

# L1 Optimization and Sparse Coding: From Sparsity-Promoting Norms to Convex Optimization

Lecture Notes

July 14, 2025

## Contents

<b>1</b>	<b>Introduction to Sparsity-Promoting Norms</b>	<b>2</b>
1.1	Motivation: From $\ell_0$ to $\ell_1$ . . . . .	2
<b>2</b>	<b>Mathematical Framework: <math>\ell_p</math> Norms</b>	<b>3</b>
2.1	Definition and Properties . . . . .	3
2.2	The Case $p < 1$ : Quasi-Norms . . . . .	3
<b>3</b>	<b>Geometric Interpretation: Unit Balls and Sparsity</b>	<b>4</b>
3.1	Visualization of $\ell_p$ Unit Balls . . . . .	4
3.2	Sparsity Promotion Through Geometric Analysis . . . . .	4
<b>4</b>	<b>Convex Optimization Theory</b>	<b>5</b>
4.1	Convex Sets and Functions . . . . .	5
4.2	Convexity of $\ell_p$ Norms . . . . .	5
4.3	Fundamental Optimization Result . . . . .	5
<b>5</b>	<b>The <math>\ell_1</math> Optimization Problem</b>	<b>6</b>
5.1	Problem Formulation . . . . .	6
5.2	Connection to LASSO . . . . .	6
5.3	Problem Components Analysis . . . . .	6
5.3.1	Data Fidelity Term . . . . .	6
5.3.2	Regularization Term . . . . .	6
<b>6</b>	<b>Optimization Theory for Non-Differentiable Functions</b>	<b>7</b>
6.1	The Descent Lemma . . . . .	7
6.2	Majorization-Minimization Approach . . . . .	7
6.3	Gradient Descent Derivation . . . . .	7

<b>7</b>	<b>Proximal Gradient Methods</b>	<b>8</b>
7.1	Proximal Operator . . . . .	8
7.2	Proximal Gradient Algorithm . . . . .	8
7.3	Proximal Operator of $\ell_1$ Norm . . . . .	8
<b>8</b>	<b>Applications and Extensions</b>	<b>9</b>
8.1	Signal Processing Applications . . . . .	9
8.2	Computational Considerations . . . . .	9
8.3	Extensions to Other Norms . . . . .	9
<b>9</b>	<b>Conclusion</b>	<b>9</b>
<b>10</b>	<b>Mathematical Appendix</b>	<b>11</b>
10.1	Subdifferential Calculus . . . . .	11
10.2	Optimality Conditions . . . . .	11
<b>11</b>	<b>Glossary of Symbols</b>	<b>11</b>

# 1 Introduction to Sparsity-Promoting Norms

The quest for sparse representations in signal processing and machine learning has led to the development of various sparsity-promoting norms. This lecture transitions from the intuitive but computationally intractable  $\ell_0$  norm to more practical alternatives, particularly the  $\ell_1$  norm, which offers a favorable balance between sparsity promotion and computational tractability.

## 1.1 Motivation: From $\ell_0$ to $\ell_1$

The  $\ell_0$  norm, defined as the number of non-zero components in a vector, provides the most intuitive measure of sparsity. However, optimization problems involving the  $\ell_0$  norm are NP-hard due to their combinatorial nature. This computational intractability necessitates the exploration of alternative sparsity-promoting norms that maintain favorable optimization properties.

**Definition 1.1** (Sparsity Measure). For a vector  $\mathbf{x} \in \mathbb{R}^n$ , the  $\ell_0$  norm is defined as:

$$\|\mathbf{x}\|_0 = \sum_{i=1}^n \mathbf{1}_{x_i \neq 0} \quad (1)$$

where  $\mathbf{1}_{x_i \neq 0}$  is the indicator function that equals 1 if  $x_i \neq 0$  and 0 otherwise.

The challenge lies in finding norms that promote sparsity while yielding tractable optimization problems. This leads us to consider the family of  $\ell_p$  norms for various values of  $p$ .

## 2 Mathematical Framework: $\ell_p$ Norms

### 2.1 Definition and Properties

The  $\ell_p$  norm extends the familiar concept of the Euclidean norm to a broader family of norms parametrized by  $p \geq 1$ .

**Definition 2.1** ( $\ell_p$  Norm). For a vector  $\mathbf{x} \in \mathbb{R}^n$  and  $p \geq 1$ , the  $\ell_p$  norm is defined as:

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \quad (2)$$

**Example 2.2** (Common  $\ell_p$  Norms).

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \quad (\text{Manhattan norm}) \quad (3)$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \quad (\text{Euclidean norm}) \quad (4)$$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| \quad (\text{Maximum norm}) \quad (5)$$

### 2.2 The Case $p < 1$ : Quasi-Norms

For  $0 < p < 1$ , the expression  $(\sum_{i=1}^n |x_i|^p)^{1/p}$  does not satisfy the triangle inequality and thus is not a norm in the mathematical sense. These are referred to as quasi-norms.

**Theorem 2.3** (Triangle Inequality Failure for  $p < 1$ ). For  $0 < p < 1$ , the triangle inequality  $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$  does not hold in general.

*Proof Sketch.* Consider the counterexample with  $\mathbf{x} = (1, 0)^T$  and  $\mathbf{y} = (0, 1)^T$  in  $\mathbb{R}^2$ . For  $p < 1$ :

$$\|\mathbf{x} + \mathbf{y}\|_p = \|(1, 1)^T\|_p = (1^p + 1^p)^{1/p} = 2^{1/p} \quad (6)$$

$$\|\mathbf{x}\|_p + \|\mathbf{y}\|_p = 1^{1/p} + 1^{1/p} = 2 \quad (7)$$

Since  $1/p > 1$  for  $p < 1$ , we have  $2^{1/p} > 2$ , violating the triangle inequality.  $\square$

### 3 Geometric Interpretation: Unit Balls and Sparsity

#### 3.1 Visualization of $\ell_p$ Unit Balls

The geometric properties of  $\ell_p$  norms can be understood through their unit balls, defined as:

$$B_p = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq 1\} \quad (8)$$

Figure 1: Unit balls for various  $\ell_p$  norms in  $\mathbb{R}^2$ . The  $\ell_1$  ball forms a diamond shape, while  $\ell_2$  forms a circle, and  $\ell_\infty$  forms a square.

#### 3.2 Sparsity Promotion Through Geometric Analysis

The sparsity-promoting properties of different norms can be understood through a geometric optimization perspective. Consider the constrained optimization problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_p \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \quad (9)$$

where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with  $m < n$  (underdetermined system).

**Theorem 3.1** (Geometric Sparsity Argument). For underdetermined linear systems, the  $\ell_1$  norm promotes sparse solutions more effectively than the  $\ell_2$  norm due to the angular structure of the  $\ell_1$  unit ball.

*Geometric Interpretation.* The solution is found by inflating the  $\ell_p$  unit ball until it touches the solution set (an affine subspace). The  $\ell_1$  ball's diamond shape with sharp corners along the coordinate axes makes it more likely to intersect the solution set at a sparse point (where many coordinates are zero) compared to the smooth  $\ell_2$  ball.  $\square$

## 4 Convex Optimization Theory

### 4.1 Convex Sets and Functions

**Definition 4.1** (Convex Set). A set  $S \subseteq \mathbb{R}^n$  is convex if and only if for any two points  $\mathbf{x}, \mathbf{y} \in S$  and any  $\alpha \in [0, 1]$ :

$$\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in S \quad (10)$$

**Definition 4.2** (Convex Function). A function  $f : S \rightarrow \mathbb{R}$  defined on a convex set  $S \subseteq \mathbb{R}^n$  is convex if for any  $\mathbf{x}, \mathbf{y} \in S$  and  $\alpha \in [0, 1]$ :

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha)f(\mathbf{y}) \quad (11)$$

### 4.2 Convexity of $\ell_p$ Norms

**Theorem 4.3** (Convexity of  $\ell_p$  Norms). For  $p \geq 1$ , the  $\ell_p$  norm is a convex function on  $\mathbb{R}^n$ .

*Proof.* For any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and  $\alpha \in [0, 1]$ , we need to show:

$$\|\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}\|_p \leq \alpha\|\mathbf{x}\|_p + (1 - \alpha)\|\mathbf{y}\|_p \quad (12)$$

Using the triangle inequality for norms:

$$\|\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}\|_p \leq \|\alpha\mathbf{x}\|_p + \|(1 - \alpha)\mathbf{y}\|_p \quad (13)$$

$$= \alpha\|\mathbf{x}\|_p + (1 - \alpha)\|\mathbf{y}\|_p \quad (14)$$

where the last equality uses the homogeneity property of norms.  $\square$

### 4.3 Fundamental Optimization Result

**Theorem 4.4** (Global Optimality in Convex Optimization). For a convex optimization problem, any local minimum is also a global minimum.

This result is crucial for sparse coding applications, as it guarantees that any convergent optimization algorithm will find the globally optimal solution.

## 5 The $\ell_1$ Optimization Problem

### 5.1 Problem Formulation

The  $\ell_1$  optimization problem for sparse coding can be formulated in two equivalent ways:

**Definition 5.1** (Constrained  $\ell_1$  Problem (P1)).

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{Ax} - \mathbf{b}\|_2 \leq \epsilon \quad (15)$$

**Definition 5.2** (Regularized  $\ell_1$  Problem (P2)).

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 \quad (16)$$

### 5.2 Connection to LASSO

The regularized formulation (P2) is known in statistics as the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Robert Tibshirani.

**Remark 5.3** (LASSO vs. Sparse Coding). While LASSO and sparse coding share the same mathematical formulation, they operate in different contexts:

- **LASSO**: Overdetermined systems ( $m > n$ ) for variable selection
- **Sparse Coding**: Underdetermined systems ( $m < n$ ) for signal representation

### 5.3 Problem Components Analysis

#### 5.3.1 Data Fidelity Term

The term  $\frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$  serves as the data fidelity term, ensuring that the solution  $\mathbf{x}$  produces a reconstruction  $\mathbf{Ax}$  that is close to the observed signal  $\mathbf{b}$ .

**Proposition 5.4** (Properties of Data Fidelity Term). The data fidelity term  $g(\mathbf{x}) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$  is:

1. Convex (as a composition of convex functions)
2. Differentiable with gradient  $\nabla g(\mathbf{x}) = \mathbf{A}^T(\mathbf{Ax} - \mathbf{b})$
3. Strongly convex if  $\mathbf{A}$  has full column rank

#### 5.3.2 Regularization Term

The term  $\lambda \|\mathbf{x}\|_1$  acts as a regularization term, promoting sparsity in the solution.

**Proposition 5.5** (Properties of  $\ell_1$  Regularization). The regularization term  $h(\mathbf{x}) = \|\mathbf{x}\|_1$  is:

1. Convex
2. Non-differentiable at  $x_i = 0$  for any component  $i$
3. Promotes sparsity through its geometric properties

## 6 Optimization Theory for Non-Differentiable Functions

### 6.1 The Descent Lemma

For smooth convex functions, we can construct quadratic majorizers that facilitate optimization.

**Lemma 6.1** (Descent Lemma). Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex, differentiable function with Lipschitz continuous gradient. Then for any  $\mathbf{x}_k \in \mathbb{R}^n$ , there exists  $L > 0$  such that:

$$f(\mathbf{x}) \leq Q_L(\mathbf{x}; \mathbf{x}_k) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2 \quad (17)$$

for all  $\mathbf{x} \in \mathbb{R}^n$ .

### 6.2 Majorization-Minimization Approach

**Definition 6.2** (Majorization-Minimization Algorithm). Given a convex function  $f$ , the majorization-minimization approach generates a sequence  $\{\mathbf{x}_k\}$  by:

$$\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} Q_L(\mathbf{x}; \mathbf{x}_k) \quad (18)$$

$$= \operatorname{argmin}_{\mathbf{x}} \left[ f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2 \right] \quad (19)$$

### 6.3 Gradient Descent Derivation

Minimizing the majorizer  $Q_L(\mathbf{x}; \mathbf{x}_k)$  with respect to  $\mathbf{x}$ :

$$\nabla_{\mathbf{x}} Q_L(\mathbf{x}; \mathbf{x}_k) = \nabla f(\mathbf{x}_k) + L(\mathbf{x} - \mathbf{x}_k) = 0 \quad (20)$$

$$\Rightarrow \mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k) \quad (21)$$

This recovers the standard gradient descent update with step size  $\gamma = 1/L$ .

**Theorem 6.3** (Gradient Descent Convergence). For a convex, differentiable function  $f$  with Lipschitz continuous gradient, the gradient descent algorithm converges to the global minimum.



## 7 Proximal Gradient Methods

### 7.1 Proximal Operator

For the composite optimization problem  $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$  where  $f$  is smooth and  $g$  is non-smooth, we introduce the proximal operator.

**Definition 7.1** (Proximal Operator). The proximal operator of a function  $g$  with parameter  $\lambda > 0$  is defined as:

$$\text{prox}_{\lambda g}(\mathbf{v}) = \underset{\mathbf{x}}{\text{argmin}} \left\{ \frac{1}{2\lambda} \|\mathbf{x} - \mathbf{v}\|_2^2 + g(\mathbf{x}) \right\} \quad (22)$$

### 7.2 Proximal Gradient Algorithm

---

For the problem  $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$ :

1. Initialize  $\mathbf{x}_0$
2. For  $k = 0, 1, 2, \dots$ :

$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k) \quad (23)$$

$$\mathbf{x}_{k+1} = \text{prox}_{\frac{1}{L}g}(\mathbf{y}_k) \quad (24)$$


---

### 7.3 Proximal Operator of $\ell_1$ Norm

**Theorem 7.2** (Soft Thresholding). The proximal operator of the  $\ell_1$  norm is given by the soft thresholding operator:

$$[\text{prox}_{\lambda \|\cdot\|_1}(\mathbf{v})]_i = \text{sign}(v_i) \max\{|v_i| - \lambda, 0\} \quad (25)$$

where  $\text{sign}(v_i)$  is the sign function.

*Proof.* The proximal operator problem becomes:

$$\min_{\mathbf{x}} \left\{ \frac{1}{2\lambda} \|\mathbf{x} - \mathbf{v}\|_2^2 + \|\mathbf{x}\|_1 \right\} \quad (26)$$

This separates into  $n$  independent scalar problems:

$$\min_{x_i} \left\{ \frac{1}{2\lambda} (x_i - v_i)^2 + |x_i| \right\} \quad (27)$$

The solution is the soft thresholding operator due to the subdifferential analysis of the absolute value function.  $\square$

## 8 Applications and Extensions

### 8.1 Signal Processing Applications

The  $\ell_1$  optimization framework finds extensive applications in:

- **Compressed Sensing:** Recovering sparse signals from undersampled measurements
- **Image Denoising:** Removing noise while preserving important features
- **Feature Selection:** Identifying relevant variables in high-dimensional data
- **Dictionary Learning:** Learning overcomplete bases for signal representation

### 8.2 Computational Considerations

**Remark 8.1** (Algorithmic Efficiency). The proximal gradient method for  $\ell_1$  optimization has several computational advantages:

1. Uses only first-order information (gradients)
2. Scales well to high dimensions
3. Produces sparse solutions automatically through soft thresholding
4. Guaranteed global convergence for convex problems

### 8.3 Extensions to Other Norms

The framework extends naturally to other sparsity-promoting norms:

- **Group LASSO:**  $\|\mathbf{x}\|_{\text{group}} = \sum_g \|\mathbf{x}_g\|_2$
- **Elastic Net:**  $\alpha\|\mathbf{x}\|_1 + (1 - \alpha)\|\mathbf{x}\|_2^2$
- **Total Variation:**  $\|\nabla\mathbf{x}\|_1$  for piecewise constant signals

## 9 Conclusion

The transition from  $\ell_0$  to  $\ell_1$  optimization represents a fundamental shift in sparse coding methodology. By leveraging the convex optimization framework, we obtain:

1. **Computational Tractability:** Polynomial-time algorithms with global optimality guarantees
2. **Sparsity Promotion:** Geometric properties that encourage sparse solutions
3. **Theoretical Foundation:** Rigorous mathematical framework for analysis

#### 4. **Practical Effectiveness:** Wide applicability across signal processing domains

The proximal gradient method provides an elegant solution to the non-differentiability challenge, enabling efficient optimization of  $\ell_1$ -regularized problems. This framework continues to be a cornerstone of modern sparse signal processing and machine learning applications.

## 10 Mathematical Appendix

### 10.1 Subdifferential Calculus

For non-differentiable convex functions, we use the concept of subdifferentials:

**Definition 10.1** (Subdifferential). The subdifferential of a convex function  $f$  at  $\mathbf{x}$  is:

$$\partial f(\mathbf{x}) = \{\mathbf{g} \in \mathbb{R}^n : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{y}\} \quad (28)$$

### 10.2 Optimality Conditions

**Theorem 10.2** (First-Order Optimality Condition). For the problem  $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$  where  $f$  is differentiable and  $g$  is convex,  $\mathbf{x}^*$  is optimal if and only if:

$$0 \in \nabla f(\mathbf{x}^*) + \partial g(\mathbf{x}^*) \quad (29)$$

## 11 Glossary of Symbols

$\mathbf{x}$	Vector in $\mathbb{R}^n$
$\ \cdot\ _p$	$\ell_p$ norm
$\mathbf{A}$	Dictionary matrix
$\mathbf{b}$	Observed signal
$\lambda$	Regularization parameter
$\epsilon$	Noise tolerance
$\text{prox}_{\lambda g}$	Proximal operator of $g$
$\partial f$	Subdifferential of $f$
$L$	Lipschitz constant
$\gamma$	Step size parameter