

Chapter 2

Sparse Solutions of Underdetermined Systems

In this chapter, we define the notions of vector sparsity and compressibility, and we establish some related inequalities used throughout the book. We will use basic results on vector and matrix norms, which can be found in Appendix A. We then investigate, in two different settings, the minimal number of linear measurements required to recover sparse vectors. We finally prove that ℓ_0 -minimization, the ideal recovery scheme, is NP-hard in general.

2.1 Sparsity and Compressibility

We start by defining the ideal notion of *sparsity*. We first introduce the notations $[N]$ for the set $\{1, 2, \dots, N\}$ and $\text{card}(S)$ for the cardinality of a set S . Furthermore, we write \overline{S} for the complement $[N] \setminus S$ of a set S in $[N]$.

Definition 2.1. The *support* of a vector $\mathbf{x} \in \mathbb{C}^N$ is the index set of its nonzero entries, i.e.,

$$\text{supp}(\mathbf{x}) := \{j \in [N] : x_j \neq 0\}.$$

The vector $\mathbf{x} \in \mathbb{C}^N$ is called *s-sparse* if at most s of its entries are nonzero, i.e., if

$$\|\mathbf{x}\|_0 := \text{card}(\text{supp}(\mathbf{x})) \leq s.$$

The customary notation $\|\mathbf{x}\|_0$ —the notation $\|\mathbf{x}\|_0^0$ would in fact be more appropriate—comes from the observation that

$$\|\mathbf{x}\|_p^p := \sum_{j=1}^N |x_j|^p \xrightarrow{p \rightarrow 0} \sum_{j=1}^N \mathbf{1}_{\{x_j \neq 0\}} = \text{card}(\{j \in [N] : x_j \neq 0\}).$$

Here, we used the notations $\mathbf{1}_{\{x_j \neq 0\}} = 1$ if $x_j \neq 0$ and $\mathbf{1}_{\{x_j \neq 0\}} = 0$ if $x_j = 0$. In other words the quantity $\|\mathbf{x}\|_0$ is the limit as p decreases to zero of the p th power of the ℓ_p -quasinorm of \mathbf{x} . It is abusively called the ℓ_0 -norm of \mathbf{x} , although it is neither a norm nor a quasinorm—see Appendix A for precise definitions of these notions. In practice, sparsity can be a strong constraint to impose, and we may prefer the weaker concept of *compressibility*. For instance, we may consider vectors that are nearly s -sparse, as measured by the *error of best s -term approximation*.

Definition 2.2. For $p > 0$, the ℓ_p -error of best s -term approximation to a vector $\mathbf{x} \in \mathbb{C}^N$ is defined by

$$\sigma_s(\mathbf{x})_p := \inf \{ \|\mathbf{x} - \mathbf{z}\|_p, \mathbf{z} \in \mathbb{C}^N \text{ is } s\text{-sparse} \}.$$

In the definition of $\sigma_s(\mathbf{x})_p$, the infimum is achieved by an s -sparse vector $\mathbf{z} \in \mathbb{C}^N$ whose nonzero entries equal the s largest absolute entries of \mathbf{x} . Hence, although such a vector $\mathbf{z} \in \mathbb{C}^N$ may not be unique, it achieves the infimum independently of $p > 0$.

Informally, we may call $\mathbf{x} \in \mathbb{C}^N$ a *compressible* vector if the error of its best s -term approximation decays quickly in s . According to the following proposition, this happens in particular if \mathbf{x} belongs to the unit ℓ_p -ball for some small $p > 0$, where the unit ℓ_p -ball is defined by

$$B_p^N := \{ \mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_p \leq 1 \}.$$

Consequently, the nonconvex balls B_p^N for $p < 1$ serve as good models for compressible vectors.

Proposition 2.3. For any $q > p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,

$$\sigma_s(\mathbf{x})_q \leq \frac{1}{s^{1/p-1/q}} \|\mathbf{x}\|_p.$$

Before proving this proposition, it is useful to introduce the notion of *nonincreasing rearrangement*.

Definition 2.4. The *nonincreasing rearrangement* of the vector $\mathbf{x} \in \mathbb{C}^N$ is the vector $\mathbf{x}^* \in \mathbb{R}^N$ for which

$$x_1^* \geq x_2^* \geq \dots \geq x_N^* \geq 0$$

and there is a permutation $\pi : [N] \rightarrow [N]$ with $x_j^* = |x_{\pi(j)}|$ for all $j \in [N]$.

Proof (of Proposition 2.3). If $\mathbf{x}^* \in \mathbb{R}_+^N$ is the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$, we have

$$\begin{aligned}
\sigma_s(\mathbf{x})_q^q &= \sum_{j=s+1}^N (x_j^*)^q \leq (x_s^*)^{q-p} \sum_{j=s+1}^N (x_j^*)^p \leq \left(\frac{1}{s} \sum_{j=1}^s (x_j^*)^p \right)^{\frac{q-p}{p}} \left(\sum_{j=s+1}^N (x_j^*)^p \right) \\
&\leq \left(\frac{1}{s} \|\mathbf{x}\|_p^p \right)^{\frac{q-p}{p}} \|\mathbf{x}\|_p^p = \frac{1}{s^{q/p-1}} \|\mathbf{x}\|_p^q.
\end{aligned}$$

The result follows by taking the power $1/q$ in both sides of this inequality. \square

We strengthen the previous proposition by finding the smallest possible constant $c_{p,q}$ in the inequality $\sigma_s(\mathbf{x})_q \leq c_{p,q} s^{-1/p+1/q} \|\mathbf{x}\|_p$. This can be skipped on first reading, but it is nonetheless informative because the proof technique, which consists in solving a convex optimization problem by hand, will reappear in Theorem 5.8 and Lemma 6.14.

Theorem 2.5. *For any $q > p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$, the inequality*

$$\sigma_s(\mathbf{x})_q \leq \frac{c_{p,q}}{s^{1/p-1/q}} \|\mathbf{x}\|_p$$

holds with

$$c_{p,q} := \left[\left(\frac{p}{q} \right)^{p/q} \left(1 - \frac{p}{q} \right)^{1-p/q} \right]^{1/p} \leq 1.$$

Let us point out that the frequent choice $p = 1$ and $q = 2$ gives

$$\sigma_s(\mathbf{x})_2 \leq \frac{1}{2\sqrt{s}} \|\mathbf{x}\|_1.$$

Proof. Let $\mathbf{x}^* \in \mathbb{R}_+^N$ be the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$. Setting $\alpha_j := (x_j^*)^p$, we will prove the equivalent statement

$$\left. \begin{aligned} \alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_N \geq 0 \\ \alpha_1 + \alpha_2 + \dots + \alpha_N \leq 1 \end{aligned} \right\} \implies \alpha_{s+1}^{q/p} + \alpha_{s+2}^{q/p} + \dots + \alpha_N^{q/p} \leq \frac{c_{p,q}^q}{s^{q/p-1}}.$$

Thus, with $r := q/p > 1$, we aim at maximizing the convex function

$$f(\alpha_1, \alpha_2, \dots, \alpha_N) := \alpha_{s+1}^r + \alpha_{s+2}^r + \dots + \alpha_N^r$$

over the convex polygon

$$\mathcal{C} := \{(\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N : \alpha_1 \geq \dots \geq \alpha_N \geq 0 \text{ and } \alpha_1 + \dots + \alpha_N \leq 1\}.$$

According to Theorem B.16, the maximum of f is attained at a vertex of \mathcal{C} . The vertices of \mathcal{C} are obtained as intersections of N hyperplanes arising by turning N

of the $(N + 1)$ inequality constraints into equalities. Thus, we have the following possibilities:

- If $\alpha_1 = \dots = \alpha_N = 0$, then $f(\alpha_1, \alpha_2, \dots, \alpha_N) = 0$.
- If $\alpha_1 + \dots + \alpha_N = 1$ and $\alpha_1 = \dots = \alpha_k > \alpha_{k+1} = \dots = \alpha_N = 0$ for some $1 \leq k \leq s$, then $f(\alpha_1, \alpha_2, \dots, \alpha_N) = 0$.
- If $\alpha_1 + \dots + \alpha_N = 1$ and $\alpha_1 = \dots = \alpha_k > \alpha_{k+1} = \dots = \alpha_N = 0$ for some $s + 1 \leq k \leq N$, then $\alpha_1 = \dots = \alpha_k = 1/k$, and consequently $f(\alpha_1, \alpha_2, \dots, \alpha_N) = (k - s)/k^r$.

It follows that

$$\max_{(\alpha_1, \dots, \alpha_N) \in \mathcal{C}} f(\alpha_1, \alpha_2, \dots, \alpha_N) = \max_{s+1 \leq k \leq N} \frac{k - s}{k^r}.$$

Considering k as a continuous variable, we now observe that the function $g(k) := (k - s)/k^r$ is increasing until the critical point $k^* = (r/(r - 1))s$ and decreasing thereafter. We obtain

$$\max_{(\alpha_1, \dots, \alpha_N) \in \mathcal{C}} f(\alpha_1, \alpha_2, \dots, \alpha_N) \leq g(k^*) = \frac{1}{r} \left(1 - \frac{1}{r}\right)^{r-1} \frac{1}{s^{r-1}} = c_{p,q}^q \frac{1}{s^{q/p-1}}.$$

This is the desired result. \square

Another possibility to define *compressibility* is to call a vector $\mathbf{x} \in \mathbb{C}^N$ *compressible* if the number

$$\text{card}(\{j \in [N] : |x_j| \geq t\})$$

of its significant—rather than nonzero—components is small. This naturally leads to the introduction of weak ℓ_p -spaces.

Definition 2.6. For $p > 0$, the weak ℓ_p space $w\ell_p^N$ denotes the space \mathbb{C}^N equipped with the quasinorm

$$\|\mathbf{x}\|_{p,\infty} := \inf \left\{ M \geq 0 : \text{card}(\{j \in [N] : |x_j| \geq t\}) \leq \frac{M^p}{t^p} \text{ for all } t > 0 \right\}.$$

To verify that the previous quantity indeed defines a quasinorm, we check, for any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and any $\lambda \in \mathbb{C}$, that $\|\mathbf{x}\| = 0 \Rightarrow \mathbf{x} = 0$, $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$, and $\|\mathbf{x} + \mathbf{y}\|_{p,\infty} \leq 2^{\max\{1, 1/p\}} (\|\mathbf{x}\|_{p,\infty} + \|\mathbf{y}\|_{p,\infty})$. The first two properties are easy, while the third property is a consequence of the more general statement below.

Proposition 2.7. Let $\mathbf{x}^1, \dots, \mathbf{x}^k \in \mathbb{C}^N$. Then, for $p > 0$,

$$\|\mathbf{x}^1 + \dots + \mathbf{x}^k\|_{p,\infty} \leq k^{\max\{1, 1/p\}} (\|\mathbf{x}^1\|_{p,\infty} + \dots + \|\mathbf{x}^k\|_{p,\infty}).$$

Proof. Let $t > 0$. If $|x_j^1 + \cdots + x_j^k| \geq t$ for some $j \in [N]$, then we have $|x_j^i| \geq t/k$ for some $i \in [k]$. This means that

$$\{j \in [N] : |x_j^1 + \cdots + x_j^k| \geq t\} \subset \bigcup_{i \in [k]} \{j \in [N] : |x_j^i| \geq t/k\}.$$

We derive

$$\begin{aligned} \text{card}(\{j \in [N] : |x_j^1 + \cdots + x_j^k| \geq t\}) &\leq \sum_{i \in [k]} \frac{\|\mathbf{x}^i\|_{p,\infty}^p}{(t/k)^p} \\ &= \frac{k^p (\|\mathbf{x}^1\|_{p,\infty}^p + \cdots + \|\mathbf{x}^k\|_{p,\infty}^p)}{t^p}. \end{aligned}$$

According to the definition of the weak ℓ_p -quasinorm of $\mathbf{x}^1 + \cdots + \mathbf{x}^k$, we obtain

$$\|\mathbf{x}^1 + \cdots + \mathbf{x}^k\|_{p,\infty} \leq k (\|\mathbf{x}^1\|_{p,\infty}^p + \cdots + \|\mathbf{x}^k\|_{p,\infty}^p)^{1/p}.$$

Now, if $p \leq 1$, comparing the ℓ_p and ℓ_1 norms in \mathbb{R}^k gives

$$(\|\mathbf{x}^1\|_{p,\infty}^p + \cdots + \|\mathbf{x}^k\|_{p,\infty}^p)^{1/p} \leq k^{1/p-1} (\|\mathbf{x}^1\|_{p,\infty} + \cdots + \|\mathbf{x}^k\|_{p,\infty}),$$

and if $p \geq 1$, comparing the ℓ_p and ℓ_1 norms in \mathbb{R}^k gives

$$(\|\mathbf{x}^1\|_{p,\infty}^p + \cdots + \|\mathbf{x}^k\|_{p,\infty}^p)^{1/p} \leq \|\mathbf{x}^1\|_{p,\infty} + \cdots + \|\mathbf{x}^k\|_{p,\infty}.$$

The result immediately follows. \square

Remark 2.8. The constant $k^{\max\{1, 1/p\}}$ in Proposition 2.7 is sharp; see Exercise 2.2.

It is sometimes preferable to invoke the following alternative expression for the weak ℓ_p -quasinorm of a vector $\mathbf{x} \in \mathbb{C}^N$.

Proposition 2.9. For $p > 0$, the weak ℓ_p -quasinorm of a vector $\mathbf{x} \in \mathbb{C}^N$ can be expressed as

$$\|\mathbf{x}\|_{p,\infty} = \max_{k \in [N]} k^{1/p} x_k^*,$$

where $\mathbf{x}^* \in \mathbb{R}_+^N$ denotes the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$.

Proof. Given $\mathbf{x} \in \mathbb{C}^N$, in view of $\|\mathbf{x}\|_{p,\infty} = \|\mathbf{x}^*\|_{p,\infty}$, we need to establish that $\|\mathbf{x}\| := \max_{k \in [N]} k^{1/p} x_k^*$ equals $\|\mathbf{x}^*\|_{p,\infty}$. For $t > 0$, we first note that either $\{j \in [N] : x_j^* \geq t\} = [k]$ for some $k \in [N]$ or $\{j \in [N] : x_j^* \geq t\} = \emptyset$. In the former case, $t \leq x_k^* \leq \|\mathbf{x}\|/k^{1/p}$, and hence, $\text{card}(\{j \in [N] : x_j^* \geq t\}) = k \leq \|\mathbf{x}\|^p/t^p$. This inequality holds trivially in the case that $\{j \in [N] : x_j^* \geq t\} = \emptyset$.

According to the definition of the weak ℓ_p -quasinorm, we obtain $\|\mathbf{x}^*\|_{p,\infty} \leq \|\mathbf{x}\|$. Let us now suppose that $\|\mathbf{x}\| > \|\mathbf{x}^*\|_{p,\infty}$, so that $\|\mathbf{x}\| \geq (1 + \epsilon)\|\mathbf{x}^*\|_{p,\infty}$ for some $\epsilon > 0$. This means that $k^{1/p}x_k^* \geq (1 + \epsilon)\|\mathbf{x}^*\|_{p,\infty}$ for some $k \in [N]$. Therefore, the set

$$\{j \in [N] : x_j^* \geq (1 + \epsilon)\|\mathbf{x}^*\|_{p,\infty}/k^{1/p}\}$$

contains the set $[k]$. The definition of the weak ℓ_p -quasinorm yields

$$k \leq \frac{\|\mathbf{x}^*\|_{p,\infty}^p}{((1 + \epsilon)\|\mathbf{x}^*\|_{p,\infty}/k^{1/p})^p} = \frac{k}{(1 + \epsilon)^p},$$

which is a contradiction. We conclude that $\|\mathbf{x}\| = \|\mathbf{x}^*\|_{p,\infty}$. \square

This alternative expression of the weak ℓ_p -quasinorm provides a slightly easier way to compare it to the ℓ_p -(quasi)norm, as follows.

Proposition 2.10. *For any $p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,*

$$\|\mathbf{x}\|_{p,\infty} \leq \|\mathbf{x}\|_p.$$

Proof. For $k \in [N]$, we write

$$\|\mathbf{x}\|_p^p = \sum_{j=1}^N (x_j^*)^p \geq \sum_{j=1}^k (x_j^*)^p \geq k(x_k^*)^p.$$

Raising to the power $1/p$ and taking the maximum over k gives the result. \square

The alternative expression of the weak ℓ_p -quasinorm also enables us to easily establish a variation of Proposition 2.3 where weak ℓ_p replaces ℓ_p .

Proposition 2.11. *For any $q > p > 0$ and $\mathbf{x} \in \mathbb{C}^N$, the inequality*

$$\sigma_s(\mathbf{x})_q \leq \frac{d_{p,q}}{s^{1/p-1/q}} \|\mathbf{x}\|_{p,\infty}$$

holds with

$$d_{p,q} := \left(\frac{p}{q-p}\right)^{1/q}.$$

Proof. We may assume without loss of generality that $\|\mathbf{x}\|_{p,\infty} \leq 1$, so that $x_k^* \leq 1/k^{1/p}$ for all $k \in [N]$. We then have

$$\begin{aligned}
\sigma_s(\mathbf{x})_q^q &= \sum_{k=s+1}^N (x_k^*)^q \leq \sum_{k=s+1}^N \frac{1}{k^{q/p}} \leq \int_s^N \frac{1}{t^{q/p}} dt = -\frac{1}{q/p-1} \frac{1}{t^{q/p-1}} \Big|_{t=s}^{t=N} \\
&\leq \frac{p}{q-p} \frac{1}{s^{q/p-1}} .
\end{aligned}$$

Taking the power $1/q$ yields the desired result. \square

Proposition 2.11 shows that vectors $\mathbf{x} \in \mathbb{C}^N$ which are compressible in the sense that $\|\mathbf{x}\|_{p,\infty} \leq 1$ for small $p > 0$ are also compressible in the sense that their errors of best s -term approximation decay quickly with s .

We close this section with a technical result on the nonincreasing rearrangement.

Lemma 2.12. *The nonincreasing rearrangement satisfies, for $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$,*

$$\|\mathbf{x}^* - \mathbf{z}^*\|_\infty \leq \|\mathbf{x} - \mathbf{z}\|_\infty . \quad (2.1)$$

Moreover, for $s \in [N]$,

$$|\sigma_s(\mathbf{x})_1 - \sigma_s(\mathbf{z})_1| \leq \|\mathbf{x} - \mathbf{z}\|_1 , \quad (2.2)$$

and for $k > s$,

$$(k-s)x_k^* \leq \|\mathbf{x} - \mathbf{z}\|_1 + \sigma_s(\mathbf{z})_1 . \quad (2.3)$$

Proof. For $j \in [N]$, the index set of j largest absolute entries of \mathbf{x} intersects the index set of $N-j+1$ smallest absolute entries of \mathbf{z} . Picking an index ℓ in this intersection, we obtain

$$x_j^* \leq |x_\ell| \leq |z_\ell| + \|\mathbf{x} - \mathbf{z}\|_\infty \leq z_j^* + \|\mathbf{x} - \mathbf{z}\|_\infty .$$

Reversing the roles of \mathbf{x} and \mathbf{z} shows (2.1).

Next, let $\mathbf{v} \in \mathbb{C}^N$ be a best s -term approximation to \mathbf{z} . Then

$$\sigma_s(\mathbf{x})_1 \leq \|\mathbf{x} - \mathbf{v}\|_1 \leq \|\mathbf{x} - \mathbf{z}\|_1 + \|\mathbf{z} - \mathbf{v}\|_1 = \|\mathbf{x} - \mathbf{z}\|_1 + \sigma_s(\mathbf{z})_1 ,$$

and again by symmetry this establishes (2.2). The inequality (2.3) follows from (2.2) by noting that

$$(k-s)x_k^* \leq \sum_{j=s+1}^k x_j^* \leq \sum_{j \geq s+1} x_j^* = \sigma_s(\mathbf{x})_1 .$$

This completes the proof. \square

2.2 Minimal Number of Measurements

The compressive sensing problem consists in reconstructing an s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ from

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{A} \in \mathbb{C}^{m \times N}$ is the so-called measurement matrix. With $m < N$, this system of linear equations is underdetermined, but the sparsity assumption hopefully helps in identifying the original vector \mathbf{x} .

In this section, we examine the question of the minimal number of linear measurements needed to reconstruct s -sparse vectors from these measurements, regardless of the practicality of the reconstruction scheme. This question can in fact take two meanings, depending on whether we require that the measurement scheme allows for the reconstruction of all s -sparse vectors $\mathbf{x} \in \mathbb{C}^N$ simultaneously or whether we require that, given an s -sparse vector $\mathbf{x} \in \mathbb{C}^N$, the measurement scheme allows for the reconstruction of this specific vector. While the second scenario seems to be unnatural at first sight because the vector \mathbf{x} is unknown a priori, it will become important later when aiming at recovery guarantees when the matrix \mathbf{A} is chosen at random and the sparse vector \mathbf{x} is fixed (so-called nonuniform recovery guarantees).

The minimal number m of measurements depends on the setting considered, namely, it equals $2s$ in the first case and $s + 1$ in the second case. However, we will see in Chap. 11 that if we also require the reconstruction scheme to be stable (the meaning will be made precise later), then the minimal number of required measurements additionally involves a factor of $\ln(N/s)$, so that recovery will never be stable with only $2s$ measurements.

Before separating the two settings discussed above, it is worth pointing out the equivalence of the following properties for given sparsity s , matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, and s -sparse $\mathbf{x} \in \mathbb{C}^N$:

- (a) The vector \mathbf{x} is the unique s -sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{y}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$, that is, $\{\mathbf{z} \in \mathbb{C}^N : \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}, \|\mathbf{z}\|_0 \leq s\} = \{\mathbf{x}\}$.
- (b) The vector \mathbf{x} can be reconstructed as the unique solution of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \quad (\mathbf{P}_0)$$

Indeed, if an s -sparse $\mathbf{x} \in \mathbb{C}^N$ is the unique s -sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{y}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$, then a solution \mathbf{x}^\sharp of (\mathbf{P}_0) is s -sparse and satisfies $\mathbf{A}\mathbf{x}^\sharp = \mathbf{y}$, so that $\mathbf{x}^\sharp = \mathbf{x}$. This shows $(a) \Rightarrow (b)$. The implication $(b) \Rightarrow (a)$ is clear.

Recovery of All Sparse Vectors

Before stating the main result for this case, we observe that the uniqueness of sparse solutions of underdetermined linear systems can be reformulated in several ways.

For a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and a subset $S \subset [N]$, we use the notation \mathbf{A}_S to indicate the column submatrix of \mathbf{A} consisting of the columns indexed by S . Similarly, for $\mathbf{x} \in \mathbb{C}^N$ we denote by \mathbf{x}_S either the subvector in \mathbb{C}^S consisting of the entries indexed by S , that is, $(\mathbf{x}_S)_\ell = x_\ell$ for $\ell \in S$, or the vector in \mathbb{C}^N which coincides with \mathbf{x} on the entries in S and is zero on the entries outside S , that is,

$$(\mathbf{x}_S)_\ell = \begin{cases} x_\ell & \text{if } \ell \in S, \\ 0 & \text{if } \ell \notin S. \end{cases} \quad (2.4)$$

It should always be clear from the context which of the two options applies.

Theorem 2.13. *Given $\mathbf{A} \in \mathbb{C}^{m \times N}$, the following properties are equivalent:*

- (a) *Every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s -sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, that is, if $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$ and both \mathbf{x} and \mathbf{z} are s -sparse, then $\mathbf{x} = \mathbf{z}$.*
- (b) *The null space $\ker \mathbf{A}$ does not contain any $2s$ -sparse vector other than the zero vector, that is, $\ker \mathbf{A} \cap \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_0 \leq 2s\} = \{\mathbf{0}\}$.*
- (c) *For every $S \subset [N]$ with $\text{card}(S) \leq 2s$, the submatrix \mathbf{A}_S is injective as a map from \mathbb{C}^S to \mathbb{C}^m .*
- (d) *Every set of $2s$ columns of \mathbf{A} is linearly independent.*

Proof. (b) \Rightarrow (a) Let \mathbf{x} and \mathbf{z} be s -sparse with $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$. Then $\mathbf{x} - \mathbf{z}$ is $2s$ -sparse and $\mathbf{A}(\mathbf{x} - \mathbf{z}) = \mathbf{0}$. If the kernel does not contain any $2s$ -sparse vector different from the zero vector, then $\mathbf{x} = \mathbf{z}$.

(a) \Rightarrow (b) Conversely, assume that for every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$, we have $\{\mathbf{z} \in \mathbb{C}^N : \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}, \|\mathbf{z}\|_0 \leq s\} = \{\mathbf{x}\}$. Let $\mathbf{v} \in \ker \mathbf{A}$ be $2s$ -sparse. We can write $\mathbf{v} = \mathbf{x} - \mathbf{z}$ for s -sparse vectors \mathbf{x}, \mathbf{z} with $\text{supp } \mathbf{x} \cap \text{supp } \mathbf{z} = \emptyset$. Then $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$ and by assumption $\mathbf{x} = \mathbf{z}$. Since the supports of \mathbf{x} and \mathbf{z} are disjoint, it follows that $\mathbf{x} = \mathbf{z} = \mathbf{0}$ and $\mathbf{v} = \mathbf{0}$.

For the equivalence of (b), (c), and (d), we observe that for a $2s$ -sparse vector \mathbf{v} with $S = \text{supp } \mathbf{v}$, we have $\mathbf{A}\mathbf{v} = \mathbf{A}_S \mathbf{v}_S$. Noting that $S = \text{supp } \mathbf{v}$ ranges through all possible subsets of $[N]$ of cardinality $\text{card}(S) \leq 2s$ when \mathbf{v} ranges through all possible $2s$ -sparse vectors completes the proof by basic linear algebra. \square

We observe, in particular, that if it is possible to reconstruct every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ from the knowledge of its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, then (a) holds and consequently so does (d). This implies $\text{rank}(\mathbf{A}) \geq 2s$. We also have $\text{rank}(\mathbf{A}) \leq m$, because the rank is at most equal to the number of rows. Therefore, the number of measurements needed to reconstruct every s -sparse vector always satisfies

$$m \geq 2s.$$

We are now going to see that $m = 2s$ measurements suffice to reconstruct every s -sparse vector—at least in theory.

Theorem 2.14. *For any integer $N \geq 2s$, there exists a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = 2s$ rows such that every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be recovered from its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ as a solution of (\mathbf{P}_0) .*

Proof. Let us fix $t_N > \dots > t_2 > t_1 > 0$ and consider the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = 2s$ defined by

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_N \\ \vdots & \vdots & \dots & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \dots & t_N^{2s-1} \end{bmatrix}. \quad (2.5)$$

Let $S = \{j_1 < \dots < j_{2s}\}$ be an index set of cardinality $2s$. The square matrix $\mathbf{A}_S \in \mathbb{C}^{2s \times 2s}$ is (the transpose of) a *Vandermonde matrix*. Theorem A.24 yields

$$\det(\mathbf{A}_S) = \begin{vmatrix} 1 & 1 & \dots & 1 \\ t_{j_1} & t_{j_2} & \dots & t_{j_{2s}} \\ \vdots & \vdots & \dots & \vdots \\ t_{j_1}^{2s-1} & t_{j_2}^{2s-1} & \dots & t_{j_{2s}}^{2s-1} \end{vmatrix} = \prod_{k < \ell} (t_{j_\ell} - t_{j_k}) > 0.$$

This shows that \mathbf{A}_S is invertible, in particular injective. Since the condition (c) of Theorem 2.13 is fulfilled, every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s -sparse vector satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, so it can be recovered as the unique solution of (\mathbf{P}_0) . \square

Many other matrices meet the condition (c) of Theorem 2.13. As an example, the integer powers of t_1, \dots, t_N in the matrix of (2.5) do not need to be the consecutive integers $0, 1, \dots, 2s - 1$. Instead of the $N \times N$ Vandermonde matrix associated with $t_N > \dots > t_1 > 0$, we can start with any matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ that is *totally positive*, i.e., that satisfies $\det \mathbf{M}_{I,J} > 0$ for any sets $I, J \subset [N]$ of the same cardinality, where $\mathbf{M}_{I,J}$ represents the submatrix of \mathbf{M} with rows indexed by I and columns indexed by J . We then select any $m = 2s$ rows of \mathbf{M} , indexed by a set I , say, to form the matrix \mathbf{A} . For an index $S \subset [N]$ of cardinality $2s$, the matrix \mathbf{A}_S reduces to $\mathbf{M}_{I,S}$; hence, it is invertible. As another example, the numbers t_N, \dots, t_1 do not need to be positive nor real, as long as $\det(\mathbf{A}_S) \neq 0$ instead of $\det(\mathbf{A}_S) > 0$. In particular, with $t_\ell = e^{2\pi i(\ell-1)/N}$ for $\ell \in [N]$, Theorem A.24 guarantees that the (rescaled) partial Fourier matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & e^{2\pi i/N} & e^{2\pi i 2/N} & \dots & e^{2\pi i(N-1)/N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{2\pi i(2s-1)/N} & e^{2\pi i(2s-1)2/N} & \dots & e^{2\pi i(2s-1)(N-1)/N} \end{bmatrix}$$

allows for the reconstruction of every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^{2s}$. In fact, an argument similar to the one we will use for Theorem 2.16 below shows that the set of $(2s) \times N$ matrices such that $\det(\mathbf{A}_S) = 0$ for some $S \subset [N]$ with $\text{card}(S) \leq 2s$ has Lebesgue measure zero; hence, most $(2s) \times N$ matrices allow the reconstruction of every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^{2s}$. In general, the reconstruction procedure consisting of solving (\mathbf{P}_0) is not feasible in practice, as will be shown in Sect. 2.3. However, in the case of Fourier measurements, a better reconstruction scheme based on the Prony method can be used.

Theorem 2.15. *For any $N \geq 2s$, there exists a practical procedure for the reconstruction of every $2s$ -sparse vector from its first $m = 2s$ discrete Fourier measurements.*

Proof. Let $\mathbf{x} \in \mathbb{C}^N$ be an s -sparse vector, which we interpret as a function x from $\{0, 1, \dots, N-1\}$ into \mathbb{C} supported on an index set $S \subset \{0, 1, \dots, N-1\}$ of size s . We suppose that this vector is observed via its first $2s$ discrete Fourier coefficients $\hat{x}(0), \dots, \hat{x}(2s-1)$, where

$$\hat{x}(j) := \sum_{k=0}^{N-1} x(k) e^{-2\pi i j k / N}, \quad 0 \leq j \leq N-1.$$

We consider the trigonometric polynomial of degree s defined by

$$p(t) := \frac{1}{N} \prod_{k \in S} (1 - e^{-2\pi i k / N} e^{2\pi i t / N}).$$

This polynomial vanishes exactly for $t \in S$, so we aim at finding the unknown set S by determining p or equivalently its Fourier transform \hat{p} . We note that, since x vanishes on the complementary set \bar{S} of S in $\{0, 1, \dots, N-1\}$, we have $p(t)x(t) = 0$ for all $0 \leq t \leq N-1$. By discrete convolution, we obtain $\hat{p} * \hat{x} = \widehat{p \cdot x} = 0$, that is to say,

$$(\hat{p} * \hat{x})(j) := \sum_{k=0}^{N-1} \hat{p}(k) \cdot \hat{x}(j-k \bmod N) = 0 \quad \text{for all } 0 \leq j \leq N-1. \quad (2.6)$$

We also note that, since $\frac{1}{N}\hat{p}(k)$ is the coefficient of $p(t)$ on the monomial $e^{2\pi i k t / N}$ and since p has degree s , we have $\hat{p}(0) = 1$ and $\hat{p}(k) = 0$ for all $k > s$. It remains to determine the s discrete Fourier coefficients $\hat{p}(1), \dots, \hat{p}(s)$. For this purpose, we write the s equations (2.6) in the range $s \leq j \leq 2s-1$ in the form

$$\begin{aligned} \hat{x}(s) &+ \hat{p}(1)\hat{x}(s-1) + \dots + \hat{p}(s)\hat{x}(0) &= 0, \\ \hat{x}(s+1) &+ \hat{p}(1)\hat{x}(s) + \dots + \hat{p}(s)\hat{x}(1) &= 0, \\ \vdots & & \ddots & \vdots \\ \hat{x}(2s-1) &+ \hat{p}(1)\hat{x}(2s-2) + \dots + \hat{p}(s)\hat{x}(s-1) &= 0. \end{aligned}$$

This translates into the system

$$\begin{bmatrix} \hat{x}(s-1) & \hat{x}(s-2) & \cdots & \hat{x}(0) \\ \hat{x}(s) & \hat{x}(s-1) & \cdots & \hat{x}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}(2s-2) & \hat{x}(2s-3) & \cdots & \hat{x}(s-1) \end{bmatrix} \begin{bmatrix} \hat{p}(1) \\ \hat{p}(2) \\ \vdots \\ \hat{p}(s) \end{bmatrix} = - \begin{bmatrix} \hat{x}(s) \\ \hat{x}(s+1) \\ \vdots \\ \hat{x}(2s-1) \end{bmatrix}.$$

Because $\hat{x}(0), \dots, \hat{x}(2s-1)$ are known, we solve for $\hat{p}(1), \dots, \hat{p}(s)$. Since the Toeplitz matrix above is not always invertible—take, e.g., $x = [1, 0, \dots, 0]^\top$, so that $\hat{x} = [1, 1, \dots, 1]^\top$ —we obtain a solution $\hat{q}(1), \dots, \hat{q}(s)$ not guaranteed to be $\hat{p}(1), \dots, \hat{p}(s)$. Appending the values $\hat{q}(0) = 1$ and $\hat{q}(k) = 0$ for all $k > s$, the linear system reads

$$(\hat{q} * \hat{x})(j) = 0 \quad \text{for all } s \leq j \leq 2s-1.$$

Therefore, the s -sparse vector $q \cdot x$ has a Fourier transform $\widehat{q \cdot x} = \hat{q} * \hat{x}$ vanishing on a set of s consecutive indices. Writing this in matrix form and using Theorem A.24, we derive that $q \cdot x = 0$, so that the trigonometric polynomial q vanishes on S . Since the degree of q is at most s , the set of zeros of q coincide with the set S , which can thus be found by solving a polynomial equation—or simply by identifying the s smallest values of $|q(j)|$, $0 \leq j \leq N-1$. Finally, the values of $x(j)$, $j \in S$, are obtained by solving the overdetermined system of $2s$ linear equations imposed by the knowledge of $\hat{x}(0), \dots, \hat{x}(2s-1)$. \square

Despite its appeal, the reconstruction procedure just described hides some important drawbacks. Namely, it is not stable with respect to sparsity defects nor is it robust with respect to measurement errors. The reader is invited to verify this statement numerically in Exercise 2.8. In fact, we will prove in Chap. 11 that any stable scheme for s -sparse reconstruction requires at least $m \approx cs \ln(eN/s)$ linear measurements, where $c > 0$ is a constant depending on the stability requirement.

Recovery of Individual Sparse Vectors

In the next setting, the s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is fixed before the measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is chosen. The conditions for the vector \mathbf{x} to be the unique s -sparse vector consistent with the measurements depend on \mathbf{A} as well as on \mathbf{x} itself. While this seems unnatural at first sight because \mathbf{x} is unknown a priori, the philosophy is that the conditions will be met for *most* $(s+1) \times N$ matrices. This setting is relevant since the measurement matrices are often chosen at random.

Theorem 2.16. *For any $N \geq s+1$, given an s -sparse vector $\mathbf{x} \in \mathbb{C}^N$, there exists a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = s+1$ rows such that the vector \mathbf{x} can be reconstructed from its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ as a solution of (\mathbf{P}_0) .*

Proof. Let $\mathbf{A} \in \mathbb{C}^{(s+1) \times N}$ be a matrix for which the s -sparse vector \mathbf{x} cannot be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ (via ℓ_0 -minimization). This means that there exists a vector $\mathbf{z} \in \mathbb{C}^N$ distinct from \mathbf{x} , supported on a set $S = \text{supp}(\mathbf{z}) = \{j_1, \dots, j_s\}$ of size at most s (if $\|\mathbf{z}\|_0 < s$, we fill up S with arbitrary elements $j_\ell \in [N]$), such that $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. If $\text{supp}(\mathbf{x}) \subset S$, then the equality $(\mathbf{A}(\mathbf{z} - \mathbf{x}))_{[s]} = 0$ shows that the square matrix $\mathbf{A}_{[s],S}$ is noninvertible; hence,

$$f(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) := \det(\mathbf{A}_{[s],S}) = 0.$$

If $\text{supp}(\mathbf{x}) \not\subset S$, then the space $V := \{\mathbf{u} \in \mathbb{C}^N : \text{supp}(\mathbf{u}) \subset S\} + \mathbb{C}\mathbf{x}$ has dimension $s + 1$, and the linear map $G : V \rightarrow \mathbb{C}^{s+1}, \mathbf{v} \mapsto \mathbf{A}\mathbf{v}$ is noninvertible, since $G(\mathbf{z} - \mathbf{x}) = 0$. The matrix of the linear map G in the basis $(\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_s}, \mathbf{x})$ of V takes the form

$$B_{\mathbf{x},S} := \begin{bmatrix} a_{1,j_1} & \cdots & a_{1,j_s} & \sum_{j \in \text{supp}(\mathbf{x})} x_j a_{1,j} \\ \vdots & \ddots & \vdots & \vdots \\ a_{s+1,j_1} & \cdots & a_{s+1,j_s} & \sum_{j \in \text{supp}(\mathbf{x})} x_j a_{s+1,j} \end{bmatrix},$$

and we have

$$g_S(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) := \det(B_{\mathbf{x},S}) = 0.$$

This shows that the entries of the matrix \mathbf{A} satisfy

$$(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) \in f^{-1}(\{0\}) \cup \bigcup_{\text{card}(S)=s} g_S^{-1}(\{0\}).$$

But since f and all g_S , $\text{card}(S) = s$, are nonzero polynomial functions of the variables $(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N})$, the sets $f^{-1}(\{0\})$ and $g_S^{-1}(\{0\})$, $\text{card}(S) = s$, have Lebesgue measure zero and so does their union. It remains to choose the entries of the matrix \mathbf{A} outside of this union of measure zero to ensure that the vector \mathbf{x} can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$. \square

2.3 NP-Hardness of ℓ_0 -Minimization

As mentioned in Sect. 2.2, reconstructing an s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ from its measurement vector $\mathbf{y} \in \mathbb{C}^m$ amounts to solving the ℓ_0 -minimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \quad (\text{P}_0)$$

Since a minimizer has sparsity at most s , the straightforward approach for finding it consists in solving every rectangular system $\mathbf{A}_S \mathbf{u} = \mathbf{y}$, or rather every square system $\mathbf{A}_S^* \mathbf{A}_S \mathbf{u} = \mathbf{A}_S^* \mathbf{y}$, for $\mathbf{u} \in \mathbb{C}^S$ where S runs through all the possible subsets of $[N]$ with size s . However, since the number $\binom{N}{s}$ of these subsets is prohibitively large, such a straightforward approach is completely unpractical. By way of illustration, for small problem sizes $N = 1000$ and $s = 10$, we would have to solve $\binom{1000}{10} \geq \left(\frac{1000}{10}\right)^{10} = 10^{20}$ linear systems of size 10×10 . Even if each such system could be solved in 10^{-10} seconds, the time required to solve (\mathbf{P}_0) with this approach would still be 10^{10} seconds, i.e., more than 300 years. We are going to show that solving (\mathbf{P}_0) in fact is intractable for any possible approach. Precisely, for any fixed $\eta \geq 0$, we are going to show that the more general problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_0 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \quad (\mathbf{P}_{0,\eta})$$

is NP-hard.

We start by introducing the necessary terminology from computational complexity. First, a polynomial-time algorithm is an algorithm performing its task in a number of steps bounded by a polynomial expression in the size of the input. Next, let us describe in a rather informal way a few classes of decision problems:

- The class \mathfrak{P} of P-problems consists of all decision problems for which there exists a polynomial-time algorithm finding a solution.
- The class \mathfrak{NP} of NP-problems consists of all decision problems for which there exists a polynomial-time algorithm certifying a solution. Note that the class \mathfrak{P} is clearly contained in the class \mathfrak{NP} .
- The class \mathfrak{NP} -hard of NP-hard problems consist of all problems (not necessarily decision problems) for which a solving algorithm could be transformed in polynomial time into a solving algorithm for any NP-problem. Roughly speaking, this is the class of problems at least as hard as any NP-problem. Note that the class \mathfrak{NP} -hard is not contained in the class \mathfrak{NP} .
- The class \mathfrak{NP} -complete of NP-complete problems consist of all problems that are both NP and NP-hard; in other words, it consists of all the NP-problems at least as hard as any other NP-problem.

The situation can be summarized visually as in Fig. 2.1. It is a common belief that \mathfrak{P} is strictly contained in \mathfrak{NP} , that is to say, that there are problems for which potential solutions can be certified, but for which a solution cannot be found in polynomial time. However, this remains a major open question to this day. There is a vast catalog of NP-complete problems, the most famous of which being perhaps the traveling salesman problem. The one we are going to use is *exact cover by 3-sets*.

Exact cover by 3-sets problem

Given a collection $\{\mathcal{C}_i, i \in [N]\}$ of 3-element subsets of $[m]$, does there exist an exact cover (a partition) of $[m]$, i.e., a set $J \subset [N]$ such that $\cup_{j \in J} \mathcal{C}_j = [m]$ and $\mathcal{C}_j \cap \mathcal{C}_{j'} = \emptyset$ for all $j, j' \in J$ with $j \neq j'$?

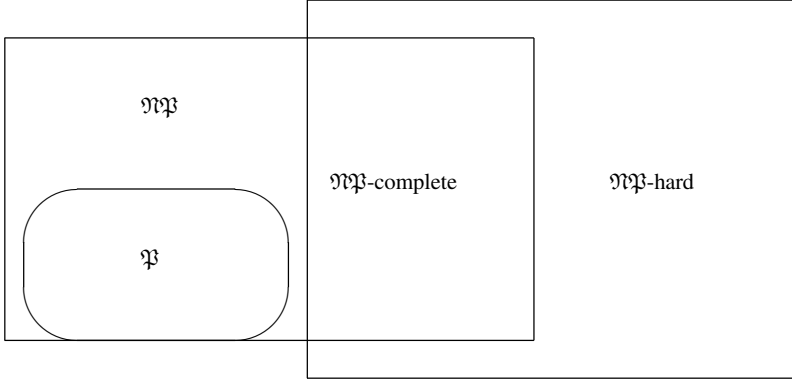


Fig. 2.1 Schematic representation of P, NP, NP-complete, and NP-hard problems

Taking for granted that this problem is NP-complete, we can now prove the main result of this section.

Theorem 2.17. *For any $\eta \geq 0$, the ℓ_0 -minimization problem $(\mathbf{P}_{0,\eta})$ for general $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\mathbf{y} \in \mathbb{C}^m$ is NP-hard.*

Proof. By rescaling, we may and do assume that $\eta < 1$. According to the previous considerations, it is enough to show that the exact cover by 3-sets problem can be reduced in polynomial time to the ℓ_0 -minimization problem. Let then $\{\mathcal{C}_i, i \in [N]\}$ be a collection of 3-element subsets of $[m]$. We define vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N \in \mathbb{C}^m$ by

$$(\mathbf{a}_i)_j = \begin{cases} 1 & \text{if } j \in \mathcal{C}_i, \\ 0 & \text{if } j \notin \mathcal{C}_i. \end{cases}$$

We then define a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and a vector $\mathbf{y} \in \mathbb{C}^m$ by

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_N \end{array} \right], \quad \mathbf{y} = [1, 1, \dots, 1]^\top.$$

Since $N \leq \binom{m}{3}$, this construction can be done in polynomial time. If a vector $\mathbf{z} \in \mathbb{C}^N$ obeys $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta$, then all the m components of the vector $\mathbf{A}\mathbf{z}$ are distant to 1 by at most η , so they are nonzero and $\|\mathbf{A}\mathbf{z}\|_0 = m$. But since each vector \mathbf{a}_i has exactly 3 nonzero components, the vector $\mathbf{A}\mathbf{z} = \sum_{j=1}^N z_j \mathbf{a}_j$ has at most $3\|\mathbf{z}\|_0$ nonzero components, $\|\mathbf{A}\mathbf{z}\|_0 \leq 3\|\mathbf{z}\|_0$. Therefore, a vector $\mathbf{z} \in \mathbb{C}^N$ obeying $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta$ must satisfy $\|\mathbf{z}\|_0 \geq m/3$. Let us now run the ℓ_0 -minimization problem, and let $\mathbf{x} \in \mathbb{C}^N$ denote the output. We separate two cases:

1. If $\|\mathbf{x}\|_0 = m/3$, then the collection $\{\mathcal{C}_j, j \in \text{supp}(\mathbf{x})\}$ forms an exact cover of $[m]$, for otherwise the m components of $\mathbf{Ax} = \sum_{j=1}^N x_j \mathbf{a}_j$ would not all be nonzero.
2. If $\|\mathbf{x}\|_0 > m/3$, then no exact cover $\{\mathcal{C}_j, j \in J\}$ can exist, for otherwise the vector $\mathbf{z} \in \mathbb{C}^N$ defined by $z_j = 1$ if $j \in J$ and $z_j = 0$ if $j \notin J$ would satisfy $\mathbf{Az} = \mathbf{y}$ and $\|\mathbf{z}\|_0 = m/3$, contradicting the ℓ_0 -minimality of \mathbf{x} .

This shows that solving the ℓ_0 -minimization problem enables one to solve the exact cover by 3-sets problem. \square

Theorem 2.17 seems rather pessimistic at first sight. However, it concerns the intractability of the problem (P_0) for general matrices \mathbf{A} and vectors \mathbf{y} . In other words, any algorithm that is able to solve (P_0) for *any* choice of \mathbf{A} and *any* choice of \mathbf{y} must necessarily be intractable (unless $P = NP$). In compressive sensing, we will rather consider special choices of \mathbf{A} and choose $\mathbf{y} = \mathbf{Ax}$ for some sparse \mathbf{x} . We will see that a variety of tractable algorithms will then provably recover \mathbf{x} from \mathbf{y} and thereby solve (P_0) for such specifically designed matrices \mathbf{A} . However, to emphasize this point once more, such algorithms will *not* successfully solve the ℓ_0 -minimization problem for *all* possible choices of \mathbf{A} and \mathbf{y} due to NP-hardness. A selection of tractable algorithms is introduced in the coming chapter.

Notes

Proposition 2.3 is an observation due to Stechkin. In the case $p = 1$ and $q = 2$, the optimal constant $c_{1,2} = 1/2$ was obtained by Gilbert, Strauss, Tropp, and Vershynin in [225]. Theorem 2.5 with optimal constants $c_{p,q}$ for all $q > p > 0$ is a particular instance of a more general result, which also contains the *shifting inequality* of Exercise 6.15; see [209].

The weak ℓ_p -spaces are weak L_p -spaces for purely atomic measures. The weak L_p -spaces are also denoted $L_{p,\infty}$ and generalize to Lorentz spaces $L_{p,q}$ [284]. Thus, weak ℓ_p -spaces are a particular instance of more general spaces equipped with the quasinorm

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{k=1}^N k^{q/p-1} (x_k^*)^q \right)^{1/q}.$$

The result of Theorem 2.16 is due to Wakin in [503]. Theorem 2.13 can be found in the article by Cohen, Dahmen, and DeVore [123]. One can also add an equivalent proposition expressed in terms of *spark* or in terms of *Kruskal rank*. The spark $\text{sp}(\mathbf{A})$ of a matrix \mathbf{A} was defined by Donoho and Elad in [155] as the minimal size of a linearly dependent set of columns of \mathbf{A} . It is related to the Kruskal rank $\text{kr}(\mathbf{A})$ of \mathbf{A} , defined in [313] as the maximal integer k such that any k columns of \mathbf{A} are

linearly independent, via $\text{sp}(\mathbf{A}) = \text{kr}(\mathbf{A}) + 1$. Thus, according to Theorem 2.13, every s -sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s -sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ if and only if $\text{kr}(\mathbf{A}) \geq 2s$ or if $\text{sp}(\mathbf{A}) > 2s$.

Totally positive matrices were extensively studied by Karlin in [298]. One can also consult the more recent book [389] by Pinkus.

The reconstruction procedure of Theorem 2.15 based on a discrete version of the Prony method was known long before the development of compressive sensing. It is also related to Reed–Solomon decoding [52, 232]. The general Prony method [402] is designed for recovering a nonharmonic Fourier series of the form

$$f(t) = \sum_{k=1}^s x_k e^{2\pi i \omega_k t}$$

from equidistant samples $f(0), f(k/\alpha), f(2k/\alpha), \dots, f(2s/\alpha)$. Here both the $\omega_k \in \mathbb{R}$ and the x_k are unknown. First the ω_k are found by solving an eigenvalue problem for a Hankel matrix associated to the samples of f . In the second step, the x_k are found by solving a linear system of equations. The difference to the method of Theorem 2.15 is due to the fact that the ω_k are not assumed to lie on a grid anymore. We refer to [344, 357] for more details. The Prony method has the disadvantage of being unstable. Several approaches have been proposed to stabilize it [14, 15, 45, 46, 401], although there seems to be a limit of how stable it can get when the number s of terms gets larger. The recovery methods in the so-called theory of *finite rate of innovation* are also related to the Prony method [55].

For an introduction to computational complexity, one can consult [19]. The NP-hardness of the ℓ_0 -minimization problem was proved by Natarajan in [359]. It was later proved by Ge, Jiang, and Ye in [220] that the ℓ_p -minimization problem is NP-hard also for any $p < 1$; see Exercise 2.10.

Exercises

2.1. For $0 < p < 1$, prove that the p th power of the ℓ_p -quasinorm satisfies the triangle inequality

$$\|\mathbf{x} + \mathbf{y}\|_p^p \leq \|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^p, \quad \mathbf{x}, \mathbf{y} \in \mathbb{C}^N.$$

For $0 < p < \infty$, deduce the inequality

$$\|\mathbf{x}_1 + \dots + \mathbf{x}_k\|_p \leq k^{\max\{0, 1/p-1\}} (\|\mathbf{x}_1\|_p + \dots + \|\mathbf{x}_k\|_p), \quad \mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{C}^N.$$

2.2. Show that the constant $k^{\max\{1, 1/p\}}$ in Proposition 2.7 is sharp.

2.3. If $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ are disjointly supported, prove that

$$\max(\|\mathbf{u}\|_{1,\infty}, \|\mathbf{v}\|_{1,\infty}) \leq \|\mathbf{u} + \mathbf{v}\|_{1,\infty} \leq \|\mathbf{u}\|_{1,\infty} + \|\mathbf{v}\|_{1,\infty}$$

and show that these inequalities are sharp.

2.4. As a converse to Proposition 2.10, prove that for any $p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,

$$\|\mathbf{x}\|_p \leq \ln(eN)^{1/p} \|\mathbf{x}\|_{p,\infty}.$$

2.5. Given $q > p > 0$ and $\mathbf{x} \in \mathbb{C}^N$, modify the proof of Proposition 2.3 to obtain

$$\sigma_s(\mathbf{x})_q \leq \frac{1}{s^{1/p-1/q}} \|\mathbf{x}\|_{p,\infty}^{1-p/q} \|\mathbf{x}\|_p^{p/q}.$$

2.6. Let $(B_0^n, B_1^n, \dots, B_n^n)$ be the *Bernstein polynomials* of degree n defined by

$$B_i^n(x) := \binom{n}{i} x^i (1-x)^{n-i}.$$

For $0 < x_0 < x_1 < \dots < x_n < 1$, prove that the matrix $[B_i^n(x_j)]_{i,j=0}^n$ is totally positive.

2.7. Prove that the product of two totally positive matrices is totally positive.

2.8. Implement the reconstruction procedure based on $2s$ discrete Fourier measurements as described in Sect. 2.2. Test it on a few random examples. Then pass from sparse to compressible vectors \mathbf{x} having small $\sigma_s(\mathbf{x})_1$ and test on perturbed measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with small $\|\mathbf{e}\|_2$.

2.9. Let us assume that the vectors $\mathbf{x} \in \mathbb{R}^N$ are no longer observed via linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$, but rather via measurements $\mathbf{y} = f(\mathbf{x})$, where $f : \mathbb{R}^N \rightarrow \mathbb{R}^m$ is a continuous map satisfying $f(-\mathbf{x}) = -f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$. Prove that the minimal number of measurements needed to reconstruct every s -sparse vector equals $2s$. You may use the *Borsuk–Ulam theorem*:

If a continuous map F from the sphere S^n —relative to an arbitrary norm—of \mathbb{R}^{n+1} into \mathbb{R}^n is antipodal, i.e.,

$$F(-\mathbf{x}) = -F(\mathbf{x}) \quad \text{for all } \mathbf{x} \in S^n,$$

then it vanishes at least once, i.e.,

$$F(\mathbf{x}) = \mathbf{0} \quad \text{for some } \mathbf{x} \in S^n.$$

2.10. *NP-Hardness of ℓ_p -minimization for $0 < p < 1$*

Given $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\mathbf{y} \in \mathbb{C}^m$, the ℓ_p -minimization problem consists in computing a vector $\mathbf{x} \in \mathbb{C}^N$ with minimal ℓ_p -quasinorm subject to $\mathbf{A}\mathbf{x} = \mathbf{y}$.

The *partition problem* consists, given integers a_1, \dots, a_n , in deciding whether there exist two sets $I, J \subset [n]$ such that $I \cap J = \emptyset$, $I \cup J = [n]$, and $\sum_{i \in I} a_i = \sum_{j \in J} a_j$. Assuming the NP-completeness of the partition problem, prove that the ℓ_p -minimization problem is NP-hard. It will be helpful to introduce the matrix \mathbf{A} and the vector \mathbf{y} defined by

$$\mathbf{A} := \begin{bmatrix} a_1 & a_2 & \cdots & a_n & -a_1 & -a_2 & \cdots & -a_n \\ 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = [0, 1, 1, \dots, 1]^\top.$$

2.11. NP-Hardness of rank minimization

Show that the rank-minimization problem

$$\underset{\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}}{\text{minimize}} \quad \text{rank}(\mathbf{Z}) \quad \text{subject to} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}.$$

is NP-hard on the set of linear measurement maps $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \rightarrow \mathbb{C}^m$ and vectors $\mathbf{y} \in \mathbb{C}^m$.

A Mathematical Introduction to Compressive Sensing

Foucart, S.; Rauhut, H.

2013, XVIII, 625 p., Hardcover

ISBN: 978-0-8176-4947-0

A product of Birkhäuser Basel