Lecture Notes (courses 1 & 2)

“Sparsity and High Dimensions”

Y. De Castro & R. Gribonval
https://statistical-learning-centrale.github.io/

draft of January 27, 2026

Contents

1 Introduction: The Geometry of Underdetermined Systemss **2**
1.1 Sparsity 2
1.2 Compressibility 3

2 The ℓ_0-Minimization Problem (P_0) **3**
2.1 The Null Space Property and Uniqueness (NSP_0) 3
2.2 Fundamental Measurement Bounds 4

3 Computational Complexity of Sparse Recovery **5**
3.1 P, NP, and NP-Hardness 5
3.2 Complexity of Sparse Recovery 6

4 Coherence and Greedy Algorithms **6**
4.1 Orthogonal Matching Pursuit (OMP) 7
4.2 The Exact Recovery Condition 8
4.3 Mutual Coherence and the Welch Bound 9

5 Basis Pursuit **10**
5.1 The Null Space Property (NSP_1) 10
5.2 Stable Sparse Recovery 11
5.3 Robustness 13

1 Introduction: The Geometry of Underdetermined Systemss

The classical paradigm of signal processing, governed largely by the Shannon-Nyquist sampling theorem, dictates that a signal must be sampled at a rate at least twice its highest frequency to be perfectly reconstructed. This principle has underpinned decades of digital data acquisition. However, the emerging theory of Compressive Sensing (CS) challenges this orthodoxy, asserting that if a signal exhibits a sparse structure—meaning it has few non-zero coefficients in some basis—it can be recovered from a number of measurements significantly smaller than its ambient dimension. Master 2 lecture notes on "Sparsity and High Dimensions" rigorously establishes the mathematical frameworks that permit such recovery.

In many areas of applied mathematics and engineering, we face the challenge of recovering a signal of interest $x \in \mathbb{C}^N$ from a set of linear measurements $y \in \mathbb{C}^m$. This process is modeled by the linear system:

$$Ax = y, \tag{1}$$

where $A \in \mathbb{C}^{m \times N}$ is the *measurement matrix* (or sensing matrix). Classical linear algebra tells us that to recover x uniquely, we generally need at least as many measurements as variables, i.e., $m \geq N$. However, in the context of **Compressive Sensing**, we operate in the regime of underdetermined linear systems:

$$m \ll N.$$

In this setting, the system (1) is *underdetermined*. The operator A has a non-trivial null space (kernel), denoted by $\ker A$. If x is a solution to $Ax = y$, then for any $v \in \ker A$, the vector $x' = x + v$ is also a solution, since

$$A(x + v) = Ax + Av = y + 0 = y.$$

Consequently, without additional information, it is impossible to distinguish the true signal x from the infinitely many other solutions $x + v$. Standard least squares solutions (minimizing the ℓ_2-norm) typically yield non-sparse results that do not recover the original signal of interest. To resolve this ambiguity, we exploit the sparsity of the signal. The central hypothesis of CS is that the true signal x resides on a low-dimensional union of subspaces within $\mathbb{C}^N$. Specifically, we assume x has at most s non-zero entries. This structural assumption regularizes the inverse problem, transforming it from an impossible algebraic task into a solvable combinatorial or geometric optimization problem. The core objectives of this report are to determine the conditions on the matrix A that guarantee unique recovery, to analyze the computational complexity of such recovery, and to develop tractable algorithms—specifically *Greedy methods* and *Convex Relaxations*—that achieve recovery guarantees.

Remark 1.1. *The central question of this course is:* Under what conditions can we recover x uniquely from y despite the system being underdetermined? *The answer relies on the concept of* **sparsity**.

1.1 Sparsity

To resolve the ill-posedness of the underdetermined system, we assume prior knowledge about the structure of the signal x. The fundamental assumption in Compressive Sensing is that x is *sparse*.

Definition 1.2 (Support)**.** *The* **support** *of a vector $x \in \mathbb{C}^N$ is the index set of its non-zero entries:*

$$\operatorname{supp}(x) := \{j \in \{1, \ldots, N\} : x_j \neq 0\}.$$

Definition 1.3 (ℓ_0-"norm")**.** *The number of non-zero entries of x is denoted by $\|x\|_0$:*

$$\|x\|_0 := \mathrm{card}(\mathrm{supp}(x)).$$

Note: Although called the ℓ_0-norm, this functional is not a norm (it is not homogeneous). It satisfies the triangle inequality $\|x+z\|_0 \leq \|x\|_0 + \|z\|_0$.

Definition 1.4 (Sparsity)**.** *A vector $x \in \mathbb{C}^N$ is called s-**sparse** if it has at most s non-zero entries, i.e.,*

$$\|x\|_0 \leq s.$$

The set of all s-sparse vectors is denoted by Σ_s:

$$\Sigma_s := \{x \in \mathbb{C}^N : \|x\|_0 \leq s\}.$$

Geometrically, Σ_s is a union of $\binom{N}{s}$ subspaces of dimension s aligned with the canonical axes of $\mathbb{C}^N$. This non-convex structure is the root of the complexity issues discussed later.

1.2 Compressibility

In practical applications, signals are rarely exactly sparse. Instead, they are compressible, meaning their sorted coefficients decay rapidly (e.g., obeying a power law). To quantify the deviation from exact sparsity, we introduce the error of best s-term approximation.

Definition 1.5 (Best s-term approximation error)**.** *For $x \in \mathbb{C}^N$ and $p > 0$, the ℓ_p-error of* **best** *s-term* **approximation** *is:*

$$\sigma_s(x)_p := \inf_{z \in \Sigma_s} \|x - z\|_p .$$

This infimum is achieved by a vector, denoted by x_{σ_s}, constructed by retaining the s largest coefficients of x in absolute value and setting the rest to zero. If $\sigma_s(x)_p$ is small relative to $\|x\|_p$, the signal is compressible. Recovery guarantees must ideally account for this "tail" energy, ensuring that the reconstruction error is proportional to $\sigma_s(x)_p$.

2 The ℓ_0-Minimization Problem (P_0)

If we assume the original signal x is sparse, the natural approach to recover x from $y = Ax$ is to search for the sparsest vector consistent with the measurements. This leads to the formulation of the problem (P_0):

$$(P_0) \quad \min_{z \in \mathbb{C}^N} \|z\|_0 \quad \text{subject to} \quad Az = y. \tag{2}$$

A crucial theoretical question is whether the solution to (P_0) corresponds to the original signal. We first establish a relationship between unique recovery and the properties of the matrix A.

2.1 The Null Space Property and Uniqueness (NSP_0)

We seek a condition on A ensuring that every s-sparse vector is the unique solution to (P_0). This condition is intimately related to the null space of A.

Theorem 2.1 (Uniqueness and NSP_0)**.** *Given a matrix $A \in \mathbb{C}^{m \times N}$ and an integer $s \geq 1$, the following properties are equivalent:*

(a) Every s-sparse vector $x \in \mathbb{C}^N$ is the unique s-sparse solution of $Az = Ax$.

(b) The null space of A contains no $2s$-sparse vector other than the zero vector:

$$\ker A \cap \Sigma_{2s} = \{0\}.$$

(c) Every set of $2s$ columns of A is linearly independent.

Proof. **(b) $\Rightarrow$ (a):** Assume (b) holds. Let x and z be two s-sparse vectors such that $Ax = Az$. Then $A(x - z) = 0$, so $v = x - z \in \ker A$. Since $x, z \in \Sigma_s$, the vector v satisfies:

$$\|v\|_0 = \|x - z\|_0 \leq \|x\|_0 + \|z\|_0 \leq s + s = 2s.$$

Thus $v \in \ker A \cap \Sigma_{2s}$. By condition (b), we must have $v = 0$, which implies $x = z$.

(a) $\Rightarrow$ (b): Assume (a) holds. Let $v \in \ker A$ be a vector with $\|v\|_0 \leq 2s$. We can decompose v as $v = x - z$ where x and z are s-sparse vectors with disjoint supports (e.g., let $S = \mathrm{supp}(v)$, partition S into $S_1 \cup S_2$ with $|S_1| \leq s, |S_2| \leq s$, and set $x = v_{S_1}, z = -v_{S_2}$). Since $v \in \ker A$, we have $0 = Av = A(x - z) \Rightarrow Ax = Az$. Since both x and z are s-sparse, assumption (a) implies $x = z$. Since they have disjoint supports, this forces $x = 0$ and $z = 0$, hence $v = 0$.

(b) $\Leftrightarrow$ (c): A vector $v \in \Sigma_{2s}$ is in $\ker A$ if and only if $Av = \sum_{j \in \mathrm{supp}(v)} v_j A_j = 0$. This is a linear dependency among at most $2s$ columns of A. If no non-trivial $2s$-sparse vector exists in the kernel, no set of $2s$ columns can be linearly dependent, and vice-versa. □

Remark 2.2 (Spark)**.** *The* **spark** *of a matrix A, denoted* $\mathrm{spark}(A)$*, is defined as the smallest number of columns of A that are linearly dependent.*

$$\mathrm{spark}(A) := \min\{\|v\|_0 : v \in \ker A \setminus \{0\}\}.$$

Condition (b) in Theorem 2.1 can be restated as:

$$\mathrm{spark}(A) > 2s.$$

Unlike the rank, which is easily computable (e.g., via SVD), calculating the spark is combinatorially difficult, as it requires checking all subsets of columns.

2.2 Fundamental Measurement Bounds

From Theorem 2.1, for unique recovery of any s-sparse vector, we need $\mathrm{spark}(A) > 2s$. Since $\mathrm{spark}(A) \leq \mathrm{rank}(A) + 1 \leq m + 1$, we derive the fundamental lower bound on the number of measurements:

$$m \geq 2s.$$

This bound is tight. Matrices (like Vandermonde matrices) exist where $m = 2s$ suffices for uniqueness, though these are often unstable numerically.

Theorem 2.3 (Minimal Measurements)**.** *For any integer $N \geq 2s$, there exists a measurement matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ rows such that every s-sparse vector $x \in \mathbb{C}^N$ can be recovered uniquely from $y = Ax$.*

Proof. Consider distinct points $t_1, t_2, \dots, t_N \in \mathbb{C}$ and form the Vandermonde matrix $A \in \mathbb{C}^{2s \times N}$ defined by:

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_N \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \dots & t_N^{2s-1} \end{bmatrix}.$$

Let $S \subset \{1, \dots, N\}$ be any index set of cardinality $2s$. The submatrix A_S consisting of the columns indexed by S is a square $2s \times 2s$ Vandermonde matrix. Its determinant is given by:

$$\det(A_S) = \prod_{j,k \in S, j>k} (t_j - t_k).$$

Since the points t_j are distinct, $\det(A_S) \neq 0$, and thus any set of $2s$ columns is linearly independent. By Theorem 2.1 (condition c), this matrix ensures unique recovery of any s-sparse vector. □

While this establishes that $m = 2s$ is theoretically sufficient, Vandermonde matrices are notoriously ill-conditioned, making them unstable for practical recovery in the presence of noise. This motivates the search for matrices with better stability properties, such as those satisfying the Restricted Isometry Property.

3 Computational Complexity of Sparse Recovery

3.1 P, NP, and NP-Hardness

We have established that unique recovery is possible via (P_0). However, computationally solving (P_0) is a distinct challenge. The feasible set is non-convex, and a brute-force approach would require checking $\binom{N}{s}$ subspaces, which is exponentially large.

Definition 3.1 (Class P)**.** *The class* **P** *(Polynomial time) consists of decision problems that can be* solved *by a deterministic Turing machine in polynomial time. Roughly speaking, there exists an algorithm that solves the problem in $O(N^k)$ operations for some constant k.*

Definition 3.2 (Class NP)**.** *The class* **NP** *(Nondeterministic Polynomial time) consists of decision problems for which a given candidate solution can be* verified *in polynomial time.*

It is immediate that P $\subseteq$ NP, since if we can find a solution in polynomial time, we can also verify it in polynomial time. The converse is the famous open problem: is P $=$ NP? The general consensus is that P $\neq$ NP.

To define NP-hardness, we need the concept of *reduction*. A problem A is polynomial-time reducible to problem B (denoted $A \leq_p B$) if an algorithm for solving B can be used to solve A with a polynomial overhead.

Definition 3.3 (NP-Hard)**.** *A problem H is called* **NP-hard** *if every problem L in* **NP** *can be polynomial-time reducible to H ($L \leq_p H$).*

Intuitively, an NP-hard problem is at least as hard as the hardest problems in NP. Note that an NP-hard problem does not necessarily have to be in NP itself (it might not even be a decision problem). To prove that a problem is NP-hard it is sufficient to show that a **NP-Complete** decision problem is polynomial-time reducible to it, the definition is given below.

Definition 3.4 (NP-Complete)**.** *A problem C is called* **NP-complete** *if:*

1. *$C \in NP$, and*
2. *C is NP-hard.*

3.2 Complexity of Sparse Recovery

While (P_0) provides the ideal theoretical solution, solving it numerically is computationally intractable.

Theorem 3.5 (NP-Hardness)**.** *For a general matrix A and vector y, the problem (P_0) is NP-hard.*

Proof. We rely on a reduction from the **Exact Cover by 3-Sets** (X3C) problem, which is known to be NP-complete.
The X3C Problem: Given a finite set $U = \{1, \dots, m\}$ and a collection of subsets $\mathcal{C} = \{C_1, \dots, C_N\}$ where each $C_j \subset U$ has exactly $|C_j| = 3$, does there exist an exact cover? That is, a sub-collection of indices $J \subset \{1, \dots, N\}$ such that $\bigcup_{j \in J} C_j = U$ and the sets are disjoint. Note that if an exact cover exists, since $|U| = m$ and each set has size 3, we must have $|J| = m/3$.
Reduction to (P_0): We construct a matrix $A \in \mathbb{R}^{m \times N}$ and a vector $y \in \mathbb{R}^m$ as follows:

- Let the columns of A, denoted $a_1, \dots, a_N$, represent the sets in $\mathcal{C}$. Specifically, set $(a_j)_i = 1$ if $i \in C_j$ and 0 otherwise.
- Let $y = (1, 1, \dots, 1)^T \in \mathbb{R}^m$.

Now consider the problem: find x such that $Ax = y$ with minimal $\|x\|_0$.

Suppose an exact cover J exists. Let x be the indicator vector of J ($x_j = 1$ if $j \in J$, 0 else). Then $Ax = \sum_{j \in J} a_j$. Since J is an exact cover, for every row index $i \in U$, exactly one C_j contains i, so $(Ax)_i = 1 = y_i$. Thus $Ax = y$. The sparsity is $\|x\|_0 = |J| = m/3$.

Conversely, suppose there exists a solution x to $Ax = y$ with $\|x\|_0 \leq m/3$. Since A and y have non-negative entries, and the target sum is 1 for each row, one can argue (in the standard binary version of the problem) that the non-zero x_j must be 1. Each column a_j has weight 3. So the total weight of Ax is $\sum x_j \|a_j\|_1 = 3 \sum x_j$. Also, the weight of y is m. Thus $3 \sum x_j = m$, which implies $\sum x_j = m/3$. If we relax to arbitrary x, any solution with $\|x\|_0 = k$ covers at most $3k$ elements. To cover m elements (entries of y), we strictly need $3k \geq m$, so $k \geq m/3$. Therefore, the minimal sparsity is exactly $m/3$. Finding the solution to (P_0) would determine if such a cover exists (sparsity $m/3$) or not (sparsity $> m/3$). Since X3C is NP-complete, solving (P_0) is NP-hard. $\square$

4 Coherence and Greedy Algorithms

Since solving (P_0) is intractable, we must look for computationally efficient alternatives. There are two main categories of tractable algorithms:

1. **Convex Relaxation:** Replacing ℓ_0 with ℓ_1 (Basis Pursuit), which will be covered in the next sections.
2. **Greedy Algorithms:** Iterative methods that select columns of A one by one.

The intuition behind greedy algorithms is to build the support of the signal iteratively. At each step, we select the column of A that correlates most strongly with the current residual.

4.1 Orthogonal Matching Pursuit (OMP)

The most fundamental greedy algorithm is Orthogonal Matching Pursuit.

Algorithm 1 Orthogonal Matching Pursuit (OMP)

Require: Matrix A, measurement vector y, sparsity level s.

Initialize: $r^0 = y$ (residual), $S^0 = \emptyset$ (support), $x^0 = 0$.

for $k = 1$ to s **do**

  Selection: Find index j_k maximizing the correlation with the residual:

$$j_k = \arg\max_j \left|\langle a_j, r^{k-1}\rangle\right|$$

  Update Support: $S^k = S^{k-1} \cup \{j_k\}$

  Projection: Compute x^k by minimizing the error over the chosen support:

$$x^k = \arg\min_{z:\mathrm{supp}(z)\subseteq S^k} \|y - Az\|_2$$

  Update Residual: $r^k = y - Ax^k$

end for

Output: x^s

OMP is computationally efficient (involving only matrix-vector products and small least-squares problems). A key question is: *Under what conditions on A does OMP recover the true s-sparse vector?*

Proposition 4.1 (Convergence of Orthogonal Matching Pursuit)**.** *Given a matrix $A \in \mathbb{C}^{m\times N}$, every nonzero vector $x \in \mathbb{C}^N$ supported on a set S of size s is recovered from $y = Ax$ after at most s iterations of Orthogonal Matching Pursuit (OMP)* if and only if *the matrix A_S is injective and*

$$\max_{j\in S} |(A^* r)_j| > \max_{\ell\in \overline{S}} |(A^* r)_\ell| \tag{3}$$

for all nonzero $r \in \{Az : \mathrm{supp}(z) \subset S\}$.

Proof. **Necessity:** Assume OMP recovers all vectors supported on S in at most s iterations. First, if two distinct vectors x, z supported on S satisfied $Ax = Az$, recovery would be impossible for at least one of them; thus A_S must be injective. Second, consider the first iteration with $y = Ax$ where $\mathrm{supp}(x) = S$. The algorithm selects the index j_1 maximizing $|(A^*y)_j|$. For the algorithm to select an index within the true support (which is required for correct recovery in s steps), we must have $\max_{j\in S} |(A^*y)_j| > \max_{\ell\in\overline{S}} |(A^*y)_\ell|$. Since y can be any vector in $\{Az : \mathrm{supp}(z) \subset S\}$, the condition (3) must hold for all such nonzero vectors.

Sufficiency: Assume the conditions hold. We prove by induction that the support set S^n selected by OMP after n iterations is a subset of S with cardinality n for all $0 \le n \le s$.

1. Base case ($n = 0$): $S^0 = \emptyset \subset S$ is trivial.

2. Inductive step: Suppose $S^n \subset S$ and $|S^n| = n$ for $n < s$. The residual is $r^n = y - Ax^n$. Since $y \in \mathrm{range}(A_S)$ and $S^n \subset S$, we have $r^n \in \{Az : \mathrm{supp}(z) \subset S\}$. If $r^n = 0$, we are done. If $r^n \neq 0$, by condition (3), the index j_{n+1} maximizing the correlation lies in S. Thus $S^{n+1} \subset S$.

Furthermore, by the properties of the OMP projection step, the residual r^n is orthogonal to the columns of A indexed by S^n, meaning $(A^* r^n)_j = 0$ for all $j \in S^n$. Since $r^n \neq 0$ and A_S is injective, the maximum correlation on S is strictly positive. Consequently, the selected index j_{n+1} cannot be in S^n. Therefore, $S^{n+1} = S^n \cup \{j_{n+1}\}$ has cardinality $n+1$.

After s iterations, we have $S^s \subset S$ with $|S^s| = s$, implying $S^s = S$. The projection step then yields $x^s = x$ due to the injectivity of A_S. □

4.2 The Exact Recovery Condition

Let $A \in \mathbb{C}^{m \times N}$ be a measurement matrix and let $S \subset \{1, \ldots, N\}$ be a support set of cardinality s. Let A_S denote the submatrix of A consisting of the columns indexed by S, and let $A_{\overline{S}}$ denote the submatrix consisting of the remaining columns. The **Exact Recovery Condition** (ERC) with respect to the set S is defined as:

$$\|A_S^\dagger A_{\overline{S}}\|_{1\to1} < 1 \tag{4a}$$

where $A_S^\dagger = (A_S^* A_S)^{-1} A_S^*$ is the Moore-Penrose pseudo-inverse of A_S (assuming A_S has full column rank), and $\|\cdot\|_{1\to1}$ denotes the operator norm induced by the ℓ_1-norm, which corresponds to the maximum absolute column sum of the matrix. Specifically, the condition requires that:

$$\max_{j \notin S} \|A_S^\dagger a_j\|_1 < 1 \tag{4b}$$

where a_j represents the j-th column of A. If this condition holds, OMP is guaranteed to recover any s-sparse signal supported on S in exactly s steps.

Lemma 4.2 (Equivalence of ERC and OMP Condition)**.** *Let $A \in \mathbb{C}^{m \times N}$ and let $S \subset \{1, \ldots, N\}$ be a support set such that A_S is injective. The condition*

$$\max_{j \in S} |(A^* r)_j| > \max_{\ell \in \overline{S}} |(A^* r)_\ell| \quad \text{for all } r \in \text{range}(A_S) \setminus \{0\} \tag{3}$$

holds if and only if the Exact Recovery Condition (ERC) is satisfied:

$$\|A_S^\dagger A_{\overline{S}}\|_{1\to1} < 1\,. \tag{4}$$

Proof. Let r be an arbitrary vector in range(A_S). We can write $r = A_S x$ for some unique vector $x \in \mathbb{C}^s$ (since A_S is injective). Let us define $z := A_S^* r = A_S^* A_S x$. Since the Gram matrix $A_S^* A_S$ is invertible, the mapping between x and z is a bijection, and z can be any vector in $\mathbb{C}^s$. We can express x as $x = (A_S^* A_S)^{-1} z$.

We examine the term on the right-hand side of condition (3). The correlations with the indices outside the support are given by:

$$(A^* r)_{\overline{S}} = A_{\overline{S}}^* r = A_{\overline{S}}^* A_S x = A_{\overline{S}}^* A_S (A_S^* A_S)^{-1} z = \left((A_S^* A_S)^{-1} A_S^* A_{\overline{S}}\right)^* z = (A_S^\dagger A_{\overline{S}})^* z.$$

The condition (3) requires that for all $z \neq 0$:

$$\|(A^* r)_{\overline{S}}\|_\infty < \|(A^* r)_S\|_\infty \iff \|(A_S^\dagger A_{\overline{S}})^* z\|_\infty < \|z\|_\infty.$$

This inequality holds for all non-zero z if and only if the operator norm of the matrix $(A_S^\dagger A_{\overline{S}})^*$ induced by the ℓ_∞-norm is strictly less than 1. Recall that the operator norm $\|M\|_{\infty\to\infty}$ is the maximum absolute row sum of M, while $\|M\|_{1\to1}$ is the maximum absolute column sum. Therefore, $\|M^*\|_{\infty\to\infty} = \|M\|_{1\to1}$. Consequently, the condition is equivalent to:

$$\|(A_S^\dagger A_{\overline{S}})^*\|_{\infty\to\infty} = \|A_S^\dagger A_{\overline{S}}\|_{1\to1} < 1.$$

□

4.3 Mutual Coherence and the Welch Bound

To enable practical recovery, we must impose stronger conditions on A than just "full spark" or ERC. We need conditions that are checkable and ensure the success of efficient algorithms. One such metric is Mutual Coherence.

Definition 4.3 (Mutual Coherence)**.** *Let A have ℓ_2-normalized columns $a_1, \ldots, a_N$. The mutual coherence $\mu(A)$ is the maximum absolute inner product between distinct columns :*

$$\mu(A) := \max_{1 \leq i \neq j \leq N} |\langle a_i, a_j \rangle|.$$

A small $\mu(A)$ implies the columns are nearly orthogonal (incoherent). For an orthonormal basis ($N = m$), $\mu(A) = 0$. For overcomplete frames ($N > m$), $\mu(A) > 0$.

There is a fundamental limit to how incoherent a redundant dictionary can be, given by the *Welch Bound* proven in the next proposition.

Proposition 4.4 (Welch Bound)**.** *For any matrix $A \in \mathbb{C}^{m \times N}$ with unit-norm columns and $N > m$:*

$$\mu(A) \geq \sqrt{\frac{N-m}{m(N-1)}}.$$

For large $N \gg m$, this bound essentially scales as $\mu(A) \gtrsim 1/\sqrt{m}$.

Proof. Let $G = A^*A$ be the Gram matrix. The diagonal entries are $G_{ii} = \|a_i\|_2^2 = 1$. The off-diagonal entries are bounded by μ. We compute the Frobenius norm squared of G in two ways. First, summing element-wise:

$$\|G\|_F^2 = \sum_i |G_{ii}|^2 + \sum_{i \neq j} |G_{ij}|^2 = N + \sum_{i \neq j} |\langle a_i, a_j \rangle|^2 \leq N + N(N-1)\mu^2.$$

Second, using eigenvalues λ_k of G. Since A has rank at most m, G has at most m non-zero eigenvalues. Also, $\mathrm{tr}(G) = \sum \lambda_k = \sum G_{ii} = N$. Minimizing $\sum \lambda_k^2$ subject to $\sum \lambda_k = N$ with m non-zero values occurs when $\lambda_k = N/m$. Thus, $\|G\|_F^2 = \sum \lambda_k^2 \geq m(N/m)^2 = N^2/m$. Combining inequalities:

$$\frac{N^2}{m} \leq N + N(N-1)\mu^2 \implies \frac{N}{m} - 1 \leq (N-1)\mu^2 \implies \mu \geq \sqrt{\frac{N-m}{m(N-1)}}.$$

□

We can derive a sufficient condition for OMP to recover the correct support based on coherence.

Theorem 4.5 (OMP Recovery Condition)**.** *OMP recovers any s-sparse signal x if the coherence satisfies:*

$$\mu(A) < \frac{1}{2s-1}.$$

This condition also ensures that the solution to (P_0) is unique.

Proof. Sketch. Suppose we are at the first step. The signal is $x = \sum_{j \in S} x_j a_j$. The correlation with a correct column a_k ($k \in S$) is $\langle a_k, y \rangle = x_k + \sum_{j \neq k} x_j \langle a_k, a_j \rangle$. The magnitude is lower bounded: $|\langle a_k, y \rangle| \geq |x_k| - (s-1)|x_{\max}|\mu$. The correlation with an incorrect column a_ℓ ($\ell \notin S$) is upper bounded: $|\langle a_\ell, y \rangle| = |\sum_{j \in S} x_j \langle a_\ell, a_j \rangle| \leq s|x_{\max}|\mu$. To ensure a correct column is picked, we need the lower bound of the "good" correlation to exceed the upper bound of the "bad" correlation. A precise worst-case analysis (see Tropp for instance) yields the condition $\mu < 1/(2s-1)$. □

While simple, the coherence condition $\mu < 1/(2s-1)$ combined with the Welch bound $\mu \geq 1/\sqrt{m}$ implies that we need $\sqrt{m} > 2s - 1$, or $m \sim O(s^2)$. This quadratic scaling is suboptimal compared to the information-theoretic limit $m \sim O(s)$. To bridge this gap, we turn to the Restricted Isometry Property (RIP) later in this course.

5 Basis Pursuit

We consider the recovery of a signal $x \in \mathbb{C}^N$ from linear measurements $y = Ax \in \mathbb{C}^m$ with $m < N$. Since the system is underdetermined, we seek the sparsest solution. The natural approach, ℓ_0-minimization, is NP-hard. Therefore, we relax the problem to ℓ_1-minimization, known as **Basis Pursuit**:

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \quad \text{subject to} \quad Az = y. \tag{P_1}$$

The geometry of this relaxation is illustrated below. The ℓ_1-ball is a polytope (cross-polytope), which tends to intersect the affine subspace $\{z : Az = y\}$ at vertices (sparse vectors), whereas the ℓ_2-ball touches the subspace at a point that is generally not sparse.

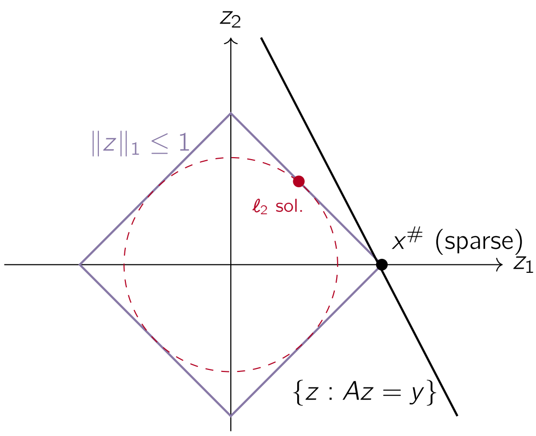

Figure 1: Geometric intuition for Basis Pursuit. The ℓ_1-ball contacts the feasible set (black line) at a coordinate axis, yielding a sparse solution. The ℓ_2-ball (dashed) contacts the line at a non-sparse point.

5.1 The Null Space Property (NSP$_1$)

A necessary and sufficient condition for the success of Basis Pursuit is the ℓ_1-Null Space Property (NSP$_1$). Intuitively, for x to be the unique minimizer, no perturbation v in the null space of A should allow us to decrease the ℓ_1-norm.

Definition 5.1 (ℓ_1-Null Space Property)**.** *A matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the ℓ_1-Null Space Property of order s* (NSP$_1$) *if for every set $S \subset \{1, \ldots, N\}$ with $|S| \leq s$, and for every nonzero vector $v \in \ker A$,*

$$\|v_S\|_1 < \|v_{\overline{S}}\|_1.$$

Remark 5.2. *The condition implies that vectors in the null space are not "concentrated" on any small support set S. In fact, at least half of their ℓ_1-mass must reside outside any set of size s.*

Remark 5.3. *It is clear that* (NSP_1) *implies* (NSP_0).

We now state and prove the fundamental characterization of exact recovery for Basis Pursuit.

Theorem 5.4 (Exact Recovery via Basis Pursuit)**.** *Given a matrix $A \in \mathbb{C}^{m\times N}$, every s-sparse vector $x \in \mathbb{C}^N$ is the unique solution of (P_1) with $y = Ax$ if and only if A satisfies the ℓ_1-Null Space Property of order s.*

Proof. **Sufficiency:** Assume A satisfies the NSP_1. Let x be an s-sparse vector supported on S, and let $y = Ax$. Let $z \in \mathbb{C}^N$ be any other feasible vector ($Az = y$) with $z \neq x$. The difference $v = z - x$ is in $\ker A \setminus \{0\}$. We compare their norms:

$$\begin{aligned}
\|z\|_1 &= \|x + v\|_1 \\
&= \|x_S + v_S\|_1 + \|v_{\overline{S}}\|_1 && (\text{since } x_{\overline{S}} = 0) \\
&\geq \|x_S\|_1 - \|v_S\|_1 + \|v_{\overline{S}}\|_1 && (\text{reverse triangle inequality}) \\
&= \|x\|_1 + (\|v_{\overline{S}}\|_1 - \|v_S\|_1).
\end{aligned}$$

By the NSP_1, $\|v_{\overline{S}}\|_1 - \|v_S\|_1 > 0$, implying $\|z\|_1 > \|x\|_1$. Thus, x is the unique minimizer.

Necessity: Assume every s-sparse vector is uniquely recovered. Suppose for contradiction that NSP_1 fails. Then there exists a set S with $|S| \leq s$ and a nonzero $v \in \ker A$ such that $\|v_S\|_1 \geq \|v_{\overline{S}}\|_1$. Define $x = v_S$. Then x is s-sparse. Let $z = -v_{\overline{S}}$. Note that $x - z = v \in \ker A$, so $Ax = Az = y$. However,

$$\|z\|_1 = \| - v_{\overline{S}}\|_1 = \|v_{\overline{S}}\|_1 \leq \|v_S\|_1 = \|x\|_1.$$

Thus, z is a solution to $Az = y$ with ℓ_1-norm no strictly greater than x. If $\|v_{\overline{S}}\|_1 < \|v_S\|_1$, x is not a minimizer. If equality holds, x is not the *unique* minimizer. Both contradict the assumption. ☐

Remark 5.5. *While* (NSP_1) *guarantees exact recovery for sparse vectors, it does not guarantee stability (recovery of approximately sparse vectors) or robustness (recovery under noise). For this, we require a stronger condition.*

5.2 Stable Sparse Recovery

To ensure that the recovery is stable with respect to the sparsity defect $\sigma_s(x)_1$, we introduce the Stable ℓ_1-Null Space Property.

Definition 5.6 (Stable ℓ_1-Null Space Property)**.** *A matrix $A \in \mathbb{C}^{m\times N}$ satisfies the Stable ℓ_1-Null Space Property of order s* (SNSP_1) *with constant* $0 < \rho < 1$ *if for every set $S \subset \{1, \ldots, N\}$ with $|S| \leq s$ and every $v \in \ker A$,*

$$\|v_S\|_1 \leq \rho\|v_{\overline{S}}\|_1.$$

It turns out that the strict inequality in the standard NSP implies the stable version with some $\rho < 1$, essentially due to the compactness of the unit sphere in finite dimensions.

Lemma 5.7 (Equivalence of NSP_1 and SNSP_1)**.** *The ℓ_1-Null Space Property of order s holds if and only if the Stable ℓ_1-Null Space Property of order s holds for some* $0 < \rho < 1$.

Proof. The implication (SNSP_1) $\Rightarrow$ (NSP_1) is trivial (since $\rho < 1$).
We prove the converse. Let $\mathcal{N} = \ker A \setminus \{0\}$. Consider the function

$$f(v) = \max_{|S|\leq s} \frac{\|v_S\|_1}{\|v\|_1}.$$

Assuming NSP holds, for any $v \in \mathcal{N}$ and any $|S| \leq s$, we have $\|v_S\|_1 < \|v_{\overline{S}}\|_1$. Adding $\|v_S\|_1$ to both sides yields $2\|v_S\|_1 < \|v\|_1$, or $\frac{\|v_S\|_1}{\|v\|_1} < \frac{1}{2}$. Thus, $f(v) < 1/2$ for all $v \in \mathcal{N}$. Now, consider the compact set $K = \{v \in \ker A : \|v\|_1 = 1\}$. Since f is continuous (being a maximum of continuous functions) and K is compact, f attains its maximum on K. Let $\gamma = \max_{v\in K} f(v)$. By NSP, $\gamma < 1/2$. Since f is scale-invariant, $f(v) \leq \gamma$ for all $v \in \ker A$. This implies $\|v_S\|_1 \leq \gamma\|v\|_1 = \gamma(\|v_S\|_1 + \|v_{\overline{S}}\|_1)$. Rearranging, we get $(1-\gamma)\|v_S\|_1 \leq \gamma\|v_{\overline{S}}\|_1$, or

$$\|v_S\|_1 \leq \frac{\gamma}{1-\gamma}\|v_{\overline{S}}\|_1.$$

Since $\gamma < 1/2$, we have $\rho := \frac{\gamma}{1-\gamma} < 1$. Thus SNSP holds. ☐

We now prove the main stability result below.

Theorem 5.8 (Stable Recovery)**.** *Suppose A satisfies the Stable ℓ_1-Null Space Property of order s with constant $0 < \rho < 1$. Let $x \in \mathbb{C}^N$ and let $x^{\#}$ be the solution to (P_1) with $y = Ax$. Then*

$$\|x - x^{\#}\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1,$$

where $\sigma_s(x)_1 = \inf\{\|x - z\|_1 : z \text{ is } s\text{-sparse}\}$ is the best s-term approximation error.

Proof. Let $v = x - x^{\#}$. Note that $v \in \ker A$. Let S be the support of the s largest entries of x (in absolute value), so $\sigma_s(x)_1 = \|x_{\overline{S}}\|_1$. Since $x^{\#}$ is the minimizer, $\|x^{\#}\|_1 \leq \|x\|_1$. Thus, $\|x + v\|_1 \leq \|x\|_1$. We decompose the norms over S and $\overline{S}$:

$$\begin{aligned}\|x\|_1 \geq \|x+v\|_1 &= \|(x_S + v_S) + (x_{\overline{S}} + v_{\overline{S}})\|_1 \\ &= \|x_S + v_S\|_1 + \|x_{\overline{S}} + v_{\overline{S}}\|_1 \\ &\geq \|x_S\|_1 - \|v_S\|_1 + \|v_{\overline{S}}\|_1 - \|x_{\overline{S}}\|_1.\end{aligned}$$

Using $\|x\|_1 = \|x_S\|_1 + \|x_{\overline{S}}\|_1$ and rearranging:

$$\|v_{\overline{S}}\|_1 - \|v_S\|_1 \leq 2\|x_{\overline{S}}\|_1 = 2\sigma_s(x)_1.$$

Using the SNSP condition $\|v_S\|_1 \leq \rho\|v_{\overline{S}}\|_1$, we substitute $\|v_S\|_1$:

$$\|v_{\overline{S}}\|_1 - \rho\|v_{\overline{S}}\|_1 \leq 2\sigma_s(x)_1 \implies \|v_{\overline{S}}\|_1 \leq \frac{2}{1-\rho}\sigma_s(x)_1.$$

Finally, the total error is:

$$\|x - x^{\#}\|_1 = \|v\|_1 = \|v_S\|_1 + \|v_{\overline{S}}\|_1 \leq (1+\rho)\|v_{\overline{S}}\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1.$$

☐

Remark 5.9. *This theorem confirms that Basis Pursuit is not only exact for s-sparse vectors (where $\sigma_s(x)_1 = 0$) but also stable: if x is close to being sparse, the recovered vector $x^{\#}$ is close to x.*

5.3 Robustness

So far, we have discussed *stable* recovery, which accounts for the fact that vectors are not exactly sparse (compressibility). We now turn to *robust* recovery, which accounts for noise in the measurements. We consider the model:

$$y = Ax + e, \quad \text{with } \|e\|_2 \le \eta,$$

where e represents measurement noise bounded by η. To recover x, we consider the quadratically constrained ℓ_1-minimization problem (often called Basis Pursuit Denoising):

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \quad \text{subject to } \|Az - y\|_2 \le \eta. \tag{$P_{1,\eta}$}$$

To guarantee recovery under these conditions, we must strengthen the Null Space Property.

Definition 5.10 (Robust Null Space Property)**.** *The matrix A satisfies the* **Robust Null Space Property** *(or ℓ_2-Robust NSP) of order s with constants $0 < \rho < 1$ and $\tau > 0$ if, for all $v \in \mathbb{C}^N$ and any $S \subset \{1, \dots, N\}$ with $|S| \le s$,*

$$\|v_S\|_1 \le \rho \|v_{\overline{S}}\|_1 + \tau \|Av\|_2. \tag{5}$$

Remark. Notice that if $\tau = 0$, we recover the standard Stable NSP. The term $\tau\|Av\|_2$ allows elements in the null space (or close to it) to have a larger support norm if they produce a non-zero measurement response, providing a margin for noise.

The following theorem establishes that the Robust NSP implies success for Basis Pursuit Denoising.

Theorem 5.11 (Stable and Robust Recovery)**.** *Suppose that A satisfies the Robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Let $x \in \mathbb{C}^N$ and $y = Ax + e$ with $\|e\|_2 \le \eta$. Let $x^\#$ be a solution to $(P_{1,\eta})$. Then,*

$$\|x - x^\#\|_1 \le \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1 + \frac{4\tau}{1-\rho}\eta.$$

Proof. Let $h = x - x^\#$ be the error vector. We aim to bound $\|h\|_1$. First, we observe that both x and $x^\#$ are feasible for the minimization problem. By hypothesis, the true noise satisfies $\|Ax - y\|_2 = \|e\|_2 \le \eta$. Since $x^\#$ is a solution to $(P_{1,\eta})$, it must satisfy the constraint $\|Ax^\# - y\|_2 \le \eta$.

Using the triangle inequality, we can bound the image of the error vector Ah:

$$\|Ah\|_2 = \|A(x - x^\#)\|_2 = \|(Ax - y) - (Ax^\# - y)\|_2 \le \|Ax - y\|_2 + \|Ax^\# - y\|_2 \le 2\eta. \tag{6a}$$

Comment on Eq. (6a)*:* This specific bound is crucial. While we do not know the noise e exactly, the feasibility of both vectors guarantees the measurement residual of the difference is bounded by 2η.

Next, we proceed with the standard norm decomposition. Let S be the set of indices of the s largest entries of x in modulus. Since $x^\#$ is the minimizer of the ℓ_1-norm, $\|x^\#\|_1 \le \|x\|_1$. Following the same derivation used in the proof of the Stable NSP:

$$\|x\|_1 \ge \|x + h\|_1 = \|x_S + h_S\|_1 + \|x_{\overline{S}} + h_{\overline{S}}\|_1 \ge \|x_S\|_1 - \|h_S\|_1 + \|h_{\overline{S}}\|_1 - \|x_{\overline{S}}\|_1.$$

Rearranging terms yields:

$$\|h_{\overline{S}}\|_1 \le \|h_S\|_1 + 2\|x_{\overline{S}}\|_1 = \|h_S\|_1 + 2\sigma_s(x)_1.$$

Now we apply the Robust NSP (Definition 5.10) to the vector h. Since A satisfies the property with constants ρ and τ:

$$\|h_S\|_1 \le \rho\|h_{\overline{S}}\|_1 + \tau\|Ah\|_2.$$

Combining this with Eq. (6a), we get:

$$\|h_S\|_1 \le \rho\|h_{\overline{S}}\|_1 + 2\tau\eta. \tag{6b}$$

The total error is $\|h\|_1 = \|h_S\|_1 + \|h_{\overline{S}}\|_1$. We substitute (6b) into the total sum, but first, let us bound $\|h_{\overline{S}}\|_1$ solely in terms of $\sigma_s(x)_1$ and η. Substituting (6b) into the earlier rearrangement $\|h_{\overline{S}}\|_1 \le \|h_S\|_1 + 2\sigma_s(x)_1$:

$$\|h_{\overline{S}}\|_1 \le \big(\rho\|h_{\overline{S}}\|_1 + 2\tau\eta\big) + 2\sigma_s(x)_1.$$

Solving for $\|h_{\overline{S}}\|_1$:

$$(1-\rho)\|h_{\overline{S}}\|_1 \le 2\sigma_s(x)_1 + 2\tau\eta \quad\implies\quad \|h_{\overline{S}}\|_1 \le \frac{2\sigma_s(x)_1 + 2\tau\eta}{1-\rho}.$$

Finally, using $\|h\|_1 \le (1+\rho)\|h_{\overline{S}}\|_1 + 2\tau\eta$ (derived from adding $\|h_{\overline{S}}\|_1$ to (6b)):

$$\begin{aligned}
\|h\|_1 &\le \frac{1+\rho}{1-\rho}(2\sigma_s(x)_1 + 2\tau\eta) + 2\tau\eta \\
&= \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1 + \left(\frac{2\tau(1+\rho)}{1-\rho} + 2\tau\right)\eta \\
&= \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1 + \frac{2\tau(1+\rho) + 2\tau(1-\rho)}{1-\rho}\eta \\
&= \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1 + \frac{4\tau}{1-\rho}\eta.
\end{aligned}$$

This concludes the proof. □

Remark 5.12 (Interpretation of the Error Bound)**.** *The error estimate in Theorem 5.11 decomposes naturally into two distinct components, reflecting the dual nature of the recovery guarantee:*

- **Stability term:** *The first term, $\frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1$, accounts for the "modeling error," i.e., the deviation of the vector x from being exactly s-sparse (or at least to share an error bound similar to the best s-term approximation). This ensures that the reconstruction is stable with respect to compressibility: if x is well-approximated by an s-sparse vector, the reconstruction error remains controlled.*

- **Robustness term:** *The second term, $\frac{4\tau}{1-\rho}\eta$, accounts for the measurement noise e where $\|e\|_2 \le \eta$. This ensures that the reconstruction is robust: the reconstruction error scales linearly with the noise magnitude η.*

Remark 5.13 (Exact Recovery)**.** *In the ideal scenario where the signal is exactly s-sparse (implying $\sigma_s(x)_1 = 0$) and the measurements are noiseless (implying $\eta = 0$), the right-hand side of the inequality vanishes. Consequently, $\left\|x - x^\#\right\|_1 = 0$, which guarantees exact recovery $x^\# = x$. This generalizes the results of the noiseless setup to the noisy case.*

Remark 5.14 (The Constants and the "Price of Robustness")**.** *The constants appearing in the bound provide insight into the quality of the recovery permitted by the matrix A:*

- *The constant $C_1 = \frac{2(1+\rho)}{1-\rho}$ governs the amplification of the approximation error. Note that as $\rho \to 0$ (stronger NSP), $C_1 \to 2$, which is the optimal factor for instance optimality in ℓ_1.*
- *The constant $C_2 = \frac{4\tau}{1-\rho}$ represents the "price of robustness." It dictates how severely the measurement noise is amplified in the recovered signal. The parameter τ (from the Robust NSP definition) acts as a scaling factor between the ℓ_2-norm of the noise and the ℓ_1-norm of the recovery error.*

Crucially, both constants diverge as $\rho \to 1$. This implies that if the matrix A only barely satisfies the Null Space Property (with ρ close to 1), the reconstruction becomes highly sensitive to both noise and sparsity defects.