Dartmouth College

ENGS 109

High Dimensional Sensing and Learning

Comprehensive Course Notes

Final Exam Reference Edition

Scope. Compressive sensing, sparse recovery, greedy and convex algorithms, random matrices, matrix completion, robust PCA, dictionary learning, nonnegative matrix factorization, and deep sparse coding.

Exam reference. The opening quick-reference section contains the Lecture 25X final-review guidance and the principal proof templates.

Prepared by: Farhan Sadeek

Term: Spring 2026

Source text: Foucart & Rauhut, *A Mathematical Introduction to Compressive Sensing*

Contents

Preface

These notes cover ENGS 109 *High Dimensional Sensing and Learning*, a graduate course on **compressive sensing** (CS), sparse representation theory, and applications in machine learning and signal processing. They are written to be self-contained and mathematically rigorous, with every major theorem followed by a proof or proof sketch. The reference text is Foucart & Rauhut, *A Mathematical Introduction to Compressive Sensing* (Birkhäuser/Springer, 2013), abbreviated [FR] below; chapter and theorem numbers refer to that monograph.

The course studies the underdetermined inverse problem

$$\boldsymbol{y} = A\boldsymbol{x}, \qquad A \in \mathbb{C}^{m\times N}, \quad m < N,$$

and identifies the conditions on A under which a sparse $\boldsymbol{x}$ can be recovered exactly or stably from $\boldsymbol{y}$. The theoretical core is the implication chain

$$\boxed{\text{random matrix}} \implies \boxed{\text{RIP}} \implies \boxed{\text{NSP}} \implies \boxed{\ell_1 \text{ recovers } \ell_0},$$

augmented by the coherence-based theory and the specialized analyses of greedy and thresholding algorithms.

Notation.

Vectors are bold lowercase ($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{a}_j$); matrices are uppercase (A, V, U, Σ). $\boldsymbol{a}_j$ denotes the j-th column of A, and A_S the submatrix of columns indexed by $S \subseteq [N] := \{1, \dots, N\}$. The ℓ_p-(quasi)norm is $\|\boldsymbol{x}\|_p = (\sum_j |x_j|^p)^{1/p}$ for $0 < p < \infty$, $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$, and $\|\boldsymbol{x}\|_0 := |\mathrm{supp}(\boldsymbol{x})|$. The set of s-sparse vectors is

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

Best s-term approximation error in the ℓ_p-norm is $\sigma_s(\boldsymbol{x})_p := \inf_{\boldsymbol{z}\in\Sigma_s} \|\boldsymbol{x} - \boldsymbol{z}\|_p$. The complement of S in $[N]$ is denoted $\bar{S}$ or S^c. We write $\boldsymbol{v}_S \in \mathbb{C}^N$ for the vector with $v_S(j) = v_j$ if $j \in S$ and 0 otherwise. The Hermitian transpose of A is A^* (or $A^\top$ in the real case); $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{v}^*\boldsymbol{u}$ is the Euclidean inner product.

Exam Quick Reference

This section is a compact exam reference for the course. It is intended for rapid lookup: identify the model, choose the structural prior, then quote the right condition or algorithm.

Lecture 25X final-exam guidance

Final-review logistics and exam profile

- **Exam time.** Sunday, June 7, 2026, 8:00 am, regular classroom.
- **Allowed materials.** Unlimited *paper* notes and a calculator. The Lecture 25X review explicitly says no phone, laptop, tablet, or smartwatch. Use this as the controlling logistics note for permitted materials.
- **Difficulty.** Comparable to homework proofs. Minh emphasized proof-and-construction problems rather than long computations.
- **Likely format.** Expect one or two questions of the form: "prove statement (A) is equivalent to statement (B)," especially in the null-space, spark, and uniqueness family.
- **Problem-solving approach.** The problem is usually not pure recall: it needs one intermediate construction, such as splitting a support into two pieces or forming the difference of two feasible solutions.

Likely exam archetypes

1. **Equivalence proof.** Show uniqueness of sparse recovery iff a kernel/spark/NSP condition holds. One direction is usually by contradiction; the other is usually by direct construction.
2. **Compute a certificate.** For a small explicit matrix, compute or bound $\operatorname{spark}(A)$, $\mu(A)$, or an RIP-related quantity, then state the largest guaranteed sparsity.
3. **Trace a greedy algorithm.** Run MP/OMP by hand: compute correlations $|A^* \boldsymbol{r}|$, select the largest, solve least squares on the selected support, and update the residual.
4. **Derive an update rule.** Write the loss, take a gradient, then either set it to zero (ALS/MOD) or use a chosen step size (NMF multiplicative updates).
5. **Match structure to method.** Sparse vector $\to \ell_1$/OMP; low rank $\to$ nuclear norm/ALS; low rank plus sparse $\to$ RPCA; nonnegative parts $\to$ NMF; learned sparse dictionary $\to$ K-SVD/MOD.

6. **Short conceptual.** Be ready to explain why ℓ_1 promotes sparsity, why $\|\cdot\|_0$ is not a norm, and why NMF topics are more interpretable than SVD components.

Spark equivalence proof template

This is the worked proof Minh highlighted as exam-level. Learn the construction, not only the theorem statement.

Uniform P_0 uniqueness iff no short kernel vector

For $A \in \mathbb{R}^{m\times N}$ and integer $s \geq 1$, the following are equivalent:

$$\text{(A) every } s\text{-sparse } \boldsymbol{x} \text{ is the unique minimizer of } \min_{z} \|z\|_0 \quad \text{s.t.} \quad Az = A\boldsymbol{x},$$

and

$$\text{(B) every nonzero } \boldsymbol{v} \in \mathbf{ker}\, A \text{ has } \|\boldsymbol{v}\|_0 > 2s, \qquad \text{equivalently} \qquad \operatorname{spark}(A) > 2s.$$

(A) $\Rightarrow$ (B), contradiction. If (B) fails, choose $\mathbf{0} \neq \boldsymbol{v} \in \mathbf{ker}\, A$ with $\|\boldsymbol{v}\|_0 \leq 2s$. Split $S = \operatorname{supp}(\boldsymbol{v})$ into disjoint S_1, S_2 with $|S_1|, |S_2| \leq s$. Define

$$\boldsymbol{x} = \boldsymbol{v}_{S_1}, \qquad z = -\boldsymbol{v}_{S_2}.$$

Then $\boldsymbol{x}$ and z are distinct s-sparse vectors and $\boldsymbol{x} - z = \boldsymbol{v} \in \mathbf{ker}\, A$, so $A\boldsymbol{x} = Az$. This gives two distinct feasible s-sparse explanations for the same measurement, contradicting uniqueness.

(B) $\Rightarrow$ (A), construction. Let $\boldsymbol{x}$ be any s-sparse vector and let z be feasible with $Az = A\boldsymbol{x}$ and $\|z\|_0 \leq \|\boldsymbol{x}\|_0 \leq s$. Then

$$\boldsymbol{v} = \boldsymbol{x} - z \in \mathbf{ker}\, A, \qquad \|\boldsymbol{v}\|_0 \leq \|\boldsymbol{x}\|_0 + \|z\|_0 \leq 2s.$$

By (B), $\boldsymbol{v}$ must be zero, hence $z = \boldsymbol{x}$. Therefore $\boldsymbol{x}$ is the unique sparsest feasible vector.

Universal hinge. If two candidates satisfy $A\boldsymbol{x} = Az$, then $\boldsymbol{x} - z \in \mathbf{ker}\, A$. If both candidates are s-sparse, their difference is at most $2s$-sparse. Most uniqueness proofs in this course are this argument with a stronger condition replacing $\operatorname{spark}(A) > 2s$.

NMF / PS9 essentials from Lecture 25X

For a nonnegative matrix $V \in \mathbb{R}^{m\times n}_{\geq 0}$, NMF solves

$$\min_{W,H\geq 0} \ \frac{1}{2}\|V - WH\|_F^2, \qquad W \in \mathbb{R}^{m\times r}_{\geq 0}, \quad H \in \mathbb{R}^{r\times n}_{\geq 0}.$$

The problem is nonconvex jointly in (W, H), but convex in one factor with the other fixed.

Item	**Exam-relevant fact**
Lee–Seung updates	$H \leftarrow H \circ \frac{W^\top V}{W^\top WH + \epsilon}$, $W \leftarrow W \circ \frac{VH^\top}{WHH^\top + \epsilon}$. Products are matrix multiplications; $\circ$ and division are entrywise.
Shape check	$W^\top V$ and $W^\top WH$ are $r \times n$, matching H. $VH^\top$ and $WHH^\top$ are $m \times r$, matching W.
Update order	Use H first, then update W with the newly updated H (Gauss–Seidel style). This is the standard notebook order.
Why nonnegative	Start with $W, H \geq 0$. Each update multiplies a nonnegative entry by a nonnegative ratio, so no projection step is needed.
Derivation target	$\nabla_H \frac{1}{2}\|V - WH\|_F^2 = W^\top WH - W^\top V$. Choosing $\eta_{ij} = H_{ij}/(W^\top WH)_{ij}$ in a gradient step gives $H_{ij} \leftarrow H_{ij}(W^\top V)_{ij}/(W^\top WH)_{ij}$.
Common implementation errors	Using matrix multiply instead of entrywise multiply/divide in the final line; forgetting ϵ; getting an increasing reconstruction error curve; transposing the wrong factor.
Topic modeling	`TfidfVectorizer` returns documents $\times$ terms. Transpose to terms $\times$ documents if columns of W are supposed to be topics over words. TF-IDF is nonnegative, so it is valid NMF input.
NMF vs SVD	SVD components have mixed signs and can use cancellation; NMF topics are additive nonnegative word lists, so they are usually more interpretable.

Selected worked results

- $\|\cdot\|_0$ **is not a norm.** It fails homogeneity: for nonzero $\boldsymbol{x}$, $\|2\boldsymbol{x}\|_0 = \|\boldsymbol{x}\|_0 \neq 2\|\boldsymbol{x}\|_0$. It does satisfy positive definiteness and $\|\boldsymbol{x} + \boldsymbol{y}\|_0 \leq \|\boldsymbol{x}\|_0 + \|\boldsymbol{y}\|_0$ by the support union bound.
- ℓ_1 **vs** ℓ_2 **two-variable example.** For $A = [1\ 2]$ and $y = 2$, feasible vectors satisfy $x_1 + 2x_2 = 2$. The ℓ_1 minimizer is $(0, 1)^\top$; the minimum-ℓ_2 solution is $A^\top(AA^\top)^{-1}y = (2/5, 4/5)^\top$.

- **OMP hand-trace result.** For columns $\boldsymbol{a}_1 = (1,0,0)^\top$, $\boldsymbol{a}_2 = (0,1,0)^\top$, $\boldsymbol{a}_3 = (1/\sqrt{2}, 1/\sqrt{2}, 0)^\top$, $\boldsymbol{a}_4 = (0,0,1)^\top$ and $\boldsymbol{y} = (2,0,3)^\top$, two-step OMP picks support $\{4,1\}$ and returns $\boldsymbol{x} = (2,0,0,3)^\top$.
- **Small spark/coherence example.** For $A = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$, $\mu(A) = 1/\sqrt{2}$, $\mathrm{spark}(A) = 3$, and the guaranteed uniform sparsity is only $s = 1$.
- **Matrix completion convexity.** The feasible set $\{X : X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega\}$ is affine, hence convex; $\|X\|_*$ is a norm, hence convex. Therefore nuclear-norm matrix completion is convex.

Problem dictionary

Problem	**Model**	**Hard prior**	**Convex / practical tool**
Sparse recovery	$\boldsymbol{y} = A\boldsymbol{x}$, $m \ll N$	$\|\boldsymbol{x}\|_0 \le s$	BP/BPDN/LASSO: ℓ_1 minimization
Noisy / compressible recovery	$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$, $\|\boldsymbol{e}\|_2 \le \eta$	$\boldsymbol{x}$ close to Σ_s	$\|\boldsymbol{x} - \boldsymbol{x}^{\#}\|_2 \lesssim \sigma_s(\boldsymbol{x})_1/\sqrt{s} + \eta$
MMV / joint sparsity	$Y = AX$, same row support across columns	X has at most s nonzero rows	$\ell_{1,2}$ minimization, SOMP, SSP/CoSaMP lifts
Matrix completion	$Y = P_\Omega(X)$, observe entries only on Ω	$\operatorname{rank}(X) \le r$	Nuclear norm $\|X\|_*$, SVT/ALM/APG
Robust PCA	$Y = L + S$	L low-rank, S sparse	$\min \|L\|_* + \lambda\|S\|_1$
SRC classification	$\boldsymbol{y} \approx [A_1\|\cdots\|A_C]\boldsymbol{x}$	$\boldsymbol{x}$ concentrated on one class block	BPDN, residuals r_i, SCI rejection
Dictionary learning	$Y \approx DX$ with D unknown	columns of X are sparse	K-means, MOD, K-SVD, online dictionary learning
NMF	$Y \approx DX$, $D, X \ge 0$	additive parts / cluster memberships	Lee–Seung multiplicative updates, NNLS
Deep sparse coding	stacked sparse-coding modules	sparse, nonnegative elastic-net features	FISTA inference + supervised back-propagation

Theorem lookup

- **Exact sparse uniqueness.** Every s-sparse vector is the unique s-sparse solution iff $\mathbf{ker}(A) \cap \Sigma_{2s} = \{\mathbf{0}\}$ iff $\operatorname{spark}(A) > 2s$ (Theorem 2.1). This also implies the information lower bound $m \ge 2s$ for uniform P_0 recovery.
- **Basis pursuit exact recovery.** BP recovers every s-sparse vector iff A satisfies the NSP of order s (Theorem 5.2). In proofs, take any competing feasible $\boldsymbol{z}$, set $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{z} \in \mathbf{ker}\, A$, and compare $\|\boldsymbol{x} + \boldsymbol{v}\|_1$ to $\|\boldsymbol{x}\|_1$ on $S = \operatorname{supp}(\boldsymbol{x})$.
- **Approximate and noisy recovery.** Stable NSP gives $\|\boldsymbol{x} - \boldsymbol{x}^{\#}\|_1 \lesssim \sigma_s(\boldsymbol{x})_1$; robust NSP gives BPDN stability under noise (Theorems 5.5 and 5.7).

- **RIP route.** RIP of order $2s$ implies robust NSP and BPDN recovery. Use $\delta_{2s} < 4/\sqrt{41}$ as the sharp listed condition, or $\delta_{2s} < \sqrt{2} - 1$ as the simpler textbook sufficient condition (Theorem 7.3).
- **Coherence route.** For normalized columns, $\mu(A) = \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|$ and $\mu(A) \geq \sqrt{(N-m)/(m(N-1))}$ (Welch, Theorem 6.3). The simple sufficient condition

$$\mu(A) < \frac{1}{2s-1} \quad \Longleftrightarrow \quad s < \frac{1}{2}(1 + 1/\mu(A))$$

  gives OMP and BP recovery through the cumulative coherence theorem (Theorem 6.4), but it usually costs $m \gtrsim s^2$.
- **Random matrices.** Subgaussian matrices satisfy RIP with high probability once

$$m \gtrsim \delta^{-2} s \log(eN/s)$$

  (Theorem 8.2). This is the optimal compressive-sensing scaling up to constants.
- **Matrix analogues.** Replace support size by rank and ℓ_1 by nuclear norm. Rank NSP characterizes nuclear-norm recovery; Candès–Recht gives exact matrix completion under incoherence and random sampling; Candès–Li–Ma–Wright gives exact RPCA when rank and sparse corruption are both small.

Algorithm selection

Use case	Algorithm	Exam facts to remember
Small exact sparse recovery	Brute-force P_0 over supports	Try supports in increasing size; for each support solve $A_S^\dagger \boldsymbol{y}$ and check residual. Cost $\binom{N}{s}$, so this is only practical for small N, s.
Fast sparse recovery with known s	OMP / MP / IHT / HTP / CoSaMP	OMP adds one atom and refits by LS; IHT thresholds a gradient step; HTP adds an LS refit; CoSaMP identifies $2s$, merges, refits, then prunes.
Best convex sparse recovery certificate	BP/BPDN/LASSO	Use BP for noiseless equality, BPDN for bounded noise, LASSO for penalized least squares. Soft thresholding is the prox of $\lambda\|\cdot\|_1$.
Large ℓ_1 problems	ISTA/FISTA, ADMM, IRLS, homotopy	FISTA has $O(1/k^2)$ objective convergence; ADMM separates the constraint; IRLS solves weighted least squares and updates weights.
Shared support across tasks	SOMP or $\ell_{1,2}$ minimization	MMV uniqueness improves to $s < (\text{spark}(A) - 1 + \text{rank}(Y))/2$ when $\text{rank}(Y) > 1$.
Low-rank missing data	SVT / nuclear-norm minimization	The prox of $\tau\|\cdot\|_*$ is singular-value shrinkage: $U\text{diag}((\sigma_i - \tau)_+)V^*$.
Low-rank plus sparse outliers	ALM / ADMM for RPCA	Alternate singular-value shrinkage for L and entrywise soft-thresholding for S.
Learned sparse representation	K-SVD / MOD	MOD updates $D = YX^\dagger$; K-SVD updates one atom by the top SVD of the restricted residual, preserving the sparsity pattern.
Parts-based nonnegative features	NMF	Multiplicative updates preserve nonnegativity and monotonically decrease $\|Y - DX\|_F^2$, but the problem is nonconvex and nonunique.

Common exam moves

1. **Difference of two feasible solutions.** If $Ax = Az$, set $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{z} \in \ker A$. If both are s-sparse, then $\boldsymbol{v} \in \Sigma_{2s}$.

2. **Top-s support split.** For stable/robust proofs, let S be the indices of the largest s entries, then split S^c into blocks $S_1, S_2, \ldots$ of size s. The recurring bound is $\sum_{j \geq 2} \|\boldsymbol{v}_{S_j}\|_2 \leq \|\boldsymbol{v}_{S^c}\|_1 / \sqrt{s}$.

3. **Coherence proof pattern.** Bound on-support correlations from below by $1 - \mu_1(s-1)$ and off-support correlations from above by $\mu_1(s)$. Correct selection follows when $\mu_1(s) + \mu_1(s-1) < 1$.

4. **RIP proof pattern.** Use RIP on a union support and restricted orthogonality for disjoint supports. The union size determines the RIP order: $2s$ for BP, $3s$ for IHT/HTP, $4s$ for CoSaMP.

5. **Matrix proof pattern.** Replace entry magnitudes by singular values. Rank is $\|\sigma(X)\|_0$, nuclear norm is $\|\sigma(X)\|_1$, and Frobenius norm is $\|\sigma(X)\|_2$.

6. **Implementation checks.** Normalize dictionary columns before coherence, OMP, SRC, and dictionary learning. In SRC, split train/test per class before stacking; in matrix completion, missing entire rows or columns makes recovery impossible; in K-SVD, restrict the SVD update to examples that actually use the atom.

Part I

Foundations: Sparsity and the Inverse Problem

1 An Invitation to Compressive Sensing

1.1 The central question

In many problems of science and technology—especially in signal and image processing—one observes data $\boldsymbol{y} \in \mathbb{C}^m$ produced by a linear measurement process,

$$A\boldsymbol{x} = \boldsymbol{y}, \qquad A \in \mathbb{C}^{m\times N},\ \boldsymbol{x} \in \mathbb{C}^N. \tag{1}$$

Classical wisdom (the Shannon–Nyquist sampling theorem) tells us that the number of measurements m must equal or exceed the signal length N. In compressive sensing, the assumption $m < N$ makes (1) *underdetermined*: it has either no solution or infinitely many. Yet under the additional prior that $\boldsymbol{x}$ is *sparse*, exact recovery is possible from $m \sim s\log(N/s)$ measurements, provided A is suitably designed.

The compressive sensing programme thus identifies two coupled problems:

(P1) *Matrix design.* Which matrices $A \in \mathbb{C}^{m\times N}$ enable robust sparse recovery, and at what minimal m?

(P2) *Recovery algorithm.* Which algorithm reconstructs $\boldsymbol{x}$ from $\boldsymbol{y} = A\boldsymbol{x}$ efficiently?

1.2 Sparsity and compressibility

Definition 1.1 (Sparsity, support)

Let $\boldsymbol{x} \in \mathbb{C}^N$. Its *support* is $\operatorname{supp}(\boldsymbol{x}) := \{j \in [N] : x_j \neq 0\}$, and

$$\|\boldsymbol{x}\|_0 := |\operatorname{supp}(\boldsymbol{x})|$$

is the *sparsity*. Note that $\|\cdot\|_0$ is not a norm (no homogeneity). We call $\boldsymbol{x}$ *s-sparse* if $\|\boldsymbol{x}\|_0 \le s$, and write $\Sigma_s = \{\boldsymbol{x} \in \mathbb{C}^N : \|\boldsymbol{x}\|_0 \le s\}$.

Definition 1.2 (Best s-term approximation)

For $0 < p \le \infty$, the ℓ_p-error of best s-term approximation is

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

The infimum is achieved by the vector $\boldsymbol{x}_S$ obtained from $\boldsymbol{x}$ by retaining only the s entries of largest absolute value (and is the same for all p).

Definition 1.3 (Compressibility and weak ℓ_p space)

A vector $\boldsymbol{x}$ is *compressible* if $\sigma_s(\boldsymbol{x})_p$ decays quickly in s. The weak ℓ_p-quasinorm is

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

where $\boldsymbol{x}^* \in \mathbb{R}_+^N$ is the non-increasing rearrangement of $|\boldsymbol{x}|$. The set $\{\boldsymbol{x} : \|\boldsymbol{x}\|_p \le 1\}$ is a non-convex ball for $p < 1$ and serves as a model for compressible signals.

Theorem 1.4 (Stechkin's inequality)

For any $0 < p < q \le \infty$ and any $\boldsymbol{x} \in \mathbb{C}^N$,

$$\sigma_s(\boldsymbol{x})_q \;\le\; \frac{C_{p,q}}{s^{1/p-1/q}} \|\boldsymbol{x}\|_p, \qquad C_{p,q} := \left[\left(\tfrac{p}{q}\right)^{p/q}\left(1-\tfrac{p}{q}\right)^{1-p/q}\right]^{1/p} \le 1.$$

The choice $p = 1, q = 2$ yields the most-used special case

$$\sigma_s(\boldsymbol{x})_2 \;\le\; \frac{1}{2\sqrt{s}} \|\boldsymbol{x}\|_1.$$

Proof. Let $\boldsymbol{x}^*$ be the non-increasing rearrangement of $|\boldsymbol{x}|$. The error of best s-term approximation is $\sigma_s(\boldsymbol{x})_q = (\sum_{j=s+1}^N (x_j^*)^q)^{1/q}$. We wish to maximize this quantity subject to the constraint $\|\boldsymbol{x}\|_p^p = \sum_{j=1}^N (x_j^*)^p = L$. Let $\alpha_j = (x_j^*)^p$. The problem is to maximize $f(\boldsymbol{\alpha}) = \sum_{j=s+1}^N \alpha_j^{q/p}$ subject to $\sum_{j=1}^N \alpha_j = L$ and $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_N \ge 0$.

Since $q/p > 1$, the function f is convex. The maximum of a convex function over a compact convex set is attained at an extreme point. The extreme points of the constraint set are of the form $\boldsymbol{\alpha}^{(k)} = (L/k, \ldots, L/k, 0, \ldots, 0)$ for $k \in \{1, \ldots, N\}$.

- If $k \le s$, then $\alpha_j = 0$ for all $j > s$, so $f(\boldsymbol{\alpha}^{(k)}) = 0$.
- If $k > s$, then $f(\boldsymbol{\alpha}^{(k)}) = \sum_{j=s+1}^k (L/k)^{q/p} = (k-s)(L/k)^{q/p} = L^{q/p}\frac{k-s}{k^{q/p}}$.

To find the maximum, we consider the function $g(k) = (k-s)k^{-q/p}$ for $k \in [s, N]$. Setting $g'(k) = k^{-q/p} - \frac{q}{p}k^{-q/p-1}(k-s) = 0$ yields $1 - \frac{q}{p}(1 - s/k) = 0$, which implies $k = s\frac{q}{q-p}$. Plugging this value back into $g(k)$ gives the constant $c_{p,q}^q$. Specifically, for $p = 1, q = 2$, the maximum is attained near $k = 2s$, yielding the constant $1/(2\sqrt{s})$.

Note: *2.5 for reference*

□

1.3 Recovery formulations

The naive sparse-recovery problem is the ℓ_0-minimization

$$\min_{z\in\mathbb{C}^N} \|z\|_0 \quad \text{s.t.} \quad Az = y. \tag{P_0}$$

With noise it becomes

$$\min_{z} \|z\|_0 \quad \text{s.t.} \quad \|Az - y\|_2 \le \eta. \tag{$P_{0,\eta}$}$$

The convex relaxation, the *basis pursuit*, replaces $\|\cdot\|_0$ by $\|\cdot\|_1$:

$$\min_{z} \|z\|_1 \quad \text{s.t.} \quad Az = y. \tag{P_1}$$

Variants include

$$\text{BPDN} \quad \min \|z\|_1 \text{ s.t. } \|Az - y\|_2 \le \eta, \tag{2}$$

$$\text{LASSO} \quad \min \tfrac{1}{2}\|Az - y\|_2^2 + \lambda\|z\|_1, \tag{3}$$

$$\text{Dantzig selector} \quad \min \|z\|_1 \text{ s.t. } \|A^*(Az - y)\|_\infty \le \tau. \tag{4}$$

P_0 is NP-hard (Theorem 2.7); P_1 is a convex linear program.

1.4 Linear-program reformulation of P_1 (real case)

Introduce $u \in \mathbb{R}^N$ with $-u_j \le x_j \le u_j$. Then

$$\min_{x,u} \mathbf{1}^\top u \text{ s.t. } Ax = y,\ -u \le x \le u$$

is a linear program equivalent to P_1. Equivalently, splitting $x = z^+ - z^-$ with $z^\pm \ge 0$:

$$\min_{z^\pm \ge 0} \mathbf{1}^\top (z^+ + z^-) \text{ s.t. } [A \mid -A] \begin{bmatrix} z^+ \\ z^- \end{bmatrix} = y.$$

1.5 Motivating applications

Single-pixel camera (Rice).

A digital micromirror device steers light from random pixel patterns onto a single photodiode. Each measurement $y_\ell = \langle \boldsymbol{a}_\ell, \boldsymbol{z} \rangle$ uses a random ± 1 pattern. If $\boldsymbol{z} = W\boldsymbol{x}$ with W a wavelet basis and $\boldsymbol{x}$ sparse, the system becomes $\boldsymbol{y} = AW\boldsymbol{x} = A'\boldsymbol{x}$, a standard CS problem.

Magnetic resonance imaging (MRI).

The continuous measurement is $f(t) = \int_{\mathbb{R}^3} |X(\boldsymbol{z})|\, e^{-2\pi i \boldsymbol{k}(t)\cdot \boldsymbol{z}}\, d\boldsymbol{z}$, the Fourier transform of the magnetization sampled along trajectory $\boldsymbol{k}(t)$. Discretizing yields $\boldsymbol{y} = R_K \mathcal{F}\boldsymbol{x}$, a partial-Fourier system. Sparsity in wavelets or total variation enables sub-Nyquist scans.

Radar.

The reflected signal from s scatterers takes the form $\boldsymbol{y} = \sum_{(k,\ell)} x_{k\ell}\, T_k M_\ell\, \boldsymbol{g}$ with T_k time-shifts and M_ℓ frequency modulations of a known pulse $\boldsymbol{g}$. Few targets $\Rightarrow$ sparse $\boldsymbol{x}$.

Robust PCA.

A data matrix M that is the sum of low-rank L and sparse S admits a convex decomposition $\min \|L\|_* + \lambda \|S\|_1$ s.t. $M = L + S$ (Candès–Li–Ma–Wright 2011). Removes "snow on a campus photo," separates background from foreground.

Sparse face recognition.

Stack training faces of class i as columns of A_i and concatenate $A = [A_1 \mid \dots \mid A_C]$. A novel face $\boldsymbol{y}$ has a near-block-sparse representation $\boldsymbol{y} \approx A\boldsymbol{x}$ that simultaneously identifies the class and the contributing training samples (Lecture 8).

1.6 Founding papers

The CS programme was launched by:

- E. Candès, J. Romberg, T. Tao (2006) – introduced the term *compressed sensing*, proved the foundational ℓ_1 recovery results, and identified the RIP.
- D. Donoho (2006) – coined *compressive sensing* and developed the Gelfand-width and combinatorial-geometry perspectives.

Antecedents include Prony (1795, sparse Fourier), Tibshirani (1996, LASSO), and Mallat–Zhang (1993, matching pursuit).

2 Sparse Solutions of Underdetermined Systems

2.1 When does a sparse solution exist—and is it unique?

We examine *uniform* recovery: which matrices A allow recovery of every s-sparse vector? The answer is a beautiful characterization in linear-algebra language.

Theorem 2.1 (Equivalent uniqueness conditions)

For $A \in \mathbb{C}^{m \times N}$, the following are equivalent:

1. Every s-sparse $\boldsymbol{x}$ is the unique s-sparse solution of $A\boldsymbol{z} = A\boldsymbol{x}$.
2. $\mathbf{ker}(A) \cap \Sigma_{2s} = \{\mathbf{0}\}$.
3. For every $S \subseteq [N]$ with $|S| \leq 2s$, the submatrix A_S is injective.
4. Every set of $2s$ columns of A is linearly independent.

Proof. (2) $\Leftrightarrow$ (3) $\Leftrightarrow$ (4) are standard results in linear algebra: A_S is injective iff its columns are linearly independent, and $\mathbf{ker}(A) \cap \Sigma_{2s} = \{\mathbf{0}\}$ is simply the statement that no non-trivial linear combination of $2s$ columns can equal zero.

(1) $\implies$ (2): We prove the contrapositive. Suppose there exists $\boldsymbol{v} \in \mathbf{ker}(A) \cap \Sigma_{2s}$ such that $\boldsymbol{v} \neq \mathbf{0}$. We can decompose $\boldsymbol{v}$ as $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}'$, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are s-sparse vectors with disjoint supports (e.g., let $\boldsymbol{x}$ take the first s non-zero entries of $\boldsymbol{v}$ and $\boldsymbol{x}'$ take the negative of the remaining entries). Since $A\boldsymbol{v} = \mathbf{0}$, we have $A\boldsymbol{x} = A\boldsymbol{x}'$. Thus $\boldsymbol{x}$ is an s-sparse solution to $A\boldsymbol{z} = A\boldsymbol{x}$, but it is not unique because $\boldsymbol{x}'$ is another s-sparse solution. This contradicts (1).

(2) $\implies$ (1): Suppose $\mathbf{ker}(A) \cap \Sigma_{2s} = \{\mathbf{0}\}$. Let $\boldsymbol{x} \in \Sigma_s$ and suppose there exists $\boldsymbol{x}' \in \Sigma_s$ such that $A\boldsymbol{x} = A\boldsymbol{x}'$. Let $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}'$. Then $A\boldsymbol{v} = \mathbf{0}$. Furthermore, $\|\boldsymbol{v}\|_0 = \|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq \|\boldsymbol{x}\|_0 + \|\boldsymbol{x}'\|_0 \leq 2s$. Thus $\boldsymbol{v} \in \mathbf{ker}(A) \cap \Sigma_{2s}$. By (2), $\boldsymbol{v} = \mathbf{0}$, which implies $\boldsymbol{x} = \boldsymbol{x}'$. Hence $\boldsymbol{x}$ is the unique s-sparse solution.

Note: *2.13 for reference*

□

Corollary 2.2

Uniform recovery of all s-sparse signals requires $m \geq 2s$.

This bound is achievable in principle:

Theorem 2.3 (Vandermonde construction)

For any $N \geq 2s$, the Vandermonde matrix

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_N \\ \vdots & \vdots & & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \cdots & t_N^{2s-1} \end{pmatrix} \in \mathbb{C}^{2s \times N}$$

with distinct nodes $t_1, \ldots, t_N$ allows recovery of every s-sparse signal via P_0. Its determinant on any $2s$-subset of columns equals $\prod_{k<\ell}(t_{j_\ell} - t_{j_k}) \neq 0$, so condition (3) of Theorem 2.1 holds.

Note: *2.14 for reference*

Theorem 2.4 (Non-uniform recovery requires only $m = s + 1$)

For any $N \geq s + 1$ and any *fixed* s-sparse $\boldsymbol{x}$, almost every $A \in \mathbb{C}^{(s+1)\times N}$ recovers $\boldsymbol{x}$ from $\boldsymbol{y} = A\boldsymbol{x}$ via P_0.

Proof. Fix an s-sparse vector $\boldsymbol{x}$ with support S. Recovery fails via P_0 if there exists $\boldsymbol{z} \in \Sigma_s$ such that $A\boldsymbol{z} = A\boldsymbol{x}$ and $\boldsymbol{z} \neq \boldsymbol{x}$. This occurs if $A(\boldsymbol{x} - \boldsymbol{z}) = \boldsymbol{0}$ for some $\boldsymbol{z} \in \Sigma_s \setminus \{\boldsymbol{x}\}$. Let $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{z}$. Then $\boldsymbol{v}$ is supported on $S \cup T$ for some T with $|T| \leq s$. The condition $A\boldsymbol{x} = A\boldsymbol{z}$ is equivalent to $A\boldsymbol{x} \in \mathrm{Range}(A_T)$.

For a fixed T with $|T| \leq s$ and $T \neq S$, the set of matrices A such that $A\boldsymbol{x} \in \mathrm{Range}(A_T)$ is defined by the vanishing of the determinant of the matrix $[A_T | A\boldsymbol{x}]$. Since $A\boldsymbol{x} = \sum_{j \in S} x_j \boldsymbol{a}_j$, the columns of $[A_T | A\boldsymbol{x}]$ are linear combinations of the columns of A. The determinant is a non-trivial polynomial in the entries of A (it is not identically zero because one can construct a matrix where $A\boldsymbol{x}$ is linearly independent of A_T columns, given $m = s + 1$).

The zero set of a non-trivial polynomial in $\mathbb{C}^{(s+1)\times N}$ has Lebesgue measure zero. Since there are only finitely many subsets $T \subseteq [N]$ with $|T| \leq s$, the exceptional set of matrices is a finite union of measure-zero sets, and thus itself has measure zero.

Note: *2.16 for reference*

□

2.2 Spark

The number "$2s$" appearing in Theorem 2.1 motivates:

Definition 2.5 (Spark, Donoho–Elad 2003)

$$\mathrm{spark}(A) := \min\{\|\boldsymbol{z}\|_0 : \boldsymbol{z} \in \mathbf{ker}\, A \setminus \{\boldsymbol{0}\}\}.$$

Theorem 2.1 can be restated: P_0 *has a unique solution for every s-sparse $\boldsymbol{x}$ iff* $\mathrm{spark}(A) > 2s$. Spark

ranges in $\{2, \dots, m+1\}$ for full-row-rank A, and equals $m+1$ a.s. for matrices with i.i.d. continuous entries.

2.3 Prony's method: practical recovery from $m = 2s$ Fourier samples

Theorem 2.6 (Prony's method)

For $N \geq 2s$, every s-sparse $\boldsymbol{x} \in \mathbb{C}^N$ can be recovered explicitly from its first $2s$ DFT coefficients $\hat{x}(0), \dots, \hat{x}(2s-1)$.

Proof. Let $S = \{j_1, \dots, j_s\} \subseteq \{0, \dots, N-1\}$ be the support of $\boldsymbol{x}$. Define the *annihilating polynomial* $p(z) = \prod_{k=1}^{s}(z - e^{2\pi i j_k/N})$. The coefficients h_k of $p(z) = \sum_{k=0}^{s} h_k z^k$ (with $h_s = 1$) satisfy:

$$\sum_{k=0}^{s} h_k e^{2\pi i j_\ell k/N} = 0 \quad \text{for each } \ell \in \{1, \dots, s\}.$$

The DFT coefficients are $\hat{x}(n) = \sum_{\ell=1}^{s} x_{j_\ell} e^{-2\pi i j_\ell n/N}$. Consider the convolution of the sequence h with $\hat{x}$:

$$\sum_{k=0}^{s} h_k \hat{x}(n-k) = \sum_{k=0}^{s} h_k \sum_{\ell=1}^{s} x_{j_\ell} e^{-2\pi i j_\ell (n-k)/N} = \sum_{\ell=1}^{s} x_{j_\ell} e^{-2\pi i j_\ell n/N} \underbrace{\sum_{k=0}^{s} h_k e^{2\pi i j_\ell k/N}}_{=0} = 0.$$

This gives a system of s linear equations for the s unknown coefficients $h_0, \dots, h_{s-1}$:

$$\begin{pmatrix} \hat{x}(s-1) & \hat{x}(s-2) & \cdots & \hat{x}(0) \\ \hat{x}(s) & \hat{x}(s-1) & \cdots & \hat{x}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}(2s-2) & \hat{x}(2s-3) & \cdots & \hat{x}(s-1) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_s \end{pmatrix} = - \begin{pmatrix} \hat{x}(s) \\ \hat{x}(s+1) \\ \vdots \\ \hat{x}(2s-1) \end{pmatrix}.$$

Once h is found, the roots of $p(z)$ reveal the support S. Finally, $\boldsymbol{x}_S$ is recovered by solving the $2s \times s$ Vandermonde system $\boldsymbol{V}_S \boldsymbol{x}_S = \boldsymbol{y}$.

Note: *2.15 for reference*

□

2.4 Computational complexity

Theorem 2.7 (NP-hardness of P_0, Natarajan 1995)

For any $\eta \geq 0$, the problem $(\mathrm{P}_{0,\eta})$ is NP-hard.

Reduction from Exact Cover by 3-Sets. Given a collection $\{C_i\}_{i=1}^{N}$ of 3-element subsets of $[m]$, set $(\boldsymbol{a}_i)_j = 1$ if $j \in C_i$, 0 otherwise; let $A = [\boldsymbol{a}_1 \mid \dots \mid \boldsymbol{a}_N]$ and $\boldsymbol{y} = \mathbf{1} \in \mathbb{R}^m$. Any $\boldsymbol{z}$ feasible for $\|A\boldsymbol{z} - \boldsymbol{y}\|_2 \leq \eta < 1$ has all entries of $A\boldsymbol{z}$ nonzero; since each column of A has exactly 3 nonzeros, $\|\boldsymbol{z}\|_0 \geq m/3$. The output $\boldsymbol{x}^*$

of $P_{0,\eta}$ has $\|\boldsymbol{x}^*\|_0 = m/3$ *iff* $\{C_j\}_{j\in\text{supp}(\boldsymbol{x}^*)}$ is an exact 3-cover. Thus P_0 would solve EXACT-COVER-3, which is NP-complete.

Note: *2.17 for reference*

□

Remark 2.8

NP-hardness pertains to general A and $\boldsymbol{y}$; for the structured A used in practice (Gaussian, partial Fourier, ETFs), tractable algorithms (greedy, ℓ_1, thresholding) succeed. The trade-off: stronger structural assumptions on A buy polynomial-time recovery.

Part II

Algorithms for Sparse Recovery

3 Sampling, Least Squares, and Greedy Pursuit

3.1 Classical sampling

Theorem 3.1 (Shannon–Nyquist)

A function $f \in L^2(\mathbb{R})$ bandlimited to $[-B, B]$ is exactly reconstructed from samples $\{f(n/(2B))\}_{n\in\mathbb{Z}}$ via

$$f(t) = \sum_{n\in\mathbb{Z}} f(n/(2B))\,\mathrm{sinc}\big(2Bt - n\big).$$

Sampling below the Nyquist rate $f_s = 2B$ produces aliasing (overlapping spectral copies in the periodic Fourier representation).

Proof. Let $f \in L^2(\mathbb{R})$ have Fourier transform $\hat{f}(\xi) = \int_{\mathbb{R}} f(t)e^{-2\pi i\xi t}dt$, supported on $[-B, B]$. We can represent $\hat{f}$ by its Fourier series on $[-B, B]$. The periodized version is $F(\xi) = \sum_{k\in\mathbb{Z}} \hat{f}(\xi - 2Bk)$, which is $2B$-periodic. Its Fourier coefficients c_n are:

$$c_n = \frac{1}{2B}\int_{-B}^{B} F(\xi)e^{-2\pi in\xi/(2B)}d\xi = \frac{1}{2B}\int_{-B}^{B} \hat{f}(\xi)e^{-2\pi in\xi/(2B)}d\xi.$$

By the inverse Fourier transform formula, $f(t) = \int_{-B}^{B} \hat{f}(\xi)e^{2\pi i\xi t}d\xi$. Comparing this with the expression for c_n, we see that $c_n = \frac{1}{2B}f(-n/(2B))$. Thus $F(\xi) = \sum_{n\in\mathbb{Z}} \frac{1}{2B}f(n/(2B))e^{2\pi in\xi/(2B)}$ on $[-B, B]$. Since $\hat{f}(\xi) = F(\xi)\cdot \mathrm{rect}(\xi/(2B))$, taking the inverse Fourier transform:

$$\begin{aligned} f(t) &= \int_{-B}^{B} \left(\sum_{n\in\mathbb{Z}} \frac{1}{2B}f(n/(2B))e^{2\pi in\xi/(2B)}\right) e^{2\pi i\xi t}d\xi \\ &= \sum_{n\in\mathbb{Z}} f(n/(2B)) \int_{-B}^{B} \frac{1}{2B}e^{2\pi i\xi(t+n/(2B))}d\xi \\ &= \sum_{n\in\mathbb{Z}} f(n/(2B))\mathrm{sinc}(2Bt + n) = \sum_{n\in\mathbb{Z}} f(n/(2B))\mathrm{sinc}(2Bt - n). \end{aligned}$$

□

CS replaces the bandlimit assumption by sparsity, enabling sub-Nyquist sampling when the signal is sparse in some basis.

3.2 Least squares geometry

For $A \in \mathbb{C}^{m\times N}$ of full column rank, the unique least-squares solution to $A\boldsymbol{x} = \boldsymbol{y}$ is

$$\boldsymbol{x}_{\text{ls}} = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 = (A^*A)^{-1}A^*\boldsymbol{y} = A^\dagger \boldsymbol{y},$$

characterized by the normal equations $A^*A\boldsymbol{x}_{\text{ls}} = A^*\boldsymbol{y}$ and the orthogonality $A^*(\boldsymbol{y} - A\boldsymbol{x}_{\text{ls}}) = 0$.

For underdetermined A ($m < N$, full row rank) the least-norm solution is $\boldsymbol{x}_2 = A^\dagger \boldsymbol{y} = A^*(AA^*)^{-1}\boldsymbol{y}$. From the SVD $A = U\Sigma V^*$, $A^\dagger = V\Sigma^\dagger U^*$ where $\Sigma^\dagger$ inverts the non-zero singular values (Appendix A).

3.3 Greedy pursuit algorithms

The combinatorial P_0 is approximated greedily: select one column (*atom*) per iteration that best correlates with the residual.

Algorithm 1 Matching Pursuit (MP)

Require: $A \in \mathbb{C}^{m\times N}$ with ℓ_2-normalized columns, $\boldsymbol{y} \in \mathbb{C}^m$, sparsity s

$\boldsymbol{r}_0 \leftarrow \boldsymbol{y}$, $\boldsymbol{x} \leftarrow \boldsymbol{0}$
for $n = 0, \dots, s-1$ **do**
  $j_{n+1} \leftarrow \arg\max_j |\langle \boldsymbol{a}_j, \boldsymbol{r}_n\rangle|$
  $c \leftarrow \langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n\rangle$
  $x_{j_{n+1}} \leftarrow x_{j_{n+1}} + c$; $\quad \boldsymbol{r}_{n+1} \leftarrow \boldsymbol{r}_n - c\,\boldsymbol{a}_{j_{n+1}}$
end for
return x

Algorithm 2 Orthogonal Matching Pursuit (OMP)

Require: $A \in \mathbb{C}^{m\times N}$, $\boldsymbol{y} \in \mathbb{C}^m$, sparsity s

$S_0 \leftarrow \emptyset$, $\boldsymbol{x}^0 \leftarrow \boldsymbol{0}$
for $n = 0, \dots, s-1$ **do**
  $j_{n+1} \leftarrow \arg\max_{j\notin S_n} |\langle \boldsymbol{a}_j, \boldsymbol{y} - A\boldsymbol{x}^n\rangle|$
  $S_{n+1} \leftarrow S_n \cup \{j_{n+1}\}$
  $\boldsymbol{x}^{n+1} \leftarrow \arg\min_{\text{supp}(\boldsymbol{z})\subseteq S_{n+1}} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$ ▷ LS re-fit
end for
return $\boldsymbol{x}^s$

Algorithm 3 Weak Matching Pursuit (WMP, parameter $\rho \in (0, 1]$)

1: Replace OMP's selection by any j_{n+1} with $|\langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n \rangle| \geq \rho \max_j |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle|$

The crucial difference between MP and OMP: OMP re-projects orthogonally after every selection, so $A_{S_n}^* \boldsymbol{r}_n = 0$ and no atom is ever selected twice.

3.4 Thresholding-based algorithms

Definition 3.2 (Hard thresholding operator)

For $\boldsymbol{z} \in \mathbb{C}^N$, $H_s(\boldsymbol{z})$ is the vector that retains the s largest-magnitude entries of $\boldsymbol{z}$ and zeros the rest. The index set of these s largest entries is denoted $L_s(\boldsymbol{z})$.

Algorithm 4 Basic Thresholding (BT)

1: $S^* \leftarrow L_s(A^* \boldsymbol{y})$
2: $\boldsymbol{x} \leftarrow \arg\min_{\mathrm{supp}(\boldsymbol{z}) \subseteq S^*} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$

Algorithm 5 Iterative Hard Thresholding (IHT)

1: $\boldsymbol{x}^0 \leftarrow \boldsymbol{0}$
2: **for** $n = 0, 1, \ldots$ **do**
3: $\quad \boldsymbol{x}^{n+1} \leftarrow H_s\big(\boldsymbol{x}^n + A^*(\boldsymbol{y} - A\boldsymbol{x}^n)\big)$
4: **end for**

Algorithm 6 Hard Thresholding Pursuit (HTP)

1: $\boldsymbol{x}^0 \leftarrow \boldsymbol{0}$
2: **for** $n = 0, 1, \ldots$ **do**
3: $\quad S^{n+1} \leftarrow L_s\big(\boldsymbol{x}^n + A^*(\boldsymbol{y} - A\boldsymbol{x}^n)\big)$
4: $\quad \boldsymbol{x}^{n+1} \leftarrow \arg\min_{\mathrm{supp}(\boldsymbol{z}) \subseteq S^{n+1}} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$
5: **end for**

Algorithm 7 CoSaMP (Compressive Sampling Matching Pursuit)

$\boldsymbol{x}^0 \leftarrow \mathbf{0}$
for $n = 0, 1, \ldots$ **do**
  $U^{n+1} \leftarrow \operatorname{supp}(\boldsymbol{x}^n) \cup L_{2s}\big(A^*(\boldsymbol{y} - A\boldsymbol{x}^n)\big)$
  $\boldsymbol{u}^{n+1} \leftarrow \arg\min_{\operatorname{supp}(\boldsymbol{z}) \subseteq U^{n+1}} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$
  $\boldsymbol{x}^{n+1} \leftarrow H_s(\boldsymbol{u}^{n+1})$
end for

Algorithm 8 Subspace Pursuit (SP)

$S_0 \leftarrow \emptyset$, $\boldsymbol{x}^0 \leftarrow \mathbf{0}$
for $n = 0, 1, \ldots$ **do**
  $U^{n+1} \leftarrow S_n \cup L_s\big(A^*(\boldsymbol{y} - A\boldsymbol{x}^n)\big)$
  $\boldsymbol{u}^{n+1} \leftarrow \arg\min_{\operatorname{supp}(\boldsymbol{z}) \subseteq U^{n+1}} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$
  $S_{n+1} \leftarrow L_s(\boldsymbol{u}^{n+1})$; $\boldsymbol{x}^{n+1} \leftarrow \arg\min_{\operatorname{supp}(\boldsymbol{z}) \subseteq S_{n+1}} \|\boldsymbol{y} - A\boldsymbol{z}\|_2$
end for

3.5 Phase transitions

Empirical performance of these algorithms is summarized by phase-transition curves in the (ρ, δ)-plane, $\rho = s/m$, $\delta = m/N$. Below each algorithm's curve, recovery succeeds with high probability; above, it fails. The Donoho–Tanner curve gives the threshold for ℓ_1-minimization in the asymptotic Gaussian setting, and is the gold standard.

4 ℓ_0 Minimization and Greedy Performance

4.1 Why ℓ_0 is hard

By Theorem 2.7, P_0 is NP-hard. Exhaustive search over supports of size s costs $\binom{N}{s}$, prohibitive even for moderate N, s. Greedy and convex-relaxation methods bypass this combinatorial barrier under structural assumptions on A.

4.2 Complexity comparison

Per iteration:

- MP: $O(mN)$ correlation, $O(m)$ residual update; total $O(smN)$.

- OMP: $O(mN)$ correlation plus an LS re-fit. Maintaining a Cholesky factor of $A_{S_n}^* A_{S_n}$ allows incremental update in $O(n^2)$, giving total $O(smN + s^3)$.
- IHT: $O(mN)$ matrix-vector product plus $O(N)$ thresholding per iteration; typically $O(\log(1/\epsilon))$ iterations.
- CoSaMP, SP: $O(mN + s^3)$ per iteration with linear convergence.

4.3 Performance theorems

Theorem 4.1 (OMP under coherence, see Lecture 5 / Section 6.3)

If the coherence $\mu(A) < \frac{1}{2s-1}$ and $\boldsymbol{x}$ is s-sparse, OMP recovers $\boldsymbol{x}$ exactly in s iterations.

Theorem 4.2 (Geometric residual decay)

Under the same hypothesis, the OMP residuals satisfy $\|\boldsymbol{r}_{n+1}\|_2^2 \leq (1-c)\|\boldsymbol{r}_n\|_2^2$ for a constant $c = c(\mu, s) > 0$.

Proof. Let S be the true support and S_n be the support after n iterations. The residual $\boldsymbol{r}_n$ is orthogonal to A_{S_n} and lies in the span of A_S. Specifically, $\boldsymbol{r}_n = A_S(\boldsymbol{x}_S - \boldsymbol{x}_S^n) = \sum_{j \in S \setminus S_n} \tilde{x}_j \boldsymbol{a}_j$. The norm of the residual satisfies:

$$\|\boldsymbol{r}_n\|_2^2 = \langle \boldsymbol{r}_n, \sum_{j \in S \setminus S_n} \tilde{x}_j \boldsymbol{a}_j \rangle = \sum_{j \in S \setminus S_n} \overline{\tilde{x}_j} \langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle \leq \|\tilde{\boldsymbol{x}}\|_1 \max_{j \in S \setminus S_n} |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle|.$$

Using the bound $\|\tilde{\boldsymbol{x}}\|_1 \leq \sqrt{s - |S_n|}\|\tilde{\boldsymbol{x}}\|_2$ and the fact that $\|\boldsymbol{r}_n\|_2^2 \geq (1 - \mu_1(s-1))\|\tilde{\boldsymbol{x}}\|_2^2$ (from Gerschgorin), we have $\|\tilde{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{r}_n\|_2 / \sqrt{1 - \mu_1(s-1)}$. Substituting:

$$\|\boldsymbol{r}_n\|_2^2 \leq \frac{\sqrt{s - |S_n|}}{\sqrt{1 - \mu_1(s-1)}} \|\boldsymbol{r}_n\|_2 \max_{j \in S} |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle|.$$

This gives $\max_{j \in S} |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle| \geq \frac{\sqrt{1-\mu_1(s-1)}}{\sqrt{s-|S_n|}} \|\boldsymbol{r}_n\|_2$. Since OMP picks $j_{n+1} = \arg\max_j |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle|$, and $\|\boldsymbol{r}_{n+1}\|_2^2 \leq \|\boldsymbol{r}_n - \langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n \rangle \boldsymbol{a}_{j_{n+1}}\|_2^2 = \|\boldsymbol{r}_n\|_2^2 - |\langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n \rangle|^2$:

$$\|\boldsymbol{r}_{n+1}\|_2^2 \leq \|\boldsymbol{r}_n\|_2^2 \left(1 - \frac{1 - \mu_1(s-1)}{s - |S_n|}\right).$$

For μ small enough, this yields the geometric decay with $c = (1 - \mu_1(s-1))/s$. □

Theorem 4.3 (CoSaMP near-optimal recovery, Needell–Tropp 2009)

If $\delta_{4s}(A) \leq 0.1$, then CoSaMP applied to $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ satisfies, for every $\boldsymbol{x} \in \mathbb{C}^N$ and noise $\|\boldsymbol{e}\|_2 \leq \eta$,

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}\|_2 + C\left(\frac{\sigma_s(\boldsymbol{x})_1}{\sqrt{s}} + \eta\right).$$

Proof. Let S be the support of the best s-term approximation of $\boldsymbol{x}$. CoSaMP iteratively updates the support Ω^{n+1} by combining the previous support with the $2s$ largest components of the proxy signal $A^*(\boldsymbol{y} - A\boldsymbol{x}^n)$. The union support $T = S \cup \Omega^{n+1} \cup \mathrm{supp}(\boldsymbol{x}^n)$ has size at most $4s$.

Restrict analysis to T. Let $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}^n$. The proxy step identifies a set J of $2s$ atoms. Using the RIP of order $4s$, one can show that the energy of $\boldsymbol{v}$ on $S \setminus \Omega^{n+1}$ is small:

$$\|\boldsymbol{x} - \boldsymbol{u}^{n+1}\|_2 \leq \mathrm{coeff} \cdot \delta_{4s} \|\boldsymbol{x} - \boldsymbol{x}^n\|_2 + \mathrm{coeff} \cdot \|\boldsymbol{e}\|_2.$$

The least-squares step on the enlarged support Ω^{n+1} (size $3s$) further reduces the error. Specifically, the solution $\boldsymbol{u}^{n+1}$ satisfies $\|\boldsymbol{u}^{n+1} - \boldsymbol{x}\|_2 \leq \frac{1+\delta_{4s}}{1-\delta_{4s}} \|\boldsymbol{x}_{T \setminus \Omega^{n+1}}\|_2$. Finally, the pruning step $\boldsymbol{x}^{n+1} = H_s(\boldsymbol{u}^{n+1})$ introduces at most a factor of 2 in the error relative to the best s-term approximation of $\boldsymbol{u}^{n+1}$. Combining these estimates, and using $\delta_{4s} \leq 0.1$, the error satisfies a contraction $\epsilon_{n+1} \leq \frac{1}{2}\epsilon_n + C\eta$. Summing the geometric series yields the result. □

4.4 Stopping rules

For known s, run s iterations. Otherwise stop when $\|\boldsymbol{r}_n\|_2 < \eta$ (matches noise level) or $|\,\|\boldsymbol{r}_n\|_2 - \|\boldsymbol{r}_{n-1}\|_2\,|$ falls below tolerance.

Part III

Theoretical Recovery Conditions

5 Spark, Null Space Property, and Basis Pursuit

5.1 The Null Space Property

Definition 5.1 (NSP)

A matrix $A \in \mathbb{C}^{m\times N}$ *satisfies the NSP relative to* $S \subseteq [N]$ if

$$\|\boldsymbol{v}_S\|_1 < \|\boldsymbol{v}_{\bar{S}}\|_1 \qquad \forall\, \boldsymbol{v} \in \ker A \setminus \{\mathbf{0}\}.$$

A satisfies the NSP of order s if it satisfies the NSP relative to every $|S| \le s$.

Note: *4.1 for reference*

Equivalent formulations:

$$\text{(i)} \quad 2\|\boldsymbol{v}_S\|_1 < \|\boldsymbol{v}\|_1 \quad \forall \boldsymbol{v} \in \ker A \setminus \{\mathbf{0}\}, \tag{5}$$

$$\text{(ii)} \quad \|\boldsymbol{v}\|_1 < 2\sigma_s(\boldsymbol{v})_1 \quad \forall \boldsymbol{v} \in \ker A \setminus \{\mathbf{0}\}. \tag{6}$$

Theorem 5.2 (NSP characterizes BP exact recovery)

A satisfies the NSP of order s if and only if every s-sparse $\boldsymbol{x} \in \mathbb{C}^N$ is the unique minimizer of P_1 with $\boldsymbol{y} = A\boldsymbol{x}$.

Proof. ($\Leftarrow$) Suppose every s-sparse vector is the unique ℓ_1-minimizer. Take $\boldsymbol{v} \in \ker A \setminus \{\mathbf{0}\}$ and any $|S| \le s$. The vector $\boldsymbol{x} = \boldsymbol{v}_S$ is s-sparse and $A\boldsymbol{x} = A\boldsymbol{x} - A\boldsymbol{v} = A(-\boldsymbol{v}_{\bar{S}})$, so $-\boldsymbol{v}_{\bar{S}}$ is feasible. By unique optimality of $\boldsymbol{x} = \boldsymbol{v}_S$, $\|\boldsymbol{v}_S\|_1 < \|\boldsymbol{v}_{\bar{S}}\|_1$.

($\Rightarrow$) Suppose NSP holds. Let $\boldsymbol{x}$ be s-sparse with $S = \mathrm{supp}(\boldsymbol{x})$, and let $\boldsymbol{z} \neq \boldsymbol{x}$ satisfy $A\boldsymbol{z} = A\boldsymbol{x}$. Then $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{z} \in \ker A \setminus \{\mathbf{0}\}$, so by NSP $\|\boldsymbol{v}_S\|_1 < \|\boldsymbol{v}_{\bar{S}}\|_1$. Hence

$$\begin{aligned}\|\boldsymbol{x}\|_1 = \|\boldsymbol{x}_S\|_1 = \|\boldsymbol{v}_S + \boldsymbol{z}_S\|_1 &\le \|\boldsymbol{v}_S\|_1 + \|\boldsymbol{z}_S\|_1 \\ &< \|\boldsymbol{v}_{\bar{S}}\|_1 + \|\boldsymbol{z}_S\|_1 = \|\boldsymbol{z}_{\bar{S}}\|_1 + \|\boldsymbol{z}_S\|_1 = \|\boldsymbol{z}\|_1,\end{aligned}$$

since $\boldsymbol{x}_{\bar{S}} = \mathbf{0}$ implies $\boldsymbol{v}_{\bar{S}} = -\boldsymbol{z}_{\bar{S}}$. Thus $\boldsymbol{x}$ is the strict ℓ_1-minimizer. □

Note: *4.4 for reference*

Theorem 5.3 (Real = complex NSP)

For $A \in \mathbb{R}^{m\times N}$, the real and complex versions of the NSP are equivalent.

Proof. (Complex NSP $\Rightarrow$ Real NSP): A real $\boldsymbol{v} \in \mathbf{ker}\, A$ is also a complex element of $\mathbf{ker}\, A$, so the complex NSP applied to it gives the real NSP inequality immediately.

(Real NSP $\Rightarrow$ Complex NSP): Let $\boldsymbol{v} \in \mathbf{ker}\, A \setminus \{0\}$ be complex. Write $v_j = r_j e^{i\theta_j}$ with $r_j = |v_j| \geq 0$. For any $\theta \in \mathbb{R}$ the rotated real vector $\boldsymbol{w}(\theta) = \Re(e^{-i\theta}\boldsymbol{v})$ has entries $w_j(\theta) = r_j \cos(\theta_j - \theta)$ and satisfies $A\boldsymbol{w}(\theta) = \mathbf{0}$ (since A is real, the real and imaginary parts of $\boldsymbol{v}$ separately lie in $\mathbf{ker}\, A$, so any real linear combination is too). The real NSP applied to $\boldsymbol{w}(\theta)$ on S gives $\sum_{j\in S} r_j|\cos(\theta_j - \theta)| < \sum_{j\in\bar{S}} r_j|\cos(\theta_j - \theta)|$ whenever $\boldsymbol{w}(\theta) \neq 0$. Integrate both sides over $\theta \in [0, 2\pi)$: by translation invariance,

$$\int_0^{2\pi} |\cos(\theta_j - \theta)|\, d\theta = 4 \quad \text{(independent of } \theta_j\text{)}.$$

So $4\sum_{j\in S} r_j \leq 4\sum_{j\in\bar{S}} r_j$, i.e. $\|\boldsymbol{v}_S\|_1 \leq \|\boldsymbol{v}_{\bar{S}}\|_1$. Strict inequality follows because the integrand inequality is strict on a set of positive measure (whenever $\boldsymbol{w}(\theta)$ is non-zero, which holds for almost every θ since $\boldsymbol{v} \neq 0$).

Note: *4.7 for reference*

□

5.2 Stable NSP and approximately sparse signals

Definition 5.4 (Stable NSP)

A satisfies the *stable NSP* of order s with constant $0 < \rho < 1$ if

$$\|\boldsymbol{v}_S\|_1 \leq \rho\, \|\boldsymbol{v}_{\bar{S}}\|_1 \qquad \forall \boldsymbol{v} \in \mathbf{ker}\, A,\ |S| \leq s.$$

Note: *4.10 for reference*

Theorem 5.5 (Stability of BP)

If A satisfies the stable NSP of order s with constant $\rho < 1$, then for every $\boldsymbol{x} \in \mathbb{C}^N$ and the BP minimizer $\boldsymbol{x}^\#$,

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

Proof. Let S index the s largest entries of $|\boldsymbol{x}|$, so $\|\boldsymbol{x}_{\bar{S}}\|_1 = \sigma_s(\boldsymbol{x})_1$. Set $\boldsymbol{v} = \boldsymbol{x}^\# - \boldsymbol{x} \in \mathbf{ker}\, A$.

Optimality of $x^\#$ gives $\|x^\#\|_1 \le \|x\|_1$. Splitting both sides by support S and the triangle inequality,

$$\begin{aligned}\|x^\#\|_1 = \|x^\#_S\|_1 + \|x^\#_{\bar S}\|_1 \;&\ge\; \|x_S\|_1 - \|v_S\|_1 + \|v_{\bar S}\|_1 - \|x_{\bar S}\|_1 \\ &= \|x\|_1 - 2\|x_{\bar S}\|_1 - \|v_S\|_1 + \|v_{\bar S}\|_1.\end{aligned}$$

Combining with $\|x^\#\|_1 \le \|x\|_1$ and rearranging,

$$\|v_{\bar S}\|_1 \;\le\; \|v_S\|_1 + 2\sigma_s(x)_1.$$

Now apply the stable NSP to $v \in \mathbf{ker}\, A$ on the set S: $\|v_S\|_1 \le \rho\|v_{\bar S}\|_1$. Substituting, $\|v_{\bar S}\|_1 \le \rho\|v_{\bar S}\|_1 + 2\sigma_s(x)_1$, so $\|v_{\bar S}\|_1 \le \frac{2}{1-\rho}\sigma_s(x)_1$. Then $\|v\|_1 = \|v_S\|_1 + \|v_{\bar S}\|_1 \le (1+\rho)\|v_{\bar S}\|_1 \le \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_1$.

Note: *4.11 for reference*

□

5.3 Robust NSP and noisy measurements

Definition 5.6 (Robust NSP)

A satisfies the *ℓ_q-robust NSP* of order s with constants $0 < \rho < 1$, $\tau > 0$ if for every $v \in \mathbb{C}^N$ and every $|S| \le s$,

$$\|v_S\|_q \le \frac{\rho}{s^{1-1/q}}\|v_{\bar S}\|_1 + \tau\,\|Av\|_2.$$

Note: *4.16/4.20 for reference*

Theorem 5.7 (BPDN under robust NSP)

Suppose A satisfies the ℓ_2-robust NSP of order s with constants $\rho < 1$, $\tau > 0$. For every $x \in \mathbb{C}^N$, e with $\|e\|_2 \le \eta$, the BPDN minimizer $x^\#$ of $y = Ax + e$ satisfies, for $1 \le p \le 2$,

$$\|x - x^\#\|_p \le \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + D\,s^{1/p-1/2}\,\eta,$$

where C, D depend only on ρ, τ.

Proof. Let $v = x^\# - x$. Since $x^\#$ is the BPDN minimizer and x is feasible, $\|x^\#\|_1 \le \|x\|_1$. Let S be the support of the s largest entries of x. Then $\|x_S + v_S\|_1 + \|x_{\bar S} + v_{\bar S}\|_1 \le \|x_S\|_1 + \|x_{\bar S}\|_1$. By the triangle inequality, $\|x_S\|_1 - \|v_S\|_1 + \|v_{\bar S}\|_1 - \|x_{\bar S}\|_1 \le \|x_S\|_1 + \|x_{\bar S}\|_1$, which rearranges to $\|v_{\bar S}\|_1 \le \|v_S\|_1 + 2\sigma_s(x)_1$.

Now apply the ℓ_2-robust NSP to v:

$$\|v_S\|_2 \le \frac{\rho}{\sqrt{s}}\|v_{\bar S}\|_1 + \tau\|Av\|_2 \le \frac{\rho}{\sqrt{s}}(\|v_S\|_1 + 2\sigma_s(x)_1) + \tau(2\eta),$$

where $\|A\boldsymbol{v}\|_2 \le \|A\boldsymbol{x}^\# - \boldsymbol{y}\|_2 + \|\boldsymbol{y} - A\boldsymbol{x}\|_2 \le 2\eta$. Using $\|\boldsymbol{v}_S\|_1 \le \sqrt{s}\|\boldsymbol{v}_S\|_2$, we obtain

$$(1-\rho)\|\boldsymbol{v}_S\|_2 \le \frac{2\rho}{\sqrt{s}}\sigma_s(\boldsymbol{x})_1 + 2\tau\eta.$$

This bounds $\|\boldsymbol{v}_S\|_2$. For $p=1$, $\|\boldsymbol{v}\|_1 = \|\boldsymbol{v}_S\|_1 + \|\boldsymbol{v}_{\bar{S}}\|_1 \le 2\|\boldsymbol{v}_S\|_1 + 2\sigma_s(\boldsymbol{x})_1 \le 2\sqrt{s}\|\boldsymbol{v}_S\|_2 + 2\sigma_s(\boldsymbol{x})_1$. For $p=2$, $\|\boldsymbol{v}_{\bar{S}}\|_2 \le \|\boldsymbol{v}_{\bar{S}}\|_1/\sqrt{s}$ by the sorting property of approximately sparse vectors. Combining these via Hölder's inequality $\|\boldsymbol{v}\|_p \le \|\boldsymbol{v}\|_1^{2/p-1}\|\boldsymbol{v}\|_2^{2-2/p}$ gives the result.

Note: *4.22 for reference*

□

5.4 Recovery of individual vectors via dual certificates

Theorem 5.8 (Dual certificate for BP)

$\boldsymbol{x} \in \mathbb{C}^N$ with support S is the unique BP minimizer iff

- A_S is injective, and
- there exists $\boldsymbol{h} \in \mathbb{C}^m$ such that $(A^*\boldsymbol{h})_j = \operatorname{sgn}(x_j)$ for $j \in S$ and $|(A^*\boldsymbol{h})_\ell| < 1$ for $\ell \notin S$.

The vector $\boldsymbol{h}$ is called a *dual certificate*.

Proof. The optimization problem $\min \|\boldsymbol{z}\|_1$ s.t. $A\boldsymbol{z} = A\boldsymbol{x}$ is convex. A feasible $\boldsymbol{x}$ is a minimizer if and only if there exists a dual vector $\boldsymbol{h} \in \mathbb{C}^m$ such that $A^*\boldsymbol{h}$ lies in the subdifferential $\partial\|\boldsymbol{x}\|_1$. The subdifferential is the set of all vectors $\boldsymbol{u}$ such that $u_j = \operatorname{sgn}(x_j)$ for $j \in S$ and $|u_j| \le 1$ for $j \notin S$.

($\Leftarrow$) Suppose such an $\boldsymbol{h}$ exists with $|(A^*\boldsymbol{h})_j| < 1$ for $j \notin S$, and A_S is injective. Let $\boldsymbol{z}$ be another feasible point, $A\boldsymbol{z} = A\boldsymbol{x}$, and let $\boldsymbol{v} = \boldsymbol{z} - \boldsymbol{x}$. Then $A\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{v} \neq \boldsymbol{0}$. Since A_S is injective, $\boldsymbol{v}$ cannot be supported on S alone, so $\boldsymbol{v}_{\bar{S}} \neq \boldsymbol{0}$. Then

$$\begin{aligned}
\|\boldsymbol{z}\|_1 = \|\boldsymbol{x}+\boldsymbol{v}\|_1 &\ge \operatorname{Re}\langle A^*\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{v}\rangle + \sum_{j\notin S}(1 - |(A^*\boldsymbol{h})_j|)|v_j| \\
&= \operatorname{Re}\langle \boldsymbol{h}, A\boldsymbol{x}\rangle + \underbrace{\operatorname{Re}\langle \boldsymbol{h}, A\boldsymbol{v}\rangle}_{=0} + \sum_{j\notin S}(1 - |(A^*\boldsymbol{h})_j|)|v_j| \\
&= \|\boldsymbol{x}\|_1 + \sum_{j\notin S}(1 - |(A^*\boldsymbol{h})_j|)|v_j|.
\end{aligned}$$

Since $|(A^*\boldsymbol{h})_j| < 1$ and $\boldsymbol{v}_{\bar{S}} \neq \boldsymbol{0}$, the last term is strictly positive, so $\|\boldsymbol{z}\|_1 > \|\boldsymbol{x}\|_1$. Thus $\boldsymbol{x}$ is the unique minimizer.

Note: *4.25 for reference*

□

Theorem 5.9 (Inexact dual certificate)

Let $\boldsymbol{x}$ be supported on S, $\|(A_S^* A_S)^{-1}\|_{2\to 2} \le \alpha$ and $\max_{\ell \notin S} \|A_S^* \boldsymbol{a}_\ell\|_2 \le \beta$. If there exists $\boldsymbol{u} = A^* \boldsymbol{h}$ with $\|\boldsymbol{u}_S - \operatorname{sgn}(\boldsymbol{x}_S)\|_2 \le \gamma$ and $\|\boldsymbol{u}_{\bar{S}}\|_\infty \le \theta$, and if $\theta + \alpha\beta\gamma < 1$, then $\boldsymbol{x}$ is the unique BP minimizer.

Proof. Let $\boldsymbol{v} \in \ker A \setminus \{0\}$. We show $\|\boldsymbol{v}_S\|_1 < \|\boldsymbol{v}_{\bar{S}}\|_1$ by first bounding $\|\boldsymbol{v}_S\|_2$. From $A\boldsymbol{v} = \boldsymbol{0}$, we have $A_S \boldsymbol{v}_S = -A_{\bar{S}} \boldsymbol{v}_{\bar{S}}$. Multiplying by A_S^*:

$$A_S^* A_S \boldsymbol{v}_S = -A_S^* A_{\bar{S}} \boldsymbol{v}_{\bar{S}} = -\sum_{\ell \notin S} v_\ell A_S^* \boldsymbol{a}_\ell.$$

Thus $\boldsymbol{v}_S = -(A_S^* A_S)^{-1} \sum_{\ell \notin S} v_\ell A_S^* \boldsymbol{a}_\ell$. Taking the ℓ_2 norm:

$$\|\boldsymbol{v}_S\|_2 \le \|(A_S^* A_S)^{-1}\|_{2\to 2} \sum_{\ell \notin S} |v_\ell| \|A_S^* \boldsymbol{a}_\ell\|_2 \le \alpha\beta \|\boldsymbol{v}_{\bar{S}}\|_1.$$

Now using the dual vector $\boldsymbol{u} = A^* \boldsymbol{h}$, we have $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{h}, A\boldsymbol{v} \rangle = 0$. So $\langle \boldsymbol{u}_S, \boldsymbol{v}_S \rangle = -\langle \boldsymbol{u}_{\bar{S}}, \boldsymbol{v}_{\bar{S}} \rangle$. Then:

$$\begin{aligned} |\langle \operatorname{sgn}(\boldsymbol{x}_S), \boldsymbol{v}_S \rangle| &= |\langle \operatorname{sgn}(\boldsymbol{x}_S) - \boldsymbol{u}_S, \boldsymbol{v}_S \rangle + \langle \boldsymbol{u}_S, \boldsymbol{v}_S \rangle| \\ &= |\langle \operatorname{sgn}(\boldsymbol{x}_S) - \boldsymbol{u}_S, \boldsymbol{v}_S \rangle - \langle \boldsymbol{u}_{\bar{S}}, \boldsymbol{v}_{\bar{S}} \rangle| \\ &\le \|\operatorname{sgn}(\boldsymbol{x}_S) - \boldsymbol{u}_S\|_2 \|\boldsymbol{v}_S\|_2 + \|\boldsymbol{u}_{\bar{S}}\|_\infty \|\boldsymbol{v}_{\bar{S}}\|_1 \\ &\le \gamma(\alpha\beta \|\boldsymbol{v}_{\bar{S}}\|_1) + \theta \|\boldsymbol{v}_{\bar{S}}\|_1 = (\theta + \alpha\beta\gamma) \|\boldsymbol{v}_{\bar{S}}\|_1. \end{aligned}$$

Since $\theta + \alpha\beta\gamma < 1$, we have $|\langle \operatorname{sgn}(\boldsymbol{x}_S), \boldsymbol{v}_S \rangle| < \|\boldsymbol{v}_{\bar{S}}\|_1$. Since $|\langle \operatorname{sgn}(\boldsymbol{x}_S), \boldsymbol{v}_S \rangle| \le \|\boldsymbol{v}_S\|_1$, and A_S is injective, this implies the unique recovery condition.

Note: *4.31 for reference*

□

5.5 Tangent cone characterization

Theorem 5.10 (Tangent cone)

For $\boldsymbol{x} \in \mathbb{R}^N$ and $A \in \mathbb{R}^{m \times N}$, define $T(\boldsymbol{x}) = \operatorname{cone}\{\boldsymbol{z} - \boldsymbol{x} : \|\boldsymbol{z}\|_1 \le \|\boldsymbol{x}\|_1\}$. Then $\boldsymbol{x}$ is the unique BP minimizer iff $\ker A \cap T(\boldsymbol{x}) = \{\boldsymbol{0}\}$.

Proof. Recall $T(\boldsymbol{x}) = \operatorname{cone}\{\boldsymbol{z} - \boldsymbol{x} : \|\boldsymbol{z}\|_1 \le \|\boldsymbol{x}\|_1\}$ consists of all positive scalar multiples of differences $\boldsymbol{z} - \boldsymbol{x}$ where $\boldsymbol{z}$ has ℓ_1-norm at most that of $\boldsymbol{x}$.

($\Rightarrow$) Suppose $\boldsymbol{x}$ is the unique BP minimizer. If $\boldsymbol{v} \in \ker A \cap T(\boldsymbol{x})$ is non-zero, then $\boldsymbol{v} = \lambda(\boldsymbol{z} - \boldsymbol{x})$ for some $\lambda > 0$ and some $\boldsymbol{z}$ with $\|\boldsymbol{z}\|_1 \le \|\boldsymbol{x}\|_1$. Since $\lambda > 0$ and $\boldsymbol{v} \neq 0$, $\boldsymbol{z} \neq \boldsymbol{x}$. Now $\boldsymbol{x} + \boldsymbol{v}/\lambda = \boldsymbol{z}$ has $A\boldsymbol{z} = A\boldsymbol{x} + \lambda^{-1} A\boldsymbol{v} = A\boldsymbol{x}$, so $\boldsymbol{z}$ is feasible. Optimality of $\boldsymbol{x}$ gives $\|\boldsymbol{z}\|_1 > \|\boldsymbol{x}\|_1$ (strict, by uniqueness), contradicting the construction. Hence $\ker A \cap T(\boldsymbol{x}) = \{0\}$.

($\Leftarrow$) Suppose $\ker A \cap T(\boldsymbol{x}) = \{0\}$. Let $\boldsymbol{z} \neq \boldsymbol{x}$ be any feasible point, $A\boldsymbol{z} = A\boldsymbol{x} = \boldsymbol{y}$. Then $\boldsymbol{v} := \boldsymbol{z} - \boldsymbol{x} \in$

ker A and $\boldsymbol{v} \neq 0$. By hypothesis $\boldsymbol{v} \notin T(\boldsymbol{x})$. The defining characterization of the descent cone says $\boldsymbol{v} \in T(\boldsymbol{x})$ iff $\|\boldsymbol{x} + t\boldsymbol{v}\|_1 \leq \|\boldsymbol{x}\|_1$ for some $t > 0$ (equivalently, $\boldsymbol{v}$ is a feasible descent direction for $\|\cdot\|_1$ at $\boldsymbol{x}$). Since $\boldsymbol{v} \notin T(\boldsymbol{x})$, $\|\boldsymbol{x} + t\boldsymbol{v}\|_1 > \|\boldsymbol{x}\|_1$ for every $t > 0$; in particular at $t = 1$, $\|\boldsymbol{z}\|_1 = \|\boldsymbol{x} + \boldsymbol{v}\|_1 > \|\boldsymbol{x}\|_1$. Hence $\boldsymbol{x}$ is the unique BP minimizer.

Note: *4.34 for reference*

□

Theorem 5.11 (Tangent-cone stability)

With $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$, $\|\boldsymbol{e}\|_2 \leq \eta$, if $\inf_{z \in T(\boldsymbol{x}),\, \|z\|_2 = 1} \|Az\|_2 \geq \tau > 0$, then the BPDN minimizer satisfies $\|\boldsymbol{x} - \boldsymbol{x}^{\#}\|_2 \leq 2\eta/\tau$.

Proof. Let $\boldsymbol{x}^{\#}$ be the BPDN minimizer. Since $\boldsymbol{x}$ is feasible for the constraint $\|A\boldsymbol{z} - \boldsymbol{y}\|_2 \leq \eta$ (assuming the noise level is correctly specified), we have $\|\boldsymbol{x}^{\#}\|_1 \leq \|\boldsymbol{x}\|_1$. By definition of the tangent cone $T(\boldsymbol{x})$, the difference $\boldsymbol{v} = \boldsymbol{x}^{\#} - \boldsymbol{x}$ lies in $T(\boldsymbol{x})$.

The distance between images satisfies:

$$\|A\boldsymbol{v}\|_2 = \|A\boldsymbol{x}^{\#} - A\boldsymbol{x}\|_2 = \|(A\boldsymbol{x}^{\#} - \boldsymbol{y}) - (A\boldsymbol{x} - \boldsymbol{y})\|_2 \leq \|A\boldsymbol{x}^{\#} - \boldsymbol{y}\|_2 + \|\boldsymbol{y} - A\boldsymbol{x}\|_2 \leq 2\eta.$$

By the hypothesis $\inf_{z \in T(\boldsymbol{x}), \|z\|_2 = 1} \|Az\|_2 \geq \tau$, it follows that for any $\boldsymbol{v} \in T(\boldsymbol{x})$, $\|A\boldsymbol{v}\|_2 \geq \tau \|\boldsymbol{v}\|_2$. Combining these inequalities:

$$\tau\|\boldsymbol{v}\|_2 \leq \|A\boldsymbol{v}\|_2 \leq 2\eta \implies \|\boldsymbol{v}\|_2 \leq \frac{2\eta}{\tau}.$$

□

Note: *4.36 for reference*

5.6 Low-rank matrix recovery

Definition 5.12 (Nuclear norm)

For $X \in \mathbb{C}^{n_1 \times n_2}$, $\|X\|_* = \sum_j \sigma_j(X)$, the ℓ_1-norm of the singular value vector.

Theorem 5.13 (Rank NSP)

A linear map $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ recovers every rank-r matrix via nuclear-norm minimization $\min \|Z\|_*$ s.t. $\mathcal{A}(Z) = \mathcal{A}(X)$ iff for every nonzero $M \in \textbf{ker}\, \mathcal{A}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n$,

$$\sum_{j=1}^{r} \sigma_j(M) < \sum_{j=r+1}^{n} \sigma_j(M).$$

Proof. This is the matrix analogue of Theorem 5.2. Let Σ_r be the set of matrices of rank at most r.

($\Rightarrow$) Suppose every rank-r matrix is the unique nuclear-norm minimizer. Let $M \in \textbf{ker}\, \mathcal{A} \setminus \{0\}$ and let

$M = U\Sigma V^*$ be its SVD. Let $M_1 = U\Sigma_r V^*$ where Σ_r contains only the r largest singular values. Then M_1 has rank at most r. Let $M_2 = M - M_1$. Since $\mathcal{A}(M) = \mathcal{A}(M_1 + M_2) = 0$, we have $\mathcal{A}(M_1) = \mathcal{A}(-M_2)$. By the unique recovery property, $\|M_1\|_* < \| - M_2\|_* = \|M_2\|_*$. Since $\|M_1\|_* = \sum_{j=1}^{r} \sigma_j(M)$ and $\|M_2\|_* = \sum_{j=r+1}^{n} \sigma_j(M)$, the condition holds.

($\Leftarrow$) Suppose the condition holds. Let X be a rank-r matrix and let Z be another matrix such that $\mathcal{A}(Z) = \mathcal{A}(X)$. Let $M = Z - X \in \mathbf{ker}\,\mathcal{A}$. By the triangle inequality for singular values (Ky Fan norms) or directly from the properties of the nuclear norm, one can show that if the condition holds, $\|X + M\|_* > \|X\|_*$ for any $M \in \mathbf{ker}\,\mathcal{A} \setminus \{0\}$. This follows from decomposing M into parts aligned and misaligned with the singular spaces of X, analogue to the support decomposition in the vector case.

Note: *4.40 for reference*

□

6 Coherence

6.1 Mutual coherence

Definition 6.1 (Coherence)

For $A \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns,

$$\mu = \mu(A) = \max_{1 \le i \ne j \le N} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|.$$

Note: *5.1 for reference*

Definition 6.2 (ℓ_1-coherence function)

For $1 \le s \le N - 1$,

$$\mu_1(s) := \max_{i \in [N]} \max_{|S| = s,\, i \notin S} \sum_{j \in S} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|.$$

Note: *5.2 for reference*

Basic bounds: $\mu \le \mu_1(s) \le s\mu$, and $\mu_1(s) \le \mu_1(s-1) + \mu$.

6.2 The Welch bound

Theorem 6.3 (Welch bound)

For $A \in \mathbb{C}^{m\times N}$ with ℓ_2-normalized columns ($m < N$),

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

Equality holds iff the columns form an Equiangular Tight Frame (ETF).

Proof. Let $G = A^*A \in \mathbb{C}^{N\times N}$. Then $\operatorname{tr}(G) = \sum_i \|\boldsymbol{a}_i\|_2^2 = N$. By Cauchy–Schwarz on traces,

$$N^2 = (\operatorname{tr} G)^2 \leq m \operatorname{tr}(G^2),$$

since G has rank $\leq m$. Now

$$\operatorname{tr}(G^2) = \sum_{i,j} |G_{ij}|^2 = \sum_i \|\boldsymbol{a}_i\|_2^4 + \sum_{i\neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|^2 = N + \sum_{i\neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|^2.$$

Since each off-diagonal term $\leq \mu^2$ and there are $N(N-1)$ of them, $\operatorname{tr}(G^2) \leq N + N(N-1)\mu^2$. Combining,

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

For $m \ll N$ the Welch bound gives $\mu \gtrsim 1/\sqrt{m}$, achieved (up to constants) by:

- **Random partial DFT**: $\mu \approx 1/\sqrt{m}$.
- **Alltop sequences (prime $m \geq 5$)**: $\mu = 1/\sqrt{m}$ for an $m \times m^2$ matrix.
- **Gabor frames** from time-frequency shifts of a window function.
- **ETFs**: achieve the Welch bound with equality (rare; require divisibility conditions).

6.3 Coherence-based recovery guarantees

Theorem 6.4 (OMP via coherence)

If $\mu_1(s) + \mu_1(s-1) < 1$, then OMP recovers every s-sparse $\boldsymbol{x}$ from $\boldsymbol{y} = A\boldsymbol{x}$ in exactly s iterations. In particular, the simpler condition $\mu(A) < \frac{1}{2s-1}$ (which implies the above) suffices.

Proof. Let $S = \operatorname{supp}(\boldsymbol{x})$ with $|S| = s$, and let $S_n \subseteq S$ be the indices selected after n iterations of OMP (we prove correctness by induction on n). At iteration n the residual is $\boldsymbol{r}_n = A\boldsymbol{x} - A\boldsymbol{x}^n$ where $\boldsymbol{x}^n$ is the LS-fit on S_n. Since OMP's LS step makes $\boldsymbol{r}_n$ orthogonal to $\{\boldsymbol{a}_j : j \in S_n\}$, we have $A_{S_n}^*\boldsymbol{r}_n = \boldsymbol{0}$. Write $\boldsymbol{r}_n = A(\boldsymbol{x} - \boldsymbol{x}^n) = \sum_{j\in S\setminus S_n} \tilde{x}_j \boldsymbol{a}_j$ for some coefficients $\tilde{x}_j$ (the residual lives in the column span of the un-selected support).

For $j \in S \setminus S_n$,

$$\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle = \tilde{x}_j + \sum_{k \in S \setminus S_n, k \neq j} \tilde{x}_k \langle \boldsymbol{a}_j, \boldsymbol{a}_k \rangle.$$

Pick $j^\star = \arg\max_{j \in S \setminus S_n} |\tilde{x}_j|$. Then $|\langle \boldsymbol{a}_{j^\star}, \boldsymbol{r}_n \rangle| \geq |\tilde{x}_{j^\star}|(1 - \mu_1(s - |S_n| - 1)) \geq |\tilde{x}_{j^\star}|(1 - \mu_1(s-1))$ by the definition of μ_1 (with $|S \setminus S_n| - 1$ summed inner products of magnitude at most $\mu_1(s-1)$).

For $\ell \notin S$, $|\langle \boldsymbol{a}_\ell, \boldsymbol{r}_n \rangle| = |\sum_{k \in S \setminus S_n} \tilde{x}_k \langle \boldsymbol{a}_\ell, \boldsymbol{a}_k \rangle| \leq \mu_1(s - |S_n|) \max_k |\tilde{x}_k| \leq \mu_1(s) |\tilde{x}_{j^\star}|$.

Hence $\max_{j \in S} |\langle \boldsymbol{a}_j, \boldsymbol{r}_n \rangle| \geq (1 - \mu_1(s-1))|\tilde{x}_{j^\star}| > \mu_1(s)|\tilde{x}_{j^\star}| \geq \max_{\ell \notin S} |\langle \boldsymbol{a}_\ell, \boldsymbol{r}_n \rangle|$ whenever $\mu_1(s) + \mu_1(s-1) < 1$. Thus OMP selects from S at iteration $n+1$. By induction $S_n \subseteq S$ for all n, so after s iterations $S_n = S$ and the LS-fit recovers $\boldsymbol{x}$ exactly.

Note: *5.14 for reference*

□

Theorem 6.5 (BP via coherence)

The same condition $\mu_1(s) + \mu_1(s-1) < 1$ implies the NSP of order s, hence ℓ_1-recovery succeeds for every s-sparse signal.

Note: *5.15 for reference*

Theorem 6.6 (BT via coherence)

If $\mu_1(s) + \mu_1(s-1) < \min_{j \in S} |x_j| / \max_{j \in S} |x_j|$, basic thresholding recovers $\boldsymbol{x}$ exactly.

Proof. Let S be the support of $\boldsymbol{x}$. Basic thresholding identifies S correctly if $\min_{j \in S} |(A^* \boldsymbol{y})_j| > \max_{\ell \notin S} |(A^* \boldsymbol{y})_\ell|$. Since $\boldsymbol{y} = \sum_{k \in S} x_k \boldsymbol{a}_k$:

$$(A^* \boldsymbol{y})_j = \langle \boldsymbol{a}_j, \sum_{k \in S} x_k \boldsymbol{a}_k \rangle = x_j + \sum_{k \in S, k \neq j} x_k \langle \boldsymbol{a}_j, \boldsymbol{a}_k \rangle.$$

For $j \in S$:

$$|(A^* \boldsymbol{y})_j| \geq |x_j| - \sum_{k \in S, k \neq j} |x_k| |\langle \boldsymbol{a}_j, \boldsymbol{a}_k \rangle| \geq \min_{k \in S} |x_k| - \mu_1(s-1) \max_{k \in S} |x_k|.$$

For $\ell \notin S$:

$$|(A^* \boldsymbol{y})_\ell| = |\sum_{k \in S} x_k \langle \boldsymbol{a}_\ell, \boldsymbol{a}_k \rangle| \leq \mu_1(s) \max_{k \in S} |x_k|.$$

Correct identification holds if:

$$\min_{k \in S} |x_k| - \mu_1(s-1) \max_{k \in S} |x_k| > \mu_1(s) \max_{k \in S} |x_k| \iff \mu_1(s) + \mu_1(s-1) < \frac{\min_{k \in S} |x_k|}{\max_{k \in S} |x_k|}.$$

Once S is identified, the LS step $\boldsymbol{x} = A_S^\dagger \boldsymbol{y}$ recovers the true coefficients exactly.

Note: *5.20 for reference*

□

6.4 The famous coherence bound

Combining $\mu_1(s) \le s\mu$ with the conditions above gives the celebrated

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right) \quad \Longleftrightarrow \quad \mu < \frac{1}{2s-1}.$$

Combined with the Welch bound $\mu \gtrsim 1/\sqrt{m}$, this yields the quadratic bottleneck:

$$m \gtrsim s^2.$$

RIP-based theory will improve this to the optimal $m \gtrsim s\log(N/s)$.

7 Restricted Isometry Property

7.1 The RIP

Definition 7.1 (Restricted Isometry Property (RIP))

The *s-th restricted isometry constant* $\delta_s = \delta_s(A)$ is the smallest $\delta \ge 0$ such that

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \le \|A\boldsymbol{x}\|_2^2 \le (1+\delta)\|\boldsymbol{x}\|_2^2 \quad \forall \boldsymbol{x} \in \Sigma_s.$$

Equivalently,

$$\delta_s = \max_{|S|\le s} \|A_S^* A_S - I\|_{2\to 2}.$$

Note: *6.1 for reference*

Monotonicity: $\delta_1 \le \delta_2 \le \cdots \le \delta_N$. Connection to coherence: $\delta_s \le (s-1)\mu(A)$.

Proposition 7.2 (Restricted orthogonality)

For disjointly supported $\boldsymbol{u} \in \Sigma_s$, $\boldsymbol{v} \in \Sigma_t$,

$$|\langle A\boldsymbol{u}, A\boldsymbol{v}\rangle| \le \delta_{s+t}\,\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2.$$

Proof. Suppose $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$. The vectors $\boldsymbol{u} \pm \boldsymbol{v}$ are $(s+t)$-sparse because their supports are contained in the union of the disjoint supports of $\boldsymbol{u}$ and $\boldsymbol{v}$. Furthermore, $\|\boldsymbol{u} \pm \boldsymbol{v}\|_2^2 = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 = 2$. Applying the RIP of order $s+t$:

$$(1-\delta_{s+t})\cdot 2 \le \|A(\boldsymbol{u} \pm \boldsymbol{v})\|_2^2 \le (1+\delta_{s+t})\cdot 2.$$

By the parallelogram identity:

$$\begin{aligned} 4\mathrm{Re}\langle A\boldsymbol{u}, A\boldsymbol{v}\rangle &= \|A(\boldsymbol{u}+\boldsymbol{v})\|_2^2 - \|A(\boldsymbol{u}-\boldsymbol{v})\|_2^2 \\ &\le 2(1+\delta_{s+t}) - 2(1-\delta_{s+t}) = 4\delta_{s+t}. \end{aligned}$$

Similarly, $4\mathrm{Re}\langle A\boldsymbol{u}, A\boldsymbol{v}\rangle \ge -4\delta_{s+t}$. The same argument applied to $i\boldsymbol{v}$ bounds the imaginary part. For general norms, rescale by $\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$. □

7.2 RIP ⇒ NSP ⇒ BP recovery

Theorem 7.3 (BPDN recovery under RIP, Candès 2008 / Foucart 2010)

If $\delta_{2s}(A) < \frac{4}{\sqrt{41}} \approx 0.6246$, then for every $\boldsymbol{x} \in \mathbb{C}^N$, $\boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \le \eta$, the BPDN solution satisfies, for $1 \le p \le 2$,

$$\|\boldsymbol{x} - \boldsymbol{x}^{\#}\|_p \le \frac{C\,\sigma_s(\boldsymbol{x})_1}{s^{1-1/p}} + D\,s^{1/p-1/2}\,\eta,$$

with C, D explicit functions of δ_{2s}. The simpler textbook condition $\delta_{2s} < \sqrt{2}-1 \approx 0.414$ suffices.

Proof. We show that RIP of order $2s$ implies the robust NSP. Let $\boldsymbol{v} \in \mathbb{C}^N$ and let S be the index set of the s largest entries of $\boldsymbol{v}$. Split the remaining indices $\bar{S}$ into blocks $S_1, S_2, \ldots$ of size s each, such that S_1 contains the s largest entries of $\boldsymbol{v}_{\bar{S}}$, S_2 the next s, and so on. By the monotonicity of sorted entries, for $j \ge 1$ and $k \in S_{j+1}$, $|v_k| \le \|\boldsymbol{v}_{S_j}\|_1/s$. Summing over $k \in S_{j+1}$ gives $\|\boldsymbol{v}_{S_{j+1}}\|_2^2 \le s(\|\boldsymbol{v}_{S_j}\|_1/s)^2$, so $\|\boldsymbol{v}_{S_{j+1}}\|_2 \le \|\boldsymbol{v}_{S_j}\|_1/\sqrt{s}$. Summing over $j \ge 1$:

$$\sum_{j\ge 2} \|\boldsymbol{v}_{S_j}\|_2 \le \frac{1}{\sqrt{s}} \sum_{j\ge 1} \|\boldsymbol{v}_{S_j}\|_1 = \frac{1}{\sqrt{s}} \|\boldsymbol{v}_{\bar{S}}\|_1.$$

Let $S_{01} = S \cup S_1$. This set has size $2s$. Applying RIP to $\boldsymbol{v}_{S_{01}}$:

$$(1-\delta_{2s})\|\boldsymbol{v}_{S_{01}}\|_2^2 \le \|A\boldsymbol{v}_{S_{01}}\|_2^2 = \langle A\boldsymbol{v}_{S_{01}}, A\boldsymbol{v} - \sum_{j\ge 2} A\boldsymbol{v}_{S_j}\rangle.$$

Using the restricted orthogonality (Proposition 7.1) between S_{01} and each S_j:

$$(1-\delta_{2s})\|\boldsymbol{v}_{S_{01}}\|_2^2 \le \|A\boldsymbol{v}_{S_{01}}\|_2\|A\boldsymbol{v}\|_2 + \sum_{j\ge 2} \delta_{2s}\|\boldsymbol{v}_{S_{01}}\|_2\|\boldsymbol{v}_{S_j}\|_2.$$

Dividing by $\|\boldsymbol{v}_{S_{01}}\|_2$ and using $\|\boldsymbol{v}_S\|_2 \le \|\boldsymbol{v}_{S_{01}}\|_2$:

$$(1-\delta_{2s})\|\boldsymbol{v}_S\|_2 \le \sqrt{1+\delta_{2s}}\|A\boldsymbol{v}\|_2 + \frac{\delta_{2s}}{\sqrt{s}}\|\boldsymbol{v}_{\bar{S}}\|_1.$$

This is the robust NSP inequality. The condition $\rho = \delta_{2s}/(1-\delta_{2s}) < 1$ is satisfied if $\delta_{2s} < 1/2$.

Note: *6.9 for reference*

□

7.3 RIP-based guarantees for thresholding and greedy algorithms

Theorem 7.4 (IHT)

If $\delta_{3s}(A) < 1/2$, IHT applied to $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ with sparsity parameter s converges to $\boldsymbol{x}$ at rate

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq \rho^n \|\boldsymbol{x}\|_2 + \tau \|\boldsymbol{e}\|_2, \quad \rho = 2\delta_{3s} < 1.$$

Proof. Let $\boldsymbol{e}^n = \boldsymbol{x}^n - \boldsymbol{x}$. The IHT update is $\boldsymbol{x}^{n+1} = H_s(\boldsymbol{x}^n + A^*(\boldsymbol{y} - A\boldsymbol{x}^n))$. Using $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$, we have:

$$\boldsymbol{x}^{n+1} = H_s(\boldsymbol{x} + (I - A^*A)\boldsymbol{e}^n + A^*\boldsymbol{e}).$$

Let $\boldsymbol{u} = \boldsymbol{x} + (I - A^*A)\boldsymbol{e}^n + A^*\boldsymbol{e}$. Both $\boldsymbol{x}$ and $\boldsymbol{x}^{n+1}$ are s-sparse. Let $T = \text{supp}(\boldsymbol{x}) \cup \text{supp}(\boldsymbol{x}^n) \cup \text{supp}(\boldsymbol{x}^{n+1})$. The size of T is at most $3s$. Since $\boldsymbol{x}^{n+1} = H_s(\boldsymbol{u})$, it is the best s-term approximation to $\boldsymbol{u}$. Thus $\|\boldsymbol{u} - \boldsymbol{x}^{n+1}\|_2 \leq \|\boldsymbol{u} - \boldsymbol{x}\|_2$. Using the triangle inequality:

$$\|\boldsymbol{x}^{n+1} - \boldsymbol{x}\|_2 \leq \|\boldsymbol{x}^{n+1} - \boldsymbol{u}\|_2 + \|\boldsymbol{u} - \boldsymbol{x}\|_2 \leq 2\|\boldsymbol{u} - \boldsymbol{x}\|_2.$$

Restricting to the support T:

$$\begin{aligned}\|\boldsymbol{e}^{n+1}\|_2 &\leq 2\|((I - A^*A)\boldsymbol{e}^n + A^*\boldsymbol{e})_T\|_2 \\ &\leq 2\|(I_T - A_T^*A_T)\boldsymbol{e}^n\|_2 + 2\|A_T^*\boldsymbol{e}\|_2 \\ &\leq 2\delta_{3s}\|\boldsymbol{e}^n\|_2 + 2\sqrt{1+\delta_{3s}}\|\boldsymbol{e}\|_2.\end{aligned}$$

If $\delta_{3s} < 1/2$, then $\rho = 2\delta_{3s} < 1$, and the error converges geometrically to a noise-dependent floor.

Note: *6.18 for reference*

□

Theorem 7.5 (HTP)

If $\delta_{3s}(A) < 1/\sqrt{3} \approx 0.577$, HTP recovers every s-sparse signal from noiseless measurements in at most s iterations, and is robust under noise.

Proof. Let S be the true support and S^{n+1} be the support identified at iteration n. The HTP algorithm combines a thresholding step to identify a support and a least squares step to estimate coefficients on that support. In the noiseless case:

$$\boldsymbol{x}^{n+1} = \text{argmin}_{\text{supp}(z) \subseteq S^{n+1}} \|Az - A\boldsymbol{x}\|_2 = A^\dagger_{S^{n+1}} A\boldsymbol{x}.$$

Let $T = S \cup S^n \cup S^{n+1}$. Then $|T| \leq 3s$. By analyzing the least squares residual on T, one can show:

$$\|\boldsymbol{x}^{n+1} - \boldsymbol{x}\|_2 \leq \sqrt{\frac{3\delta_{3s}^2}{1 - \delta_{3s}^2}} \|\boldsymbol{x}^n - \boldsymbol{x}\|_2.$$

If $\delta_{3s} < 1/\sqrt{3}$, the contraction factor $\rho = \sqrt{3\delta_{3s}^2/(1-\delta_{3s}^2)} < 1$. Since there are only $\binom{N}{s}$ possible supports, and each iteration with a change must strictly reduce the error, the support must eventually stabilize. Once the correct support is found ($S^{n+1} = S$), the LS step yields $\boldsymbol{x}^{n+1} = \boldsymbol{x}$ immediately.

Note: *6.21 for reference*

□

Theorem 7.6 (CoSaMP)

If $\delta_{4s}(A) < (\sqrt{11/3}-1)/2 \approx 0.4782$, CoSaMP achieves $\|\boldsymbol{x}-\boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}\|_2 + \tau\eta$, where $\eta = \|\boldsymbol{e}\|_2$.

Proof. The proof proceeds in four steps. Let $\boldsymbol{x}$ be s-sparse and $\boldsymbol{v}^n = \boldsymbol{x} - \boldsymbol{x}^n$. *Step 1: Signal proxy.* The identified atoms $J = L_{2s}(A^*\boldsymbol{r}_n)$ contain most of the energy of $\boldsymbol{v}^n$. By RIP: $\|\boldsymbol{v}_J^n\|_2 \leq 0.5\delta_{4s}\|\boldsymbol{v}^n\|_2 + C\eta$. *Step 2: LS estimation.* On the enlarged support $\Omega = \operatorname{supp}(\boldsymbol{x}^n) \cup J$, the LS solution $\boldsymbol{u}^{n+1}$ satisfies: $\|\boldsymbol{u}^{n+1} - \boldsymbol{x}\|_2 \leq \frac{1}{\sqrt{1-\delta_{4s}^2}}\|\boldsymbol{x} - \boldsymbol{x}_\Omega\|_2 + C'\eta$. *Step 3: Pruning.* $\boldsymbol{x}^{n+1} = H_s(\boldsymbol{u}^{n+1})$. By the triangle inequality: $\|\boldsymbol{x}^{n+1} - \boldsymbol{x}\|_2 \leq 2\|\boldsymbol{u}^{n+1} - \boldsymbol{x}\|_2$. Combining steps: $\|\boldsymbol{x}^{n+1} - \boldsymbol{x}\|_2 \leq \rho\|\boldsymbol{x}^n - \boldsymbol{x}\|_2 + \tau\eta$. For $\delta_{4s} < 0.4782$, $\rho < 0.5$. Iterating gives geometric convergence.

Note: *6.27 for reference*

□

Theorem 7.7 (OMP via RIP, Zhang 2011)

If $\delta_{31s}(A) < 1/3$, OMP recovers every s-sparse $\boldsymbol{x}$ in $30s$ iterations.

Proof. Standard OMP analysis under coherence requires $\mu(A) < 1/(2s-1)$. Under RIP, Zhang (2011) used a potential function argument. Let $\boldsymbol{r}_n$ be the residual after n steps. The energy decrease is $\|\boldsymbol{r}_n\|_2^2 - \|\boldsymbol{r}_{n+1}\|_2^2 = |\langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n\rangle|^2$. By picking the atom with largest correlation, we ensure:

$$|\langle \boldsymbol{a}_{j_{n+1}}, \boldsymbol{r}_n\rangle|^2 \geq \frac{1-\delta_{s+|S_n|}}{s}\|\boldsymbol{r}_n\|_2^2.$$

However, OMP might pick "wrong" atoms. Zhang showed that for $\delta_{31s} < 1/3$, the number of times OMP can pick an atom outside the true support S is bounded. Suppose n steps have been taken and $S \not\subseteq S_n$. The residual $\boldsymbol{r}_n$ is still large. Using restricted orthogonality, one can bound the maximum correlation outside S relative to that inside S. A counting argument shows that after at most $30s$ steps, all s true atoms must have been selected. Once $S \subseteq S_n$, the LS step ensures $\boldsymbol{r}_n = \boldsymbol{0}$ and exact recovery is achieved. □

Remark 7.8 (The RIP threshold table)

Algorithm	RIP order	Threshold
Basis Pursuit (sharp)	δ_{2s}	$< 4/\sqrt{41} \approx 0.6246$
Basis Pursuit (textbook)	δ_{2s}	$< \sqrt{2} - 1 \approx 0.414$
Iterative Hard Thresholding	δ_{3s}	$< 1/2$
Hard Thresholding Pursuit	δ_{3s}	$< 1/\sqrt{3} \approx 0.577$
CoSaMP	δ_{4s}	< 0.478
Orthogonal Matching Pursuit	δ_{31s}	$< 1/3$

The greedy bound (δ_{31s}) is uniform and conservative; coherence-based analysis gives sharper guarantees in the high-coherence regime.

7.4 Information-theoretic lower bound

Theorem 7.9 (Information-theoretic lower bound on m)

If A satisfies $\delta_{2s}(A) < 1/2$ for some $A \in \mathbb{C}^{m\times N}$ then $m \geq C\, s \log(N/s)$ for an absolute constant $C > 0$.

Proof. For each subset $S \subseteq [N]$ with $|S| = s$ and each sign pattern $\boldsymbol{\sigma} \in \{\pm 1\}^s$, form the vector $\boldsymbol{x}_{S,\boldsymbol{\sigma}} \in \mathbb{R}^N$ with $\boldsymbol{\sigma}$ on S and 0 elsewhere. The collection

$$\mathcal{P} \;=\; \{\boldsymbol{x}_{S,\boldsymbol{\sigma}} \,:\, |S| = s,\ \boldsymbol{\sigma} \in \{\pm 1\}^s\}$$

has cardinality $|\mathcal{P}| = \binom{N}{s} 2^s$, and any two distinct elements $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{P}$ satisfy $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq 2s$ and $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \sqrt{2}$. (If they share a support and differ in k signs, $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 = 2\sqrt{k} \geq 2$; if their supports differ, then some coordinate is ± 1 in one and 0 in the other, contributing at least 1, and there are at least two such coordinates by symmetry.)

For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{P}$, the difference $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}'$ is $2s$-sparse. By RIP with $\delta_{2s} < 1/2$,

$$\|A\boldsymbol{v}\|_2 \;\geq\; \sqrt{1 - \delta_{2s}}\, \|\boldsymbol{v}\|_2 \;>\; \tfrac{1}{\sqrt{2}} \|\boldsymbol{v}\|_2 \;\geq\; 1.$$

Conversely $\|A\boldsymbol{x}\|_2 \leq \sqrt{1 + \delta_{2s}}\sqrt{s} < \sqrt{3s/2}$, so all images $\{A\boldsymbol{x} : \boldsymbol{x} \in \mathcal{P}\}$ lie in the ball of radius $\sqrt{3s/2}$ in $\mathbb{R}^m$ and are pairwise at distance > 1. Hence the images form a $\frac{1}{2}$-packing of that ball.

If K disjoint open balls of radius r fit inside a ball of radius R in $\mathbb{R}^m$, then $K\, r^m \leq R^m$, so $K \leq (R/r)^m$. With $r = 1/2$ and $R \leq \sqrt{3s/2} + 1/2 \leq 2\sqrt{s}$ (for $s \geq 1$), we get

$$\binom{N}{s} 2^s \;=\; |\mathcal{P}| \;\leq\; (4\sqrt{s})^m.$$

Taking logarithms and using $\binom{N}{s} \geq (N/s)^s$,

$$s\,\log(N/s) + s\log 2 \;\leq\; m\,\log(4\sqrt{s}) \;=\; m\big(\log 4 + \tfrac{1}{2}\log s\big)\,.$$

Provided $s \leq N/2$ (otherwise the bound is trivial), the left side is at least $\frac{s}{2}\log(N/s)$ and the right side is $O(m \log s)$; dividing gives $m \geq C\, s \log(N/s)/\log s$. The sharper bound without the $\log s$ factor follows from a standard refinement: replace $\mathcal{P}$ with a Gilbert–Varshamov subcode of $\{\pm 1, 0\}^N$ of cardinality $\binom{N}{s}/4$ all of whose pairwise differences have weight $\geq s$, recovering $m \geq C\, s \log(N/s)$. □

This matches the upper bound proved next, so $m \asymp s \log(N/s)$ is optimal.

Part IV

Random Matrices, Sensing, and Algorithms

8 Random Matrices and Sensing-Matrix Design

8.1 Subgaussian random matrices

Definition 8.1 (Subgaussian random variable)

X is *subgaussian* with parameter $c > 0$ if $\mathbb{E}\, e^{\theta X} \le e^{c\theta^2}$ for all $\theta \in \mathbb{R}$. Examples: Gaussian $\mathcal{N}(0,1)$ ($c = 1/2$), Rademacher (uniform on $\{\pm 1\}$, $c = 1/2$), bounded $|X| \le K$ ($c = K^2/2$).

Theorem 8.2 (RIP for subgaussian matrices)

Let $A \in \mathbb{R}^{m\times N}$ have i.i.d. mean-zero subgaussian entries with variance $1/m$. There exists a universal constant $C > 0$ depending only on the subgaussian parameter such that, for any $\delta, \varepsilon \in (0,1)$, if

$$m \ge C\delta^{-2}\big(s\,\ln(eN/s) + \ln(2/\varepsilon)\big),$$

then with probability $\ge 1-\varepsilon$, $\delta_s(A) \le \delta$.

Proof. The proof follows a standard concentration-and-covering strategy. *Step 1: Concentration for a fixed vector.* For any $\boldsymbol{x} \in \mathbb{R}^N$, the random variable $\|A\boldsymbol{x}\|_2^2 = \frac{1}{m}\sum_{i=1}^m |(A\sqrt{m}\boldsymbol{x})_i|^2$ is a sum of independent subgaussian squares. By Bernstein's inequality:

$$\mathbf{Pr}\Big(\big|\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\big| \ge t\|\boldsymbol{x}\|_2^2\Big) \le 2e^{-ct^2 m}.$$

Step 2: Covering the s-sparse sphere. Fix a support S with $|S| = s$. Let $B_S = \{\boldsymbol{z} \in \mathbb{R}^N : \mathrm{supp}(\boldsymbol{z}) \subseteq S, \|\boldsymbol{z}\|_2 \le 1\}$. A ρ-net $\mathcal{U}_S$ for B_S exists with $|\mathcal{U}_S| \le (1+2/\rho)^s$. Pick $\rho = \delta/4$. By a union bound over $\mathcal{U}_S$:

$$\mathbf{Pr}\Big(\max_{\boldsymbol{u}\in\mathcal{U}_S}\big|\|A\boldsymbol{u}\|_2^2 - 1\big| \ge \delta/2\Big) \le 2(1+8/\delta)^s e^{-c\delta^2 m/4}.$$

Step 3: Extending to the whole sphere. If the bound holds on the net, then for any $\boldsymbol{x} \in B_S$, let $\boldsymbol{u} \in \mathcal{U}_S$ be the closest point. One can show $|\langle A\boldsymbol{x}, A\boldsymbol{x}\rangle - 1| \le \frac{1}{1-2\rho}\max_{\boldsymbol{u}\in\mathcal{U}}|\|A\boldsymbol{u}\|_2^2 - 1| \le \delta$. *Step 4: Union bound over supports.* There are $\binom{N}{s} \le (eN/s)^s$ supports. Total probability of failure:

$$P_{fail} \le (eN/s)^s \cdot 2(1+8/\delta)^s e^{-c\delta^2 m/4} \le 2e^{s(\ln(eN/s)+\ln(9))-c\delta^2 m/4}.$$

Setting this $\le \varepsilon$ and solving for m yields the required bound. □

Theorem 8.3 (Theorem 8.2 ⇒ uniform recovery)

Under the same hypothesis, with the same probability A allows uniform recovery of every s-sparse signal via BP, IHT, HTP, or CoSaMP, by combining with the respective RIP thresholds (Section 7).

Proof. Each algorithm in the list has a deterministic recovery guarantee that holds whenever an RIP constant of suitable order is below an explicit threshold:

- BP succeeds for every s-sparse signal whenever $\delta_{2s} < \sqrt{2} - 1$ (Theorem 6.9).
- IHT converges with $\delta_{3s} < 1/2$ (Theorem 6.18).
- HTP converges with $\delta_{3s} < 1/\sqrt{3}$ (Theorem 6.21).
- CoSaMP recovers up to a constant factor of $\sigma_s(\boldsymbol{x})_1$ when $\delta_{4s} < 0.4782$ (Theorem 6.27).

Fix the algorithm of interest and let (δ^*, k^*) be the corresponding threshold/order pair from this list. Pick any $\delta < \delta^*$.

Apply Theorem 8.2 with sparsity level k^*s and tolerance δ: the bound becomes

$$m \;\geq\; C\,\delta^{-2}\big(k^*s\,\ln(eN/(k^*s)) + \ln(2/\varepsilon)\big),$$

i.e. the same shape with $s \mapsto k^*s$. With probability $\geq 1 - \varepsilon$ the realized matrix A then satisfies $\delta_{k^*s}(A) \leq \delta < \delta^*$, so its restricted-isometry constant lies in the recovery regime of the chosen algorithm. The deterministic recovery theorem applies pointwise to A and yields uniform recovery of every s-sparse signal on the same probability event. □

8.2 Gaussian matrices: explicit constants

Theorem 8.4 (Gaussian basis pursuit, explicit constants)

Let $A \in \mathbb{R}^{m\times N}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries. For any $\rho \in (0,1)$ and $\varepsilon \in (0,1)$, if

$$\frac{m^2}{m+1} \geq 2s\left[\rho^{-1} + D(s/N) + \sqrt{\frac{\ln(\varepsilon^{-1})}{s\ln(eN/s)}}\right]^2,$$

where $D(\alpha)$ is bounded by 2.05 for $\alpha \in (0,1)$, then with probability $\geq 1 - \varepsilon$, BP recovers every $\boldsymbol{x}$ from $\boldsymbol{y} = A\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}^\#\|_1 \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(\boldsymbol{x})_1$. In particular, setting $\rho = 1$, $m \geq 8s\log(eN/s)$ suffices for exact uniform recovery.

Proof. This proof utilizes Gordon's *Escape-through-the-mesh* theorem. Recovery via BP is exact if $\mathbf{ker}\, A \cap T(\boldsymbol{x}) = \{\mathbf{0}\}$, where $T(\boldsymbol{x})$ is the tangent cone of the ℓ_1-norm at $\boldsymbol{x}$. For Gaussian A, this holds with high probability if $m \geq w(T(\boldsymbol{x}) \cap S^{N-1})^2 + 1$. The Gaussian width w of the intersection of the cone with

the sphere can be bounded using the distance to the dual cone (the normal cone $N_{\|\cdot\|_1}(\boldsymbol{x})$):

$$w(T(\boldsymbol{x}) \cap S^{N-1})^2 \le \mathbb{E}\mathrm{dist}(\boldsymbol{g}, N_{\|\cdot\|_1}(\boldsymbol{x}))^2.$$

For an s-sparse vector $\boldsymbol{x}$, the normal cone consists of vectors $\boldsymbol{z}$ with $z_j = \lambda \mathrm{sgn}(x_j)$ for $j \in S$ and $|z_j| \le \lambda$ for $j \notin S$. Optimizing over λ and summing independent Gaussian components yields the $w^2 \approx 2s\ln(N/s)$ estimate. The stable NSP extension follows by considering the larger cone $T_{\rho,s} = \{\boldsymbol{v} : \|\boldsymbol{v}_S\|_1 \ge \rho\|\boldsymbol{v}_{\bar{S}}\|_1\}$, whose width is slightly larger but of the same order. Concentration of the singular values of A completes the probability bound. □

8.3 Singular-value concentration of Gaussian matrices

Theorem 8.5 (Singular-value concentration for Gaussian matrices)

For $A \in \mathbb{R}^{m\times s}$ Gaussian with $m > s$, the singular values $\sigma_{\min}, \sigma_{\max}$ of $\frac{1}{\sqrt{m}}A$ satisfy, for $t > 0$,

$$\mathbf{Pr}(\sigma_{\max} \ge 1 + \sqrt{s/m} + t) \le e^{-mt^2/2},$$
$$\mathbf{Pr}(\sigma_{\min} \le 1 - \sqrt{s/m} - t) \le e^{-mt^2/2}.$$

Proof. Write $G = A$ and view the entries of G as a vector $g \in \mathbb{R}^{ms}$ with $g \sim \mathcal{N}(0, I_{ms})$. The maps

$$F_+(g) \;=\; \sigma_{\max}(G), \qquad F_-(g) \;=\; -\sigma_{\min}(G)$$

are 1-Lipschitz with respect to the Frobenius (Euclidean) norm on g, because for any unit $\boldsymbol{x}$ and any perturbation $G \mapsto G + H$,

$$\big|\|(G+H)\boldsymbol{x}\|_2 - \|G\boldsymbol{x}\|_2\big| \;\le\; \|H\boldsymbol{x}\|_2 \;\le\; \|H\|_F,$$

and singular values are extrema of $\boldsymbol{x} \mapsto \|G\boldsymbol{x}\|_2$.

Apply the Gaussian Lipschitz concentration inequality (Borell–TIS): for any 1-Lipschitz F on $\mathbb{R}^d$ with $g \sim \mathcal{N}(0, I_d)$, $\mathbf{Pr}(F(g) \ge \mathbb{E}F + t) \le e^{-t^2/2}$ for all $t > 0$. This gives

$$\mathbf{Pr}(\sigma_{\max} \ge \mathbb{E}\sigma_{\max} + t) \le e^{-t^2/2}, \qquad \mathbf{Pr}(\sigma_{\min} \le \mathbb{E}\sigma_{\min} - t) \le e^{-t^2/2}.$$

It remains to bound the two means. By the Gordon–Slepian comparison theorem, for the Gaussian process $X_{\boldsymbol{x},\boldsymbol{y}} = \langle G\boldsymbol{x}, \boldsymbol{y}\rangle$ on $S^{s-1} \times S^{m-1}$,

$$\mathbb{E}\sup_{\boldsymbol{x},\boldsymbol{y}} X_{\boldsymbol{x},\boldsymbol{y}} \;\le\; \mathbb{E}\sup_{\boldsymbol{x},\boldsymbol{y}}\big(\langle \boldsymbol{g}, \boldsymbol{x}\rangle + \langle \boldsymbol{h}, \boldsymbol{y}\rangle\big) \;=\; \mathbb{E}\|\boldsymbol{g}\|_2 + \mathbb{E}\|\boldsymbol{h}\|_2 \;\le\; \sqrt{s} + \sqrt{m},$$

with independent $\boldsymbol{g} \sim \mathcal{N}(0, I_s)$, $\boldsymbol{h} \sim \mathcal{N}(0, I_m)$. The left side equals $\mathbb{E}\,\sigma_{\max}(G)$. A symmetric Gordon argument using the min-max representation $\sigma_{\min} = \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \langle G\boldsymbol{x}, \boldsymbol{y}\rangle$ (with the signs of $\boldsymbol{x}, \boldsymbol{y}$ exchanged) yields $\mathbb{E}\,\sigma_{\min}(G) \ge \sqrt{m} - \sqrt{s}$.

Combining the mean bounds with the deviation bounds and rescaling $G \mapsto \frac{1}{\sqrt{m}}G$ (which divides every singular value by $\sqrt{m}$ and the deviation t becomes $t\sqrt{m}$) gives

$$\mathbf{Pr}\big(\sigma_{\max}(\tfrac{1}{\sqrt{m}}G) \geq 1 + \sqrt{s/m} + t\big) \leq e^{-mt^2/2},$$

and analogously for $\sigma_{\min}$.

Note: *9.24 for reference*

□

8.4 Johnson–Lindenstrauss embeddings and the JL⇔RIP duality

Theorem 8.6 (Johnson–Lindenstrauss lemma)

For any finite set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\} \subset \mathbb{R}^N$ and $\delta \in (0, 1/2)$, if $m \geq C\delta^{-2} \ln M$ (universal C), there exists a linear map $f : \mathbb{R}^N \to \mathbb{R}^m$ (e.g., a Gaussian random projection) such that

$$(1-\delta)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \leq \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|_2 \leq (1+\delta)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \quad \forall i, j.$$

Proof. Let $A \in \mathbb{R}^{m \times N}$ have i.i.d. $\mathcal{N}(0,1)$ entries and put $f(\boldsymbol{x}) = \frac{1}{\sqrt{m}}A\boldsymbol{x}$. By linearity it suffices to show that for any fixed $\boldsymbol{u} \in \mathbb{R}^N$,

$$\mathbf{Pr}\Big((1-\delta)\|\boldsymbol{u}\|_2 \leq \|f(\boldsymbol{u})\|_2 \leq (1+\delta)\|\boldsymbol{u}\|_2\Big) \;\geq\; 1 - 2e^{-cm\delta^2},$$

for an absolute constant $c > 0$.

By homogeneity assume $\|\boldsymbol{u}\|_2 = 1$. Then $A\boldsymbol{u} \sim \mathcal{N}(0, I_m)$, so $\|A\boldsymbol{u}\|_2^2$ is a χ^2_m random variable with mean m. Standard χ^2-concentration (Laurent–Massart) gives, for $t \in (0,1)$,

$$\mathbf{Pr}\Big(\big|\tfrac{1}{m}\|A\boldsymbol{u}\|_2^2 - 1\big| \geq t\Big) \;\leq\; 2\exp\!\big(-mt^2/8\big).$$

Setting $t = \delta(2-\delta)/(1+\delta) \geq \delta/2$ converts the squared-norm deviation into the desired norm deviation: if $|\,m^{-1}\|A\boldsymbol{u}\|_2^2 - 1| \leq t$ then $1-\delta \leq m^{-1/2}\|A\boldsymbol{u}\|_2 \leq 1+\delta$.

Apply this to each difference vector $\boldsymbol{u}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$. There are $\binom{M}{2} \leq M^2/2$ such pairs, so a union bound gives a failure probability of at most $M^2 e^{-cm\delta^2}$. Choosing $m \geq C\delta^{-2}\ln M$ with C large enough makes this < 1, proving existence. □

Theorem 8.7 (Krahmer–Ward 2011 – RIP ⇒ JL)

If A has $\delta_{2s}(A) \leq \eta/4$ for some $s \geq 16\ln(4M/\varepsilon)$, then the randomized matrix AD_ϵ, where $D_\epsilon = \mathrm{diag}(\epsilon_j)$ with $\epsilon_j \in \{\pm 1\}$ random, satisfies the JL bound on E with probability $\geq 1 - \varepsilon$.

Proof. Fix $\boldsymbol{x} \in E$ with $\|\boldsymbol{x}\|_2 = 1$. Partition the coordinates of $\boldsymbol{x}$ into blocks $S_0, S_1, \dots$ of size s, where S_0 indexes the s largest entries by magnitude, S_1 the next s, and so on. Write $\boldsymbol{x}_k = \boldsymbol{x}_{S_k}$ and $\boldsymbol{y}_k = AD_\epsilon \boldsymbol{x}_k =$

$A\boldsymbol{x}_k \odot \boldsymbol{\epsilon}_{S_k}$ (with the abusive notation that the sign acts on the columns).

Expand the squared norm

$$\|AD_\epsilon \boldsymbol{x}\|_2^2 \;=\; \sum_k \|A\boldsymbol{x}_k\|_2^2 \;+\; 2\sum_{k<\ell} \big\langle AD_\epsilon \boldsymbol{x}_k,\, AD_\epsilon \boldsymbol{x}_\ell \big\rangle.$$

The diagonal terms are deterministic and well-controlled: each $\boldsymbol{x}_k$ is s-sparse, so RIP gives $(1-\delta_{2s})\|\boldsymbol{x}_k\|_2^2 \le \|A\boldsymbol{x}_k\|_2^2 \le (1+\delta_{2s})\|\boldsymbol{x}_k\|_2^2$. Summing over k and using $\sum_k \|\boldsymbol{x}_k\|_2^2 = \|\boldsymbol{x}\|_2^2 = 1$,

$$1-\delta_{2s} \;\le\; \sum_k \|A\boldsymbol{x}_k\|_2^2 \;\le\; 1+\delta_{2s}.$$

For the cross-terms, fix $k<\ell$ and write the inner product as a Rademacher chaos in the signs $\boldsymbol{\epsilon}$:

$$\big\langle AD_\epsilon \boldsymbol{x}_k, AD_\epsilon \boldsymbol{x}_\ell \big\rangle \;=\; \sum_{i\in S_k}\sum_{j\in S_\ell} \epsilon_i\epsilon_j\, x_i x_j\, \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle.$$

This is a centered second-order Rademacher chaos $\boldsymbol{\epsilon}^\top M\boldsymbol{\epsilon}$ with off-diagonal matrix $M_{ij} = x_i x_j \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle$ supported on $S_k \times S_\ell$. The Hanson–Wright inequality bounds its tail by $\mathbf{Pr}(|\boldsymbol{\epsilon}^\top M\boldsymbol{\epsilon}| \ge t) \le 2\exp\!\big(-c\min(t^2/\|M\|_F^2,\, t/\|M\|_{op})\big)$. Both norms are estimable via the RIP of A on $S_k \cup S_\ell$ (size $\le 2s$): $\|M\|_{op} \le \delta_{2s}\|\boldsymbol{x}_k\|_2\|\boldsymbol{x}_\ell\|_2$ and $\|M\|_F^2 \le \delta_{2s}^2\|\boldsymbol{x}_k\|_2^2\|\boldsymbol{x}_\ell\|_2^2$, since the cross-Gram matrix $A_{S_k}^* A_{S_\ell}$ has operator norm $\le \delta_{2s}$.

Summing the chaos bounds over the $\binom{N/s}{2}$ pairs (k,ℓ) and using Cauchy–Schwarz to control $\sum_{k<\ell}\|\boldsymbol{x}_k\|_2\|\boldsymbol{x}_\ell\|_2 \le 1$, the total cross-term contribution is bounded by $C\delta_{2s}\eta$ in absolute value with probability $\ge 1-2e^{-cs}$ for $s \ge 16\ln(4M/\varepsilon)$.

Combining diagonal and off-diagonal estimates, $\big|\|AD_\epsilon \boldsymbol{x}\|_2^2 - 1\big| \le \eta$ with probability $\ge 1-2e^{-cs}$. A union bound over the $|E| = M$ points gives failure probability $\le 2Me^{-cs} \le \varepsilon$ under the stated s-bound, proving the JL guarantee. □

This says: *RIP and JL are essentially the same property*; one implies the other up to a column-sign randomization.

8.5 Fast and structured constructions

Dense Gaussian matrices cost $O(mN)$ to apply. Faster alternatives:

- **Fast JL Transform (Ailon–Chazelle 2006)**: $A =$ (sparse Gaussian) $\cdot$ (Hadamard) $\cdot$ (random sign flips), applied in $O(N\log m)$ time, achieves $m = O(\delta^{-2}\log P)$ JL embedding.
- **Structurally Random Matrices**: $A = DFR$, R a randomizer (sign flips or permutation), F a fast transform (FFT, Walsh–Hadamard, DCT), D a random down-sampler. Cost $O(N\log N)$ with RIP guarantees $m = O(s\log^c N)$.
- **Partial Fourier matrices**: pick m rows uniformly at random from the $N \times N$ DFT. Then RIP with

$m \geq Cs\log^4 N$ holds with high probability (

Note: *12.31 for reference*

, restricted-isometry result for bounded orthonormal systems).

- **2D-SME**: separable measurement ensembles for images. $Y = D_r F R X R^\top F^\top D_c^\top$, sampling rows and columns separately.

8.6 Non-uniform recovery

Theorem 8.8 (Gaussian non-uniform recovery via BP)

Fix an s-sparse $\boldsymbol{x} \in \mathbb{R}^N$. For $\varepsilon \in (0,1)$, if

$$\frac{m^2}{m+1} \geq 2s\Big[\sqrt{\ln(2.34N/s)} + \sqrt{\ln(\varepsilon^{-1})/s}\Big]^2,$$

then with probability $\geq 1-\varepsilon$, BP recovers $\boldsymbol{x}$ from $\boldsymbol{y} = A\boldsymbol{x}$.

Proof. By the tangent-cone characterization (Theorem 4.34), $\boldsymbol{x}$ is the unique BP solution of $A\boldsymbol{z} = A\boldsymbol{x}$ iff

$$\mathbf{ker}\, A \cap T(\boldsymbol{x}) \;=\; \{\mathbf{0}\},$$

where $T(\boldsymbol{x})$ is the tangent cone of the ℓ_1-ball at $\boldsymbol{x}/\|\boldsymbol{x}\|_1$ (equivalently, the cone of feasible descent directions for the ℓ_1-norm at $\boldsymbol{x}$).

Let $E = T(\boldsymbol{x}) \cap S^{N-1}$. Gordon's theorem gives, for $A \in \mathbb{R}^{m\times N}$ with i.i.d. $\mathcal{N}(0,1)$ entries,

$$\mathbb{E}\inf_{\boldsymbol{v}\in E}\|A\boldsymbol{v}\|_2 \;\geq\; \lambda_m - w(E), \qquad \lambda_m \;=\; \mathbb{E}\|\boldsymbol{g}\|_2 \;=\; \sqrt{2}\,\Gamma(\tfrac{m+1}{2})/\Gamma(\tfrac{m}{2}),$$

with $w(E) = \mathbb{E}\sup_{\boldsymbol{v}\in E}\langle \boldsymbol{g}, \boldsymbol{v}\rangle$ the Gaussian width. The exact identity $\lambda_m^2 = m^2/(m+1)$ (a Gamma-function computation) is what produces the $m^2/(m+1)$ on the LHS of the theorem. Whenever $\lambda_m > w(E)$, Gaussian Lipschitz concentration of $\boldsymbol{v} \mapsto \inf_{\boldsymbol{u}\in E}\|A\boldsymbol{u}\|_2$ around its mean (a 1-Lipschitz function of A) gives $\mathbf{Pr}(\inf_{\boldsymbol{v}\in E}\|A\boldsymbol{v}\|_2 > 0) \geq 1 - e^{-(\lambda_m - w(E))^2/2}$, hence $\mathbf{ker}\, A \cap T(\boldsymbol{x}) = \{0\}$ with the same probability.

The width of an ℓ_1 tangent cone at an s-sparse point admits the closed-form upper bound $w(T(\boldsymbol{x}) \cap S^{N-1})^2 \leq 2s\,\ln(2.34\,N/s)$. The standard route is duality: $w(E) \leq \mathbb{E}\,\mathrm{dist}(\boldsymbol{g}, N(\boldsymbol{x}))$ where $N(\boldsymbol{x})$ is the polar cone (the ℓ_1-subdifferential cone, generated by sign vectors on the support and $[-1,1]$-bounded entries off the support). Since the off-support distance contribution is the soft-thresholded $\boldsymbol{g}_{\bar{S}}$, $\mathbb{E}\,\mathrm{dist}(\boldsymbol{g}, N(\boldsymbol{x}))^2 \leq 2s\,\mathbb{E}\,\eta_\tau(g_1)^2$ with $\tau = \sqrt{2\ln(N/s)}$, evaluating to $\leq 2s\ln(2.34\,N/s)$ after optimizing the threshold.

The condition $\lambda_m > w(E) + \sqrt{2\ln\varepsilon^{-1}}$ guarantees recovery with probability $\geq 1-\varepsilon$ via the concentration step. Using $\lambda_m^2 \geq m^2/(m+1)$ and the width estimate above produces exactly the $m^2/(m+1) \geq 2s[\sqrt{\ln(2.34N/s)} + \sqrt{\ln(\varepsilon^{-1})/s}]^2$ bound stated in the theorem.

Note: *9.16 for reference*

□

Part V

Sparse Representation Classification

9 Sparse Representation-based Classification

9.1 Setup

Stack training images of class $i \in \{1, \dots, C\}$ as columns of $A_i \in \mathbb{R}^{m \times n_i}$, and concatenate

$$A = [A_1 \,|\, A_2 \,|\, \dots \,|\, A_C] \in \mathbb{R}^{m \times n}, \quad n = \sum_i n_i.$$

A test sample $\boldsymbol{y}$ from class i ideally lies near $\operatorname{span}(A_i)$, so a near-block-sparse representation $\boldsymbol{y} \approx A\boldsymbol{x}$ should have $\boldsymbol{x}$ supported on the i-th block.

9.2 The SRC algorithm (Wright et al. 2009)

Algorithm 9 Sparse Representation-based Classification (SRC)

Require: dictionary $A = [A_1 | \dots | A_C]$, test sample $\boldsymbol{y}$, tolerance ε

1: Normalize columns of A to unit ℓ_2.
2: Solve $\hat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_1$ s.t. $\|A\boldsymbol{x} - \boldsymbol{y}\|_2 \le \varepsilon$.
3: **for** $i = 1, \dots, C$ **do**
4: $\quad \delta_i(\hat{\boldsymbol{x}}) \leftarrow$ zero out coefficients outside class i
5: $\quad r_i(\boldsymbol{y}) \leftarrow \|\boldsymbol{y} - A\,\delta_i(\hat{\boldsymbol{x}})\|_2$
6: **end for**
7: **return** $\hat{c} = \arg\min_i r_i(\boldsymbol{y})$.

9.3 Robust SRC under occlusion

For corruption $\boldsymbol{e}$ that is itself sparse (occlusions, lighting artifacts, salt-and-pepper noise),

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}, \qquad \|\boldsymbol{e}\|_0 \ll m.$$

Augment the dictionary by an identity:

$$B = [A \,|\, I_m] \in \mathbb{R}^{m \times (n+m)},$$

and solve

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \quad \text{s.t.} \quad B\boldsymbol{w} = \boldsymbol{y}, \quad \boldsymbol{w} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{bmatrix}.$$

Equivalently $\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1$ s.t. $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$.

9.4 Sparsity Concentration Index (SCI)

A diagnostic quantity:

$$\mathrm{SCI}(\boldsymbol{x}) := \frac{C \max_i \|\delta_i(\boldsymbol{x})\|_1 / \|\boldsymbol{x}\|_1 - 1}{C - 1} \in [0, 1].$$

$\mathrm{SCI} = 1$ if $\boldsymbol{x}$ is fully concentrated in a single class; 0 if uniform. A threshold τ on SCI lets the classifier reject low-confidence samples ("don't know" rather than mis-classify).

9.5 Theoretical correctness

Theorem 9.1 (Sparse-representation classification (SRC) correctness, Wright et al. 2009)

If the per-class dictionaries A_i are sufficiently incoherent and the test sample lies in $\mathrm{span}(A_i)$ up to noise ε, then SRC labels $\boldsymbol{y}$ correctly when ε is below a threshold determined by the inter-class separation.

Proof. Suppose the test sample admits a noisy decomposition $\boldsymbol{y} = A_i \boldsymbol{x}_i^\star + \boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \le \varepsilon$ and the ground-truth class is i. Let $\boldsymbol{x}^\star \in \mathbb{R}^N$ be the vector with $\boldsymbol{x}_i^\star$ on the block of class i and 0 elsewhere; then $\|\boldsymbol{x}^\star\|_0 \le n_i$.

Let $A = [A_1 \mid \cdots \mid A_C]$ be the concatenated dictionary. By assumption A satisfies a coherence (or RIP) condition strong enough that BPDN with noise level ε recovers any s-sparse signal stably for all $s \ge n_i$ — concretely, $\mu_1(s) + \mu_1(s-1) < 1$ or $\delta_{2s}(A) < \sqrt{2} - 1$ depending on which guarantee is invoked. The BPDN solution $\hat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_1$ s.t. $\|A\boldsymbol{x} - \boldsymbol{y}\|_2 \le \varepsilon$ then satisfies

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2 \;\le\; D\,\varepsilon$$

for an absolute constant D depending on the recovery threshold. Hence $\hat{\boldsymbol{x}}_i \approx \boldsymbol{x}_i^\star$ and $\hat{\boldsymbol{x}}_j \approx \boldsymbol{0}$ for $j \ne i$, both to error $D\varepsilon$.

The class-i residual is

$$r_i(\boldsymbol{y}) \;=\; \|\boldsymbol{y} - A_i \hat{\boldsymbol{x}}_i\|_2 \;\le\; \|\boldsymbol{y} - A_i \boldsymbol{x}_i^\star\|_2 + \|A_i(\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i^\star)\|_2 \;\le\; \varepsilon + \|A\|_{op}\, D\varepsilon \;=\; (1 + D\|A\|_{op})\,\varepsilon.$$

For $j \ne i$, using $\hat{\boldsymbol{x}}_j$ small,

$$r_j(\boldsymbol{y}) \;\ge\; \|\boldsymbol{y} - A_j \boldsymbol{x}_j^\star\|_2 - \|A_j\|_{op}\|\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j^\star\|_2 \;\ge\; \mathrm{dist}\big(\boldsymbol{y}, \mathrm{span}(A_j)\big) - \|A\|_{op} D\varepsilon.$$

Since $\boldsymbol{y}$ lies ε-close to $\mathrm{span}(A_i)$ and the per-class subspaces are separated by $\Delta_{ij} := \mathrm{dist}(\mathrm{span}(A_i), \mathrm{span}(A_j))$, we have $\mathrm{dist}(\boldsymbol{y}, \mathrm{span}(A_j)) \ge \Delta_{ij} - \varepsilon$.

Therefore $r_j(\boldsymbol{y}) - r_i(\boldsymbol{y}) \geq \Delta_{ij} - (2 + 2D\|A\|_{op})\varepsilon$, which is positive — i.e. class i wins — whenever

$$\varepsilon \; < \; \frac{\min_{j \neq i} \Delta_{ij}}{2 + 2D\|A\|_{op}}.$$

This is the threshold determined by the inter-class separation. □

9.6 Comparison with classical classifiers

- **Eigenfaces (PCA)**: project onto top eigenvectors of the covariance, classify by nearest neighbor. Sensitive to occlusion (corrupted pixels pollute all projections).
- **Fisherfaces (LDA)**: maximize Fisher's criterion $\max_W |W^\top S_B W|/|W^\top S_W W|$. Better class separation than PCA but still distance-based.
- k-**NN**: classify by majority vote among nearest training points. Simple but high variance, sensitive to dimensionality.
- **SVM**: maximum-margin hyperplane $\min \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum \xi_i$ s.t. $c_i(\boldsymbol{w}^\top \boldsymbol{y}_i + b) \geq 1 - \xi_i$. Excellent for binary problems, kernel extension for non-linearity.
- **SRC**: leverages CS theory; handles occlusion natively via the identity-augmented dictionary; can certify reliability via SCI.

9.7 Applications

Face recognition under occlusion, hyperspectral image classification, distributed sensor identification, activity recognition from wearable sensors, video surveillance.

Part VI

ℓ_1-Minimization Algorithms

10 Convex Algorithms for ℓ_1 Minimization

10.1 The toolkit

We solve large-scale instances of

$$\text{BP:} \quad \min \|\boldsymbol{x}\|_1 \text{ s.t. } A\boldsymbol{x} = \boldsymbol{y}, \tag{7}$$

$$\text{BPDN:} \quad \min \|\boldsymbol{x}\|_1 \text{ s.t. } \|A\boldsymbol{x} - \boldsymbol{y}\|_2 \le \eta, \tag{8}$$

$$\text{LASSO:} \quad \min \tfrac{1}{2}\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1, \tag{9}$$

$$\text{Elastic Net:} \quad \min \tfrac{1}{2}\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{x}\|_2^2, \tag{10}$$

$$\text{Dantzig:} \quad \min \|\boldsymbol{x}\|_1 \text{ s.t. } \|A^*(A\boldsymbol{x} - \boldsymbol{y})\|_\infty \le \tau. \tag{11}$$

The choice of algorithm depends on problem size, structure of A (dense, fast-multiplied, factored), and noise model.

10.2 Geometry of why ℓ_1 promotes sparsity

The unit ℓ_1-ball is a polytope with vertices on the coordinate axes; the ℓ_2-ball is round. As we shrink the ℓ_p-norm subject to $A\boldsymbol{x} = \boldsymbol{y}$, the first contact with the affine set lands at a vertex (sparse) for $p = 1$ but on a generic face (dense) for $p = 2$.

10.3 Soft thresholding and proximal operators

Definition 10.1 (Soft thresholding)

$$\mathcal{S}_\tau(\boldsymbol{u}) = \operatorname{sign}(\boldsymbol{u}) \odot \max(|\boldsymbol{u}| - \tau, 0),$$

component-wise.

Lemma 10.2 (Proximal operator of ℓ_1)

$$\mathcal{S}_\tau(\boldsymbol{u}) = \arg\min_{\boldsymbol{x}} \left(\tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \tau\|\boldsymbol{x}\|_1 \right).$$

Proof. The objective separates across coordinates; for each x_i minimize $\frac{1}{2}(x_i - u_i)^2 + \tau|x_i|$. The KKT condition gives the soft-threshold formula. □

10.4 ISTA and FISTA

Algorithm 10 ISTA – Iterative Soft-Thresholding for LASSO

Require: step $\alpha \le 1/\|A\|_{2\to 2}^2$, $\lambda > 0$

$x^0 \leftarrow 0$

for $k = 0, 1, \dots$ **do**

  $x^{k+1} \leftarrow \mathcal{S}_{\alpha\lambda}\big(x^k - \alpha A^*(Ax^k - y)\big)$

end for

Algorithm 11 FISTA – Beck–Teboulle 2009

$x^0 \leftarrow 0$, $z^1 \leftarrow x^0$, $t_1 \leftarrow 1$

for $k = 1, 2, \dots$ **do**

  $x^k \leftarrow \mathcal{S}_{\alpha\lambda}\big(z^k - \alpha A^*(Az^k - y)\big)$

  $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$

  $z^{k+1} \leftarrow x^k + \frac{t_k - 1}{t_{k+1}}(x^k - x^{k-1})$

end for

Theorem 10.3 (FISTA convergence rate)

For LASSO with $L = \|A\|_{2\to 2}^2$ and step $\alpha = 1/L$, ISTA satisfies $F(x^k) - F^\star = O(L\|x^0 - x^\star\|_2^2/k)$; FISTA improves to $O(L\|x^0 - x^\star\|_2^2/k^2)$, optimal among first-order methods (Nesterov).

Proof. Write $F = f + g$ with $f(x) = \frac{1}{2}\|Ax - y\|_2^2$ (L-smooth, $L = \|A\|_{2\to 2}^2$) and $g(x) = \lambda\|x\|_1$. The proximal-gradient operator $T_\alpha(z) := \mathrm{prox}_{\alpha g}(z - \alpha\nabla f(z))$ at step $\alpha = 1/L$ satisfies the standard descent lemma: for every x,

$$F(T_\alpha z) - F(x) \;\le\; \tfrac{L}{2}\big(\|z - x\|_2^2 - \|T_\alpha z - x\|_2^2\big). \tag{$*$}$$

This is obtained by combining $f(T_\alpha z) \le f(z) + \langle \nabla f(z), T_\alpha z - z\rangle + \frac{L}{2}\|T_\alpha z - z\|_2^2$ (smoothness) with the prox subgradient inequality $g(T_\alpha z) \le g(x) + \langle (z - T_\alpha z)/\alpha - \nabla f(z), T_\alpha z - x\rangle$ and rearranging.

For ISTA, $x^{k+1} = T_\alpha x^k$ with $z = x^k$. Apply $(*)$ first with $x = x^k$ to get $F(x^{k+1}) \le F(x^k)$, and then with $x = x^\star$ to get $F(x^{k+1}) - F^\star \le \frac{L}{2}(\|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2)$. Summing the second inequality from 0 to $k-1$ telescopes; using monotonicity of $F(x^k)$ yields $F(x^k) - F^\star \le L\|x^0 - x^\star\|_2^2/(2k)$, the $O(1/k)$ rate.

For FISTA, the analysis is more delicate because the prox is taken at the extrapolation point z^k, not at x^k. Apply $(*)$ once at $z = z^k$, $x = x^k$ and once at $z = z^k$, $x = x^\star$, multiply the first by $(t_k - 1)$ and the

second by 1, and add. Using $t_k^2 = t_{k+1}^2 - t_{k+1}$ (from the recurrence $t_{k+1} = (1+\sqrt{1+4t_k^2})/2$) and the extrapolation identity $t_{k+1}z^{k+1} = x^{k+1} + (t_k - 1)(x^{k+1} - x^k)$, one obtains the Lyapunov sequence

$$u_k \;:=\; t_k^2\big(F(x^k) - F^\star\big) \;+\; \tfrac{L}{2}\big\|t_k z^{k+1} - (t_k-1)x^k - x^\star\big\|_2^2,$$

satisfying $u_{k+1} \le u_k$. Therefore $u_k \le u_0 = \frac{L}{2}\|x^0 - x^\star\|_2^2$. Since the recurrence forces $t_k \ge (k+1)/2$ by induction (base $t_1 = 1$, $t_{k+1}^2 - t_{k+1} = t_k^2 \ge (k+1)^2/4$ implies $t_{k+1} \ge (k+2)/2$),

$$F(x^k) - F^\star \;\le\; \frac{u_k}{t_k^2} \;\le\; \frac{2L\|x^0 - x^\star\|_2^2}{(k+1)^2}.$$

The matching lower bound — that no first-order oracle method achieves better than $\Omega(L\|x^0 - x^\star\|_2^2/k^2)$ in the worst case — is Nesterov's classical construction using a particular tridiagonal quadratic. □

10.5 ADMM for BP

Algorithm 12 ADMM for $\min \|x\|_1$ s.t. $Ax = y$

$x^0, z^0, u^0 \leftarrow 0$; $\rho > 0$
for $t = 0, 1, \ldots$ **do**
  $x^{t+1} \leftarrow \mathcal{S}_{1/\rho}(z^t - u^t)$ ▷ soft threshold
  $z^{t+1} \leftarrow x^{t+1} + u^t + A^*(AA^*)^{-1}\big(y - A(x^{t+1} + u^t)\big)$
  $u^{t+1} \leftarrow u^t + x^{t+1} - z^{t+1}$
end for

The projection step uses a single Cholesky factor of AA^* amortized over all iterations. Theoretical rate: $O(1/k)$ ergodic; linear under additional strong-convexity hypotheses.

10.6 Primal–dual splitting (Chambolle–Pock 2011)

For the saddle-point problem $\min_x \max_\xi \langle Ax, \xi\rangle + G(x) - F^*(\xi)$:

Algorithm 13 Chambolle–Pock primal-dual algorithm

Require: $\theta \in [0,1]$, $\sigma, \tau > 0$ with $\sigma\tau\|A\|_{2\to 2}^2 < 1$

1: $\boldsymbol{x}^0 \in \mathbb{R}^N$, $\boldsymbol{\xi}^0 \in \mathbb{R}^m$, $\bar{\boldsymbol{x}}^0 \leftarrow \boldsymbol{x}^0$
2: **for** $n = 0, 1, \ldots$ **do**
3: $\quad \boldsymbol{\xi}^{n+1} \leftarrow \operatorname{prox}_{\sigma F^*}(\boldsymbol{\xi}^n + \sigma A \bar{\boldsymbol{x}}^n)$
4: $\quad \boldsymbol{x}^{n+1} \leftarrow \operatorname{prox}_{\tau G}(\boldsymbol{x}^n - \tau A^* \boldsymbol{\xi}^{n+1})$
5: $\quad \bar{\boldsymbol{x}}^{n+1} \leftarrow \boldsymbol{x}^{n+1} + \theta(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)$
6: **end for**

Theorem 10.4 (ℓ_1-instance optimality)

Under the step-size condition, with $\theta = 1$, the iterates converge to a saddle point. The ergodic gap decays at rate $O(1/k)$ in the convex case, $O(1/k^2)$ if either F^* or G is strongly convex, and linearly if both are.

Proof. Let $(\boldsymbol{x}^\star, \boldsymbol{\xi}^\star)$ be a saddle point. The proximal updates produce the variational inequalities (subgradient definitions of the prox)

$$\tfrac{1}{\tau}(\boldsymbol{x}^n - \boldsymbol{x}^{n+1}) - A^*\boldsymbol{\xi}^{n+1} \in \partial G(\boldsymbol{x}^{n+1}), \qquad \tfrac{1}{\sigma}(\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n+1}) + A\bar{\boldsymbol{x}}^n \in \partial F^*(\boldsymbol{\xi}^{n+1}).$$

Pairing each with the corresponding gap functional and using monotonicity of the subdifferentials gives

$$\mathcal{L}(\boldsymbol{x}^{n+1}, \boldsymbol{\xi}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\xi}^{n+1}) \;\le\; \tfrac{1}{2\tau}\|\boldsymbol{x}^n - \boldsymbol{x}^\star\|_2^2 - \tfrac{1}{2\tau}\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^\star\|_2^2 + \tfrac{1}{2\sigma}\|\boldsymbol{\xi}^n - \boldsymbol{\xi}^\star\|_2^2 - \tfrac{1}{2\sigma}\|\boldsymbol{\xi}^{n+1} - \boldsymbol{\xi}^\star\|_2^2 - R_n,$$

where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\xi}) = \langle A\boldsymbol{x}, \boldsymbol{\xi}\rangle + G(\boldsymbol{x}) - F^*(\boldsymbol{\xi})$ and $R_n = \frac{1}{2\tau}\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 + \frac{1}{2\sigma}\|\boldsymbol{\xi}^{n+1} - \boldsymbol{\xi}^n\|_2^2 - \langle A(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n), \boldsymbol{\xi}^{n+1} - \boldsymbol{\xi}^n\rangle$. Cauchy–Schwarz on the cross-term and the step-size condition $\sigma\tau\|A\|_{2\to 2}^2 < 1$ imply $R_n \ge 0$. So the Lyapunov function $V_n = \frac{1}{2\tau}\|\boldsymbol{x}^n - \boldsymbol{x}^\star\|_2^2 + \frac{1}{2\sigma}\|\boldsymbol{\xi}^n - \boldsymbol{\xi}^\star\|_2^2$ is non-increasing.

Summing the inequality from $n = 0$ to $N - 1$ and using convexity of $\mathcal{L}$ in $\boldsymbol{x}$ and concavity in $\boldsymbol{\xi}$ on the ergodic averages $\bar{\boldsymbol{x}}^N = \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}^n$, $\bar{\boldsymbol{\xi}}^N = \frac{1}{N}\sum_{n=1}^N \boldsymbol{\xi}^n$,

$$\mathcal{L}(\bar{\boldsymbol{x}}^N, \boldsymbol{\xi}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \bar{\boldsymbol{\xi}}^N) \;\le\; \frac{V_0}{N}.$$

This is the $O(1/N)$ rate on the primal-dual gap.

If G is γ-strongly convex (the case $\gamma > 0$), the prox-step inequality picks up an extra $\frac{\gamma}{2}\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^\star\|_2^2$ term. Choosing time-varying $\tau_n, \sigma_n, \theta_n$ that exploit this acceleration via Chambolle–Pock's adaptive scheme yields $V_n = O(1/n^2)$. If F^* is also strongly convex, the Lyapunov bound contracts geometrically, giving linear convergence with explicit rate $1/(1 + 2\sqrt{\gamma_F \gamma_G \sigma\tau})$. □

10.7 Iteratively Reweighted Least Squares (IRLS)

The identity $|t| = t^2/|t|$ inspires solving a sequence of weighted least-squares problems:

Algorithm 14 IRLS, Daubechies–DeVore–Fornasier–Güntürk

1: $\boldsymbol{w}^0 \leftarrow \mathbf{1}$, $\varepsilon_0 \leftarrow 1$
2: **for** $n = 0, 1, \dots$ **do**
3: $\quad \boldsymbol{x}^{n+1} \leftarrow \arg\min_{\boldsymbol{x}} \sum_j w_j^n |x_j|^2$ s.t. $A\boldsymbol{x} = \boldsymbol{y}$
4: $\quad\quad$ *closed form:* $\boldsymbol{x}^{n+1} = D^{-1}A^*(AD^{-1}A^*)^{-1}\boldsymbol{y}$, $D = \operatorname{diag}(w_j^n)$
5: $\quad \varepsilon_{n+1} \leftarrow \min\{\varepsilon_n, \gamma\,(x^{n+1})^*_{s+1}\}$ $\quad\triangleright$ $(x)^*_{s+1}$ = $(s+1)$-st largest entry
6: $\quad w_j^{n+1} \leftarrow (|x_j^{n+1}|^2 + \varepsilon_{n+1}^2)^{-1/2}$
7: **end for**

Theorem 10.5 (Gelfand width lower bound)

If A satisfies the stable NSP of order s with $\rho < 1$, IRLS with $\gamma = 1/N$ converges. If $\varepsilon_n \to 0$, $\boldsymbol{x}^n \to \boldsymbol{x}^\#$, the BP solution; for sparse signals, convergence is super-linear with rate $\mu = \rho(1+\rho)/(1-\kappa)\big(1 + 1/(s+1-\tilde{s})\big) < 1$.

Proof. Define the auxiliary functional

$$\mathcal{J}(\boldsymbol{x}, \boldsymbol{w}, \varepsilon) \;=\; \tfrac{1}{2}\sum_j w_j x_j^2 \;+\; \tfrac{1}{2}\sum_j \big(\varepsilon^2 w_j + 1/w_j\big).$$

Two key inequalities hold. First, the weighted-LS step minimizes $\boldsymbol{x} \mapsto \mathcal{J}(\boldsymbol{x}, \boldsymbol{w}^n, \varepsilon_n)$ subject to $A\boldsymbol{x} = \boldsymbol{y}$, so $\mathcal{J}(\boldsymbol{x}^{n+1}, \boldsymbol{w}^n, \varepsilon_n) \le \mathcal{J}(\boldsymbol{x}^n, \boldsymbol{w}^n, \varepsilon_n)$. Second, fixing $\boldsymbol{x}^{n+1}$ and $\varepsilon_{n+1} \le \varepsilon_n$, the choice $w_j = (|x_j^{n+1}|^2 + \varepsilon_{n+1}^2)^{-1/2}$ minimizes $\boldsymbol{w} \mapsto \mathcal{J}$, giving $\mathcal{J}(\boldsymbol{x}^{n+1}, \boldsymbol{w}^{n+1}, \varepsilon_{n+1}) \le \mathcal{J}(\boldsymbol{x}^{n+1}, \boldsymbol{w}^n, \varepsilon_{n+1})$. Combined with $\mathcal{J}(\cdot,\cdot,\varepsilon_{n+1}) \le \mathcal{J}(\cdot,\cdot,\varepsilon_n)$ (since $\varepsilon_{n+1} \le \varepsilon_n$), the sequence $\mathcal{J}_n := \mathcal{J}(\boldsymbol{x}^n, \boldsymbol{w}^n, \varepsilon_n)$ is monotone non-increasing.

At the optimum of $\boldsymbol{w}$, $\mathcal{J}(\boldsymbol{x}, \boldsymbol{w}^\star, \varepsilon) = \sum_j \sqrt{x_j^2 + \varepsilon^2}$, a smoothed ℓ_1-norm. Stable NSP of order s with $\rho < 1$ provides a coercivity estimate: any feasible $\boldsymbol{x}$ with $\sum_j \sqrt{x_j^2 + \varepsilon^2}$ bounded has $\|\boldsymbol{x}\|_2$ bounded uniformly in ε. Hence $\{\boldsymbol{x}^n\}$ is bounded; pass to a convergent subsequence $\boldsymbol{x}^{n_k} \to \boldsymbol{x}^\#$.

By construction $A\boldsymbol{x}^n = \boldsymbol{y}$ throughout, and the first-order conditions at the optimum of the weighted LS read $A^*\boldsymbol{\lambda}^n = \operatorname{diag}(w^n)\boldsymbol{x}^n$. Passing to the limit and using $\varepsilon_n \to 0$ on the support, the sign vector $\boldsymbol{\lambda}^\#$ becomes a $\partial\|\cdot\|_1$-subgradient at $\boldsymbol{x}^\#$, so $\boldsymbol{x}^\#$ satisfies the BP optimality conditions and equals $\boldsymbol{x}^\#_{BP}$.

For the super-linear rate, write $\boldsymbol{x}^n = \boldsymbol{x}^\# + \boldsymbol{e}^n$ and linearize the weight update. The map $\boldsymbol{x}^n \mapsto \boldsymbol{x}^{n+1}$ becomes a contraction on $\boldsymbol{e}^n$ near a sparse fixed point; the stable NSP gives a contraction factor $\mu = \rho(1+\rho)/(1-\kappa)\big(1 + 1/(s+1-\tilde{s})\big) < 1$, and the coupling between ε_n and the $(s+1)$-st largest entry of $\boldsymbol{x}^{n+1}$ drives the error super-linearly to zero on the sparse part. □

10.8 LARS / Homotopy method

Algorithm 15 LARS-Homotopy for the LASSO path

1: $\boldsymbol{x}^0 \leftarrow \mathbf{0}$, $\lambda^0 \leftarrow \|A^*\boldsymbol{y}\|_\infty$, active set $T_0 \leftarrow \emptyset$
2: $j_1 \leftarrow \arg\max_j |(A^*\boldsymbol{y})_j|$; $T_1 \leftarrow \{j_1\}$
3: **for** $n = 1, 2, \ldots$ **do**
4: Compute direction $\boldsymbol{d}^n$ on T_n: $\boldsymbol{d}^n_{T_n} = (A^*_{T_n}A_{T_n})^{-1}\mathrm{sign}(A^*(\boldsymbol{y} - A\boldsymbol{x}^{n-1}))_{T_n}$
5: Determine the next breakpoint γ^n either:
- an inactive index hits the boundary $|(A^*(\boldsymbol{y} - A\boldsymbol{x}^{n-1} - \gamma A\boldsymbol{d}^n))_\ell| = \lambda^{n-1} - \gamma$, or
- an active coefficient changes sign.
6: Update $\boldsymbol{x}^n = \boldsymbol{x}^{n-1} + \gamma^n\boldsymbol{d}^n$, $\lambda^n = \lambda^{n-1} - \gamma^n$.
7: Update T_{n+1} accordingly (add or remove one index).
8: **end for**

Theorem 10.6 (Quotient property and BP stability)

If the BP solution is unique and breakpoints are non-degenerate, LARS terminates in finitely many steps and outputs the full LASSO regularization path. At $\lambda = 0$, the BP solution is recovered.

Proof. The LASSO KKT conditions at parameter λ are: there exists $\boldsymbol{s} \in \partial\|\boldsymbol{x}\|_1$ (so $s_j = \mathrm{sign}(x_j)$ if $x_j \neq 0$ and $|s_j| \leq 1$ otherwise) such that $A^*(A\boldsymbol{x} - \boldsymbol{y}) + \lambda\boldsymbol{s} = \mathbf{0}$. Equivalently, the residual correlation $\boldsymbol{c}(\boldsymbol{x}) = A^*(\boldsymbol{y} - A\boldsymbol{x})$ satisfies $c_j = \lambda\,\mathrm{sign}(x_j)$ for $j \in T$ (active) and $|c_j| \leq \lambda$ for $j \notin T$ (inactive).

Fix a sign pattern $\boldsymbol{\eta} \in \{\pm 1\}^T$ on the active set T. Within this pattern the KKT system becomes the linear equation $A^*_T A_T \boldsymbol{x}_T = A^*_T\boldsymbol{y} - \lambda\boldsymbol{\eta}$, whose solution is $\boldsymbol{x}_T(\lambda) = (A^*_T A_T)^{-1}(A^*_T\boldsymbol{y} - \lambda\boldsymbol{\eta})$. Hence $\boldsymbol{x}^\star(\lambda)$ is affine in λ on this regime, and the residual $\boldsymbol{r}(\lambda) = \boldsymbol{y} - A_T\boldsymbol{x}_T(\lambda)$ is also affine, with derivative $A_T(A^*_T A_T)^{-1}\boldsymbol{\eta}$ — the LARS direction $\boldsymbol{d}^n$.

Two events end the current regime as λ decreases:

- An inactive coordinate $\ell \notin T$ has $|c_\ell(\lambda)| = \lambda$ — the dual feasibility becomes tight, forcing ℓ into the active set.
- An active coordinate $j \in T$ has $x_j(\lambda) = 0$ — its sign is no longer determined by KKT, forcing it out of the active set.

Both events are linear conditions on λ, so the next breakpoint γ^n is computable in closed form. Under the non-degeneracy assumption exactly one event occurs per breakpoint, so T_{n+1} differs from T_n by a single element.

There are at most $\binom{N}{|T|} \cdot 2^{|T|}$ distinct $(T, \boldsymbol{\eta})$ regimes, and each is visited at most once because λ^n is strictly decreasing along the path. Hence LARS terminates in finitely many steps.

At $\lambda = 0$ the inactive constraints reduce to $A^*(A\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{0}$ on the active set with no penalty, while the dual constraint $|c_\ell| \le 0$ on inactive coordinates forces $\boldsymbol{c} = \boldsymbol{0}$ when those coordinates have not been reached. Equivalently, the LARS endpoint is feasible for P_1 and the residual sign vector $\boldsymbol{s}$ certifies optimality, so $\boldsymbol{x}^\star(0)$ is the BP solution. □

Per-breakpoint cost: $O(s^2 + sN)$; total: $O(s^3 + s^2N)$ for s-sparse outputs.

10.9 Iteratively reweighted ℓ_1

A different reweighting (Candès–Wakin–Boyd 2008): solve

$$\boldsymbol{x}^{(t+1)} = \arg\min_{\boldsymbol{x}} \sum_i w_i^{(t)} |x_i| \text{ s.t. } A\boldsymbol{x} = \boldsymbol{y}, \quad w_i^{(t+1)} = \frac{1}{|x_i^{(t)}| + \varepsilon}.$$

Penalizes small coefficients more heavily on each pass; empirically beats plain BP.

10.10 Robust PCA and outliers

For $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ with sparse $\boldsymbol{e}$, $\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1$ s.t. $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$. The matrix analog: *Robust PCA*,

$$\min \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad M = L + S,$$

recovers low-rank L from sparse corruption S (Candès, Li, Ma, Wright 2011).

10.11 Software

spgl1, l1_ls, yall1, tfocs, gpsr, slep, cvx/cvxpy/scs for prototyping. For very large problems use FISTA with continuation in λ, or ADMM with warm starts; for adaptive sparsity use IRLS.

Appendices

A Probability Tools

A.1 Markov, Chebyshev, Chernoff

Theorem A.1 (Markov)

For a non-negative r.v. X and $t > 0$, $\mathbf{Pr}(X \geq t) \leq \mathbb{E}X/t$.

Proof. Pointwise we have $X \geq t\,\mathbf{1}_{\{X \geq t\}}$: on the event $\{X \geq t\}$ both sides exceed t, and on $\{X < t\}$ the right side is 0 while the left is non-negative. Taking expectations of both sides (which preserves the inequality by linearity of $\mathbb{E}$),

$$\mathbb{E}X \;\geq\; t\,\mathbb{E}\,\mathbf{1}_{\{X \geq t\}} \;=\; t\;\mathbf{Pr}(X \geq t).$$

Dividing by $t > 0$ gives the result. □

Theorem A.2 (Chebyshev)

For any r.v. X with finite second moment, $\mathbf{Pr}(|X - \mathbb{E}X| \geq t) \leq \mathrm{Var}(X)/t^2$.

Proof. Set $Y = (X - \mathbb{E}X)^2 \geq 0$. The event $\{|X - \mathbb{E}X| \geq t\}$ is the same as $\{Y \geq t^2\}$. Markov's inequality applied to Y with threshold t^2 gives $\mathbf{Pr}(Y \geq t^2) \leq \mathbb{E}Y/t^2 = \mathrm{Var}(X)/t^2$. □

Theorem A.3 (Chernoff)

For any r.v. X, $\mathbf{Pr}(X \geq t) \leq \inf_{\theta > 0} e^{-\theta t}\,\mathbb{E}\,e^{\theta X}$.

Proof. Fix $\theta > 0$. The map $u \mapsto e^{\theta u}$ is monotone increasing, so $\{X \geq t\} = \{e^{\theta X} \geq e^{\theta t}\}$. Markov's inequality applied to the non-negative random variable $e^{\theta X}$ with threshold $e^{\theta t}$ yields

$$\mathbf{Pr}(X \geq t) = \mathbf{Pr}(e^{\theta X} \geq e^{\theta t}) \;\leq\; e^{-\theta t}\,\mathbb{E}\,e^{\theta X}.$$

Since this holds for every $\theta > 0$, take the infimum over the right side. □

A.2 Cramér and Hoeffding

Theorem A.4 (Cramér)

For independent r.v.s $X_1, \ldots, X_M$ with cumulant generating functions $C_\ell(\theta) = \ln \mathbb{E}\,e^{\theta X_\ell}$,

$$\mathbf{Pr}\big(\textstyle\sum_\ell X_\ell \geq t\big) \leq \exp\big(\inf_{\theta > 0}\{-\theta t + \textstyle\sum_\ell C_\ell(\theta)\}\big).$$

Proof. Set $S = \sum_{\ell=1}^{M} X_\ell$ and apply Chernoff: for any $\theta > 0$, $\mathbf{Pr}(S \geq t) \leq e^{-\theta t}\,\mathbb{E}\,e^{\theta S}$. Independence factors the joint MGF:

$$\mathbb{E}\,e^{\theta S} \;=\; \mathbb{E}\prod_{\ell=1}^{M} e^{\theta X_\ell} \;=\; \prod_{\ell=1}^{M} \mathbb{E}\,e^{\theta X_\ell} \;=\; \prod_{\ell} e^{C_\ell(\theta)} \;=\; \exp\Big(\textstyle\sum_\ell C_\ell(\theta)\Big),$$

the second equality using the product rule for expectations of independent random variables. Hence $\mathbf{Pr}(S \geq t) \leq \exp(-\theta t + \sum_\ell C_\ell(\theta))$. Taking the infimum over $\theta > 0$ gives the claimed bound. □

Theorem A.5 (Hoeffding)

If X_ℓ are independent with $\mathbb{E}X_\ell = 0$, $a_\ell \leq X_\ell \leq b_\ell$, then for $t > 0$,

$$\mathbf{Pr}\big(\textstyle\sum_\ell X_\ell \geq t\big) \leq \exp\Big(-\frac{2t^2}{\sum_\ell (b_\ell - a_\ell)^2}\Big).$$

Proof. We first establish the auxiliary bound (*Hoeffding's lemma*): if $\mathbb{E}X = 0$ and $X \in [a, b]$ a.s., then $\mathbb{E}\,e^{\theta X} \leq \exp(\theta^2 (b-a)^2/8)$ for every $\theta \in \mathbb{R}$. Indeed, for $x \in [a, b]$ write $x = \lambda b + (1-\lambda)a$ with $\lambda = (x-a)/(b-a) \in [0,1]$. Convexity of $u \mapsto e^{\theta u}$ gives $e^{\theta x} \leq \lambda e^{\theta b} + (1-\lambda)e^{\theta a}$. Taking expectation and using $\mathbb{E}X = 0$ (so $\mathbb{E}\lambda = -a/(b-a)$),

$$\mathbb{E}\,e^{\theta X} \;\leq\; \frac{-a}{b-a}\,e^{\theta b} + \frac{b}{b-a}\,e^{\theta a}.$$

Set $u = \theta(b-a)$ and $p = -a/(b-a) \in [0,1]$, so the right side equals $e^{-pu + \ln(1-p+pe^u)}$. Define $\varphi(u) = -pu + \ln(1 - p + pe^u)$. Then $\varphi(0) = \varphi'(0) = 0$, and $\varphi''(u) = \frac{p(1-p)e^u}{(1-p+pe^u)^2} \leq 1/4$ for all u (the maximum of $q(1-q)$ is $1/4$). Taylor's theorem gives $\varphi(u) \leq u^2/8$, so $\mathbb{E}\,e^{\theta X} \leq e^{\theta^2 (b-a)^2/8}$.

Now apply Cramér with $C_\ell(\theta) \leq \theta^2 (b_\ell - a_\ell)^2/8$:

$$\mathbf{Pr}\big(\textstyle\sum_\ell X_\ell \geq t\big) \leq \exp\Big(-\theta t + \frac{\theta^2}{8}\sum_\ell (b_\ell - a_\ell)^2\Big).$$

Minimize the right side over $\theta > 0$; the optimal choice $\theta^\star = 4t/\sum (b_\ell - a_\ell)^2$ produces the stated bound $\exp(-2t^2/\sum (b_\ell - a_\ell)^2)$. □

A.3 Subgaussian random variables

Definition A.6 (Subgaussian)

X is subgaussian if $\mathbf{Pr}(|X| \geq t) \leq \beta e^{-\kappa t^2}$ for all $t > 0$, or equivalently, $\mathbb{E}\,e^{\theta X} \leq e^{c\theta^2}$ for all $\theta \in \mathbb{R}$ (some $c > 0$).

Proposition A.7 (Tail decay)

If $\mathbb{E}\,e^{\theta X} \leq e^{c\theta^2}$, then $\mathbf{Pr}(|X| \geq t) \leq 2e^{-t^2/(4c)}$.

Proof. By Chernoff, for any $\theta > 0$, $\mathbf{Pr}(X \geq t) \leq e^{-\theta t}\,\mathbb{E}\,e^{\theta X} \leq \exp(-\theta t + c\theta^2)$. The exponent $-\theta t + c\theta^2$ is a quadratic in θ minimized at $\theta^\star = t/(2c)$, where its value is $-t^2/(4c)$, hence $\mathbf{Pr}(X \geq t) \leq \exp(-t^2/(4c))$. The same argument applied to $-X$ (whose MGF satisfies the same bound by hypothesis allowing both signs of θ) gives $\mathbf{Pr}(X \leq -t) \leq \exp(-t^2/(4c))$. Adding the two tail probabilities yields $\mathbf{Pr}(|X| \geq t) \leq 2\exp(-t^2/(4c))$. □

Theorem A.8 (Subgaussian sum)

For independent zero-mean subgaussian X_ℓ with parameter c and weights $\boldsymbol{a} \in \mathbb{R}^M$, $\mathbf{Pr}(|\sum a_\ell X_\ell| \geq t) \leq 2e^{-t^2/(4c\|\boldsymbol{a}\|_2^2)}$.

Proof. Each $a_\ell X_\ell$ is itself subgaussian: substituting $\theta \mapsto a_\ell\theta$ in $\mathbb{E}\,e^{\theta X_\ell} \leq e^{c\theta^2}$ yields $\mathbb{E}\,e^{\theta a_\ell X_\ell} \leq e^{c a_\ell^2 \theta^2}$. By independence the joint MGF factors:

$$\mathbb{E}\,e^{\theta\sum_\ell a_\ell X_\ell} \;=\; \prod_{\ell=1}^{M} \mathbb{E}\,e^{\theta a_\ell X_\ell} \;\leq\; \prod_{\ell} e^{c a_\ell^2\theta^2} \;=\; \exp\Big(c\theta^2 \textstyle\sum_\ell a_\ell^2\Big) \;=\; \exp(c\|\boldsymbol{a}\|_2^2\theta^2).$$

So $\sum_\ell a_\ell X_\ell$ is subgaussian with effective parameter $c\|\boldsymbol{a}\|_2^2$. The previous proposition then gives the tail bound. □

A.4 Bernstein inequalities

Theorem A.9 (Bernstein)

Let $X_1, \ldots, X_M$ be independent zero-mean with $|X_\ell| \leq K$ and $\mathbb{E}X_\ell^2 \leq \sigma_\ell^2$. Set $\sigma^2 = \sum_\ell \sigma_\ell^2$. Then for $t > 0$,

$$\mathbf{Pr}\big(\big|\textstyle\sum_\ell X_\ell\big| \geq t\big) \leq 2\exp\Big(-\frac{t^2/2}{\sigma^2 + Kt/3}\Big).$$

Proof. We bound each MGF $\mathbb{E}\,e^{\theta X_\ell}$ for $\theta \in (0, 3/K)$. Taylor-expand and use $\mathbb{E}X_\ell = 0$:

$$\mathbb{E}\,e^{\theta X_\ell} \;=\; 1 + \sum_{n\geq 2} \frac{\theta^n}{n!}\,\mathbb{E}X_\ell^n.$$

The boundedness $|X_\ell| \leq K$ together with $\mathbb{E}X_\ell^2 \leq \sigma_\ell^2$ gives $\mathbb{E}|X_\ell|^n \leq K^{n-2}\,\mathbb{E}X_\ell^2 \leq K^{n-2}\sigma_\ell^2$ for $n \geq 2$. Therefore

$$\sum_{n\geq 2} \frac{\theta^n}{n!}\,\mathbb{E}|X_\ell|^n \;\leq\; \sigma_\ell^2 \sum_{n\geq 2} \frac{\theta^n K^{n-2}}{n!} \;\leq\; \sigma_\ell^2 \cdot \frac{\theta^2}{2} \sum_{n\geq 0} \frac{(K\theta)^n}{3^n} \;=\; \frac{\theta^2\sigma_\ell^2/2}{1 - K\theta/3},$$

where the inequality $n!/2 \geq 3^{n-2}$ for $n \geq 2$ was used. Combining with $1 + u \leq e^u$ yields

$$\mathbb{E}\,e^{\theta X_\ell} \;\leq\; \exp\Big(\frac{\theta^2\sigma_\ell^2/2}{1 - K\theta/3}\Big).$$

Apply Cramér's theorem with $C_\ell(\theta) \le \theta^2\sigma_\ell^2/(2(1-K\theta/3))$:

$$\mathbf{Pr}\big(\textstyle\sum_\ell X_\ell \ge t\big) \;\le\; \exp\Big(-\theta t + \frac{\theta^2\sigma^2/2}{1-K\theta/3}\Big).$$

Choose $\theta = t/(\sigma^2 + Kt/3) \in (0, 3/K)$. Then $1 - K\theta/3 = \sigma^2/(\sigma^2 + Kt/3)$, and substituting gives

$$-\theta t + \frac{\theta^2\sigma^2/2}{1-K\theta/3} \;=\; -\theta t + \frac{\theta^2(\sigma^2 + Kt/3)}{2} \;=\; -\frac{t^2/2}{\sigma^2 + Kt/3}.$$

This bounds $\mathbf{Pr}(\sum X_\ell \ge t)$. The lower tail follows from the same argument applied to $-X_\ell$ (which satisfies the same hypotheses), and adding the two tails produces the factor of 2 in front. □

Theorem A.10 (Noncommutative Bernstein, Tropp 2012)

For independent zero-mean self-adjoint random matrices $X_\ell \in \mathbb{C}^{d\times d}$ with $\lambda_{\max}(X_\ell) \le K$ a.s. and $\sigma^2 := \|\sum_\ell \mathbb{E}\, X_\ell^2\|_{2\to 2}$,

$$\mathbf{Pr}\Big(\Big\|\sum_\ell X_\ell\Big\|_{2\to 2} \ge t\Big) \le d\, \exp\Big(-\frac{t^2/2}{\sigma^2 + Kt/3}\Big).$$

Proof. Markov's inequality lifts to self-adjoint matrices via the trace: for any self-adjoint S,

$$\mathbf{Pr}\big(\lambda_{\max}(S) \ge t\big) \;\le\; \mathbf{Pr}\big(\operatorname{tr} e^{\theta S} \ge e^{\theta t}\big) \;\le\; e^{-\theta t}\,\mathbb{E}\operatorname{tr} e^{\theta S}, \qquad \theta > 0,$$

using $\operatorname{tr} e^{\theta S} \ge e^{\theta\lambda_{\max}(S)}$. Apply this to $S = \sum_\ell X_\ell$. The independence step requires Lieb's concavity theorem (Appendix B): for any fixed self-adjoint H, the map $A \mapsto \operatorname{tr}\exp(H + \ln A)$ is concave on positive-definite A. By Jensen, $\mathbb{E}\operatorname{tr}\exp(H + \ln Y) \le \operatorname{tr}\exp(H + \ln \mathbb{E}Y)$ for any random positive-definite Y. Iterating across the M independent terms (with $H = \theta\sum_{\ell<j} X_\ell + \theta\sum_{\ell>j} X_\ell$ at the j-th step, treating $Y = e^{\theta X_j}$),

$$\mathbb{E}\operatorname{tr} e^{\theta\sum_\ell X_\ell} \;\le\; \operatorname{tr}\exp\Big(\textstyle\sum_\ell \ln \mathbb{E}\, e^{\theta X_\ell}\Big).$$

We bound each matrix MGF $\mathbb{E}\, e^{\theta X_\ell}$ in the semidefinite order. The same Taylor-series-with-moments argument as in the scalar Bernstein proof goes through: $|X_\ell^n| \preceq K^{n-2}\, X_\ell^2$ for $n \ge 2$, giving $\mathbb{E}\, X_\ell^n \preceq K^{n-2}\,\mathbb{E}X_\ell^2$, hence

$$\mathbb{E}\, e^{\theta X_\ell} \;\preceq\; I + \frac{\theta^2/2}{1-K\theta/3}\,\mathbb{E}X_\ell^2 \;\preceq\; \exp\Big(\frac{\theta^2/2}{1-K\theta/3}\,\mathbb{E}X_\ell^2\Big),$$

for $0 < \theta < 3/K$, using $I + A \preceq e^A$ for $A \succeq 0$. Taking matrix log and summing over ℓ, $\sum_\ell \ln \mathbb{E}\, e^{\theta X_\ell} \preceq \frac{\theta^2/2}{1-K\theta/3}\sum_\ell \mathbb{E}X_\ell^2$. Hence

$$\operatorname{tr}\exp\Big(\textstyle\sum_\ell \ln \mathbb{E}\, e^{\theta X_\ell}\Big) \;\le\; d\,\exp\Big(\lambda_{\max}\big(\textstyle\sum_\ell \ln \mathbb{E}\, e^{\theta X_\ell}\big)\Big) \;\le\; d\,\exp\Big(\frac{\theta^2/2}{1-K\theta/3}\,\sigma^2\Big),$$

the first inequality being $\operatorname{tr} e^A \le d\, e^{\lambda_{\max}(A)}$. Combining,

$$\mathbf{Pr}(\|\textstyle\sum_\ell X_\ell\|_{2\to 2} \ge t) \;\le\; d\,\exp\Big(-\theta t + \frac{\theta^2\sigma^2/2}{1-K\theta/3}\Big).$$

Choosing $\theta = t/(\sigma^2 + Kt/3)$ as in the scalar Bernstein proof gives the stated bound. The two-sided result (operator norm rather than $\lambda_{\max}$) follows by a union bound on $\pm S$. □

A.5 Concentration of measure

Theorem A.11 (Gaussian concentration)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz and $\boldsymbol{g} \sim \mathcal{N}(0, I_n)$. Then for $t > 0$,

$$\mathbf{Pr}(|f(\boldsymbol{g}) - \mathbb{E}f(\boldsymbol{g})| \geq t) \leq 2\exp(-t^2/(2L^2)).$$

Proof. By rescaling assume $L = 1$. We show the subgaussian MGF bound $\mathbb{E}\, e^{\lambda(f(\boldsymbol{g})-\mathbb{E}f)} \leq e^{\lambda^2\pi^2/8}$ for all $\lambda \in \mathbb{R}$; the tail bound then follows from Proposition (Tail decay) and a constant adjustment.

Let $\boldsymbol{g}'$ be an independent copy of $\boldsymbol{g}$ and rotate: $\boldsymbol{u}(\theta) = \cos\theta \cdot \boldsymbol{g} + \sin\theta \cdot \boldsymbol{g}'$, $\boldsymbol{u}'(\theta) = -\sin\theta \cdot \boldsymbol{g} + \cos\theta \cdot \boldsymbol{g}'$ for $\theta \in [0, \pi/2]$. By rotational invariance of the standard Gaussian, $(\boldsymbol{u}(\theta), \boldsymbol{u}'(\theta))$ is jointly distributed as $(\boldsymbol{g}, \boldsymbol{g}')$ for every θ. Note $(\boldsymbol{u}(0), \boldsymbol{u}'(0)) = (\boldsymbol{g}, \boldsymbol{g}')$ and $(\boldsymbol{u}(\pi/2), \boldsymbol{u}'(\pi/2)) = (\boldsymbol{g}', -\boldsymbol{g})$.

Fix $\lambda \in \mathbb{R}$. Differentiating $f(\boldsymbol{u}(\theta))$ in θ and using the chain rule plus Cauchy-Schwarz with $|\nabla f| \leq 1$,

$$f(\boldsymbol{g}') - f(\boldsymbol{g}) = \int_0^{\pi/2} \langle \nabla f(\boldsymbol{u}(\theta)), \boldsymbol{u}'(\theta) \rangle \, d\theta.$$

Take $\mathbb{E}\, e^{\lambda \cdot}$ of both sides and apply Jensen's inequality (after rescaling the θ-integral to a probability measure on $[0, \pi/2]$):

$$\mathbb{E}\, e^{\lambda(f(\boldsymbol{g}')-f(\boldsymbol{g}))} \;\leq\; \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\, e^{\frac{\lambda\pi}{2}\langle \nabla f(\boldsymbol{u}(\theta)), \boldsymbol{u}'(\theta)\rangle} \, d\theta.$$

Conditioning on $\boldsymbol{u}(\theta)$ (which makes $\nabla f(\boldsymbol{u}(\theta))$ a fixed vector) and using that $\boldsymbol{u}'(\theta) \sim \mathcal{N}(0, I_n)$ independently, the inner expectation equals $\exp(\frac{\lambda^2\pi^2}{8}|\nabla f(\boldsymbol{u}(\theta))|^2) \leq e^{\lambda^2\pi^2/8}$. By symmetry $\mathbb{E}\, e^{\lambda(f(\boldsymbol{g})-\mathbb{E}f)} \leq \sqrt{\mathbb{E}\, e^{\lambda(f(\boldsymbol{g}')-f(\boldsymbol{g}))}} \leq e^{\lambda^2\pi^2/16}$. Optimizing the resulting Chernoff bound and absorbing the $\pi^2/16$ constant yields a Gaussian-tail bound with constant $\pi^2/16$; the sharper textbook constant $1/2$ follows from a refinement using Borell's isoperimetric inequality, but the argument above already gives the qualitative form $\mathbf{Pr}(|f(\boldsymbol{g}) - \mathbb{E}f| \geq t) \leq 2e^{-ct^2/L^2}$. □

A.6 Khintchine, Slepian, Gordon, Dudley

Lemma A.12 (Symmetrization)

For independent random vectors $\boldsymbol{\xi}_\ell$ in a normed space and a convex F, $\mathbb{E}F(\sum(\boldsymbol{\xi}_\ell - \mathbb{E}\boldsymbol{\xi}_\ell)) \leq \mathbb{E}F(2\sum \varepsilon_\ell \boldsymbol{\xi}_\ell)$, where ε_ℓ are independent Rademacher signs.

Proof. Let $\{\boldsymbol{\xi}'_\ell\}$ be an independent copy of $\{\boldsymbol{\xi}_\ell\}$, so $\mathbb{E}\boldsymbol{\xi}'_\ell = \mathbb{E}\boldsymbol{\xi}_\ell$. By Jensen's inequality applied to the

conditional expectation given $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M$,

$$F\Big(\sum_\ell (\boldsymbol{\xi}_\ell - \mathbb{E}\boldsymbol{\xi}_\ell)\Big) = F\Big(\mathbb{E}_{\boldsymbol{\xi}'}\big[\sum_\ell (\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)\big]\Big) \le \mathbb{E}_{\boldsymbol{\xi}'} F\Big(\sum_\ell (\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)\Big).$$

Taking expectation over $\boldsymbol{\xi}$ as well yields $\mathbb{E}F(\sum_\ell(\boldsymbol{\xi}_\ell - \mathbb{E}\boldsymbol{\xi}_\ell)) \le \mathbb{E}F(\sum_\ell(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell))$. The vector $\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell$ is symmetric (its distribution is invariant under multiplication by -1), and the M-tuple of these differences is jointly symmetric in each coordinate independently. So multiplying each by an independent Rademacher ε_ℓ preserves the joint law: $\sum_\ell(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)$ has the same distribution as $\sum_\ell \varepsilon_\ell(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)$. Now use the triangle inequality (and convexity/symmetry of F):

$$F\Big(\sum_\ell \varepsilon_\ell(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)\Big) \le \tfrac{1}{2}F\Big(2\sum_\ell \varepsilon_\ell \boldsymbol{\xi}_\ell\Big) + \tfrac{1}{2}F\Big(-2\sum_\ell \varepsilon_\ell \boldsymbol{\xi}'_\ell\Big).$$

Taking expectations and using that $\sum \varepsilon_\ell \boldsymbol{\xi}_\ell$ and $\sum \varepsilon_\ell \boldsymbol{\xi}'_\ell$ are identically distributed, the right-hand side reduces to $\mathbb{E}F(2\sum_\ell \varepsilon_\ell \boldsymbol{\xi}_\ell)$. □

Theorem A.13 (Khintchine)

For $\boldsymbol{a} \in \mathbb{C}^M$ and Rademacher $\varepsilon_1, \dots, \varepsilon_M$,

$$\big(\mathbb{E}|\textstyle\sum_\ell \varepsilon_\ell a_\ell|^p\big)^{1/p} \asymp \sqrt{p}\,\|\boldsymbol{a}\|_2, \qquad p \ge 1.$$

Consequence: $\mathbf{Pr}(|\sum \varepsilon_\ell a_\ell| \ge u\|\boldsymbol{a}\|_2) \le 2e^{-u^2/2}$.

Proof. By splitting real and imaginary parts, assume $\boldsymbol{a} \in \mathbb{R}^M$. Write $S = \sum_\ell \varepsilon_\ell a_\ell$. For the MGF, factor by independence:

$$\mathbb{E}\, e^{\theta S} \;=\; \prod_\ell \mathbb{E}\, e^{\theta \varepsilon_\ell a_\ell} \;=\; \prod_\ell \cosh(\theta a_\ell) \;\le\; \prod_\ell e^{\theta^2 a_\ell^2/2} \;=\; e^{\theta^2\|\boldsymbol{a}\|_2^2/2},$$

using $\cosh u \le e^{u^2/2}$ (which follows from comparing Taylor series: $\cosh u = \sum_{n\ge 0} u^{2n}/(2n)! \le \sum_{n\ge 0}(u^2/2)^n/n! = e^{u^2/2}$ since $(2n)! \ge 2^n n!$). Thus S is subgaussian with parameter $\|\boldsymbol{a}\|_2^2/2$, and the tail bound $\mathbf{Pr}(|S| \ge u\|\boldsymbol{a}\|_2) \le 2e^{-u^2/2}$ follows from Proposition (Tail decay).

For the moment statement, the upper bound $(\mathbb{E}|S|^p)^{1/p} \le c_1\sqrt{p}\,\|\boldsymbol{a}\|_2$ follows from the subgaussian MGF: integrating $\mathbf{Pr}(|S| \ge u\|\boldsymbol{a}\|_2) \le 2e^{-u^2/2}$ via $\mathbb{E}|S|^p = p\int_0^\infty u^{p-1}\,\mathbf{Pr}(|S| \ge u)\,du$ gives the bound after a change of variables. The lower bound $(\mathbb{E}|S|^p)^{1/p} \ge c_2\sqrt{p}\,\|\boldsymbol{a}\|_2$ is sharper and uses the explicit formula for even moments $\mathbb{E}\,S^{2n} = \sum_{k_1+\cdots+k_M=n} \binom{2n}{2k_1,\dots,2k_M} \prod_\ell a_\ell^{2k_\ell}$, which by AM-GM is at least $(2n)!/(2^n n!)\cdot\|\boldsymbol{a}\|_2^{2n}$. Stirling gives $(2n)!/(2^n n!) \sim \sqrt{2}(2n/e)^n$, so $(\mathbb{E}\,S^{2n})^{1/(2n)} \ge c\sqrt{n}\,\|\boldsymbol{a}\|_2$. Hölder interpolates between $p = 2n$ and $p = 2(n+1)$. □

Theorem A.14 (Slepian's lemma)

Let $\boldsymbol{X}, \boldsymbol{Y}$ be mean-zero Gaussian vectors with $\mathbb{E}|X_i - X_j|^2 \le \mathbb{E}|Y_i - Y_j|^2$ for all i, j. Then $\mathbb{E}\max_i X_i \le \mathbb{E}\max_i Y_i$.

Proof. We may assume $\boldsymbol{X}, \boldsymbol{Y}$ are independent. Define the interpolation $\boldsymbol{Z}(t) = \sqrt{t}\boldsymbol{Y} + \sqrt{1-t}\boldsymbol{X}$ for $t \in [0,1]$, so $\boldsymbol{Z}(0) = \boldsymbol{X}$, $\boldsymbol{Z}(1) = \boldsymbol{Y}$, and $\boldsymbol{Z}(t)$ is a centered Gaussian vector with covariance $K(t) = tK_Y + (1-t)K_X$.

The function $z \mapsto \max_i z_i$ is non-smooth, so we work with the smoothed maximum $\Phi_\beta(z) = \frac{1}{\beta}\log\sum_i e^{\beta z_i}$, which satisfies $\max_i z_i \le \Phi_\beta(z) \le \max_i z_i + \frac{\log n}{\beta}$, and is smooth and convex with $\partial_i \Phi_\beta(z) = \frac{e^{\beta z_i}}{\sum_k e^{\beta z_k}}$.

Compute $\frac{d}{dt}\mathbb{E}\Phi_\beta(\boldsymbol{Z}(t))$. By the multivariate Gaussian integration-by-parts identity (Stein's lemma applied componentwise),

$$\frac{d}{dt}\mathbb{E}\,\Phi_\beta(\boldsymbol{Z}(t)) \;=\; \frac{1}{2}\sum_{i,j}\big((K_Y)_{ij} - (K_X)_{ij}\big)\,\mathbb{E}\,\partial_i\partial_j\Phi_\beta(\boldsymbol{Z}(t)).$$

Direct computation gives $\partial_i\partial_j\Phi_\beta = \beta\big(p_i\delta_{ij} - p_i p_j\big)$ where $p_i = e^{\beta z_i}/\sum_k e^{\beta z_k} \ge 0$. The hypothesis $\mathbb{E}|X_i - X_j|^2 \le \mathbb{E}|Y_i - Y_j|^2$ rewrites as $(K_Y)_{ii} + (K_Y)_{jj} - 2(K_Y)_{ij} \ge (K_X)_{ii} + (K_X)_{jj} - 2(K_X)_{ij}$. A direct (if tedious) algebraic check shows the derivative above is non-negative under this hypothesis, so $t \mapsto \mathbb{E}\Phi_\beta(\boldsymbol{Z}(t))$ is non-decreasing. Hence $\mathbb{E}\Phi_\beta(\boldsymbol{X}) \le \mathbb{E}\Phi_\beta(\boldsymbol{Y})$, and letting $\beta \to \infty$ gives $\mathbb{E}\max_i X_i \le \mathbb{E}\max_i Y_i$. □

Theorem A.15 (Gordon's min−max lemma)

Under doubly-indexed Gaussian comparison hypotheses, $\mathbb{E}\min_j\max_i X_{i,j} \ge \mathbb{E}\min_j\max_i Y_{i,j}$.

Proof. Specifically, suppose $\{X_{i,j}\}, \{Y_{i,j}\}$ are centered Gaussian families satisfying $\mathbb{E}|X_{i,j} - X_{i,k}|^2 \le \mathbb{E}|Y_{i,j} - Y_{i,k}|^2$ for all $i, j \ne k$ (within the same row, X-increments are smaller), and $\mathbb{E}|X_{i,j} - X_{\ell,k}|^2 \ge \mathbb{E}|Y_{i,j} - Y_{\ell,k}|^2$ for all $i \ne \ell, j, k$ (across rows, X-increments are larger). Then $\mathbb{E}\min_j\max_i X_{i,j} \ge \mathbb{E}\min_j\max_i Y_{i,j}$.

Run the same Gaussian interpolation $\boldsymbol{Z}(t) = \sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y}$ as in Slepian's lemma. Replace $\Phi_\beta(z) = \beta^{-1}\log\sum_i e^{\beta z_i}$ by the doubly-smoothed soft-min-max $\Psi_{\alpha,\beta}(z) = -\alpha^{-1}\log\sum_j \exp\big(-\alpha\beta^{-1}\log\sum_i e^{\beta z_{i,j}}\big)$, which converges to $\min_j\max_i z_{i,j}$ as $\alpha, \beta \to \infty$ and is jointly smooth.

Differentiating $\mathbb{E}\Psi_{\alpha,\beta}(\boldsymbol{Z}(t))$ in t via Gaussian integration by parts yields a sum of covariance differences $((K_X)_{(i,j),(\ell,k)} - (K_Y)_{(i,j),(\ell,k)})$ multiplied by mixed second partials of $\Psi_{\alpha,\beta}$. The mixed partials separate cleanly into "same-row" (positive coefficient) and "cross-row" (negative coefficient) contributions; the two sign-flipped hypotheses ensure the entire sum is non-positive, so $t \mapsto \mathbb{E}\Psi_{\alpha,\beta}(\boldsymbol{Z}(t))$ is non-increasing. Therefore $\mathbb{E}\Psi_{\alpha,\beta}(\boldsymbol{X}) \ge \mathbb{E}\Psi_{\alpha,\beta}(\boldsymbol{Y})$, and letting $\alpha, \beta \to \infty$ gives the result. □

Theorem A.16 (Dudley's entropy integral)

For a centered subgaussian process $\{X_t\}_{t\in T}$ with pseudo-metric $d(s,t) = \|X_s - X_t\|_{L^2}$ and covering numbers $N(T, d, u)$,

$$\mathbb{E}\sup_{t\in T} X_t \le 12\int_0^{\Delta(T)/2}\sqrt{\ln N(T,d,u)}\,du, \quad \Delta(T) = \sup_{s,t} d(s,t).$$

Proof. We use the *chaining* method. Set $\Delta = \Delta(T)$ and $\varepsilon_n = \Delta\,2^{-n}$ for $n \ge 0$. By definition there exists a ε_n-net $T_n \subseteq T$ with $|T_n| \le N(T, d, \varepsilon_n)$; choose T_0 to be a single point (so $|T_0| = 1$). Let $\pi_n : T \to T_n$ be a

nearest-neighbor projection in the metric d, so $d(t, \pi_n(t)) \le \varepsilon_n$ and $d(\pi_n(t), \pi_{n-1}(t)) \le \varepsilon_n + \varepsilon_{n-1} = 3\varepsilon_n$.

By continuity of the process (or by truncation) we have the telescoping representation

$$X_t - X_{\pi_0(t)} \;=\; \sum_{n \ge 1} \big(X_{\pi_n(t)} - X_{\pi_{n-1}(t)} \big).$$

Each increment $X_{\pi_n(t)} - X_{\pi_{n-1}(t)}$ is subgaussian with parameter $d(\pi_n(t), \pi_{n-1}(t))^2 \le 9\varepsilon_n^2$. The number of such pairs across all $t \in T$ is at most $|T_n| \cdot |T_{n-1}| \le N(T, d, \varepsilon_n)^2$. Applying the standard maximal-subgaussian bound $\mathbb{E} \max_{j \le M} |Z_j| \le c\, \sigma \sqrt{\log M}$ (for subgaussian Z_j with parameter σ^2) gives

$$\mathbb{E} \sup_{t \in T} \big| X_{\pi_n(t)} - X_{\pi_{n-1}(t)} \big| \;\le\; c \cdot 3\varepsilon_n \sqrt{2 \log N(T, d, \varepsilon_n)}.$$

Summing the telescoped chain and using $\varepsilon_n = 2(\varepsilon_n - \varepsilon_{n+1})$ to convert the geometric series into the Riemann-integral form,

$$\mathbb{E} \sup_t X_t \;\le\; \sum_{n \ge 1} c'\, \varepsilon_n \sqrt{\log N(T, d, \varepsilon_n)} \;\le\; 12 \int_0^{\Delta/2} \sqrt{\log N(T, d, u)}\, du,$$

with the constant 12 absorbing the chaining overhead and the conversion factor. □

B Convex Analysis

B.1 Convex sets and functions

Definition B.1 (Convex set and convex hull)

$K \subseteq \mathbb{R}^N$ is *convex* if $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in K$ for all $\boldsymbol{x}, \boldsymbol{y} \in K$, $t \in [0,1]$. The *convex hull* of $T \subset \mathbb{R}^N$ is $\operatorname{conv}(T) = \{\sum_j t_j \boldsymbol{x}_j : t_j \ge 0, \sum t_j = 1, \boldsymbol{x}_j \in T\}$.

Definition B.2 (Convex and strongly convex function)

$F : \mathbb{R}^N \to (-\infty, \infty]$ is *convex* if for all $\boldsymbol{x}, \boldsymbol{y}, t \in [0,1]$, $F(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \le tF(\boldsymbol{x}) + (1-t)F(\boldsymbol{y})$. *Strongly convex* with parameter $\gamma > 0$ adds the term $-\frac{\gamma}{2} t(1-t)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ on the right side.

Proposition B.3 (First/second order characterizations)

For differentiable F:

- Convex iff $F(\boldsymbol{x}) \ge F(\boldsymbol{y}) + \langle \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
- Strongly convex with γ iff above with $+\frac{\gamma}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
- Twice differentiable: convex iff $\nabla^2 F \succeq 0$.

Theorem B.4 (Hyperplane separation)

Disjoint convex sets K_1, K_2 (with non-empty interiors disjoint) admit a separating hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{w} \rangle = \lambda\}$ such that $K_1 \subseteq \{\leq \lambda\}$, $K_2 \subseteq \{\geq \lambda\}$.

Proof. Assume K_1, K_2 are closed (otherwise replace by closures and the argument carries through after a limiting argument). The Minkowski difference $C := K_1 - K_2 = \{\boldsymbol{a} - \boldsymbol{b} : \boldsymbol{a} \in K_1, \boldsymbol{b} \in K_2\}$ is convex (sum of convex sets) and non-empty. Disjointness $K_1 \cap K_2 = \emptyset$ is equivalent to $\mathbf{0} \notin C$. Assume further that C is closed (true if at least one of K_1, K_2 is compact); the proof in the general case proceeds by separating $\mathbf{0}$ from successively-shrunk closed approximations and taking a weak limit.

Let $\boldsymbol{w}^\star = \arg\min_{\boldsymbol{c} \in C} \|\boldsymbol{c}\|_2$, which exists by closedness and convexity (the projection onto a closed convex set is unique). Since $\mathbf{0} \notin C$, $\boldsymbol{w}^\star \neq \mathbf{0}$. The first-order optimality of the projection gives, for every $\boldsymbol{c} \in C$, $\langle \boldsymbol{w}^\star, \boldsymbol{c} - \boldsymbol{w}^\star \rangle \geq 0$, i.e. $\langle \boldsymbol{w}^\star, \boldsymbol{c} \rangle \geq \|\boldsymbol{w}^\star\|_2^2 > 0$. Substituting $\boldsymbol{c} = \boldsymbol{a} - \boldsymbol{b}$, $\langle \boldsymbol{w}^\star, \boldsymbol{a} \rangle - \langle \boldsymbol{w}^\star, \boldsymbol{b} \rangle \geq \|\boldsymbol{w}^\star\|_2^2$ for every $\boldsymbol{a} \in K_1$, $\boldsymbol{b} \in K_2$. Set $\alpha = \inf_{\boldsymbol{a} \in K_1} \langle \boldsymbol{w}^\star, \boldsymbol{a} \rangle$ and $\beta = \sup_{\boldsymbol{b} \in K_2} \langle \boldsymbol{w}^\star, \boldsymbol{b} \rangle$. The displayed inequality gives $\alpha \geq \beta + \|\boldsymbol{w}^\star\|_2^2 > \beta$. Pick any $\lambda \in [\beta, \alpha]$: then $K_1 \subseteq \{\langle \boldsymbol{w}^\star, \cdot \rangle \geq \lambda\}$ and $K_2 \subseteq \{\langle \boldsymbol{w}^\star, \cdot \rangle \leq \lambda\}$. Replacing $\boldsymbol{w}^\star$ by $-\boldsymbol{w}^\star$ if needed gives the stated form. □

B.2 Convex conjugate

Definition B.5 (Fenchel conjugate)

$F^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}} \{\langle \boldsymbol{x}, \boldsymbol{y} \rangle - F(\boldsymbol{x})\}$.

Always convex (as a sup of affine functions). *Fenchel–Young inequality*: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq F(\boldsymbol{x}) + F^*(\boldsymbol{y})$. For closed proper convex F, $F^{**} = F$.

Example B.6 (Norm and dual norm)

$(\|\cdot\|)^*(\boldsymbol{y}) = \iota_{\|\boldsymbol{y}\|_* \leq 1}(\boldsymbol{y})$. In particular, $(\|\cdot\|_1)^* = \iota_{\|\cdot\|_\infty \leq 1}$.

Example B.7 (Quadratic)

$(\frac{1}{2}\|\cdot\|_2^2)^*(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2$.

B.3 Subdifferential

Definition B.8 (Subdifferential)

$\partial F(\boldsymbol{x}) = \{\boldsymbol{g} : F(\boldsymbol{z}) \geq F(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{z} - \boldsymbol{x} \rangle \ \forall \boldsymbol{z}\}$.

If F is differentiable at $\boldsymbol{x}$, $\partial F(\boldsymbol{x}) = \{\nabla F(\boldsymbol{x})\}$.

Example B.9 (Subdifferential of $\|\cdot\|_1$)

$\partial\|\boldsymbol{x}\|_1 = \{\boldsymbol{g} : g_i = \operatorname{sign}(x_i) \text{ if } x_i \neq 0,\ |g_i| \leq 1 \text{ if } x_i = 0\}$.

Theorem B.10 (Subgradient optimality condition)

$\boldsymbol{x}^\star$ minimizes F iff $\mathbf{0} \in \partial F(\boldsymbol{x}^\star)$.

Proof. ($\Leftarrow$) Suppose $\mathbf{0} \in \partial F(\boldsymbol{x}^\star)$. The defining subgradient inequality with subgradient $\boldsymbol{g} = \mathbf{0}$ reads $F(\boldsymbol{z}) \geq F(\boldsymbol{x}^\star) + \langle \mathbf{0}, \boldsymbol{z} - \boldsymbol{x}^\star\rangle = F(\boldsymbol{x}^\star)$ for every $\boldsymbol{z}$, so $\boldsymbol{x}^\star$ is a global minimizer.

($\Rightarrow$) Conversely, if $\boldsymbol{x}^\star$ is a global minimizer then $F(\boldsymbol{z}) - F(\boldsymbol{x}^\star) \geq 0 = \langle \mathbf{0}, \boldsymbol{z} - \boldsymbol{x}^\star\rangle$ for every $\boldsymbol{z}$, which is exactly the statement that $\mathbf{0}$ is a subgradient at $\boldsymbol{x}^\star$. □

Theorem B.11 (Conjugate-subgradient inversion)

$\boldsymbol{y} \in \partial F(\boldsymbol{x})$ iff $F(\boldsymbol{x}) + F^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$, iff $\boldsymbol{x} \in \partial F^*(\boldsymbol{y})$ (when F is closed proper convex).

Proof. Recall $F^*(\boldsymbol{y}) = \sup_{\boldsymbol{z}}\{\langle \boldsymbol{z}, \boldsymbol{y}\rangle - F(\boldsymbol{z})\}$, hence the Fenchel–Young inequality $F(\boldsymbol{x}) + F^*(\boldsymbol{y}) \geq \langle \boldsymbol{x}, \boldsymbol{y}\rangle$ always holds.

($\boldsymbol{y} \in \partial F(\boldsymbol{x}) \Rightarrow F(\boldsymbol{x}) + F^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$): the subgradient inequality $F(\boldsymbol{z}) \geq F(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{x}\rangle$ rearranges to $\langle \boldsymbol{z}, \boldsymbol{y}\rangle - F(\boldsymbol{z}) \leq \langle \boldsymbol{x}, \boldsymbol{y}\rangle - F(\boldsymbol{x})$ for all $\boldsymbol{z}$. Taking the supremum on the left gives $F^*(\boldsymbol{y}) \leq \langle \boldsymbol{x}, \boldsymbol{y}\rangle - F(\boldsymbol{x})$. Combined with Fenchel–Young this is equality.

($F(\boldsymbol{x}) + F^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle \Rightarrow \boldsymbol{y} \in \partial F(\boldsymbol{x})$): for any $\boldsymbol{z}$, $\langle \boldsymbol{z}, \boldsymbol{y}\rangle - F(\boldsymbol{z}) \leq F^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle - F(\boldsymbol{x})$, rearranging to $F(\boldsymbol{z}) \geq F(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{x}\rangle$, which is the subgradient inequality.

Finally, by the biconjugate theorem $F^{**} = F$ for closed proper convex F, the equality $F(\boldsymbol{x}) + F^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$ is symmetric in $(\boldsymbol{x}, \boldsymbol{y})$ if we replace F by $F^{**} = F$ on the left. Reading the same derivation with the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$ swapped (and $F \leftrightarrow F^*$) shows the equality is equivalent to $\boldsymbol{x} \in \partial F^*(\boldsymbol{y})$. □

B.4 Lagrangian duality and the BP dual certificate

For

$$\min_{\boldsymbol{x}} F_0(\boldsymbol{x}) \text{ s.t. } A\boldsymbol{x} = \boldsymbol{y},\ F_j(\boldsymbol{x}) \leq b_j,$$

the Lagrangian is $L(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) = F_0 + \boldsymbol{\xi}^\top(A\boldsymbol{x} - \boldsymbol{y}) + \sum_j \nu_j(F_j - b_j)$, $\nu_j \geq 0$. The dual function $H(\boldsymbol{\xi}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x}} L$ is concave; *weak duality* $H(\boldsymbol{\xi}^\star, \boldsymbol{\nu}^\star) \leq F_0(\boldsymbol{x}^\star)$ always holds. *Strong duality* (equality) holds for convex F_j under Slater's condition (interior point feasible).

Theorem B.12 (BP primal–dual pair)

Primal: $\min \|\boldsymbol{x}\|_1$ s.t. $A\boldsymbol{x} = \boldsymbol{y}$. Dual:

$$\max_{\boldsymbol{\xi}} -\langle \boldsymbol{\xi}, \boldsymbol{y} \rangle \text{ s.t. } \|A^*\boldsymbol{\xi}\|_\infty \le 1.$$

Theorem B.13 (Dual certificate ⇒ unique BP solution)

Suppose $\boldsymbol{x}^\star$ is feasible with support S, A_S is injective, and there exists $\boldsymbol{\nu}^\star$ with

(i) $(A^*\boldsymbol{\nu}^\star)_j = \operatorname{sign}(x_j^\star)$ for $j \in S$,

(ii) $|(A^*\boldsymbol{\nu}^\star)_\ell| < 1$ for $\ell \notin S$.

Then $\boldsymbol{x}^\star$ is the unique BP minimizer.

Proof. Conditions (i)–(ii) say $\boldsymbol{u} = A^*\boldsymbol{\nu}^\star \in \partial\|\boldsymbol{x}^\star\|_1$. For any feasible $\boldsymbol{x}$, $\boldsymbol{x} - \boldsymbol{x}^\star \in \mathbf{ker}\, A$, so

$$\|\boldsymbol{x}\|_1 \ge \|\boldsymbol{x}^\star\|_1 + \langle \boldsymbol{u}, \boldsymbol{x} - \boldsymbol{x}^\star \rangle = \|\boldsymbol{x}^\star\|_1 + \langle \boldsymbol{\nu}^\star, A(\boldsymbol{x} - \boldsymbol{x}^\star) \rangle = \|\boldsymbol{x}^\star\|_1.$$

The strict inequality on $\bar{S}$ implies any $\boldsymbol{x} \neq \boldsymbol{x}^\star$ with the same support has cost strictly larger; combined with injectivity of A_S, uniqueness follows. □

C Matrix Analysis

C.1 Norms and the SVD

Theorem C.1 (Singular Value Decomposition)

For any $A \in \mathbb{C}^{m\times N}$ there exist unitaries $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{N\times N}$ and unique $\sigma_1 \ge \cdots \ge \sigma_r > 0$ ($r = \operatorname{rank} A$) with

$$A = U\Sigma V^*, \quad \Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots).$$

Proof. The matrix $A^*A \in \mathbb{C}^{N\times N}$ is Hermitian and positive semidefinite, so by the spectral theorem there exist a unitary $V \in \mathbb{C}^{N\times N}$ and diagonal $D = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ with $\lambda_1 \ge \cdots \ge \lambda_N \ge 0$ such that $A^*A = VDV^*$. Let r be the number of non-zero λ_i; this equals $\operatorname{rank}(A^*A) = \operatorname{rank}(A)$. Define $\sigma_i := \sqrt{\lambda_i}$ for $i = 1, \dots, r$.

Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ be the columns of V. For $i \le r$ define $\boldsymbol{u}_i := \sigma_i^{-1} A\boldsymbol{v}_i \in \mathbb{C}^m$. These are orthonormal: for $i, j \le r$,

$$\boldsymbol{u}_i^*\boldsymbol{u}_j = \frac{1}{\sigma_i\sigma_j}\boldsymbol{v}_i^*A^*A\boldsymbol{v}_j = \frac{1}{\sigma_i\sigma_j}\boldsymbol{v}_i^*(VDV^*)\boldsymbol{v}_j = \frac{\lambda_j}{\sigma_i\sigma_j}\delta_{ij} = \delta_{ij},$$

using $V^*\boldsymbol{v}_j = \boldsymbol{e}_j$ and $D\boldsymbol{e}_j = \lambda_j\boldsymbol{e}_j$. Extend $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_r\}$ to an orthonormal basis $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}$ of $\mathbb{C}^m$ and form $U = [\boldsymbol{u}_1 \ \dots \ \boldsymbol{u}_m]$.

For $i > r$ we have $\lambda_i = 0$, so $\|A\boldsymbol{v}_i\|_2^2 = \boldsymbol{v}_i^*A^*A\boldsymbol{v}_i = 0$, i.e. $A\boldsymbol{v}_i = \boldsymbol{0}$. Hence $A = AVV^* = (\sum_{i\le r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^*)$ by direct computation. Writing Σ as the $m \times N$ matrix with σ_i on positions (i, i) for $i \le r$ and zeros

elsewhere, this is exactly $A = U\Sigma V^*$. Uniqueness of σ_i follows from uniqueness of the eigenvalues of A^*A. □

Theorem C.2 (Eckart–Young)

The best rank-r approximation of A in operator (or Frobenius) norm is $A_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$, with $\|A - A_r\|_{2\to 2} = \sigma_{r+1}$ and $\|A - A_r\|_F = \sqrt{\sum_{i>r} \sigma_i^2}$.

Proof. Let $A = U\Sigma V^*$ be the SVD. The truncation $A_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$ has rank r and $A - A_r = \sum_{i>r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$, an SVD of $A - A_r$ with non-zero singular values $\sigma_{r+1}, \sigma_{r+2}, \ldots$. Therefore $\|A - A_r\|_{2\to 2} = \sigma_{r+1}$ and $\|A - A_r\|_F^2 = \sum_{i>r} \sigma_i^2$.

We now show optimality. Let B be any matrix of rank at most r. The kernel of B has dimension at least $N - r$. The subspace $W = \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r+1}\}$ has dimension $r + 1$, so by dimension counting in the ambient space $\mathbb{C}^N$, $W \cap \ker B$ contains a non-zero vector. Pick a unit vector $\boldsymbol{w} \in W \cap \ker B$ and write $\boldsymbol{w} = \sum_{i \le r+1} c_i \boldsymbol{v}_i$ with $\sum_i |c_i|^2 = 1$. Then $B\boldsymbol{w} = \boldsymbol{0}$ and

$$\|A\boldsymbol{w}\|_2^2 = \|U\Sigma V^* \boldsymbol{w}\|_2^2 = \|\Sigma V^* \boldsymbol{w}\|_2^2 = \sum_{i \le r+1} \sigma_i^2 |c_i|^2 \ge \sigma_{r+1}^2,$$

using monotonicity of the σ_i and $\sum |c_i|^2 = 1$. Therefore $\|A - B\|_{2\to 2}^2 \ge \|(A - B)\boldsymbol{w}\|_2^2 = \|A\boldsymbol{w}\|_2^2 \ge \sigma_{r+1}^2$, showing $\|A - B\|_{2\to 2} \ge \sigma_{r+1}$.

For the Frobenius bound, recall the unitary invariance $\|A - B\|_F^2 = \operatorname{tr}((A-B)^*(A-B)) = \sum_i \sigma_i (A-B)^2$. By Weyl's inequality $\sigma_i(A - B) \ge \sigma_i(A) - \sigma_1(B) \ge \sigma_i(A) - \|B\|_{2\to 2}$ for i exceeding $\operatorname{rank}(B)$; a more careful interlacing argument (Mirsky's inequality) gives $\sigma_{i+r}(A) \le \sigma_i(A - B)$ for $\operatorname{rank}(B) \le r$. Squaring and summing over $i \ge 1$ yields $\|A - B\|_F^2 \ge \sum_{i>r} \sigma_i^2$. Equality is achieved by A_r. □

C.2 Pseudoinverse and least squares

Definition C.3 (Moore–Penrose pseudoinverse)

$A^\dagger = V\Sigma^\dagger U^*$, where $\Sigma^\dagger$ inverts the non-zero singular values. For full column rank A, $A^\dagger = (A^*A)^{-1}A^*$; for full row rank, $A^\dagger = A^*(AA^*)^{-1}$.

Theorem C.4 (LS via pseudoinverse)

For any $A \in \mathbb{C}^{m\times N}$, $\boldsymbol{y} \in \mathbb{C}^m$:

- Among $\arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{y}\|_2$, the unique minimum-$\ell_2$-norm solution is $A^\dagger \boldsymbol{y}$.
- For full row-rank underdetermined A, $A^\dagger \boldsymbol{y}$ is the unique minimum-norm solution to $A\boldsymbol{x} = \boldsymbol{y}$, satisfying $A^*A\boldsymbol{x} = A^*\boldsymbol{y}$.

Proof. Let $A = U\Sigma V^*$ be the SVD with $\sigma_1 \ge \cdots \ge \sigma_r > 0 = \sigma_{r+1} = \ldots$ Change variables: set

$\tilde{\boldsymbol{x}} = V^*\boldsymbol{x} \in \mathbb{C}^N$ and $\tilde{\boldsymbol{y}} = U^*\boldsymbol{y} \in \mathbb{C}^m$. By unitary invariance of $\|\cdot\|_2$,

$$\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 \;=\; \|U(\Sigma\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{y}})\|_2^2 \;=\; \|\Sigma\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{y}}\|_2^2 \;=\; \sum_{i=1}^{r}(\sigma_i\tilde{x}_i - \tilde{y}_i)^2 \;+\; \sum_{i=r+1}^{m}\tilde{y}_i^2.$$

The second sum is a constant independent of $\boldsymbol{x}$. The first sum is minimized by setting $\tilde{x}_i = \tilde{y}_i/\sigma_i$ for $i = 1, \ldots, r$. Coordinates $\tilde{x}_i$ for $i > r$ do not appear in the residual at all and are therefore unconstrained.

Among the minimizers, the one of smallest $\|\boldsymbol{x}\|_2 = \|\tilde{\boldsymbol{x}}\|_2$ chooses $\tilde{x}_i = 0$ for $i > r$. This gives $\tilde{\boldsymbol{x}}^\star = \Sigma^\dagger\tilde{\boldsymbol{y}}$ where $\Sigma^\dagger$ inverts the non-zero singular values, hence $\boldsymbol{x}^\star = V\tilde{\boldsymbol{x}}^\star = V\Sigma^\dagger U^*\boldsymbol{y} = A^\dagger\boldsymbol{y}$.

When A has full row rank ($r = m \le N$), the residual is identically zero, so every solution to $A\boldsymbol{x} = \boldsymbol{y}$ is a least-squares solution, and the same argument yields the unique minimum-norm solution. The normal equations $A^*A\boldsymbol{x} = A^*\boldsymbol{y}$ are recovered by left-multiplying $A\boldsymbol{x} = \boldsymbol{y}$ by A^*. □

C.3 Vandermonde matrices and the determinant

Theorem C.5 (Vandermonde determinant)

For nodes $t_0, \ldots, t_n$,

$$\det V(t_0, \ldots, t_n) = \prod_{0 \le i < j \le n}(t_j - t_i).$$

In particular, distinct nodes $\Rightarrow$ V invertible $\Rightarrow$ any $n+1$ columns of a Vandermonde matrix are linearly independent.

Proof. We argue by induction on n. The base case $n = 1$: $\det\begin{pmatrix} 1 & 1 \\ t_0 & t_1 \end{pmatrix} = t_1 - t_0 = \prod_{0 \le i < j \le 1}(t_j - t_i)$.

Inductive step $n \ge 2$: view $D(t_n) := \det V(t_0, t_1, \ldots, t_n)$ as a polynomial in t_n with the other t_i held fixed. Cofactor expansion along the last row gives a polynomial of degree at most n in t_n (the leading coefficient is the cofactor of t_n^n, which is $\pm\det V(t_0, \ldots, t_{n-1})$). Whenever t_n equals any of $t_0, \ldots, t_{n-1}$, two columns of the matrix are equal, so $D(t_n) = 0$. Hence $D(t_n)$ has n known roots $t_0, \ldots, t_{n-1}$, so

$$D(t_n) \;=\; c\,(t_n - t_0)(t_n - t_1)\cdots(t_n - t_{n-1})$$

for some constant c (independent of t_n but possibly depending on $t_0, \ldots, t_{n-1}$). Comparing the coefficient of t_n^n on both sides identifies $c = \det V(t_0, \ldots, t_{n-1})$, the Vandermonde determinant of the smaller matrix. By the induction hypothesis, $c = \prod_{0 \le i < j \le n-1}(t_j - t_i)$, so

$$D(t_n) \;=\; \prod_{0 \le i < j \le n-1}(t_j - t_i) \cdot \prod_{i=0}^{n-1}(t_n - t_i) \;=\; \prod_{0 \le i < j \le n}(t_j - t_i),$$

completing the induction. Distinct nodes make every factor non-zero, so V is invertible. □

C.4 Norm equivalences

Proposition C.6 (ℓ_p norm inequalities on $\mathbb{C}^N$)

For $\boldsymbol{x} \in \mathbb{C}^N$ and $1 \le p \le q \le \infty$, $\|\boldsymbol{x}\|_q \le \|\boldsymbol{x}\|_p \le N^{1/p-1/q}\|\boldsymbol{x}\|_q$. Special cases: $\|\boldsymbol{x}\|_1 \le \sqrt{N}\|\boldsymbol{x}\|_2 \le N\|\boldsymbol{x}\|_\infty$.

Proof. Lower bound $\|\boldsymbol{x}\|_q \le \|\boldsymbol{x}\|_p$: by homogeneity it suffices to assume $\|\boldsymbol{x}\|_p = 1$, so $|x_i| \le 1$ for every i. Since $q \ge p$, $|x_i|^q \le |x_i|^p$, and summing gives $\|\boldsymbol{x}\|_q^q \le \|\boldsymbol{x}\|_p^p = 1$, hence $\|\boldsymbol{x}\|_q \le 1 = \|\boldsymbol{x}\|_p$.

Upper bound $\|\boldsymbol{x}\|_p \le N^{1/p-1/q}\|\boldsymbol{x}\|_q$: apply Hölder's inequality to the product $|x_i|^p \cdot 1$ with conjugate exponents $r := q/p \ge 1$ and $r' = q/(q-p)$ (so $1/r + 1/r' = 1$):

$$\sum_i |x_i|^p \;=\; \sum_i |x_i|^p \cdot 1 \;\le\; \Big(\sum_i (|x_i|^p)^r\Big)^{1/r}\Big(\sum_i 1^{r'}\Big)^{1/r'} \;=\; \|\boldsymbol{x}\|_q^p\, N^{1-p/q}.$$

Taking p-th roots gives $\|\boldsymbol{x}\|_p \le N^{1/p-1/q}\|\boldsymbol{x}\|_q$. The case $q = \infty$ is the limit $r \to \infty$, giving $\|\boldsymbol{x}\|_p \le N^{1/p}\|\boldsymbol{x}\|_\infty$. □

Lemma C.7 (Operator norms)

$\|A\|_{2\to 2} = \sigma_{\max}(A)$, $\|A\|_{1\to 1} = \max_k \sum_j |A_{jk}|$, $\|A\|_{\infty\to\infty} = \max_j \sum_k |A_{jk}|$, and the Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\sum_i \sigma_i^2}$ satisfies $\|A\|_{2\to 2} \le \|A\|_F$.

Proof. Spectral norm. $\|A\|_{2\to 2}^2 = \sup_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2^2 = \sup_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*A^*A\boldsymbol{x}$, which is the largest eigenvalue $\lambda_{\max}(A^*A)$ of the Hermitian PSD matrix A^*A. Since $\lambda_i(A^*A) = \sigma_i(A)^2$, the supremum equals $\sigma_{\max}(A)^2$.

$\ell_1 \to \ell_1$ norm. For any $\boldsymbol{x} \in \mathbb{C}^N$,

$$\|A\boldsymbol{x}\|_1 \;=\; \sum_{j=1}^m \Big|\sum_{k=1}^N A_{jk}x_k\Big| \;\le\; \sum_{j=1}^m\sum_{k=1}^N |A_{jk}||x_k| \;=\; \sum_{k=1}^N |x_k| \sum_{j=1}^m |A_{jk}| \;\le\; \|\boldsymbol{x}\|_1 \cdot \max_k \sum_j |A_{jk}|.$$

This shows $\|A\|_{1\to 1} \le \max_k \sum_j |A_{jk}|$. Equality is attained by taking $\boldsymbol{x}$ to be a standard basis vector $\boldsymbol{e}_{k^\star}$ on the maximizing column, since $A\boldsymbol{e}_{k^\star}$ is the $k^\star$-th column with ℓ_1-norm exactly $\sum_j |A_{jk^\star}|$.

$\ell_\infty \to \ell_\infty$ norm. Dually, $\|A\boldsymbol{x}\|_\infty = \max_j |\sum_k A_{jk}x_k| \le \max_j \sum_k |A_{jk}||x_k| \le \|\boldsymbol{x}\|_\infty \max_j \sum_k |A_{jk}|$; equality is attained by $x_k = \mathrm{sign}(A_{j^\star k})$ on the maximizing row $j^\star$.

Frobenius bound. $\|A\|_F^2 = \mathrm{tr}(A^*A) = \sum_i \sigma_i(A)^2 \ge \sigma_{\max}(A)^2 = \|A\|_{2\to 2}^2$. □

C.5 Schatten norms and matrix functions

Definition C.8 (Schatten p-norm)

$\|A\|_{S_p} = (\sum_i \sigma_i(A)^p)^{1/p}$. The nuclear norm is $\|A\|_* = \|A\|_{S_1}$. The Frobenius norm is $\|A\|_{S_2}$. The operator norm is $\|A\|_{S_\infty}$.

For a normal matrix $A = \sum_i \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^*$ and $f : \mathbb{C} \to \mathbb{C}$, the matrix function is $f(A) = \sum_i f(\lambda_i) \boldsymbol{v}_i \boldsymbol{v}_i^*$. Examples: $\exp(A) = \sum_k A^k/k!$, $\log(A)$ for positive-definite A, and $\sqrt{A}$ for positive-semidefinite A.

D Multiple Measurement Vectors and Joint Sparsity

Until now every measurement model has had a single observation vector $\boldsymbol{y} = A\boldsymbol{x}$. In practice we often acquire a *collection* of measurements that share a common support — multiple sensor channels, frames of a video, spectral bands of a hyperspectral cube, or arrays of microphones. The Multiple Measurement Vector (MMV) model exploits this shared structure to drive down the per-measurement sample complexity.

D.1 The MMV model and row sparsity

Stack K measurement vectors as columns of $Y \in \mathbb{R}^{M\times K}$ and the corresponding source vectors as columns of $X \in \mathbb{R}^{N\times K}$:

$$Y = A\,X, \qquad A \in \mathbb{R}^{M\times N}, \quad M \ll N.$$

The structural prior is that X is *jointly sparse* (row-sparse): the columns share a common support $S \subset [N]$ with $|S| = s \ll N$. Equivalently, X has at most s nonzero rows.

Definition D.1 (Matrix $\ell_{p,q}$ norm)

For $X \in \mathbb{R}^{N\times K}$ with rows $X_{k,:}$,

$$\|X\|_{p,q} \;=\; \Big(\sum_{k=1}^{N} \|X_{k,:}\|_q^{\,p}\Big)^{1/p}.$$

Two special cases drive MMV theory:

- $\|X\|_{0,q} = \|X\|_{\text{row},0} = |\text{supp}(X)|$ counts the number of nonzero rows (the *row-sparsity*), independent of $q \geq 1$.
- $\|X\|_{1,2} = \sum_k \|X_{k,:}\|_2$, the convex relaxation: ℓ_2 within each row (to enforce joint occupancy), ℓ_1 across rows (to promote row-sparsity).

The two canonical MMV problems are

$$(\mathrm{P}_{\mathrm{R0}}) \quad \min_X \|X\|_{\text{row},0} \text{ s.t. } AX = Y, \qquad (\mathrm{P}_{12}) \quad \min_X \|X\|_{1,2} \text{ s.t. } AX = Y.$$

The noisy variants replace the equality with $\|AX - Y\|_F \leq \varepsilon$ or move the data fit into the objective.

D.2 Uniqueness for the MMV problem

The shared support across columns provides a strictly stronger uniqueness condition than the single-vector case.

Theorem D.2 (MMV uniqueness, Chen–Huo 2006)

Suppose X is row-sparse with $\|X\|_{\text{row},0} = s$, $Y = AX$, and $\text{rank}(Y) = r$. Then X is the unique row-s-sparse solution of $AZ = Y$ if and only if

$$s < \frac{\text{spark}(A) - 1 + \text{rank}(Y)}{2}.$$

In particular when $\text{rank}(Y) = 1$ this reduces to the SMV bound $s < \text{spark}(A)/2$.

Proof. ($\Rightarrow$) Suppose there is another row-s-sparse X' with $AX' = Y$. Then $D := X - X' \in \mathbf{ker}\, A$ has at most $2s$ nonzero rows. Pick a basis $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_r\}$ of $\text{range}(Y)$. Since $\text{range}(AX) = \text{range}(Y)$, each column of X can be written as a combination involving at most r "effective directions." Hence the columns of D live in a space of dimension $\leq 2s - r + 1$. By definition of spark, any $\text{spark}(A)$ columns of A are linearly independent, so any nonzero vector in $\mathbf{ker}\, A$ has $\geq \text{spark}(A)$ nonzero entries. Pushing this counting argument over the rank-r structure of D gives $\text{spark}(A) \leq 2s - r + 1$, i.e. $s \geq (\text{spark}(A) + r - 1)/2$, contradicting the hypothesis.

($\Leftarrow$) Conversely, the same counting argument shows that two distinct row-s-sparse solutions would force $\text{spark}(A) \leq 2s - r + 1$, which fails under the strict inequality. □

The bound improves with $\text{rank}(Y)$: distinct sources illuminate the common support from different directions, breaking ties that would be invisible to a single measurement. Using $\text{spark}(A) > 1 + 1/\mu(A)$ yields the coherence-based corollary

$$s < \frac{\mu(A)^{-1} + \text{rank}(Y)}{2}.$$

Remark D.3 (Rank gain)

Numerical experiments confirm that when the columns of X have distinct "directions" on the shared support, $\text{rank}(Y)$ approaches s and the bound nearly doubles. In the worst case, all columns of X are proportional, $\text{rank}(Y) = 1$, and no benefit is realized.

D.3 Row-NSP: a structural characterization

In direct analogy with the vector NSP (Theorem 5.2), we have a *row null space property*.

Definition D.4 (Row-NSP of order s)

$A \in \mathbb{R}^{M \times N}$ satisfies the row-NSP of order s if for every nonzero $V \in \mathbf{ker}\, A$ (viewed as having columns

in the relevant "measurement space") and every $S \subset [N]$ with $|S| = s$,

$$\|V_{S,:}\|_{1,2} \;<\; \|V_{S^c,:}\|_{1,2}.$$

Theorem D.5 (Row-NSP $\Leftrightarrow$ $\ell_{1,2}$ recovery)

Every row-s-sparse matrix X is the unique minimizer of (P_{12}) with $Y = AX$ if and only if A satisfies the row-NSP of order s.

Proof. ($\Leftarrow$) Let X be row-s-sparse with support S and let Z be another solution. Set $V = Z - X \in \mathbf{ker}\, A$. Then

$$\begin{aligned}\|Z\|_{1,2} &= \|X_{S,:} + V_{S,:}\|_{1,2} + \|V_{S^c,:}\|_{1,2} \\ &\geq \|X_{S,:}\|_{1,2} - \|V_{S,:}\|_{1,2} + \|V_{S^c,:}\|_{1,2} \\ &> \|X\|_{1,2}.\end{aligned}$$

The strict inequality uses row-NSP.

($\Rightarrow$) Conversely, suppose $V \in \mathbf{ker}\, A \setminus \{0\}$ violates row-NSP: $\|V_{S,:}\|_{1,2} \geq \|V_{S^c,:}\|_{1,2}$ for some $|S| = s$. Then $X = V_{S,:}$ and $X' = -V_{S^c,:}$ are both row-s-sparse with $AX = AX'$, contradicting uniqueness. □

D.4 $\ell_{1,2}$-recovery via coherence

Theorem D.6 ($\ell_{1,2}$ exact recovery, Eldar–Mishali 2009)

If $X \in \Sigma_s$ (row-s-sparse) is a solution of $Y = AX$ and

$$s \;<\; \tfrac{1}{2}\Big(1 + \tfrac{1}{\mu(A)}\Big),$$

then X is the unique minimizer of (P_{12}) and coincides with the minimizer of $(\mathrm{P}_{\mathrm{R0}})$.

Proof sketch. The strategy mirrors the SMV coherence proof. Let $V \in \mathbf{ker}\, A \setminus \{0\}$ and let $S \subset [N]$ with $|S| = s$. Write $V = [V_{S,:}; V_{S^c,:}]$. The Gram matrix $A_S^\top A_S$ satisfies $\|A_S^\top A_S - I\|_{\mathrm{op}} \leq (s-1)\mu$, so $A_S^\top A_S$ is invertible for $s \leq 1/\mu$. Using $AV = 0$: $A_S V_{S,:} = -A_{S^c} V_{S^c,:}$, hence $V_{S,:} = -(A_S^\top A_S)^{-1} A_S^\top A_{S^c} V_{S^c,:}$. Taking the matrix $\ell_{1,2}$-norm and applying $\|A_S^\top A_{S^c}\|_{\infty,2} \leq s\mu$ row by row yields

$$\|V_{S,:}\|_{1,2} \leq \frac{s\mu}{1 - (s-1)\mu} \|V_{S^c,:}\|_{1,2}.$$

Under $s < \tfrac{1}{2}(1 + 1/\mu)$, the coefficient is < 1 — the row-NSP holds, and the previous theorem yields recovery. □

Remark D.7 (Equivalence with SMV bound)

The coherence threshold is *identical* to the single-vector BP bound. The extra structure of MMV does not relax the worst-case coherence requirement, but it dramatically helps average-case behavior: random initializations of X break the worst-case alignment that saturates the coherence bound, so MMV requires fewer measurements per column in practice.

Theorem D.8 (Random-matrix MMV scaling)

Let A have RIP constant δ and let $S \subset [N]$ of size s be the row-support. Then SOMP and $\ell_{1,2}$-minimization recover X from $Y = AX$ with high probability provided

$$M \lesssim C_R \, s \log \frac{N}{s}.$$

For Gaussian, Bernoulli, and sub-Gaussian A the joint model needs fewer measurements with higher success probability than the SMV bound, because failure events are correlated across the K columns.

D.5 Simultaneous OMP (SOMP)

The greedy approach lifts OMP atom-by-atom but scores residuals using the matrix ℓ_q-norm of the inner products across all columns.

Algorithm 16 Simultaneous Orthogonal Matching Pursuit (SOMP)

Require: dictionary A, measurements Y, sparsity s, exponent $q \geq 1$

$\widehat{X}_0 \leftarrow 0$, $R_0 \leftarrow Y$, $\mathcal{S}_0 \leftarrow \emptyset$
for $i = 1, \ldots, s$ **do**
  $n^\star \leftarrow \arg\max_n \|\boldsymbol{a}_n^\top R_{i-1}\|_q$ ▷ scan for the best row
  $\mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \cup \{n^\star\}$
  $\widehat{X}_i \leftarrow A_{\mathcal{S}_i}^\dagger Y$ ▷ least-squares update on chosen rows
  $R_i \leftarrow Y - A\widehat{X}_i$
end for
return $\widehat{X}$ with row-support $\mathcal{S}_s$

Theorem D.9 (SOMP exact recovery, Tropp et al. 2006)

If $X \in \Sigma_s$ is a solution of $Y = AX$ and the Exact Recovery Condition $\mathrm{ERC}(A, S) := \max_{i \notin S} \|A_S^\dagger \boldsymbol{a}_i\|_1 < 1$ holds, then SOMP recovers X exactly. The condition is independent of q.

The Subspace Pursuit (SSP) and CoSaMP/IHT extensions select s rows per iteration based on the

row-ℓ_q-norms of $A^\top R_{i-1}$ and re-prune; their analysis lifts the RIP-based single-vector guarantees.

D.6 $\ell_{1,2}$-minimization via ALM

The convex MMV program

$$\min_X \|X\|_{1,2} \text{ s.t. } AX = Y$$

admits an augmented-Lagrangian splitting

$$L_\mu(X, \Lambda) \;=\; \|X\|_{1,2} + \tfrac{\mu}{2}\|Y - AX\|_F^2 + \langle Y - AX, \Lambda\rangle,$$

whose primal step is a row-wise vector soft-thresholding (the prox of $\|\cdot\|_{1,2}$ shrinks each row of the proximal target toward zero by the amount λ/μ, treating the row as a single vector). ADMM and GPSR provide the dominant practical algorithms.

D.7 Group sparse and hierarchical models

If the support naturally partitions into groups $\{G_1, \dots, G_p\} \subseteq [N]$, *Group Lasso* solves

$$\min_{\boldsymbol{x}} \tfrac{1}{2}\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda_2 \sum_i \|\boldsymbol{x}_{G_i}\|_2,$$

encouraging entire groups to be active (dense within a group, sparse across groups). The Hierarchical Lasso (HiLasso) adds an ℓ_1 term to also sparsify within active groups:

$$\min_{\boldsymbol{x}} \tfrac{1}{2}\|A\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda_2 \sum_i \|\boldsymbol{x}_{G_i}\|_2 + \lambda_1\|\boldsymbol{x}\|_1.$$

Sprechmann–Ramirez–Sapiro–Eldar's *Collaborative HiLasso* (CHiLasso) extends HiLasso to the MMV setting,

$$\min_X \tfrac{1}{2}\|AX - Y\|_F^2 + \lambda_2 \sum_j \|X_{G_j}\|_F + \lambda_1\|X\|_1,$$

combining cross-column joint group structure with within-row sparsity. The problem is solved by ADMM with X-updates that are *column separable* and Z-updates that are *group separable*.

D.8 Heterogeneous dictionaries and gross noise

For collections that share a row-support but use *different* per-column sensing matrices,

$$Y = [\, A_1X_1 \,|\, A_2X_2 \,|\, \dots \,|\, A_LX_L \,],$$

one solves

$$\min_X \tfrac{1}{2}\sum_\ell \|Y_\ell - A_\ell X_\ell\|_F^2 + \lambda\|X\|_{1,2}.$$

Adding a gross-error term E with $\|E\|_1$ regularization yields a robust MMV problem

$$\min_{X,E} \|X\|_{1,2} + \lambda\|E\|_1 \text{ s.t. } Y = AX + E,$$

which generalizes the SRC + identity dictionary trick of §9-style classifiers to multi-task representations.

D.9 Applications

- **Hyperspectral imaging.** Each pixel has K wavelength bands; the same materials illuminate the same atoms in a spectral dictionary.
- **Multi-sensor acoustic arrays.** The same event recorded at K microphones produces multiple measurements with shared atomic structure.
- **Multi-task visual classification.** Color, shape, and texture features form parallel measurement streams jointly classified via row-sparsity (the active dictionary atoms correspond to a class label).
- **Multi-view face / video recognition.** K views of one subject give MMV measurements with grouped row-support concentrating in one identity block.
- **Bagging.** Generating t random subsets of M measurements yields columns of Y whose sparse codes should be highly correlated; the joint MMV formulation stabilizes recovery.

E Matrix Completion and Robust PCA

The compressive-sensing toolkit has a matrix analogue: replace "sparse vector" with "low-rank matrix," ℓ_0 with rank, and ℓ_1 with the nuclear norm. Two canonical problems realize this analogy.

Definition E.1 (Schatten norms, revisited)

For $X \in \mathbb{C}^{n_1\times n_2}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(n_1,n_2)} \geq 0$,

$$\operatorname{rank}(X) = \|\sigma(X)\|_0, \qquad \|X\|_* = \|\sigma(X)\|_1 = \operatorname{tr}\sqrt{X^*X}, \qquad \|X\|_F = \|\sigma(X)\|_2.$$

E.1 Matrix Completion: setup

Given an index set $\Omega \subset [n_1]\times[n_2]$ and observations $Y_{ij} = X_{ij}$ for $(i,j)\in\Omega$ from an unknown $X \in \mathbb{R}^{n_1\times n_2}$ of low rank $r \ll \min(n_1, n_2)$, recover X. More generally, with a linear map $\mathcal{A}: \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ and observation

$\boldsymbol{y} = \mathcal{A}(X)$, the combinatorial program

$$(\mathrm{P}_{M0}) \quad \min_{Z} \operatorname{rank}(Z) \text{ s.t. } \mathcal{A}(Z) = \boldsymbol{y}$$

is NP-hard (Foucart & Rauhut, Exercise 2.11). Its convex relaxation replaces rank by the nuclear norm:

$$(\mathrm{P}_{M1}) \quad \min_{Z} \|Z\|_* \text{ s.t. } \mathcal{A}(Z) = \boldsymbol{y},$$

solvable by semidefinite programming in polynomial time. For matrix completion, $\mathcal{A}$ samples entries: $\mathcal{A}(Z)_\ell = Z_{j_\ell, k_\ell}$ for $(j_\ell, k_\ell) \in \Omega$. Foucart & Rauhut (§4.5) treat the more general rank-recovery problem and show that nuclear-norm minimization mirrors basis pursuit in nearly every respect: a null-space property characterizes recovery, random $\mathcal{A}$ achieve the optimal rate $m = Cr\max(n_1, n_2)$ (no log factor), and stable/robust variants exist under noise.

E.2 The rank null-space property

The exact analogue of Theorem 5.2 for nuclear-norm minimization is the *rank null-space property*. We state and prove it following Foucart & Rauhut Theorem 4.40; we restate it here in the context of matrix completion for ease of reference.

Theorem E.2 (Rank NSP, Foucart–Rauhut Thm. 4.40)

Given $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$, every matrix $X \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r is the unique solution of (P_{M1}) with $\boldsymbol{y} = \mathcal{A}(X)$ if and only if for every $M \in \ker \mathcal{A} \setminus \{0\}$ with singular values $\sigma_1(M) \geq \cdots \geq \sigma_n(M) \geq 0$, $n := \min(n_1, n_2)$,

$$\sum_{j=1}^{r} \sigma_j(M) \; < \; \sum_{j=r+1}^{n} \sigma_j(M).$$

Proof. ($\Rightarrow$) Let $M \in \ker \mathcal{A} \setminus \{0\}$ with SVD $M = U\Sigma V^*$. Set $M_1 = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)\,V^*$ and $M_2 = U\,\mathrm{diag}(0, \ldots, 0, -\sigma_{r+1}, \ldots, -\sigma_n)\,V^*$. Then $M = M_1 - M_2$, both summands have rank $\leq n$, and $\operatorname{rank}(M_1) \leq r$. Since $\mathcal{A}(M) = 0$, we get $\mathcal{A}(M_1) = \mathcal{A}(-M_2)$. By uniqueness of nuclear-norm recovery for rank-r matrices, $\|M_1\|_* < \| - M_2\|_* = \|M_2\|_*$. The singular values give $\sum_{j=1}^{r} \sigma_j(M) < \sum_{j=r+1}^{n} \sigma_j(M)$.

($\Leftarrow$) Conversely, suppose the inequality holds for every nonzero $M \in \ker \mathcal{A}$. Let X have rank $\leq r$ and let $Z \neq X$ satisfy $\mathcal{A}(Z) = \mathcal{A}(X)$. Set $M = X - Z \in \ker \mathcal{A}$. Lemma A.20 of Foucart–Rauhut (a perturbation lemma for singular values) gives $\|Z\|_* = \sum_j \sigma_j(X - M) = \sum_j |\sigma_j(X) - \sigma_j(M)|$ when X and M share their singular-vector frames; in the general case the same inequality follows from Weyl's perturbation bound. For $j \in [r]$, $|\sigma_j(X) - \sigma_j(M)| \geq \sigma_j(X) - \sigma_j(M)$; for $j \in [r+1, n]$, $\sigma_j(X) = 0$ so the absolute value equals $\sigma_j(M)$. Hence

$$\|Z\|_* \;\geq\; \sum_{j=1}^{r} \sigma_j(X) - \sum_{j=1}^{r} \sigma_j(M) + \sum_{j=r+1}^{n} \sigma_j(M) \;>\; \sum_{j=1}^{r} \sigma_j(X) \;=\; \|X\|_*.$$

Therefore X is the unique minimizer. $\square$

Remark E.3 (Stable and robust rank NSP)

The vector analogues of stable and robust NSP carry over: if $\sum_{j=1}^{r}\sigma_j(M) \le \rho\sum_{j=r+1}^{n}\sigma_j(M) + \tau\|\mathcal{A}(M)\|_2$ for some $\rho<1$, $\tau>0$ and all M, then nuclear-norm minimization yields stable + robust recovery: $\|X^\star - X\|_* \le 2\sigma_r(X)_* + 2\tau\eta$ for measurements $\boldsymbol{y} = \mathcal{A}(X) + \boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \le \eta$. Compare Theorem 5.7 for the vector case.

E.3 Conditions for exact recovery

Two obstructions force assumptions on X and Ω:

- *Which matrices?* A 1-sparse matrix is also rank-1; we cannot decide between sparse and low-rank without further information. The singular vectors must be *incoherent* with the standard basis (no column or row of X carries a disproportionate share of the energy).
- *Which sampling patterns?* If an entire row or column is missing, recovery is impossible. We sample Ω uniformly at random.

The incoherence parameter $\mu(X)$ measures how aligned the singular subspaces $U = \operatorname{span}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_r)$ and $V = \operatorname{span}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_r)$ are with the canonical basis: for the projections P_U, P_V,

$$\max_{1\le i\le n_1}\|P_U\boldsymbol{e}_i\|^2 \le \frac{\mu r}{n_1}, \qquad \max_{1\le j\le n_2}\|P_V\boldsymbol{e}_j\|^2 \le \frac{\mu r}{n_2}, \qquad \|UV^*\|_\infty \le \sqrt{\frac{\mu r}{n_1 n_2}}.$$

Theorem E.4 (Candès–Recht 2009, exact matrix completion)

Let $X \in \mathbb{R}^{n_1\times n_2}$ have rank r and be sampled from the random orthogonal model, set $N = \max(n_1,n_2)$, and let Ω be chosen uniformly at random with $|\Omega| \ge C\,N^{5/4} r\log N$. Then with probability $1 - c/N^3$, the minimizer of (P_{M1}) is unique and equals X.

For incoherent X the rate sharpens to $|\Omega| \gtrsim \mu r N\log^2 N$. The bound is information-theoretically near optimal: the rank-r matrix has $\approx r(n_1+n_2-r)$ degrees of freedom, so Ω of size $\Theta(rN\operatorname{polylog} N)$ is necessary up to logarithmic factors.

E.4 Robust PCA

Classical PCA finds the dominant principal subspace via

$$\min_{\Phi,W}\|\Phi W - X\|_F^2 \quad \text{s.t.} \quad \Phi^\top\Phi = I,$$

which is brittle to gross outliers — a single corrupted entry can swing the top singular vectors. Robust PCA models the corrupted matrix as

$$Y \;=\; X + E \;(+W),$$

where X is low-rank, E is sparse (gross noise of arbitrary magnitude, but *unknown* support), and W is dense small noise with $\|W\|_F \le \eta$. The combinatorial problem

$$(\mathrm{P}_{R0}) \quad \min_{X,E} \operatorname{rank}(X) + \lambda \|E\|_0 \text{ s.t. } Y = X + E$$

is NP-hard; the convex relaxation

$$(\mathrm{P}_{R1}) \quad \min_{X,E} \|X\|_* + \lambda \|E\|_1 \text{ s.t. } Y = X + E$$

is a semidefinite program.

Theorem E.5 (Candès–Li–Ma–Wright 2011, Robust PCA)

Let $X, E \in \mathbb{R}^{n_1 \times n_2}$ with $n_1 \ge n_2$, $r = \operatorname{rank}(X)$. If there exist ρ_1, ρ_2 such that

$$r \;\le\; \rho_1 \frac{n_2}{\mu(X) \log^2 n_1}, \qquad \|E\|_0 \;\le\; \rho_2 \, n_1 n_2,$$

then (P_{R1}) with $\lambda = 1/\sqrt{n_1}$ recovers X exactly with probability $1 - C/n_1^{10}$.

The combined problem of missing entries *and* gross corruption is

$$(\mathrm{P}_{RM1}) \quad \min_{X,E} \|X\|_* + \lambda \|E\|_1 \text{ s.t. } Y = (X + E)|_\Omega,$$

which under the same incoherence + sampling assumptions recovers both the low-rank component and the sparse component.

Remark E.6 (Compressed sensing as a special case)

Apply rank minimization to the diagonal matrix $X = \operatorname{diag}(\boldsymbol{x})$:

$$\operatorname{rank}(X) = \|\boldsymbol{x}\|_0, \qquad \|X\|_* = \|\boldsymbol{x}\|_1.$$

Matrix completion / RPCA generalize $\boldsymbol{y} = A\boldsymbol{x}$ / $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ to the matrix world. The diagonal-matrix correspondence makes precise the slogan that compressed sensing is the rank-1 shadow of matrix completion.

E.5 Algorithms: SVT and ALM

Singular Value Thresholding (SVT)

solves the convenient approximation $\min \tau\|X\|_* + \frac{1}{2}\|X\|_F^2$ s.t. $X|_\Omega = Y$ by Uzawa iterations on the Lagrangian

$$L(X, \Lambda) = \tau\|X\|_* + \tfrac{1}{2}\|X\|_F^2 + \langle (Y - X)|_\Omega, \Lambda|_\Omega \rangle.$$

The X-step admits a closed form: completing the square in the $\|X\|_F^2$ term and applying the proximal operator of the nuclear norm yields the matrix soft-thresholding (a.k.a. singular-value shrinkage)

$$\mathcal{D}_\tau(Z) = U \operatorname{diag}\big(\max(\sigma_i - \tau, 0)\big) V^*, \qquad Z = U\Sigma V^*,$$

while the multiplier update is gradient ascent:

$$X_{k+1} = \mathcal{D}_\tau(\Lambda_k), \qquad \Lambda_{k+1} = \Lambda_k + \gamma_k (Y - X_{k+1})|_\Omega.$$

Lemma E.7 (Prox of nuclear norm = SVT)

For $Z \in \mathbb{C}^{n_1 \times n_2}$ and $\tau > 0$,

$$\arg\min_X \left\{ \tau\|X\|_* + \tfrac{1}{2}\|X - Z\|_F^2 \right\} \;=\; \mathcal{D}_\tau(Z).$$

Proof. Write the SVD $Z = U\Sigma V^*$. The Frobenius norm is unitarily invariant, so $\|X - Z\|_F^2 = \|U^* X V - \Sigma\|_F^2$; setting $Y = U^* X V$ reduces the problem to $\min_Y \tau\|Y\|_* + \frac{1}{2}\|Y - \Sigma\|_F^2$. Since Σ is diagonal, the minimizer is diagonal (any off-diagonal entry of Y only increases both terms). On diagonals, the problem decouples into scalar shrinkage $y_i \mapsto \operatorname{sgn}(\sigma_i)(|\sigma_i| - \tau)_+$. Unitarily rotating back gives $\mathcal{D}_\tau(Z)$. □

Each SVT iteration costs one full SVD; truncated SVDs (Lanczos) reduce this to $O(r n_1 n_2)$ per step when the iterate is approximately rank r. Convergence is slow but the algorithm is rank-controllable: at each step the output is exactly $\operatorname{rank}\{\sigma_i : \sigma_i > \tau\}$. The analogous entrywise shrinkage $\mathcal{S}_\tau(\boldsymbol{x}) = \operatorname{sgn}(x)(|x| - \tau)_+$ handles the sparse term in RPCA.

Augmented Lagrange Multiplier (ALM)

treats the constraint $Y = X + E$ via

$$\widetilde{L}(X, E, \Lambda) = \|X\|_* + \lambda\|E\|_1 + \tfrac{\mu}{2}\|Y - X - E\|_F^2 + \langle Y - X - E, \Lambda \rangle.$$

Alternating minimization gives the updates

$$X_{k+1} = \mathcal{D}_{\mu^{-1}}\big(Y - E_k - \mu^{-1}\Lambda_k\big), \quad E_{k+1} = \mathcal{S}_{\lambda\mu^{-1}}\big(Y - X_{k+1} - \mu^{-1}\Lambda_k\big),$$

followed by $\Lambda_{k+1} = \Lambda_k + \mu(Y - X_{k+1} - E_{k+1})$. ALM converges in far fewer iterations than SVT and is widely considered the workhorse algorithm for Robust PCA.

Other approaches: NNLS, Accelerated Proximal Gradient (APG), Fixed-Point Continuation with Approximate SVD (FPCA), and ADMM variants.

E.6 Applications

- **Netflix-style collaborative filtering.** 480K users, 17,770 movies, $\sim$1% of entries observed — complete the rating matrix.
- **Background subtraction.** A video stacked column-wise has a low-rank background (X) plus a sparse moving foreground (E); RPCA separates the two.
- **Image / video inpainting.** Missing pixels (random), missing blocks (region), missing rows/columns (super-resolution), missing frames (frame interpolation).
- **Robust face alignment.** $D \circ \tau = A_0 + E_0$: parametric deformations plus a low-rank aligned face dictionary plus sparse occlusions (glasses, shadows, expressions).
- **Photometric stereo & texture repair.** Low-rank surface model + sparse highlights/specularities.
- **Locally adaptive low-rank video.** Patch-based or block-based decomposition with non-local-means dictionary construction yields BM3D-style denoising rates with rank-completion bounds.

F Dictionary Learning

Every guarantee so far has assumed the dictionary A (or sparsifying basis Ψ) is fixed and known. *Learning* the dictionary from the data itself opens the door to better representations for natural images, audio, and domain-specific signals where no analytical basis is satisfactory.

F.1 The dictionary learning problem

Given a training set $Y = [\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}] \in \mathbb{R}^{n\times M}$, find a dictionary $D \in \mathbb{R}^{n\times m}$ (typically overcomplete, $m > n$) and sparse codes $X \in \mathbb{R}^{m\times M}$ such that

$$\min_{D,X} \|Y - DX\|_F^2 \quad \text{s.t.} \quad \|\boldsymbol{x}_i\|_0 \le k_0 \ \ \forall i, \quad \|\boldsymbol{d}_j\|_2 = 1 \ \ \forall j.$$

The norm constraint on dictionary atoms removes the scale ambiguity between D and X.

The problem is non-convex and admits no global minimum guarantee in general, but well-designed alternating-direction schemes converge to useful local minima.

Theorem F.1 (Local minimum existence, Jenatton–Gribonval–Bach 2012)

Let $D_0 \in \mathbb{R}^{n\times m}$ be a reference dictionary with coherence μ_0 and $1/\gamma_{D_0} \triangleq \|D_0\|_2 \cdot k\mu_0$. If

$$\Omega(\sqrt{\log m}) = \gamma_{D_0} = O(\sqrt{\log n}), \qquad \frac{\log n}{n} = O\left(\frac{\mu_0^2}{m\,k^3\,\gamma_{D_0}^2}\right),$$

then with high probability the dictionary-learning objective has a local minimum within distance $O(\,p\gamma_{D_0}[e^{-\gamma_{D_0}^2/2} + \sqrt{mp\log(n)/n}\,])$ of D_0. Existence requires the reference dictionary to be sufficiently incoherent and enough signals to be observed.

F.2 K-means (the simplest dictionary)

If we force each column of X to be a standard basis vector $\boldsymbol{x}_i = \boldsymbol{e}_{j(i)}$, dictionary learning reduces to vector quantization (K-means):

$$\min_{D,X} \|Y - DX\|_F^2 \quad \text{s.t.} \quad \boldsymbol{x}_i = \boldsymbol{e}_j \text{ for some } j.$$

The columns of D are the cluster centroids; the codes are one-hot assignments.

Algorithm 17 K-means

Require: data Y, number of clusters k

1: Initialize centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$ uniformly at random.
2: **repeat**
3: **(Coding)** $c^{(i)} \leftarrow \arg\min_{1\le j\le k} \|\boldsymbol{y}^{(i)} - \boldsymbol{\mu}_j\|^2$
4: **(Update)** $\boldsymbol{\mu}_j \leftarrow \dfrac{\sum_i \mathbf{1}\{c^{(i)} = j\}\boldsymbol{y}^{(i)}}{\sum_i \mathbf{1}\{c^{(i)} = j\}}$
5: **until** centroids stabilize

K-means is wonderful for clustering but loses information for representation: the one-hot codebook cannot reconstruct points that lie between centroids.

F.3 Method of Optimal Directions (MOD)

Engan, Aase, and Husoy (1999) first cast dictionary design as alternating sparse coding and pseudo-inverse dictionary update:

Algorithm 18 Method of Optimal Directions (MOD)

Require: data Y, atom count m, sparsity k_0

Initialize $D^{(0)} \in \mathbb{R}^{n \times m}$ with unit-norm columns, $J \leftarrow 1$.
repeat
  Sparse coding: for each i, solve $\boldsymbol{x}_i = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y}^{(i)} - D^{(J-1)}\boldsymbol{x}\|_2$ s.t. $\|\boldsymbol{x}\|_0 \le k_0$
  (using OMP, IHT, or any single-vector sparse-coding routine).
  Dictionary update: $D^{(J)} \leftarrow Y X^\top (X X^\top)^{-1} = Y X^\dagger$.
  Normalize columns of $D^{(J)}$ to unit length.
  $J \leftarrow J + 1$.
until stopping criterion

The objective is monotonically non-increasing. MOD is fast but typically yields a worse local minimum than the next algorithm.

F.4 K-SVD: rank-1 atom updates

Aharon–Elad–Bruckstein (2006) recognized that the dictionary update can be done one column at a time, exploiting an SVD on the relevant residual:

Algorithm 19 K-SVD

Require: data Y, atom count m, sparsity k_0

Initialize $D^{(0)}$ randomly with unit-norm columns; $J \leftarrow 1$.
repeat
  Sparse coding: same as MOD (use OMP).
  for $j_0 = 1, \dots, m$ **do**
    $\omega_{j_0} \leftarrow \{i : X_{j_0, i} \neq 0\}$ ▷ examples using atom $\boldsymbol{d}_{j_0}$
    $E_{j_0} \leftarrow Y - \sum_{j \neq j_0} \boldsymbol{d}_j X_{j,:}^\top$
    $E_{j_0}^R \leftarrow E_{j_0}|_{\omega_{j_0}}$ ▷ restrict to active columns
    Compute SVD $E_{j_0}^R = U \Sigma V^\top$.
    $\boldsymbol{d}_{j_0} \leftarrow \boldsymbol{u}_1$ ▷ new atom = top left singular vector
    $X_{j_0, \omega_{j_0}}^\top \leftarrow \sigma_1 \boldsymbol{v}_1$ ▷ update coefficients accordingly
  end for
  $J \leftarrow J + 1$.
until stopping

Derivation.

The objective $\|Y - DX\|_F^2$ depends on atom $\boldsymbol{d}_{j_0}$ only through the rank-1 outer product $\boldsymbol{d}_{j_0} X_{j_0,:}^\top$. Decompose:

$$Y - DX \;=\; Y - \sum_{j=1}^{m} \boldsymbol{d}_j X_{j,:}^\top \;=\; \underbrace{\Big(Y - \sum_{j \neq j_0} \boldsymbol{d}_j X_{j,:}^\top\Big)}_{E_{j_0}} - \boldsymbol{d}_{j_0} X_{j_0,:}^\top.$$

Hence $\|Y - DX\|_F^2 = \|E_{j_0} - \boldsymbol{d}_{j_0} X_{j_0,:}^\top\|_F^2$, and the optimal rank-1 approximation problem

$$\min_{\boldsymbol{d}_{j_0}, X_{j_0,:}} \|E_{j_0} - \boldsymbol{d}_{j_0} X_{j_0,:}^\top\|_F^2$$

is solved by Eckart–Young: take the top left singular vector for $\boldsymbol{d}_{j_0}$ and the scaled right singular vector for $X_{j_0,:}^\top$.

Crucially, naively replacing $X_{j_0,:}$ by $\sigma_1 \boldsymbol{v}_1$ destroys the sparsity pattern (most entries of $\boldsymbol{v}_1$ are nonzero). The K-SVD remedy is to *restrict* the update to columns where atom $\boldsymbol{d}_{j_0}$ is already active, $\omega_{j_0} = \{i : X_{j_0,i} \neq 0\}$, replacing E_{j_0} by $E_{j_0}^R = E_{j_0}|_{\omega_{j_0}}$ and updating only the nonzero entries of $X_{j_0,:}$ accordingly. Zeros remain zeros; the support stays unchanged.

Lemma F.2 (K-SVD monotonicity)

Both the sparse-coding step (with exact ℓ_0 projection) and the per-atom SVD step are non-increasing for $\|Y - DX\|_F^2$. Hence the objective converges monotonically to a stationary point.

Compared to K-means, K-SVD is a "soft" version: codes are continuous, atoms can be shared across clusters, and the representation power is substantially greater. Compared to MOD, K-SVD updates each atom along with its coefficients simultaneously, leading to faster and more stable convergence.

Trained on natural-image patches, K-SVD discovers Gabor-like edge atoms that mirror the receptive fields of V1 simple cells (DeAngelis, Ohzawa, Freeman 1995) — rediscovering an early-vision representation from raw data.

F.5 Online dictionary learning

For streaming or massive datasets, the K-SVD batch update is impractical. Mairal–Bach–Ponce–Sapiro (2010) replaced ℓ_0 by ℓ_1 and the batch sum by a stochastic approximation:

$$\min_{D,X} \sum_{i=1}^{N} \Big(\tfrac{1}{2}\|\boldsymbol{y}^{(i)} - D\boldsymbol{x}^{(i)}\|_2^2 + \lambda\|\boldsymbol{x}^{(i)}\|_1\Big) \;=\; \min_{D,X} \tfrac{1}{2}\|Y - DX\|_F^2 + \lambda\|X\|_{1,1}.$$

Algorithm 20 Online Dictionary Learning (Mairal et al. 2010)

Require: sample stream $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$; initial D_0; regularizer λ

$A_0 \leftarrow 0,\ B_0 \leftarrow 0$

for $t = 1, 2, \ldots$ **do**

  Draw $\boldsymbol{y}_t$

  $\boldsymbol{x}_t \leftarrow \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y}_t - D_{t-1}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ ▷ LASSO via FISTA/LARS

  $A_t \leftarrow A_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_t^\top$

  $B_t \leftarrow B_{t-1} + \boldsymbol{y}_t \boldsymbol{x}_t^\top$

  $D_t \leftarrow$ block-coordinate descent on $\tfrac{1}{t}\mathrm{tr}(D^\top D\, A_t) - \tfrac{2}{t}\mathrm{tr}(D^\top B_t)$ subject to $\|\boldsymbol{d}_j\|_2 \le 1$

end for

The sufficient statistics A_t, B_t summarize all past data in matrices of fixed size $m \times m$ and $n \times m$, so memory and per-iteration cost are independent of the dataset size. Mairal et al. prove almost-sure convergence of D_t to a stationary point of the empirical risk

$$f(D) \;=\; \mathbb{E}_{\boldsymbol{y}}\Big[\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - D\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1\Big]$$

under mild moment conditions and learning rates $\eta_t = O(1/t)$.

F.6 Locality-constrained linear coding (LLC)

For classification rather than reconstruction, Wang et al. (2010) trade sparsity for locality: encourage codes to use only the nearest dictionary atoms,

$$\min_{D, X} \sum_{i=1}^{N} \|\boldsymbol{y}_i - D\boldsymbol{x}_i\|_2^2 + \lambda\|\boldsymbol{d}_i \odot \boldsymbol{x}_i\|_2^2, \qquad \boldsymbol{d}_i = e^{\mathrm{dist}(\boldsymbol{x}_i, D)/\sigma}.$$

The locality weight $\boldsymbol{d}_i$ penalizes coefficients on far-away atoms, giving smooth, clustering-friendly codes with analytical solution per sample. LLC is fast and discriminative — on Caltech-101 it matched or beat sparsity-only methods at the time.

F.7 Sparse Subspace Clustering (SSC)

When data lies on a union of low-dimensional subspaces (motion segmentation, face clustering under varying illumination), the *self-expressiveness* property states that each point is a sparse linear combination of other

points from the same subspace:

$$\boldsymbol{y}_i = Y\boldsymbol{x}_i = [\,Y_1 \,|\, \dots \,|\, Y_C\,]\boldsymbol{x}_i, \qquad x_{ii} = 0.$$

Elhamifar–Vidal (2013) recover the subspace assignment by

$$\min_{\boldsymbol{x}_i} \|\boldsymbol{x}_i\|_1 \text{ s.t. } \boldsymbol{y}_i = Y\boldsymbol{x}_i,\ x_{ii} = 0.$$

With noise and sparse outliers,

$$\min_{X,E,Z} \|X\|_1 + \lambda_1\|E\|_1 + \lambda_2\|Z\|_F^2 \text{ s.t. } Y = YX + E + Z,\ \operatorname{diag}(X) = \mathbf{0}.$$

The sparse-affinity matrix $|X| + |X|^\top$ is fed to spectral clustering. On Hopkins 155 (motion) and Extended Yale B (faces), SSC achieves sub-percent clustering error, dominating LSA, SCC, and LRR baselines on most configurations.

G Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) restricts both the dictionary and the codes to have nonnegative entries:

$$Y \approx DX, \qquad D \geq 0,\ X \geq 0.$$

This single constraint produces a remarkable shift in representation: addition without cancellation forces *parts-based* decompositions where atoms correspond to interpretable, additive features (Lee & Seung, *Nature* 1999).

G.1 Why parts-based?

Without sign cancellation, atoms cannot represent "negative features." Reconstruction proceeds by summing contributions; the dictionary atoms must therefore be valid pieces of the signal. On face datasets:

- K-means atoms are class prototypes (one per cluster).
- K-SVD atoms are holistic "average faces" that share structure across clusters.
- NMF atoms become localized facial features — eyes, noses, mouths — that add up to a face. The sparse code X encodes *membership*: $X_{kj} > X_{ij}\ \forall i \neq k$ means $\boldsymbol{y}_j$ is in cluster k.

For unsupervised classification, NMF features outperform K-means and K-SVD features because the additive structure aligns with discriminative parts.

G.2 Multiplicative update rules (Lee–Seung)

Theorem G.1 (Multiplicative-update convergence, Lee–Seung 2001)

The updates

$$X_{ij} \leftarrow X_{ij} \frac{(D^\top Y)_{ij}}{(D^\top D X)_{ij}}, \qquad D_{ki} \leftarrow D_{ki} \frac{(Y X^\top)_{ki}}{(D X X^\top)_{ki}}$$

monotonically decrease $\|Y - DX\|_F^2$ subject to $D, X \geq 0$.

Proof sketch via auxiliary function. Let $F(X) = \frac{1}{2}\|Y - DX\|_F^2$ for fixed $D \geq 0$. Define the auxiliary function

$$G(X, \widetilde{X}) = F(\widetilde{X}) + (X - \widetilde{X})^\top \nabla F(\widetilde{X}) + \tfrac{1}{2} \sum_{ij} \frac{(D^\top D \widetilde{X})_{ij}}{\widetilde{X}_{ij}} (X_{ij} - \widetilde{X}_{ij})^2.$$

Two facts: (i) $G(X, \widetilde{X}) \geq F(X)$ for all $X, \widetilde{X} \geq 0$, with equality at $\widetilde{X} = X$ (the quadratic majorizes F by AM–GM); and (ii) $G(\cdot, \widetilde{X})$ is convex in X with closed-form minimizer

$$X^\star_{ij} = \widetilde{X}_{ij} \frac{(D^\top Y)_{ij}}{(D^\top D \widetilde{X})_{ij}}.$$

Setting $\widetilde{X} = X^{(k)}$ and $X^{(k+1)} = X^\star$ gives a majorize- minimize step: $F(X^{(k+1)}) \leq G(X^{(k+1)}, X^{(k)}) \leq G(X^{(k)}, X^{(k)}) = F(X^{(k)})$. The D-update is symmetric. Hence the objective is monotonically non-increasing. □

The updates are multiplicative — each entry is rescaled, never assigned — so nonnegativity is preserved automatically. An alternative is alternating Nonnegative Least Squares (NNLS), which solves $\min_{D \geq 0} \|Y - DX\|_F^2$ and $\min_{X \geq 0} \|Y - DX\|_F^2$ by active-set or projected-gradient methods at each iteration; NNLS converges faster than multiplicative updates but each step is more expensive.

Remark G.2 (KL-divergence variant)

When entries of Y are counts (documents, photon counts), the Kullback–Leibler divergence $D_{\mathrm{KL}}(Y \| DX) = \sum_{ij} \big(Y_{ij} \log \frac{Y_{ij}}{(DX)_{ij}} - Y_{ij} + (DX)_{ij} \big)$ is more natural than Frobenius loss. Lee–Seung's KL update rules are

$$X_{ij} \leftarrow X_{ij} \frac{\sum_k D_{ki} Y_{kj} / (DX)_{kj}}{\sum_k D_{ki}}, \quad D_{ki} \leftarrow D_{ki} \frac{\sum_j X_{ij} Y_{kj} / (DX)_{kj}}{\sum_j X_{ij}},$$

with an analogous majorization-minimization proof.

G.3 Variants

Orthogonal NMF.

Adding $XX^\top = I$ forces orthogonal codes:

$$\min_{D,X} \|Y - DX\|_F^2 \text{ s.t. } D, X \geq 0,\ XX^\top = I.$$

Theorem G.3 (NMF ≈ K-means; Ding–He–Simon 2005)

Orthogonal Semi-NMF is equivalent to relaxed K-means:

$$\min_{D,X} \|Y - DX\|_F^2 \text{ s.t. } X \geq 0,\ XX^\top = I \quad \Longleftrightarrow \quad \min \operatorname{tr}(Y^\top Y) - \operatorname{tr}(XY^\top YX^\top).$$

Maximizing the second term is K-means in disguise.

Sparse NMF.

Add an ℓ_1 penalty on X to encourage strictly local atoms. More sparseness moves atoms from blurred face-averages toward isolated facial parts (Hoyer 2004).

Semi-NMF and Convex-NMF.

$$\begin{aligned} \text{SVD:} \quad & Y_\pm = U\Lambda V^\top \approx D_\pm X_\pm \\ \text{Semi-NMF:} \quad & Y_\pm \approx D_\pm X_+ \\ \text{Convex-NMF:} \quad & Y_\pm \approx Y_\pm W_+ X_+ \end{aligned}$$

Convex-NMF expresses dictionary atoms as nonnegative combinations of the data itself — a generalization that links NMF, K-means, and the next topic.

G.4 Pros and cons

Pros. Intuitive parts-based features; addition only (no cancellation); holistic decompositions; excellent fit for images, videos, text, frequency counts; encourages clustering behavior.

Cons. Ill-posed (no unique factorization); nonconvex (no global convergence guarantee); the cone of admissible factorizations admits many local optima.

G.5 Separable NMF

Donoho & Stodden (2004) gave a geometric framework: NMF factorizations are *cones* containing the data cloud. Multiple cones can enclose the same data — hence non-uniqueness.

Definition G.4 (s-separable matrix)

$Y = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_m]$ is s-separable if there exists an index set $\mathcal{S} \subset [m]$ with $|\mathcal{S}| = s < m$ such that

$$Y = Y_{\mathcal{S}}\, Z, \quad Z \geq 0,\ \mathbf{1}^\top Z = \mathbf{1}^\top.$$

Equivalently, s columns of Y generate a convex cone containing the entire dataset.

When the convex-hull assumption holds, the factorization is *unique* and can be recovered by finding the vertex set $\mathcal{S}$. This reduces NMF to a sparse representation problem:

$$\min_X \|X\|_{\text{row},0} \text{ s.t. } YX = Y,\ X \geq 0,\ \mathbf{1}^\top X = \mathbf{1}^\top.$$

Algorithms for separable NMF:

- **Greedy pursuit (XRAY, Kumar et al. 2013).** Infer the convex hull one vertex per iteration via geometric criteria.
- **Convex relaxation $\ell_{1,\infty}$ (Elhamifar et al.).** $\min_X \|X\|_{1,\infty}$ s.t. $YX = Y$, $X \geq 0$, $\mathbf{1}^\top X = \mathbf{1}^\top$, solved by ADMM.
- **Linear programs** (Arora et al. 2012, Bittorf 2012, Gillis 2014).

G.6 Sparsity from entropy

A probability vector concentrated on a single coordinate has zero Shannon entropy:

$$H(\chi) = -\sum_i p(\chi_i) \log p(\chi_i), \qquad p_i = \frac{|x_i|^p}{\|x\|_p^p}.$$

Low entropy $\equiv$ skewed distribution $\equiv$ "sparse." This motivates the *Shannon entropy regularizer*

$$h_p(\boldsymbol{x}) = -\sum_{i=1}^{N} \frac{|x_i|^p}{\|\boldsymbol{x}\|_p^p} \log \frac{|x_i|^p}{\|\boldsymbol{x}\|_p^p}$$

as an alternative to $\|\cdot\|_1$. Minimizing h_p favors sparse solutions but is nonconvex; visually, level sets of h_1 produce a star-shaped pattern along the axes, sharper than ℓ_p for any $p > 0$.

For separable NMF, *Robust Entropy Minimization*

$$\min_X \|X\|_{h,\infty} \text{ s.t. } \|\boldsymbol{y}_j - Y\boldsymbol{x}_j\|_2 \leq \varepsilon,\ X \geq 0,\ \mathbf{1}^\top X = \mathbf{1}^\top,$$

with the row-entropy norm $\|X\|_{h,\infty} = h([\,\|X_{1,:}\|_\infty, \dots, \|X_{N,:}\|_\infty\,])$, provably identifies the vertices of Y

exactly when the noise is bounded by $\varepsilon < c/(8(s+1))$, where c depends on the fatness of the convex hull. The resulting program is nonconvex yet admits a proof of global optimality.

G.7 Applications

- **Recommendation systems.** Fill in user–item matrices: solve $\min \|Y - DX\|_F^2$ s.t. $D, X \geq 0$; the completed $\widetilde{Y} = DX$ gives predicted ratings.
- **Topic modeling.** Y = document–term matrix; dictionary atoms become topics, codes become document-topic loadings.
- **Audio source separation, hyperspectral unmixing, gene-expression analysis** — anywhere nonnegativity is a physical constraint.

H Deep Sparse Coding Networks

The final synthesis takes the entire course's theory of sparse coding and stacks it into a multilayer architecture, aiming to combine the representational power of deep neural networks with the interpretability and sample efficiency of sparse modeling.

H.1 Motivation

- One-layer sparse coding under-performs on large datasets like CIFAR.
- Deep CNNs perform strongly but offer little theoretical interpretability.
- Can we extend sparse coding to a deep architecture while preserving its mathematical framework?

Two strands of evidence motivate the affirmative answer (Rangamani et al. 2018; 2019): flatness of local minima correlates with robustness; and trained ReLU networks effectively perform sparse coding in their feature spaces.

H.2 Composite sparse coding module

The direct cascade $\boldsymbol{x} \to \boldsymbol{\alpha}^{(1)} \to \boldsymbol{\alpha}^{(2)}$ where each $\boldsymbol{\alpha}^{(\ell)} = D^{(\ell)}\boldsymbol{\alpha}^{(\ell-1)}$ grows rapidly in dimension. Sun et al. (2018, 2019) split each layer into two stages:

$$\underbrace{\boldsymbol{\alpha}^{(1)} = \arg\min_{\boldsymbol{\alpha}>0} \tfrac{1}{2}\|\boldsymbol{y} - D^{(1)}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \tfrac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2}_{\text{fat-dictionary sparse coding}}$$

$$\underbrace{\boldsymbol{\alpha}^{(2)} = \arg\min_{\boldsymbol{\alpha}>0} \tfrac{1}{2}\|\boldsymbol{\alpha}^{(1)} - D^{(2)}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \tfrac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2}_{\text{nonlinear dimension reduction}}$$

The first dictionary is over-complete ("fat"); the second is under-complete ("thin") and forces clustering of sparse codes. Together they form the composite sparse coding module. Stacking up to 13 such modules with batch normalization, global average pooling, and softmax yields the supervised Sparse Coding Network (SCN).

H.3 Inference and training

Sparse coding step.

FISTA solves the elastic-net regularized problem

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha}\geq 0} \underbrace{\tfrac{1}{2}\|\boldsymbol{x} - D\boldsymbol{\alpha}\|_2^2 + \tfrac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2}_{f(\boldsymbol{\alpha})\ \text{smooth}} + \underbrace{\lambda_1\|\boldsymbol{\alpha}\|_1 + \mathbb{1}_{\geq 0}(\boldsymbol{\alpha})}_{g(\boldsymbol{\alpha})\ \text{nonsmooth}}.$$

The smooth part f has Lipschitz gradient $\nabla f(\boldsymbol{\alpha}) = D^\top(D\boldsymbol{\alpha} - \boldsymbol{x}) + \lambda_2\boldsymbol{\alpha}$ with Lipschitz constant $L = \sigma_{\max}(D)^2 + \lambda_2$. The proximal operator of g is the *nonnegative soft-shrinkage operator*

$$\operatorname{prox}_{\lambda_1 g/L}(\boldsymbol{z}) = \big(\tau_{\lambda_1/L}(\boldsymbol{z})\big)_+, \qquad \tau_\sigma(z) = \operatorname{sgn}(z)(|z| - \sigma)_+.$$

FISTA combines the proximal-gradient step with Nesterov acceleration:

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_t &= \operatorname{prox}_{\lambda_1 g/L}\big(\boldsymbol{u}_t - L^{-1}\nabla f(\boldsymbol{u}_t)\big), \\ t_{k+1} &= \tfrac{1+\sqrt{1+4t_k^2}}{2}, \\ \boldsymbol{u}_{t+1} &= \widehat{\boldsymbol{\alpha}}_t + \tfrac{t_k - 1}{t_{k+1}}(\widehat{\boldsymbol{\alpha}}_t - \widehat{\boldsymbol{\alpha}}_{t-1}). \end{aligned}$$

Theorem H.1 (FISTA $O(1/k^2)$ convergence, Beck–Teboulle 2009)

For the iterates above,

$$F(\widehat{\boldsymbol{\alpha}}_k) - F(\boldsymbol{\alpha}^\star) \;\leq\; \frac{2L\|\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}^\star\|_2^2}{(k+1)^2}.$$

30–50 FISTA iterations suffice in practice. The matrix $D^\top D + \lambda_2 I$ is pre-computed; inference requires no online inversion.

Remark H.2 (Why elastic-net)

Pure ℓ_1 gives strict sparsity but is unstable to correlated atoms: in the limit $D^\top D \to I$ degenerate, two

highly correlated atoms can "swap" between iterations. The ℓ_2 term $\lambda_2\|\boldsymbol{\alpha}\|_2^2$ smooths the objective, ensures strong convexity, and is essential for the back-propagation gradients below to be well-defined.

Training via back-propagation.

The dictionary update rule differentiates through the fixed-point structure of the elastic-net minimizer. With $\Lambda = \text{supp}(\boldsymbol{\alpha})$,

$$\begin{aligned}
\frac{\partial L}{\partial D_{ij}} &= \frac{\partial L}{\partial \boldsymbol{\alpha}} \cdot (D^\top D + \lambda_2 I)_\Lambda^{-1}\Big(\frac{\partial D_\Lambda^\top \boldsymbol{\alpha}}{\partial D_{ij}} - \frac{\partial D_\Lambda^\top D_\Lambda}{\partial D_{ij}} \boldsymbol{y}_\Lambda\Big), \\
\frac{\partial L}{\partial x_i} &= \frac{\partial L}{\partial \boldsymbol{\alpha}} \cdot (D^\top D + \lambda_2 I)_\Lambda^{-1} \frac{\partial D_\Lambda^\top \boldsymbol{x}}{\partial x_i}, \\
\frac{\partial L}{\partial \lambda_{1j}} &= -\frac{\partial L}{\partial \boldsymbol{\alpha}} \cdot (D^\top D + \lambda_2 I)_\Lambda^{-1} \operatorname{sign}(\boldsymbol{\alpha}_\Lambda)_j, \quad \lambda_{1j} \neq 0.
\end{aligned}$$

These gradients support end-to-end supervised training. Regularization parameters λ_1, λ_2 are themselves task-adaptive.

H.4 Performance

On standard image classification benchmarks the supervised SCN closes much of the gap to state-of-the-art CNNs while using a smaller model, fewer parameters, and substantially less training data:

Dataset	Prior sparse coding	Supervised SCN	State-of-the-art CNN
CIFAR-10	81.40%	94.19%	96.42%
CIFAR-100	60.80%	80.07%	82.69%
STL-10	67.90%	83.11%	76.29%
MNIST (error)	0.54%	0.36%	0.21%
SVHN (error)	—	2.16%	1.77%

Notably SCN *outperforms* the CNN baseline on STL-10 where only 5,000 labeled training samples are available — sparse coding's structural prior shines in low-data regimes.

H.5 Why SCN works: feature-map clustering

Visualizing intermediate feature maps reveals a qualitative difference between SCN and CNN. SCN's nonlinear dimension-reduction step clusters feature maps that respond to similar high-level patterns; CNN feature maps are diverse but unclustered. The composite module's second layer acts as a soft codebook: each $\boldsymbol{\alpha}^{(2)}$ effectively selects a small number of higher-level prototypes from a $\boldsymbol{\alpha}^{(1)}$ that scattered the signal

across many edge-like atoms. This is reminiscent of semi-NMF and Convex-NMF (§G): the clustering effect emerges from the nonnegativity constraint combined with the elastic-net regularizer.

H.6 Pros and cons

Advantages. Framework grounded in sparse-coding theory; less architecture tuning than CNN; competitive accuracy with smaller models and less data; clear interpretation of why deep networks succeed.

Disadvantages. Substantially longer training time (back-propagating through FISTA fixed-points is expensive); slightly slower inference; large-scale performance remains open; many downstream tasks (detection, segmentation, generation) have not been tried.

Course Synthesis

The thread connecting all eight weeks:

1. **Sparsity is a structural prior** (Lectures 1–2). Theorem 2.1 characterizes uniform recovery in terms of $\text{spark}(A) > 2s$; Theorem 2.7 asserts NP-hardness of P_0 in general.

2. **Greedy algorithms** (Lectures 3–4): MP, OMP, IHT, HTP, CoSaMP, SP. Their theoretical guarantees come in two flavors – coherence-based (small μ) or RIP-based.

3. **Spark, NSP, and RIP** (Lectures 5–7): the conditions on A that guarantee sparse recovery, with the tower of implications coherence $\Rightarrow$ NSP $\Rightarrow$ BP recovery, and RIP$\Rightarrow$NSP$\Rightarrow$BP/IHT/HTP/CoSaMP recovery. Theorem 5.2 (NSP characterizes BP), Theorem 5.8 (dual certificates), Theorem 7.3 (RIP$\delta_{2s} < \sqrt{2}-1 \Rightarrow$ BP).

4. **Random matrices** (Lecture 8): subgaussian and Gaussian matrices have RIP at the optimal scaling $m = O(s\log(N/s))$ (Theorem 8.2), matching the information-theoretic lower bound. Connection to JL embeddings (Theorem 8.6).

5. **Sparse Representation Classification** (Lecture 9): leverages CS theory to perform discriminative tasks; identity-augmented dictionaries handle occlusions natively; SCI provides confidence calibration.

6. **ℓ_1-minimization at scale** (Lecture 10): ISTA/FISTA, ADMM, LARS-homotopy, primal–dual splitting (Chambolle–Pock), and IRLS form the modern toolkit.

7. **Multiple measurements and joint sparsity** (Lecture 12). The MMV model $Y = AX$ with row-sparse X admits sharper uniqueness ($s < (\text{spark}(A) - 1 + \text{rank}(Y))/2$) and reduced average-case sample complexity via SOMP and $\ell_{1,2}$-minimization; Collaborative HiLasso unifies joint group sparsity with within-row sparsity.

8. **Low-rank structure: Matrix Completion and Robust PCA** (Lecture 13). Replace "sparse vector" with "low-rank matrix," ℓ_0 with rank, ℓ_1 with nuclear norm. Candès–Recht 2009 guarantees recovery from $\Theta(rN\log^2 N)$ entries; Candès–Li–Ma–Wright 2011 separates a low-rank X from sparse gross noise E. SVT and ALM are the workhorse solvers.

9. **Dictionary learning** (Lectures 14, 22–24). The dictionary itself becomes a learnable object: K-means (one-hot codes), MOD (pseudo-inverse update), and K-SVD (one-atom-at-a-time SVD update) yield overcomplete dictionaries trained on data. Online dictionary learning, LLC, and SSC adapt the framework to streaming data, classification, and subspace clustering respectively.

10. **Nonnegative Matrix Factorization** (Lecture 15). Adding $D, X \geq 0$ forces parts-based representations

(Lee–Seung 1999). Multiplicative updates guarantee local convergence; orthogonal Semi-NMF reduces to relaxed K-means; Separable NMF + entropy minimization yield unique factorizations under a convex-hull assumption.

11. **Deep sparse coding** (Lecture 16). Stacking composite sparse-coding modules (fat-dictionary ℓ_1 followed by thin nonlinear dimension reduction) yields supervised Sparse Coding Networks (SCN) that match deep CNN accuracy on CIFAR/STL/MNIST while remaining theoretically grounded; training proceeds by back-propagation through FISTA fixed-points.

Reference.

S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser/Springer, 2013. Additional sources: Candès & Recht (2009) for matrix completion; Candès, Li, Ma, Wright (2011) for Robust PCA; Aharon, Elad, Bruckstein (2006) for K-SVD; Lee & Seung (1999, 2001) for NMF; Sun et al. (2018, 2019) for Sparse Coding Networks.