A Detailed Introduction to the LASSO and ℓ_1–Regularized Least Squares

Contents

1 Overview

The goal of this document is to give a detailed, math–heavy treatment of the LASSO (Least Absolute Shrinkage and Selection Operator) and its close relative in signal processing, the *basis pursuit denoising* (BPD) formulation.

We start from a generic linear inverse problem

$$y = Ax + w,$$

compare standard least squares to ℓ_1–based methods, and then study:

- Penalized and constrained LASSO formulations.
- Convexity and optimality conditions (via subgradients and KKT).
- Closed–form solutions in simple cases (soft–thresholding).
- Examples: sparse denoising, deconvolution, missing data, etc.
- First–order algorithms: ISTA, FISTA, and splitting / SALSA–type ideas.

Throughout, we interpret BPD,

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 ,$$

as the canonical *LASSO* problem in the signal–processing setting.

2 Notation and quick reminders

We use standard finite–dimensional linear algebra:

- Vectors are columns in $\mathbb{R}^N$; matrices are real $M \times N$ unless otherwise stated.
- For a matrix A, A^T is the transpose.
- For $x \in \mathbb{R}^N$, the ℓ_p–norms for $p \geq 1$ are

$$\|x\|_p = \Big(\sum_{i=1}^N |x_i|^p \Big)^{1/p}, \quad \|x\|_\infty = \max_i |x_i|.$$

- The ℓ_0 *pseudo–norm* is

$$\|x\|_0 = \#\{i : x_i \neq 0\},$$

  i.e., the number of nonzero entries (*sparsity*).

Below are one–line reminders for all mathematical terms beyond basic calculus (derivatives, integrals, limits):

- **Norm:** A function $\|\cdot\|$ on $\mathbb{R}^N$ that is nonnegative, positively homogeneous, zero only at 0, and satisfies the triangle inequality.

- **Inner product:** A map $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^N$ that is bilinear, symmetric, and positive definite; it induces a norm via $\|x\|_2 = \sqrt{\langle x, x \rangle}$.
- **Convex set:** A set C such that for any $x, y \in C$ and $\theta \in [0,1]$, $\theta x + (1-\theta)y \in C$.
- **Convex function:** A function f with domain a convex set such that $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for all x, y, $\theta \in [0,1]$.
- **Subgradient:** For convex f, a vector g is a subgradient at x if $f(z) \geq f(x) + \langle g, z - x \rangle$ for all z; the set of all such g is the *subdifferential* $\partial f(x)$.
- **Proximal operator:** For proper closed convex g and $\tau > 0$,

  $$\operatorname{prox}_{\tau g}(v) = \arg\min_x \left(\frac{1}{2} \|x - v\|_2^2 + \tau g(x) \right),$$

  a "regularized projection" of v.
- **Lipschitz continuous gradient:** A differentiable function f has Lipschitz continuous gradient with constant L if $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2$ for all x, y.
- **Spectral norm:** For a matrix A, $\|A\|_2$ is the largest singular value, equivalently $\max_{\|x\|_2=1} \|Ax\|_2$.
- **Eigenvalue:** A scalar λ such that $Av = \lambda v$ for some nonzero vector v (an eigenvector).
- **Parseval frame (tight frame):** A matrix A whose columns satisfy $AA^T = pI$ for some $p > 0$, a generalization of an orthonormal basis.
- **KKT conditions:** Necessary and often sufficient optimality conditions for constrained convex optimization, involving primal feasibility, dual feasibility, and complementary slackness.
- **Restricted isometry property (RIP):** A property of a matrix A stating that all s–sparse vectors have nearly preserved ℓ_2–norm under A.
- **Mutual coherence:** For a matrix A with normalized columns a_i, $\mu(A) = \max_{i \neq j} |\langle a_i, a_j \rangle|$, measuring how correlated different columns are.
- **Support of a vector:** $\operatorname{supp}(x) = \{i : x_i \neq 0\}$, the index set of nonzeros.
- **Soft–thresholding operator:** $S_\alpha(t) = \operatorname{sgn}(t) \max(|t| - \alpha, 0)$, applied componentwise to vectors; it is the proximal operator of $\alpha \|\cdot\|_1$.
- **Iterative shrinkage/thresholding algorithm (ISTA):** A proximal gradient method for problems with smooth + ℓ_1 terms, using gradient descent plus soft–thresholding.
- **FISTA (Fast ISTA):** An accelerated version of ISTA that uses a momentum term to improve convergence rates from $O(1/k)$ to $O(1/k^2)$ in objective value.
- **Augmented Lagrangian / ADMM:** Splitting methods that solve constrained problems by iteratively minimizing an augmented Lagrangian and updating dual variables.

3 From least squares to the LASSO

3.1 Linear inverse problem and least squares

Consider a linear model

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

where

- $y \in \mathbb{R}^M$ is the observed data,
- $A \in \mathbb{R}^{M\times N}$ is a known sensing / design matrix,
- $x \in \mathbb{R}^N$ is an unknown parameter vector or signal,
- $w \in \mathbb{R}^M$ is additive noise or modeling error.

When $M \geq N$ and A has full column rank, the classical least squares estimate is

$$\hat{x}_{\mathrm{LS}} = \arg\min_x \|y - Ax\|_2^2 = (A^T A)^{-1} A^T y. \tag{2}$$

When $M < N$ (underdetermined system), $A^T A$ is singular and there are infinitely many solutions to $y = Ax$. A standard choice is the minimum–norm solution:

$$\hat{x}_{\mathrm{MN}} = \arg\min_x \|x\|_2^2 \quad \text{s.t. } y = Ax = A^T(AA^T)^{-1}y, \tag{3}$$

assuming AA^T is invertible.

Reminder: In underdetermined problems, least squares (or minimum ℓ_2–norm) does *not* encourage sparsity; it spreads energy among coordinates to reduce the ℓ_2 norm.

3.2 Sparsity and the ℓ_0 formulation

In many signal processing and statistical settings, we believe that x is *sparse*, meaning that only a small number of entries in x are nonzero. A natural formulation is

$$\min_x \ \|x\|_0 \quad \text{s.t.} \quad y = Ax, \tag{4}$$

or in the noisy case,

$$\min_x \ \|x\|_0 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \varepsilon. \tag{5}$$

Reminder: Minimizing $\|x\|_0$ is combinatorial and NP–hard in general, because it essentially searches over subsets of columns of A.

3.3 Relaxation: from ℓ_0 to ℓ_1

The LASSO replaces the nonconvex ℓ_0 objective by its convex surrogate ℓ_1. Two common forms appear:

Penalized (LASSO / BPD) form.

$$\min_{x\in\mathbb{R}^N} \ \frac{1}{2}\|y - Ax\|_2^2 + \lambda \|x\|_1, \qquad \lambda > 0. \tag{6}$$

In signal processing this is frequently called *basis pursuit denoising (BPD).*

Constrained (classic LASSO) form.

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\|y-Ax\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \le \tau, \tag{7}$$

for some radius $\tau > 0$.

For every τ under mild conditions, there exists a λ such that the constrained and penalized forms have the same solution set (this follows from convex duality and KKT).

Reminder: The ℓ_1 norm is the tightest convex lower bound of $\|\cdot\|_0$ on the unit ℓ_∞ ball, which explains why it is the standard convex surrogate for sparsity.

3.4 Geometric intuition: ℓ_1 vs. ℓ_2

To understand sparsity promotion, consider a very small case $N = 2$, $M = 2$. The least squares solution minimizes $\|y - Ax\|_2^2$ with an implicit quadratic regularizer, while the constrained LASSO solves

$$\min_x \|y-Ax\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \le \tau.$$

The feasible set $\{x : \|x\|_1 \le \tau\}$ is a diamond in $\mathbb{R}^2$, while $\{x : \|x\|_2 \le r\}$ is a disk. The optimum of a convex quadratic over a diamond tends to occur at a vertex, i.e., at points where one coordinate is exactly zero. Over a disk, the optimum rarely occurs on coordinate axes.

Reminder: This geometric picture explains why ℓ_1 regularization naturally yields sparse solutions (many coordinates pinned to zero), whereas ℓ_2 regularization does not.

4 Optimality conditions for LASSO

We now analyze the penalized LASSO / BPD problem

$$\min_{x\in\mathbb{R}^N} F(x) := \frac{1}{2}\|y-Ax\|_2^2 + \lambda\|x\|_1. \tag{8}$$

The objective F is the sum of a smooth convex function and a nonsmooth convex function.

4.1 Convexity and existence of a minimizer

The function $f(x) = \frac{1}{2}\|y - Ax\|_2^2$ is convex because it is a composition of affine and convex functions:

$$f(x) = \frac{1}{2}\|Ax-y\|_2^2 = \frac{1}{2}(Ax-y)^T(Ax-y),$$

and its Hessian is $A^TA \succeq 0$ (positive semidefinite). The function $g(x) = \lambda\|x\|_1$ is also convex as a norm scaled by $\lambda > 0$. Therefore, $F = f + g$ is convex.

If A has full column rank (or more generally if $\ker(A)$ intersects the level sets of g in a nice way), $F(x) \to \infty$ as $\|x\|_2 \to \infty$ (coercivity), so a minimizer exists and the set of minimizers is nonempty and convex.

4.2 Subgradient optimality conditions

Because g is nonsmooth, we use subgradients. For

$$f(x) = \frac{1}{2} \|y - Ax\|_2^2,$$

the gradient is

$$\nabla f(x) = A^T(Ax - y). \tag{9}$$

For $g(x) = \lambda \|x\|_1$, the subdifferential $\partial g(x)$ can be described componentwise:

$$\partial(\lambda|x_i|) = \begin{cases} \{\lambda \operatorname{sgn}(x_i)\}, & x_i \neq 0, \\ [-\lambda, \lambda], & x_i = 0, \end{cases}$$

and $\partial g(x) = \{v \in \mathbb{R}^N : v_i \in \partial(\lambda|x_i|)\}$.

A point $x^\star$ is optimal if and only if

$$0 \in \nabla f(x^\star) + \partial g(x^\star) \quad \Longleftrightarrow \quad -A^T(Ax^\star - y) \in \partial g(x^\star). \tag{10}$$

Componentwise, this means:

$$-[A^T(Ax^\star - y)]_i = \begin{cases} \lambda \operatorname{sgn}(x_i^\star), & x_i^\star \neq 0, \\ \in [-\lambda, \lambda], & x_i^\star = 0. \end{cases}$$

Reminder: Subgradient optimality conditions generalize the condition $\nabla F(x^\star) = 0$ to nonsmooth convex problems.

5 Soft–thresholding examples

5.1 Scalar LASSO: closed form solution

Consider the 1D problem

$$\min_{x \in \mathbb{R}} \frac{1}{2}(y - x)^2 + \lambda|x|. \tag{11}$$

This is LASSO with $A = 1$ and $\lambda > 0$. We can solve it explicitly.

For $x > 0$, the objective is $F(x) = \frac{1}{2}(y - x)^2 + \lambda x$. Differentiating (ordinary derivative) and setting to zero:

$$F'(x) = -(y - x) + \lambda = x - y + \lambda = 0 \quad \Rightarrow \quad x = y - \lambda.$$

This candidate must satisfy $x > 0$, so $y - \lambda > 0 \Rightarrow y > \lambda$.

Similarly, for $x < 0$, we write $|x| = -x$, so $F(x) = \frac{1}{2}(y - x)^2 + \lambda(-x)$ and

$$F'(x) = x - y - \lambda = 0 \quad \Rightarrow \quad x = y + \lambda,$$

which must satisfy $x < 0$, so $y + \lambda < 0 \Rightarrow y < -\lambda$.

For $x = 0$, we use subgradient optimality. The subdifferential of $\lambda|x|$ at $x = 0$ is $[-\lambda, \lambda]$. The derivative of $\frac{1}{2}(y - x)^2$ at 0 is $-(y - 0) = -y$. So $x = 0$ is optimal if

$$0 \in -y + [-\lambda, \lambda] \quad \Leftrightarrow \quad y \in [-\lambda, \lambda] \quad \Leftrightarrow \quad |y| \leq \lambda.$$

Putting these cases together,

$$x^\star = \begin{cases} y - \lambda, & y > \lambda, \\ 0, & |y| \le \lambda, \\ y + \lambda, & y < -\lambda. \end{cases}$$

This is exactly the scalar soft–thresholding operator:

$$x^\star = S_\lambda(y) := \operatorname{sgn}(y) \max(|y| - \lambda, 0).$$

Reminder: Soft–thresholding shrinks y toward zero by λ and sets it to zero if it is within a λ–sized deadzone around zero.

5.2 Orthogonal design: componentwise soft–thresholding

If A has orthonormal columns, $A^T A = I$, we can express LASSO in a diagonalized form. Let $z = A^T y$ (the least squares coefficients), and consider

$$F(x) = \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 .$$

Using $A^T A = I$ and $AA^T = I$ (square orthonormal case),

$$\|y - Ax\|_2^2 = \left\|A^T y - A^T A x\right\|_2^2 = \|z - x\|_2^2 .$$

Thus the problem becomes

$$\min_x \frac{1}{2} \|z - x\|_2^2 + \lambda \|x\|_1 = \sum_{i=1}^N \left[\frac{1}{2}(z_i - x_i)^2 + \lambda |x_i|\right],$$

which decouples componentwise. Each coordinate solves a scalar LASSO as in (11) with $y = z_i$, so the solution is

$$x_i^\star = S_\lambda(z_i) = S_\lambda((A^T y)_i), \quad i = 1, \ldots, N.$$

Equivalently,

$$x^\star = S_\lambda(A^T y), \tag{12}$$

where S_λ acts componentwise.

Reminder: An orthonormal design makes LASSO completely separable across coordinates, yielding an exact closed form in terms of soft–thresholding of the least squares coefficients.

6 LASSO as basis pursuit denoising and sparse signal models

6.1 Basis pursuit and BPD

Recall the constrained *basis pursuit* (BP) problem:

$$\min_x \|x\|_1 \quad \text{s.t. } y = Ax, \tag{13}$$

an ℓ_1–based replacement for (4). When y is noisy, it is more reasonable to allow a misfit and consider

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 , \tag{14}$$

which is exactly (6). This is called *basis pursuit denoising (BPD)* in the sparse signal processing literature and *LASSO* in statistics.

6.2 Sparse representation model

In many examples, x is not itself the signal of interest, but rather represents coefficients in some transform domain. We write

$$s = Ax, \tag{15}$$

where

- $s \in \mathbb{R}^M$ is the signal (e.g., a short speech frame),
- $A \in \mathbb{R}^{M \times N}$ is a dictionary (e.g., DFT, wavelets, time–frequency atoms),
- $x \in \mathbb{R}^N$ is the coefficient vector, hoped to be sparse.

Given data y that is a noisy version of s, or data that is s passed through a linear system, we can formulate BPD / LASSO in the coefficient domain. Examples include:

- Denoising: $y = s + w = Ax + w$; recover sparse x via LASSO, then reconstruct s.
- Deconvolution: $y = Hs + w = HAx + w$, where H is convolution; LASSO on x.
- Inpainting / missing data: $y = Ss = SAx$, where S selects observed samples.

Reminder: In all these cases, the success of LASSO critically depends on the assumption that x is sparse (or compressible) in the chosen dictionary A.

7 Example: LASSO for denoising

Consider a 1D signal $s \in \mathbb{R}^M$ that is sparsely represented in some dictionary A, so $s = Ax^\star$ for a sparse vector $x^\star$. We observe

$$y = s + w = Ax^\star + w,$$

with w i.i.d. Gaussian noise.

We solve

$$\hat{x}_\lambda = \arg\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 , \tag{16}$$

and reconstruct $\hat{s} = A\hat{x}_\lambda$.

7.1 Choice of λ and bias–variance tradeoff

For Gaussian noise with variance σ^2, a common heuristic is

$$\lambda \propto \sigma\sqrt{2 \log N}$$

to balance false alarm and detection probabilities of nonzero coefficients. As λ increases:

- More coefficients are driven to zero $\Rightarrow$ stronger denoising, but more bias.
- Fewer nonzeros $\Rightarrow$ less variance due to noise in the coefficients.

Thus λ controls a bias–variance tradeoff and is typically chosen by cross–validation or by analytic risk estimates (Stein's unbiased risk estimate, etc., in some settings).

7.2 Special case: Parseval frame

If A is a tight frame with $AA^T = pI$, LASSO structure changes slightly but maintains nice properties. In particular, $f(x) = \frac{1}{2}\|y - Ax\|_2^2$ has a Lipschitz gradient with

$$L = \left\|A^T A\right\|_2 = p.$$

This is useful for algorithm design (step–size selection in ISTA / FISTA).

Reminder: A tight frame behaves like an orthonormal basis up to a scalar factor, so many proofs and algorithm analyses carry over with only minor modifications.

8 Example: LASSO for sparse deconvolution

Let $x^\star \in \mathbb{R}^N$ be a sparse spike train, and $h \in \mathbb{R}^L$ be a known impulse response. We observe

$$y = h * x^\star + w,$$

where $*$ denotes linear convolution, and w is noise. Writing this as $y = Hx^\star + w$ with Toeplitz convolution matrix $H \in \mathbb{R}^{M\times N}$ ($M = N + L - 1$), we consider

$$\hat{x}_\lambda = \arg\min_x \frac{1}{2}\|y - Hx\|_2^2 + \lambda\|x\|_1. \tag{17}$$

8.1 Comparison with ℓ_2–regularized deconvolution

The classical Tikhonov–regularized least squares deconvolution

$$\hat{x}_{\text{Tik}} = \arg\min_x \frac{1}{2}\|y - Hx\|_2^2 + \frac{\gamma}{2}\|x\|_2^2 \tag{18}$$

has the closed form

$$\hat{x}_{\text{Tik}} = (H^T H + \gamma I)^{-1} H^T y,$$

and tends to produce a *smoothed* version of $x^\star$. When $x^\star$ is sparse (spikes), the LASSO formulation (17) tends to produce sharp, well–localized spikes matching the locations of the true impulses, whereas Tikhonov spreads them.

Reminder: In deconvolution, ℓ_1 regularization can recover sparse structures that are heavily blurred in the observations, provided that the blur h is not too ill–conditioned.

9 Iterative algorithms for LASSO

9.1 Proximal gradient / ISTA

We split $F(x) = f(x) + g(x)$ with

$$f(x) = \frac{1}{2}\|y - Ax\|_2^2, \qquad g(x) = \lambda\|x\|_1.$$

The gradient of f is Lipschitz with constant $L = \left\|A^T A\right\|_2$:

$$\nabla f(x) = A^T(Ax - y).$$

Proximal gradient (ISTA) iterates

$$x^{k+1} = \operatorname{prox}_{\frac{\lambda}{L}\|\cdot\|_1}\Big(x^k - \frac{1}{L}\nabla f(x^k)\Big), \tag{19}$$

for any $L \geq \left\|A^T A\right\|_2$. Using the fact that

$$\operatorname{prox}_{\alpha\|\cdot\|_1}(v) = S_\alpha(v),$$

we get the explicit iteration

$$x^{k+1} = S_{\lambda/L}\Big(x^k - \frac{1}{L}A^T(Ax^k - y)\Big), \tag{20}$$

where $S_{\lambda/L}$ is applied componentwise.

Reminder: ISTA is guaranteed to converge to a minimizer of F for any $L \geq \left\|A^T A\right\|_2$, with convergence rate $O(1/k)$ in objective value.

9.2 FISTA: accelerated proximal gradient

FISTA introduces an extrapolated variable z^k and a momentum parameter t_k:

$$x^{k+1} = S_{\lambda/L}\Big(z^k - \frac{1}{L}A^T(Az^k - y)\Big), \tag{21}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \tag{22}$$

$$z^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}}(x^{k+1} - x^k), \tag{23}$$

with initial x^0, $z^0 = x^0$, $t_0 = 1$.

FISTA enjoys a faster $O(1/k^2)$ rate in objective value:

$$F(x^k) - F(x^\star) \leq \frac{C}{k^2}$$

for some constant C.

Reminder: FISTA is often preferred in practice over ISTA because it achieves much faster decrease in objective for essentially the same per–iteration cost.

9.3 Splitting and SALSA–type methods

Another class of algorithms, often used in signal processing, introduces an auxiliary variable z so that the ℓ_1 term is separated. For example, rewrite

$$\min_{x,z} \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|z\|_1 \quad \text{s.t.} \quad x = z.$$

The augmented Lagrangian for this constrained problem leads to ADMM–type iterations:

$$\begin{aligned}
x^{k+1} &:= \arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \frac{\rho}{2}\left\|x - z^k + u^k\right\|_2^2, \\
z^{k+1} &:= \arg\min_z \lambda\|z\|_1 + \frac{\rho}{2}\left\|x^{k+1} - z + u^k\right\|_2^2 = S_{\lambda/\rho}(x^{k+1} + u^k), \\
u^{k+1} &:= u^k + x^{k+1} - z^{k+1},
\end{aligned}$$

where u is a dual / scaled Lagrange multiplier, and $\rho > 0$ is a penalty parameter.

Reminder: In many signal processing problems A has structure (convolution, transforms) that allows efficient solution of the quadratic x–update, making such splitting approaches very competitive for large–scale LASSO / BPD.

10 Conditions for sparse recovery

LASSO not only produces sparse solutions; under conditions on A and noise level, it recovers the true sparse $x^\star$ exactly or approximately.

10.1 Noiseless exact recovery

In the noiseless case $y = Ax^\star$, under appropriate conditions, the BP problem (13) has a unique solution equal to $x^\star$ if $\|x^\star\|_0$ is small enough relative to properties of A.

Two standard sufficient conditions:

Mutual coherence condition. If the columns of A are ℓ_2–normalized, and $x^\star$ is s–sparse, one condition is

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

where $\mu(A)$ is the mutual coherence.

Restricted isometry property (RIP). If A satisfies RIP of order $2s$ with constant $\delta_{2s} < \sqrt{2}-1$, then every s–sparse $x^\star$ is the unique minimizer of $\min_x \|x\|_1$ s.t. $y = Ax$.

Reminder: These conditions guarantee that A does not mix different sparse supports too much, so that sparsity and data consistency together identify a unique solution.

10.2 Noisy case and LASSO

In the noisy case $y = Ax^\star + w$, under RIP and appropriate choice of λ depending on $\|A^T w\|_\infty$, LASSO solutions obey bounds of the form

$$\|\hat{x}_\lambda - x^\star\|_2 \le C_0 \frac{\sigma_s(x^\star)_1}{\sqrt{s}} + C_1 \frac{\lambda}{\phi},$$

where $\sigma_s(x^\star)_1$ is the best s–term approximation error in ℓ_1 and ϕ is a stability constant derived from RIP. Such results formalize that LASSO recovers sparse signals stably even in noise.

Reminder: The precise constants and assumptions vary by theorem, but the qualitative picture is that ℓ_1–based methods are near–optimal for sparse recovery under randomness or RIP.

11 LASSO in the context of Selesnick–style problems

Here we explicitly connect the generic LASSO formulation to some canonical sparse signal processing problems of the type discussed in sparsity notes.

11.1 Sparse Fourier coefficients

Let $y \in \mathbb{R}^M$ be a short segment of a real signal, and $A \in \mathbb{C}^{M\times N}$ be the first M rows of an inverse N–point DFT matrix. We model

$$y = Ac + w,$$

where $c \in \mathbb{C}^N$ are Fourier coefficients. If y consists of a small number of pure tones that do not align with the discrete frequency grid, then c is still sparse if we allow sufficiently fine N (oversampled grid). The BPD / LASSO problem

$$\min_c \frac{1}{2} \|y - Ac\|_2^2 + \lambda \|c\|_1$$

selects a sparse set of Fourier components that explain the data, avoiding spectral leakage and yielding frequency estimates sharper than standard DFT of the raw samples.

Reminder: Here, LASSO is performing *sparse spectral estimation* on an oversampled dictionary of sinusoids.

11.2 Denoising in the Fourier domain

For noisy speech (or any signal) that is sparse in the Fourier domain, the same formulation applies: given y as noisy time–domain samples, we choose A as partial inverse DFT, solve LASSO for coefficients c, and reconstruct $\hat{s} = A\hat{c}$. The sparsity of $\hat{c}$ reflects the limited number of dominant frequencies; noise components are suppressed by soft–thresholding.

11.3 Deconvolution with sparse spikes

Let $x^\star$ be a sparse spike train and h a short blur kernel; then $H \in \mathbb{R}^{M\times N}$ is a banded convolution matrix. LASSO

$$\hat{x} = \arg\min_x \frac{1}{2} \|y - Hx\|_2^2 + \lambda \|x\|_1$$

recovers spike locations and amplitudes that are much closer to the true ones than a least–squares or ℓ_2–penalized approach, exploiting the sparsity structure.

11.4 Missing data / inpainting

Let S be a selection matrix that picks observed samples of s, so $y = Ss$, and suppose $s = Ac$ with sparse c. Then $y = SAc$ and we solve

$$\hat{c} = \arg\min_c \frac{1}{2} \|y - SAc\|_2^2 + \lambda \|c\|_1 ,$$

and define $\hat{s} = A\hat{c}$. The missing entries (unobserved components of S) are filled in by the sparse model. This is a LASSO instance in the coefficient domain.

11.5 Morphological component separation

For a signal y consisting of two components y_1 and y_2, each sparse in different dictionaries A_1 and A_2, respectively,

$$y = y_1 + y_2 = A_1c_1 + A_2c_2,$$

we solve the joint LASSO–type problem

$$\min_{c_1,c_2} \frac{1}{2} \|y - A_1 c_1 - A_2 c_2\|_2^2 + \lambda_1 \|c_1\|_1 + \lambda_2 \|c_2\|_1 \,,$$

and set $y_i = A_i \hat{c}_i$. This is sometimes called "dual basis pursuit" but is simply a multi–block extension of LASSO.

Reminder: The key idea is sparsity in different transform domains (morphologies) that allow the optimization to allocate each component to the dictionary in which it is sparse.

12 Summary

We have presented an extensive overview of the LASSO / BPD problem

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1$$

from a signal–processing and inverse–problems perspective. The main messages are:

- LASSO replaces an intractable ℓ_0 objective with a convex ℓ_1 surrogate, yielding computationally feasible and theoretically well–understood optimization problems.
- The geometric shape of the ℓ_1–ball favors sparse solutions, in contrast to ℓ_2–based methods which tend to produce dense, low–energy solutions.
- In special cases (scalar, orthogonal design) LASSO admits closed–form solutions via soft–thresholding, illuminating the role of shrinkage and sparsity.
- In general, LASSO is efficiently solvable via first–order methods (ISTA, FISTA, and splitting / ADMM–type algorithms), especially when A has structure (frames, convolutions, transforms).
- Under appropriate conditions (RIP, coherence bounds), LASSO provides strong guarantees for sparse recovery, both in noiseless and noisy settings.
- Many practical signal processing problems — including denoising, deconvolution, inpainting, and component separation — can be expressed as LASSO or multi–block LASSO problems by modeling signals as sparse in suitable dictionaries.