Chapter 3

Convex Methods for Sparse Signal Recovery

In the previous chapter, we saw many problems for which the goal is to find a sparse solution to an underdetermined linear system of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. This problem is NP-hard in general. However, we also observed that certain well-structured instances *can* be solved efficiently: in experiments, when $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$ and $\boldsymbol{x}_o$ was *sufficiently sparse*, tractable ℓ^1-minimization

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \end{array} \tag{3.0.1}$$

exactly recovered $\boldsymbol{x}_o$: $\boldsymbol{x}_o$ was the unique optimal solution to this optimization problem.

The experiments in the previous chapter are inspiring, and perhaps surprising. In this chapter, we will study this phenomenon mathematically, and try to precisely characterize the behavior of (3.0.1). The engineering motivation is simple: we would like to know whether the behavior in the previous chapter is expected in general, and whether we can it use to build reliable systems.

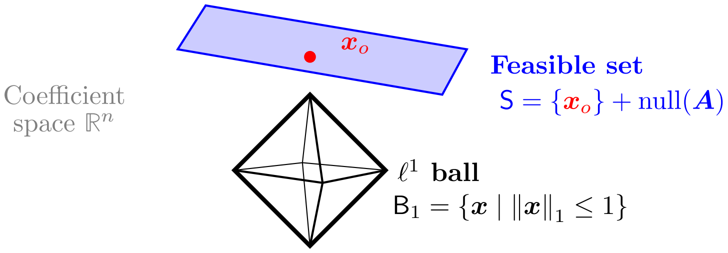

$\boldsymbol{x}_o$
Feasible set
$\mathsf{S} = \{\boldsymbol{x}_o\} + \mathrm{null}(\boldsymbol{A})$
Coefficient space $\mathbb{R}^n$
ℓ^1 ball
$\mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \le 1\}$

Figure 3.1. **Coefficient-space picture.** The set of all solutions $\boldsymbol{x}$ to the equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is an affine subspace S of the coefficient space $\mathbb{R}^n$. The ℓ^1 ball B_1 consists of all coefficient vectors $\boldsymbol{x}$ whose objective function is at most one.

3.1 *Why* Does ℓ^1 Minimization Succeed? Geometric Intuitions

Before diving into a formal proof that the ℓ^1 minimization (3.0.1) correctly recovers sparse signals, we describe two intuitive, geometric pictures of why this is the case.

Coefficient space picture.

We first visualize the problem in the space $\mathbb{R}^n$ of coefficient vectors $\boldsymbol{x}$. The set of vectors $\boldsymbol{x}$ that satisfy the constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ in (3.0.1) is an *affine subspace*[1]

$$\mathsf{S} = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\} = \{\boldsymbol{x}_o\} + \mathrm{null}(\boldsymbol{A}). \tag{3.1.1}$$

Figure 3.1 visualizes this set. The ℓ^1 minimization problem (3.0.1) picks, out of all of the points in the set S, the one (or ones) with smallest ℓ^1 norm. This can be visualized as follows. Consider the ℓ^1 ball of radius one

$$\mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \le 1\} \tag{3.1.2}$$

in $\mathbb{R}^n$. This contains all the vectors $\boldsymbol{x}$ with objective function at most one. Scaling this object by $t \ge 0$ produces the set of vectors $\boldsymbol{x}$ with objective function at most t:

$$t \cdot \mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \le t\}\,. \tag{3.1.3}$$

If we first scale B_1 down to zero, by setting $t = 0$, and then slowly expand it, by increasing t, the ℓ^1 minimizer is obtained when $t \cdot \mathsf{B}_1$ first touches the affine subspace S. This contact point is the solution to (3.0.1) – see Figure 3.2. From the geometry of the ball, it seems that these contact points will

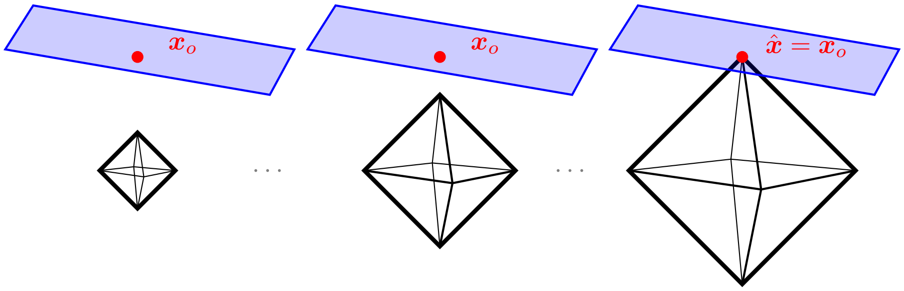


Figure 3.2. ℓ^1 **minimization in the coefficient-space picture.** ℓ^1 minimization can be visualized geometrically as follows: we squeeze the ℓ^1 ball down to zero, and then slowly expand it until it first touches the feasible set S. The point (or points) at which it first touches S is the ℓ^1 minimizer $\hat{\boldsymbol{x}}$.

tend to be the vertices or edges of B_1, which precisely correspond to the sparse vectors!

Observation space picture.

We can also visualize ℓ^1 minimization in the space $\mathbb{R}^m$ of observation vectors $\boldsymbol{y}$. This picture is slightly more complicated, but turns out to be very useful. The $m \times n$ matrix $\boldsymbol{A}$ maps n-dimensional vectors $\boldsymbol{x}$ to $m \ll n$ dimensional vectors $\boldsymbol{y}$. Let us consider how the matrix $\boldsymbol{A}$ acts on the ℓ^1 ball $\mathsf{B}_1 \subset \mathbb{R}^n$. Applying $\boldsymbol{A}$ to each of the vectors $\boldsymbol{x} \in \mathsf{B}_1$, we obtain a lower-dimensional object $\mathsf{P} = \boldsymbol{A}(\mathsf{B}_1)$, which we visualize in Figure 3.3 (right). The lower-dimensional set P is a *convex polytope*. Every vertex $\boldsymbol{v}$ of P is the image $\boldsymbol{A}\boldsymbol{\nu}$ of some vertex $\boldsymbol{\nu} = \pm \boldsymbol{e}_i$ of B_1. More generally, every k-dimensional face of P is the image of some face of B_1.

The polytope P consist of all points $\boldsymbol{y}'$ of the form $\boldsymbol{A}\boldsymbol{x}'$ for some $\boldsymbol{x}'$ with objective function $\|\boldsymbol{x}'\|_1 \leq 1$. ℓ^1 minimization corresponds to squeezing B_1 down to the origin, and then slowly expanding it until it first touches $\boldsymbol{y}$. The touching point is the image $\boldsymbol{A}\boldsymbol{x}_\star$ of the ℓ^1 minimizer – see Figure 3.4.

So, ℓ^1 will correctly recover $\boldsymbol{x}_o$ whenever $\boldsymbol{A}\boldsymbol{x}_o$ is on the outside of $\mathsf{P} = \boldsymbol{A}(\mathsf{B}_1)$. For example, in Figure 3.3, all of the vertices of B_1 map to the outside of $\boldsymbol{A}(\mathsf{B}_1)$, and so ℓ^1 recovers any 1-sparse $\boldsymbol{x}_o$. However, certain edges (one-dimensional faces) of B_1 map to the inside of $\boldsymbol{A}(\mathsf{B}_1)$. ℓ^1 minimization will not recover these $\boldsymbol{x}_o$.

From this picture, it may be very surprising that ℓ^1 works as well as it does. However, as we will see in the remainder of this chapter, the high-dimensional picture differs significantly from the low-dimensional picture

[1] In (3.1.1), the set addition $\{\boldsymbol{x}_o\} + \text{null}(\boldsymbol{A})$ is in the sense of Minkowski, i.e., for sets S and T, $\mathsf{S} + \mathsf{T} = \{\boldsymbol{s} + \boldsymbol{t} \mid \boldsymbol{s} \in \mathsf{S}, \boldsymbol{t} \in \mathsf{T}\}$.

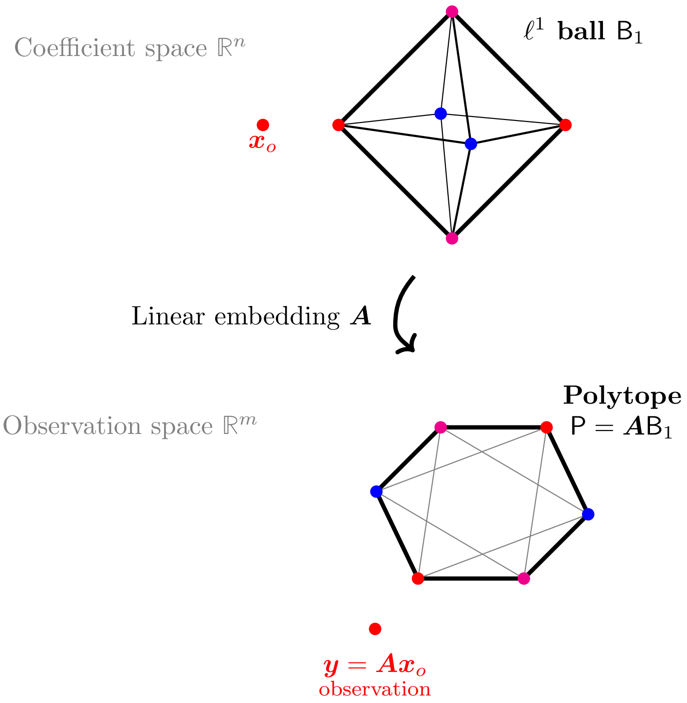

Coefficient space $\mathbb{R}^n$
ℓ^1 **ball** B_1
$\boldsymbol{x}_o$
Linear embedding $\boldsymbol{A}$
Observation space $\mathbb{R}^m$
Polytope
$\mathsf{P} = \boldsymbol{A}\mathsf{B}_1$
$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$
observation

Figure 3.3. **Observation-space picture.** The ℓ^1 ball is a convex polytope B_1 in the coefficient space $\mathbb{R}^n$. The linear map $\boldsymbol{A}$ projects this down to a lower dimensional set $\mathsf{P} = \boldsymbol{A}(\mathsf{B}_1)$ in the observation space $\mathbb{R}^m$. The vertices $\boldsymbol{v}_i$ of P are subsets of the projections $\boldsymbol{A}\boldsymbol{\nu}_j$ of B_1.

(and our low-dimensional intuition!) in ways that are very useful – a "blessing of dimensionality." In particular, if we are in m dimensions and n is proportional to m, not only do all of the vertices of B_1 map to the outside of $\boldsymbol{A}(\mathsf{B}_1)$, so do all the one-dimensional faces, and all of the two-dimensional faces, and so on, all the way up to k-dimensional faces with k proportional to m!

3.2 A First Correctness Result for Incoherent Matrices

With solid empirical evidence and a bit of geometric intuition at hand, our next task is to try to give some rigorous understanding of this phenomenon.

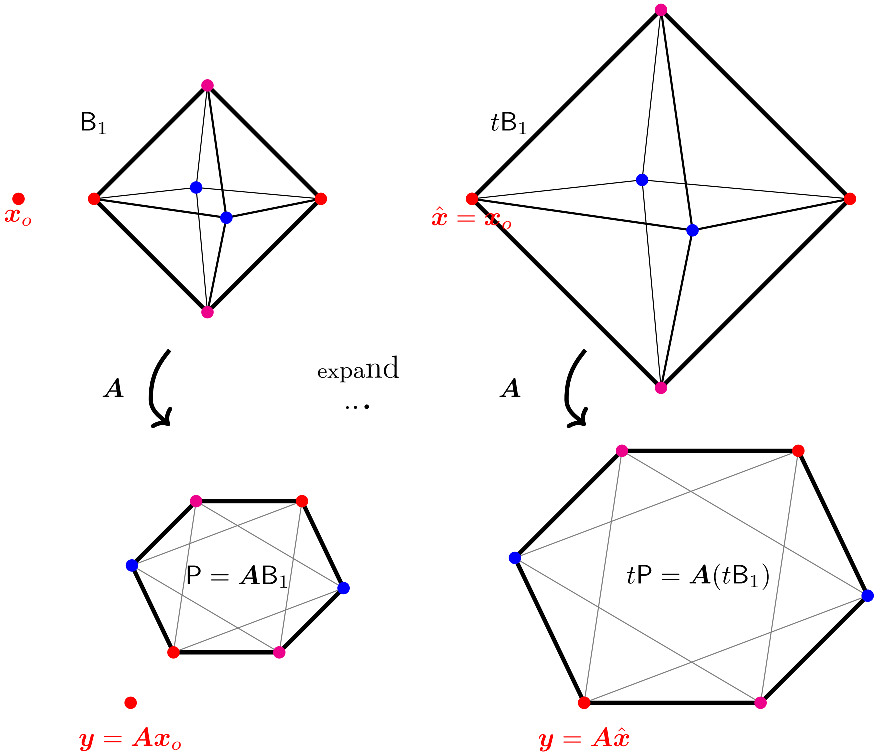

B_1
$\boldsymbol{x}_o$
$t\mathsf{B}_1$
$\hat{\boldsymbol{x}} = \boldsymbol{x}_o$
$\boldsymbol{A}$
expand
…
$\boldsymbol{A}$
$\mathsf{P} = \boldsymbol{A}\mathsf{B}_1$
$t\mathsf{P} = \boldsymbol{A}(t\mathsf{B}_1)$
$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$
$\boldsymbol{y} = \boldsymbol{A}\hat{\boldsymbol{x}}$

Figure 3.4. ℓ^1 **minimization in the observation-space picture.** ℓ^1 minimization corresponds to scaling B_1 down to zero, and then slowly expanding it. As B_1 expands, so does $\mathsf{P} = \boldsymbol{A}\mathsf{B}_1$. The optimal value for the ℓ^1 minimization problem is the first scalar t such that $t\mathsf{P} = \boldsymbol{A}(t\mathsf{B}_1)$ touches the observation vector $\boldsymbol{y}$. The first point that touches $\boldsymbol{y}$ is the image $\boldsymbol{A}\hat{\boldsymbol{x}}$ of the ℓ^1 minimizer $\hat{\boldsymbol{x}}$. This means that *ℓ^1 minimization recovers point $\boldsymbol{x}_o$ if and only if $\boldsymbol{A}\frac{\boldsymbol{x}_o}{\|\boldsymbol{x}_o\|_1}$ lies on the boundary of P.*

3.2.1 Coherence of a Matrix

What determines whether ℓ^1 minimization can recover a target sparse solution $\boldsymbol{x}_o$? Our discussion on ℓ^0 minimization isolated two key factors: how structured the target $\boldsymbol{x}_o$ is (i.e., how many nonzero entries) and how nice the map $\boldsymbol{A}$ is (measured there through the Kruskal rank). Moreover, there was a tradeoff between the two factors: *the nicer $\boldsymbol{A}$ is, the denser $\boldsymbol{x}_o$ we can recover.*

In fact, this qualitative tradeoff carries over to tractable algorithms such as the ℓ^1 relaxation as well. However, we need a slightly stronger notion of the "niceness" of $\boldsymbol{A}$ to guarantee that the tractable relaxation succeeds. Our first notion measures how "spread out" the columns of $\boldsymbol{A}$ are in the high dimensional space $\mathbb{R}^m$:

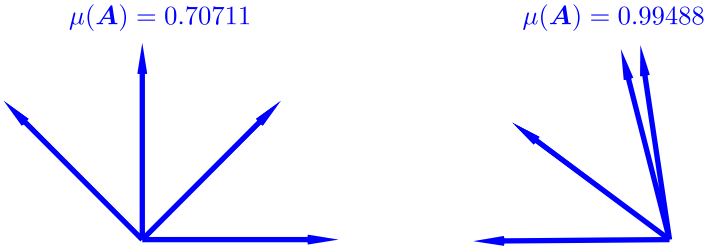

$\mu(\boldsymbol{A}) = 0.70711$
$\mu(\boldsymbol{A}) = 0.99488$

Figure 3.5. **Mutual coherence for two configurations of the columns of $\boldsymbol{A}$.** Left: well-spread vectors in $\mathbb{S}^2$: $\mu(\boldsymbol{A}) \approx 0.707$. This is the smallest achievable μ for four vectors in two dimensions. In higher dimensions, the mutual coherence can be *much* smaller: for example, a random $m \times 2m$ dimensional matrix has coherence on the order of $\sqrt{\log(m)/m}$, which diminishes to zero as m increases. Right: $\mu(\boldsymbol{A}) \approx 0.995$. Mutual coherence depends on the closest pair $\boldsymbol{a}_i, \boldsymbol{a}_j$, and so in this example it is very large.

Definition 3.2.1 (Mutual Coherence)**.** *For a matrix*

$$\boldsymbol{A} = \Big[\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \cdots \mid \boldsymbol{a}_n \Big] \in \mathbb{R}^{m \times n}$$

with nonzero columns, the mutual coherence $\mu(\boldsymbol{A})$ *is the largest normalized inner product between two distinct columns:*

$$\mu(\boldsymbol{A}) = \max_{i \neq j} \left| \left\langle \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}, \frac{\boldsymbol{a}_j}{\|\boldsymbol{a}_j\|_2} \right\rangle \right|. \tag{3.2.1}$$

The mutual coherence takes values in $[0, 1]$. If the columns of $\boldsymbol{A}$ are orthogonal, $\mu(\boldsymbol{A})$ is zero. If $n > m$, the columns of $\boldsymbol{A}$ cannot be orthogonal. The quantity $\mu(\boldsymbol{A})$ captures how close they are to orthogonal, in the worst case sense. Matrices with small $\mu(\boldsymbol{A})$ have columns that are more spread out; we will see that such matrices tend to be better for sparse recovery, in the sense that ℓ^1 succeeds in recovering denser $\boldsymbol{x}_o$. Figure 3.5 visualizes the columns $\boldsymbol{A}$ and displays the coherence, for two examples of $\boldsymbol{A} \in \mathbb{R}^{2 \times n}$.

One intuition for why small $\mu(\boldsymbol{A})$ is helpful is the following: suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, with $\boldsymbol{x}_o$ sparse, and I the support of $\boldsymbol{x}_o$. Then $\boldsymbol{y} = \sum_{i \in I} \boldsymbol{a}_i \boldsymbol{x}_o(i)$. Intuitively speaking, it should be easier to "guess" which columns $\boldsymbol{a}_i$ participate in this linear combination if distinct columns are not too similar to each other.

To connect the mutual coherence more formally to sparse recovery, we will show that whenever $\mu(\boldsymbol{A})$ is small, the Kruskal rank $\mathrm{krank}(\boldsymbol{A})$ is large. Recall that $\mathrm{krank}(\boldsymbol{A}) \geq k$ if and only if every subset of k columns of $\boldsymbol{A}$ is linearly independent, i.e., every k-column submatrix $\boldsymbol{A}_I$ has full column rank. In fact, if the coherence $\mu(\boldsymbol{A})$ is small, then column submatrices of $\boldsymbol{A}$ not only have full column rank – they are even *well-conditioned*, in the sense that their smallest singular value is not far from their largest singular value. To see this, let $I \subset [n]$ with $k = |I|$. Write

$$\boldsymbol{A}_I^* \boldsymbol{A}_I = \boldsymbol{I} + \boldsymbol{\Delta}. \tag{3.2.2}$$

Because $\|\boldsymbol{\Delta}\| \le \|\boldsymbol{\Delta}\|_F < k \|\boldsymbol{\Delta}\|_\infty \le k\mu(\boldsymbol{A})$,[2] we have

$$1 - k\mu(\boldsymbol{A}) \;<\; \sigma_{\min}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;\le\; \sigma_{\max}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;<\; 1 + k\mu(\boldsymbol{A}). \tag{3.2.3}$$

In particular, if $k\mu(\boldsymbol{A}) \le 1$, $\boldsymbol{A}_I$ has full column rank. Combining this observation with our previous discussion of the Kruskal rank, we obtain:

Proposition 3.2.2 (Coherence controls Kruskal rank)**.** *For any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,*

$$\operatorname{krank}(\boldsymbol{A}) \;\ge\; \frac{1}{\mu(\boldsymbol{A})}. \tag{3.2.4}$$

In particular, if $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$ and

$$\|\boldsymbol{x}_o\|_0 \;\le\; \frac{1}{2\mu(\boldsymbol{A})}, \tag{3.2.5}$$

then $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^0 minimization problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.2.6}$$

Thus, provided $\mu(\boldsymbol{A})$ is small enough, ℓ^0 minimization will uniquely recover $\boldsymbol{x}_o$.

3.2.2 Correctness of ℓ^1 Minimization

The previous result showed that if $\mu(\boldsymbol{A})$ is small, then ℓ^0 minimization recovers sufficiently sparse $\boldsymbol{x}_o$. The next result shows that under the same hypotheses, if $\mu(\boldsymbol{A})$ is small, the *tractable* ℓ^1 minimization heuristic also recovers $\boldsymbol{x}_o$. This implies that sparse solutions can be reliably obtained using efficient algorithms! The result is as follows:

Theorem 3.2.3 (ℓ^1 succeeds under incoherence)**.** *Let $\boldsymbol{A}$ be a matrix whose columns have unit ℓ^2 norm, and let $\mu(\boldsymbol{A})$ denote its mutual coherence. Suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, with*

$$\|\boldsymbol{x}_o\|_0 \le \frac{1}{2\mu(\boldsymbol{A})}. \tag{3.2.7}$$

Then $\boldsymbol{x}_o$ is the unique optimal solution to the problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \end{array} \tag{3.2.8}$$

[2] The first inequality comes because the operator norm is always bounded by the Frobenius norm. The second inequality arises because $\|\boldsymbol{\Delta}\|_F^2 = \sum_{ij} |\boldsymbol{\Delta}_{ij}|^2$. The diagonal entries of $\boldsymbol{\Delta}$ are zero, and so in this case, $\|\boldsymbol{\Delta}\|_F^2 = \sum_{i \ne j} |\boldsymbol{\Delta}_{ij}|^2 \le k(k-1) \|\boldsymbol{\Delta}\|_\infty^2$.

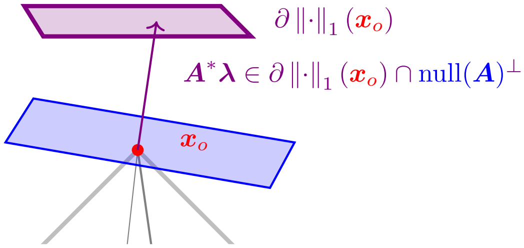

$\partial \left\|\cdot\right\|_1(\boldsymbol{x}_o)$
$\boldsymbol{A}^*\boldsymbol{\lambda} \in \partial \left\|\cdot\right\|_1(\boldsymbol{x}_o) \cap \mathrm{null}(\boldsymbol{A})^\perp$
$\boldsymbol{x}_o$

Figure 3.6. **Geometry of the proof of ℓ^1 recovery.** We prove that $\boldsymbol{x}_o$ is an optimal solution to the ℓ^1 minimization problem, by demonstrating that there exists $\boldsymbol{\lambda}$ such that $\boldsymbol{A}^*\boldsymbol{\lambda}$ is in the subdifferential of $\partial \left\|\cdot\right\|_1(\boldsymbol{x}_o)$. In this picture, there is a *subgradient* of the objective which is orthogonal to $\mathrm{null}(\boldsymbol{A})$. This generalizes the condition for projecting onto an affine subspace (Figure 2.13), in which the gradient of the approximation error is orthogonal to $\mathrm{null}(\boldsymbol{A})$.

Remark 3.2.4. *It is possible to improve the condition of Theorem 3.2.3 slightly, to allow recovery of $\boldsymbol{x}_o$ satisfying*

$$\left\|\boldsymbol{x}_o\right\|_0 \le \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{A})}\right). \tag{3.2.9}$$

This is the best possible statement of this form: there exist examples of $\boldsymbol{A}$ and $\boldsymbol{x}_o$ with $\left\|\boldsymbol{x}_o\right\|_0 > \frac{1}{2}\left(1+\frac{1}{\mu(\boldsymbol{A})}\right)$ for which ℓ^1 minimization does not recover $\boldsymbol{x}_o$. Nevertheless, we will see later in this chapter that for certain classes of $\boldsymbol{A}$ of practical importance, far better guarantees are possible, and that this has important implications for sensing, error correction, and a number of related problems.

Proof ideas for ℓ^1 recovery.

Before embarking on a rigorous proof of Theorem 3.2.3, we sketch our approach. Recall from the previous chapter that for any $\boldsymbol{v} \in \partial \left\|\cdot\right\|_1(\boldsymbol{x}_o)$ and $\boldsymbol{x}' \in \mathbb{R}^n$, the subgradient inequality,

$$\left\|\boldsymbol{x}'\right\|_1 \ge \left\|\boldsymbol{x}_o\right\|_1 + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}_o\rangle \tag{3.2.10}$$

lower bounds the ℓ^1 norm of $\boldsymbol{x}'$. Notice that if $\boldsymbol{x}'$ is feasible for (3.2.8), then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}'$ and so $\boldsymbol{A}(\boldsymbol{x}' - \boldsymbol{x}_o) = \boldsymbol{0}$. Hence, for any $\boldsymbol{\lambda} \in \mathbb{R}^n$,

$$\langle \boldsymbol{A}^*\boldsymbol{\lambda}, \boldsymbol{x}' - \boldsymbol{x}_o\rangle = \langle \boldsymbol{\lambda}, \boldsymbol{A}(\boldsymbol{x}' - \boldsymbol{x}_o)\rangle = 0. \tag{3.2.11}$$

So *if* we can produce a $\boldsymbol{\lambda}$ such that $\boldsymbol{A}^*\boldsymbol{\lambda} \in \partial \left\|\cdot\right\|_1(\boldsymbol{x}_o)$, plugging into (3.2.10) we necessarily have

$$\left\|\boldsymbol{x}'\right\|_1 \ge \left\|\boldsymbol{x}_o\right\|_1 \tag{3.2.12}$$

for every $\boldsymbol{x}' \in \mathbb{R}^n$. This implies that $\boldsymbol{x}_o$ is *an* optimal solution. Figure 3.6 visualizes this construction geometrically.

Let I denote the support of $\boldsymbol{x}_o$, and $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x}_{oI}) \in \{\pm 1\}^k$. Recall that the subdifferential $\partial \left\|\cdot\right\|_1 (\boldsymbol{x}_o)$ consists of those vectors $\boldsymbol{v}$ such that

$$\begin{aligned} \boldsymbol{v}_I &= \boldsymbol{\sigma} && (3.2.13) \\ \left\|\boldsymbol{v}_{I^c}\right\|_\infty &\le 1. && (3.2.14) \end{aligned}$$

Hence, the condition $\boldsymbol{A}^*\boldsymbol{\lambda} \in \partial \left\|\cdot\right\|_1 (\boldsymbol{x}_o)$ places two conditions on the vector $\boldsymbol{A}^*\boldsymbol{\lambda}$:

$$\begin{aligned} \boldsymbol{A}_I^*\boldsymbol{\lambda} &= \boldsymbol{\sigma} && (3.2.15) \\ \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty &\le 1. && (3.2.16) \end{aligned}$$

The first condition is a linear system of k equations, in m unknowns $\boldsymbol{\lambda}$. The second is a system of $n-k$ inequality constraints. The system of equations (3.2.15) is underdetermined. Our approach will be to look at the simplest possible solution to this underdetermined system,

$$\hat{\boldsymbol{\lambda}}_{\ell^2} = \boldsymbol{A}_I(\boldsymbol{A}_I^*\boldsymbol{A}_I)^{-1}\boldsymbol{\sigma}. \tag{3.2.17}$$

This putative solution automatically satisfies the equality constraints (3.2.15). Moreover, $\hat{\boldsymbol{\lambda}}_{\ell^2}$ is a superposition of the columns of $\boldsymbol{A}_I$. Because $\mu(\boldsymbol{A})$ is small, the columns of $\boldsymbol{A}_{I^c}$ are almost orthogonal to the columns of $\boldsymbol{A}_I$, and so $\left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty$ is also small.

Below, we make the above discussion rigorous. The details are slightly more complicated than the above sketch, because we wish to prove that $\boldsymbol{x}_o$ is not just *an* optimal solution, but actually *the unique* optimal solution. We will see that if we can ensure that $\boldsymbol{A}_I$ has full column rank and $\left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty$ is strictly smaller than one, this follows.

Proof of Theorem 3.2.3. Let $I = \operatorname{supp}(\boldsymbol{x}_o)$ and $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x}_{oI}) \in \{\pm 1\}^k$. Notice that $\sigma_{\min}(\boldsymbol{A}_I^*\boldsymbol{A}_I) > 1 - k\mu(A)$, and so under our assumption $\boldsymbol{A}_I$ has full column rank. Suppose that there exists $\boldsymbol{\lambda}$ such that

$$\begin{aligned} \boldsymbol{A}_I^*\boldsymbol{\lambda} &= \boldsymbol{\sigma} && (3.2.18) \\ \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty &< 1. && (3.2.19) \end{aligned}$$

Consider any $\boldsymbol{x}'$ which is feasible, i.e., satisfies $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{y}$. Let $\boldsymbol{v} \in \mathbb{R}^n$ be a vector such that $\boldsymbol{v}_I = \boldsymbol{\sigma}$, and $\boldsymbol{v}_{I^c} = \operatorname{sign}([\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c})$. Notice that $\boldsymbol{v} \in \partial \left\|\cdot\right\|_1 (\boldsymbol{x}_o)$, and so by the subgradient inequality,

$$\left\|\boldsymbol{x}'\right\|_1 \ge \left\|\boldsymbol{x}_o\right\|_1 + \left\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}_o \right\rangle. \tag{3.2.20}$$

Since $\boldsymbol{x}' - \boldsymbol{x}_o \in \operatorname{null}(\boldsymbol{A})$, $\left\langle \boldsymbol{A}^*\boldsymbol{\lambda}, \boldsymbol{x}' - \boldsymbol{x}_o \right\rangle = 0$, and the above equation implies that

$$\begin{aligned} \left\|\boldsymbol{x}'\right\|_1 &\ge \left\|\boldsymbol{x}_o\right\|_1 + \left\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}_o \right\rangle \\ &= \left\|\boldsymbol{x}_o\right\|_1 + \left\langle \boldsymbol{v} - \boldsymbol{A}^*\boldsymbol{\lambda}, \boldsymbol{x}' - \boldsymbol{x}_o \right\rangle \\ &= \left\|\boldsymbol{x}_o\right\|_1 + \left\langle \boldsymbol{v}_{I^c} - \boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}, [\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c} \right\rangle \\ &\ge \left\|\boldsymbol{x}_o\right\|_1 + \left\|[\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c}\right\|_1 - \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty \left\|[\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c}\right\|_1 \\ &= \left\|\boldsymbol{x}_o\right\|_1 + \left(1 - \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{\lambda}\right\|_\infty\right) \left\|[\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c}\right\|_1. && (3.2.21) \end{aligned}$$

Since $\|\boldsymbol{A}_{I^c}^* \boldsymbol{\lambda}\|_\infty < 1$, either $\|\boldsymbol{x}'\|_1 > \|\boldsymbol{x}_o\|_1$, or $\|[\boldsymbol{x}' - \boldsymbol{x}_o]_{I^c}\|_1 = 0$. In the latter case, this means that $\mathrm{supp}(\boldsymbol{x}') \subseteq I$, and $\boldsymbol{x}'_I - \boldsymbol{x}_{oI} \in \mathrm{null}(\boldsymbol{A}_I)$. Since $\boldsymbol{A}_I$ has full column rank, this implies that $\boldsymbol{x}'_I = \boldsymbol{x}_{oI}$, and so $\boldsymbol{x}' = \boldsymbol{x}$.

Hence, if we can construct a $\boldsymbol{\lambda}$ satisfying (3.2.18)-(3.2.19), then any alternative feasible solution $\boldsymbol{x}'$ has larger ℓ^1-norm than $\boldsymbol{x}_o$. Let us try to produce such a $\boldsymbol{\lambda}$. The first equation (3.2.18) above is an underdetermined linear system of equations, with k equations and $m > k$ unknowns $\boldsymbol{\lambda}$. Let us write down one particular solution to this system of equations:

$$\hat{\boldsymbol{\lambda}}_{\ell^2} = \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \boldsymbol{\sigma}. \tag{3.2.22}$$

By construction, $\boldsymbol{A}_I^* \hat{\boldsymbol{\lambda}}_{\ell^2} = \boldsymbol{\sigma}$. We are just left to verify (3.2.19), by calculating

$$\left\| \boldsymbol{A}_{I^c}^* \hat{\boldsymbol{\lambda}}_{\ell^2} \right\|_\infty = \left\| \boldsymbol{A}_{I^c}^* \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \boldsymbol{\sigma} \right\|_\infty . \tag{3.2.23}$$

Consider a single element of this vector, which has the form (for some $j \in I^c$) of

$$\begin{aligned}
|\boldsymbol{a}_j^* \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \boldsymbol{\sigma}| &\le \underbrace{\|\boldsymbol{A}_I^* \boldsymbol{a}_j\|_2}_{\le \sqrt{k}\mu} \underbrace{\left\|(\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1}\right\|_{2,2}}_{< \frac{1}{1 - k\mu(\boldsymbol{A})}} \underbrace{\|\boldsymbol{\sigma}\|_2}_{= \sqrt{k}} && (3.2.24) \\
&< \frac{k\mu(\boldsymbol{A})}{1 - k\mu(\boldsymbol{A})} && (3.2.25) \\
&\le \underset{\text{Provided } k\mu(\boldsymbol{A}) \le 1/2.}{1} && (3.2.26)
\end{aligned}$$

In (3.2.25), we have used that for any invertible $\boldsymbol{M}$, $\left\|\boldsymbol{M}^{-1}\right\| = 1/\sigma_{\min}(\boldsymbol{M})$ and our previous calculation that $\sigma_{\min}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \ge 1 - k\mu(\boldsymbol{A})$ to bound $\left\|(\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1}\right\|_{2,2}$. This calculation shows that under our assumptions, condition (3.2.19) is verified. □

Many extensions and variants of the above result are known. Historically, results of this nature were first proved for special $\boldsymbol{A}$, which consisted of a concatenation of two orthobases

$$\boldsymbol{A} = [\boldsymbol{\Phi} \mid \boldsymbol{\Psi}], \tag{3.2.27}$$

with $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \mid \cdots \mid \boldsymbol{\phi}_n]$, $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \mid \cdots \mid \boldsymbol{\psi}_n] \in O(n)$. In this case, it is possible to prove a sharper bound based on the cross-coherence

$$\max_{ij} \left| \langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_j \rangle \right| . \tag{3.2.28}$$

Another case which is of great interest is when the matrix $\boldsymbol{A}$ has the form $\boldsymbol{A} = \boldsymbol{\Phi}_I^* \boldsymbol{\Psi}$, where $I \subset [n]$, and $\boldsymbol{\Phi}_I \in \mathbb{R}^{n \times |I|}$ is a submatrix of an orthogonal base. For example, in the MRI problem in the previous chapter, $\boldsymbol{\Phi}$ would correspond to the Fourier transform, while $\boldsymbol{\Psi}$ was the basis of sparsity (e.g., wavelets).

3.2.3 Constructing an Incoherent Matrix

In Theorem 3.2.3, we showed that if $\|\boldsymbol{x}_o\|_0 \le 1/2\mu(\boldsymbol{A})$, $\boldsymbol{x}_o$ is correctly recovered by ℓ_1 minimization. Hence, matrices with smaller coherence admit better bounds. The easiest way to build a matrix $\boldsymbol{A}$ with small $\mu(\boldsymbol{A})$ is simply to choose the matrix at random:

Theorem 3.2.5. *Let $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \cdots \mid \boldsymbol{a}_n]$ with columns $\boldsymbol{a}_i \sim \mathrm{uni}(\mathbb{S}^{m-1})$ chosen independently according to the uniform distribution on the sphere. Then with probability at least $3/4$,*

$$\mu(\boldsymbol{A}) \;\le\; C\sqrt{\frac{\log n}{m}}, \tag{3.2.29}$$

where $C > 0$ is a numerical constant.

This result is essentially just a calculation. The main tool needed is the following result, which observes that a Lipschitz function on the sphere concentrates sharply about its median:

Theorem 3.2.6 (Spherical measure concentration)**.** *Let $\boldsymbol{u} \sim \mathrm{uni}(\mathbb{S}^{m-1})$ be distributed according to the uniform distribution on the sphere. Let $f : \mathbb{S}^{m-1} \to \mathbb{R}$ be an 1-Lipschitz function:*

$$\forall\, \boldsymbol{u},\, \boldsymbol{u}', \quad |f(\boldsymbol{u}) - f(\boldsymbol{u}')| \;\le\; 1 \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2\,, \tag{3.2.30}$$

and let $\mathrm{med}(f)$ denote any median of the random variable $Z = f(\boldsymbol{u})$. Then

$$\mathbb{P}\left[\, f(\boldsymbol{u}) > \mathrm{med}(f) + t\,\right] \quad\le\quad 2\exp\left(-\frac{mt^2}{2}\right), \tag{3.2.31}$$

$$\mathbb{P}\left[\, f(\boldsymbol{u}) < \mathrm{med}(f) - t\,\right] \quad\le\quad 2\exp\left(-\frac{mt^2}{2}\right). \tag{3.2.32}$$

For a more detailed introduction to measure concentration, the reader may refer to the book of [Matousek, 2002b]. We will also lay out some of the basic facts of measure concentration and their proofs in the book appendix. For now, we will take it for granted and use it to prove our Theorem 3.2.5.

Proof of Theorem 3.2.5. For any fixed $\boldsymbol{v} \in \mathbb{S}^{m-1}$, we have

$$||\boldsymbol{v}^*\boldsymbol{a}| - |\boldsymbol{v}^*\boldsymbol{a}'|| \;\le\; |\boldsymbol{v}^*(\boldsymbol{a} - \boldsymbol{a}')| \;\le\; \|\boldsymbol{a} - \boldsymbol{a}'\|_2\,. \tag{3.2.33}$$

So, the function $f(\boldsymbol{a}) = |\boldsymbol{v}^*\boldsymbol{a}|$ is 1-Lipschitz. A quick calculation shows that for $\boldsymbol{a} \sim \mathrm{uni}(\mathbb{S}^{m-1})$, we have

$$\mathbb{E}[(\boldsymbol{v}^*\boldsymbol{a})^2] = \frac{1}{m}. \tag{3.2.34}$$

As x^2 is convex, $\mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{a})|\right]^2 \le \mathbb{E}\left[(\boldsymbol{v}^*\boldsymbol{a})^2\right]$. So, we have $\mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{a}|\right] \le \frac{1}{\sqrt{m}}$.

Applying the Markov inequality $\mathbb{P}\left[\, X \geq a \,\right] \leq \frac{\mathbb{E}[X]}{a}$ to f with $a = \mathrm{med}(f)$, then any median of f satisfies

$$\mathrm{med}(f) \;\leq\; 2\mathbb{E}[f] \;\leq\; \frac{2}{\sqrt{m}}. \tag{3.2.35}$$

Finally applying the measure concentration fact from Theorem 3.2.6, we have

$$\mathbb{P}\left[|\boldsymbol{v}^* \boldsymbol{a}| > \frac{2+t}{\sqrt{m}} \right] \leq 2\exp\left(-\frac{t^2}{2} \right). \tag{3.2.36}$$

Since this holds for every fixed $\boldsymbol{v}$, it also holds if $\boldsymbol{v}$ is an independent random vector uniformly distributed on $\mathbb{S}^{m-1}$. So,

$$\mathbb{P}\left[|\boldsymbol{a}_i^* \boldsymbol{a}_j| > \frac{2+t}{\sqrt{m}} \right] \;\leq\; 2\exp\left(-\frac{t^2}{2} \right). \tag{3.2.37}$$

Summing the failure probability over all $n(n-1)/2$ pairs of distinct $(\boldsymbol{a}_i, \boldsymbol{a}_j)$, we have

$$\mathbb{P}\left[\exists\,(i,j) \;:\; |\boldsymbol{a}_i^* \boldsymbol{a}_j| > \frac{2+t}{\sqrt{m}} \right] \;\leq\; n(n-1)\exp\left(-\frac{t^2}{2} \right). \tag{3.2.38}$$

Setting $t = 2\sqrt{\log 2n}$, the above probability is less than $1/4$ and we obtain the result. □

There are several points about Theorem 3.2.5 that are worth remarking on here. First, there is nothing particularly special about the success probability $3/4$. By a slightly different choice of t (which affects the constant C), one can make the success probability arbitrarily close to 1. Second, there is nothing particularly special about the uniform distribution on $\mathbb{S}^{m-1}$ – many distributions will produce similar results, although this one is especially convenient to analyze.

Figure 3.7 plots the average mutual coherence of matrices sampled according to Theorem 3.2.5, for various values of n and $m = n/8$. The observations seem to agree with the predictions of the theorem: the average observed mutual coherence is very close to $1.75\sqrt{\frac{\log n}{m}}$.

3.3 Towards Stronger Correctness Results

3.3.1 Limitations of Incoherence

Theorem 3.2.3 gives a quantitative tradeoff between niceness of $\boldsymbol{A}$ and sparsity of $\boldsymbol{x}_o$, which asserts that when $\boldsymbol{x}_o$ is sparse enough: $\|\boldsymbol{x}_o\|_0 \leq 1/2\mu(\boldsymbol{A})$, then $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^1 minimization problem. This gives a sufficient condition for the ℓ^1 minimization to be correct.

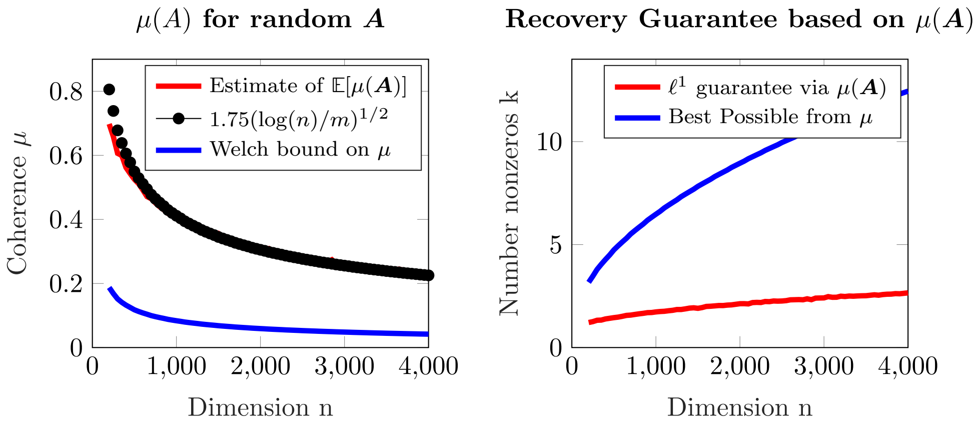


Figure 3.7. **How does coherence decay with dimension?** Left: Average mutual coherence across 50 trials, for $\boldsymbol{A}$ with columns $\boldsymbol{a}_i \sim_{iid} \mathrm{uni}(\mathbb{S}^{m-1})$, for various values of n and $m = n/8$. The black curve, given for reference, is $1.75\sqrt{\frac{\log n}{m}}$. The blue curve is the Welch lower bound $\mu_{\min}$ on the smallest achievable mutual coherence for an $m \times n$ matrix (see Theorem 3.3.1). Right: Average number of nonzeros k which can we can guarantee to reconstruct using the observe $\mu(\boldsymbol{A})$ and Theorem 3.2.3 (red). The blue curve bounds the best possible number of nonzero entries using Theorem 3.2.3, for *any* matrix $\boldsymbol{A}$ of size $m \times n$, using the Welch bound.

But how sharp is this result? According to Theorem 3.2.5, a random matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with high probability has its coherence bounded from above as $\mu(\boldsymbol{A}) \le C\sqrt{\frac{\log n}{m}}$. So, for a "generic" $\boldsymbol{A}$, the above recovery guarantee implies correct recovery of $\boldsymbol{x}_o$ with $O(\sqrt{m/\log n})$ nonzeros. If we turn this around, and think of the matrix multiplication $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ as a sampling procedure, then for appropriately distributed random $\boldsymbol{A}$, we can recover k-sparse $\boldsymbol{x}_o$ from

$$m \ge C' k^2 \log n \tag{3.3.1}$$

observations. When k is small, this is substantially better than simply sampling all n entries of $\boldsymbol{x}$. On the other hand, the measurement burden $m = \Omega(k^2)$ seems a little too high – to specify a k-sparse $\boldsymbol{x}$, we only need to specify its k nonzero entries, …and yet the theory demands k^2 samples!

One might naturally guess that the choice of $\boldsymbol{A}$ as a random matrix was a poor one – perhaps some delicate deterministic construction can yield a better performance guarantee, by making $\mu(\boldsymbol{A})$ smaller. How small can the coherence $\mu(\boldsymbol{A})$ be? We already noted that if $\boldsymbol{A}$ is a square matrix with orthogonal columns, $\mu(\boldsymbol{A}) = 0$. However, if we fix m and allow the number of columns, n, to grow, we are forced to pack more and more vectors $\boldsymbol{a}_j$ into a compact set $\mathbb{S}^{m-1}$. As we increase n, the minimum achievable coherence μ increases.

As it turns out in this case, no matter what we do, we cannot construct a matrix whose coherence is significantly smaller than a randomly chosen one: the coherence of the random matrix $\boldsymbol{A}$ is within $C \log n$ of optimal. The following theorem makes this precise:

Theorem 3.3.1 (Welch bound)**.** *For any matrix* $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \cdots \mid \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$*,* $m \le n$*, and suppose that the columns* $\boldsymbol{a}_i$ *have unit* ℓ^2 *norm. Then*

$$\mu(\boldsymbol{A}) \;=\; \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle| \;\ge\; \sqrt{\frac{n-m}{m(n-1)}}. \tag{3.3.2}$$

Proof. Let $\boldsymbol{G} = \boldsymbol{A}^* \boldsymbol{A} \in \mathbb{R}^{n \times n}$, and let $\lambda_1 \ge \cdots \ge \lambda_m \ge 0$ denote its nonzero eigenvalues.[3] Notice that

$$\sum_{i=1}^{m} \lambda_i(\boldsymbol{G}) = \operatorname{trace}(\boldsymbol{G}) = \sum_{i=1}^{n} \|\boldsymbol{a}_i\|_2^2 = n. \tag{3.3.3}$$

Using this fact, we obtain that

$$\begin{aligned}
\frac{n^2}{m} \;&\le\; \frac{n^2}{m} + \sum_{i=1}^{m} \left(\lambda_i(\boldsymbol{G}) - \frac{n}{m} \right)^2 && (3.3.4)\\
&=\; \frac{n^2}{m} + \sum_{i=1}^{m} \left\{ \lambda_i^2(\boldsymbol{G}) + \frac{n^2}{m^2} - 2\frac{n}{m}\lambda_i(\boldsymbol{G}) \right\} && (3.3.5)\\
&=\; \sum_{i=1}^{m} \lambda_i^2(\boldsymbol{G}) \;=\; \|\boldsymbol{G}\|_F^2 && (3.3.6)\\
&=\; \sum_{i,j} |\boldsymbol{a}_i^* \boldsymbol{a}_j|^2 \;=\; n + \sum_{i \neq j} |\boldsymbol{a}_i^* \boldsymbol{a}_j|^2 && (3.3.7)\\
&\le\; n + n(n-1) \left(\max_{i \neq j} |\boldsymbol{a}_i^* \boldsymbol{a}_j| \right)^2. && (3.3.8)
\end{aligned}$$

Simplifying, we obtain the desired result.

In the above sequence of inequalities, we have used in (3.3.6) the fact that for any symmetric matrix $\boldsymbol{G}$, $\|\boldsymbol{G}\|_F^2 = \sum_i \lambda_i(\boldsymbol{G})^2$, which follows from the eigenvector decomposition $\boldsymbol{G} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$ and the fact that for any matrix $\boldsymbol{M}$ and orthogonal matrices $\boldsymbol{P}, \boldsymbol{Q}$ of appropriate size, $\|\boldsymbol{M}\|_F = \|\boldsymbol{P}\boldsymbol{M}\boldsymbol{Q}\|_F$. □

The important thing to notice here is that if we take n proportional to m, i.e., $n = \beta m$ for some $\beta > 1$, then the bound says that for *any* $m \times n$ matrix $\boldsymbol{A}$,

$$\mu(\boldsymbol{A}) \;\ge\; \Omega\left(\frac{1}{\sqrt{m}} \right). \tag{3.3.9}$$

[3] Because $\operatorname{rank}(\boldsymbol{G}) \le m$, it has at most m nonzero eigenvalues.

Hence, in the best possible case, Theorem 3.2.3 guarantees we can recover $\boldsymbol{x}_o$ with about $\sqrt{m}$ nonzero entries. Or equivalently, no matter how well we choose $\boldsymbol{A}$, to guarantee success Theorem 3.2.3 would demand

$$m \geq C'' k^2 \tag{3.3.10}$$

samples to reconstruct a k-sparse vector, which is only $\log n$ factor better than the previous bound (3.3.1) for a randomly chosen $\boldsymbol{A}$.

Does this behavior reflect a fundamental limitation of the ℓ^1 relaxation? Or is our analysis loose? It turns out that for generic matrices, the situation is much better than the bounds (3.3.1)-(3.3.10) seem to suggest. Again, the easiest way to see this is to do an experiment! We can try solving problems with constant aspect ratio (say, $m = n/2$), and n growing. Try to set $k = \|\boldsymbol{x}_o\|_0$ proportional to m – say, $k = m/4$ (a much better scaling than $k \sim \sqrt{m}$!). Now, try different aspect ratios $m = \alpha n$ and sparsity ratios $k = \beta m$. You may notice something intriguing:

> *In a proportional growth setting $m \propto n$, $k \propto m$, ℓ^1 minimization succeeds with very high probability whenever the constants of proportionality n/m and k/m are small enough.*

This is a very important observation, since it implies that

- **More error correction**: *we can correct constant fractions of errors, using an efficient algorithm.*
- **Compressive sampling**: *we can sense sparse vectors using a number of measurements that is proportional to the intrinsic "information content" of the signal – the number of nonzero entries.*

However, to have a theory that could explain such observation, we will need a more refined measure of the goodness of $\boldsymbol{A}$ than the (rather crude) coherence. In addition, we are going to need to sharpen our theoretical tools too.

3.3.2 The Restricted Isometry Property (RIP)

In the previous section, we saw that the ℓ^1 minimization problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \end{array} \tag{3.3.11}$$

correctly recovers a sparse $\boldsymbol{x}_o$ from observation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, provided two conditions are in force:

- **$\boldsymbol{x}_o$ is structured**: $k = \|\boldsymbol{x}_o\|_0 \ll n$.
- **$\boldsymbol{A}$ is "nice"**: its coherence $\mu(\boldsymbol{A})$ is small.

The intuition provided by incoherence is qualitatively very suggestive, but it does not provide a quantitative explanation for the good behavior we have seen in our experiments so far. How can we strengthen the condition? Suppose that $\boldsymbol{A}$ has unit norm columns. Then it is easy to calculate that for every two-column submatrix $\boldsymbol{A}_I = [\boldsymbol{a}_i \mid \boldsymbol{a}_j] \in \mathbb{R}^{m \times 2}$,

$$\boldsymbol{A}_I^* \boldsymbol{A}_I = \begin{bmatrix} 1 & \boldsymbol{a}_i^* \boldsymbol{a}_j \\ \boldsymbol{a}_j^* \boldsymbol{a}_i & 1 \end{bmatrix}. \tag{3.3.12}$$

Exercise **??** asks you to show that since $|\boldsymbol{a}_i^* \boldsymbol{a}_j| \le \mu(\boldsymbol{A})$, this matrix is well conditioned:

$$1 - \mu(\boldsymbol{A}) \;\le\; \sigma_{\min}(\boldsymbol{A}_I)^2 \;\le\; \sigma_{\max}(\boldsymbol{A}_I)^2 \;\le\; 1 + \mu(\boldsymbol{A}). \tag{3.3.13}$$

This property holds simultaneously for every two-column submatrix $\boldsymbol{A}_I$. So, the property that the columns of $\boldsymbol{A}$ are well-spread implies that *the column submatrices of* $\boldsymbol{A}$ *are well-conditioned.*

We can generalize both properties by taking the set I to be larger than 2. Indeed, we can demand that all k-column submatrices of $\boldsymbol{A}$ are well-conditioned: For every $I \subset \{1, \dots, n\}$ of size k, we have

$$1 - k\mu(\boldsymbol{A}) \;\le\; \sigma_{\min}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;\le\; \sigma_{\max}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;\le\; 1 + k\mu(\boldsymbol{A}), \;\forall\, I \text{ of size} \le k. \tag{3.3.14}$$

This controls the Kruskal rank: if $1 - k\mu(\boldsymbol{A}) > 0$, then $\operatorname{krank}(\boldsymbol{A}) \ge k$. This implies that an incoherent matrix with small μ tends to have large Kruskal rank. Hence according to Theorem 2.2.6, any sufficiently sparse $\boldsymbol{x}_o$ is *the sparsest* solution to the observation equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$.

In (3.3.14), we saw that the coherence $\mu(\boldsymbol{A})$ controls the conditioning of the column submatrices $\boldsymbol{A}_I$ – if $\mu(\boldsymbol{A})$ is small, every submatrix spanned by just a few columns of $\boldsymbol{A}$ is well-conditioned:

$$1 - \delta \;\le\; \sigma_{\min}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;\le\; \sigma_{\max}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \;\le\; 1 + \delta, \tag{3.3.15}$$

with δ small. This turned out to be critical in our proof of Theorem 3.2.3. In fact, we will see that for certain well-structured matrices $\boldsymbol{A}$, including random matrices, the bounds in (3.3.15) hold with δ far smaller than would be predicted using only the coherence.[4] They also holds for far larger $k = |I|$ than might have been predicted from coherence alone. We will see that this leads (via different and slightly more complicated arguments), to substantially tighter guarantees for the performance of both ℓ^0 and ℓ^1 minimization.

The bounds in (3.3.15) hold uniformly over sets I of size k if and only if

$$\forall\, \boldsymbol{x} \;\, k\text{-sparse}, \quad (1 - \delta) \left\| \boldsymbol{x} \right\|_2^2 \;\le\; \left\| \boldsymbol{A}\boldsymbol{x} \right\|_2^2 \;\le\; (1 + \delta) \left\| \boldsymbol{x} \right\|_2^2. \tag{3.3.16}$$

That is to say, the mapping $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ approximately preserves the norm of sparse vectors $\boldsymbol{x}$. Informally, we call such a mapping a *restricted isometry*:

[4] For example, if $\boldsymbol{A}_I$ is a large $m \times k$ $(k < m)$ matrix with entries independent $\mathcal{N}(0, 1/m)$, $\sigma_{min}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \approx (\sqrt{1} - \sqrt{k/m})^2 \ge 1 - 2\sqrt{k/m}$, and $\sigma_{max}(\boldsymbol{A}_I^* \boldsymbol{A}_I) \approx (\sqrt{1} + \sqrt{k/m})^2 \le 1 + 3\sqrt{k/m}$. You can check these values numerically; the aforementioned bounds can be made into rigorous statements using tools for Gaussian processes.

it is (nearly) an isometry[5], *if* we restrict our attention to the sparse vectors $\boldsymbol{x}$.

Definition 3.3.2 (Restricted Isometry Property [Candès and Tao, 2005b])**.** *The matrix* $\boldsymbol{A}$ *satisfies the* restricted isometry property (RIP) *of* order k, *with constant* $\delta \in [0,1)$, *if*

$$\forall\ \boldsymbol{x}\ \ k\text{-sparse}, \quad (1-\delta)\left\|\boldsymbol{x}\right\|_2^2 \ \le\ \left\|\boldsymbol{A}\boldsymbol{x}\right\|_2^2 \ \le\ (1+\delta)\left\|\boldsymbol{x}\right\|_2^2. \tag{3.3.17}$$

The order-k restricted isometry constant $\delta_k(\boldsymbol{A})$ *is the smallest number* δ *such that the above inequality holds.*

Whenever $\delta_k(\boldsymbol{A}) < 1$, every k-column submatrix has full column rank k. This implies that ℓ^0 recovery succeeds under RIP:

Theorem 3.3.3 (ℓ^0 recovery under RIP [Candès et al., 2006, Candès, 2008])**.** *Suppose that* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, *with* $k = \left\|\boldsymbol{x}_o\right\|_0$. *If* $\delta_{2k}(\boldsymbol{A}) < 1$, *then* $\boldsymbol{x}_o$ *is the unique optimal solution to*

$$\begin{array}{ll} \text{minimize} & \left\|\boldsymbol{x}\right\|_0 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.3.18}$$

Proof. Suppose on the contrary that there exists $\boldsymbol{x}' \neq \boldsymbol{x}_o$ with $\left\|\boldsymbol{x}'\right\|_0 \le k$. Then $\boldsymbol{x}_o - \boldsymbol{x}' \in \operatorname{null}(\boldsymbol{A})$, and $\left\|\boldsymbol{x}_o - \boldsymbol{x}'\right\|_0 \le 2k$. This implies that $\delta_{2k}(\boldsymbol{A}) \ge 1$, contradicting our assumption. □

So, provided the RIP constant of order $2k$ is bounded away from one, ℓ^0 minimization successfully recovers $\boldsymbol{x}_o$. If we tighten our demand to $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$, ℓ^1 minimization succeeds as well:

Theorem 3.3.4 (ℓ^1 recovery under RIP)**.** *Suppose that* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, *with* $k = \left\|\boldsymbol{x}_o\right\|_0$. *If* $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$, *then* $\boldsymbol{x}_o$ *is the unique optimal solution to*

$$\begin{array}{ll} \text{minimize} & \left\|\boldsymbol{x}\right\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.3.19}$$

The significance of this result comes from the fact that for "generic" matrices, the condition $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$ holds even when k is nearly proportional to m:

Theorem 3.3.5 (RIP of Gaussian matrices [Candès et al., 2006, Baraniuk et al., 2008])**.** *There exists a numerical constant* $C > 0$ *such that if* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *is a random matrix with entries independent* $\mathcal{N}\left(0, \frac{1}{m}\right)$ *random variables, with high probability,* $\delta_k(\boldsymbol{A}) < \delta$, *provided*

$$m \ \ge\ Ck \log(n/k)/\delta^2. \tag{3.3.20}$$

This implies that recovery of k-sparse $\boldsymbol{x}$ is possible from about $m \ge Ck\log(n/k)$ random measurements. This is a substantial improvement over

[5] An isometry is a mapping that preserves the norm of every vector.

our previous estimate of $m \sim k^2$. In particular, it allows (k, m, n) to scale proportionally [Donoho, 2006, Candès and Tao, 2005b]. This improvement has stimulated a lot of work on efficient sensing and sampling schemes in various application domains.

3.3.3 Restricted Strong Convexity Condition

We have stated the above two theorems without proof. We will prove Theorem 3.3.4 in several stages. In this section, we introduce two intermediate properties of the sensing matrix $\boldsymbol{A}$, which turn out to be very useful in their own right. In the next section, we prove Theorem 3.3.4 by proving that when $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$, these intermediate properties obtain, and hence ℓ^1 minimization succeeds.

As above, suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, for some $\|\boldsymbol{x}_o\|_0 \le k$. We hope that under certain conditions, $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^1 minimization program

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.3.21}$$

Let $\boldsymbol{x}'$ be any feasible point, i.e., any point satisfying $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{y}$. Because $\boldsymbol{A}\boldsymbol{x}_o = \boldsymbol{y}$ as well, *the difference* $\boldsymbol{h} = \boldsymbol{x}' - \boldsymbol{x}_o$ *belongs to the nullspace* $\operatorname{null}(\boldsymbol{A})$.

Let I denote the support of $\boldsymbol{x}_o$, and I^c its complement. Then

$$\begin{aligned} \|\boldsymbol{x}'\|_1 &= \|\boldsymbol{x}_o + \boldsymbol{h}\|_1 && (3.3.22) \\ &\ge \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{h}_I\|_1 + \|\boldsymbol{h}_{I^c}\|_1 . && (3.3.23) \end{aligned}$$

Hence, if $\|\boldsymbol{h}_{I^c}\|_1 > \|\boldsymbol{h}_I\|_1$, $\boldsymbol{x}'$ has strictly larger objective function than $\boldsymbol{x}_o$ and so $\boldsymbol{x}'$ is not optimal. Conversely, if the nullspace of $\boldsymbol{A}$ contains no vectors $\boldsymbol{h} \neq \boldsymbol{0}$ for which $\|\boldsymbol{h}_I\|_1 \ge \|\boldsymbol{h}_{I^c}\|_1$, then $\boldsymbol{x}_o$ must be the unique optimal solution to (3.3.21).

It is helpful to ask what if this were not to be true? What happens if the optimal solution to the above program, say $\hat{\boldsymbol{x}}_{\ell^1}$, was not $\boldsymbol{x}_o$. Under what conditions could their difference $\boldsymbol{h} \doteq \hat{\boldsymbol{x}}_{\ell^1} - \boldsymbol{x}_o$ be nonzero? Let us denote I as the support of $\boldsymbol{x}_o$ and I^c its complement.

Since $\hat{\boldsymbol{x}}$ is the optimal solution to the above program, we must have

$$\begin{aligned} 0 &\ge \|\hat{\boldsymbol{x}}_{\ell^1}\|_1 - \|\boldsymbol{x}_o\|_1 \\ &\ge \|\boldsymbol{x}_o + \boldsymbol{h}\|_1 - \|\boldsymbol{x}_o\|_1 \\ &\ge \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{h}_I\|_1 + \|\boldsymbol{h}_{I^c}\|_1 - \|\boldsymbol{x}_o\|_1 \\ &\ge -\|\boldsymbol{h}_I\|_1 + \|\boldsymbol{h}_{I^c}\|_1. \end{aligned} \tag{3.3.24}$$

That is, we have

$$\|\boldsymbol{h}_{I^c}\|_1 \le \|\boldsymbol{h}_I\|_1. \tag{3.3.25}$$

Meanwhile, since $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o = \boldsymbol{A}\hat{\boldsymbol{x}}_{\ell^1}$, we also have

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{0}. \tag{3.3.26}$$

In other words, in order for the ℓ^1 program to admit a better solution $\hat{\boldsymbol{x}}_{\ell^1}$ than the original sparse solution $\boldsymbol{x}_o$, we must have the above two conditions (3.3.25) and (3.3.26) hold simultaneously. Therefore, in order to show that $\boldsymbol{x}_o$ is the unique optimal solution for the ℓ^1 program, we only have to show that these conditions cannot be all true for any such $\boldsymbol{h}$.

Null Space Property

The above discussion suggests that the null space of $\boldsymbol{A}$ is very important for understanding when we can recover $\boldsymbol{x}_o$. Previous ℓ^0 recovery results all come by showing that the null space does not contain sparse vectors. The condition that for every nonzero $\boldsymbol{h} \in \operatorname{null}(\boldsymbol{A})$, $\|\boldsymbol{h}_I\|_1 < \|\boldsymbol{h}_{I^c}\|_1$, can be interpreted as saying that the null space does not contain any vector that is concentrated on the (small) set of coordinates I. This is sufficient for ℓ^1 minimization to recover $\boldsymbol{x}_o$ with support I. If we want to guarantee recover *any* k-sparse $\boldsymbol{x}_o$, we can ask that for every set I of k coordinates and every nonzero null vector $\boldsymbol{h}$, $\|\boldsymbol{h}_I\|_1 < \|\boldsymbol{h}_{I^c}\|_1$:

Definition 3.3.6 (Null space property)**.** *The matrix $\boldsymbol{A}$ satisfies the* null space property *of order k if for every $\boldsymbol{h} \in \operatorname{null}(\boldsymbol{A}) \setminus \{\boldsymbol{0}\}$ and every I of size at most k,*

$$\|\boldsymbol{h}_I\|_1 < \|\boldsymbol{h}_{I^c}\|_1 . \tag{3.3.27}$$

This can be interpreted as saying that the nullspace does not contain any near-sparse vectors, where sparsity is measured via the ℓ^1 norm. If $\boldsymbol{A}$ satisfies the null space property, then ℓ^1 succeeds in recovering any k-sparse $\boldsymbol{x}_o$:

Lemma 3.3.7. *Suppose that $\boldsymbol{A}$ satisfies the null space property of order k. Then for any $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, with $\|\boldsymbol{x}_o\|_0 \le k$, $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^1 problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.3.28}$$

Proof. Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, with $\|\boldsymbol{x}_o\|_0 \le k$, and let $I = \operatorname{supp}(\boldsymbol{x}_o)$. Let $\hat{\boldsymbol{x}}_{\ell^1}$ be the optimal solution, so $\boldsymbol{h} = \hat{\boldsymbol{x}}_{\ell^1} - \boldsymbol{x}_o \in \operatorname{null}(\boldsymbol{A})$. If $\boldsymbol{h} \neq 0$, then $\|\hat{\boldsymbol{x}}_{\ell^1}\|_1 = \|\boldsymbol{x}_o + \boldsymbol{h}\|_1 \ge \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{h}_I\|_1 + \|\boldsymbol{h}_{I^c}\|_1 > \|\boldsymbol{x}_o\|_1$, contradicting the optimality of $\hat{\boldsymbol{x}}_{\ell^1}$. □

In the viewpoint of the coefficient space picture for ℓ^1 minimization introduced in Section 3.2.2, the nullspace condition asserts that when $\operatorname{null}(\boldsymbol{A})$ is translated to any k-sparse point $\boldsymbol{x}_o$ on the boundary of the ℓ^1 ball B_1, the translate $\boldsymbol{x}_o + \operatorname{null}(\boldsymbol{A})$ does not intersect the interior of B_1. Figure 3.8 visualizes this condition for the special case in which $n = 3$, and $\operatorname{null}(\boldsymbol{A})$ is one-dimensional. In the literature, the null space property has been used to establish various sufficient conditions for the success of ℓ^1 minimization

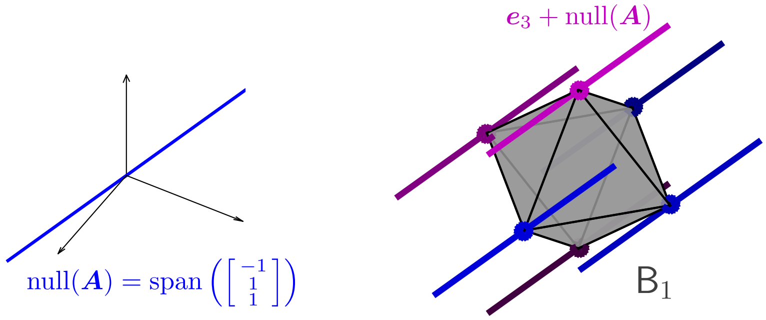

$\boldsymbol{e}_3 + \mathrm{null}(\boldsymbol{A})$
$\mathrm{null}(\boldsymbol{A}) = \mathrm{span}\left(\begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}\right)$
B_1

Figure 3.8. **Visualizing the nullspace property** in three dimensions. Left: the sensing matrix $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ has nullspace spanned by $[-1, 1, 1]^*$. This matrix satisfies the nullspace property of order $k = 1$. Right: Geometrically this implies that any translate $\pm\boldsymbol{e}_j + \mathrm{null}(\boldsymbol{A})$ to a vertex of the ℓ^1 ball B_1 intersects B_1 only at the vertex $\pm\boldsymbol{e}_j$.

for sparse recovery. In fact, Theorem 3.3.4 can be proved by showing that the RIP condition on the matrix $\boldsymbol{A}$ implies the null space property.

Restricted Strong Convexity Condition

Alternatively and equivalently, we can study the success of ℓ^1 minimization by considering possible perturbations $\boldsymbol{h}$ that could reduce the value of the objective function. According to condition (3.3.25), they must satisfy

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1 . \tag{3.3.29}$$

To ensure the original k-sparse $\boldsymbol{x}_o$ is the unique optimal solution, we can require that for any nonzero perturbation $\boldsymbol{h}$ satisfying (3.3.29), $\boldsymbol{A}\boldsymbol{h} \neq \boldsymbol{0}$:

$$\|\boldsymbol{A}\boldsymbol{h}\|_2^2 > 0. \tag{3.3.30}$$

Since the set $\mathsf{S} = \cup_I \{\boldsymbol{h} : \|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1, \|\boldsymbol{h}\|_2^2 = 1\}$ is compact, $\|\boldsymbol{A}\boldsymbol{h}\|_2^2$ must attain its minimum $\mu > 0$. The above condition is therefore equivalent to:

$$\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq \mu\|\boldsymbol{h}\|_2^2, \quad \forall \boldsymbol{h} \ \ \|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1 \tag{3.3.31}$$

for some $\mu > 0$.

If we consider the quadratic loss, $L(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, the second derivative in the $\boldsymbol{h}$ direction is $\boldsymbol{h}^*\nabla^2 L(\boldsymbol{x})\boldsymbol{h} = \|\boldsymbol{A}\boldsymbol{h}\|_2^2 > 0$. The above condition can be interpreted as saying that the function $L(\boldsymbol{x})$ is *strongly convex* when restricted to directions $\boldsymbol{h}$ satisfying (3.3.29) – see Figure 3.9 for a visualization of this interpretation. We term this *(uniform) restricted strong convexity*:

Definition 3.3.8 (Restricted strong convexity)**.** *The matrix* $\boldsymbol{A}$ *satisfies the* restricted strong convexity *(RSC) condition of order* k*, with parameters*

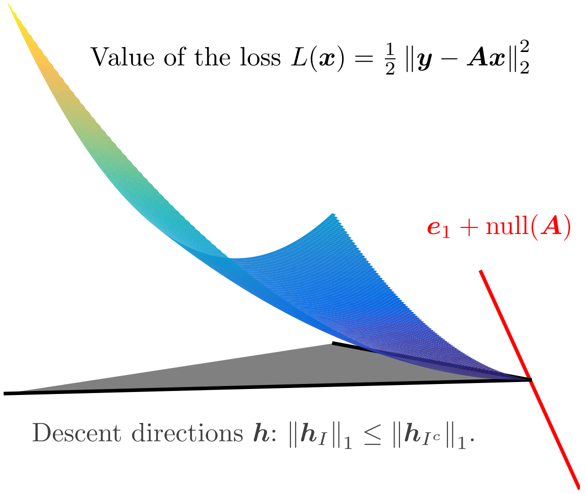

Value of the loss $L(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2$
$\boldsymbol{e}_1 + \mathrm{null}(\boldsymbol{A})$
Descent directions $\boldsymbol{h}$: $\left\|\boldsymbol{h}_I\right\|_1 \le \left\|\boldsymbol{h}_{I^c}\right\|_1$.

Figure 3.9. **Restricted Strong Convexity** implies that the loss $L(\boldsymbol{x})$ exhibits positive curvature along the potential **descent directions** $\boldsymbol{h}$ satsifying $\left\|\boldsymbol{h}_I\right\|_1 \le \left\|\boldsymbol{h}_{I^c}\right\|_1$. Here, $\boldsymbol{x}_o = \boldsymbol{e}_1$. Red: the **feasible set** of $\boldsymbol{x}$ that satisfy $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$. Under RSC, the loss strictly positive at any point $\boldsymbol{x}$ whose ℓ^1 norm is smaller than $\left\|\boldsymbol{x}_o\right\|_1$.

$\mu > 0$, $\alpha \ge 1$, *if for every* I *of size at most* k *and for all nonzero* $\boldsymbol{h}$ *satisfying* $\|\boldsymbol{h}_{I^c}\|_1 \le \alpha \|\boldsymbol{h}_I\|_1$,

$$\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \ge \mu \|\boldsymbol{h}\|_2^2. \tag{3.3.32}$$

In this definition, we have generalized the condition (3.3.29) to consider $\left\|\boldsymbol{h}_{I^c}\right\|_1 \le \alpha \left\|\boldsymbol{h}_I\right\|_1$. This generalization will be used in an essential way later when we study sparse recovery from noisy measurements. For now, we note that for noiseless measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, restricted strong convexity indeed implies that ℓ^1 minimization succeeds:

Lemma 3.3.9. *Suppose that* $\boldsymbol{A}$ *satisfies the restricted strong convexity condition of order* k *with constant* $\alpha \ge 1$, *for some* $\mu > 0$. *Then for any* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, *with* $\left\|\boldsymbol{x}_o\right\|_0 \le k$, $\boldsymbol{x}_o$ *is the unique optimal solution to the* ℓ^1 *problem*

$$\begin{array}{ll} \text{minimize} & \left\|\boldsymbol{x}\right\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{3.3.33}$$

Proof. We leave it as an exercise for the reader to prove this result by verifying that Restricted Strong Convexity implies the nullspace property. □

3.3.4 Success of ℓ^1 Minimization under RIP

In this section we prove Theorem 3.3.4. Earlier, in Section 3.2.2, we have followed a fairly simple path to prove Theorem 3.2.3: write down an optimality condition, and then construct a dual certificate using a bit of cleverness. This approach can be used to prove a variant of Theorem 3.3.4 [Candès and Tao, 2005b]. However, the argument is more delicate than before.

So here, to prove Theorem 3.3.4, we will take a slightly different path, which utilizes properties of "good" sensing matrices $\boldsymbol{A}$ that we have introduced in the previous section. As we have discussed there, to prove that RIP implies correct recovery, it suffices to show that RIP implies the restricted strong convexity (RSC) condition. Our proof here follows close to that of [Candès et al., 2006, Candès, 2008].[6] In doing so, we will use the following property of the restricted isometry constants:

Lemma 3.3.10. *If $\boldsymbol{x}$, $\boldsymbol{z}$ are vectors with disjoint support, and $|\mathrm{supp}(\boldsymbol{x})| + |\mathrm{supp}(\boldsymbol{z})| \le k$, then*

$$|\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{z}\rangle| \;\le\; \delta_k(\boldsymbol{A})\, \|\boldsymbol{x}\|_2 \, \|\boldsymbol{z}\|_2 \,. \tag{3.3.34}$$

Proof. Because the expression is invariant to scaling $\boldsymbol{x}$ and $\boldsymbol{z}$, we lose no generality in assuming that $\|\boldsymbol{x}\|_2 = \|\boldsymbol{z}\|_2 = 1$. Notice that

$$\|\boldsymbol{p} + \boldsymbol{q}\|_2^2 \;=\; \|\boldsymbol{p}\|_2^2 + \|\boldsymbol{q}\|_2^2 + 2\langle \boldsymbol{p}, \boldsymbol{q}\rangle \tag{3.3.35}$$

$$\|\boldsymbol{p} - \boldsymbol{q}\|_2^2 \;=\; \|\boldsymbol{p}\|_2^2 + \|\boldsymbol{q}\|_2^2 - 2\langle \boldsymbol{p}, \boldsymbol{q}\rangle\,, \tag{3.3.36}$$

Hence,

$$|\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{z}\rangle| \;\le\; \frac{1}{4}\left| \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{z}\|_2^2 - \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{z}\|_2^2 \right| \tag{3.3.37}$$

$$\le\; \frac{1}{4}\left| (1+\delta_k)\, \|\boldsymbol{x} + \boldsymbol{z}\|_2^2 - (1-\delta_k)\, \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 \right| \tag{3.3.38}$$

Because $\boldsymbol{x}$ and $\boldsymbol{z}$ have disjoint support, $\|\boldsymbol{x} + \boldsymbol{z}\|_2^2 = \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 = 2$, and the result follows. □

We are now ready to prove the following theorem.

Theorem 3.3.11 (RIP Implies RSC)**.** *If a matrix $\boldsymbol{A}$ satisfies RIP with $\delta_{2k}(\boldsymbol{A}) < \frac{1}{1+\alpha\sqrt{2}}$, then $\boldsymbol{A}$ satisfies the RSC condition of order k with constant α.*

Proof. Let I be any set of size k and let $\boldsymbol{h} \in \mathbb{R}^n$ any vector that satisfies the restriction

$$\|\boldsymbol{h}_{I^c}\|_1 \le \alpha \cdot \|\boldsymbol{h}_I\|_1. \tag{3.3.39}$$

Form disjoint subsets $J_1, J_2, \dots \subseteq I^c$ as follows:

[6] We have modified the original proof that shows RIP implies the null space property to RSC.

J_1 indexes the k largest (in magnitude) elements of $\boldsymbol{h}_{I^c}$

J_2 indexes the k largest (in magnitude) elements of $\boldsymbol{h}_{(I \cup J_1)^c}$

J_3 indexes the k largest (in magnitude) elements of $\boldsymbol{h}_{(I \cup J_1 \cup J_2)^c}$

$$\vdots$$

Notice that because every entry of J_i is at least as large as every entry of J_{i+1}, the average magnitude of an entry in J_i is at least as large as the largest entry in J_{i+1}:

$$\forall\, i \geq 1, \qquad \left\| \boldsymbol{h}_{J_{i+1}} \right\|_\infty \;\leq\; \frac{\| \boldsymbol{h}_{J_i} \|_1}{k}. \tag{3.3.40}$$

We also note that for any vector $\boldsymbol{z}$ with $\|\boldsymbol{z}\|_0 \leq k$, $\|\boldsymbol{z}\|_1 \leq \sqrt{k}\, \|\boldsymbol{z}\|_2$ and $\|\boldsymbol{z}\|_2 \leq \sqrt{k}\, \|\boldsymbol{z}\|_\infty$.

Using the RIP with the $2k$-sparse vector $\boldsymbol{h}_{I \cup J_1}$ and the fact

$$\boldsymbol{A}\boldsymbol{h}_I + \boldsymbol{A}\boldsymbol{h}_{J_1} = \boldsymbol{A}\boldsymbol{h} - \boldsymbol{A}\boldsymbol{h}_{J_2} - \boldsymbol{A}\boldsymbol{h}_{J_3} - \dots, \tag{3.3.41}$$

we have that

$$\begin{aligned}
(1-\delta_{2k})\|\boldsymbol{h}_{I\cup J_1}\|_2^2 \;&\leq\; \|\boldsymbol{A}\boldsymbol{h}_{I\cup J_1}\|_2^2 \\
&= \langle \boldsymbol{A}\boldsymbol{h}_I + \boldsymbol{A}\boldsymbol{h}_{J_1}, -\boldsymbol{A}\boldsymbol{h}_{J_2} - \boldsymbol{A}\boldsymbol{h}_{J_3} - \dots \rangle + \langle \boldsymbol{A}\boldsymbol{h}_I + \boldsymbol{A}\boldsymbol{h}_{J_1}, \boldsymbol{A}\boldsymbol{h} \rangle \\
&\leq \sum_{j=2}^{\infty} \left(\left| \left\langle \boldsymbol{A}\boldsymbol{h}_I, \boldsymbol{A}\boldsymbol{h}_{J_j} \right\rangle \right| + \left| \left\langle \boldsymbol{A}\boldsymbol{h}_{J_1}, \boldsymbol{A}\boldsymbol{h}_{J_j} \right\rangle \right| \right) + \|\boldsymbol{A}\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2 \\
&\leq \delta_{2k}(\|\boldsymbol{h}_I\|_2 + \|\boldsymbol{h}_{J_1}\|_2) \sum_{j=2}^{\infty} \left\| \boldsymbol{h}_{J_j} \right\|_2 + (1+\delta_{2k})^{1/2} \|\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2 \\
&\leq \delta_{2k}\sqrt{2}\, \|\boldsymbol{h}_{I\cup J_1}\|_2 \sum_{j=2}^{\infty} \left\| \boldsymbol{h}_{J_j} \right\|_2 + (1+\delta_{2k})^{1/2} \|\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2 \\
&\leq \delta_{2k}\sqrt{2}\, \|\boldsymbol{h}_{I\cup J_1}\|_2 \sum_{j=2}^{\infty} \left\| \boldsymbol{h}_{J_j} \right\|_\infty \sqrt{k} + (1+\delta_{2k})^{1/2} \|\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2 \\
&\leq \delta_{2k}\sqrt{2}\, \|\boldsymbol{h}_{I\cup J_1}\|_2 \sum_{j=1}^{\infty} \left\| \boldsymbol{h}_{J_j} \right\|_1 / \sqrt{k} + (1+\delta_{2k})^{1/2} \|\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2 \\
&= \delta_{2k}\sqrt{2}\, \|\boldsymbol{h}_{I\cup J_1}\|_2 \, \|\boldsymbol{h}_{I^c}\|_1 / \sqrt{k} + (1+\delta_{2k})^{1/2} \|\boldsymbol{h}_{I\cup J_1}\|_2 \|\boldsymbol{A}\boldsymbol{h}\|_2.
\end{aligned} \tag{3.3.42}$$

After dividing through by $\|\boldsymbol{h}_{I\cup J_1}\|_2$, we have

$$(1-\delta_{2k})\|\boldsymbol{h}_{I\cup J_1}\|_2 \leq \delta_{2k}\sqrt{2}\, \|\boldsymbol{h}_{I^c}\|_1 / \sqrt{k} + (1+\delta_{2k})^{1/2} \, \|\boldsymbol{A}\boldsymbol{h}\|_2 \,. \tag{3.3.43}$$

Since $\boldsymbol{h}$ satisfies the restricted cone condition, we have

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \alpha \|\boldsymbol{h}_I\|_1 \leq \alpha\sqrt{k} \|\boldsymbol{h}_I\|_2 \leq \alpha\sqrt{k} \|\boldsymbol{h}_{I\cup J_1}\|_2. \tag{3.3.44}$$

Substituting this into the previous inequality, we obtain:

$$(1-\delta_{2k})\|\boldsymbol{h}_{I\cup J_1}\|_2 \leq \alpha\delta_{2k}\sqrt{2}\,\|\boldsymbol{h}_{I\cup J_1}\|_2 + (1+\delta_{2k})^{1/2}\,\|\boldsymbol{A}\boldsymbol{h}\|_2\,. \tag{3.3.45}$$

This gives

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 \geq \frac{1-\delta_{2k}(1+\alpha\sqrt{2})}{(1+\delta_{2k})^{1/2}}\,\|\boldsymbol{h}_{I\cup J_1}\|_2\,. \tag{3.3.46}$$

Since the i-th element of $\boldsymbol{h}_{(I\cup J_1)^c}$ is no larger than the mean of the first i elements of $\boldsymbol{h}_{I^c}$, we have

$$|\boldsymbol{h}_{(I\cup J_1)^c}|_{(i)} \leq \|\boldsymbol{h}_{I^c}\|_1/i. \tag{3.3.47}$$

Combining with the restriction (3.3.39), we have

$$\begin{aligned}
\|\boldsymbol{h}_{(I\cup J_1)^c}\|_2^2 &\leq \|\boldsymbol{h}_{I^c}\|_1^2\sum_{i=k+1}^{\infty}\frac{1}{i^2} &(3.3.48)\\
&\leq \frac{\|\boldsymbol{h}_{I^c}\|_1^2}{k} \leq \frac{\alpha^2\|\boldsymbol{h}_I\|_1^2}{k} &(3.3.49)\\
&\leq \alpha^2\|\boldsymbol{h}_I\|_2^2 \leq \alpha^2\|\boldsymbol{h}_{I\cup J_1}\|_2^2. &(3.3.50)
\end{aligned}$$

So we have

$$\|\boldsymbol{h}\|_2^2 \leq (1+\alpha^2)\|\boldsymbol{h}_{I\cup J_1}\|_2^2. \tag{3.3.51}$$

Combining this with the previous condition on $\|\boldsymbol{A}\boldsymbol{h}\|_2$, we get

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 \geq \frac{1-\delta_{2k}(1+\alpha\sqrt{2})}{(1+\delta_{2k})^{1/2}\sqrt{1+\alpha^2}}\,\|\boldsymbol{h}\|_2\,. \tag{3.3.52}$$

So as long as $\delta_{2k} < \frac{1}{1+\alpha\sqrt{2}}$, $\boldsymbol{A}$ satisfies the RSC condition of order k with the constant

$$\mu = \frac{\big(1-\delta_{2k}(1+\alpha\sqrt{2})\big)^2}{(1+\delta_{2k})(1+\alpha^2)}, \tag{3.3.53}$$

as claimed. □

Theorem 3.3.4 then becomes a corollary to this theorem for the case $\alpha = 1$ since the restriction set we need to consider is $\|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1$ for the ℓ^1 minimization in Theorem 3.3.4 and that gives the RIP constant $\delta_{2k} = \frac{1}{1+\sqrt{2}} = \sqrt{2}-1$.

3.4 Matrices with Restricted Isometry Property

The RIP gives a useful tool for analyzing the performance of sparse recovery with random matrices $\boldsymbol{A}$. Below, we will prove the probabilistic result, Theorem 3.3.5, which asserts that Gaussian random matrix $\boldsymbol{A}$ has RIP

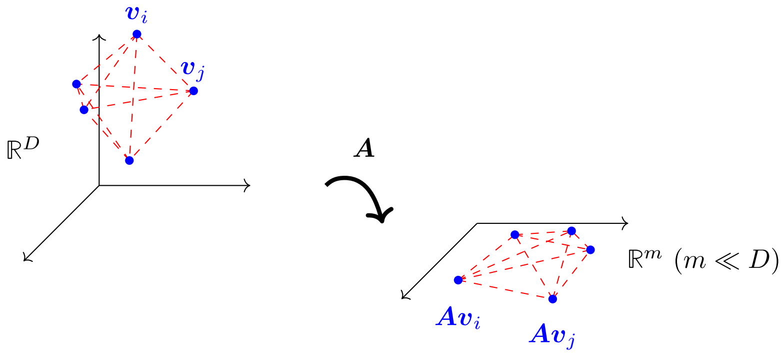

$\boldsymbol{v}_i$
$\boldsymbol{v}_j$
$\mathbb{R}^D$
$\boldsymbol{A}$
$\mathbb{R}^m$ $(m \ll D)$
$\boldsymbol{A}\boldsymbol{v}_i$
$\boldsymbol{A}\boldsymbol{v}_j$

Figure 3.10. **The Johnson-Lindenstrauss Lemma.** Given a fixed collection of points $\boldsymbol{v}_1 \dots \boldsymbol{v}_n$ in a high-dimensional space $\mathbb{R}^D$, with high probability a random mapping into $m \sim \log n$ dimensions approximately preserves the distances between all pairs of points.

when $m > Ck\log(n/k)$. We will make heavy use of the following simple inequality:

Lemma 3.4.1. *Let $\boldsymbol{g} = (g_1, \dots, g_m)$ be an m-dimensional random vector whose entries are iid $\mathcal{N}(0, 1/m)$. Then for any $t \in [0,1]$,*

$$\mathbb{P}\left[\left|\|\boldsymbol{g}\|_2^2 - 1\right| > t\right] \;\le\; 2\exp\left(-\frac{t^2 m}{8}\right). \tag{3.4.1}$$

This result can be obtained via the exponential moment method (in a similar fashion to the Chernoff bound).

3.4.1 The Johnson-Lindenstrauss Lemma

Before proving Theorem 3.3.5, we first state and prove a simpler result, as an illustration of the basic approach we will take to this result, and is very useful in its own right:

Theorem 3.4.2 (Johnson-Lindenstrauss Lemma)**.** *Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^D$ for some D. Let $\boldsymbol{A} \in \mathbb{R}^{m \times D}$ be a random matrix whose entries are independent $\mathcal{N}(0, 1/m)$ random variables. Then for any $\varepsilon \in (0,1)$, with probability at least $1 - 1/n^2$, the following holds:*

$$\forall\, i \ne j, \qquad (1-\varepsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \;\le\; \|\boldsymbol{A}\boldsymbol{v}_i - \boldsymbol{A}\boldsymbol{v}_j\|_2^2 \;\le\; (1+\varepsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2, \tag{3.4.2}$$

provided $m > 32\frac{\log n}{\varepsilon^2}$.

This result can be thought of as follows: we have a large database $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ of very high-dimensional vectors. We would like to embed them in a lower-dimensional space ($m \ll D$) such that the pairwise distances between the vectors are preserved. This is useful, for example, if we think of these as points in a database, and we imagine that we would like to be able to query the database to find points that are close to a given input $\boldsymbol{q}$ in norm – a good embedding will reduce both the storage and computation requirements for achieving this. If you think carefully, it should be clear that we can achieve a perfect (norm-preserving) embedding into $m = n$ dimensional space – simply project each point onto the span of the n points $\boldsymbol{v}_i$. The surprise in the Johnson-Lindenstrauss lemma is that actually, if we allow some slack ε, the dimension can be much lower – only logarithmic in the number of points, and completely independent of the ambient data dimension D. It should not be too surprising that approaches (loosely) inspired by this result have significant applications in search problems. Interestingly, with some additional clever ideas, it is possible to give algorithms that can find approximate nearest neighbors in a database of points in a search time that depends sublinearly on the size of the dataset. Time permitting, we will return to this phenomenon at the end of the semester.

Proof. Set $\boldsymbol{g}_{ij} = \boldsymbol{A}\frac{\boldsymbol{v}_i - \boldsymbol{v}_j}{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2}$. Notice that for any $\boldsymbol{v}_i \neq \boldsymbol{v}_j$, $\boldsymbol{g}_{ij}$ is distributed as an iid Gaussian vector, with entries $\mathcal{N}(0, 1/m)$. Applying Lemma 3.4.1, for each $i \neq j$, we have

$$\mathbb{P}\left[\left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| > t\right] \;\le\; 2\exp\left(-t^2 m/8\right). \tag{3.4.3}$$

Summing the probability of failure over all $i \neq j$, and then plugging in $t = \varepsilon$ and $m \geq 32 \log n/\varepsilon^2$, we get

$$\begin{aligned}\mathbb{P}\left[\exists\,(i,j)\;:\;\left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| > t\right] \;&\le\; \frac{n(n-1)}{2} \times 2\exp\left(-t^2 m/8\right) \\ &\le\; n^{-2}. \end{aligned} \tag{3.4.4}$$

Whenever $\left|\|\boldsymbol{g}_{ij}\|_2^2 - 1\right| \leq \varepsilon$, we have

$$(1-\varepsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \;\le\; \|\boldsymbol{A}\boldsymbol{v}_i - \boldsymbol{A}\boldsymbol{v}_j\|_2^2 \;\le\; (1+\varepsilon)\,\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2\,, \tag{3.4.5}$$

as desired. □

Thus, the fairly powerful embedding result (Theorem 3.4.2) follows from a fairly straightforward pattern:

- **Discretization**: Argue that if $\boldsymbol{A}$ respects the norms of some finite set of vectors (here $\{\boldsymbol{v}_i - \boldsymbol{v}_j \mid i \neq j\}$), the desired property holds.
- **Tail bound**: Develop an upper bound on the probability that $\boldsymbol{A}$ fails to respect the norm of a single vector (here, this is Lemma 3.4.1).
- **Union bound**: Sum the failure probabilities over all of the finite set. Choose the embedding dimension m large enough that the total failure probability is small.

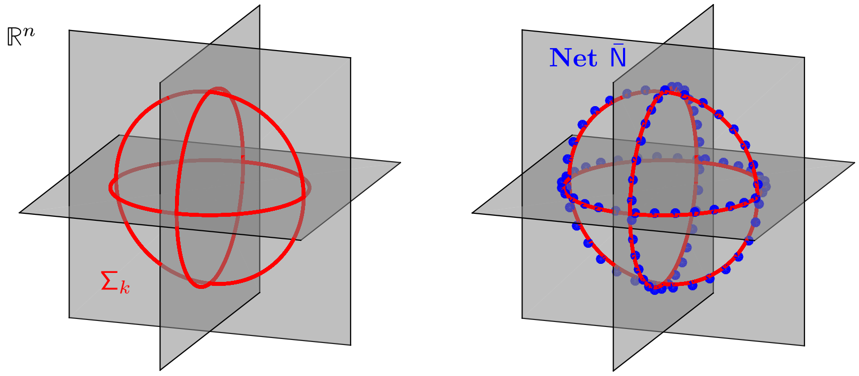


Figure 3.11. **The set Σ_k of unit norm sparse vectors.** Left: visualization of the set $\Sigma_k = \left\{ \boldsymbol{x} \mid \|\boldsymbol{x}\|_0 \le k, \|\boldsymbol{x}\|_2 = 1 \right\}$ of unit norm sparse vectors. Here, $k = 2$ and $n = 3$. Right: an ε-net $\bar{\mathsf{N}}$ for this set.

3.4.2 RIP of Gaussian Matrices

To prove Theorem 3.3.5, we follow exactly the same pattern as we did for Johnson-Lindenstrauss. However, we will need to work a little bit harder in the discretization stage, since unlike the JL Lemma, which was a statement about n (or $n(n-1)/2$) vectors, the RIP is a statement about an infinite family of vectors – all of the sparse vectors.

Discretization.

Let

$$\Sigma_k = \{ \boldsymbol{x} \mid \|\boldsymbol{x}\|_0 \le k, \; \|\boldsymbol{x}\|_2 = 1 \} . \tag{3.4.6}$$

Notice that $\delta_k(\boldsymbol{A}) \; \le \; \delta$ if and only if

$$\sup_{\boldsymbol{x} \in \Sigma_k} \left| \|\boldsymbol{A}\boldsymbol{x}\|_2^2 - 1 \right| \; \le \; \delta. \tag{3.4.7}$$

This is equivalent to

$$\sup_{\boldsymbol{x} \in \Sigma_k} \left| \langle \boldsymbol{A}^* \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle - 1 \right| \; \le \; \delta. \tag{3.4.8}$$

Lemma 3.4.3 (Discretization)**.** *Suppose we have a set $\bar{\mathsf{N}} \subseteq \Sigma_k$ with the following property: for all $\boldsymbol{x} \in \Sigma_k$, there exists $\bar{\boldsymbol{x}} \in \bar{\mathsf{N}}$ such that*

- $|\mathrm{supp}\,(\bar{\boldsymbol{x}}) \cup \mathrm{supp}\,(\boldsymbol{x})| \; \le \; k$
- $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \le \varepsilon.$

set

$$\delta_{\bar{\mathsf{N}}} = \max_{\bar{\boldsymbol{x}} \in \bar{\mathsf{N}}} \left| \|\boldsymbol{A}\bar{\boldsymbol{x}}\|_2^2 - 1 \right| . \tag{3.4.9}$$

Then

$$\delta_k(\boldsymbol{A}) \;\le\; \frac{\delta_{\bar{\mathsf{N}}} + 2\varepsilon}{1 - 2\varepsilon}. \tag{3.4.10}$$

So, provided ε is small, not much changes if we restrict our calculation to the finite set $\bar{\mathsf{N}}$. The proof of this result uses the fact that if $\boldsymbol{x}$ and $\boldsymbol{z}$ are k-sparse vectors,

$$\langle \boldsymbol{Ax}, \boldsymbol{Az} \rangle \;\le\; \sqrt{\|\boldsymbol{Ax}\|_2^2 \, \|\boldsymbol{Az}\|_2^2} \;\le\; (1 + \delta_k(\boldsymbol{A})) \, \|\boldsymbol{x}\|_2 \, \|\boldsymbol{z}\|_2 \,. \tag{3.4.11}$$

Proof. Take any $\boldsymbol{x} \in \Sigma_k$ and choose $\bar{\boldsymbol{x}} \in \bar{\mathsf{N}}$ such that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_0 \le k$ and $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \le \varepsilon$. We have

$$\begin{aligned}
\left| \|\boldsymbol{Ax}\|_2^2 - 1 \right| &= |\langle \boldsymbol{Ax}, \boldsymbol{Ax} \rangle - 1| && (3.4.12)\\
&= |\langle \boldsymbol{Ax}, \boldsymbol{Ax} \rangle - \langle \boldsymbol{A\bar{x}}, \boldsymbol{A\bar{x}} \rangle + \langle \boldsymbol{A\bar{x}}, \boldsymbol{A\bar{x}} \rangle - 1| && (3.4.13)\\
&\le |\langle \boldsymbol{Ax}, \boldsymbol{Ax} \rangle - \langle \boldsymbol{A\bar{x}}, \boldsymbol{A\bar{x}} \rangle| + \delta_{\bar{\mathsf{N}}} && (3.4.14)\\
&= |\langle \boldsymbol{Ax}, \boldsymbol{A}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \rangle - \langle \boldsymbol{A\bar{x}}, \boldsymbol{A}(\bar{\boldsymbol{x}} - \boldsymbol{x}) \rangle| + \delta_{\bar{\mathsf{N}}} && (3.4.15)\\
&\le 2\,(1 + \delta_k(\boldsymbol{A}))\,\varepsilon + \delta_{\bar{\mathsf{N}}}. && (3.4.16)
\end{aligned}$$

Since this inequality holds for all $\boldsymbol{x} \in \Sigma_k$, we obtain that

$$\delta_k(\boldsymbol{A}) \;\le\; 2\,(1 + \delta_k(\boldsymbol{A}))\,\varepsilon + \delta_{\bar{\mathsf{N}}}, \tag{3.4.17}$$

from which the target inequality follows. □

The next task is to construct a set $\bar{\mathsf{N}}$ which has the desired good properties. We call a set N an ε-net for a given set S if

$$\forall\, \boldsymbol{x} \in \mathsf{S}, \quad \exists\, \bar{\boldsymbol{x}} \in \mathsf{N} \quad \text{such that} \quad \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \le \varepsilon. \tag{3.4.18}$$

Let

$$\mathsf{B}(\boldsymbol{x}, r) = \left\{ \boldsymbol{z} \in \mathbb{R}^d \mid \|\boldsymbol{z} - \boldsymbol{x}\|_2 \le r \right\} \tag{3.4.19}$$

denote the ℓ^2 ball of center $\boldsymbol{z}$ and radius r, in $\mathbb{R}^d$. The following clever argument shows that there exists an ε net for the ℓ^2 ball $B(0,1)$ of size at most $(3/\varepsilon)^d$. It uses the fact that if $\mathsf{S} \subset \mathbb{R}^d$ is a set, and

$$\alpha \mathsf{S} = \{ \alpha \boldsymbol{s} \mid \boldsymbol{s} \in \mathsf{S} \} \tag{3.4.20}$$

denotes its α dilation, then

$$\mathrm{vol}(\alpha \mathsf{S}) \;\le\; \alpha^d \mathrm{vol}(\mathsf{S}). \tag{3.4.21}$$

See Figure 3.12 for a visualization of this.

Lemma 3.4.4 (ε-nets for the ball)**.** *There exists an ε-net for the unit ball $\mathsf{B}(0,1) \subset \mathbb{R}^d$ of size at most $(3/\varepsilon)^d$.*

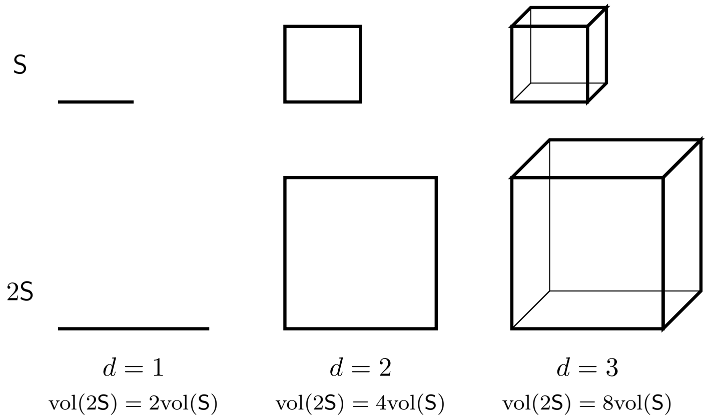


Figure 3.12. **Volumes scale as** α^d**.**

Proof. Call a set ε-separated if every pair of distinct points in M has distance at least ε. Let $\mathsf{N} \subset \mathsf{B}(0,1)$ be a *maximal* ε-separated set. Here, maximal means that it is not contained in any larger ε-separated set.

We claim that N is an ε-net for $\mathsf{B}(0,1)$. Indeed, if it is not an ε-net, then there exists some point $\boldsymbol{x} \in \mathsf{B}(0,1)$ with distance greater than ε to each element of N. Adding $\boldsymbol{x}$ to N, we obtain a larger ε-separated set, contradicting maximality of N.

Since N is ε-separated, the balls $\mathsf{B}(\boldsymbol{x}, \varepsilon/2)$ and $\mathsf{B}(\boldsymbol{x}', \varepsilon/2)$ are disjoint, for any pair of distinct elements $\boldsymbol{x} \neq \boldsymbol{x}' \in \mathsf{N}$. Moreover, the union of these balls is contained in $\mathsf{B}(\mathbf{0}, 1+\varepsilon/2)$. Thus,

$$|\mathsf{N}| \operatorname{vol}(\mathsf{B}(\mathbf{0}, \varepsilon/2)) \;\le\; \operatorname{vol}(\mathsf{B}(\mathbf{0}, 1+\varepsilon/2)). \tag{3.4.22}$$

Hence,

$$\begin{aligned} |\mathsf{N}| &\le \frac{\operatorname{vol}(\mathsf{B}(\mathbf{0}, 1+\varepsilon/2))}{\operatorname{vol}(\mathsf{B}(\mathbf{0}, \varepsilon/2))} && (3.4.23)\\ &= \left(\frac{1+\varepsilon/2}{\varepsilon/2}\right)^d && (3.4.24)\\ &= (1+2/\varepsilon)^d && (3.4.25)\\ &\le (3/\varepsilon)^d && (3.4.26) \end{aligned}$$

as desired. □

To construct our target set $\bar{\mathsf{N}}$, we simply consider each support pattern I of size k individually. There are $\binom{n}{k}$ such patterns. For each pattern, we use the previous lemma to build an ε-net N for the unit ball of vectors of ℓ^2 norm at most one, whose support is contained in I. Each of these nets has size at most $(3/\varepsilon)^k$. So, finally, we obtain

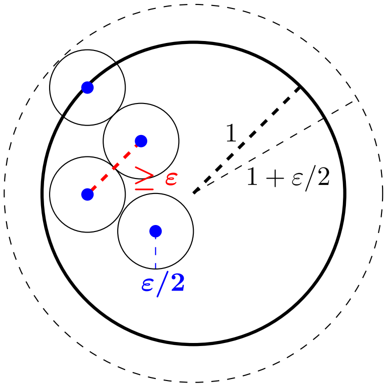


Figure 3.13. **Volume Calculation for an ε-net**. An ε-separated set. The interiors of $\varepsilon/2$ balls around the points do not intersect. The union of the $\varepsilon/2$ balls is contained in an $(1+\varepsilon/2)$-ball.

Lemma 3.4.5. *There exists an ε-net $\bar{\mathsf{N}}$ for Σ_k satisfying the two properties required in Lemma 3.4.3, with*

$$\left|\bar{\mathsf{N}}\right| \;\le\; \exp\Big(k\log(3/\varepsilon) + k\log(n/k) + k\Big). \tag{3.4.27}$$

Proof. The construction follows the above discussion. Using the Stirling's formula,[7] we can estimate

$$\begin{aligned}\left|\bar{\mathsf{N}}\right| &\le (3/\varepsilon)^k \binom{n}{k} &(3.4.28)\\ &\le (3/\varepsilon)^k \left(\frac{ne}{k}\right)^k, &(3.4.29)\end{aligned}$$

as desired. □

Union bound.

Proof. For each $\boldsymbol{x} \in \bar{\mathsf{N}}$, $\boldsymbol{Ax}$ is a random vector with entries independent $\mathcal{N}(0, 1/m)$. We have

$$\mathbb{P}\left[\Big|\|\boldsymbol{Ax}\|_2^2 - 1\Big| > t\right] \;\le\; 2\exp(-mt^2/8). \tag{3.4.30}$$

Hence, summing over all elements of $\bar{\mathsf{N}}$, we have

$$\begin{aligned}\mathbb{P}\left[\delta_{\bar{\mathsf{N}}} > t\right] &\le 2\left|\bar{\mathsf{N}}\right|\exp\left(-mt^2/8\right) &(3.4.31)\\ &\le 2\exp\Big(-\frac{mt^2}{8} + k\log\big(\frac{n}{k}\big) + k(\log(\frac{3}{\varepsilon})+1)\Big). &(3.4.32)\end{aligned}$$

[7] Stirling's formula gives the bounds for factorials: $\sqrt{2\pi k}\big(\frac{k}{e}\big)^k \le k! \le e\sqrt{k}\big(\frac{k}{e}\big)^k$.

On the complement of the event $\delta_{\bar{\mathsf{N}}} > t$, we have

$$\delta_k(\boldsymbol{A}) < \frac{2\varepsilon + t}{1 - 2\varepsilon}. \tag{3.4.33}$$

Setting $\varepsilon = \delta/8$, $t = \delta/4$, and ensuring that $m \geq Ck \log(n/k)/\delta^2$ for sufficiently large numerical constant C, we obtain the result. □

In the above derivation, especially from equation (4.2.67), we see that a slight more tight bound for m is of the form

$$m \geq 128k \log(n/k)/\delta^2 + (\log(24/\delta) + 1)k/\delta^2 \doteq C_1 k \log(n/k) + C_2 k.$$

However, for a small δ, the constants C_1 and C_2 can be rather large. Although qualitatively this bound is in the right form, it however does not reflect exactly when ℓ^1 minimization works. In the work of [Rudelson and Vershynin, 2008], a much tighter bound for m is given as:

$$m \geq 8k \log(n/k) + 12k.$$

This is one of the best known bounds given through the RIP properties of Gaussian matrices. Nevertheless, as we will see later, using more advanced tools, ultimately we will be able to derive for Gaussian matrices a precise condition that characterizes the "phase-transition" behavior for the success of ℓ^1 minimization that we can observe through simulations.

3.4.3 RIP of Non-Gaussian Matrices

In many applications of interest, the matrix $\boldsymbol{A}$ cannot be assumed to be iid Gaussian. Perhaps surprisingly, often the theory developed for the Gaussian model is predictive of the behavior of ℓ^1 minimization in other models. However, it is still desirable to have a precise understanding (and corresponding mathematical guarantees) to describe what happens when the model is not so homogeneous.

Random submatrices of a unitary matrix.

One model that occurs quite often posits that we generate $\boldsymbol{A}$ by randomly sampling some rows of an orthognonal matrix (in the real case) or a unitary matrix (in the complex case). Actually, we have already seen such a model in our brief discussion of MRI applications. There, we generated $\boldsymbol{A}$ as a row submatrix of $\boldsymbol{F\Psi}$, where $\boldsymbol{F}$ was the DFT matrix, and $\boldsymbol{\Psi} \in \mathbb{C}^{n \times n}$ was a matrix whose columns formed an orthonormal wavelet basis for $\mathbb{C}^{n \times n}$. Since both $\boldsymbol{F}$ and $\boldsymbol{\Psi}$ were unitary, their product is unitary. In the work [Candès et al., 2006], it has been shown that for a given k-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$, if A randomly takes $m = O(k \log n)$ rows of a unitary matrix, then with high probability the ℓ^1 minimization $\min \|\boldsymbol{x}\|_1$ s.t. $\boldsymbol{y} = \boldsymbol{Ax}$ recovers the sparse vector. However, this result does not implies that with the same $\boldsymbol{A}$, the ℓ^1 minimization succeeds for all k-sparse vectors.[8]

The following theorem, according to [Rudelson and Vershynin, 2008], shows that if we sample a random row submatrix from a unitary matrix, it also has RIP with high probability, provided enough rows are chosen. We know that for a matrix satisfying the RIP condition, it is guaranteed that the associated ℓ^1 minimization succeeds for all k-sparse vectors.

Theorem 3.4.6. *Let $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ be unitary ($\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$) and Ω is a random subset of m elements from $\{1,\dots,n\}$. Suppose that*

$$\|\boldsymbol{U}\|_\infty \;\le\; \zeta/\sqrt{n}. \tag{3.4.34}$$

If

$$m \;\ge\; \frac{C\zeta^2}{\delta^2} k \log^4 n, \tag{3.4.35}$$

then with high probability, $\boldsymbol{A} = \sqrt{\frac{n}{m}}\boldsymbol{U}_{\Omega,\bullet}$ satisfies the RIP of order k, with constant $\delta_k(\boldsymbol{A}) \le \delta$.

For simplicity, we here do not give a proof to this theorem and interested readers may refer to the work of [Rudelson and Vershynin, 2008].

In our context, there are two very salient points about this result. The first is the dependence on $\|\boldsymbol{U}\|_\infty$. It is worth noting that for any unitary matrix $\boldsymbol{U}$, $\|\boldsymbol{U}\|_\infty \;\ge\; 1/\sqrt{n}$. So, the parameter ζ measures how much we lose with respect to this optimal bound. The bound is clearly achievable in some cases – our DFT matrix $\boldsymbol{F}$ has $\|\boldsymbol{F}\|_\infty = 1/\sqrt{n}$. If we are willing to interpret the result a bit, the idea that $\boldsymbol{U}$ should have uniformly bounded elements leads to a very nice intuition about sampling. Namely, if we wish to reconstruct an element that is sparse in some basis $\boldsymbol{\Psi}$, and we can take whatever linear samples $\langle \boldsymbol{f}_i, \boldsymbol{y}\rangle$ we want, we should take samples that are as *incoherent* with the basis of sparsity as possible, in the sense that

$$\left\langle \boldsymbol{f}_i, \boldsymbol{\psi}_j \right\rangle \tag{3.4.36}$$

is uniformly small. This is in contrast to our usual intuition from signal processing, which might suggest that some sort of matched filter is best here. The challenge is that there are actually an exponentially large number of potential support patterns for $\boldsymbol{x}$, and hence an exponentially large number of signals to match. If, instead, we let each (incoherent) measurement collect information across all of the basis elements, we can then, using efficient computation, decide which elements of $\boldsymbol{\Psi}$ are active.

[8] To see the difference, one can recall in the Johnson-Lindenstrauss Lemma, the task is not just to show that given any pair of points, with high probability there exists a projection that approximately preserves the distance. We need to use the union bound to show that with high probability there exists a projection that approximately preserves the distance between all pairs simultaneously.

The second salient point is that the number of measurements, $k \log^4 n$ is visually similar to the $k \log n/k$ that we saw for the Gaussian ensemble. It is currently conjectured that here $k \log n$ measurements suffice. It is currently an open problem to show this; it is considered hard, and known to connect to a number of interesting questions in probability and functional analysis. In fact, in [Rudelson and Vershynin, 2008] a more precise expression is given as: $m = O(k \log(n) \log^2(k) \log(k \log n))$. This bound, against the conjectured optimal bound, is within a $\log \log n$ factor for n and within a $\log^3 k$ factor for k.

Random convolutions.

Another model that occurs quite frequently in engineering practice involves sampling the convolution of the input signal $\boldsymbol{x}$ with some filter $\boldsymbol{r}$. Formally, we can imagine that

$$\boldsymbol{y} = \mathcal{P}_\Omega[\boldsymbol{r} * \boldsymbol{x}] = \boldsymbol{A}\boldsymbol{x}, \tag{3.4.37}$$

where $\boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{r} \in \mathbb{C}^n$, and $\Omega \subseteq [n]$ is our collection of sampling locations. Here, $*$ denotes circular convolution[9]

$$(\boldsymbol{r} * \boldsymbol{x})_i = \sum_{j=0}^{n-1} x_j r_{n-j \bmod n}. \tag{3.4.38}$$

This leads to a highly structured linear operator on $\boldsymbol{x}$ since we can represent the convolution in matrix form as

$$\boldsymbol{r} * \boldsymbol{x} \quad = \quad \begin{bmatrix} r_0 & r_{n-1} & \dots & r_2 & r_1 \\ r_1 & r_0 & r_{n-1} & & r_2 \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{n-2} & & \ddots & \ddots & r_{n-1} \\ r_{n-1} & r_{n-2} & \dots & r_1 & r_0 \end{bmatrix} \boldsymbol{x} \quad \doteq \quad \boldsymbol{R}\boldsymbol{x}. \tag{3.4.39}$$

Such a matrix $\boldsymbol{R}$ is called a circulant matrix. Hence we can view the sampling matrix A as taking a subset of rows of the circulant matrix $\boldsymbol{R}$, that is $\boldsymbol{A} = \boldsymbol{R}_{\Omega,\bullet}$.

The filter $\boldsymbol{r}$ can be rather general as well. For instance, it could as simple as a random Rademacher vector, i.e.a random vector with independent entries distributed according to $\mathbb{P}(r_i = \pm 1) = 0.5$, or it could be a random vector with independent zero-mean, subgaussian random variables of variance one. The exact randomness of $\boldsymbol{r}$ is not critical.

For this model, the work of [Krahmer et al., 2012] has shown that essentially the following statement is true:

[9] This is not really essential.

Theorem 3.4.7. *Let $\Omega \subseteq \{1, \ldots, n\}$ be any fixed subset of size $|\Omega| = m$. Then if*

$$m \geq \frac{Ck \log^2 k \log^2 n}{\delta^2}, \tag{3.4.40}$$

then with high probability, $\boldsymbol{A}$ has RIP of order k with $\delta_k(\boldsymbol{A}) \leq \delta$.

Notice that the above statement is rather strong in the following sense: Firstly, it states that even for a highly structured sampling matrix (a circulant matrix versus a random Gaussian matrix studied in previous section), we only lose a small factor of $\log^2 k \log n$ in the required number of samples. Secondly, it claims that any subset of rows of $\boldsymbol{R}$ has the RIP property, not just a random subset with high probability. Thirdly, the RIP property ensures recoverability of any k-sparse vectors $\boldsymbol{x}$ uniformly not just for a fixed k-sparse vector. It has been shown in [Rauhut, 2009] that, if one relaxes the uniform recoverability requirement, considering only a fixed k-sparse vector, it can be recovered via ℓ^1-minimization from a partial random circulant matrix with $m \geq Ck \log^2 n$ measurements. This bound is slightly better than the one given in the theorem, but it is not uniform for all k-sparse vectors.

3.5 Noisy Observations or Approximate Sparsity

Thus far, we have been very idealistic in our model. We have assumed that the target $\boldsymbol{x}_o$ is perfectly sparse, and that there is no noise in the measurements, so $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$ exactly. These assumptions are clearly violated in many practical applications. In practice, the observation $\boldsymbol{y}$ is usually perturbed by some amount of noise $\boldsymbol{z}$, which we assume to be small:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}, \qquad \|\boldsymbol{z}\|_2 \leq \varepsilon. \tag{3.5.1}$$

In other practical scenarios, the ground truth signal $\boldsymbol{x}_o$ may not be perfectly sparse and may be only approximately so.

This motivates two natural questions. First, on the practical side, is it possible to modify our approaches to be stable under noise or for imperfect sparse signals? Second, what should we expect of their performance? Do the conditions and guarantees we introduced in previous sections remain meaningful?

To clearly state our assumptions and goals, we can consider the following three scenarios (or some combination of them):

- **Deterministic (worst case) noise**: $\boldsymbol{z}$ is bounded: $\|\boldsymbol{z}\|_2 \leq \varepsilon$, and ε is known.

- **Stochastic noise**: $\boldsymbol{z} \sim_{iid} \mathcal{N}(0, \frac{\sigma^2}{m})$. Notice that under this random model, a typical noise vector $\boldsymbol{z}$ is of norm $\|\boldsymbol{z}\|_2 \approx \sigma$.[10] Gaussian noise is a very natural assumption; the results obtained here also extend to other noise models.
- **Inexact sparsity**: $\boldsymbol{x}_o$ is not perfectly sparse. Technically speaking, this is not noise, but rather a violation of our sparse modeling assumption. In this scenario, it may be meaningful to assume that $\boldsymbol{x}_o$ is *close* to a k-sparse vector. We can formalize this by letting $[\boldsymbol{x}_o]_k$ denote a best k-term approximation to $\boldsymbol{x}_o$:

  $$[\boldsymbol{x}_o]_k \in \arg \min_{\|\boldsymbol{z}\|_0 \le k} \|\boldsymbol{x}_o - \boldsymbol{z}\|_2^2 . \qquad (3.5.2)$$

  This just keeps the k largest elements of $\boldsymbol{x}_o$. $\boldsymbol{x}_o$ is said to be "approximately sparse" if $\|\boldsymbol{x}_o - [\boldsymbol{x}_o]_k\|$ is small.

In all of these scenarios, we might hope to still "recover" a sparse estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}_o$ in some sense. There are (perhaps) three natural senses to consider:

- **Estimation**: Is $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ small?
- **Prediction**: Is $\boldsymbol{A}\hat{\boldsymbol{x}} \approx \boldsymbol{A}\boldsymbol{x}_o$?
- **Support recovery**: Is $\operatorname{supp}(\hat{\boldsymbol{x}}) = \operatorname{supp}(\boldsymbol{x}_o)$?

For engineering practice, we often care about either estimating the signal $\boldsymbol{x}_o$ (for sensing problems) or recovering its support $\operatorname{supp}(\boldsymbol{x}_o)$ (for recognition problems). Nevertheless, statisticians sometimes also care about the prediction error $\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)$.

In the following subsections, discuss results on stable estimation under (i) deterministic noise, (ii) stochastic noise and (iii) deterministic noise *and* inexact sparsity. Results on support recovery are discussed briefly in Section 3.6 and in the Notes and References to this chapter.

3.5.1 Stable Recovery of Sparse Signals

In the ideal sensing model, the observation equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$ holds exactly for a sparse signal $\boldsymbol{x}_o$. In this subsection, we consider a more practical situation in which the observation $\boldsymbol{y}$ is perturbed by some amount of noise. For simplicity, we still assume the signal $\boldsymbol{x}_o$ is perfectly sparse. We can model the noise as an additive error $\boldsymbol{z}$, which we will assume to have a small magnitude:[11]

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}, \qquad \|\boldsymbol{z}\|_2 \le \varepsilon. \qquad (3.5.3)$$

[10] We scale the variance of the normal distribution by $1/m$ on purpose, so that σ is directly comparable to ϵ in the deterministic noise case.

[11] This is similar to the setting in conventional signal processing problems where we typically assume the signal to noise ratio (SNR) is large.

To recover a sparse solution from the above observation, we may extend ℓ^1 minimization to this new setting and solve

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \le \varepsilon. \end{array} \tag{3.5.4}$$

In words, this program asks us to (try to) find the sparest $\boldsymbol{x}$ that agrees with the observation up to the noise level. Almost equally popular is the Lagrangian relaxation of this problem, which introduces a penalty parameter $\lambda \ge 0$, and solves the unconstrained optimization problem

$$\text{minimize} \;\; \lambda \|\boldsymbol{x}\|_1 + \tfrac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 . \tag{3.5.5}$$

The optimization (3.5.4) is almost uniformly referred to as “Basis Pursuit Denoising” [Chen et al., 2001], while the problem (3.5.5) is almost uniformly referred to as the “Lasso (Least absolute shrinkage and selection operator)” [Tibshirani, 1996]. These two problems are completely equivalent, in the sense that there is a calibration $\lambda \leftrightarrow \varepsilon$ such that if $\boldsymbol{x}$ is a solution to the Lasso problem for some choice of λ, then there exists an ε such that $\boldsymbol{x}$ is also a solution to the BPDN problem with parameter ε, and conversely, whenever $\boldsymbol{x}$ is a solution to BPDN with parameter ε, there exists a corresponding λ such that $\boldsymbol{x}$ also solves the Lasso problem with parameter λ. So, from a theoretical perspective, these two problems are completely equivalent.

On the other hand, from a practical perspective, they may be quite different, since the calibration $\lambda \leftrightarrow \varepsilon$ depends on the problem data $(\boldsymbol{y}, \boldsymbol{A})$, and no explicit form is known. In some situations, it may be easier to tune λ than ε, or vice versa. In particular, in situations in which the norm of the noise is known or can be estimated, the BPDN formulation may be more attractive, since its parameter can be set to be the noise level.[12] The optimal choice of the regularization parameter λ (or ε) is a surprisingly tricky issue in practice. In general, we have to either use generic statistical rules such as cross validation, or resort to theoretical analysis to get some insight into what scalings make sense.

Despite their conceptual equivalence, these problems may require rather different optimization techniques. In Part II, we will discuss in more details about how to solve both (and how to solve many related problems!).

Deterministic Noise.

To account for measurement noise, we can simply solve one of (3.5.4) or (3.5.5). Both are convex problems. Any global minimizer gives an estimate $\hat{\boldsymbol{x}}$. Unlike the previous two sections, under noise we cannot expect $\hat{\boldsymbol{x}} =$

[12] Historically, the Lasso is preferred by statisticians, and BPDN by engineers, although confusingly, in the original papers the names Lasso and BPDN are not used to refer to these problems, but rather different equivalent problems!

$\boldsymbol{x}_o$ exactly. However, we *can* hope that if the noise level ε is small, the estimation error $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ will also be small.

How well do we expect to do? Imagine that we somehow knew the support I of $\boldsymbol{x}_o$. In this situation, we could form another estimate $\hat{\boldsymbol{x}}'$, by setting

$$\begin{cases} \hat{\boldsymbol{x}}'(I) = (\boldsymbol{A}_I^T \boldsymbol{A}_I)^{-1} \boldsymbol{A}_I^T \boldsymbol{y}, \\ \hat{\boldsymbol{x}}'(I^c) = \mathbf{0}. \end{cases} \tag{3.5.6}$$

This is just the least squares estimate, restricted to the set I. It is not difficult to argue that it is optimal, in the sense that it minimizes over all estimators, the worst error $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ over all $\boldsymbol{x}_o$ supported on I and $\boldsymbol{z}$ of norm at most ε. This “oracle” estimator produces an estimate $\hat{\boldsymbol{x}}'$ that satisfies

$$\left\|\hat{\boldsymbol{x}}' - \boldsymbol{x}_o\right\|_2 \;\le\; \frac{\varepsilon}{\sigma_{\min}(\boldsymbol{A}_I)}, \tag{3.5.7}$$

and this bound can be tight.

So, the best we can possibly hope for in general is

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \sim c\varepsilon$$

with $c = \sigma_{\min}(\boldsymbol{A}_I)^{-1}$. As above, if we restrict ourselves to efficient algorithms, this is too much to hope for in general. However, can we still hope that under the same hypotheses as above,

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \le C\varepsilon? \tag{3.5.8}$$

That is to say, the solution is at least *stable*: the error in estimating $\boldsymbol{x}$ is proportional to the size ε of the perturbation, even though the constant might not be as small as when we know the oracle of the correct support of $\boldsymbol{x}_o$.

The theorem below, which is similar to that in [Candès et al., 2006],[13] makes this precise:

Theorem 3.5.1 (Stable Sparse Recovery via BPDN)**.** *Suppose that* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$*, with* $\|\boldsymbol{z}\|_2 \le \varepsilon$*, and let* $k = \|\boldsymbol{x}_o\|_0$*. If* $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$*, then any solution* $\hat{\boldsymbol{x}}$ *to the optimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \le \varepsilon. \end{array} \tag{3.5.9}$$

satisfies

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \le C\varepsilon. \tag{3.5.10}$$

Here, C *is a constant which depends only on* $\delta_{2k}(\boldsymbol{A})$ *(and not on the noise level* ε*).*

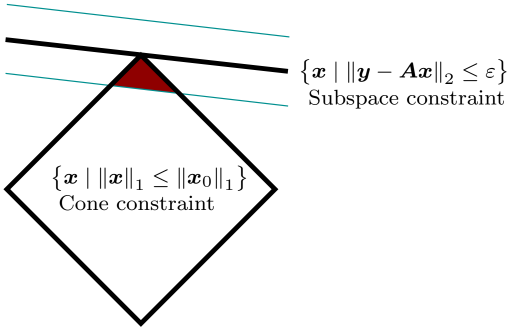

$\{\boldsymbol{x} \mid \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \varepsilon\}$
Subspace constraint
$\{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}_0\|_1\}$
Cone constraint

Figure 3.14. Geometry of the proof of Theorem 3.5.1.

Proof. Because $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o\|_2 = \|\boldsymbol{z}\|_2 \leq \varepsilon$. Since $\hat{\boldsymbol{x}}$ is feasible, we have $\|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 \leq \varepsilon$ as well. Using the triangle inequality,

$$\begin{aligned} \|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2 &= \|(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}) - (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o)\|_2 \\ &\leq \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 + \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o\|_2 \\ &\leq 2\varepsilon. \end{aligned}$$

Let $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o$, we have $\|\boldsymbol{A}\boldsymbol{h}\|_2 \leq 2\epsilon$. Geometrically, this means that the perturbation $\boldsymbol{h}$ must be close to the null space of $\boldsymbol{A}$.

Because $\boldsymbol{x}_o$ is feasible for the optimization problem, and $\hat{\boldsymbol{x}}$ is optimal, $\hat{\boldsymbol{x}}$ must have a lower objective function value than $\boldsymbol{x}_o$:

$$\|\hat{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}_o\|_1 . \tag{3.5.11}$$

Let I denote the support of $\boldsymbol{x}_o$. We have

$$\begin{aligned} \|\boldsymbol{x}_o\|_1 &\geq \|\boldsymbol{x}_o + \boldsymbol{h}\|_1 \\ &\geq \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{h}_I\|_1 + \|\boldsymbol{h}_{I^c}\|_1 , \end{aligned}$$

and so

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1 . \tag{3.5.12}$$

Geometrically, this means that $\hat{\boldsymbol{x}}$ lives in an ℓ^1 ball of radius $\|\boldsymbol{x}_o\|_1$, centered at the origin. Locally, this set looks like a convex cone (the "descent cone" of the ℓ^1 norm), hence the constraint $\|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1$ is also known as a "cone constraint". It describes the set of all possible perturbations of $\hat{\boldsymbol{x}}$ from $\boldsymbol{x}_o$ that would decrease the value of the objective function. The geometric intuition behind the two constraints on the perturbation $\boldsymbol{h}$ is shown in Figure 3.14.

Note that the matrix $\boldsymbol{A}$ satisfies RIP. According to Theorem 3.3.11, we know that if $\delta_{2k} < \sqrt{2} - 1$, $\boldsymbol{A}$ satisfies the restricted strong convexity

[13] The condition on RIP constant in [Candès et al., 2006] was $\delta_{4k}(\boldsymbol{A}) < 1/4$, which is more restrictive than the one shown here.

property with constant $\alpha = 1$ (which is the case for the restriction condition (3.5.12) on $\boldsymbol{h}$ above). Therefore, we have

$$\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq \mu \|\boldsymbol{h}\|_2^2 \tag{3.5.13}$$

for some $\mu > 0$. Combining this with $\|\boldsymbol{A}\boldsymbol{h}\|_2 \leq 2\epsilon$, we have

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 = \|\boldsymbol{h}\|_2 \leq \frac{2}{\sqrt{\mu}}\epsilon. \tag{3.5.14}$$

Choosing $C = \frac{2}{\sqrt{\mu}}$ completes the proof. □

Notice that in the above proof, the constant C can be rather large if μ is very small. According to the proof of Theorem 3.3.11, we know

$$\sqrt{\mu} = \frac{1 - \delta_{2k}(1 + \sqrt{2})}{\sqrt{2(1 + \delta_{2k})}}.$$

The quantity μ becomes small if δ_{2k} is close to $\sqrt{2} - 1$. Therefore, if we do not want the constant C in the above theorem to be too large, we need to ensure that δ_{2k} is significantly small than $\sqrt{2} - 1$. However, no matter how small δ_{2k} is, we always have $\sqrt{\mu} < 1/\sqrt{2}$. Hence, based on this proof, the smallest that the constant C can be in the theorem is $2\sqrt{2}$.

Random Noise.

Above, we have shown that for any additive noise $\boldsymbol{z}$ in the observation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, we can estimate $\boldsymbol{x}_o$ with an error of size controlled by $C\,\|\boldsymbol{z}\|_2$ for some constant C. Based on our discussion before the theorem, this error bound is already close to the best possible.

For random noise, we might hope that if $m \gg k$, most of the energy of $\boldsymbol{z}$ would "miss" the k-dimensional subspace range$(\boldsymbol{A}_I)$. If so, the accuracy in the estimated $\hat{\boldsymbol{x}}$ can improve as m grows. More precisely, the coefficient C in the error bound $C\,\|\boldsymbol{z}\|_2$ decreases as m increases. This turns out to be the case. For simplicity, we here state a theorem for random $\boldsymbol{A}$.[14] More precisely, we assume that the measurement model:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}. \tag{3.5.15}$$

where $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{x}_o$ k-sparse, and the matrix $\boldsymbol{A} \sim_{iid} \mathcal{N}(0, \frac{1}{m})$ and $\boldsymbol{z} \sim_{iid} \mathcal{N}(0, \frac{\sigma^2}{m})$. Notice that in the study of the deterministic case, we have assumed the measurement matrix $\boldsymbol{A}$ is a matrix that satisfies RIP conditions. Hence the norm of the columns of $\boldsymbol{A}$ there is typically normalized to one. Here the scaling factor $\frac{1}{m}$ in the variance is to ensure the columns of $\boldsymbol{A}$ is typically of length one and the noise vector of length σ so that the model

[14] An analogous result holds for $\boldsymbol{A}$ satisfying the RIP.

and the results will be directly comparable to those for the deterministic case.[15]

As we have discussed earlier, with noisy measurements, we could find an estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}_o$ that strikes a balance between sparsity and minimizing the error. In particular, we would like to solve the following Lasso program for $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \tfrac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda_m \left\| \boldsymbol{x} \right\|_1 . \tag{3.5.16}$$

As usual, for convenience, we let $I = \operatorname{supp}(\boldsymbol{x}_o)$, let I^c denote its complement, and $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o \in \mathbb{R}^n$ the difference between the estimate and the ground truth. We also define $L(\boldsymbol{x}) = \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2$. Notice that $\nabla L(\boldsymbol{x}) = -\boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$ and in particular $\nabla L(\boldsymbol{x}_o) = -\boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o) = -\boldsymbol{A}^*\boldsymbol{z}$ according to (3.5.15).

We want to know how small the difference $\|\boldsymbol{h}\| = \|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|$ is for a given λ_m. First we show that for a properly chosen λ_m, the difference vector $\boldsymbol{h}$ is highly *restricted* in the way that $\|\boldsymbol{h}_{I^c}\|_1 \le \alpha \|\boldsymbol{h}_I\|_1$ for some constant α, i.e., the error off the support I of $\boldsymbol{x}_o$ is controlled by that on I.[16] More precisely, we have the following lemma.

Lemma 3.5.2. *For the optimization problem (3.5.16), if we choose the regularization parameter $\lambda_m \ge c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$, then $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o$ satisfies the cone condition:*

$$\|\boldsymbol{h}_{I^c}\|_1 \le \frac{c+1}{c-1} \cdot \|\boldsymbol{h}_I\|_1 \tag{3.5.17}$$

where I is the support of the sparse $\boldsymbol{x}_o$.

Proof. Note that the difference between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}_o$ is related to the difference between the values of the objective function in (3.5.16). Since $\hat{\boldsymbol{x}}$ minimizes the objective function, we have:

$$\begin{aligned} 0 &\ge L(\hat{\boldsymbol{x}}) + \lambda_m \|\hat{\boldsymbol{x}}\|_1 - L(\boldsymbol{x}_o) - \lambda_m \|\boldsymbol{x}_o\|_1 \\ &\ge \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o \rangle + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\ge -\left| \langle \boldsymbol{A}^*\boldsymbol{z}, \boldsymbol{h} \rangle \right| + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\ge -\|\boldsymbol{A}^*\boldsymbol{z}\|_\infty \|\boldsymbol{h}\|_1 + \lambda_m (\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1), \end{aligned} \tag{3.5.18}$$

where the second inequality we used the fact that $L(\boldsymbol{x})$ is a convex function. It remains to see how the two terms in the last inequality interact. Obviously we need to have a good idea about the value of $\|\boldsymbol{A}^*\boldsymbol{z}\|_\infty$. This is where we need to resort to results about measure concentration of high-dimensional statistics.

[15] The variance σ replaces the role of ϵ in Theorem 3.5.1.

[16] Notice that a similar restriction on $\boldsymbol{h}$ was derived in (3.5.12). There the constant is $\alpha = 1$ and as we will soon see, here the constant needs to be 3.

Notice that the column $\boldsymbol{a}_i$ of $\boldsymbol{A}$ is typically of norm $\|\boldsymbol{a}_i\|_2 \approx 1$. Hence we here may assume the columns of $\boldsymbol{A}$ are all normalized to one. Therefore $\boldsymbol{a}_i^* \boldsymbol{z}$ is a Gaussian random variable of variance σ^2/m. We have

$$\mathbb{P}\left[|\boldsymbol{a}_i^* \boldsymbol{z}| \geq t\right] \leq 2 \exp\left(-\frac{mt^2}{2\sigma^2}\right). \tag{3.5.19}$$

By union bound on the n columns, we have

$$\mathbb{P}\left[\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \geq t\right] \leq 2 \exp\left(-\frac{mt^2}{2\sigma^2} + \log n\right). \tag{3.5.20}$$

As we may see, as long as we choose t^2 to be in the order of $C\frac{\sigma^2 \log n}{m}$ for a large enough constant C, the exponent will be negative and the event $\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \geq t$ will be of low probability. In particular we may choose $t^2 = 4\frac{\sigma^2 \log n}{m}$, and we know that with high probability at least $1 - cn^{-1}$, we have

$$\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \leq 2\sigma\sqrt{\frac{\log n}{m}}.$$

So to make the two terms in (3.5.18) comparable in scale, it is natural to choose λ_m of the scale $\sigma\sqrt{\frac{\log n}{m}}$. In particular, we choose $\lambda_m \geq c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$. Then from the last inequality of (3.5.18), we have

$$\begin{aligned}
0 &\geq -\|\boldsymbol{A}^* \boldsymbol{z}\|_\infty \|\boldsymbol{h}\|_1 + \lambda_m(\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\
&\geq -\frac{\lambda_m}{c}\|\boldsymbol{h}\|_1 + \lambda_m(\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\
&\geq -\frac{\lambda_m}{c}\|\boldsymbol{h}_I\|_1 - \frac{\lambda_m}{c}\|\boldsymbol{h}_{I^c}\|_1 + \lambda_m\|\boldsymbol{h}_{I^c}\|_1 - \lambda_m\|\boldsymbol{h}_I\|_1 \\
&= \lambda_m\left(\left(1 - \frac{1}{c}\right)\|\boldsymbol{h}_{I^c}\|_1 - \left(1 + \frac{1}{c}\right)\|\boldsymbol{h}_I\|_1\right),
\end{aligned} \tag{3.5.21}$$

where in the second to last inequality we used the fact that $\boldsymbol{x}_o$ is zero on I^c and $\|\hat{\boldsymbol{x}}_I\|_1 - \|\boldsymbol{x}_{oI}\|_1 \geq -\|\boldsymbol{h}_I\|_1$. Therefore we have

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \frac{c+1}{c-1} \cdot \|\boldsymbol{h}_I\|_1. \tag{3.5.22}$$

Notice that if we choose c to be large, $\frac{c+1}{c-1}$ can be arbitrarily close to 1. □

As we have discussed in the deterministic case, since $\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2 \leq \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 + \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o\|_2$, it suggests that $\|\boldsymbol{A}\boldsymbol{h}\|_2$ is typically very small and of the scale $C\sigma$. If the norm $\|\boldsymbol{A}\boldsymbol{h}\|_2$ upper bounds the norm $\|\boldsymbol{h}\|_2$, then the estimate is stable. Of course, this cannot be true for any $\boldsymbol{h} \in \mathbb{R}^n$ since the matrix $\boldsymbol{A}$ is typically severely under-determined and for any $\boldsymbol{h}$ in the null space of $\boldsymbol{A}$, the norm $\|\boldsymbol{A}\boldsymbol{h}\|$ is zero but the norm $\|\boldsymbol{h}\|$ can be arbitrarily large.

Nevertheless, due to the above lemma, we could hope that for $\boldsymbol{h}$ that satisfies the cone restriction $\|\boldsymbol{h}_{I^c}\|_1 \leq \alpha\|\boldsymbol{h}_I\|_1$ for $\alpha = \frac{c+1}{c-1}$, $\|\boldsymbol{A}\boldsymbol{h}\|_2$ controls

$\|\boldsymbol{h}\|_2$. Due to Theorem 3.3.5, we know that with high probability, $\boldsymbol{A}$ as a random Gaussian matrix satisfies RIP. Then Theorem 3.3.11 ensures that when $\boldsymbol{h}$ is restricted in such a cone, $\|\boldsymbol{A}\boldsymbol{h}\|_2$ controls the norm $\|\boldsymbol{h}\|_2$. This leads to the following theorem.[17]

Theorem 3.5.3 (Stable Sparse Recovery via Lasso)**.** *Suppose that* $\boldsymbol{A} \sim_{iid} \mathcal{N}(0, \frac{1}{m})$*, and* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$*, with* $\boldsymbol{x}_o$ *k-sparse and* $\boldsymbol{z} \sim_{iid} \mathcal{N}(0, \frac{\sigma^2}{m})$*. Solve the Lasso*

$$\text{minimize} \;\; \tfrac{1}{2} \left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda_m \left\|\boldsymbol{x}\right\|_1, \tag{3.5.23}$$

with regularization parameter $\lambda_m = c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$ *for a large enough c. Then with high probability,*

$$\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\right\|_2 \;\le\; C'\sigma\sqrt{\frac{k\log n}{m}}. \tag{3.5.24}$$

Generally, we are interested in the regime $m \geq k\log n$, because this is when the measurement matrix $\boldsymbol{A}$ satisfies RIP (due to Theorem 3.3.5. The above theorem indicates that in this case, we actually do much better under random noise than under deterministic noise: the estimation error in the random case can be the noise norm σ scaled by a diminishing factor[18] whereas in the deterministic case the error is the noise norm ϵ scaled by a constant factor (see Theorem 3.5.1 for comparison).

Proof. With $L(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2$, we have

$$L(\hat{\boldsymbol{x}}) = L(\boldsymbol{x}_o) + \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o\rangle + \frac{1}{2}\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2^2.$$

We now use this equality to better estimate the difference between the values of the objective function at $\hat{\boldsymbol{x}}$ and at $\boldsymbol{x}_o$ than that done in (3.5.18):

$$\begin{aligned} 0 \;&\geq\; L(\hat{\boldsymbol{x}}) + \lambda_m\|\hat{\boldsymbol{x}}\|_1 - L(\boldsymbol{x}_o) - \lambda_m\|\boldsymbol{x}_o\|_1 \\ &\geq\; \frac{1}{2}\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2^2 + \langle \nabla L(\boldsymbol{x}_o), \hat{\boldsymbol{x}} - \boldsymbol{x}_o\rangle + \lambda_m(\|\hat{\boldsymbol{x}}\|_1 - \|\boldsymbol{x}_o\|_1) \\ &\geq\; \frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda_m\left(\left(1 - \frac{1}{c}\right)\|\boldsymbol{h}_{I^c}\|_1 - \left(1 + \frac{1}{c}\right)\|\boldsymbol{h}_I\|_1\right) \qquad (3.5.25) \end{aligned}$$

where the last inequality follows exactly the same derivation that we have done in (3.5.18) and (3.5.21) for other terms without the term $\frac{1}{2}\|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2^2 = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2$.

From the last inequality we have

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \leq \lambda_m\left(1 + \frac{1}{c}\right)\|\boldsymbol{h}_I\|_1.$$

[17] This result and its proof essentially follows that of Candes+Tao'07, Bickel+Ritov+Tsybakov'07, and Negahban+Ravikumar+Wainwright+Yu'12.

[18] As $\sqrt{\frac{k\log n}{m}}$ can be chosen to be arbitrarily small

According to Theorem 3.3.5 and Theorem 3.3.11, with high probability, the random Gaussian matrix $\boldsymbol{A}$ satisfies the restricted strong convexity property, we have $\|\boldsymbol{A}\boldsymbol{h}\|_2^2 \geq \mu\|\boldsymbol{h}\|_2^2$ for some constant μ.[19] Also from the relationship between 1-norm and 2-norm, we have $\|\boldsymbol{h}_I\|_1 \leq \sqrt{k}\|\boldsymbol{h}_I\|_2 \leq \sqrt{k}\|\boldsymbol{h}\|_2$. Finally, with the choice $\lambda_m = c \cdot 2\sigma\sqrt{\frac{\log n}{m}}$, the above inequality leads to:

$$\begin{aligned} &\frac{\mu}{2}\|\boldsymbol{h}\|_2^2 \leq 2(c+1)\sigma\sqrt{\frac{k\log n}{m}}\|\boldsymbol{h}\|_2 \\ \Rightarrow \quad &\|\boldsymbol{h}\|_2 \leq C'\sigma\sqrt{\frac{k\log n}{m}} \end{aligned} \tag{3.5.26}$$

for some constant $C' = \frac{4(c+1)}{\mu} \in \mathbb{R}_+$. □

The error bound given in the above theorem is actually nearly optimal as it is close to the best error that one can achieve by considering all possible estimators:

Theorem 3.5.4 (Candes+Davenport'12)**.** *Suppose that we will observe* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$*. Set*

$$M^\star(\boldsymbol{A}) = \inf_{\hat{\boldsymbol{x}}} \sup_{\|\boldsymbol{x}\|_0 \leq k} \mathbb{E}\,\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\,. \tag{3.5.27}$$

Then for any $\boldsymbol{A}$ *with* $\|\boldsymbol{e}_i^*\boldsymbol{A}\|_2 \leq \sqrt{n}$ *for each* i*, we have*

$$M^\star(\boldsymbol{A}) \;\geq\; C\sigma^2\frac{k\log(n/k)}{m}. \tag{3.5.28}$$

Proof of this theorem is beyond the scope of this book; we refer interested readers to the original paper for a proof. According to Theorem 3.5.3, the error bound $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 \sim O(\sigma^2\frac{k\log n}{m})$ achieved by Lasso is within a difference of $O(\sigma^2\frac{k\log k}{m})$ from the best achievable bound above. When $m \gg k$, such a difference is negligible.

3.5.2 Recovery of Inexact Sparse Signals

In all the above analysis, we have assumed that in the observation model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, the signal $\boldsymbol{x}_o$ is perfectly k-sparse. In many cases, $\boldsymbol{x}_o$ might not be so sparse and even all entries could be nonzero. Then a question naturally arises: for $\boldsymbol{x}_o$ is close to a k-sparse signal, can we still expect good recovery performance in some sense?

Let $[\boldsymbol{x}_o]_k$ be the best k-sparse signal that approximates $\boldsymbol{x}_o$. Then we can rewrite the observation model as:

$$\boldsymbol{y} = \boldsymbol{A}[\boldsymbol{x}_o]_k + \boldsymbol{A}(\boldsymbol{x}_o - [\boldsymbol{x}_o]_k) + \boldsymbol{z}.$$

[19] Notice that μ depends on the RIP constant $\delta_{2k}(\boldsymbol{A})$ and the constant $C = \frac{c+1}{c-1}$ of the cone restriction.

Strictly speaking the term $\boldsymbol{w} = \boldsymbol{A}(\boldsymbol{x}_o - [\boldsymbol{x}_o]_k)$ is not noise. It is more of a deviation from our idealistic sparse signal assumption. But we may view it as introducing a deterministic error to the observation. Hence, if the norm of $\boldsymbol{w}$ is small, we should expect to obtain an estimate $\hat{\boldsymbol{x}}$ whose error from $\boldsymbol{x}_o$ is proportional to this norm.

The following is a typical result on estimation with inexact sparsity, which also allows deterministic noise.[20]

Theorem 3.5.5 (Candes'09,Candes+Romberg+Tao'06)**.** *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, with $\|\boldsymbol{z}\|_2 \le \varepsilon$. Let $\hat{\boldsymbol{x}}$ solve the basis pursuit denoising problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \varepsilon. \end{array} \tag{3.5.29}$$

Then for any k such that $\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$,

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \;\le\; C\frac{\|\boldsymbol{x}_o - [\boldsymbol{x}_o]_k\|_1}{\sqrt{k}} + C'\varepsilon \tag{3.5.30}$$

for some constants C and C' which only depend on $\delta_{2k}(\boldsymbol{A})$.

How should we interpret this result? One way of reading it is to say that if we are working in a regime where noise-free sparse recovery would have succeeded ($\delta_{2k}(\boldsymbol{A}) < \sqrt{2} - 1$), then even if our modeling assumptions are violated (due to the introduction of noise and inexact sparsity), we can still *stably* estimate $\boldsymbol{x}_o$. Moreover, the error in our estimate is proportional to the degree to which our assumptions are violated and proportional to the noise level. When the original signal $\boldsymbol{x}_o$ is indeed k-sparse, we have $\boldsymbol{x}_o - [\boldsymbol{x}_o]_k = 0$ and the above result reduces to the deterministic noise case, i.e. Theorem 3.5.1.

Proof. As usual, we denote $\boldsymbol{h} = \hat{\boldsymbol{x}} - \boldsymbol{x}_o$. We also denote the support of $[\boldsymbol{x}_o]_k$ as I so that we have $[\boldsymbol{x}_o]_k = \boldsymbol{x}_{oI}$. Because $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_o\|_2 = \|\boldsymbol{z}\|_2 \le \varepsilon$. Since $\hat{\boldsymbol{x}}$ is feasible, we have $\|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 \le \varepsilon$ as well. Using the triangle inequality,

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 = \|\boldsymbol{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_o)\|_2 \le 2\varepsilon.$$

Therefore, in the inexact sparse case, the prediction error $\|\boldsymbol{A}\boldsymbol{h}\|_2$ is again bounded by the noise level.

Since $\hat{\boldsymbol{x}}$ minimizes the objective function, we have

$$\begin{array}{rcl} 0 & \le & \|\boldsymbol{x}_o\|_1 - \|\hat{\boldsymbol{x}}\|_1 \\ & = & \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{x}_{oI} + \boldsymbol{h}_I\|_1 - \|\boldsymbol{x}_{oI^c} + \boldsymbol{h}_{I^c}\|_1 \\ & \le & \|\boldsymbol{x}_o\|_1 - \|\boldsymbol{x}_{oI}\|_1 + \|\boldsymbol{h}_I\|_1 + \|\boldsymbol{x}_{oI^c}\|_1 - \|\boldsymbol{h}_{I^c}\|_1 \,. \end{array}$$

[20] In fact, similar statements hold for random noise. The proof requires slight modification to that of Theorem 3.5.3. We leave the details to the reader as an exercise.

Thus we have

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \|\boldsymbol{h}_I\|_1 + 2\|\boldsymbol{x}_{oI^c}\|_1, \tag{3.5.31}$$

where $\boldsymbol{x}_{oI^c} = \boldsymbol{x}_o - \boldsymbol{x}_{oI}$. So in the inexact sparse case, the feasible perturbation $\boldsymbol{h}$ no longer satisfies the cone condition as that in the exact sparse case (see Theorem 3.5.1). Therefore, to establish the result of this theorem, we need to modify the proof of Theorem 3.3.11 to accommodate the extra term $2\|\boldsymbol{x}_{oI^c}\|_1$ in estimating the bounds for $\|\boldsymbol{A}\boldsymbol{h}\|_2$ and $\|\boldsymbol{h}\|_2$.

The proof essentially follows the same steps as in the proof for Theorem 3.3.11. The only difference is that in places where we used to apply the cone condition $\|\boldsymbol{h}_{I^c}\|_1 \leq \alpha\|\boldsymbol{h}_I\|_1$, we now need to replace it with the new condition (3.5.31). Therefore, instead of (3.3.44), the new condition (3.5.31) implies

$$\|\boldsymbol{h}_{I^c}\|_1 \leq \sqrt{k}\|\boldsymbol{h}_I\|_2 + 2\|\boldsymbol{x}_{oI^c}\|_1 \leq \sqrt{k}\|\boldsymbol{h}_{I\cup J_1}\|_2 + 2\|\boldsymbol{x}_{oI^c}\|_1. \tag{3.5.32}$$

Substituting this into (3.3.43) to establish a bound for $\|\boldsymbol{A}\boldsymbol{h}\|_2$, we obtain

$$(1-\delta_{2k})\|\boldsymbol{h}_{I\cup J_1}\|_2 \leq \sqrt{2}\delta_{2k}\|\boldsymbol{h}_{I\cup J_1}\|_2 + 2\sqrt{2}\delta_{2k}\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}} + (1+\delta_{2k})^{1/2}\|\boldsymbol{A}\boldsymbol{h}\|_2. \tag{3.5.33}$$

This gives

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 \geq \frac{1-(1+\sqrt{2})\delta_{2k}}{(1+\delta_{2k})^{1/2}}\|\boldsymbol{h}_{I\cup J_1}\|_2 - \frac{2\sqrt{2}\delta_{2k}}{(1+\delta_{2k})^{1/2}}\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}}. \tag{3.5.34}$$

Now, to establish a bound for $\|\boldsymbol{h}\|_2$, in (3.3.50) where we have applied the cone condition in the second inequality, we also need to replace the cone condition with the new condition (3.5.31) and that gives:

$$\begin{aligned}
\|\boldsymbol{h}_{(I\cup J_1)^c}\|_2 &\leq \frac{\|\boldsymbol{h}_{I^c}\|_1}{\sqrt{k}} \leq \frac{\|\boldsymbol{h}_I\|_1 + 2\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}} && (3.5.35)\\
&\leq \|\boldsymbol{h}_I\|_2 + 2\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}} && (3.5.36)\\
&\leq \|\boldsymbol{h}_{I\cup J_1}\|_2 + 2\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}}. && (3.5.37)
\end{aligned}$$

This gives

$$\|\boldsymbol{h}\|_2 \leq \|\boldsymbol{h}_{I\cup J_1}\|_2 + \|\boldsymbol{h}_{(I\cup J_1)^c}\|_2 \leq 2\|\boldsymbol{h}_{I\cup J_1}\|_2 + 2\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}}. \tag{3.5.38}$$

Combining this with (3.5.34) and the fact that $\|\boldsymbol{A}\boldsymbol{h}\|_2 \leq 2\epsilon$, we get

$$\|\boldsymbol{h}\|_2 \leq \left(\frac{2+2(\sqrt{2}-1)\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}\right)\frac{\|\boldsymbol{x}_{oI^c}\|_1}{\sqrt{k}} + \left(\frac{4(1+\delta_{2k})^{1/2}}{1-(1+\sqrt{2})\delta_{2k}}\right)\epsilon, \tag{3.5.39}$$

where we note $\boldsymbol{x}_{oI^c} = \boldsymbol{x}_o - [\boldsymbol{x}_o]_k$. Therefore, as long as $1-(1+\sqrt{2})\delta_{2k} > 0$ or equivalently $\delta_{2k} < \sqrt{2}-1$, the conclusion of the theorem holds. □

Note that from the above proof, we know that the two constants in the Theorem can be chosen to be:

$$C = \frac{2 - 2(\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2})\delta_{2k}} \quad \text{and} \quad C' = \frac{4(1 + \delta_{2k})}{1 - (1 + \sqrt{2})\delta_{2k}}. \tag{3.5.40}$$

If δ_{2k} is very small, say approaching to zero, then C approaches to 2 and C' to 4. Those constants give the smallest possible bound for the error $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ based on this proof.

3.6 Phase Transitions in Sparse Recovery

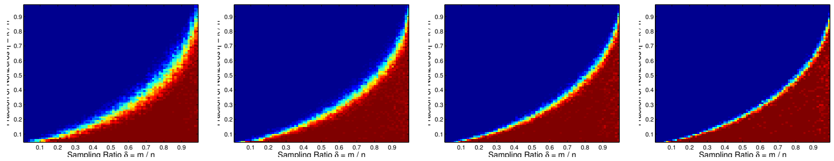
Above, we showed that sparse vectors $\boldsymbol{x}_o$ can be accurately estimated from linear observations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$. One of the surprises was that in the noise-free case ($\boldsymbol{z} = \boldsymbol{0}$), k-sparse vectors could be exactly recovered from just slightly more than k measurements – to be precise, $m \geq Ck\log(n/k)$ measurements, where C is a constant. The key technical tool for doing this was the restricted isometry property (RIP). The RIP and related properties enable simple proofs, with correct orders of growth (i.e., $m \sim k\log(n/k)$), but are not intended to give precise estimates of the constant C.

For applications, it can be important to know C. In sampling and reconstruction, this tells us precisely how many samples we need to acquire to accurately estimate a sparse signal; in error correction, this tells us precisely how many errors the system can tolerate.

Put another way, we would like to obtain precise relationships between the dimensionality n, the number of measurements m, and the number of nonzero entries k that we can recover. We would like these relationships to be as sharp and explicit as possible. To get some intuition for what to expect, we again resort to numerical simulation. We fix n, and consider different levels of sparsity k, and numbers of measurements m. For each pair (k, m), we generate a number of random ℓ^1 minimization problems, with noiseless Gaussian measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$, and ask *"For what fraction of these problems does ℓ^1 minimization correctly recover $\boldsymbol{x}_o$?"*

Figure 3.15 displays the result as a two dimensional image. Here, the horizontal axis is the sampling ratio $\delta = m/n$. This ranges from zero on the left (a very short, wide $\boldsymbol{A}$) to one on the right (a nearly square $\boldsymbol{A}$). The vertical axis is the fraction of nonzeros $\eta = k/n$. Again, this ranges from zero at the bottom (very sparse problems) to one at the top (denser problems). For each pair (η, δ), we generate 200 random problems. The intensity is the fraction of problems for which ℓ^1 minimization succeeds. The four graphs, from left to right, show the result for $n = 50, 100, 200, 400$.

This figure conveys several important pieces of information. First, as expected, when m is large and k is small (the lower right corner of each graph), ℓ^1 minimization always succeeds. Conversely, when m is small and



$n = 50$ $n = 100$ $n = 200$ $n = 400$

Figure 3.15. **Phase transition in sparse recovery with Gaussian matrices.** Each display plots the fraction of correct recoveries using ℓ^1 minimization, over a suite of randomly generated problems. The vertical axis represents the fraction of nonzero entries $\eta = k/n$ in the target vector $\boldsymbol{x}_o$ – the bottom corresponds to very sparse vectors, while the top corresponds to fully dense vectors. The horizontal axis represents the sampling ratio $\delta = m/n$ – the left corresponds to drastically under sampled problems ($m \ll n$), while the right corresponds to almost fully observed problems. For each (η, δ) pair, we generate 200 random problems, which we solve using CVX. We declare success if the recovered vector is accurate up to a relative error $\leq 10^{-6}$. Several salient features emerge: first, there is an easy regime (lower right corner) in which ℓ^1 minimization always succeeds. Second, there is a hard regime (upper left corner) in which ℓ^1 minimization always fails. Finally, as n increases, this transition between success and failure becomes increasingly sharp.

k is large (the upper left corner of each graph), ℓ^1 minimization always fails. Moreover, as n grows, the transition between success and failure becomes increasingly abrupt. Put another way, for high-dimensional problems, the behavior of ℓ^1 minimization is surprisingly predictable: it either almost always succeeds, or almost always fails. The line demarcating the sharp boundary between success and failure is known as a *phase transition*.

3.6.1 Phase Transitions: Main Conclusions

In this section, we state a result that precisely specifies the location of the phase transition. Namely, we will show that a sharp transition from failure to success occurs when the sampling ratio $\delta = m/n$ exceeds a certain function $\psi(\eta)$ of the sparsity ratio $\eta = k/n$. This result will be sharper than the ones we stated above using incoherence and RIP, in the sense that it identifies the precise number of measurements $m^\star = \psi(k/n)n$ required for success. To obtain such sharp results, we need to make two changes to our setting. First, we will make stronger assumptions on the matrix $\boldsymbol{A}$. Second, we will weaken the goal of our performance guarantee.

Random vs. deterministic $\boldsymbol{A}$*.*

Thus far, we have focused on deterministic properties of the matrix $\boldsymbol{A}$, such as (in)-coherence and the RIP. These properties do not depend on any random model for the matrix $\boldsymbol{A}$, although they are easiest to verify for random $\boldsymbol{A}$. Obtaining sharp estimates on the location of the phase transition requires more sophisticated probabilistic tools, which intrinsically require $\boldsymbol{A}$ to be a random matrix. We will sketch this theory under the assumption that $\boldsymbol{A}_{ij} \sim_{iid} \mathcal{N}(0, \frac{1}{m})$, i.e., $\boldsymbol{A}$ is a standard Gaussian random matrix. We will also briefly describe experiments and theoretical results which show that the results we will obtain for Gaussian $\boldsymbol{A}$ are "universal", in the sense that they precisely describe the behavior of ℓ^1 minimization for a fairly broad family of matrices $\boldsymbol{A}$. Nevertheless, all currently known theory which is sharp enough precisely characterize the phase transition requires $\boldsymbol{A}$ to be a random matrix.

Recovering a particular sparse $\boldsymbol{x}_o$ vs. recovering all sparse $\boldsymbol{x}_o$.

Incoherence and RIP allow one to prove "for all" results, which say that for a given matrix $\boldsymbol{A}$, ℓ^1 minimization recovers *every* sparse $\boldsymbol{x}_o$ from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$. The strongest and most general known results for phase transitions pertain to a slightly weaker statement: for a given, *fixed* $\boldsymbol{x}_o$, with high probability in the random matrix $\boldsymbol{A}$, ℓ^1 minimization recovers that particular $\boldsymbol{x}_o$ from the measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$.

Phase transition in ℓ^1 recovery.

Consider a sparse vector $\boldsymbol{x}_o \in \mathbb{R}^n$. We assume that $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a standard Gaussian random matrix, and ask whether ℓ^1 minimization recovers $\boldsymbol{x}_o$. We will show that the probability of correct recovery undergoes a sharp transition at

$$m^\star = \psi\left(\frac{k}{n}\right) n. \tag{3.6.1}$$

Here, $\psi : [0,1] \to [0,1]$ a function which takes as input the fraction $\eta = k/n$ of nonzeros, and describes the ratio $m^\star/n$ of number of measurements to the ambient dimension. The precise location ψ of the transition is given by the expression:

$$\psi(\eta) = \min_{t \geq 0} \left\{ \eta(1+t^2) + (1-\eta)\sqrt{\frac{2}{\pi}} \int_t^\infty (s-t)^2 \exp\left(-\frac{s^2}{2}\right) ds \right\}. \tag{3.6.2}$$

The function ψ is somewhat complicated; below, we will describe how it arises naturally from the geometry of ℓ^1 minimization. While there is no closed form solution for the minimization over t in this formula, it can be calculated numerically. Figure 3.16 displays this curve (red) superimposed over the empirical fraction of successes (grayscale) in our experiment. Clearly, there is a very good agreement between this theoretical prediction

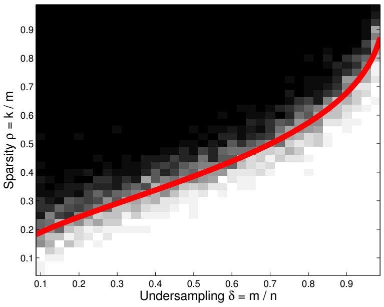


Figure 3.16. **Phase transitions: Agreement between theory and practice.** Theoretical phase transition predicted by (3.6.1) and (3.6.2), overlaid on fraction of successes in 200 experiments, for varying sparsities $\rho = k/m$ and aspect ratios $\delta = n/m$.

and our previous experiment: the empirical fraction of successes transitions rapidly from 0 to 1 as m/n exceeds $\psi(k/n)$.[21]

In fact, one can do slightly more: in addition to showing that $\psi(t)$ determines a point of transition between likely success and likely failure, we can give lower bounds on the probability of success (below the phase transition) and failure (above the phase transition) which quantify how sharp the transition is, for finite n. The following theorem makes all of this precise:

Theorem 3.6.1. *Let $\boldsymbol{x}_o \in \mathbb{R}^n$ be k-sparse, and suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o \in \mathbb{R}^{m \times n}$, with $\boldsymbol{A} \sim_{\text{iid}} \mathcal{N}(0, \frac{1}{m})$. Let $m^\star = \psi(k/n)n$, with ψ as in (3.6.2). Then*

$$\begin{aligned} \mathbb{P}\left[\ell^1 \textit{ recovers } \boldsymbol{x}_o\right] &\geq 1 - C\exp\left(-c\frac{(m-m^\star)^2}{n}\right), \qquad m > m^\star \\ \mathbb{P}\left[\ell^1 \textit{ does not recover } \boldsymbol{x}_o\right] &\geq 1 - c'\exp\left(-C'\frac{(m^\star-m)^2}{n}\right), \qquad m < m^\star, \end{aligned}$$

where C, c, c', C' are positive numerical constants.

This result implies that a sharp transition indeed occurs at $m^\star$ measurements: when $m/n > m^\star/n + C''/\sqrt{n}$, the probability of failure is bounded by a small constant (which can be made arbitrarily small by choosing C'' large). Conversely, when $m/n < m^\star/n - C''/\sqrt{n}$, the probability of success is bounded by a small constant. Hence, the transition region observed in Figure 3.15 has width $O(1/\sqrt{n})$ – in particular, it vanishes as $n \to \infty$.

A variety of mathematical tools have been brought to bear on the analysis of phase transitions in ℓ^1 minimization. Parts of Theorem 3.6.1 have been obtained using a several different approaches, by different sets of au-

[21] Figure 3.16 displays the same phase transition as in Figure 3.15 in a different parameterization, in which the vertical axis is $\rho = k/m$ and the horizontal axis is $\delta = m/n$.

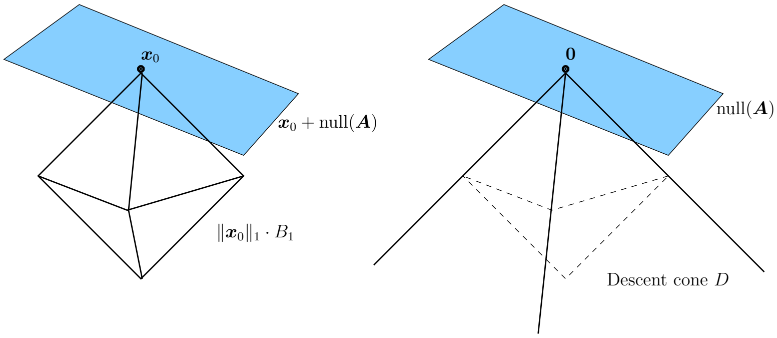


Figure 3.17. **Cones and the coefficient space geometry.** ℓ^1 minimization uniquely recovers x_o if and only if the intersection of the descent cone D with $\mathrm{null}(A)$ is $\{\mathbf{0}\}$.

thors. In the following two sections, we describe two such approaches, which correspond roughly to the two geometric pictures in Section 3.1, which describe the behavior of ℓ^1 minimization in terms of the space $\mathbb{R}^n$ of coefficient vectors x and the space $\mathbb{R}^m$ of observation vectors y.

3.6.2 Phase Transitions via Coefficient-Space Geometry

Suppose that $y = Ax_o$. Recall the geometric picture in Figure 3.17 (left), which we introduced in Section 3.1. There, we argued that x_o is the unique optimal solution to the ℓ^1 minimization problem if and only if the affine subspace

$$x_o + \mathrm{null}(A) \tag{3.6.3}$$

of feasible solutions x intersects the scaled ℓ^1 ball

$$\mathsf{B}_1 = \{x \mid \|x\|_1 \le \|x_o\|_1\} \tag{3.6.4}$$

only at x_o.

We can express the same geometry more cleanly in terms of the *descent cone*:

$$\mathsf{D} = \{v \mid \|x_o + tv\|_1 \le \|x_o\|_1 \text{ for some } t > 0\}\,. \tag{3.6.5}$$

This is the set of directions v for which a small (but nonzero) perturbation of x_o in the v direction does not increase the objective function $\|\cdot\|_1$. The descent cone D is visualized in Figure 3.17 (right).

Notice that the perturbation $x_o + tv$ is feasible for $t \neq 0$ if and only if $v \in \mathrm{null}(A)$. The feasible perturbations which do not increase the objective function reside in the intersection $\mathsf{D} \cap \mathrm{null}(A)$. Because D is a convex cone and $\mathrm{null}(A)$ is a subspace, D and $\mathrm{null}(A)$ always intersect at $\mathbf{0}$. It is not difficult to see that $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^1 problem if and only if $\mathbf{0}$ is the only point of intersection between $\mathrm{null}(\boldsymbol{A})$ and D:

Lemma 3.6.2. *Suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$. Then $\boldsymbol{x}_o$ is the unique optimal solution to the ℓ^1 minimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \end{array} \tag{3.6.6}$$

if and only if $\mathsf{D} \cap \mathrm{null}(\boldsymbol{A}) = \{\mathbf{0}\}$.

Proof. First, suppose that $\mathsf{D} \cap \mathrm{null}(\boldsymbol{A}) = \{\mathbf{0}\}$. Consider any alternative solution $\boldsymbol{x}'$. Then $\boldsymbol{x}' - \boldsymbol{x}_o \in \mathrm{null}(\boldsymbol{A}) \setminus \{\mathbf{0}\}$. Since $\mathsf{D} \cap \mathrm{null}(\boldsymbol{A}) = \{\mathbf{0}\}$, $\boldsymbol{x}' - \boldsymbol{x}_o \notin \mathsf{D}$, and so $\|\boldsymbol{x}'\|_1 > \|\boldsymbol{x}_o\|_1$, and $\boldsymbol{x}'$ is not an optimal solution. Since this holds for any feasible $\boldsymbol{x}'$, $\boldsymbol{x}_o$ is the unique optimal solution.

Conversely, suppose $\boldsymbol{x}_o$ is not the unique optimal solution. Then there exists $\boldsymbol{x}' \neq \boldsymbol{x}_o$ with $\|\boldsymbol{x}'\|_1 \leq \|\boldsymbol{x}_o\|_1$. Thus $\boldsymbol{x}' - \boldsymbol{x}_o \in \mathsf{D}$. By feasibility, $\boldsymbol{x}' - \boldsymbol{x}_o \in \mathrm{null}(\boldsymbol{A})$, and so $\mathsf{D} \cap \mathrm{null}(\boldsymbol{A}) \neq \{\mathbf{0}\}$. □

Hence, to study whether ℓ^1 minimization succeeds, we may equivalently study whether the subspace $\mathrm{null}(\boldsymbol{A})$ has nontrivial intersection with the cone D. Because $\boldsymbol{A}$ is a random matrix, $\mathrm{null}(\boldsymbol{A})$ is a random subspace, of dimension $n-m$. If $\boldsymbol{A}$ is Gaussian, then $\mathrm{null}(\boldsymbol{A})$ follows the uniform distribution on the set of subspaces $S \subset \mathbb{R}^n$ of dimension $n-m$.[22] Clearly, the probability that the random subspace $\mathrm{null}(\boldsymbol{A})$ intersects the descent cone D depends on properties of D. Intuitively, we would expect intersections to be more likely if D is "big" in some sense.

Measuring the "size" of a convex cone.

We describe a way of measuring the "size" of a convex cone, which provides very precise control on the probability that a randomly chosen subspace has nontrivial intersection with it. Recall that for a closed convex cone $\mathsf{C} \subseteq \mathbb{R}^n$ and a vector $\boldsymbol{z}$, there is a unique nearest vector to $\boldsymbol{z}$ in C, denoted $\mathcal{P}_{\mathsf{C}}\boldsymbol{z}$:

$$\mathcal{P}_{\mathsf{C}}[\boldsymbol{z}] \doteq \arg\min_{\boldsymbol{x} \in \mathsf{C}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 . \tag{3.6.7}$$

Figure 3.18 shows the projections $\mathcal{P}_{\mathsf{C}_i}[\boldsymbol{z}]$ of a vector $\boldsymbol{z}$ onto two convex cones C_1 and C_2. Notice that it is always true that

$$\|\mathcal{P}_{\mathsf{C}}[\boldsymbol{z}]\|_2 \leq \|\boldsymbol{z}\|_2 . \tag{3.6.8}$$

Moreover, in Figure 3.18, the norm of the projection is larger for the wider C_i. Thus, we could take $\|\mathcal{P}_{\mathsf{C}}[\boldsymbol{z}]\|_2^2$ as an indication of the "size" of C. To

[22] To be more precise, $\mathrm{null}(\boldsymbol{A})$ is distributed according to the Haar measure on the Grassmannian $\mathsf{G}_{n,m-n}$.

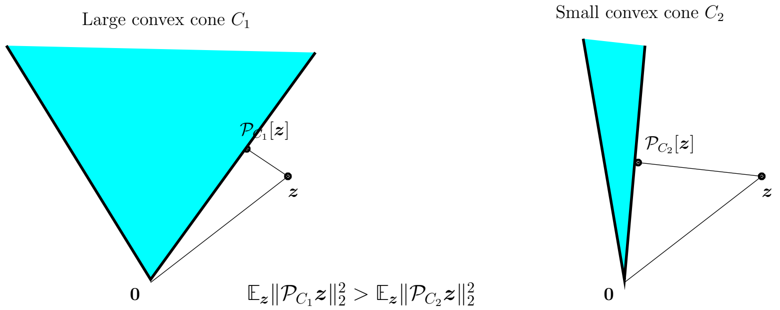

Large convex cone C_1
$\mathcal{P}_{C_1}[\boldsymbol{z}]$
$\boldsymbol{z}$
$\mathbf{0}$
$\mathbb{E}_{\boldsymbol{z}}\|\mathcal{P}_{C_1}\boldsymbol{z}\|_2^2 > \mathbb{E}_{\boldsymbol{z}}\|\mathcal{P}_{C_2}\boldsymbol{z}\|_2^2$
Small convex cone C_2
$\mathcal{P}_{C_2}[\boldsymbol{z}]$
$\boldsymbol{z}$
$\mathbf{0}$

Figure 3.18. **Projections onto a closed convex cone.** For a closed convex cone C, $\mathcal{P}_{\mathsf{C}}[\boldsymbol{z}]$ is the nearest point to $\boldsymbol{z}$ in C. Notice that in this case, the projection of $\boldsymbol{z}$ onto the larger cone C_1 has greater norm than the projection of $\boldsymbol{z}$ onto the smaller cone C_2: $\|\mathcal{P}_{\mathsf{C}_1}\boldsymbol{z}\|_2^2 > \|\mathcal{P}_{\mathsf{C}_2}\boldsymbol{z}\|_2^2$. We can measure the "size" of a convex cone C by averaging $\|\mathcal{P}_{\mathsf{C}}\boldsymbol{z}\|_2^2$ over all directions $\boldsymbol{z}$; this average is known as the *statistical dimension* of the cone, denoted $\delta(\mathsf{C})$.

obtain a quantity that does not depend on our particular choice of $\boldsymbol{z}$, we average over all possible vectors $\boldsymbol{z}$ and define

$$\delta(\mathsf{C}) \doteq \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,1)}\left[\|\mathcal{P}_{\mathsf{C}}\boldsymbol{g}\|_2^2\right]. \tag{3.6.9}$$

This quantity is sometimes referred to as the *statistical dimension* of the cone C.

Why is this a dimension?

To understand why it is reasonable to refer to $\delta(\mathsf{C})$ as a dimension, let us first consider the special case when the cone C is a linear subspace. Suppose that $\mathsf{C} = \mathsf{S}$ for some linear subspace S of $\mathbb{R}^n$, of dimension d. Let $U \in \mathbb{R}^{n\times d}$ denote a matrix whose columns form an orthonormal basis for S, and notice that for any $\boldsymbol{z} \in \mathbb{R}^n$, $\|\mathcal{P}_{\mathsf{S}}\boldsymbol{z}\|_2^2 = \|\boldsymbol{U}^*\boldsymbol{z}\|_2^2$. If $\boldsymbol{g} \sim \mathcal{N}(0,1)$, the entries of the d-dimensional random vector $\boldsymbol{U}^*\boldsymbol{g}$ are independent $\mathcal{N}(0,1)$ random variables, and so

$$\delta(\mathsf{C}) = \mathbb{E}_{\boldsymbol{g}}\left[\|\mathcal{P}_{\mathsf{C}}\boldsymbol{g}\|_2^2\right] = \mathbb{E}_{\boldsymbol{g}}\left[\|\boldsymbol{U}^*\boldsymbol{g}\|_2^2\right] = d. \tag{3.6.10}$$

Hence, *for a linear subspace, the statistical dimension* $\delta(\mathsf{S})$ *is exactly the same as the usual dimension:* $\delta(\mathsf{S}) = d$.

For more general convex cones C, the statistical dimension $\delta(\mathsf{C})$ does not coincide with any familiar geometric or algebraic notion of dimension. However, we will see that for the purposes of studying intersections with random subspaces, a cone C with $\delta(\mathsf{C}) = d$ "acts like" a subspace of dimension d. In making this precise, we will obtain the key probabilistic result needed to prove Theorem 3.6.1.

Probability that a random subspace intersects a convex cone.

Recall that ℓ^1 minimization uniquely recovers $\boldsymbol{x}_o$ if and only if the random subspace $\mathsf{S} = \mathrm{null}(\boldsymbol{A})$ has only trivial intersection with the descent cone D: $\mathsf{S} \cap \mathsf{D} = \{\mathbf{0}\}$. Thus, we need to determine how likely it is that a random subspace intersects a given convex cone.

We will show that the statistical dimension provides very precise bounds on the probability that this event occurs. To do so, we will state a general result controlling the probability that a random subspace S of dimension d intersects a convex cone C of statistical dimension $\delta(\mathsf{C})$. To get intuition for what to expect for general cones C, we can again start with simplest case: when C is a linear subspace S' of dimension d'. In this situation $\delta(\mathsf{C}) = d'$.

Does the randomly chosen subspace S intersect S'? If the sum of the dimensions $d + d'$ is greater than the ambient dimension n, then S and S' necessarily have a nontrivial intersection. Conversely, if $d + d' \le n$, the probability that S intersects S' is zero:

Proposition 3.6.3. *Let S' be any linear subspace of $\mathbb{R}^n$, and let S be a uniform random subspace. Then*

$$\mathbb{P}\left[\mathsf{S} \cap \mathsf{S}' = \{\mathbf{0}\}\right] = 0, \quad \delta(\mathsf{S}) + \delta(\mathsf{S}') > n; \tag{3.6.11}$$

$$\mathbb{P}\left[\mathsf{S} \cap \mathsf{S}' = \{\mathbf{0}\}\right] = 1, \quad \delta(\mathsf{S}) + \delta(\mathsf{S}') \le n. \tag{3.6.12}$$

Thus, for linear subspaces, the sum of the statistical dimensions precisely controls whether S and S' have nontrivial intersection: once $\delta(\mathsf{S}) + \delta(\mathsf{S}') > n$, the probability of nontrivial intersection goes from one to zero. For general convex cones, there is a similar phenomenon: if S is a random subspace of $\mathbb{R}^n$, and C a closed convex cone, then

$$\begin{aligned} \delta(\mathsf{S}) + \delta(\mathsf{C}) \gg n \quad &\Longrightarrow \quad \mathsf{S} \cap \mathsf{C} \neq \{\mathbf{0}\} \text{ with high probability;} \\ \delta(\mathsf{S}) + \delta(\mathsf{C}) \ll n \quad &\Longrightarrow \quad \mathsf{S} \cap \mathsf{C} = \{\mathbf{0}\} \text{ with high probability.} \end{aligned}$$

The following theorem makes this precise:

Theorem 3.6.4. *Let C denote any closed convex cone in $\mathbb{R}^n$, and let S be a uniformly distributed random subspace of dimension $\delta(\mathsf{S})$. Then*

$$\mathbb{P}\left[\mathsf{S} \cap \mathsf{C} = \{\mathbf{0}\}\right] \le C \exp\left(-c\frac{(n - \delta(\mathsf{S}) - \delta(\mathsf{C}))^2}{n}\right), \quad \delta(\mathsf{S}) + \delta(\mathsf{C}) \ge n;$$

$$\mathbb{P}\left[\mathsf{S} \cap \mathsf{C} = \{\mathbf{0}\}\right] \ge 1 - C \exp\left(-c\frac{(\delta(\mathsf{S}) + \delta(\mathsf{C}) - n)^2}{n}\right), \quad \delta(\mathsf{S}) + \delta(\mathsf{C}) \le n.$$

This theorem is a special case of a somewhat more general result controlling the probability that two randomly oriented convex cones intersect. The proof relies on technical results in spherical integral geometry. We refer the interested reader to Theorem 1 of [**?**], its proof, and references therein.

For now, we will show how to use Theorem 3.6.4 to prove our main claims about the phase transition in ℓ^1 minimization. In our situation, the cone

C of interest is the descent cone D of the ℓ^1 norm at $\boldsymbol{x}_o$. We wish to know whether $\mathsf{S} = \mathrm{null}(\boldsymbol{A})$ has nontrivial intersection with C. The dimension of S is $n - m$, and so the above heuristics become

$$\begin{array}{lll} \textbf{FAILURE:} & \delta(\mathsf{D}) \gg m \implies & \mathrm{null}(\boldsymbol{A}) \cap \mathsf{D} \neq \{\mathbf{0}\} \text{ with high probability;} \\ \textbf{SUCCESS:} & \delta(\mathsf{D}) \ll m \implies & \mathrm{null}(\boldsymbol{A}) \cap \mathsf{D} = \{\mathbf{0}\} \text{ with high probability.} \end{array}$$

In the first case, ℓ^1 fails to recover $\boldsymbol{x}_o$; in the second case it succeeds. Using Theorem 3.6.4 to make this precise, we obtain:

Corollary 3.6.5. *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a Gaussian matrix, and suppose that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$. Let D denote the descent cone of the ℓ^1 norm at $\boldsymbol{x}_o$. Then*

$$\mathbb{P}\left[\,\ell^1 \textit{ uniquely recovers } \boldsymbol{x}_o\,\right] \;\leq\; C\exp\left(-c\frac{(\delta(\mathsf{D}) - m)^2}{n}\right), \quad m \leq \delta(\mathsf{D});$$

$$\mathbb{P}\left[\,\ell^1 \textit{ uniquely recovers } \boldsymbol{x}_o\,\right] \;\geq\; 1 - C\exp\left(-c\frac{(m - \delta(\mathsf{D}))^2}{n}\right), \quad m \geq \delta(\mathsf{D}).$$

Thus, when m is substantially smaller than $\delta(\mathsf{D})$, recovery fails with high probability; when m is substantially larger than $\delta(\mathsf{D})$ recovery succeeds with high probability.

The statistical dimension of the descent cone.

To prove Theorem 3.6.1, we need to show that the statistical dimension $\delta(\mathsf{D})$ of the descent cone D is very close to $n\psi(k/n)$. We state this result as a lemma below:

Lemma 3.6.6. *Let D be the descent cone of the ℓ^1 norm at any $\boldsymbol{x}_o \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x}_o\|_0 = k$. Then*

$$n\psi\left(\frac{k}{n}\right) - 4\sqrt{n/k} \leq \delta(\mathsf{D}) \leq n\psi\left(\frac{k}{n}\right). \tag{3.6.13}$$

Proof. For this, we will need two basic facts about projections onto convex cones. The first is the generalized pythagorean formula, which implies that for a closed convex cone D with polar cone

$$\mathsf{D}^\circ = \{\boldsymbol{v} \mid \langle \boldsymbol{v}, \boldsymbol{x}\rangle \leq 0\; \forall\, \boldsymbol{x} \in \mathsf{D}\}, \tag{3.6.14}$$

for any $\boldsymbol{z} \in \mathbb{R}^n$,

$$\begin{aligned} \|\mathcal{P}_{\mathsf{D}}\boldsymbol{z}\|_2^2 &= \|\boldsymbol{z} - \mathcal{P}_{\mathsf{D}^\circ}\boldsymbol{z}\|_2^2 & (3.6.15) \\ &= \mathrm{dist}^2(\boldsymbol{z}, \mathsf{D}^\circ). & (3.6.16) \end{aligned}$$

This allows us to replace the norm of the projection of $\boldsymbol{z}$ onto D with the distance of $\boldsymbol{z}$ to the polar cone D°. The second fact is that the polar of the descent cone is the conic hull of the subdifferential

$$\mathsf{S} \doteq \partial\left\|\cdot\right\|_1(\boldsymbol{x}_o) = \{\boldsymbol{v} \mid \boldsymbol{v}_I = \mathrm{sign}(\boldsymbol{x}_I),\ \|\boldsymbol{v}_{I^c}\|_\infty \leq 1\}. \tag{3.6.17}$$

Namely,

$$\begin{aligned} \mathsf{D}^\circ &= \operatorname{cone}(\mathsf{S}) \\ &= \bigcup_{t\geq 0} t\,\mathsf{S} \\ &= \left\{ t\boldsymbol{v} \mid t \geq 0,\ \boldsymbol{v}_I = \boldsymbol{\sigma}_I,\ \|\boldsymbol{v}_{I^c}\|_\infty \leq 1 \right\}. \end{aligned} \tag{3.6.18}$$

For any vector $\boldsymbol{z}$, the nearest vector $\hat{\boldsymbol{z}} \in tS$ satisfies

$$\hat{\boldsymbol{z}}_i = \begin{cases} t\,\operatorname{sign}(\boldsymbol{z}_i) & i \in I, \\ \boldsymbol{z}_i & i \in I^c,\ |\boldsymbol{z}_i| \leq t, \\ t\,\operatorname{sign}(\boldsymbol{z}_i) & i \in I^c,\ |\boldsymbol{z}_i| > t, \end{cases} \tag{3.6.19}$$

and the distance is given by

$$\operatorname{dist}^2(\boldsymbol{z}, t\mathsf{S}) = \|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2^2 \tag{3.6.20}$$

$$= \|\boldsymbol{z}_I - t\boldsymbol{\sigma}_I\|_2^2 + \sum_{j\in I^c} \max\left\{|z_j| - t, 0\right\}^2. \tag{3.6.21}$$

Hence, for any vector $\boldsymbol{z}$,

$$\operatorname{dist}^2(\boldsymbol{z}, \mathsf{D}^\circ) = \min_{t\geq 0} \operatorname{dist}^2(\boldsymbol{z}, t\mathsf{S}) \tag{3.6.22}$$

$$= \min_{t\geq 0} \left\{ \|\boldsymbol{z}_I - t\boldsymbol{\sigma}_I\|_2^2 + \sum_{j\in I^c} \max\left\{|z_j| - t, 0\right\}^2 \right\}. \tag{3.6.23}$$

Using these facts, we calculate

$$\begin{aligned} \delta(\mathsf{D}) &= \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,1)} \left[\|\mathcal{P}_{\mathsf{D}}\boldsymbol{g}\|_2^2 \right] \\ &= \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,1)} \left[\operatorname{dist}^2(\boldsymbol{g}, \mathsf{D}^\circ) \right] \\ &= \mathbb{E}_{\boldsymbol{g}} \left[\min_{t\geq 0} \operatorname{dist}^2\left(\boldsymbol{g}, t\,\mathsf{S}\right) \right] \\ &\leq \min_{t\geq 0}\ \mathbb{E}_{\boldsymbol{g}} \left[\operatorname{dist}^2\left(\boldsymbol{g}, t\,\mathsf{S}\right) \right] \\ &= \min_{t\geq 0}\ \mathbb{E}_{\boldsymbol{g}} \left[\|\boldsymbol{g}_I - t\boldsymbol{\sigma}_I\|_2^2 + \sum_{j\in I^c} \max\left\{|g_j| - t, 0\right\}^2 \right] \\ &= \min_{t\geq 0} \left\{ |I|(1+t^2) + 2|I^c| \int_{s=t}^{\infty} (s-t)^2 \varphi(s) ds \right\} \\ &= n\,\psi(k/n), \end{aligned} \tag{3.6.24}$$

where $\varphi(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2}$ is the Gaussian density. Thus, we have established $n\psi(k/n)$ as an upper bound on the statistical dimension; and hence $m^\star = n\psi(k/n)$ as a lower bound on the phase transition.

To finish, we show that this upper bound on $\delta(\mathsf{D})$ is tight, by establishing a (nearly) matching lower bound. Let $\hat{t}$ minimize $\mathbb{E}_{\boldsymbol{g}}\left[\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\right]$. Then

$$\begin{aligned} 0 &= \frac{d}{dt}\mathbb{E}_{\boldsymbol{g}}\left[\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\right]\Big|_{t=\hat{t}} \\ &= \mathbb{E}_{\boldsymbol{g}}\left[\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right]. \end{aligned} \tag{3.6.25}$$

Let $t_{\boldsymbol{g}}$ minimize $\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})$ with respect to t. By convexity of this function in t,

$$\operatorname{dist}^2(\boldsymbol{g}, t_{\boldsymbol{g}}\mathsf{S}) \quad \geq \quad \operatorname{dist}^2(\boldsymbol{g}, \hat{t}\mathsf{S}) + \left(t_{\boldsymbol{g}} - \hat{t}\right)\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\big|_{t=\hat{t}}. \tag{3.6.26}$$

Notice that by (3.6.25),

$$0 = \hat{t}\,\mathbb{E}_{\boldsymbol{g}}\left[\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right] = \mathbb{E}\left[t_{\boldsymbol{g}}\right]\mathbb{E}_{\boldsymbol{g}}\left[\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right], \tag{3.6.27}$$

and so

$$\begin{aligned} \mathbb{E}_{\boldsymbol{g}}\left[\min_t \operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\right] &= \mathbb{E}_{\boldsymbol{g}}\left[\operatorname{dist}^2(\boldsymbol{g}, t_{\boldsymbol{g}}\mathsf{S})\right] \\ &\geq \mathbb{E}_{\boldsymbol{g}}\left[\operatorname{dist}^2(\boldsymbol{g}, \hat{t}\mathsf{S})\right] + \mathbb{E}_{\boldsymbol{g}}\left[\left(t_{\boldsymbol{g}} - \mathbb{E}_{\boldsymbol{g}}\left[t_{\boldsymbol{g}}\right]\right)\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right], \\ &\geq \mathbb{E}_{\boldsymbol{g}}\left[\operatorname{dist}^2(\boldsymbol{g}, \hat{t}\mathsf{S})\right] - \operatorname{var}(t_{\boldsymbol{g}})^{1/2}\operatorname{var}\left(\frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right)^{1/2}. \end{aligned} \tag{3.6.28}$$

In the last line we have used the Cauchy-Schwarz inequality for random variables.

To conclude, we bound the variance of the two terms. For $t_{\boldsymbol{g}}$, let $\boldsymbol{v}_{\boldsymbol{g}} \in \mathsf{S}$ be such that $t_{\boldsymbol{g}}\boldsymbol{v}_{\boldsymbol{g}}$ is the nearest element to $\boldsymbol{g}$ in D°. Notice that

$$\begin{aligned} \|\boldsymbol{g} - \boldsymbol{g}'\|_2 &\geq \|t_{\boldsymbol{g}}\boldsymbol{v}_{\boldsymbol{g}} - t_{\boldsymbol{g}'}\boldsymbol{v}_{\boldsymbol{g}'}\|_2 && (3.6.29) \\ &\geq \|t_{\boldsymbol{g}}\boldsymbol{\sigma}_I - t_{\boldsymbol{g}'}\boldsymbol{\sigma}_I\|_2 && (3.6.30) \\ &= |t_{\boldsymbol{g}} - t_{\boldsymbol{g}'}|\sqrt{k}, && (3.6.31) \end{aligned}$$

whence $t_{\boldsymbol{g}}$ is a $1/\sqrt{k}$-Lipschitz function of $\boldsymbol{g}$. By the Gaussian Poincare inequality,[23] its variance is bounded as $\operatorname{var}(t_{\boldsymbol{g}}) \leq 1/k$.

Meanwhile, by Danskin's theorem,

$$\begin{aligned} \frac{d}{dt}\operatorname{dist}^2(\boldsymbol{g}, t\mathsf{S}) &= \frac{d}{dt}\|\boldsymbol{g} - t\boldsymbol{v}_g\|_2^2 && (3.6.32) \\ &= 2\boldsymbol{v}_{\boldsymbol{g}}^*\left(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}\right). && (3.6.33) \end{aligned}$$

[23] which states that if f is an L-Lipschitz function and $\boldsymbol{g}$ a Gaussian vector, then $\operatorname{var}(f(\boldsymbol{g})) \leq L^2$.

Note that because $t\boldsymbol{v}_{\boldsymbol{g}'}$ is the projection of $\boldsymbol{g}'$ onto the convex set S, for any other $\boldsymbol{v} \in \mathsf{S}$,

$$(t\boldsymbol{v}_{\boldsymbol{g}} - t\boldsymbol{v})^*(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}) \le 0, \tag{3.6.34}$$

whence

$$\boldsymbol{v}_{\boldsymbol{g}}^*(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}) \le \boldsymbol{v}_{\boldsymbol{g}'}^*(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}), \tag{3.6.35}$$

and

$$\begin{aligned}
\frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S}) - \frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}', t\mathsf{S}) \quad &= \quad 2\boldsymbol{v}_{\boldsymbol{g}}^*(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}) - 2\boldsymbol{v}_{\boldsymbol{g}'}^*(t\boldsymbol{v}_{\boldsymbol{g}'} - \boldsymbol{g}') \\
&\le \quad 2\boldsymbol{v}_{\boldsymbol{g}'}^*(t\boldsymbol{v}_{\boldsymbol{g}} - \boldsymbol{g}) - 2\boldsymbol{v}_{\boldsymbol{g}'}^*(t\boldsymbol{v}_{\boldsymbol{g}'} - \boldsymbol{g}') \\
&\le \quad 2\left\|\boldsymbol{v}_{\boldsymbol{g}'}\right\|_2 \left(\left\|t\boldsymbol{v}_{\boldsymbol{g}} - t\boldsymbol{v}_{\boldsymbol{g}'}\right\|_2 + \left\|\boldsymbol{g} - \boldsymbol{g}'\right\|_2\right) \\
&\le \quad 4\left\|\boldsymbol{v}_{\boldsymbol{g}'}\right\|_2 \left\|\boldsymbol{g} - \boldsymbol{g}'\right\|_2 \\
&\le \quad 4\sqrt{n}\left\|\boldsymbol{g} - \boldsymbol{g}'\right\|_2 .
\end{aligned} \tag{3.6.36}$$

By the same reasoning,

$$\begin{aligned}
\frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S}) - \frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}', t\mathsf{S}) \quad &\ge \quad 2\boldsymbol{v}_{\boldsymbol{g}}^*\left(t\boldsymbol{v}_g - \boldsymbol{g} - t\boldsymbol{v}_{\boldsymbol{g}'} + \boldsymbol{g}'\right) \\
&\ge \quad -4\sqrt{n}\left\|\boldsymbol{g} - \boldsymbol{g}'\right\|_2 ,
\end{aligned} \tag{3.6.37}$$

whence

$$\left|\frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S}) - \frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}', t\mathsf{S})\right| \;\le\; 4\sqrt{n}\left\|\boldsymbol{g} - \boldsymbol{g}'\right\|_2 , \tag{3.6.38}$$

and $\frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S})$ is $4\sqrt{n}$-Lipschitz. By the Gaussian Poincare inequality,

$$\mathrm{var}\left(\frac{d}{dt}\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S})\Big|_{t=\hat{t}}\right) \;\le\; 4\sqrt{n}, \tag{3.6.39}$$

and so

$$\mathbb{E}_{\boldsymbol{g}}\left[\min_t \mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S})\right] \;\ge\; \min_t \mathbb{E}_{\boldsymbol{g}}\left[\mathrm{dist}^2(\boldsymbol{g}, t\mathsf{S})\right] - 4\sqrt{n/k}. \tag{3.6.40}$$

Thus,

$$n\psi(k/n) - 4\sqrt{n/k} \;\le\; \delta(\mathsf{D}) \;\le\; n\psi(k/n). \tag{3.6.41}$$

Combining this bound with the above results proves that the phase transition occurs within $O(\sqrt{n})$ of $m^\star = n\psi(k/n)$. □

3.6.3 Phase Transitions via Observation-Space Geometry

Historically, the first sharp estimates of the location of the phase transition were derived using the "observation space" geometric picture of ℓ^1 minimization, which we reproduce in Figure 3.4. In this picture, ℓ^1 minimization

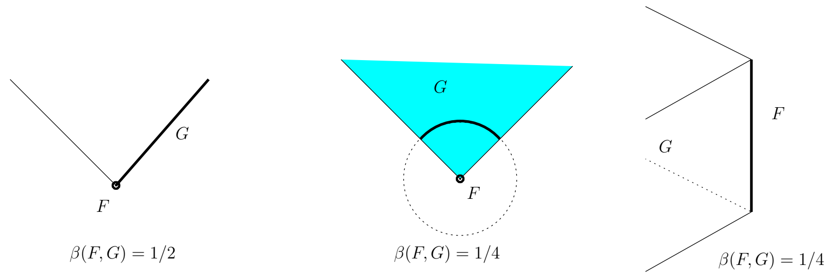


Figure 3.19. **Internal angles of convex polytopes.** The internal angle $\beta(\mathsf{F},\mathsf{G})$ of a face $\mathsf{F} \subseteq \mathsf{G}$ with respect to another face G containing it is the fraction of the linear span of $\mathsf{G} - \boldsymbol{x}$ occupied by $\mathsf{G} - \boldsymbol{x}$, where $\boldsymbol{x}$ is any point in the relative interior of F.

is visualized through the relationship between two convex polytopes, the unit ℓ^1 ball

$$\mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \le 1\} \tag{3.6.42}$$

and its projection into $\mathbb{R}^m$,

$$\mathsf{P} = \boldsymbol{A}\mathsf{B}_1 = \{\boldsymbol{A}\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \le 1\}\,. \tag{3.6.43}$$

Namely, ℓ^1 minimization uniquely recovers any $\boldsymbol{x}$ with support $\boldsymbol{I}$ and signs $\boldsymbol{\sigma}$ if and only if

$$\mathsf{F} = \mathrm{conv}(\{\sigma_i \boldsymbol{a}_i \mid i \in I\}) \tag{3.6.44}$$

forms a face of the polytope P. Conversely, if F intersects the interior of P, then ℓ^1 minimization does not recover $\boldsymbol{x}_o$ with support I and signs $\boldsymbol{\sigma}$.

The first results bounding the phase transition derived from remarkable results in stochastic geometry, which give exact formulas for the expected number of k-dimensional faces of a randomly projected polytope $\mathsf{P} = \boldsymbol{A}\mathsf{Q}$. This expectation depends two notions of the "size" of the polytope Q: the *internal angle* and *external angle.*

Definition 3.6.7 (Internal angle)**.** *The internal angle* $\beta(\mathsf{F},\mathsf{G})$ *of a face* F *of a polytope* G *is the fraction of* $\mathrm{span}(\mathsf{G} - \boldsymbol{x})$ *occupied by* $\mathsf{G} - \boldsymbol{x}$*, where* $\mathrm{span}(\cdot)$ *denotes the linear span, and* $\boldsymbol{x}$ *is any point in* $\mathrm{relint}(\mathsf{F})$*.*

The internal angle is visualized for several examples in Figure 3.19. Informally speaking, the internal angle measures the fraction of the space cut out by G, when viewed from F. There is a complementary notion of angle, called the external angle, which captures the fraction of the space cut out by the *normal cone* to G at a point in the relative interior of F:

Definition 3.6.8 (External angle)**.** *The external angle* $\beta(\mathsf{F},\mathsf{G})$ *of a face* $\mathsf{F} \subseteq \mathsf{G}$ *is the fraction of* $\mathrm{span}(\mathsf{G}-\boldsymbol{x})$ *occupied by the normal cone* $\mathsf{N}(\mathsf{F},\mathsf{G}) =$

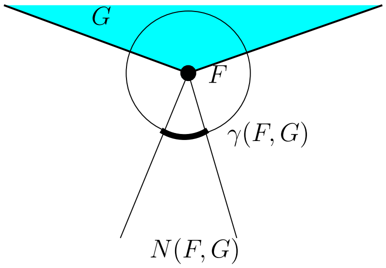


Figure 3.20. **External angles of convex polytopes.** The external angle $\gamma(\mathsf{F},\mathsf{G})$ of a face $\mathsf{F} \subseteq \mathsf{G}$ with respect to another face G containing it is the fraction of the linear span of $\mathsf{G} - \boldsymbol{x}$ occupied by the normal cone $\mathsf{N}(\mathsf{F},\mathsf{G})$.

$\{\boldsymbol{v} \in \operatorname{span}(\mathsf{G} - \boldsymbol{x}) \mid \langle v - \boldsymbol{x}, \boldsymbol{x}' - \boldsymbol{x}\rangle \le 0 \;\forall\, \boldsymbol{x}' \in \mathsf{G}\}$, *where* $\boldsymbol{x}$ *is any point in* $\operatorname{relint}(\mathsf{F})$.

Figure 3.20 visualizes the external angle. There is an exquisite characterization of the expected number of k-dimensional faces of a random projection of a convex polytope P, in terms of its internal and external angles. Let $f_k(\mathsf{P})$ denote the number of k-dimensional faces of a polytope P, and let F_k denote the collection of such faces. Then for an $m \times n$ Gaussian matrix $\boldsymbol{A}$,

$$\mathbb{E}_{\boldsymbol{A}}[f_k(\boldsymbol{A}\mathsf{P})] = f_k(\mathsf{P}) - 2 \underbrace{\sum_{\ell=m+1,m+3,\dots} \sum_{\mathsf{F}\in\mathsf{F}_k(\mathsf{P})} \sum_{\mathsf{G}\in\mathsf{F}_\ell(\mathsf{P})} \beta(\mathsf{F},\mathsf{G})\gamma(\mathsf{G},\mathsf{P})}_{\Delta=\text{Expected number of faces lost}}.$$

This formula arises out of a line of work in discrete geometry, which aims at understanding the behavior of "typical" point clouds, and studying the simplex method for linear programming for "typical" inputs. One remarkable aspect is that it gives the *exact* value of the expected face count. The connection to ℓ^1 minimization is that ℓ^1 successfully recovers every $k+1$-sparse vector $\boldsymbol{x}_o$ from measurements $\boldsymbol{A}\boldsymbol{x}_o$ if and only if $f_k(\boldsymbol{A}\mathsf{P}) = f_k(\mathsf{P})$. This can be observed from the observation-space geometry described above. This event can be studied through the quantity Δ – the expected number of faces lost. Whenever $\Delta < 1$, there exists an $\boldsymbol{A}$ such that $f_k(\boldsymbol{A}\mathsf{P}) = f_k(\mathsf{P})$; when Δ is substantially smaller than one, we can use the Markov inequality to argue that the probability that any face is lost in the projection is small.

3.6.4 Phase Transitions in Support Recovery

Thus far, we have focused on the problem of *estimating* a sparse vector $\boldsymbol{x}_o$. We showed that from noisy observations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, convex optimization produces a vector $\hat{\boldsymbol{x}}$ such that $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2$ is small. For many engineering applications, where $\boldsymbol{x}_o$ is represents a signal to be sensed or an error to be corrected, this is exactly what we need. However, in some applications, the goal is not so much to estimate $\boldsymbol{x}_o$ as to determine which of the entries of $\boldsymbol{x}_o$ are nonzero. A good example, which we will revisit in later chapters, is in spectrum sensing for wireless communications. Here, the entries of $\boldsymbol{x}_o$ represent frequency bands which might be available for transmission, or which might be occupied. The goal is know which frequency bands are available, so that we can avoid interfering with other users. In this setting, it is much more important to know which entries of $\boldsymbol{x}_o$ are nonzero, than to estimate the particular values.

Support Recovery: Desiderata.

In this section, we consider the problem of estimating the signed support

$$\boldsymbol{\sigma}_o = \operatorname{sign}(\boldsymbol{x}_o), \tag{3.6.45}$$

from noisy observations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}. \tag{3.6.46}$$

We will derive theory under the assumptions that the noise $\boldsymbol{z}$ is iid $\mathcal{N}(0, \frac{\sigma^2}{m})$. Let $\hat{\boldsymbol{x}}$ solve the Lasso problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \tfrac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1 . \tag{3.6.47}$$

We can distinguish between two conclusions:

- **Partial support recovery**: $\operatorname{supp}(\hat{\boldsymbol{x}}) \subseteq \operatorname{supp}(\boldsymbol{x}_o)$. Our estimator exhibits no "false positives": every element of the estimated support is an element of the true support.
- **Signed support recovery**: $\operatorname{sign}(\hat{\boldsymbol{x}}) = \boldsymbol{\sigma}_o$. Our estimator correctly determines the nonzero entries of $\boldsymbol{x}_o$ and their signs.

Signed support recovery is clearly more desirable than partial support recovery. Signed support recovery requires stronger assumptions of the signal $\boldsymbol{x}_o$ than partial support recovery – if the nonzero entries of $\boldsymbol{x}_o$ are too small relative to the noise level σ, no method of any kind will be able to reliably determine the support.

In contrast, partial support recovery can be studied without additional assumptions on the signal $\boldsymbol{x}_o$. We will assume that $\boldsymbol{A} \sim_{\text{iid}} \mathcal{N}(0, \frac{1}{m})$. We will first derive a sharp phase transition for partial support recovery, at

$$m_\star = 2k \log(n - k) \tag{3.6.48}$$

measurements. The main result of this section will show that when m significantly exceeds this threshold, partial support recovery obtains with high probability. Moreover, through further analysis, we will show that when m significantly exceeds $m_\star$, *and* all of the nonzero entries of $\boldsymbol{x}_o$ are significantly larger than λ, *signed support recovery* also obtains with high probability. Conversely, if m is significantly smaller than $m_\star$, the probability of signed support recovery is vanishingly small. Thus, $m_\star$ indeed gives a sharp threshold for support recovery. Notice that (3.6.48) grows roughly as $k \log n$, rather than $k \log(n/k)$. So, if m, n, k, grow in fixed ratios, support recovery is unlikely. In this sense, support recovery is a more challenging problem than estimation.

The following theorem makes the above discussion precise:

Theorem 3.6.9 (Phase transition in partial support recovery)**.** *Suppose that $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with entries iid $\mathcal{N}(0, \frac{1}{m})$ random variables, and let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, with $\boldsymbol{x}_o$ a k-sparse vector and $\boldsymbol{z} \sim_{\mathrm{iid}} \mathcal{N}\left(0, \frac{\sigma^2}{m}\right)$. If*

$$m \geq \left(1 + \frac{\sigma^2}{\lambda^2 k} + \varepsilon\right) 2k \log(n-k), \tag{3.6.49}$$

then with probability at least $1 - Cn^{-\varepsilon}$, any solution $\hat{\boldsymbol{x}}$ to the Lasso problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \tfrac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \tag{3.6.50}$$

satisfies $\mathrm{supp}(\hat{\boldsymbol{x}}) \subseteq \mathrm{supp}(\boldsymbol{x}_o)$*. Conversely, if*

$$m < \left(1 + \frac{\sigma^2}{\lambda^2 k} - \varepsilon\right) 2k \log(n-k) \tag{3.6.51}$$

then the probability that there exists a solution $\hat{\boldsymbol{x}}$ of the Lasso which satisfies $\mathrm{sign}(\hat{\boldsymbol{x}}) = \mathrm{sign}(\boldsymbol{x}_o)$ *is at most $Cn^{-\varepsilon}$. Above, $C > 0$ is a positive numerical constant.*

Partial vs. (Exact) Signed Support Recovery.

The notion of support recovery in Theorem 3.6.9 is somewhat weak: it only demands that

$$\mathrm{supp}(\hat{\boldsymbol{x}}) \subseteq \mathrm{supp}(\boldsymbol{x}_o). \tag{3.6.52}$$

Put another way, *the support contains no false positives*. In many applications, we would like to *exactly* recover the support – i.e., we would like

$$\mathrm{supp}(\hat{\boldsymbol{x}}) = \mathrm{supp}(\boldsymbol{x}_o). \tag{3.6.53}$$

For this, we need that the nonzero entries of $\boldsymbol{x}_o$ are not too small, so that they do not become "lost" in the noise. Under (3.6.49), it is possible to show that exact support recovery occurs, as long as the smallest nonzero entry of $\boldsymbol{x}_o$ is larger than λ: if

$$\min_{i \in I} |\boldsymbol{x}_{oi}| \; > \; C\lambda, \tag{3.6.54}$$

then $\mathrm{sign}(\hat{\boldsymbol{x}}) = \boldsymbol{\sigma}_o$ with high probability. In the remainder of this section, we will prove Theorem 3.6.9. Exercise **??** guides the reader through an extension of this argument, which shows that under the same assumptions,

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_\infty < C\lambda. \tag{3.6.55}$$

When the nonzero entries of $\boldsymbol{x}_o$ have magnitude at least $C\lambda$, this implies that $\operatorname{sign}(\hat{\boldsymbol{x}}) = \boldsymbol{\sigma}_o$, as desired.

Main ideas of the proof of Theorem 3.6.9.

The phase transition in Theorem 3.6.9 has a strikingly simple formula: $m_\star = 2k\log(n-k)$. The proof of this result is similar in spirit to our first proof of the correctness of ℓ^1-minimization, Theorem 3.2.3, which directly manipulated the optimality conditions for the recovery program.

Optimality conditions.

By differentiating the objective function (3.6.50), we can show that a given vector $\hat{\boldsymbol{x}}$ is optimal if and only if

$$\boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}) \in \lambda\partial\|\cdot\|_1(\hat{\boldsymbol{x}}). \tag{3.6.56}$$

Let $J = \operatorname{supp}(\hat{\boldsymbol{x}})$. Recall that the subdifferential $\partial\|\cdot\|_1(\hat{\boldsymbol{x}})$ consists of those vectors $\boldsymbol{v} \in \mathbb{R}^n$ such that $\boldsymbol{v}_J = \operatorname{sign}(\hat{\boldsymbol{x}}_J)$ and $\|\boldsymbol{v}_{J^c}\|_\infty \le 1$. Hence, the condition (3.6.56) decomposes into two conditions:

$$\boldsymbol{A}_J^*(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}) = \lambda\operatorname{sign}(\hat{\boldsymbol{x}}_J), \tag{3.6.57}$$

$$\|\boldsymbol{A}_{J^c}^*(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}})\|_\infty \le \lambda. \tag{3.6.58}$$

Much like the proof of Theorem 3.2.3, we will proceed as follows: we will construct a guess at a solution vector $\boldsymbol{x}_\star$ such that the equality constraints in (3.6.57) are automatically satisfied. We will then be left to check the inequality constraints (3.6.58). In particular, we will construct our guess $\boldsymbol{x}_\star$ at the solution by solving a *restricted* Lasso problem

$$\boldsymbol{x}_\star \in \arg\min_{\operatorname{supp}(\boldsymbol{x})\subseteq I}\left\{\tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1\right\}, \tag{3.6.59}$$

where $I = \operatorname{supp}(\boldsymbol{x}_o)$.

Recall that that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$. We can write

$$\boldsymbol{r} \doteq \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{A}_I(\boldsymbol{x}_{oI} - \boldsymbol{x}_{\star I}) + \boldsymbol{z}. \tag{3.6.60}$$

Notice that $\boldsymbol{r}$ depends only on $\boldsymbol{A}_I$ and $\boldsymbol{z}$; it is probabilistically independent of $\boldsymbol{A}_{I^c}$. The key work that we will do in proving Theorem 3.6.9 is to determine whether the ℓ^∞ norm constraint is satisfied on I^c. That is to say, we need to study

$$\|\boldsymbol{A}_{I^c}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star)\|_\infty = \|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty. \tag{3.6.61}$$

The matrix $\boldsymbol{A}_{I^c}$ is a Gaussian matrix; moreover, it is probabilistically independent of $\boldsymbol{r}$. Conditioned on $\boldsymbol{r}$, $\boldsymbol{A}_{I^c}^*\boldsymbol{r}$ is distributed as an $(n-k)$-dimensional iid $\mathcal{N}\left(0, \frac{\|\boldsymbol{r}\|_2^2}{m}\right)$ random vector. We will see that the ℓ^∞ norm of such a vector is sharply concentrated about $\|r\|_2 \sqrt{\frac{2\log(n-k)}{m}}$. The following lemma provides the control that we need:

Lemma 3.6.10. *Suppose that $\boldsymbol{q} = (q_1, \dots, q_d)$ is a $d \geq 2$-dimensional random vector, whose elements are independent $\mathcal{N}(0, \xi^2)$ random variables. Then, for any $\varepsilon \in [0, 1)$,*

$$\mathbb{P}\left[\|\boldsymbol{q}\|_\infty < \xi\sqrt{(2-\varepsilon)\log d}\right] \leq \exp\left(-\frac{d^{\varepsilon/2}}{4\sqrt{2\log d}}\right), \tag{3.6.62}$$

$$\mathbb{P}\left[\|\boldsymbol{q}\|_\infty > \xi\sqrt{(2+\varepsilon)\log d}\right] \leq 2d^{-\varepsilon/2}. \tag{3.6.63}$$

This lemma can be proved using relatively elementary ideas (the union bound for the upper bound, a direct calculation for the lower bound). Using this lemma, we conclude that, conditioned on $\boldsymbol{r}$ (i.e., with high probability in $\boldsymbol{A}_{I^c}$), $\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty$ is very close to $\|\boldsymbol{r}\|_2 \sqrt{\frac{2\log(n-k)}{m}}$. To understand whether this quantity is smaller than λ (and hence recovery succeeds) or larger than λ (and hence recovery fails), we will need to study the norm of $\boldsymbol{r}$.

Notice that $\boldsymbol{r} = \boldsymbol{A}_I(\boldsymbol{x}_{oI} - \boldsymbol{x}_{\star I}) + \boldsymbol{z}$. To study the size of $\boldsymbol{r}$ it will be important to understand the properties of the random matrix $\boldsymbol{A}_I$ and the random vector $\boldsymbol{z}$. Because $\boldsymbol{A}_I \in \mathbb{R}^{m\times k}$ is a "tall", random matrix, it is well-conditioned, in a sense that the following lemma makes precise:

Lemma 3.6.11. *Let $\boldsymbol{G} \in \mathbb{R}^{m\times k}$ be a random matrix whose entries are iid $\mathcal{N}(0, \frac{1}{m})$ random variables. Then, with high probability*

$$\|\boldsymbol{G}^*\boldsymbol{G} - \boldsymbol{I}\|_{\ell^2\to\ell^2} \leq C\sqrt{\frac{k}{m}} \tag{3.6.64}$$

The proof of this lemma follows similar lines to our proof of the RIP property of Gaussian matrices (discretization, tail bound, union bound). Using this lemma, we can control $\|\boldsymbol{r}\|_2$; combining with the above calculations, we obtain control on $\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty$. The prescription for the required number of measurements m follows by demanding that this quantity be smaller than λ. To formally prove Theorem 3.6.9, we need to do a bit more. First, we need to formally control $\|\boldsymbol{r}\|_2$ and $\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty$. This is sufficient to show that our putative solution $\boldsymbol{x}_\star$ is indeed optimal. Second, we need to argue that under the same conditions, *every* solution $\hat{\boldsymbol{x}}$ indeed satisfies $\text{supp}(\hat{\boldsymbol{x}}) \subseteq \text{supp}(\boldsymbol{x}_o)$. This will follow from some auxiliary reasoning about the subdifferential of the ℓ^1 norm. Finally, we obtain the converse portion of Theorem 3.6.9 by showing that when the number of measurements $m \ll m_\star$, with high probability $\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty > \lambda$, and hence no putative solution $\boldsymbol{x}_\star$ with $\text{sign}(\boldsymbol{x}_\star) = \boldsymbol{\sigma}_o$ can be optimal. We carry through all of this reasoning in the next section.

Proof of Theorem 3.6.9. We proceed as follows.

Sufficient condition for partial support recovery.

Let $I = \mathrm{supp}(\boldsymbol{x}_o)$. We wish to show that every solution $\hat{\boldsymbol{x}}$ to the Lasso problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; \varphi(\boldsymbol{x}) \doteq \tfrac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1, \tag{3.6.65}$$

satisfies $\mathrm{supp}(\boldsymbol{x}) \subseteq I$. To do this, we will generate a vector $\boldsymbol{x}_\star$ with $\mathrm{supp}(\boldsymbol{x}_\star) \subseteq I$, such that the residual

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star \tag{3.6.66}$$

satisfies

$$\begin{aligned} \boldsymbol{A}^*\boldsymbol{r} \quad &\in \quad \lambda\partial\left\|\cdot\right\|_1(\boldsymbol{x}_\star), && (3.6.67)\\ \text{and} \qquad \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\right\|_\infty \quad &< \quad \lambda. && (3.6.68)\end{aligned}$$

The first property implies that $\boldsymbol{x}_\star$ is optimal for the Lasso problem, since it implies that

$$\begin{aligned} \boldsymbol{0} \in \partial\varphi(\boldsymbol{x}_\star) \quad &= \quad \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{y}) + \lambda\partial\left\|\cdot\right\|_1(\boldsymbol{x}_\star)\\ &= \quad -\boldsymbol{r} + \lambda\partial\left\|\cdot\right\|_1(\boldsymbol{x}_\star). && (3.6.69)\end{aligned}$$

The property $\left\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\right\|_\infty < \lambda$ implies that any other optimal solution *also* has support contained in I. The reason is as follows: let $\lambda' = \lambda - \left\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\right\|_\infty > 0$. Then for any vector $\boldsymbol{v}$ supported on I^c, with $\left\|\boldsymbol{v}\right\|_\infty < \lambda'$, we have that

$$\boldsymbol{v} \in \partial\varphi_{\mathrm{Lasso}}(\boldsymbol{x}_\star). \tag{3.6.70}$$

For any $\boldsymbol{x}'$ with $\boldsymbol{x}'_{I^c} \neq \boldsymbol{0}$, set $\boldsymbol{v} = \lambda'\mathrm{sign}(\boldsymbol{x}'_{I^c})/2$ and note that by the subgradient inequality,

$$\begin{aligned} \varphi(\boldsymbol{x}') \quad &\geq \quad \varphi(\boldsymbol{x}_\star) + \langle \boldsymbol{x}' - \boldsymbol{x}_\star, \boldsymbol{v}\rangle\\ &= \quad \varphi(\boldsymbol{x}_\star) + \tfrac{\lambda'}{2}\left\|\boldsymbol{x}'_{I^c}\right\|_1\\ &> \quad \varphi(\boldsymbol{x}_\star), && (3.6.71)\end{aligned}$$

and hence, $\boldsymbol{x}'$ is not optimal. Thus, if there exists an $\boldsymbol{x}_\star$ satisfying (3.6.67)-(3.6.68), then every solution $\hat{\boldsymbol{x}}$ to the Lasso problem satisfies $\mathrm{supp}(\hat{\boldsymbol{x}}) \subseteq \mathrm{supp}(\boldsymbol{x}_o)$.

Constructing the putative solution $\boldsymbol{x}_\star$.

Let

$$\boldsymbol{x}_\star \in \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{x})\subseteq I} \; \tfrac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1. \tag{3.6.72}$$

Let $J = \mathrm{supp}(\boldsymbol{x}_\star) \subseteq I$. The KKT optimality conditions for this problem give that

$$\begin{aligned} \boldsymbol{A}_J^*(\boldsymbol{y} - \boldsymbol{A}_I\boldsymbol{x}_{\star I}) \quad &= \quad \lambda\,\mathrm{sign}(\boldsymbol{x}_{\star J}), && (3.6.73)\\ \left\|\boldsymbol{A}_{I\setminus J}^*(\boldsymbol{y} - \boldsymbol{A}_I\boldsymbol{x}_{\star I})\right\|_\infty \quad &\leq \quad \lambda. && (3.6.74)\end{aligned}$$

An equivalent way of expressing these conditions is to say that

$$\boldsymbol{A}_I^*(\boldsymbol{y} - \boldsymbol{A}_I \boldsymbol{x}_{\star I}) = \lambda \boldsymbol{\nu}, \tag{3.6.75}$$

for some $\boldsymbol{\nu} \in \partial \left\| \cdot \right\|_1 (\boldsymbol{x}_{\star I})$.

Because $\boldsymbol{y} = \boldsymbol{A}_I \boldsymbol{x}_{oI} + \boldsymbol{z}$, we can use (3.6.75) to express the difference $\boldsymbol{x}_{oI} - \boldsymbol{x}_{\star I}$ in terms of the subgradient $\boldsymbol{\nu}$ and the noise $\boldsymbol{z}$:

$$\boldsymbol{x}_{oI} - \boldsymbol{x}_{\star I} = (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} (\lambda \boldsymbol{\nu} - \boldsymbol{A}_I^* \boldsymbol{z}). \tag{3.6.76}$$

Notice that since $m > k$, with probability one, $\boldsymbol{A}_I^* \boldsymbol{A}_I$ is invertible, and so this expression indeed makes sense.

Verifying the KKT conditions.

We will prove that the restricted solution $\boldsymbol{x}_\star$ is indeed optimal for the full problem (3.6.65). The KKT conditions for *this problem* give that $\boldsymbol{x}_\star$ is optimal if and only if

$$\boldsymbol{A}^* (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star) \in \lambda \partial \left\| \cdot \right\|_1 (\boldsymbol{x}_\star). \tag{3.6.77}$$

Let $J = \mathrm{supp}(\boldsymbol{x}_\star)$. The above expression can be broken into two parts as

$$\begin{aligned} \boldsymbol{A}_J^* (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star) &= \lambda \mathrm{sign}(\boldsymbol{x}_{\star J}), & (3.6.78)\\ \left\| \boldsymbol{A}_{I \cap J^c}^* (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star) \right\|_\infty &\le \lambda, & (3.6.79)\\ \left\| \boldsymbol{A}_{I^c}^* (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star) \right\|_\infty &\le \lambda. & (3.6.80) \end{aligned}$$

Because $\breve{\boldsymbol{x}}_I$ satisfies the restricted KKT conditions, the first two conditions are automatically satisfied; to complete the proof, we establish the stronger version

$$\left\| \boldsymbol{A}_{I^c}^* (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star) \right\|_\infty < \lambda \tag{3.6.81}$$

of the third – this is the condition (3.6.68) that $\left\| \boldsymbol{r} \right\|_\infty < \lambda$. Using (3.6.76), we can express the residual $\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star$ as

$$\begin{aligned} \boldsymbol{r} &\doteq \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star \\ &= \left[\boldsymbol{I} - \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \boldsymbol{A}_I^* \right] \boldsymbol{z} + \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \lambda \breve{\boldsymbol{\nu}}. \end{aligned} \tag{3.6.82}$$

The two components of $\boldsymbol{r}$ are orthogonal, and so

$$\begin{aligned} \left\| \boldsymbol{r} \right\|_2 &= \sqrt{ \left\| \left[\boldsymbol{I} - \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \boldsymbol{A}_I^* \right] \boldsymbol{z} \right\|_2^2 + \left\| \boldsymbol{A}_I (\boldsymbol{A}_I^* \boldsymbol{A}_I)^{-1} \lambda \boldsymbol{\nu} \right\|_2^2 } \\ &\le \sqrt{ \left\| \boldsymbol{z} \right\|_2^2 + \lambda^2 \frac{\left\| \boldsymbol{\nu} \right\|_2^2}{\sigma_{\min}(\boldsymbol{A}_I^* \boldsymbol{A}_I)} } \\ &\le \sqrt{ \sigma^2 + \frac{\lambda^2 k}{1 - Ck/m} } \qquad \text{with high probability} \\ &\le \sqrt{ \sigma^2 + \lambda^2 k + C' \lambda^2 k^2 / m }. \end{aligned} \tag{3.6.83}$$

Applying the above lemma, with high probability in $\boldsymbol{A}_{I^c}$,

$$\begin{aligned} \|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty &< \sqrt{\frac{(2+\varepsilon)\log(n-k)}{m}}\,\|\boldsymbol{r}\|_2 \\ &\leq \lambda\left(\frac{(2k\log(n-k)\left(1+\frac{\sigma^2}{\lambda^2 k}+\varepsilon\right)}{m}\right)^{1/2}. \end{aligned} \tag{3.6.84}$$

Under our hypothesis on m, this is strictly smaller than λ, and so indeed (3.6.68) is verified.

No Signed Support Recovery when $m \ll m_\star$.

We next prove that when m is significantly smaller than $2k\log(n-k)$, no vector $\boldsymbol{x}$ satisfying

$$\operatorname{sign}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{x}_o) \tag{3.6.85}$$

can be a solution to the Lasso problem. Without loss of generality, we can assume that $m \geq k$.[24] Suppose on the contrary that $\boldsymbol{x}$ was the solution to the Lasso problem. Then $\boldsymbol{x}$ is also the solution to the restricted Lasso problem. Moreover, since $\operatorname{sign}(\boldsymbol{x}_I) = \boldsymbol{\sigma}_I$ has no zero entries, we have

$$\boldsymbol{r} = \left[\boldsymbol{I} - \boldsymbol{A}_I(\boldsymbol{A}_I^*\boldsymbol{A}_I)^{-1}\boldsymbol{A}_I^*\right]\boldsymbol{z} + \lambda\boldsymbol{A}_I(\boldsymbol{A}_I^*\boldsymbol{A}_I)^{-1}\boldsymbol{\sigma}_I. \tag{3.6.87}$$

With high probability,

$$\left\|\left[\boldsymbol{I} - \boldsymbol{A}_I(\boldsymbol{A}_I^*\boldsymbol{A}_I)^{-1}\boldsymbol{A}_I^*\right]\boldsymbol{z}\right\|_2^2 > (1-\varepsilon)(n-k)\sigma^2 \tag{3.6.88}$$

and

$$\left\|\boldsymbol{A}_I(\boldsymbol{A}_I^*\boldsymbol{A}_I)^{-1}\lambda\boldsymbol{\sigma}_I\right\|_2^2 > \frac{\lambda^2 k}{1+Ck/m} \tag{3.6.89}$$

whence, with high probability,

$$\|\boldsymbol{A}_{I^c}^*\boldsymbol{r}\|_\infty \;>\; \sqrt{\frac{(2-\varepsilon)\log(n-k)}{m}}\,\|\boldsymbol{r}\|_2 \tag{3.6.90}$$

and

$$\|\boldsymbol{r}\|_2 \;\geq\; \sqrt{\sigma^2(1-ck/m)+\lambda^2 k(1-c'k/m)}, \tag{3.6.91}$$

[24] If on the contrary, $m < k$, then the KKT conditions for the restricted problem become

$$\underbrace{\boldsymbol{A}_I^*\boldsymbol{A}_I}_{\text{Rank deficient}}\quad \boldsymbol{x}_I = \boldsymbol{A}_I^*\boldsymbol{y} - \lambda\boldsymbol{\sigma}_I. \tag{3.6.86}$$

This equation admits a solution if and only if $\boldsymbol{\sigma}_I \in \operatorname{range}(\boldsymbol{A}_I^*)$. Because $\boldsymbol{A}_I^*$ is a tall Gaussian matrix, the probability that its range contains the fixed vector $\boldsymbol{\sigma}_I$ is zero. So, when $m < k$, the probability that the Lasso problem admits a solution $\hat{\boldsymbol{x}}$ with $\operatorname{sign}(\hat{\boldsymbol{x}}) = \boldsymbol{\sigma}_o$ is zero.

Combining, we obtain

$$\begin{aligned} \|\boldsymbol{A}_{I^c}^* \boldsymbol{r}\|_\infty &> \lambda \sqrt{\frac{(2-\varepsilon) k \log(n-k) \left(1 + \frac{\sigma^2}{\lambda^2 k} + \varepsilon\right)}{m}} \\ &\geq \lambda. \end{aligned} \tag{3.6.92}$$

Hence, the putative solution $\boldsymbol{x}$ *is not* optimal for the full Lasso problem, with high probability in the matrix $\boldsymbol{A}$ and the noise $\boldsymbol{z}$. The above argument depends on $\boldsymbol{x}$ only through its sign and support pattern, and so on the same (large probability) bad event, *every* $\boldsymbol{x}$ having this sign and support pattern is suboptimal for the full Lasso problem. □

3.7 Notes and References

To our knowledge, historically the first result on exact recovery was obtained by B. Logan [Logan, 1965]. Analyses of sparse recovery with incoherent dictionaries are due to [Gribonval and Nielsen, 2003, Donoho and Elad, 2003]. The proof approach described here is due to [Fuchs, 2004]. The analysis of phase transitions via observation space geometry is developed in [Donoho, 2005, Donoho and Tanner, 2009, Donoho and Tanner, 2010]. The approach to phase transitions via coefficient space geometry follows [Amelunxen et al., 2014]. The analysis of phase transitions in support recovery is due to [Wainwright, 2009].