# Sparse Representation and Redundant Dictionaries: From Orthonormal Bases to Overcomplete Systems

Advanced Signal Processing Lecture

July 7, 2025

## Contents

# 1 Introduction and Motivation

The theory of sparse representation has emerged as a fundamental paradigm in modern signal processing, machine learning, and data analysis. This lecture establishes the theoretical foundations for understanding when and why we must abandon the comfort of orthonormal bases in favor of overcomplete dictionaries to achieve sparse representations.

## 1.1 Historical Context

The development of sparse representation theory represents a significant departure from classical linear algebra approaches. While traditional methods rely on orthonormal bases that guarantee unique, easily computable representations, real-world signals often exhibit structure that cannot be efficiently captured by any single orthonormal basis. This observation has led to the development of overcomplete dictionary methods, which sacrifice uniqueness and computational simplicity for the ability to provide sparse representations of complex signals.

## 1.2 Lecture Overview

This lecture covers the fundamental transition from orthonormal bases to overcomplete dictionaries, examining:

- The limitations of orthonormal bases for sparse representation
- The mathematical foundations of linear independence and span
- The construction of overcomplete dictionaries
- The resulting computational challenges and their solutions

# 2 Fundamental Concepts in Linear Algebra

## 2.1 Vector Spaces and Linear Combinations

**Definition 2.1** (Span of Vectors)**.** Given a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^m$, the *span* of these vectors is defined as:

$$\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} = \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{v}_i : \lambda_i \in \mathbb{R} \right\} \tag{1}$$

The span represents the set of all possible linear combinations of the given vectors, forming a vector subspace of $\mathbb{R}^m$. This concept is fundamental to understanding how different sets of vectors can generate different subspaces.

## 2.2 Linear Independence and Basis

**Definition 2.2** (Linear Independence)**.** A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is *linearly independent* if and only if:

$$\sum_{i=1}^{n} \lambda_i \mathbf{v}_i = \mathbf{0} \quad \Rightarrow \quad \lambda_i = 0 \text{ for all } i = 1, 2, \ldots, n \tag{2}$$

This definition captures the fundamental property that no vector in the set can be expressed as a linear combination of the others. The importance of linear independence becomes clear when we consider the uniqueness of representations.

**Theorem 2.3** (Uniqueness of Representation)**.** Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^m$ be a linearly independent set of vectors. If $\mathbf{s} \in \text{span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, then there exists a unique representation:

$$\mathbf{s} = \sum_{i=1}^{n} x_i \mathbf{e}_i \tag{3}$$

where the coefficients $x_i \in \mathbb{R}$ are uniquely determined.

*Proof.* Suppose $\mathbf{s}$ admits two different representations:

$$\mathbf{s} = \sum_{i=1}^{n} x_i \mathbf{e}_i \tag{4}$$

$$\mathbf{s} = \sum_{i=1}^{n} y_i \mathbf{e}_i \tag{5}$$

Subtracting these equations:

$$\mathbf{0} = \sum_{i=1}^{n} (x_i - y_i) \mathbf{e}_i \tag{6}$$

By linear independence, $(x_i - y_i) = 0$ for all $i$, implying $x_i = y_i$ for all $i$. Therefore, the representation is unique. □

## 2.3 Orthonormal Bases and Their Properties

**Definition 2.4** (Orthonormal Basis)**.** A set of vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^n$ forms an *orthonormal basis* if:

1. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \delta_{ij}$ (orthonormality condition)
2. $\text{span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} = \mathbb{R}^n$ (spanning condition)

where $\delta_{ij}$ is the Kronecker delta.

The power of orthonormal bases lies in their computational convenience. For any signal $\mathbf{s} \in \mathbb{R}^n$ and orthonormal basis $\mathbf{D} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, the coefficient computation is straightforward:

$$\mathbf{x} = \mathbf{D}^T \mathbf{s} \tag{7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $x_i = \langle \mathbf{e}_i, \mathbf{s} \rangle$.

# 3 Limitations of Orthonormal Bases

## 3.1 The Sparsity Problem

While orthonormal bases provide computational convenience and guarantee unique representations, they suffer from a fundamental limitation: *no single orthonormal basis can provide sparse representations for all signals of interest.*

**Example 3.1** (DCT Basis Limitation)**.** Consider a signal $\mathbf{s}_0 \in \mathbb{R}^n$ that admits a sparse representation with respect to the Discrete Cosine Transform (DCT) basis $\mathbf{D}_{DCT}$:

$$\mathbf{s}_0 = \mathbf{D}_{DCT}\mathbf{x}_0 \tag{8}$$

where $\mathbf{x}_0$ is sparse (most entries are zero).

Now consider the modified signal:

$$\mathbf{s} = \mathbf{s}_0 + \lambda\mathbf{e}_j \tag{9}$$

where $\mathbf{e}_j$ is the $j$-th canonical basis vector and $\lambda \in \mathbb{R}$ is a scaling factor.

The DCT representation of $\mathbf{s}$ becomes:

$$\mathbf{x} = \mathbf{D}_{DCT}^T\mathbf{s} = \mathbf{D}_{DCT}^T\mathbf{s}_0 + \lambda\mathbf{D}_{DCT}^T\mathbf{e}_j = \mathbf{x}_0 + \lambda\mathbf{D}_{DCT}^T\mathbf{e}_j \tag{10}$$

Since $\mathbf{D}_{DCT}^T\mathbf{e}_j$ is typically dense (all entries are non-zero), the addition of a single spike destroys the sparsity of the representation.

## 3.2 Mathematical Analysis of the Limitation

The fundamental issue can be understood through the lens of mutual coherence between different bases. The DCT basis and the canonical basis are *maximally incoherent*, meaning that any vector sparse in one basis becomes dense in the other.

**Definition 3.2** (Mutual Coherence)**.** Given two orthonormal bases $\mathbf{D}_1$ and $\mathbf{D}_2$, their mutual coherence is defined as:

$$\mu(\mathbf{D}_1, \mathbf{D}_2) = \max_{i,j} |\langle \mathbf{d}_{1,i}, \mathbf{d}_{2,j} \rangle| \tag{11}$$

where $\mathbf{d}_{1,i}$ and $\mathbf{d}_{2,j}$ are columns of $\mathbf{D}_1$ and $\mathbf{D}_2$ respectively.

For the DCT and canonical bases, $\mu(\mathbf{D}_{DCT}, \mathbf{I}) = 1/\sqrt{n}$, which is the maximum possible coherence for orthonormal bases in $\mathbb{R}^n$.

## 3.3 Experimental Demonstration

The experimental verification of this limitation involves:

1. Generate a sparse signal $\mathbf{s}_0$ with respect to DCT basis
2. Add a single spike: $\mathbf{s} = \mathbf{s}_0 + \lambda\mathbf{e}_j$

3. Compute DCT coefficients of both signals

4. Observe the loss of sparsity in the modified signal

The results consistently show that the addition of a single spike causes all DCT coefficients to become significant, effectively destroying the sparse structure that denoising algorithms rely upon.

# 4 Overcomplete Dictionaries: The Solution

## 4.1 Motivation for Redundancy

The solution to the sparsity limitation lies in abandoning the constraint of orthonormality and embracing redundancy. Instead of using a single $n \times n$ orthonormal basis, we construct an $n \times m$ dictionary matrix $\mathbf{D}$ where $m > n$.

**Definition 4.1** (Overcomplete Dictionary)**.** An *overcomplete dictionary* is a matrix $\mathbf{D} \in \mathbb{R}^{n \times m}$ with $m > n$ such that:

$$\text{span}\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_m\} = \mathbb{R}^n \tag{12}$$

where $\mathbf{d}_i$ are the columns of $\mathbf{D}$.

## 4.2 Construction of Overcomplete Dictionaries

For the DCT-spike example, we construct the overcomplete dictionary by concatenating the DCT basis with the canonical basis:

$$\mathbf{D} = [\mathbf{D}_{DCT} \mid \mathbf{I}] \in \mathbb{R}^{n \times 2n} \tag{13}$$

This construction ensures that:

- Signals sparse in DCT domain remain sparse
- Signals sparse in canonical domain remain sparse
- Mixed signals (DCT-sparse + spikes) admit sparse representations

**Example 4.2** (Sparse Representation with Overcomplete Dictionary)**.** Consider the signal $\mathbf{s} = \mathbf{s}_0 + \lambda \mathbf{e}_j$ where $\mathbf{s}_0 = \mathbf{D}_{DCT}\mathbf{x}_0$ with sparse $\mathbf{x}_0$.

The representation with respect to the overcomplete dictionary is:

$$\mathbf{s} = \mathbf{D} \begin{pmatrix} \mathbf{x}_0 \\ \lambda \mathbf{e}_j \end{pmatrix} \tag{14}$$

The coefficient vector $\begin{pmatrix} \mathbf{x}_0 \\ \lambda \mathbf{e}_j \end{pmatrix} \in \mathbb{R}^{2n}$ is sparse, containing only the non-zero entries of $\mathbf{x}_0$ plus the single entry $\lambda$ at position $j$ in the second block.

## 4.3 Theoretical Properties of Overcomplete Systems

**Theorem 4.3** (Rouché-Capelli Theorem)**.** Consider the linear system $\mathbf{Dx} = \mathbf{s}$ where $\mathbf{D} \in \mathbb{R}^{n \times m}$ and $\mathbf{s} \in \mathbb{R}^n$. The system admits a solution if and only if:

$$\text{rank}(\mathbf{D}) = \text{rank}([\mathbf{D} \mid \mathbf{s}]) \tag{15}$$

When $m > n$ and $\text{rank}(\mathbf{D}) = n$, the system has infinitely many solutions forming an affine subspace of dimension $m - n$.

**Corollary 4.4** (Solution Space Dimension)**.** If $\mathbf{D} \in \mathbb{R}^{n \times m}$ with $m > n$ and $\text{rank}(\mathbf{D}) = n$, then for any $\mathbf{s} \in \mathbb{R}^n$, the solution set of $\mathbf{Dx} = \mathbf{s}$ forms an affine subspace of dimension $m - n$.

# 5 Regularization and Sparse Recovery

## 5.1 The Ill-Posed Nature of Overcomplete Systems

The abundance of solutions in overcomplete systems necessitates additional criteria for solution selection. This is where regularization theory becomes essential.

**Definition 5.1** (Regularization)**.** Given an ill-posed problem $\mathbf{Dx} = \mathbf{s}$ with multiple solutions, *regularization* involves solving:

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} J(\mathbf{x}) \quad \text{subject to} \quad \mathbf{Dx} = \mathbf{s} \tag{16}$$

where $J : \mathbb{R}^m \to \mathbb{R}_+$ is a regularization functional encoding our prior knowledge about the desired solution.

## 5.2 $\ell_2$ Regularization: Ridge Regression

The most mathematically tractable regularization is the $\ell_2$ norm:

$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} x_i^2 \tag{17}$$

This leads to the constrained optimization problem:

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{Dx} = \mathbf{s} \tag{18}$$

Alternatively, we can formulate the unconstrained version:

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} \frac{1}{2}\|\mathbf{Dx} - \mathbf{s}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 \tag{19}$$

## 5.3 Analytical Solution via Matrix Calculus

The unconstrained $\ell_2$ regularization problem admits a closed-form solution. To derive this, we use matrix calculus.

**Theorem 5.2** (Ridge Regression Solution)**.** The solution to the ridge regression problem:

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} \frac{1}{2}\|\mathbf{Dx} - \mathbf{s}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 \tag{20}$$

is given by:

$$\hat{\mathbf{x}} = (\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{s} \tag{21}$$

where $\lambda > 0$ ensures the matrix $(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})$ is invertible.

*Proof.* Define the objective function:

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{Dx} - \mathbf{s}\|_2^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 \tag{22}$$

Expanding the squared norms:

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{D}\mathbf{x} - \mathbf{s})^T(\mathbf{D}\mathbf{x} - \mathbf{s}) + \frac{\lambda}{2}\mathbf{x}^T\mathbf{x} \tag{23}$$

$$= \frac{1}{2}\mathbf{x}^T\mathbf{D}^T\mathbf{D}\mathbf{x} - \mathbf{s}^T\mathbf{D}\mathbf{x} + \frac{1}{2}\mathbf{s}^T\mathbf{s} + \frac{\lambda}{2}\mathbf{x}^T\mathbf{x} \tag{24}$$

Taking the gradient with respect to $\mathbf{x}$:

$$\nabla f(\mathbf{x}) = \mathbf{D}^T\mathbf{D}\mathbf{x} - \mathbf{D}^T\mathbf{s} + \lambda\mathbf{x} \tag{25}$$

Setting $\nabla f(\mathbf{x}) = \mathbf{0}$:

$$(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})\mathbf{x} = \mathbf{D}^T\mathbf{s} \tag{26}$$

Since $\lambda > 0$, the matrix $(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})$ is positive definite and therefore invertible, yielding the stated solution. □

## 5.4 Limitations of $\ell_2$ Regularization

While $\ell_2$ regularization provides a computationally efficient solution, it does not promote sparsity. The solution $\hat{\mathbf{x}}$ typically has all non-zero entries, which contradicts our goal of sparse representation.

**Remark 5.3** (Sparsity vs. $\ell_2$ Regularization)**.** The $\ell_2$ norm penalizes large coefficients but does not drive small coefficients to zero. For sparse recovery, we need regularization functionals that promote sparsity, such as the $\ell_1$ norm or $\ell_0$ pseudo-norm.

# 6 Towards Sparsity: $\ell_0$ and $\ell_1$ Regularization

## 6.1 The $\ell_0$ "Norm" and True Sparsity

The most natural regularization for sparse recovery is the $\ell_0$ "norm" (technically a pseudo-norm):

$$\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}| \tag{27}$$

This counts the number of non-zero entries in $\mathbf{x}$. The corresponding optimization problem:

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{D}\mathbf{x} = \mathbf{s} \tag{28}$$

directly seeks the sparsest representation.

## 6.2 Computational Challenges

The $\ell_0$ minimization problem is NP-hard in general, making it computationally intractable for large-scale problems. This has led to the development of convex relaxations and approximation algorithms.

## 6.3 Future Directions

The next lectures will cover:

- Sparse coding algorithms for $\ell_1$ regularization
- Dictionary learning methods
- Compressed sensing theory
- Applications to signal processing and machine learning

# 7 Summary and Conclusions

## 7.1 Key Insights

This lecture has established several fundamental principles:

1. **Orthonormal Basis Limitations**: No single orthonormal basis can provide sparse representations for all signals of interest.

2. **Overcomplete Dictionary Benefits**: Redundant dictionaries sacrifice uniqueness and computational simplicity but enable sparse representations for broader signal classes.

3. **Regularization Necessity**: The ill-posed nature of overcomplete systems requires regularization for meaningful solutions.

4. **Trade-offs**: We exchange the computational convenience of orthonormal bases for the representational power of overcomplete dictionaries.

## 7.2 Mathematical Framework Summary

The transition from orthonormal to overcomplete systems can be summarized as follows:

| **Property** | **Orthonormal Basis** | **Overcomplete Dictionary** |
|---|---|---|
| Matrix Size | $n \times n$ | $n \times m$ (where $m > n$) |
| Linear Independence | Yes | No |
| Uniqueness | Yes | No |
| Computational Complexity | Low | High |
| Sparsity Guarantee | No | Possible |
| Perfect Reconstruction | Yes | Yes |

## 7.3 Practical Implications

The theoretical framework developed here has immediate practical applications:

- **Signal Denoising**: Overcomplete dictionaries can preserve both smooth regions and sharp features.

- **Image Processing**: Different image structures (edges, textures, smooth regions) can be sparsely represented by different dictionary elements.

- **Machine Learning**: Overcomplete representations can lead to better feature extraction and classification performance.

## 7.4 Looking Forward

The mathematical foundations established in this lecture form the basis for understanding:

- Advanced sparse coding algorithms
- Dictionary learning methods
- Compressed sensing theory
- Applications in signal processing, image processing, and machine learning

The journey from orthonormal bases to overcomplete dictionaries represents a paradigm shift in signal representation theory, trading computational simplicity for representational power and opening new avenues for signal processing and analysis.