximated by a convex program. In this setting, doing so leads to the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \qquad \text{such that } \mathbf{X}\theta = y. \tag{7.10}$$

Unlike the ℓ_0-version, this is now a convex program, since the constraint set is a subspace (hence convex), and the cost function is piecewise linear and thus convex as well. More precisely, the problem (7.10) is a linear program, since any piecewise linear cost can always be re-formulated as the maximum of a collection of linear functions. We refer to the optimization problem (7.10) as the *basis pursuit linear program*, after Chen, Donoho and Saunders [CDS98].

7.2.2 Exact recovery and restricted nullspace

We now turn to an interesting theoretical question: when is solving the basis pursuit program (7.10) equivalent to solving the original ℓ_0-problem (7.9)? More concretely, let us suppose that there is a vector $\theta^* \in \mathbb{R}^d$ such that $y = \mathbf{X}\theta^*$, and moreover, the vector θ^* has support $S \subset \{1, 2, \ldots, d\}$, meaning that $\theta^*_j = 0$ for all $j \in S^c$ (where S^c denotes the complement of S). Intuitively, the success of basis pursuit should depend on how the nullspace of $\mathbf{X}$ is related to this support. To make this concrete, let us define the subset

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

corresponding to the cone of vectors whose ℓ_1-norm off the support is dominated by the ℓ_1-norm on the support. We now define a property that links the nullspace of a matrix $\mathbf{X}$ to this set:

Definition 7.1. The matrix $\mathbf{X}$ satisfies the *restricted nullspace property* with respect to S if $\mathbb{C}(S) \cap \text{null}(\mathbf{X}) = \{0\}$.

The set $\mathbb{C}(S)$ is relevant because, as shown in the proof of Theorem 7.1 to follow, the difference $\widehat{\Delta} = \widehat{\theta} - \theta^*$ between the basis pursuit solution $\widehat{\theta}$ and the unknown vector $\theta^* \in \mathbb{R}^d$ is always contained within it. In fact, more deeply, the restricted nullspace property is equivalent to the success of the basis pursuit LP in the following sense:

Theorem 7.1. The following two properties are equivalent:

(a) For any vector $\theta^* \in \mathbb{R}^d$ with support S, the basis pursuit program (7.10) has unique solution $\widehat{\theta} = \theta^*$.

(b) The matrix $\mathbf{X}$ satisfies the restricted nullspace property with respect to S.

Proof. We first show that $(b) \Longrightarrow (a)$. 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 $\widehat{\Delta} = \widehat{\theta} - \theta^*$, 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 facts that $\theta^*_{S^c} = 0$, and applied the triangle inequality. Rearranging this inequality, we conclude that the error $\widehat{\Delta} \in \mathbb{C}(S)$. However, by construc-

tion, we also have $\mathbf{X}\widehat{\Delta} = 0$, so $\widehat{\Delta} \in \text{null}(\mathbf{X})$ as well. By our assumption, this implies that $\widehat{\Delta} = 0$, or that $\widehat{\theta} = \theta^*$.

In order to establish the implication $(a) \Longrightarrow (b)$, it suffices to show that if the ℓ_1-relaxation succeeds for all S-sparse vectors, then the set $\text{null}(\mathbf{X})\backslash\{0\}$ has no intersection with $\mathbb{C}(S)$. For a given vector $\theta^* \in \text{null}(\mathbf{X})\backslash\{0\}$, 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 $\widehat{\beta} = \begin{bmatrix} \theta^*_S & 0 \end{bmatrix}^T$. Since $\mathbf{X}\theta^* = 0$ by assumption, the vector $\begin{bmatrix} 0 & -\theta^*_{S^c} \end{bmatrix}^T$ 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. □

■ 7.2.3 Sufficient conditions for restricted nullspace

In order for Theorem 7.1 to be a useful result in practice, one requires a certificate that the restricted nullspace property holds. The earliest sufficient conditions were based on the incoherence parameter of the design matrix, namely the quantity

$$\delta_{\text{PW}}(\mathbf{X}) := \max_{j \neq k} \Big| \frac{|\langle X_j, X_k \rangle|}{n} \Big| \tag{7.12}$$

where X_j denotes the j^{th} column of $\mathbf{X}$, and $\mathbb{I}[j = k]$ denotes the 0-1-valued indicator for the event $\{j = k\}$. Here we have chosen to rescale matrix columns by $1/\sqrt{n}$, as it makes results for random designs more readily interpretable.

The following result shows that a small pairwise incoherence is sufficient to guarantee a uniform version of the restricted nullspace property.

Proposition 7.1. If the pairwise incoherence satisfies the bound

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

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

We guide the reader through the steps involved in the proof of this claim in Exercise 7.3. Moreover, as we explore in this same exercise, the pairwise incoherence bound (7.13) holds with high probability for sub-Gaussian random matrices with i.i.d. elements as long as $n = \Omega(s^2 \log d)$. As an immediate corollary, if we choose such a random matrix $\mathbf{X}$ as our design, then we are assured that the basis pursuit LP will exactly recover for all vectors with sparsity at most s.

A related but more sophisticated sufficient condition is the restricted isometry property (RIP). It can be understood as a natural generalization of the pairwise incoherence condition, based on looking at conditioning of larger subsets of columns.

Definition 7.2. For each $s = 1, \ldots, d$, the restricted isometry constant of $\mathbf{X} \in \mathbb{R}^{n \times d}$ of order s is the smallest quantity $\delta_s(\mathbf{X}) > 0$ such that

$$\|\!| \frac{\mathbf{X}_S^T \mathbf{X}_S}{n} - \mathbf{I}_s \|\!|_{\mathrm{op}} \leq \delta_s(\mathbf{X}) \quad \text{for all subsets } S \text{ of size at most } s. \tag{7.14}$$

In this definition, we recall that $\|\!| \cdot |\!\|_{\mathrm{op}}$ denotes the ℓ_2-operator norm of a matrix, corresponding to its maximum singular value. For $s = 1$, the RIP condition implies that the rescaled columns of $\mathbf{X}$ are near unit-norm—that is, we are guaranteed that $\frac{\|X_j\|_2^2}{n} \in [1 - \delta_1, 1 + \delta_1]$ for all $j = 1, 2, \ldots, d$. For $s = 2$, the RIP constant δ_2 is very closely related to the pairwise incoherence parameter $\delta_{\mathrm{PW}}(\mathbf{X})$. This connection is most apparent when the matrix $\mathbf{X}/\sqrt{n}$ has unit-norm columns, in which case, for any pair of columns $\{j, k\}$, we have

$$\frac{\mathbf{X}_{\{j,k\}}^T \mathbf{X}_{\{j,k\}}}{n} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{\|X_j\|_2^2}{n} - 1 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & \frac{\|X_k\|_2^2}{n} - 1 \end{bmatrix} \stackrel{(i)}{=} \begin{bmatrix} 0 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & 0 \end{bmatrix},$$

where the final equality (i) uses the column normalization condition. Consequently, we find that

$$\delta_2(\mathbf{X}) = \|\!| \frac{\mathbf{X}_{\{j,k\}}^T \mathbf{X}_{\{j,k\}}}{n} - \mathbf{I}_2 \|\!|_{\mathrm{op}} = \max_{j \neq k} |\frac{\langle X_j, X_k \rangle}{n}| \; = \; \delta_{\mathrm{PW}}(\mathbf{X}).$$

More generally, as we show in Exercise 7.4, for any matrix $\mathbf{X}$ and sparsity level $s \in \{2, \ldots, d\}$, we have the sandwich relation

$$\delta_{\mathrm{PW}}(\mathbf{X}) \stackrel{(i)}{\leq} \delta_s(\mathbf{X}) \stackrel{(ii)}{\leq} s \, \delta_{\mathrm{PW}}(\mathbf{X}), \tag{7.15}$$

and neither bound can be improved in general. (We also show that there exist matrices for which $\delta_s(\mathbf{X}) = \sqrt{s} \, \delta_{\mathrm{PW}}(\mathbf{X})$.) Although RIP imposes constraints on much larger sub-matrices than pairwise incoherence, the magnitude of the constraints required to guarantee the uniform restricted nullspace property can be milder.

The following result shows that suitable control on the RIP constants implies that the restricted nullspace property holds:

Proposition 7.2. If the RIP constant of order $2s$ satisfies $\delta_{2s} < 1/3$, then the uniform restricted nullspace property holds for any subset S of cardinality $|S| \leq s$.

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

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

We frequently use the fact that $\|\widetilde{\theta}_A\| = \|\theta_A\|$ for any elementwise separable norm, such as the ℓ_1 or ℓ_2 norms.

Let S_0 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 $\|\theta_{S^c}\|_1 > \|\theta_{S_0}\|_1$ for 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 $\widetilde{\theta}_{S^c}$; 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 $\theta = \widetilde{\theta}_{S_0} + \sum_{k \geq 1} \widetilde{\theta}_{S_k}$.

The RIP property guarantees that $\|\widetilde{\theta}_{S_0}\|_2^2 \leq \frac{1}{1-\delta_{2s}} \|\mathbf{X}\widetilde{\theta}_{S_0}\|_2^2$. Moreover, since $\theta \in \text{null}(\mathbf{X})$, we have $\mathbf{X}\widetilde{\theta}_{S_0} = -\sum_{j \geq 1} \mathbf{X}\widetilde{\theta}_{S_j}$, and hence

$$\|\widetilde{\theta}_{S_0}\|_2^2 \leq \frac{1}{1-\delta_{2s}} \Big| \sum_{j \geq 1} \frac{\langle \mathbf{X}\widetilde{\theta}_{S_0}, \mathbf{X}\widetilde{\theta}_{S_j} \rangle}{n} \Big| \stackrel{(i)}{=} \frac{1}{1-\delta_{2s}} \Big| \sum_{j \geq 1} \widetilde{\theta}_{S_0} \big[\frac{\mathbf{X}^T\mathbf{X}}{n} - I \big] \widetilde{\theta}_{S_j} \Big|,$$

where equality (i) uses the fact that $\langle \widetilde{\theta}_{S_0}, \widetilde{\theta}_{S_j} \rangle = 0$.

By the RIP property, for each $j \geq 1$, the $\ell_2 \to \ell_2$ operator norm satisfies the bound $\|\!| n^{-1} \mathbf{X}_{S_0 \cup S_j}^T \mathbf{X}_{S_0 \cup S_j} - I \|\!|_2 \leq \delta_{2s}$, and hence we have

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

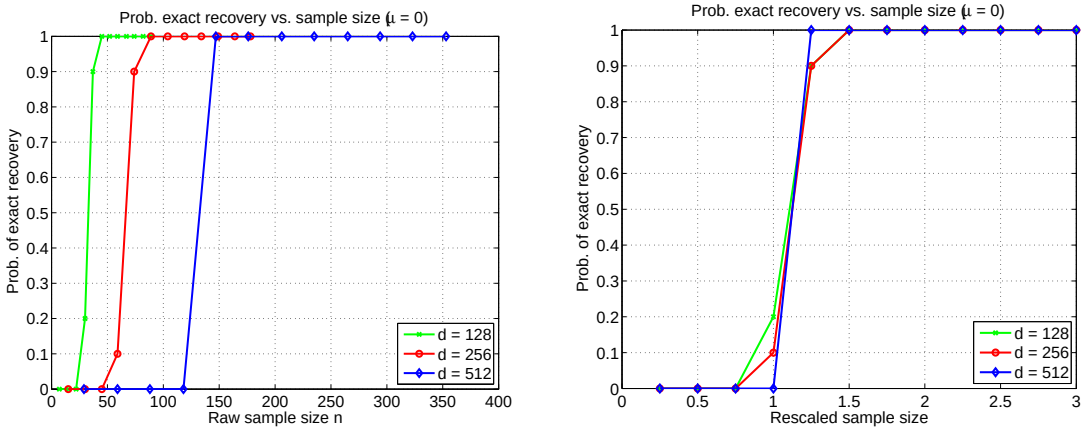
where we have canceled out a factor of $\|\widetilde{\theta}_{S_0}\|_2$ from each side. Finally, by construction of the sets S_j, for each $j \geq 1$, we have $\|\widetilde{\theta}_{S_j}\|_\infty \leq \frac{1}{s} \|\widetilde{\theta}_{S_{j-1}}\|_1$, which implies that $\|\widetilde{\theta}_{S_j}\|_2 \leq \frac{1}{\sqrt{s}} \|\widetilde{\theta}_{S_{j-1}}\|_1$. Applying these upper bounds to the inequality (7.16), we obtain

$$\|\widetilde{\theta}_{S_0}\|_1 \leq \sqrt{s}\, \|\widetilde{\theta}_{S_0}\|_2 \;\leq\; \frac{\delta_{2s}}{(1-\delta_{2s})} \Big\{ \|\widetilde{\theta}_{S_0}\|_1 + \sum_{j \geq 1} \|\widetilde{\theta}_{S_j}\|_1 \Big\},$$

or equivalently $\|\widetilde{\theta}_{S_0}\|_1 \leq \frac{\delta_{2s}}{(1-\delta_{2s})} \Big\{ \|\widetilde{\theta}_{S_0}\|_1 + \|\widetilde{\theta}_{S^c}\|_1 \Big\}$. Some simple algebra verifies that this inequality implies that $\|\widetilde{\theta}_{S_0}\|_1 < \|\widetilde{\theta}_{S^c}\|_1$ as long as $\delta_{2s} < 1/3$. □

Like the pairwise incoherence constant, control on the RIP constants is a sufficient

conditions for the basis pursuit LP to succeed. A major advantage of the RIP approach is that for various classes of random design matrices, of particular interest in compressed sensing (see Example 7.5), 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. As we explore in Exercise 7.7, for sub-Gaussian random matrices with i.i.d. elements, the pairwise incoherence is bounded by $\frac{1}{3s}$ with high probability as long as $n \gtrsim s^2 \log d$. By contrast, this same exercise also shows that the RIP constants for certain classes of random design matrices $\mathbf{X}$ are well-controlled as long as $n \gtrsim s \log(d/s)$. Consequently, the RIP approach overcomes the "quadratic barrier"—namely, the requirement that the sample size n scale quadratically in the sparsity s, as in the pairwise incoherence approach.



(a) (b)

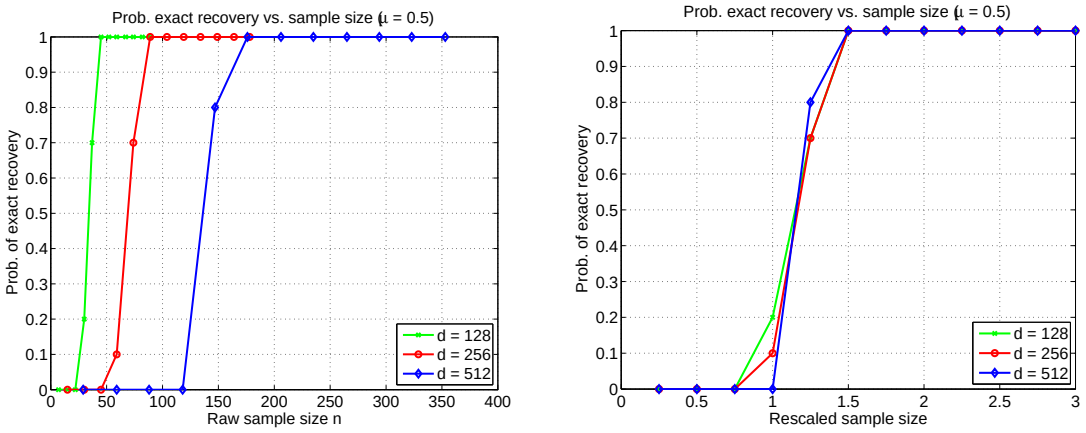
Figure 7-2. (a) Probability of basis pursuit success versus the raw sample size n for random design matrices drawn with i.i.d. $\mathcal{N}(0,1)$ entries. Each curve corresponds to a different problem size $d \in \{128, 256, 512\}$ with sparsity $s = \lceil 0.1d \rceil$. (b) Same results re-plotted versus the rescaled sample size $n/\big(s \log(d/s)\big)$. Curves exhibit a phase transition at the same value of this rescaled sample size.

It should be noted that, unlike the restricted nullspace property, neither the pairwise incoherence condition nor the RIP condition are necessary conditions. Indeed, the basis pursuit LP succeeds for many classes of matrices for which both pairwise incoherence and RIP conditions are violated. For example, consider a random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with i.i.d. rows $X_i \sim \mathcal{N}(0, \Sigma)$. Letting $\mathbb{1} \in \mathbb{R}^d$ denote the all-ones vector, consider the family of covariance matrices

$$\Sigma := (1 - \mu)\mathbf{I}_d + \mu \, \mathbb{1}\mathbb{1}^T, \tag{7.17}$$

for a parameter $\mu \in [0, 1)$. In Exercise 7.8, we show that for any fixed $\mu \in (0, 1)$, the pairwise incoherence bound (7.13) is violated with high probability for large s, and

moreover, that the condition number of any $2s$-sized subset grows at the rate $\mu\sqrt{s}$ with high probability, so that the RIP constants will (w.h.p.) grow unboundedly as $s \to +\infty$ for any fixed $\mu \in (0, 1)$. Nonetheless, for any $\mu \in [0, 1)$, the basis pursuit LP relaxation still succeeds with high probability with sample size $n \gtrsim s \log(d/s))$, as illustrated in Figure 7-3. Later in the chapter, we provide a result on random matrices that allows



(a) (b)

Figure 7-3. (a) Probability of basis pursuit success versus the raw sample size n for random design matrices drawn with i.i.d. rows $X_i \sim \mathcal{N}(0, \Sigma)$, where $\mu = 0.5$ in the model (7.17). Each curve corresponds to a different problem size $d \in \{128, 256, 512\}$ with sparsity $s = \lceil 0.1d \rceil$. (b) Same results re-plotted versus the rescaled sample size $n/\big(s \log(d/s)\big)$. Curves exhibit a phase transition at the same value of this rescaled sample size.

for direct verification of the restricted nullspace property for various families, including (among others) the family (7.17). See Theorem 7.3 and the associated discussion for further details.

■ 7.3 Estimation in noisy settings

Let us now turn to the noisy setting, in which we observe the pair $(y, \mathbf{X}) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$ linked by the observation model $y = \mathbf{X}\theta^* + w$. The new ingredient here is the noise vector $w \in \mathbb{R}^n$. A natural extension of the basis pursuit program is based on minimizing a weighted combination of the data-fidelity term $\|y - \mathbf{X}\theta\|_2^2$ with the ℓ_1-norm penalty, say of the form

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

Here $\lambda_n > 0$ is a *regularization parameter* to be chosen by the user. Following Tibshirani [Tib96], we refer to it as the *Lasso program*.

Alternatively, one can consider different constrained forms of the Lasso, that is either

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

for some radius $R > 0$, or

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

for some noise tolerance $b > 0$. The constrained version (7.20) is referred to as *relaxed basis pursuit* by Chen et al. [CDS98]. By Lagrangian duality theory, all three families of convex programs are equivalent. More precisely, for any choice of radius $R > 0$ in the constrained variant (7.19), there is a regularization parameter $\lambda \geq 0$ such that solving the Lagrangian version (7.18) is equivalent to solving the constrained version (7.19). Similar statements apply to choices of $b > 0$ in the constrained variant (7.20).

■ 7.3.1 Restricted eigenvalue condition

In the noisy setting, we can no longer expect to achieve perfect recovery. Instead, we focus on bounding the ℓ_2 error $\|\widehat{\theta} - \theta^*\|_2$ between a Lasso solution $\widehat{\theta}$ and the unknown regression vector θ^*. In the presence of noise, we require a condition that is closely related to but slightly stronger than the restricted nullspace property—namely, that the restricted eigenvalues of $\frac{\mathbf{X}^T\mathbf{X}}{n}$ are lower bounded over a cone. In particular, for a constant $\alpha \geq 1$, let us define the set

$$\mathbb{C}_\alpha(S) := \big\{ \Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1 \big\}. \tag{7.21}$$

This definition generalizes the set $\mathbb{C}(S)$ used in our definition of the restricted nullspace property, which corresponds to the special case $\alpha = 1$.

Definition 7.3. The matrix $\mathbf{X}$ satisfies the *restricted eigenvalue* (RE) condition over S with parameters (κ, α) if

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

Note that the RE condition is a strengthening of the restricted nullspace property. In particular, if the RE condition holds with parameters $(\kappa, 1)$ for any $\kappa > 0$, then the restricted nullspace property holds. Moreover, we will prove that under the RE condition, the error $\|\widehat{\theta} - \theta^*\|_2$ in the Lasso solution is well-controlled.

From where does the need for the RE condition arise? To provide some intuition, let us consider the constrained version (7.19) of the Lasso, with radius $R = \|\theta^*\|_1$. With this setting, the true parameter vector θ^* is feasible for the problem. By definition, the Lasso estimate $\widehat{\theta}$ minimizes the quadratic cost function $\mathcal{L}_n(\theta) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$ over the

ℓ_1-ball of radius R. As the amount of data increases, we expect that θ^* should become a near-minimizer of the same loss, so that $\mathcal{L}_n(\widehat{\theta}) \approx \mathcal{L}_n(\theta^*)$. But when does closeness in the loss imply that the error vector $\Delta := \widehat{\theta} - \theta^*$ is small? As illustrated in Figure 7-4,

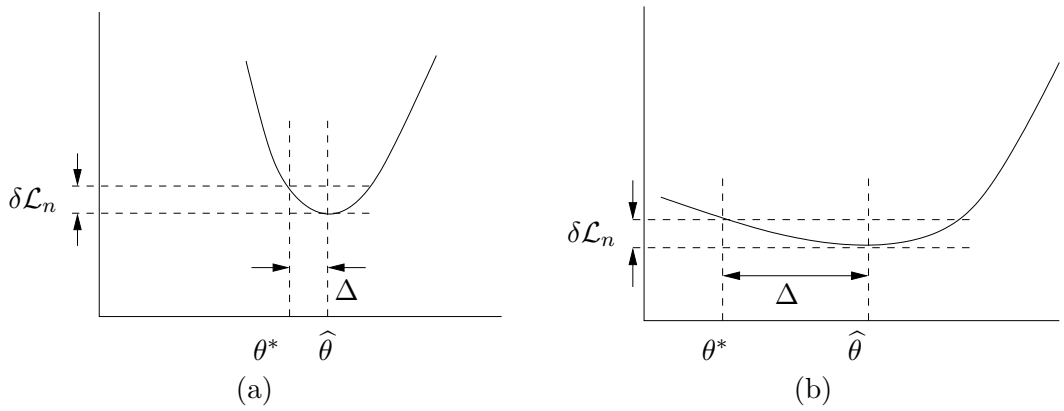


Figure 7-4. Illustration of the connection between curvature (strong convexity) of the loss function, and estimation error. (a) In a favorable setting, the loss function is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ in the loss implies that the error vector $\Delta = \widehat{\theta} - \theta^*$ is not too large. (b) In an unfavorable setting, the loss is very flat, so that a small loss difference $\delta\mathcal{L}_n$ need not imply small error.

the link between the excess loss $\mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ and the size of the error $\Delta = \widehat{\theta} - \theta^*$ is controlled by the curvature of the loss function. In the favorable setting of panel (a), the loss has a high curvature around its optimum $\widehat{\theta}$, so that a small excess loss $\delta\mathcal{L}_n$ implies that the error vector Δ is small. This curvature no longer holds for the loss function in panel (b), for which it is possible that $\delta\mathcal{L}_n$ could be small while the error Δ is relatively large.

Figure 7-4 illustrates a 1-dimensional function, in which case the curvature can be captured by a scalar. For a function in d dimensions, the curvature of a loss function is captured by the structure of its Hessian matrix $\nabla^2\mathcal{L}_n(\theta)$, which is a symmetric positive semi-definite matrix. In the special case of the quadratic loss that underlies the Lasso, the Hessian is easily calculated

$$\nabla^2\mathcal{L}_n(\theta) = \frac{1}{n}\mathbf{X}^T\mathbf{X}. \tag{7.23}$$

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

$$\frac{\|\mathbf{X}\Delta\|_2^2}{n} \geq \kappa\,\|\Delta\|_2^2 > 0 \qquad \text{for all } \Delta \in \mathbb{R}^d\backslash\{0\}, \tag{7.24}$$

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

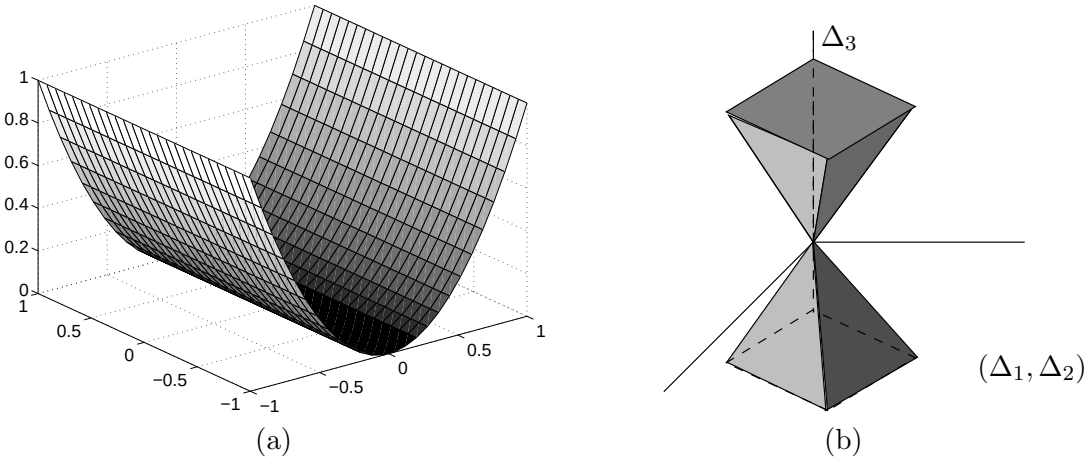


Figure 7-5. (a) A convex loss function in high-dimensional settings (with $d \gg n$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. (b) The Lasso error $\widehat{\Delta}$ must lie in the restricted subset $\mathtt{C}_\alpha(S)$ of $\mathtt{R}^d$. For this reason, it is only necessary that the loss function be curved in certain directions of space.

In the high-dimensional setting with $d > n$, this Hessian is a $d \times d$ matrix with rank at most n, so that it is impossible to guarantee that it has a positive curvature in all directions. Rather, the quadratic loss always has the form illustrated in Figure 7-5(a): although it may be curved in some directions, there is always a $d - n$ dimensional subspace of directions in which it is completely flat! Consequently, the uniform lower bound (7.24) is never satisfied. For this reason, we need to relax the stringency of the uniform curvature condition, and require that it only holds for a subset $\mathbb{C}_\alpha(S)$ of vectors, as illustrated in Figure 7-5(b). If we can be assured that the subset $\mathbb{C}_\alpha(S)$ is well-aligned with the curved directions of the Hessian, then a small excess loss will translate into bounds on the difference between $\widehat{\theta}$ and θ^*.

■ 7.3.2 Bounds on ℓ_2-error for hard sparse models

With this intuition in place, we now state a result that provides a bound on the error $\|\widehat{\theta} - \theta^*\|_2$ in the case of a "hard sparse" vector θ^*. In particular, let us impose the following conditions:

(A1) The vector θ^* 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 $(\kappa, 3)$.

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

Theorem 7.2. Under assumptions (A1) and (A2):

(a) Any solution of the Lagrangian Lasso (7.18) with regularization parameter $\lambda_n \geq 2\big\|\frac{\mathbf{X}^T w}{n}\big\|_\infty$ satisfies the bound

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

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

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

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

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{8}{\kappa}\sqrt{s}\,\big\|\frac{\mathbf{X}^T w}{n}\big\|_\infty + \frac{2}{\sqrt{\kappa}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}. \tag{7.25c}$$

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

In order to develop intuition for these claims, we first discuss them at a high level, and then illustrate them with some concrete examples. First, it is important to note that these results are deterministic, and apply to any set of linear regression equations. As stated, however, the results involve unknown quantities stated in terms of w and/or θ^*. Obtaining results for specific statistical models—as determined by assumptions on the noise vector w and/or the design matrix—involves bounding or approximating these quantities. Based on our earlier discussion of the role of strong convexity, it is natural that all three upper bounds are inversely proportional to the restricted eigenvalue constant $\kappa > 0$. 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 in the bound involve the unknown noise vector, either via the quantity $\|\frac{\mathbf{X}^T w}{n}\|_\infty$ in parts (a) and (b), or via $\frac{\|w\|_2^2}{n}$ in part (c).

Let us illustrate some concrete consequences of Theorem 7.2 for some linear regression models that are commonly used and studied.

Example 7.7 (Classical linear Gaussian model)**.** We begin with the classical linear Gaussian model from statistics, for which the observation noise $w \in \mathbb{R}^n$ is Gaussian, with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Let us consider the case of deterministic design, meaning that that 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 $\max_{j=1,\ldots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$. (Here we use

$X_j \in \mathbb{R}^n$ to denote the j^{th} column of $\mathbf{X}$.) With this set-up, the random variable $\big\|\frac{\mathbf{X}^T w}{n}\big\|_\infty$ corresponds to the absolute maximum of d zero-mean Gaussian variables, each with variance at most $\frac{C^2\sigma^2}{n}$. Consequently, from standard Gaussian tail bounds (Exercise 2.12), we have

$$\mathbb{P}\Big[\big\|\frac{\mathbf{X}^T w}{n}\big\|_\infty \geq C\,\sigma\,\big(\sqrt{\frac{2\log d}{n}} + \delta\big)\Big] \leq 2\,e^{-\frac{n\delta^2}{2}} \quad \text{for all } \delta > 0.$$

Consequently, if we set $\lambda_n = 2\,C\,\sigma\,\big(\sqrt{\frac{2\log d}{n}} + \delta\big)$, then part (a) of Theorem 7.2 implies that any optimal solution of the Lagrangian Lasso (7.18) satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{6\,C\,\sigma}{\kappa}\,\sqrt{s}\,\big\{\sqrt{\frac{2\log d}{n}} + \delta\big\} \tag{7.26}$$

with probability at least $1-2e^{-\frac{n\delta^2}{2}}$. Similarly, part (b) implies that any optimal solution of the constrained Lasso (7.19) satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4\,C\,\sigma}{\kappa}\,\sqrt{s}\,\big\{\sqrt{\frac{2\log d}{n}} + \delta\big\} \tag{7.27}$$

with the same probability. Apart from constant factors, these two bounds are equivalent. Perhaps the most significant difference is the constrained Lasso (7.19) assumes exact knowledge of the ℓ_1-norm $\|\theta^*\|_1$, whereas the Lagrangian Lasso only requires knowledge of the noise variance σ^2. In practice, it is relatively straightfoward to estimate the noise variance, whereas the ℓ_1-norm is a more delicate object.

Turning to part (c), given the Gaussian noise vector w, the rescaled variable $\frac{\|w\|_2^2}{\sigma^2 n}$ is χ^2 with n degrees of freedom. From Example 2.5, we have

$$\mathbb{P}\Big[\big|\frac{\|w\|_2^2}{n} - \sigma^2\big| \geq \sigma^2\delta\Big] \leq 2e^{-n\delta^2/8}, \qquad \text{for all } \delta \in (0,1).$$

Consequently, part (c) implies that any optimal solution of the relaxed basis pursuit program (7.20) with $b^2 = \frac{\sigma^2}{2}(1+\delta)$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{8C\,\sigma}{\kappa}\sqrt{s}\,\big\{\sqrt{\frac{2\log d}{n}} + \delta\big\} + \frac{2\,\sigma}{\sqrt{\kappa}}\,\sqrt{\delta} \qquad \text{for all } \delta \in (0,1),$$

with probability at least $1 - 4e^{-\frac{n\delta^2}{8}}$. ♣

Example 7.8 (Compressed sensing)**.** In the domain of compressed sensing, the design matrix $\mathbf{X}$ can be chosen by the user, and one standard choice is the standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Suppose that the noise vector $w \in \mathbb{R}^n$ is determin-

istic, say with bounded entries ($\|w\|_\infty \leq \sigma$.) Under these assumptions, each variable $X_j^T w/\sqrt{n}$ 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 the Lasso estimates will again satisfy the bounds (7.26) and (7.27), this time with $C = 1$. Similarly, if we set $b^2 = \frac{\sigma^2}{2}$, then the relaxed basis pursuit program (7.19) will satisfy the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{8\,\sigma}{\kappa}\sqrt{s}\Big\{\sqrt{\frac{2\log d}{n}} + \delta\Big\} + \frac{2\,\sigma}{\sqrt{\kappa}}$$

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

With these examples in hand, we now turn to the proof of Theorem 7.2.

Proof. Part (b): We begin by proving the error bound (7.25b) for the constrained Lasso (7.19). Given the choice $R = \|\theta^*\|_1$, the target vector θ^* 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 Hölder's inequality to the right-hand side yields $\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leq 2\big\|\frac{\mathbf{X}^T w}{n}\big\|_\infty \|\widehat{\Delta}\|_1$. As shown in the proof of Theorem 7.1, whenever $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1$ 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 $\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \geq \kappa\|\widehat{\Delta}\|_2^2$. Putting together the pieces yields the claimed bound.

Part (c): Next we prove the error bound (7.25c) for the relaxed basis pursuit (RBP) program. 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 vector θ^* 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

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

Re-arranging yields the modified basic inequality

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

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

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

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

Part (a): Finally, we prove the bound (7.25a) for the Lagrangian Lasso (7.18). 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

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

Re-arranging yields the *Lagrangian basic inequality*

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

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

$$\begin{aligned} 0 \leq \frac{1}{n}\|\mathbf{X}\widehat{\Delta}\|_2^2 \;\; &\leq \;\; 2\,\frac{w^T\mathbf{X}\widehat{\Delta}}{n} + 2\lambda_n\big\{\|\theta^*_S\|_1 - \|\theta^*_S + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\} \\ &\stackrel{(i)}{\leq} \;\; 2\|\mathbf{X}^T w/n\|_\infty\|\widehat{\Delta}\|_1 + 2\lambda_n\big\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\} \\ &\stackrel{(ii)}{\leq} \;\; \lambda_n\big\{3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\}, \end{aligned} \tag{7.30}$$

where step (i) follows from a combination of Hölder's and the triangle inequality, whereas step (ii) follows from the choice of λ_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 $\kappa\|\widehat{\Delta}\|_2^2 \leq 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2$, which implies the claim (7.25a). □

7.3.3 Restricted nullspace and eigenvalues for random designs

Theorem 7.2 is based on assuming that the design matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition (7.22). In practice, it is difficult to verify that a given design matrix $\mathbf{X}$ satisfies this condition. Indeed, developing methods to "certify" design matrices in this way is one line of on-going research. 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.

The following theorem provides a result along these lines. It involves the maximum diagonal entry $\rho^2(\Sigma)$ of a covariance matrix Σ.

Theorem 7.3. 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

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

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

Remark: The proof of this result is provided in the Appendix. It makes use of techniques discussed in other chapters, including the Gordon-Slepian inequalities (Chapters 5 and Chapter 6) and concentration of measure for Gaussian functions (Chapter 2). Concretely, we show that the bound (7.31) holds with $c_1 = \frac{1}{8}$ and $c_2 = 50$, but sharper constants can be obtained with a more careful argument. It can be shown (Exercise 7.11) that a lower bound of the form (7.31) implies that an RE condition (and hence a restricted nullspace condition) holds over $\mathbb{C}_3(S)$, uniformly over all subsets of cardinality $|S| \leq \frac{c_1}{32 c_2} \frac{\lambda_{\min}(\Sigma)}{\rho^2(\Sigma)} \frac{n}{\log d}$.

Theorem 7.3 can be used to establish restricted nullspace and eigenvalue conditions for various matrix ensembles that do not satisfy incoherence or RIP conditions. Let us consider a few examples to illustrate.

Example 1 (Geometric decay)**.** Consider a covariance matrix with the Toeplitz structure $\Sigma_{ij} = \nu^{|i-j|}$ for some parameter $\nu \in [0, 1)$. This type of geometrically decaying covariance structure arises naturally from autoregressive processes, where the parameter ν allows for tuning of the memory in the process. By classical results on eigenvalues of

Toeplitz matrices, we have $\lambda_{\min}(\Sigma) \geq (1-\nu)^2 > 0$ and $\rho^2(\Sigma) = 1$, independently of the dimension d. Consequently, Theorem 7.3 implies that, with high probability, the sample covariance matrix $\widehat{\Sigma} = \frac{\mathbf{X}^T\mathbf{X}}{n}$ obtained by sampling from this distribution will satisfy the RE condition for all subsets S of cardinality at most $|S| \leq \frac{c_1}{32c_2}(1-\nu)^2 \frac{n}{\log d}$. This provides an example of a matrix family with substantial correlation between covariates for which the RE property still holds. ♣

We now consider a matrix family with an even higher amount of dependency among the covariates.

Example 2 (Spiked identity model)**.** For a parameter $\mu \in [0, 1)$, recall the spiked identity family (7.17) of covariance matrices. For this family, we have $\lambda_{\min}(\Sigma) = 1 - \mu$ and $\rho^2(\Sigma) = 1$, again independent of the dimension. Consequently, Theorem 7.3 implies that with high probability, the sample covariance based on i.i.d. draws from this ensemble satisfies the RE and RN conditions uniformly over all subsets of cardinality at most $|S| \leq \frac{c_1}{32c_2}(1-\mu)\frac{n}{\log d}$.

Check the fraction constant here.

However, for any $\mu \neq 0$, the spiked identity matrix is very poorly conditioned, and also has poorly conditioned sub-matrices. This fact implies that both the pairwise incoherence and restricted isometry property will be violated w.h.p., regardless of how large the sample size is taken. To see this, for an arbitrary subset S of size s, consider the associated $s \times s$ submatrix of Σ, which we denote by Σ_{SS}. The maximal eigenvalue of Σ_{SS} scales as $1 + \mu(s-1)$, which diverges as s increases for any fixed $\mu > 0$. As we explore in Exercise 7.8, this fact implies that both pairwise incoherence and RIP will be violated with high probability. ♣

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*. 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. The flexibility in tuning this parameter is akin to that of an oracle, which would have access to the ordered coefficients of θ^*. In order to minimize notational clutter, we introduce the convenient shorthand notation $\bar{\kappa} := \lambda_{\min}(\Sigma)$.

Theorem 7.4 (Lasso oracle inequality)**.** Under the condition (7.31), consider the Lagrangian Lasso (7.18) with regularization parameter $\lambda_n \geq 2\|\mathbf{X}^T w/n\|_\infty$. For any $\theta^* \in \mathbb{R}^d$, any optimal solution $\widehat{\theta}$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \underbrace{\frac{144}{c_1^2}\frac{\lambda_n^2}{\bar{\kappa}^2}|S|}_{\text{Estimation error}} + \underbrace{\frac{16}{c_1}\frac{\lambda_n}{\bar{\kappa}}\|\theta^*_{S^c}\|_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma)}{\bar{\kappa}}\frac{\log d}{n}\|\theta^*_{S^c}\|_1^2}_{\text{Approximation error}}. \quad (7.32)$$

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

Note that inequality (7.32) actually provides a family of upper bounds, one for each valid choice of the subset S. The optimal choice of S is based on trading off the two sources of error. The first term grows linearly with the cardinality $|S|$, and corresponds to the error associated with estimating a total of $|S|$ unknown coefficients. The second term corresponds to approximation error, and depends on the unknown regression vector via the tail sum $\|\theta^*_{S^c}\|_1 = \sum_{j \notin S} |\theta^*_j|$. An optimal bound is obtained by choosing S to balance these two terms. We illustrate an application of this type of trade-off in Exercise 7.12.

Proof. Throughout the proof, we use ρ^2 as a shorthand for $\rho^2(\Sigma)$. 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 except that additional terms involving $\|\theta^*_{S^c}\|_1$ must be tracked. Doing so yields that

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

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

$$\|\widehat{\Delta}\|_1^2 \leq \big(4\|\widehat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1\big)^2 \;\leq 32\,|S|\,\|\widehat{\Delta}\|_2^2 + 8\|\theta^*_{S^c}\|_1^2. \tag{7.34}$$

Combined with the bound (7.31), we find that

$$\begin{aligned}\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} &\geq \big\{c_1\bar{\kappa} - 32c_2\rho^2|S|\frac{\log d}{n}\big\}\|\widehat{\Delta}\|_2^2 - 8c_2\rho^2\frac{\log d}{n}\|\theta^*_{S^c}\|_1^2 \\ &\geq c_1\frac{\bar{\kappa}}{2}\,\|\widehat{\Delta}\|_2^2 - 8c_2\rho^2\frac{\log d}{n}\|\theta^*_{S^c}\|_1^2,\end{aligned} \tag{7.35}$$

where the final inequality uses the condition $32c_2\rho^2|S|\frac{\log d}{n} \leq c_1\frac{\bar{\kappa}}{2}$. We split the remainder of the analysis into two cases.

Case 1: First suppose that $c_1\,\frac{\bar{\kappa}}{4}\|\widehat{\Delta}\|_2^2 \geq 8c_2\rho^2\frac{\log d}{n}\|\theta^*_{S^c}\|_1^2$. Combining the bounds (7.35) and (7.33) yields

$$c_1\,\frac{\bar{\kappa}}{4}\|\widehat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2}\big\{3\sqrt{|S|}\|\widehat{\Delta}\|_2 + 2\|\theta^*_{S^c}\|_1\big\}. \tag{7.36}$$

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

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

Case 2: Otherwise, we must have $c_1 \frac{\bar{\kappa}}{4} \|\widehat{\Delta}\|_2^2 < 8 c_2 \rho^2 \frac{\log d}{n} \|\theta^*_{S^c}\|_1^2$.

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

7.4 Bounds on prediction error

In the previous analysis, we have focused exclusively on the problem of parameter recovery, either in noiseless or noisy settings. In other applications, the actual value of the regression vector θ^* may not be of primary interest; rather, we might be interested in finding a good predictor, meaning a vector $\widehat{\theta} \in \mathbb{R}^d$ such that *mean-squared prediction error*

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

is small. To understand why the quantity (7.37) is a measure of prediction error, suppose that we estimate $\widehat{\theta}$ on the basis of the response vector $y = \mathbf{X}\theta^* + w$. Suppose that we then receive a "fresh" vector of responses, say $\widetilde{y} = \mathbf{X}\theta^* + \widetilde{w}$, where $\widetilde{w} \in \mathbb{R}^n$ 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 $\widetilde{y}$ in terms of squared error, taking averages over instantiations of the noise vector $\widetilde{w}$. Following some algebra, we find that

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

so that apart from the constant additive factor of σ^2, the quantity (7.37) measures how well we can predict a new vector of responses (with the design matrix held fixed).

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. Indeed, the prediction problem does not require that θ^* even be identifiable: unlike in parameter recovery, the problem can still be solved if two columns of the design matrix $\mathbf{X}$ are identical.

Theorem 7.5 (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 S and the design matrix satisfies the $(\kappa; 3)$-RE condition over S, then any optimal solution satisfies the bound

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

Remarks: As previously discussed in Example 7.7, when the noise vector w has i.i.d. zero-mean σ-sub-Gaussian entries and the design matrix is C-column normalized, the choice $\lambda_n = 2\,C\,\sigma\,\big(\sqrt{\frac{2\log d}{n}} + \delta\big)$ is valid with probability at least $1 - 2e^{-\frac{n\delta^2}{2}}$. In this case, part (a) implies the upper bound

$$\frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \leq 24\,\|\theta^*\|_1\,C\,\sigma\,\big(\sqrt{\frac{2\log d}{n}} + \delta\big) \tag{7.40}$$

with the same high probability. For this bound, the requirements on the design matrix are extremely mild—only the column normalization condition $\max_{j=1,\ldots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$. Thus, the matrix X could have many identical columns, and this would have no effect on the prediction error. In fact, when the only constraint on θ^* is the ℓ_1-norm bound $\|\theta^*\|_1 \leq R$, then the bound (7.40) is unimprovable (see the bibliographic section for further discussion).

On the other hand, when θ^* is $|S|$-sparse and in addition, the design matrix satisfies an RE condition, then part (b) guarantees the bound

$$\frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{72}{\kappa} C^2\,\sigma^2\,\big(\frac{2|S|\log d}{n} + |S|\delta^2\big) \tag{7.41}$$

with the same high probability. This error bound can be significantly smaller than the $\sqrt{\frac{\log d}{n}}$ error bound (7.40) guaranteed under weaker assumptions. For this reason, the bounds (7.38) and (7.39) are often referred to as the *slow rates* and *fast rates*, respectively, for prediction error.

Proof. Throughout the proof, we adopt the usual notation $\widehat{\Delta} = \widehat{\theta} - \theta^*$ for the error vector.

(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

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

By Hölder's inequality and our choice of λ_n, we have

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

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

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

which (for $\lambda_n > 0$) implies that $\|\widehat{\theta}\|_1 \leq 3\|\theta^*\|_1$. Consequently, a final application of the triangle inequality yields $\|\widehat{\Delta}\|_1 \leq \|\theta^*\|_1 + \|\widehat{\theta}\|_1 \leq 4\|\theta^*\|_1$, as claimed.

We can now complete the proof. Returning to our earlier inequality (7.42), we have

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{2n} \leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_1 + \lambda_n\big\{\|\theta^*\|_1 - \|\theta^* + \widehat{\Delta}\|_1\big\} \stackrel{(i)}{\leq} \frac{3\lambda_n}{2}\|\widehat{\Delta}\|_1$$

where step (i) is based on the triangle inequality bound $\|\theta^* + \widehat{\Delta}\|_1 \geq \|\theta^*\|_1 - \|\widehat{\Delta}\|_1$. Combined with the upper bound $\|\widehat{\Delta}\|_1 \leq 4\|\theta^*\|_1$, the proof is complete.

(b) In this case, the same argument as the proof of Theorem 7.2(a) leads to the basic inequality

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leq 3\lambda_n\|\widehat{\Delta}_S\|_1 \;\leq\; 3\lambda_n\sqrt{s}\|\widehat{\Delta}\|_2.$$

Similarly, the proof of Theorem 7.2(a) shows that the error vector $\widehat{\Delta}$ belongs to $\mathbb{C}_3(S)$, whence the $(\kappa; 3)$-RE condition can be applied, this time to the right-hand side of the basic inequality. Doing so yields $\|\widehat{\Delta}\|_2^2 \leq \frac{1}{\kappa}\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n}$, and hence that $\frac{\|\mathbf{X}\widehat{\Delta}\|_2}{\sqrt{n}} \leq \frac{3}{\sqrt{\kappa}}\sqrt{s}\,\lambda_n$, as claimed. □

■ 7.5 Variable or subset selection

Thus far, we have focused on results that guarantee that either the ℓ_2-error or the prediction error of the Lasso is small. In other settings, we are interested in a 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 vector θ^*. More precisely, suppose that the true regression vector θ^* is s-sparse, meaning that it is supported on a subset $S(\theta^*)$

of cardinality $s = |S(\theta^*)|$. In such a setting, a natural goal is to correctly identify the subset $S(\theta^*)$ of relevant variables. In terms of the Lasso, we ask the following question: given an optimal Lasso solution $\widehat{\theta}$, when is its support set—denoted by $S(\widehat{\theta})$—exactly equal to the true support $S(\theta^*)$? We refer to this property as *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, as long as $\widehat{\theta}$ is non-zero for all "suitably large" entries of θ^*, and not too large in positions where θ^* is zero. On the other hand, as we discuss in the sequel, given an estimate $\widehat{\theta}$ that correctly recovers the support of θ^*, we can estimate θ^* very well (in ℓ_2-norm, or other metrics) simply by performing an ordinary least-squares regression restricted to this subset.

■ 7.5.1 Variable selection consistency for the Lasso

We begin by addressing the issue of variable selection in the context of deterministic design matrices $\mathbf{X}$. (Such a result can be extended to random design matrices, albeit with additional effort.) It turns out that variable selection requires some assumptions that are related to but distinct from the restricted eigenvalue condition (7.22). In particular, consider the following conditions:

(A3) *Lower eigenvalue:* The smallest eigenvalue of the sample covariance sub-matrix indexed by S is bounded below:

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

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

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

To provide some intuition, the first condition (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. To gain intuition for it, note that for an index j in the complement $S^c = \{1, 2, \ldots, d\} \backslash S$, the quantity

$$\|X_j^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1}\|_1$$

is a measure of how well the column X_j aligns with the sub-matrix $\mathbf{X}_S$. In the most desirable case, the columns $\{X_j, j \in S^c\}$ would all be orthogonal to the columns of $\mathbf{X}_S$,

and we would be guaranteed that $\alpha = 0$. Of course, in the high-dimensional setting ($d \gg n$), this complete orthogonality is not possible, but one can still hope for a type of "near orthogonality" to hold.

The following result applies to the Lagrangian Lasso (7.18) when applied to an instance of the linear observation model such that the true parameter θ^* is supported on a subset S with cardinality s. In order to state the result, we introduce the convenient shorthand $\Pi_{S^\perp}(\mathbf{X}) = \mathbf{I}_n - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$, a type of orthogonal projection matrix.

Theorem 7.6. Suppose that the design matrix satisfies conditions (A3) and (A4). Then for any choice of regularization parameter

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

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 $\widehat{S}$ contained within the true support S.

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

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\big\|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\mathbf{X}_S^T\frac{w}{n}\big\|_\infty + \|\!|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\|\!|_\infty\,\lambda_n}_{B(\lambda_n;\mathbf{X})}. \tag{7.45}$$

(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 $\min_{i\in S}|\theta_i^*| > B(\lambda_n;\mathbf{X})$.

Before proving this result, let us try to interpret its main claims. First, the uniqueness claim in part (a) is not trivial in the high-dimensional setting, because as discussed previously, although the Lasso objective is convex, it can never be strictly convex when $d > n$. Based on the uniqueness claim, we can talk unambiguously about the support of the Lasso estimate $\widehat{\theta}$. Part (b) guarantees that the Lasso does not falsely include variables that are not in the support of θ^*, or equivalently that $\widehat{\theta}_{S^c} = 0$, whereas part (d) is a consequence of the sup-norm bound from part (c): as long as the minimum value of $|\theta_i^*|$ over indices $i \in S$ is not too small, then the Lasso is variable-selection consistent in the full sense.

As with our earlier result (Theorem 7.2) on ℓ_2-error bounds, Theorem 7.6 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, as we

show here.

Corollary 7.1. Consider an instance of the linear model in which the noise vector w has zero-mean i.i.d. σ-sub-Gaussian entries, and the deterministic design matrix $\mathbf{X}$ satisfies assumptions (A3) and (A4), as well as the C-column normalization ($\max_{j=1,\ldots,d} \|X_j\|_2/\sqrt{n} \leq C$). Suppose that we solve the Lagrangian Lasso (7.18) with regularization parameter

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

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

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\Big\{\sqrt{\frac{2\log s}{n}} + \epsilon\Big\} + \Big|\!\Big|\!\Big|\Big(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\Big)^{-1}\Big|\!\Big|\!\Big|_\infty \lambda_n, \tag{7.47}$$

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

Proof. We first verify that the given choice (7.46) of regularization parameter satisfies the bound (7.44) with high probability. It suffices to bound the maximum absolute value of the random variables

$$Z_j := X_j^T \underbrace{\Big[I - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\Big]}_{\Pi_{S^\perp}(\mathbf{X})}\big(\frac{w}{n}\big), \quad \text{for } j \in S^c.$$

Since $\Pi_{S^\perp}(\mathbf{X})$ is an orthogonal projection matrix, we have $\|\Pi_{S^\perp}(\mathbf{X})\, X_j\|_2 \leq \|X_j\|_2 \leq C\sqrt{n}$, where we have used the column normalization assumption. Therefore, each variable Z_j is sub-Gaussian with parameter at most $C^2\,\sigma^2/n$. From standard tail bounds (Chapter 2), we have

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

from which we see that the choice (7.46) satisfies the bound (7.44) with the claimed probability.

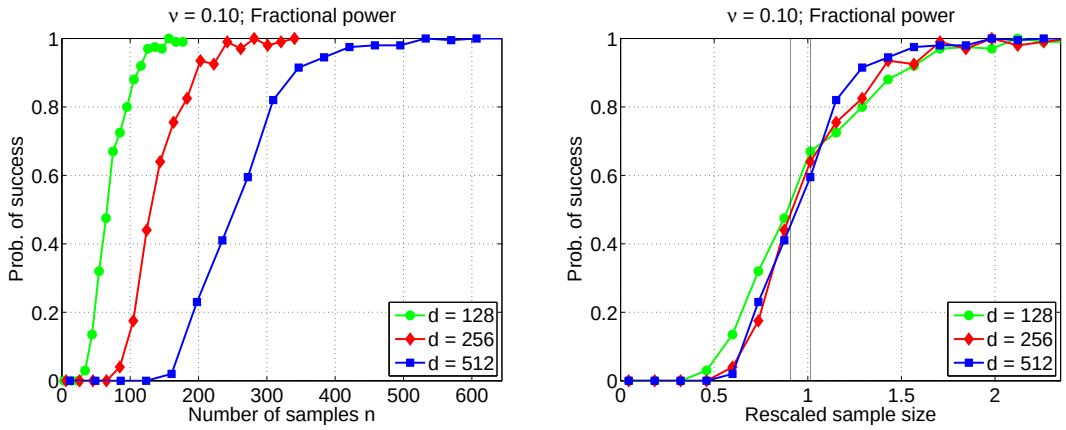
The only remaining step is to control the first (stochastic) term in the ℓ_∞-bound (7.45). The second term is a deterministic quantity, so that it remains to bound the first term. For each $i = 1, \ldots, s$, consider the random variable $\widetilde{Z}_i := e_i^T\big(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S\big)^{-1}\mathbf{X}_S^T w/n$. Since the elements of w are i.i.d. σ-sub-Gaussian, the variable $\widetilde{Z}_i$ is zero-mean and sub-

Gaussian with parameter at most

$$\frac{\sigma^2}{n}\|(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S)^{-1}\|_2 \leq \frac{\sigma^2}{c_{\min} n},$$

where we have used the eigenvalue condition (7.43a). Consequently, for any $\epsilon > 0$, we have $\mathbb{P}\big[\max_{i=1,\ldots,s} |\widetilde{Z}_i| > \frac{\sigma}{\sqrt{c_{\min}}}\big\{\sqrt{\frac{2\log s}{n}} + \epsilon\big\}\big] \leq 2\,e^{-\frac{n\epsilon^2}{2}}$, from which the claim follows. $\square$

Corollary 7.1 applies to sub-Gaussian noise with a fixed design matrix $\mathbf{X}$. An analogous result—albeit with a more involved proof—can be established for sub-Gaussian noise with a Gaussian random design matrix. Doing so involves showing that a random matrix from the Σ-Gaussian ensemble, with rows sampled i.i.d. from a $\mathcal{N}(0, \Sigma)$ distribution, satisfies the α-incoherence condition with high probabability (whenever the population matrix Σ satisfies this condition). We work through a version of this result in Exercise 7.18, showing that the incoherence condition holds with high probability with $n \gtrsim s\log(d-s)$ samples. Figure 7-6 shows that this theoretical prediction is actually sharp, in that the Lasso undergoes a phase transition as a function of the control parameter $\frac{n}{s\log(d-s)}$. See the bibliographic section for further discussion of this phenomenon.



(a) (b)

Figure 7-6. Thresholds for support set recovery using the Lasso. (a) Probability of success $\mathbb{P}[\widehat{S} = S]$ versus the raw sample size n for three different problem sizes $d \in \{128, 256, 512\}$ and square-root sparsity $s = \lceil\sqrt{d}\rceil$. Each point corresponds to the average of 20 random trials, using a random design matrix drawn from the Toeplitz ensemble of Example 1 with $\nu = 0.1$. Note that larger problems require more samples before the Lasso is able to recover the correct support. (b) The same simulation results replotted versus the rescaled sample size $\frac{n}{s\log(d-s)}$. Notice how all three curves are now well-aligned, and show a threshold behavior, consistent with theoretical predictions.

7.5.2 Proof of Theorem 7.6

We begin by developing the necessary and sufficient conditions for optimality in the Lasso. A minor complication arises because the ℓ_1-norm is not differentiable, due to its sharp point at the origin. Instead, we need to work in terms of the sub-differential of the ℓ_1-norm. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$, we say that $z \in \mathbb{R}^d$ is a subgradient at θ, denoted by $z \in \partial f(\theta)$, if we have

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

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 $j = 1, 2, \ldots, d$, where 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

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

which is the analogue of a zero gradient condition in the non-differentiable setting.

Our proof of Theorem 7.6 is based on a constructive procedure, known as a *primal-dual witness method*, which constructs a pair $(\widehat{\theta}, \widehat{z})$. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support. The procedure is as follows:

Primal-dual witness (PDW) construction:

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

$$\widehat{\theta}_S \in \arg\min_{\theta_S \in \mathbb{R}^s} \big\{ \frac{1}{2n}\|y - \mathbf{X}_S\theta_S\|_2^2 + \lambda_n\|\theta_S\|_1 \big\}, \tag{7.48}$$

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

Note that the vector $\widehat{\theta}_{S^c} \in \mathbb{R}^{d-s}$ is determined in Step 1, whereas the remaining three sub-vectors are determined in Steps 2 and 3. By construction, the sub-vectors $\widehat{\theta}_S$, $\widehat{z}_S$ and $\widehat{z}_{S^c}$ satisfy the zero-subgradient condition. By using the fact that $\widehat{\theta}_{S^c} = \theta^*_{S^c} = 0$

and writing out this condition in block matrix form, we obtain

$$\frac{1}{n}\begin{bmatrix}\mathbf{X}_S^T\mathbf{X}_S & \mathbf{X}_S^T\mathbf{X}_{S^c}\\ \mathbf{X}_{S^c}^T\mathbf{X}_S & \mathbf{X}_{S^c}^T\mathbf{X}_{S^c}\end{bmatrix}\begin{bmatrix}\widehat{\theta}_S-\theta^*_S\\ 0\end{bmatrix}-\frac{1}{n}\begin{bmatrix}\mathbf{X}_S^T w\\ \mathbf{X}_{S^c}^T w\end{bmatrix}+\lambda_n\begin{bmatrix}\widehat{z}_S\\ \widehat{z}_{S^c}\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}. \tag{7.49}$$

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 7.1. If the PDW construction succeeds, then under condition (A3), the vector $(\widehat{\theta}_S, 0) \in \mathbb{R}^d$ is the unique optimal solution of the Lasso.

Proof. When the PDW construction succeeds, then $\widehat{\theta} = (\widehat{\theta}_S, 0)$ 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. If we introduce the shorthand notation $F(\theta) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$, then we are guaranteed that $F(\widehat{\theta}) + \lambda_n\langle \widehat{z}, \widehat{\theta}\rangle = F(\widetilde{\theta}) + \lambda_n\|\widetilde{\theta}\|_1$, and hence

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

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

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

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 (7.48). Given the assumption (A3), this subproblem is strictly convex, and so has a unique minimizer. □

Thus, in order to prove parts (a) and (b) of Theorem 7.6, it suffices to show that $\widehat{z}_{S^c}$ from Step 3 satisfies the strict dual feasibility condition. Using the zero-subgradient conditions (7.49), we can solve for the vector $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$, thereby finding that

$$\widehat{z}_{S^c} = -\frac{1}{\lambda_n n}\mathbf{X}_{S^c}^T\mathbf{X}_S\big(\widehat{\theta}_S - \theta^*_S\big) - \mathbf{X}_{S^c}^T\big(\frac{w}{\lambda_n n}\big). \tag{7.50}$$

Similarly, using the assumed invertibility of $\mathbf{X}_S^T\mathbf{X}_S$ in order to solve for the difference

$\widehat{\theta}_S - \theta^*_S$ yields

$$\underbrace{\widehat{\theta}_S - \theta^*_S}_{U_S} = -(\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T w - \lambda_n n (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \widehat{z}_S. \tag{7.51}$$

Substituting this expression back into equation (7.50) and simplifying yields

$$\widehat{z}_{S^c} = \underbrace{\mathbf{X}_{S^c}^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \widehat{z}_S}_{\mu} + \underbrace{\mathbf{X}_{S^c}^T \Big[I - \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \Big] (\frac{w}{\lambda_n \, n})}_{V_{S^c}}. \tag{7.52}$$

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 (7.51) and the triangle inequality, we have

$$\|U_S\|_\infty \leq \big\| (\frac{\mathbf{X}_S^T \mathbf{X}_S}{n})^{-1} \mathbf{X}_S^T \frac{w}{n} \big\|_\infty + \|\!| (\frac{\mathbf{X}_S^T \mathbf{X}_S}{n})^{-1} |\!\|_\infty \, \lambda_n, \tag{7.53}$$

which completes the proof.

Appendix: Proof of Theorem 7.3

By a rescaling argument, it suffices to restrict attention to vectors belonging to the ellipse $\mathbb{S}^{d-1}(\Sigma) = \{\theta \in \mathbb{R}^d \mid \|\sqrt{\Sigma}\theta\|_2 = 1\}$. Define the function $g(t) := \rho(\Sigma) \sqrt{\frac{2 \log d}{n}} \, t$, and the associated "bad" event

$$\mathcal{E} := \Big\{ \inf_{\theta \in \mathbb{S}^{d-1}(\Sigma)} \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2 g(\|\theta\|_1) \Big\}. \tag{7.54}$$

We first claim that on the complementary set $\mathcal{E}^c$, the lower bound (7.31) holds. Let $\theta \in \mathbb{S}^{d-1}(\Sigma)$ be arbitrary. Defining $a = \frac{1}{4}$, $b = 2g(\|\theta\|_1)$ and $c = \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}}$, we have $c \geq \max\{a - b, 0\}$ on the event $\mathcal{E}^c$. We claim that this lower bound implies that $c^2 \geq (1-\delta)^2 a^2 - \frac{1}{\delta^2} b^2$ for any $\delta \in (0,1)$. Indeed, if $\frac{b}{\delta} \geq a$, then the claimed lower bound is trivial. Otherwise, we may assume that $b \leq \delta a$, in which case the bound $c \geq a - b$ implies that $c \geq (1-\delta)a$, and hence that $c^2 \geq (1-\delta)^2 a^2$. Setting $(1-\delta)^2 = \frac{1}{2}$ then yields the claim. Thus, the remainder of our proof is devoted to upper bounding $\mathbb{P}[\mathcal{E}]$.

For radii $0 \leq r_\ell < r_u$, define the sets $\mathbb{K}(r_\ell, r_u) := \Big\{ \theta \in \mathbb{S}^{d-1}(\Sigma) \;\big|\; g(\|\theta\|_1) \in [r_\ell, r_u] \Big\}$,

and the events

$$\mathcal{A}(r_\ell, r_u) := \Big\{ \inf_{\theta \in \mathbb{K}(r_\ell, r_u)} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{2} - 2r_u \Big\}. \tag{7.55}$$

The following lemma is the central technical result in the proof:

Lemma 7.2. For any pair of radii $0 \leq r_\ell < r_u$, we have

$$\mathbb{P}\big[\mathcal{A}(r_\ell, r_u)\big] \leq e^{-\frac{n}{32}} \, e^{-\frac{n}{2} r_u^2}. \tag{7.56}$$

Moreover, for $\mu = 1/4$, we have

$$\mathcal{E} \subseteq \mathcal{A}(0, \mu) \cup \Big(\bigcup_{\ell=1}^{\infty} \mathcal{A}(2^{\ell-1}\mu, 2^\ell \mu) \Big). \tag{7.57}$$

Based on this lemma, the remainder of the proof is straightforward. By the inclusion (7.57) and the union bound, we have

$$\mathbb{P}[\mathcal{E}] \;\leq\; \mathbb{P}[\mathcal{A}(0,\mu)] + \sum_{\ell=1}^{\infty} \mathbb{P}[\mathcal{A}(2^{\ell-1}\mu, 2^\ell \mu)] \leq e^{-\frac{n}{32}} \Big\{ \sum_{\ell=0}^{\infty} e^{-\frac{n}{2} 2^{2\ell} \mu^2} \Big\}.$$

Since $\mu = 1/4$ and $2^{2\ell} \geq 2\ell$, we have

$$\mathbb{P}[\mathcal{E}] \leq e^{-\frac{n}{32}} \sum_{\ell=0}^{\infty} e^{-\frac{n}{2} 2^{2\ell} \mu^2} \leq e^{-\frac{n}{32}} \sum_{\ell=0}^{\infty} \big(e^{-n\mu^2}\big)^\ell \;\leq\; \frac{e^{-\frac{n}{32}}}{1 - e^{-\frac{n}{32}}}$$

It remains to prove the lemma.

Proof of Lemma 7.2: We begin with the inclusion (7.57). Let $\theta \in \mathbb{S}^{d-1}(\Sigma)$ be a vector that certifies the event $\mathcal{E}$; then it must belong to either to the set $\mathbb{K}(0, \mu)$ or to a set $\mathbb{K}(2^{\ell-1}\mu, 2^\ell \mu)$ for some $\ell = 1, 2, \ldots$.

Case 1: First suppose that $\theta \in \mathbb{K}(0, \mu)$, so that $g(\|\theta\|_1) \leq \mu = 1/4$. Since θ certifies the event $\mathcal{E}$, we have

$$\frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2\, g(\|\theta\|_1) \;\leq \frac{1}{4} \;=\; \frac{1}{2} - \mu$$

showing that event $\mathcal{A}(0, \mu)$ must happen.

Case 2: Otherwise, we must have $\theta \in \mathbb{K}(2^{\ell-1}\mu, 2^\ell \mu)$ for some $\ell = 1, 2, \ldots$, and moreover

$$\frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta\|_1) \;\leq \frac{1}{2} - 2\big(2^{\ell-1}\mu\big) \;\leq\; \frac{1}{2} - 2^\ell \mu,$$

which shows that the event $\mathcal{A}(2^{\ell-1}\mu, 2^{\ell}\mu)$ must happen.

We now establish the tail bound (7.56). It is equivalent to upper bound the random variable $T(r_\ell, r_u) := -\inf_{\theta \in \mathbb{K}(r_\ell, r_u)} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}}$. By the variational representation of the ℓ_2-norm, we have

$$T(r_\ell, r_u) = -\inf_{\theta \in \mathbb{K}(r_\ell, r_u)} \sup_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \mathbf{X}\theta \rangle}{\sqrt{n}} = \sup_{\theta \in \mathbb{K}(r_\ell, r_u)} \inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \mathbf{X}\theta \rangle}{\sqrt{n}}.$$

Consequently, if we write $\mathbf{X} = \mathbf{W}\sqrt{\Sigma}$, where $\mathbf{W} \in \mathbb{R}^{n \times d}$ is a standard Gaussian matrix and define the transformed vector $v = \sqrt{\Sigma}\theta$,

$$-\inf_{\theta \in \mathbb{K}(r_\ell, r_u)} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} = \sup_{v \in \widetilde{\mathbb{K}}(r_\ell, r)} \inf_{u \in \mathbb{S}^{n-1}} \underbrace{\frac{\langle u, \mathbf{W}v \rangle}{\sqrt{n}}}_{Z_{u,v}} \tag{7.58}$$

where $\widetilde{\mathbb{K}}(r_\ell, r_u) = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1, \quad g(\Sigma^{-\frac{1}{2}} v) \in [r_\ell, r_u]\}$.

Since (u, v) range over a subset of $\mathbb{S}^{n-1} \times \mathbb{S}^{d-1}$, each variable $Z_{u,v}$ is zero-mean Gaussian with variance n^{-1}. Furthermore, the Gaussian comparison principle due to Gordon, previously used in in the proof of Theorem 6.1, may be applied. More precisely, we may compare the Gaussian process $\{Z_{u,v}\}$ to the zero-mean Gaussian process with elements

$$Y_{u,v} := \frac{\langle g, u \rangle}{\sqrt{n}} + \frac{\langle h, v \rangle}{\sqrt{n}}, \quad \text{where } g \in \mathbb{R}^n, \, h \in \mathbb{R}^d \text{ have i.i.d. } \mathcal{N}(0,1) \text{ entries.}$$

Applying Gordon's inequality (6.60), we find that

$$\begin{aligned}
\mathbb{E}[T(r_\ell, r_u)] \;=\; \mathbb{E}\Big[\sup_{v \in \widetilde{\mathbb{K}}(r_\ell, r_u)} \inf_{u \in \mathbb{S}^{n-1}} Z_{u,v}\Big] &\leq \mathbb{E}\Big[\sup_{v \in \widetilde{\mathbb{K}}(r_\ell, r_u)} \inf_{u \in \mathbb{S}^{n-1}} Y_{u,v}\Big] \\
&= \mathbb{E}\Big[\sup_{v \in \widetilde{\mathbb{K}}(r_\ell, r_u)} \frac{\langle h, v \rangle}{\sqrt{n}}\Big] + \mathbb{E}\Big[\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle g, u \rangle}{\sqrt{n}}\Big] \\
&= \mathbb{E}\Big[\sup_{\theta \in \mathbb{K}(r_\ell, r_u)} \frac{\langle \sqrt{\Sigma}\, h, \theta \rangle}{\sqrt{n}}\Big] - \mathbb{E}\Big[\frac{\|g\|_2}{\sqrt{n}}\Big].
\end{aligned}$$

On one hand, we have $\mathbb{E}[\|g\|_2] \geq \sqrt{n}\sqrt{\frac{2}{\pi}}$. On the other hand, we have

$$\mathbb{E}\Big[\sup_{\theta \in \mathbb{K}(r_\ell, r_u)} \frac{\langle \sqrt{\Sigma}\, h, \theta \rangle}{\sqrt{n}}\Big] \leq \mathbb{E}\Big[\sup_{\theta \in \mathbb{K}(r_\ell, r_u)} \|\theta\|_1 \, \frac{\|\sqrt{\Sigma} h\|_\infty}{\sqrt{n}}\Big] \;\leq\; r_u,$$

since $\mathbb{E}[\frac{\|\sqrt{\Sigma}h\|_\infty}{\sqrt{n}}] \leq \rho(\Sigma)\sqrt{\frac{2\log d}{n}}$. Putting together the pieces, we have shown that

$$\mathbb{E}[T(r_\ell, r_u)] \leq -\sqrt{\frac{2}{\pi}} + r_u. \tag{7.59}$$

From the representation (7.58), we see that the random variable $\sqrt{n}\,T(r_\ell, r_u)$ is a 1-Lipschitz function of the standard Gaussian matrix $\mathbf{W}$, so that Theorem 2.4 implies the upper tail bound $\mathbb{P}\big[T(r_\ell, r_u) \geq \mathbb{E}[T(r_\ell, r_u)] + \delta\big] \leq e^{-n\delta^2/2}$ for all $\delta > 0$. Define the constant $C = \sqrt{\frac{2}{\pi}} - \frac{1}{2} \geq \frac{1}{4}$. Setting $\delta = C + r_u$ and using our upper bound on the mean (7.59) yields

$$\mathbb{P}\big[T(r_\ell, r_u) \geq -\frac{1}{2} + 2\,r_u\big] \leq e^{-\frac{n}{2}C^2}\; e^{-\frac{n}{2}\,r_u^2} \;\leq\; e^{-\frac{n}{32}}\; e^{-\frac{n}{2}\,r_u^2},$$

as claimed.

■ 7.6 Bibliographic details and background

The Gaussian sequence model discussed briefly in Example 7.1 has the been the subject of intensive study. Among other reasons, it is of interest because many non-parametric estimation problems can be "reduced" to equivalent versions in the (infinite-dimensional) normal sequence model. The book by Johnstone [Johar] provides a comprehensive introduction; see also the references therein. Donoho and Johnstone [DJ94] derive sharp upper and lower bounds on the minimax risk in ℓ_p-norm for a vector belonging to an ℓ_q-ball, $q \in [0, 1]$, for the case of the Gaussian sequence model. The problem of bounding the in-sample prediction error for non-parametric least-squares, as studied in Chapter 13, can also be understand as a special case of the Gaussian sequence model.

The use of ℓ_1-regularization for ill-posed inverse problems has a lengthy history, with early work in geophysics (e.g., [LF81, OSL83, SS86]); see Donoho and Stark [DS89] for further discussion. It became the subject of more intensive study in statistics and applied mathematics following the seminal papers of Chen, Donoho and Saunders [CDS98] on the basis pursuit linear program (7.10) as well as its relaxed form, and Tibshirani [Tib96] on the Lasso (7.18); see also the concurrent paper []. Other authors have also studied various forms of non-convex regularization for enforcing sparsity (e.g., see the papers [FL01, ZL08, Zha12, FXZ13, ZZ12, LW13] and references therein).

Early work on the basis pursuit linear program (7.10) focused on the the problem of representing a signal in a pair of bases, in which n is the signal length, and $p = 2n$ indexes the union of the two bases of $\mathbb{R}^n$. The incoherence condition arose from this line of work (e.g., [DH01, EB02]); the necessary and sufficient conditions that constitute the restricted nullspace property seem to have been isolated for the first time by Feuer

and Nemirovski [FN03]. However, the terminology and precise definition of restricted nullspace used here was given by Cohen et al. [CDD08].

Greenshtein and Ritov [GR04] were some of the first authors to provide a high-dimensional analysis of the Lasso, in particular providing bounds on the prediction error. In independent work, Candes and Tao [CT05] and Donoho [Don06a, Don06b] analyzed the basis pursuit method for the case of random Gaussian or unitary matrices, and showed that it can succeed with $n \gtrsim s \log(e\, d/s)$ samples. Donoho and Tanner [DT08] provided a sharp analysis of this threshold phenomenon in the noiseless case, with connections to the structure of random polytopes. The restricted isometry property was introduced by Candes and Tao [CT05, CT07]. They also proposed the Dantzig selector, an alternative ℓ_1-based relaxation closely related to the Lasso, and proved bounds on noisy recovery for ensembles that satisfy the RIP condition. Bickel et al. [BRT09] introduced the weaker restricted eigenvalue (RE) condition, slightly different than but essentially equivalent to the version stated here, and provided a unified way to derive ℓ_2-error and prediction error bounds for both the Lasso and the Dantzig selector. van de Geer and Bühlmann [vdGB09] provide a comprehensive overview of different types of RE conditions, and the relationships among them; see also their book [BvdG11]. The proof of Theorem 7.2(a) was inspired by Bickel et al. [BRT09]; see also the material in Chapter 9, and the paper by Negahban et al. [NRWY12] for a general viewpoint on regularized M-estimators. There are many variants and extensions of the basic Lasso, including the elastic net [ZH05], the fused Lasso [TSR+05], the adaptive Lasso [Zou06], and the group Lasso [YL06]. We discuss some of these extensions in Chapter 9.

Theorem 7.3 was proved by Raskutti et al. [RWY10]. Rudelson and Zhou [RZ13] prove an analogous result for more general ensembles of sub-Gaussian random matrices; this analysis requires substantially different techniques, since Gaussian comparison results are no longer available. Both of these results apply to a very broad class of random matrices; for instance, it is even possible to sample the rows of the random matrix $X \in \mathbb{R}^{n \times d}$ from a distribution with a degenerate covariance matrix, and/or with its maximum eigenvalue diverging with the problem size, and these results can still be applied to show that a (lower) restricted eigenvalue condition holds with high probability. Exercise 7.10 is based on results of Loh and Wainwright [LW12].

Irrepresentable conditions for variable selection consistency were introduced independently by Tropp [Tro06] and Fuchs [Fuc04] in signal processing, and Meinshausen and Buhlmann [MB06] and Zhao and Yu [ZY06] in statistics. The primal-dual witness proof of Theorem 7.6 follows the argument of Wainwright [Wai09b]; see also this paper for extensions to general random Gaussian designs. The proof of Lemma 7.1 was suggested by Caramanis [Car10]. The primal-dual witness method that underlies the proof of Theorem 7.6 has been applied in a variety of other settings, including analysis of group Lasso [OWJ11, WLX13] and related relaxations [JRSR10, NW11b], graphical Lasso [RWRY11], and methods for Gaussian graph selection with hidden vari-

ables [CPW12]. Lee et al. [LST13] describe a general framework for deriving consistency results using the primal-dual witness method.

7.7 Exercises

Exercise 7.1 (Optimization and threshold estimators). (a) Show that the hard thresholding estimator (7.6) corresponds to the optimal solution $\widehat{\theta}$ of the non-convex program

$$\min_{\theta \in \mathbb{R}^n} \Big\{ \frac{1}{2} \|y - \theta\|_2^2 + \frac{1}{2} \lambda^2 \|\theta\|_0 \Big\}.$$

(b) Show that the soft-thresholding estimator (7.7) corresponds to the optimal solution $\widehat{\theta}$ of the ℓ_1-regularized quadratic program

$$\min_{\theta \in \mathbb{R}^n} \Big\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \|\theta\|_1 \Big\}.$$

Exercise 7.2 (Properties of ℓ_q-balls). For a given $q \in (0, 1]$, recall the (strong) ℓ_q-ball

$$\mathbb{B}_q(R_q) := \big\{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q \big\}. \tag{7.60}$$

The weak ℓ_q-ball with parameters (C, α) is defined as

$$\mathbb{B}_{w(\alpha)}(C) := \big\{ \theta \in \mathbb{R}^d \mid |\theta|_{(j)} \leq C j^{-\alpha} \quad \text{for } j = 1, \ldots, d \big\}. \tag{7.61}$$

Here $|\theta|_{(j)}$ denote the order statistics of θ^* in absolute value, ordered from largest to smallest (so that $|\theta|_{(1)} = \max_{j=1,2,\ldots,d} |\theta_j|$ and $|\theta|_{(d)} = \min_{j=1,2,\ldots,d} |\theta_j|$.)

(a) Show that the set $\mathbb{B}_q(R_q)$ is star-shaped around the origin. (A set $\mathcal{C} \subseteq \mathbb{R}^d$ is star-shaped around the origin if $\theta \in \mathcal{C} \Rightarrow t\theta \in \mathcal{C}$ for all $t \in [0, 1]$.)

(b) For any $\alpha > 1/q$, show that there is a radius R_q depending on (C, α) such that $\mathbb{B}_{w(\alpha)}(C) \subseteq \mathbb{B}_q(R_q)$. This inclusion underlies the terminology "strong" and "weak" respectively.

(c) For a given integer $s \in \{1, 2, \ldots, d\}$, the best s-term approximation to a vector $\theta^* \in \mathbb{R}^d$ is given by

$$\Pi_s(\theta^*) := \arg \min_{\|\theta\|_0 \leq s} \|\theta - \theta^*\|_2^2. \tag{7.62}$$

Give a closed form expression for $\Pi_s(\theta^*)$.

(d) When $\theta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0,1]$, show that the best s-term approximation satisfies

$$\|\Pi_s(\theta^*) - \theta^*\|_2^2 \leq \big(R_q\big)^{2/q} \big(\frac{1}{s}\big)^{\frac{2}{q}-1}. \tag{7.63}$$

Exercise 7.3 (Pairwise incoherence)**.** Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, suppose that it has normalized columns ($\|X_j\|_2/\sqrt{n} = 1$ for all $j = 1, \ldots, d$) and pairwise incoherence (7.12) upper bounded as $\delta_{\mathrm{PW}}(\mathbf{X}) < \frac{\gamma}{s}$.

(a) Let $S \subset \{1, 2, \ldots, d\}$ be any subset of size s. Show that $\lambda_{\min}\big(\frac{\mathbf{X}_S^T \mathbf{X}_S}{n}\big) \geq c(\gamma) > 0$, as long as γ is sufficiently small.

(b) Prove that $\mathbf{X}$ satisfies the restricted nullspace property with respect to S as long as $\gamma < 1/3$. (Do this from first principles, without using any results on restricted isometry.)

Exercise 7.4 (RIP and pairwise incoherence)**.** (a) Prove the sandwich relation (7.15) for the pairwise incoherence and RIP constants. Give a matrix for which inequality (i) is tight, and another matrix for which inequality (ii) is tight.

(b) Construct a matrix such that $\delta_s(\mathbf{X}) = \sqrt{s}\, \delta_{\mathrm{PW}}(\mathbf{X})$.

Exercise 7.5 (ℓ_2-RE $\Rightarrow$ ℓ_1-RE)**.** Let $S \subset \{1, 2, \ldots, d\}$ be a subset of cardinality s. A matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ satisfies an ℓ_1-RE condition over S with parameters (γ_1, α_1) if

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq \gamma_1 \frac{\|\theta\|_1^2}{s} \quad \text{for all } \theta \in \mathbb{C}(S; \alpha_1).$$

Show that any matrix satisfying the ℓ_2-RE condition (7.22) with parameters (γ_2, α_2) satisfies the ℓ_1-RE condition with parameters $\gamma_1 = \frac{\gamma_2}{(1+\alpha_2^2)}$ and $\alpha_1 = \alpha_2$.

Exercise 7.6 (Weighted ℓ_1-norms)**.** In many applications, one has additional information about the relative scalings of different predictors, so that it is natural to use a weighted ℓ_1-norm, of the form $\|\theta\|_{\nu(1)} := \sum_{j=1}^d \omega_j |\theta_j|$ where $\omega \in \mathbb{R}^d$ is a vector of strictly positive weights. In the case of noiseless observations, this leads to the weighted basis pursuit LP

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_{\nu(1)} \qquad \text{such that } \mathbf{X}\theta = y.$$

(a) State and prove necessary and sufficient conditions on $\mathbf{X}$ for the weighted basis pursuit LP to (uniquely) recover all k-sparse vectors θ^*.

(b) Suppose that θ^* is supported on a subset S of cardinality s, and the weight vector ω satisfies

$$\omega_j = \begin{cases} 1 & \text{if } j \in S \\ t & \text{otherwise,} \end{cases}$$

for some $t \geq 1$. State and prove a sufficient condition for recovery in terms of $c_{\min} = \lambda_{\min}(\mathbf{X}_S^T \mathbf{X}_S / n)$, the pairwise incoherence $\delta_{\mathrm{PW}}(\mathbf{X})$ and t. How do the conditions on $\mathbf{X}$ behave as $t \to +\infty$?

Exercise 7.7 (Pairwise incoherence and RIP for isotropic ensembles)**.** Consider a random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries.

(a) For any $s \in \{1, 2, \ldots, d\}$, show that the pairwise incoherence satisfies the bound $\delta_{\mathrm{PW}}(\mathbf{X}) < \frac{1}{3s}$ with high probability as long as $n \gtrsim s^2 \log d$.

(b) Show that the RIP constant satisfies the bound $\delta_{2s} < 1/3$ with high probability as long as $n \gtrsim s \log\left(\frac{es}{d}\right)$.

Exercise 7.8 (Violations of pairwise incoherence and RIP)**.** Recall the ensemble of spiked identity covariance matrices from Example 2 with a constant $\mu > 0$, and consider an arbitrary sparsity level $s \in \{1, 2, \ldots, d\}$.

(a) Violation of pairwise incoherence: show that

$$\mathbb{P}[\delta_{\mathrm{PW}}(\mathbf{X}) > \mu - 3\delta] \geq 1 - 6e^{-n\delta^2/8} \quad \text{for all } \delta \in (0, 1/\sqrt{2}).$$

Consequently, a pairwise incoherence condition cannot hold unless $\mu \ll \frac{1}{s}$.

(b) Violation of RIP: Show that

$$\mathbb{P}\big[\delta_{2s}(\mathbf{X}) \geq \big(1 + (\sqrt{2s} - 1)\mu\big)\delta\big] \geq 1 - e^{-n\delta^2/8} \quad \text{for all } \delta \in (0, 1).$$

Consequently, a RIP condition cannot hold unless $\mu \ll \frac{1}{\sqrt{s}}$.

Exercise 7.9 (ℓ_0 and ℓ_1 balls)**.** For an integer $k \in \{1, \ldots, d\}$, consider the following two subsets:

$$\mathbb{L}_0(k) := \mathbb{B}_2(1) \cap \mathbb{B}_0(k) \;=\; \big\{\theta \in \mathbb{R}^d \;\mid\; \|\theta\|_2 \leq 1, \text{ and } \|\theta\|_0 \leq k\big\},$$
$$\mathbb{L}_1(k) := \mathbb{B}_2(1) \cap \mathbb{B}_1(\sqrt{k}) \;=\; \big\{\theta \in \mathbb{R}^d \;\mid\; \|\theta\|_2 \leq 1, \text{ and } \|\theta\|_1 \leq \sqrt{k}\big\}.$$

Let $\overline{\mathrm{conv}}$ denote the closure of the convex hull (when applied to a set).

(a) Prove that $\overline{\text{conv}}\big(\mathbb{L}_0(k)\big) \subseteq \mathbb{L}_1(k)$.

(b) Prove that $\mathbb{L}_1(k) \subseteq 3\,\overline{\text{conv}}\big(\mathbb{L}_0(k)\big)$.

(*Hint:* For part (b), you may find it useful to consider the support functions of the two sets.)

Exercise 7.10 (Sufficient conditions for RE)**.** Consider an arbitrary symmetric matrix Γ for which there is a scalar $\delta > 0$ such that

$$\big|\theta^T \Gamma \theta\big| \leq \delta \qquad \text{for all } \theta \in \mathbb{L}_0(s),$$

where the set $\mathbb{L}_0(s)$ was defined in Exercise 7.9.

(a) Show that

$$\big|\theta^T \Gamma \theta\big| \leq 27\delta\Big\{\|\theta\|_2^2 + \frac{1}{s}\|\theta\|_1^2\Big\} \qquad \text{for all } \theta \in \mathbb{R}^d.$$

(*Hint:* Part (b) of Exercise 7.9 could be useful.)

(b) Use part (a) to show that RIP implies the RE condition.

(c) Give an example of a matrix family that violates RIP for which part (a) can be used to guarantee the RE condition.

Exercise 7.11 (Weaker sufficient conditions for RE)**.** Consider a covariance matrix Σ with minimum eigenvalue $\lambda_{\min}(\Sigma) > 0$ and maximum variance $\rho^2(\Sigma)$.

(a) Show that the lower bound (7.31) implies that the RE condition (7.22) holds with parameter $\kappa = \frac{c_1}{2}\lambda_{\min}(\Sigma)$ over $\mathbb{C}_\alpha(S)$, uniformly for all subsets S of cardinality $|S| \leq \frac{c_1}{2c_2}\frac{\lambda_{\min}(\Sigma)}{\rho^2(\Sigma)}(1+\alpha)^{-2}\frac{n}{\log d}$.

(b) Give a sequence of covariance matrices $\{\Sigma^{(d)}\}$ for which $\lambda_{\max}(\Sigma^{(d)})$ diverges, but part (a) can still be used to guarantee the RE condition.

Exercise 7.12 (Estimation over ℓ_q-"balls")**.** In this problem, we consider linear regression with a vector $\theta^* \in \mathbb{B}_q(R_q)$ for some radius $R_q \geq 1$ and parameter $q \in (0, 1]$ under the following conditions: (a) the design matrix $\mathbf{X}$ satisfies the lower bound (7.31) and uniformly bounded columns ($\|X_j\|_2/\sqrt{n} \leq 1$ for all $j = 1, \ldots, d$); (b) the noise vector $w \in \mathbb{R}^n$ has i.i.d. zero-mean entries that are sub-Gaussian with parameter σ.

Using Theorem 7.4 and under an appropriate lower bound on the sample size n in terms of (d, R_q, σ, q), show that there is a universal constant c such that any Lasso solution $\widehat{\theta}$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq c\, R_q\Big(\frac{\sigma^2 \log d}{n}\Big)^{1-\frac{q}{2}}$$

for universal constant c. (This rate is known to be minimax-optimal; see the paper [RWY11] for details.)

Exercise 7.13 (ℓ_∞-bounds for the Lasso)**.** Consider the sparse linear regression model $y = \mathbf{X}\theta^* + w$, where $w \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ and $\theta^* \in \mathbb{R}^d$ is supported on a subset S. Suppose that the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ has its diagonal entries uniformly upper bounded by one, and that for some parameter $\gamma > 0$, it also satisfies an ℓ_∞-curvature condition of the form

$$\big\|\widehat{\Sigma}_n \Delta\big\|_\infty \geq \gamma \|\Delta\|_\infty \quad \text{for all } \Delta \in \mathbb{C}_3(S). \tag{7.64}$$

Show that with the regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$, any Lasso solution satisfies the ℓ_∞-bound

$$\|\widehat{\theta} - \theta^*\|_\infty \leq \frac{6\sigma}{\gamma}\sqrt{\frac{\log d}{n}}$$

with high probability.

Exercise 7.14 (Verifying ℓ_∞-curvature conditions)**.** This problem is a continuation of Exercise 7.13. Suppose that we form a random design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with rows drawn i.i.d. from a $\mathcal{N}(0, \Sigma)$ distribution, and moreover that

$$\|\Sigma\Delta\|_\infty \geq \gamma\|\Delta\|_\infty \qquad \text{for all vectors } \Delta \in \mathbb{C}_3(S).$$

Show that, with high probability, the sample covariance $\widehat{\Sigma} := \frac{1}{n}\mathbf{X}^T\mathbf{X}$ satisfies this same property with $\gamma/2$ as long as $n \gtrsim s^2 \log d$.

Exercise 7.15 (Sharper bounds for Lasso)**.** Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a fixed design matrix such that $\frac{|||\mathbf{X}_S|||_{\mathrm{op}}}{\sqrt{n}} \leq C$ for all subsets S of cardinality at most s. In this exercise, we show that any solution of the constrained Lasso (7.19) with $R = \|\theta^*\|_1$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \precsim \frac{\sigma}{\kappa}\sqrt{\frac{s\log(e\,d/s)}{n}}, \tag{7.65}$$

where $s = \|\theta^*\|_0$. Note that this bound provides an improvement for linear sparsity (i.e., whenever $s = \alpha d$ for some constant $\alpha \in (0, 1)$).

(a) Define the random variable

$$Z := \sup_{\Delta\in\mathbb{R}^d} \big|\langle \Delta, \frac{1}{n}\mathbf{X}^T w\rangle\big| \quad \text{such that } \|\Delta\|_2 \leq 1 \text{ and } \|\Delta\|_1 \leq \sqrt{s}, \tag{7.66}$$

where $w \sim \mathcal{N}(0, \sigma^2 I)$. Show that

$$\mathbb{P}\Big[Z \geq c_1 \, C \, \sigma \, \big\{\sqrt{s \log \frac{ed}{s}} + \delta\big\}\Big] \leq c_2 e^{-c_3 \, n\delta^2}$$

for universal constants (c_1, c_2, c_3). (*Hint:* The result of Exercise 7.9 may be useful here.)

(b) Use part (a) and results from the chapter to show that if $\mathbf{X}$ satisfies an RE condition, then any optimal Lasso solution $\widehat{\theta}$ satisfies the bound (7.65) with high probability.

Exercise 7.16 (Analysis of weighted Lasso)**.** In this exercise, we analyze the weighted Lasso estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_{\nu(1)} \Big\},$$

where $\|\theta\|_{\nu(1)} := \sum_{j=1}^d \nu_j |\theta_j|$ denotes the *weighted* ℓ_1*-norm* defined by a positive weight vector $\nu \in \mathbb{R}^d$. Define $C_j = \frac{\|X_j\|_2}{\sqrt{n}}$, where $X_j \in \mathbb{R}^n$ denotes the j^{th} column of the design matrix, and let $\widehat{\Delta} = \widehat{\theta} - \theta^*$ be the error vector associated with an optimal solution $\widehat{\theta}$.

(a) Suppose that we choose a regularization parameter $\lambda_n \geq 2 \max\limits_{j=1,\ldots,d} \frac{|\langle X_j, w \rangle|}{n\nu_j}$. Show that the vector $\widehat{\Delta}$ belongs to the modified cone set

$$\mathbb{C}_3(S; \nu) := \big\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_{\nu(1)} \leq 3 \|\Delta_S\|_{\nu(1)}\big\}. \tag{7.67}$$

(b) Assuming that $\mathbf{X}$ satisfies a κ-RE condition over $\mathbb{C}_\nu(S; 3)$, show that

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa} \lambda_n \max_{j \in S} \nu_j.$$

(c) For a general design matrix, the rescaled column norms $C_j = \|X_j\|_2/\sqrt{n}$ may vary widely. Give a choice of weights for which the weighted Lasso error bound is superior to the ordinary Lasso bound. (*Hint:* You should be able to show an improvement by a factor of $\frac{\max_{j \in S} C_j}{\max_{j=1,\ldots,d} C_j}$.)

Exercise 7.17 (From pairwise incoherence to irrepresentability)**.** Consider a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ whose pairwise incoherence (7.12) satisfies the bound $\delta_{\mathrm{PW}}(\mathbf{X}) < \frac{1}{2s}$. Show that the mutual incoherence condition (7.43b) holds for any subset S of cardinality at most s.

Exercise 7.18 (Irrepresentable condition for random designs)**.** Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a random matrix with rows $\{x_i\}_{i=1}^n$ sampled i.i.d. according to a $\mathcal{N}(0, \Sigma)$ distribution. Suppose that the diagonal entries of Σ are at most 1, and that it satisfies the incoherence condition with parameter $\alpha \in [0, 1)$—that is,

$$\max_{j \in S^c} \|\Sigma_{jS}(\Sigma_{SS})^{-1}\|_1 \leq \alpha < 1.$$

Let $z \in \mathbb{R}^s$ be a random vector that depends only on the sub-matrix $\mathbf{X}_S$. 1

(a) Show that, for each $j \in S^c$,

$$\big|X_j^T(\mathbf{X}_S^T\mathbf{X}_S)^{-1}z\big| \leq \alpha + \big|W_j^T(\mathbf{X}_S^T\mathbf{X}_S)^{-1}z\big|,$$

where $W_j \in \mathbb{R}^n$ is a Gaussian random vector, independent of $\mathbf{X}_S$. 2

(b) Use part (a) and random matrix/vector tail bounds to show that

$$\max_{j \in S^c} \big|X_j^T(\mathbf{X}_S^T\mathbf{X}_S)^{-1}z\big| \leq \alpha' = \frac{1}{2}\big(1 + \alpha\big),$$

with probability at least $1 - 4e^{-c \log d}$, as long as $n > \frac{16}{(1-\alpha)\sqrt{c_{\min}}} s \log(d - s)$, where 3
$c_{\min} = \lambda_{\min}(\Sigma_{SS})$. 4

Exercise 7.19 (Analysis of ℓ_0-regularization)**.** Consider a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ satisfying the ℓ_0-based upper/lower RE condition

$$\gamma_\ell \|\Delta\|_2^2 \leq \frac{\|\mathbf{X}\Delta\|_2^2}{n} \leq \gamma_u \|\Delta\|_2^2 \qquad \text{for all } \|\Delta\|_0 \leq 2s, \tag{7.68}$$

Suppose that we observe noisy samples $y = \mathbf{X}\theta^* + w$ for some s-sparse vector θ^*, where the noise vector has i.i.d. entries distributed as $\mathcal{N}(0, \sigma^2)$. In this exercise, we analyze an estimator based on the ℓ_0-constrained quadratic program

$$\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \Big\} \quad \text{such that } \|\theta\|_0 \leq s. \tag{7.69}$$

(a) Show that the non-convex program (7.69) has a unique optimal solution $\widehat{\theta} \in \mathbb{R}^d$. 5

(b) Using the "basic inequality" proof technique, show that

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{\sigma^2 \gamma_u}{\gamma_\ell^2} \frac{s \log(e\, d/s)}{n}.$$

(*Hint:* The result of Exercise 5.7 could be useful to you.) 6