

Lecture 20: Basis pursuit, compressed sensing, and LASSO

Lecturer: Yiqiao Zhong

Report an issue

Last update: October 19, 2026

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Optional part. This lecture belongs to optional Part VI and is not examined. Its high, mid and low labels organize a suggested reading path within the optional material; they do not add exam requirements. See the Part VI overview.

High priority *Overview*

An underdetermined system can identify a sparse signal if its null space avoids the relevant error directions. Basis pursuit makes this geometric idea into a convex optimization problem. We compare pairwise incoherence with restricted isometry, then add noise and derive the LASSO error bound from a basic inequality and restricted curvature. The main route keeps the assumptions and normalization of each guarantee explicit.

Recap from earlier lectures

From Lectures 13, 14 and 19: earlier ideas used here

Reminder only: this block introduces no new exam requirements.

- **Sparsity.** Exact sparsity counts nonzero coordinates; the support records their locations. Used in basis pursuit.
- **Orthogonal LASSO.** The loss normalization and column normalization determine the soft threshold. Used to extend LASSO to general designs.
- **Linear regression geometry.** A null vector changes coefficients without changing fitted values. Used in exact recovery.
- **Concentration.** Tail bounds control averages and can be combined across many events. Used for random-design incoherence.

An asterisk in a topic heading marks optional material; the priority labels specify the reading scope.

Learning goals

- Interpret basis pursuit using its feasible space and the geometry of the norm ball.
- Distinguish target-specific recovery from a uniform support guarantee.
- Compare incoherence, restricted isometry and restricted eigenvalues.
- Follow the basic inequality that controls LASSO estimation error.

High priority *Introduction*

We will explore several topics in statistics that exploit sparsity, spanning the 1990s to the 2000s. The conceptual framework also underlies recent progress in deep neural networks.

1 Basis pursuit

High priority *Basis pursuit*

Basis pursuit (BP) is a precursor of LASSO where we want to recover a sparse solution from an underdetermined linear system of equations. Consider $y = X\theta$ where $X \in \mathbb{R}^{n \times p}$ and $n < p$. We want to find a sparse solution θ . Suppose that there is a true sparse signal (measured by the ℓ_0 count) θ^* satisfying $y = X\theta^*$. Our goal is to recover this vector from measurement data X and y . BP solves the following convex problem.

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & \|\theta\|_1 \\ \text{subject to} \quad & y = X\theta, \end{aligned} \tag{BP}$$

The solution need not be unique: for $X = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $y = 1$, every $(t, 1 - t)$ with $0 \leq t \leq 1$ minimizes the objective. Uniqueness requires an additional condition. The feasible set (parameters satisfying the linear constraints) is

$$\theta^* + \text{Ker}(X), \quad \text{where } \text{Ker}(X) = \{\theta : X\theta = 0\}$$

an affine subspace of dimension $p - \text{rank}(X)$; it is a hyperplane only when $\text{rank}(X) = 1$.

High priority *Practice: Basis pursuit with one measurement*

Exercise prompt. For $X = \begin{pmatrix} 1 & 2 \end{pmatrix}$ and $y = 2$, find the unique basis-pursuit solution.

See Exercise 1 in Section 6 for the full statement and complete solution.

1.1 Restricted null space property

High priority *Restricted null space property*

We ask the question: What does the solution $\arg \min\{\|\theta\|_1 : \theta \in \theta^* + \text{Ker}(X)\}$ look like?

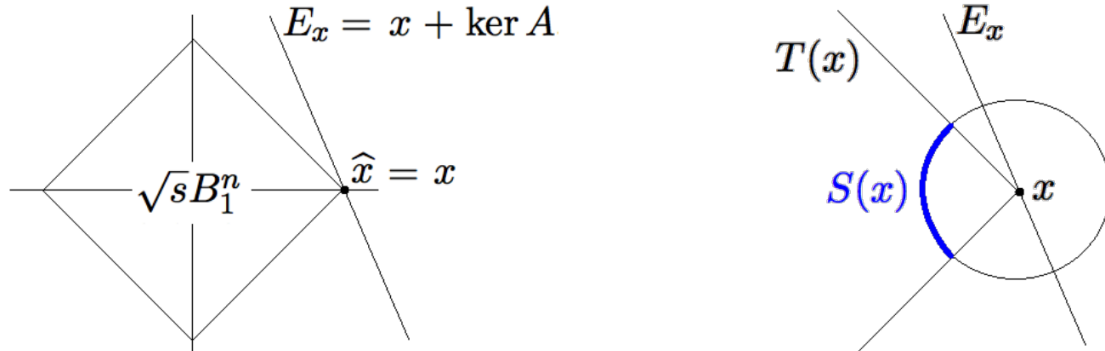


Figure 1: Left: the feasible affine space touches the ℓ_1 ball at the target. Right: directions that enter the ball must be excluded from the null space. The source uses $A = X$, $x = \theta^*$ and $E_x = \theta^* + \text{ker } X$; its displayed ball radius is a normalization for its unit target. The labels $T(x)$ and $S(x)$ mark the cone into the ball and its unit directions, drawn at x . Figure: Vershynin, Figure 10.5, diagrams only.

Let $B_1(r) = \{\theta : \|\theta\|_1 \leq r\}$. Increase r from zero until this ball first intersects the feasible affine subspace. The first contact radius is the minimum objective value. A single contact point gives a unique optimizer; contact along a face gives multiple optimizers. The ℓ_1 norm is convex, not strictly convex: the midpoint of two equal-norm points can have the same norm. Thus the number of intersections with a level surface alone does not determine whether the prescribed θ^* is recovered.

Blessing of high dimensions. The ℓ_1 ball in p dimensions, namely

$$B_1^p(r) = \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq r\}$$

is intuitively very pointy, meaning that the “angle” is very sharp at all corner points. This is in contrast with low dimensions such as 2D. More concretely, consider the simple case $\theta^* = (1, 0, 0, \dots, 0)^\top$. BP recovers this sparse signal θ^* iff the feasible affine subspace first touches the ℓ_1 ball at θ^* alone, i.e.,

$$B_1(\|\theta^*\|_1) \cap (\theta^* + \text{Ker}(X)) = \{\theta^*\}.$$

Instructor's Note *Pointy corners and enough measurements*

Keep the geometric picture: the ℓ_1 ball has corners on coordinate axes, so a feasible set can touch a sparse point first. Each measurement removes a direction of freedom. Recovery succeeds when the remaining null directions cannot move us away from the sparse corner without increasing the ℓ_1 norm.

For one Gaussian measurement, $X = a^\top \in \mathbb{R}^{1 \times p}$, and target e_1 , basis pursuit recovers e_1 uniquely exactly when $|a_1| > \max_{j>1} |a_j|$. Indeed the cheapest representation of $y = a_1$ uses the column with largest absolute value. The continuous i.i.d. entries are exchangeable, so this success probability is $1/p$. One measurement leaves too much freedom; the bounds below quantify how additional measurements improve recovery.

RNS: formalizing geometric intuitions. Denote the support set of the true signal as $S = \{j \leq p : \theta_j^* \neq 0\}$. Define the scale-invariant set (not generally a convex cone) as

$$C(S) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\},$$

Here $\Delta_S = (\Delta_j)_{j \in S}$ is the coordinate subvector; thus $X_S \Delta_S$ has the usual matrix dimensions. We say the design matrix X satisfies the *restricted null space (RNS) property* if

$$\text{Ker}(X) \cap C(S) = \{0\}.$$

Proposition 1. *If X satisfies the RNS property, then BP recovers θ^* .*

Proof. Suppose, for a contradiction, that we can find another θ in the feasible set with an ℓ_1 norm no larger than that of θ^* . Then, there exists $\Delta \neq 0$, such that

$$X\Delta = 0, \quad \|\theta^* + \Delta\|_1 \leq \|\theta^*\|_1.$$

By our definition, $\|u\|_1 = \|u_S\|_1 + \|u_{S^c}\|_1$ for any vector u , so

$$\begin{aligned} \|\theta^* + \Delta\|_1 &= \|\theta_S^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \quad (\text{since } \theta_{S^c}^* = 0 \text{ by definition}) \\ &\geq \|\theta_S^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \quad (\text{by triangle inequality}) \\ &> \|\theta_S^*\|_1 \quad (\text{by RNS property}) \\ &= \|\theta^*\|_1, \end{aligned}$$

which leads to a contradiction! □

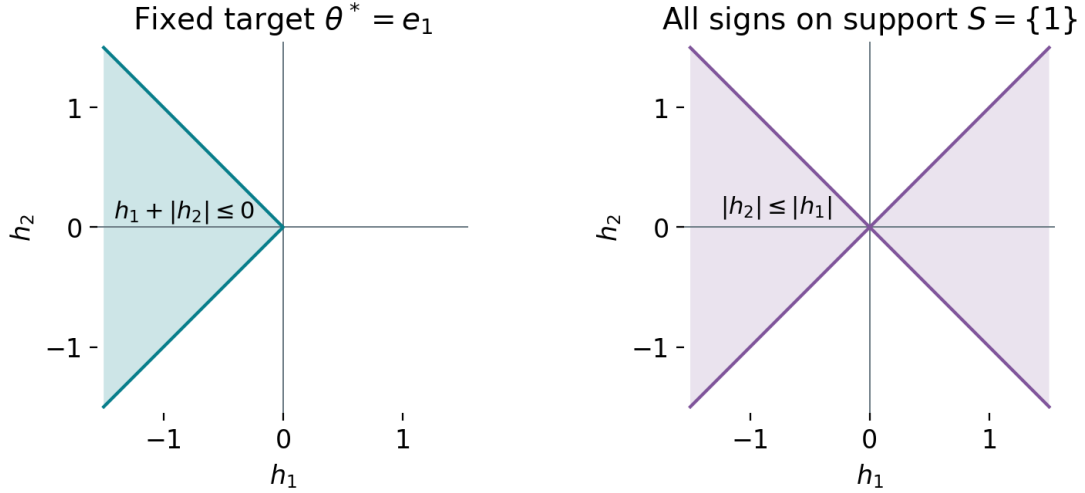
From an optimization perspective, $C(S)$ contains all displacements that can avoid increasing $\|\theta^*\|_1$, but it can be larger than the descent cone for a fixed sign pattern. RNS is a sufficient condition for this target and characterizes uniform recovery over every vector supported on S . For a fixed target, unique recovery means no nonzero feasible displacement has a nonincreasing objective.

Restricted null space. If no nonzero null vector satisfies $\|h_{S^c}\|_1 \leq \|h_S\|_1$, then a signal supported on S is the unique basis-pursuit solution. This is a restricted condition, not full column rank.

For a concrete comparison, take $\theta^* = e_1 \in \mathbb{R}^2$. For sufficiently small $t > 0$,

$$\|e_1 + th\|_1 = 1 + t(h_1 + |h_2|).$$

Thus the fixed-sign descent cone is $h_1 + |h_2| \leq 0$. Allowing either sign of the signal produces the larger set $|h_2| \leq |h_1|$ used in RNS.



A null direction must avoid the relevant set (except at zero) for unique recovery.

Figure 2: For a fixed target $\theta^* = e_1$, the local nonincreasing directions satisfy $h_1 + |h_2| \leq 0$ (left). Uniform recovery on $S = \{1\}$ uses the larger, two-sided set $C(S) = \{h : |h_2| \leq |h_1|\}$ (right). A null direction must avoid the relevant set except at zero.

1.2 Pairwise incoherence: a key condition on the design matrix

Mid priority *Pairwise incoherence: a key condition on the design matrix*

The RNS property looks a bit abstract. So we want to understand what kind of X satisfies this property.

Definition 1 (Pairwise incoherence, aka mutual incoherence). Let $X = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$. The pairwise incoherence (PI) is

$$\delta_{\text{pw}}(X) = \left\| \frac{1}{n} X^\top X - I_p \right\|_{\max} = \max_{k, j \leq p} \left| \frac{1}{n} \langle X_k, X_j \rangle - \mathbb{1}(k = j) \right|.$$

For every n, p , this measures the largest entrywise error in $X^\top X/n$ from I_p . It controls column-length errors as well as correlations; when all columns have norm $\sqrt{n}$, only the off-diagonal correlations remain. A remarkable fact is that we can have a small $\delta_{\text{pw}}(X)$ even when $p \gg n$.

Proposition 2 (Exponentially many random vectors have small PI). Suppose $X \in \mathbb{R}^{n \times p}$ has i.i.d. $N(0, 1)$ entries, with $p \geq 2$. For $0 < \alpha < 1$, let $u = \log(2p^2/\alpha)$. With probability at least $1 - \alpha$,

$$\delta_{\text{pw}}(X) \leq C \left(\sqrt{u/n} + u/n \right).$$

When $u \leq n$, this reduces to a constant times $\sqrt{u/n}$.

Remark 1. Even exponentially many columns can have small normalized pairwise inner products, provided their exponential growth rate is sufficiently small relative to the requested incoherence tolerance. The diagonal terms also control their lengths.

Proof. For fixed j, k , the centered products $X_{ij}X_{ik} - I\{j = k\}$ are independent across rows, with uniformly bounded subexponential norm. Bernstein gives

$$\mathbb{P}\left(\left|n^{-1}\langle X_j, X_k \rangle - I\{j = k\}\right| > t\right) \leq 2e^{-cn \min(t^2, t)}.$$

A union bound over p^2 ordered pairs yields

$$\mathbb{P}(\delta_{\text{pw}}(X) > t) \leq 2p^2 e^{-cn \min(t^2, t)}.$$

Choosing $t = C(\sqrt{u/n} + u/n)$ with a sufficiently large absolute constant makes the right side at most α . The linear term accounts for very small failure probabilities or large $\log p/n$. $\square$

Theorem 1. If X has pairwise incoherence $\delta_{\text{pw}}(X) < \frac{1}{2k}$, and $|S| \leq k$, then X satisfies the RNS property.

A short coordinate proof is in Exercise 5; the later low-priority proof gives a complementary matrix argument.

By combining the above deterministic result with Proposition 2, we get the following recovery guarantee.

Corollary 1. Assume $X_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ with $p \geq 2$ and $1 \leq k \leq p$. Then BP recovers every k -sparse θ^* simultaneously with probability at least $1 - p^{-2}$ if

$$n \geq Ck^2 \log p.$$

where C is an absolute constant.

The remarkable aspect of the corollary is that we can recover θ^* even $p \gg n$, namely $X\theta^* = y$ highly underdetermined due to the blessing of high dimensions. However, it turns out that the requirement $n \gg k^2$ —also known as the sample complexity—is still suboptimal because pairwise incoherence is not the “right” notion. We need to view X holistically (as an operator)!

Mid priority Practice: Use an incoherence certificate

Exercise prompt. If $\delta_{\text{pw}}(X) = 0.04$, what positive integer sparsity levels are certified by $\delta_{\text{pw}} < 1/(2k)$? See Exercise 2 in Section 6 for the full statement and complete solution.

1.3 Proof of Theorem 1

Low priority Proof of Theorem 1

The proof relies on a simple matrix norm bound. **Fact:** For a symmetric $A \in \mathbb{R}^{d \times d}$, $\|A\|_{\text{op}} \leq d\|A\|_{\text{max}}$.

Proof of the fact. Recall $\|A\|_{\text{op}} = \max_{u: \|u\|_2=1} \|Au\|_2$. Let A_i be the i -th column of A . By the Cauchy–Schwarz inequality, for any vector u ,

$$\|Au\|_2^2 = \sum_{i=1}^d \langle A_i, u \rangle^2 \leq \sum_{i=1}^d \|A_i\|_2^2 \|u\|_2^2 = \|A\|_F^2 \|u\|_2^2 \leq d^2 \|A\|_{\text{max}}^2 \|u\|_2^2.$$

So maximizing over u yields $\|A\|_{\text{op}} \leq d\|A\|_{\text{max}}$. $\square$

Proof of Theorem 1. We take any nonzero $\Delta \in \text{Ker}(X)$, i.e., $X\Delta = 0$, $\Delta \neq 0$. We partition X such that

$$X = \begin{bmatrix} X_S & X_{S^c} \end{bmatrix}, \quad X\Delta = X_S\Delta_S + X_{S^c}\Delta_{S^c},$$

$n \times |S| \quad n \times (p - |S|)$

so we have

$$X_S\Delta_S = -X_{S^c}\Delta_{S^c} \implies X_S^\top X_S\Delta_S = -X_S^\top X_{S^c}\Delta_{S^c}.$$

From the assumption, we have

$$(1) \lambda_{\min}(\frac{1}{n}X_S^\top X_S) > 1/2,$$

$$(2) \|\frac{1}{n}X_S^\top X_{S^c}\|_{\max} < \frac{1}{2k},$$

since

$$\begin{aligned} \left\| \frac{1}{n}X_S^\top X_S - I_{S \times S} \right\|_{\text{op}} &\leq |S| \left\| \frac{1}{n}X_S^\top X_S - I_{S \times S} \right\|_{\max} \leq k\delta_{\text{pw}}(X), \\ \left\| \frac{1}{n}X_S^\top X_{S^c} \right\|_{\max} &\leq \left\| \frac{1}{n}X^\top X - I_{p \times p} \right\|_{\max} \leq \delta_{\text{pw}}(X). \end{aligned}$$

Thus, we have

$$\Delta_S = -\left(\frac{1}{n}X_S^\top X_S\right)^{-1} \left(\frac{1}{n}X_S^\top X_{S^c}\right) \Delta_{S^c}.$$

So,

$$\begin{aligned} \|\Delta_S\|_1 &\leq \sqrt{|S|}\|\Delta_S\|_2 \quad (\text{by Cauchy-Schwarz inequality.}) \\ &\leq \frac{\sqrt{|S|}}{\lambda_{\min}(\frac{1}{n}X_S^\top X_S)} \left\| \frac{1}{n}X_S^\top X_{S^c}\Delta_{S^c} \right\|_2 \\ &\leq \frac{|S|}{\lambda_{\min}(\frac{1}{n}X_S^\top X_S)} \left\| \frac{1}{n}X_S^\top X_{S^c}\Delta_{S^c} \right\|_{\infty} \\ &\leq \frac{|S|}{\lambda_{\min}(\frac{1}{n}X_S^\top X_S)} \left\| \frac{1}{n}X_S^\top X_{S^c} \right\|_{\max} \|\Delta_{S^c}\|_1 \quad (\text{by } |\langle u, v \rangle| \leq \|u\|_{\infty}\|v\|_1.) \\ &< 2k \frac{1}{2k} \|\Delta_{S^c}\|_1 \\ &= \|\Delta_{S^c}\|_1. \end{aligned}$$

So, we conclude that $\Delta \notin C(S)$ and BP recovers θ^* . □

2 Compressed sensing

Mid priority *Compressed sensing*

Definition 2 (Restricted isometry). *The restricted isometry constant δ_k of X is the smallest nonnegative number such that*

$$(1 - \delta_k)\|u\|_2^2 \leq \|Xu\|_2^2 \leq (1 + \delta_k)\|u\|_2^2,$$

holds for all k -sparse $u \in \mathbb{R}^p$.

If $\delta_k = 0$, $\|Xu\| = \|u\|$, for every k -sparse u ; this is an isometry on each coordinate subspace of dimension at most k . It does not imply global isometry unless $k = p$. In particular, unit-length columns already give $\delta_1 = 0$, even if two columns coincide. The above definition focuses on all k -sparse vectors and relaxes isometry to allow a tolerance δ_k .

The restricted isometry property (RIP) means that this constant is small. We use X itself in the RIP definition above. For an unscaled Gaussian design, apply RIP to $X/\sqrt{n}$ and rescale y by the same factor; the BP feasible set is unchanged.

Theorem 2. Assume $y = X\theta^*$.

(1) Suppose θ^* is k -sparse and $\delta_{2k} < 1$. Then θ^* is the unique solution to

$$\min_{\theta \in \mathbb{R}^p} \|\theta\|_0, \quad \text{s.t. } X\theta = y.$$

(2) Suppose θ^* is k -sparse and $\delta_{2k} < \sqrt{2} - 1$. Then θ^* is the unique solution to

$$\min_{\theta \in \mathbb{R}^p} \|\theta\|_1, \quad \text{s.t. } X\theta = y.$$

Remark 2 (Convex vs. nonconvex). (1) Recall $\|\theta\|_0 = \sum_{j=1}^p \mathbb{1}(\theta_j \neq 0)$. The count $\|\theta\|_0$ is nonconvex. Finding the sparsest solution is NP-hard in general¹; nonconvexity alone is not a proof of computational hardness.

(2) Basis pursuit is a linear program and has polynomial-time algorithms in the usual finite-precision model. Exercise 8 gives the reformulation.

Theorem 3. Let $0 < \delta < 1$, $1 \leq k \leq p$, and $X_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. With probability at least $1 - 2\exp(-c\delta^2 n)$, the normalized matrix $X/\sqrt{n}$ has restricted isometry constant at most δ , provided

$$n \geq C\delta^{-2}k \log(ep/k).$$

Here $c, C > 0$ are absolute constants. A net argument over all k -coordinate subspaces proves this result; see [1, Section 10.5].

The basis-pursuit recovery implication of Theorem 2 is proved in Candès (2008). Finally, we have a tight sample complexity bound in terms of both sparsity k and dimension p .

Mid priority Practice: Sparse injectivity from RIP

Exercise prompt. Prove that $\delta_{2k} < 1$ rules out two distinct k -sparse vectors with the same measurements. See Exercise 3 in Section 6 for the full statement and complete solution.

2.1 Algorithmic perspective*

Low priority Algorithmic perspective

Convexity is useful because every local minimizer is global. For basis pursuit, the linear-program formulation gives a concrete algorithmic guarantee, with the computational cost depending on dimension, encoding length and requested accuracy.

For sparsest recovery, searching over supports is a useful mental picture of the difficulty. In the problem

$$\min \|\theta\|_0, \quad \text{s.t. } X\theta = y,$$

a direct approach searches over candidate supports S . If $|S| = k$, we need to consider $\binom{p}{k}$ possible choices of S , which is bad if p and k are large. It is connected to the NP-hardness in complexity theory.

¹NP-hard is a worst-case complexity classification: a polynomial-time algorithm for every instance would also solve a broad class of combinatorial search problems efficiently. It does not say that every dataset is difficult; the structural assumptions in this lecture identify cases where a convex method succeeds.

Convex relaxation generally refers to solving a convex surrogate problem. For example, BP replaces $\|\theta\|_0$ with $\|\theta\|_1$. A takeaway from convex relaxation is that, in suitable statistical settings, solving a convex problem is the same as a nonconvex one with high probability.

3 Statistical theory for LASSO

High priority *Statistical theory for LASSO*

So far, we have studied two problems related to LASSO.

- (i) Orthogonal design: LASSO = coordinatewise shrinkage.
- (ii) Noiseless data: as the positive penalty tends to zero, LASSO approaches basis-pursuit solutions.

In this section, we provide a theoretical analysis of the general LASSO. Assume the linear model $y = X\theta^* + \varepsilon$ with a k -sparse θ^* . We solve the convex optimization problem

$$\min_{\theta \in \mathbb{R}^p} \underbrace{\frac{1}{2n} \|y - X\theta\|_2^2}_{L_n(\theta)} + \underbrace{\lambda_n \|\theta\|_1}_{p_n(\theta)} =: \ell_n(\theta). \quad (1)$$

Write $\lambda = \lambda_n$ in the following arguments. The solution is not necessarily unique. LASSO generalizes denoising and BP in two ways.

- We use the ℓ_1 norm with a regularization parameter λ_n to encourage sparse solutions and reduce noise. In comparison, when $p > n$, ordinary least squares has no unique coefficient solution, and a minimum-norm choice does not by itself exploit coordinate sparsity.
- Instead of the linear constraint in BP, we use a quadratic loss $L_n(\theta)$ to accommodate noisy measurements.

Why the noiseless limit is basis pursuit. Let θ_{BP} be any basis-pursuit minimizer for feasible data y . For $\lambda > 0$, comparison with that exact fit gives

$$\frac{\|y - X\hat{\theta}_\lambda\|_2^2}{2n} + \lambda \|\hat{\theta}_\lambda\|_1 \leq \lambda \|\theta_{\text{BP}}\|_1.$$

Consequently $\|\hat{\theta}_\lambda\|_1 \leq \|\theta_{\text{BP}}\|_1$ and the residual tends to zero as $\lambda \downarrow 0$. The bounded estimates have convergent subsequences; every subsequential limit is feasible and has minimum ℓ_1 norm. If BP is unique, the whole family converges to it. For a fixed positive λ , LASSO generally need not interpolate exactly.

The basic LASSO inequality. Optimality compares the fitted objective with the objective at the true signal. A noise bound then puts the error in a cone, where restricted curvature can control its Euclidean norm.

3.1 Geometric intuition

High priority *Geometric intuition*

Due to the noise ε and a continuous parameter space $\mathbb{R}^p$, we can't expect the solution $\hat{\theta}$ to LASSO to be exactly θ^* , so we want to have guarantees on the estimation error $\|\hat{\theta} - \theta^*\|_2$.

Difficulty. Note that the quadratic loss in (1) is strongly convex if and only if X has full column rank. If $p < n$, we have seen in earlier lectures that $X^\top X$ is a key quantity in the Fisher information of the linear model. However, if $p > n$, the Hessian $\nabla_\theta^2 L_n(\theta) = X^\top X/n \in \mathbb{R}^{p \times p}$ must be degenerate because

$$\text{rank}(\nabla_\theta^2 L_n(\theta)) \leq n < p. \quad (2)$$

We have seen that this singularity is the root of many issues such as non-identifiability, numerical instability, singular Fisher information, etc.

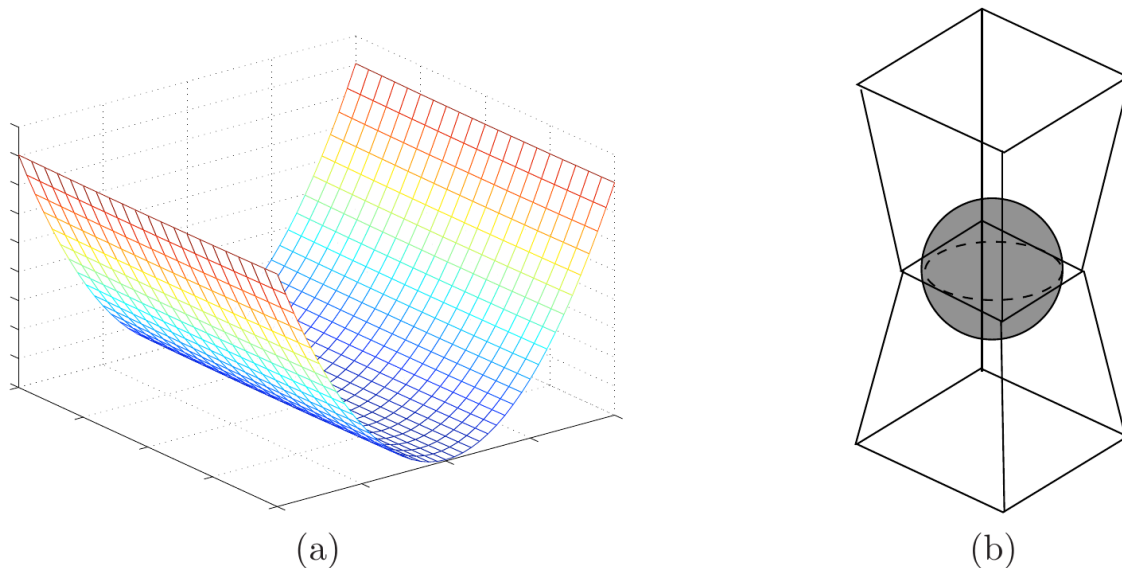


Figure 3: Left: a high-dimensional loss can be flat in some directions. Right: only the error directions allowed by the regularizer need curvature; the shaded portion shows such directions at a fixed radius. Figure: Negahban, Ravikumar, Wainwright and Yu, Figure 3 (2010 draft).

Key insights. Our strategy is to show that the Hessian is non-degenerate when restricted to the allowed error directions in $C_3(S) \subset \mathbb{R}^p$ (defined below) instead of the entire space $\mathbb{R}^p$. Sparsity restricts the error directions, so we do not need curvature along every vector in $\mathbb{R}^p$.

As before, let the support set be $S := \text{supp}(\theta^*) = \{j \leq p : \theta_j^* \neq 0\}$, and denote $\Delta = \theta - \theta^*$. The question is when $\ell_n(\theta^* + \Delta) \leq \ell_n(\theta^*)$? The full objective, not its penalty alone, determines a descent direction. The basic inequality shows that a minimizing error lies in the enlarged set $C_3(S)$ when λ dominates the noise correlation. Restricted curvature is required on that set.

Instructor's Note *Only the relevant directions need curvature*

A flat valley is a problem only if an estimation error can travel along it. Sparsity and the penalty constrain the possible error directions. Restricted curvature asks the loss to resist motion along those directions, even though it may remain flat elsewhere.

3.2 Estimation error of LASSO

High priority *Estimation error of LASSO*

The BP feasible-error set is generally not a cone. For LASSO we instead control error directions through the homogeneous set below; it is closed under nonnegative scaling but need not be convex. Denote

$$C_r(S) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq r\|\Delta_S\|_1\}.$$

Definition 3 (Restricted eigenvalue condition). *We say X satisfies the restricted eigenvalue (RE) condition over S with parameter $\kappa > 0$ if*

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2, \quad \text{for any } \Delta \in C_3(S).$$

This is the restricted-curvature convention of [2, Definition 7.12]: κ bounds the normalized quadratic form itself, rather than its square root. We state a deterministic result assuming the RE condition.

Theorem 4. *Assume that X satisfies κ -restricted eigenvalue condition, and*

$$\lambda > 0, \quad \lambda \geq 2 \left\| \frac{X^T \varepsilon}{n} \right\|_\infty.$$

Denote any LASSO solution by $\hat{\theta}$ and $\Delta = \hat{\theta} - \theta^*$. Then

$$(i) \quad \Delta \in C_3(S).$$

$$(ii) \quad \|\Delta\|_2 \leq \frac{3\lambda\sqrt{|S|}}{\kappa}.$$

The complete argument is in Exercise 6; the optional proof subsection expands the same steps. For independent Gaussian noise $\varepsilon \sim N(0, \sigma^2 I_n)$, conditional on a fixed design with $\max_j \|X_j\|_2 \leq L\sqrt{n}$, a Gaussian tail and union bound give

$$\mathbb{P}\left(\|X^T \varepsilon / n\|_\infty > \sigma L \sqrt{2 \log(2p/\alpha)/n} \mid X\right) \leq \alpha.$$

Choose λ twice this radius. For an independent isotropic Gaussian design, restricted-eigenvalue theory gives a constant lower bound on κ with high probability once $n \geq Ck \log(ep/k)$, and the column bound holds in the corresponding sample regime. Therefore

$$\|\hat{\theta} - \theta^*\|_2 \leq C\sigma \sqrt{k \log(2p/\alpha)/n}$$

on the joint good event. Independence, noise scale and sample size are part of the guarantee; Gaussian design alone is not enough.

Error requires scale and curvature. The LASSO error bound is proportional to $\lambda\sqrt{|S|}/\kappa$. The tuning parameter must dominate the noise correlation, and the restricted-eigenvalue constant must be positive.

High priority *Practice: Read a LASSO error guarantee*

Exercise prompt. If $|S| = 4$, $\kappa = 0.5$, $\lambda = 0.1$, and the noise correlation is at most 0.04, what error bound does the theorem provide?

See Exercise 4 in Section 6 for the full statement and complete solution.

3.3 Proof of Theorem 4

Low priority *Proof of Theorem 4*

Proof of Theorem 4.

$$\begin{aligned}\ell_n(\hat{\theta}) \leq \ell_n(\theta^*) &\Leftrightarrow \frac{1}{2n} \|X\Delta - \varepsilon\|_2^2 + \lambda \|\theta^* + \Delta\|_1 \leq \frac{1}{2n} \|\varepsilon\|_2^2 + \lambda \|\theta^*\|_1 \\ &\Leftrightarrow \frac{1}{2n} \|X\Delta\|_2^2 \leq \frac{1}{n} \langle X\Delta, \varepsilon \rangle + \lambda \|\theta^*\|_1 - \lambda \|\theta^* + \Delta\|_1\end{aligned}$$

By definition, $\theta_{S^c}^* = 0$, so $\|\theta^*\|_1 = \|\theta_S^*\|_1$. Using triangle inequality, and that

$$|\langle X\Delta, \varepsilon \rangle| = |\langle \Delta, X^T \varepsilon \rangle| \leq \|X^T \varepsilon\|_\infty \|\Delta\|_1,$$

we have

$$\frac{1}{2n} \|X\Delta\|_2^2 \leq \frac{1}{n} \|X^T \varepsilon\|_\infty \|\Delta\|_1 - \lambda \|\Delta_{S^c}\|_1 + \lambda \|\Delta_S\|_1. \quad (3)$$

(i) Without using restricted eigenvalue condition,

$$0 \leq \frac{1}{2n} \|X\Delta\|_2^2 \leq \frac{\lambda}{2} \|\Delta\|_1 - \lambda \|\Delta_{S^c}\|_1 + \lambda \|\Delta_S\|_1,$$

which implies that

$$0 \leq \frac{1}{2} (\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) - \|\Delta_{S^c}\|_1 + \|\Delta_S\|_1,$$

thus

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1.$$

(ii) In Equation (3), use

$$\frac{1}{2n} \|X\Delta\|_2^2 \geq \frac{\kappa}{2} \|\Delta\|^2,$$

by κ -restricted eigenvalues,

$$\begin{aligned}\frac{\kappa}{2} \|\Delta\|^2 &\leq \frac{\lambda}{2} \|\Delta\|_1 - \lambda \|\Delta_{S^c}\|_1 + \lambda \|\Delta_S\|_1 \leq \frac{3\lambda}{2} \|\Delta_S\|_1 \\ \Rightarrow \kappa \|\Delta\|_2^2 &\leq 3\lambda \|\Delta_S\|_1 \leq 3\lambda \sqrt{|S|} \|\Delta_S\|_2 \leq 3\lambda \sqrt{|S|} \|\Delta\|_2 \\ \Rightarrow \|\Delta\|_2 &\leq \frac{3\lambda \sqrt{|S|}}{\kappa}.\end{aligned}$$

□

4 Further reading

Further reading *Further reading*

The main reference for sparse linear regression is [2]. For the geometric recovery viewpoint, see Vershynin, Section 10.5. For restricted curvature and regularized estimation, see Negahban et al., Sections 2–4.

Key ideas

- Exact sparse recovery requires the null space to avoid directions that do not increase the norm.
- Pairwise incoherence gives a convenient sufficient condition; restricted isometry can use fewer measurements.
- For noisy LASSO, the penalty constrains the error directions and restricted curvature bounds their length.
- The tuning parameter must be calibrated to the noise correlations under the stated design normalization.

5 Detailed reference

High priority *Restricted null space*

If no nonzero null vector satisfies $\|h_{S^c}\|_1 \leq \|h_S\|_1$, then a signal supported on S is the unique basis-pursuit solution. This is a restricted condition, not full column rank. Return to the earlier explanation.

High priority *The basic LASSO inequality*

Optimality compares the fitted objective with the objective at the true signal. A noise bound then puts the error in a cone, where restricted curvature can control its Euclidean norm. Return to the earlier explanation.

High priority *Error requires scale and curvature*

The LASSO error bound is proportional to $\lambda\sqrt{|S|}/\kappa$. The tuning parameter must dominate the noise correlation, and the restricted-eigenvalue constant must be positive. Return to the earlier explanation.

Looking ahead

This completes the optional shrinkage and sparsity sequence. The planned Part VII: Testing and confidence sets turns from estimation to decisions and uncertainty; its lecture materials are not yet available.

6 Exercises and complete solutions

High priority *Basis pursuit with one measurement*

Exercise 1 (Basis pursuit with one measurement).

For $X = \begin{pmatrix} 1 & 2 \end{pmatrix}$ and $y = 2$, find the unique basis-pursuit solution.

Solution.

The constraint is $b_1 + 2b_2 = 2$. Since $2 = |b_1 + 2b_2| \leq |b_1| + 2|b_2| \leq 2(|b_1| + |b_2|)$, the objective is at least 1. Equality in the second inequality forces $b_1 = 0$, then the constraint gives $b_2 = 1$. Thus $(0, 1)$ is the unique minimizer.

□

Mid priority *Use an incoherence certificate*

Exercise 2 (Use an incoherence certificate).

If $\delta_{\text{pw}}(X) = 0.04$, what positive integer sparsity levels are certified by $\delta_{\text{pw}} < 1/(2k)$?

Solution.

The inequality is $0.08k < 1$, or $k < 12.5$. Hence every integer k from 1 through 12 is certified. The condition is sufficient: lack of this certificate for $k = 13$ does not prove recovery fails. $\square$

Mid priority *Sparse injectivity from RIP*

Exercise 3 (Sparse injectivity from RIP).

Prove that $\delta_{2k} < 1$ rules out two distinct k -sparse vectors with the same measurements.

Solution.

If $Xu = Xv$ and u, v are k -sparse, then $h = u - v$ has support size at most $2k$. RIP gives $0 = \|Xh\|^2 \geq (1 - \delta_{2k})\|h\|^2$, so $h = 0$. Any solution of the sparsest-representation problem has support size no larger than the true k -sparse target, proving its uniqueness. $\square$

High priority *Read a LASSO error guarantee*

Exercise 4 (Read a LASSO error guarantee).

If $|S| = 4$, $\kappa = 0.5$, $\lambda = 0.1$, and the noise correlation is at most 0.04, what error bound does the theorem provide?

Solution.

We have $0.1 \geq 2(0.04)$, so the penalty condition holds. The Euclidean bound is $3(0.1)\sqrt{4}/0.5 = 1.2$. It is an upper guarantee, not a claim that the actual error equals 1.2. $\square$

Mid priority *A complete incoherence-to-RNS argument*

Exercise 5 (A complete incoherence-to-RNS argument).

Prove RNS for $|S| \leq k$ when $\delta = \|X^\top X/n - I\|_{\max} < 1/(2k)$ using a coordinate argument.

Solution.

Let $G = X^\top X/n$ and $Xh = 0$. Then $h = (I - G)h$, giving $|h_j| \leq \delta\|h\|_1$ for every j . Thus $\|h_S\|_1 \leq k\delta(\|h_S\|_1 + \|h_{S^c}\|_1)$, or $\|h_S\|_1 \leq [k\delta/(1 - k\delta)]\|h_{S^c}\|_1$. The factor is strictly less than one. If the outside norm were zero, the same inequality would force $h = 0$. Every nonzero null vector therefore lies outside $C(S)$. $\square$

High priority *The basic inequality and restricted curvature*

Exercise 6 (The basic inequality and restricted curvature).

Under the assumptions of Theorem 4, prove both the cone inclusion and the Euclidean error bound.

Solution.

Optimality and expansion of squared loss give

$$\frac{\|X\Delta\|^2}{2n} \leq \frac{\varepsilon^\top X\Delta}{n} + \lambda(\|\theta^*\|_1 - \|\theta^* + \Delta\|_1).$$

The noise bound and triangle inequality give at most $\frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1$. Since the left side is nonnegative, $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. Apply RE to obtain $\kappa\|\Delta\|_2^2 \leq 3\lambda\|\Delta_S\|_1 \leq 3\lambda\sqrt{|S|}\|\Delta\|_2$. The zero-error case is immediate; otherwise divide by its norm.

□

High priority *Calibrate lambda conditional on the design*

Exercise 7 (Calibrate lambda conditional on the design).

Derive the displayed conditional Gaussian noise-correlation bound when $\|X_j\| \leq L\sqrt{n}$.

Solution.

Conditional on X , each $X_j^\top \varepsilon / n$ is mean-zero Gaussian with variance at most $\sigma^2 L^2 / n$. A union bound gives $2p \exp[-nt^2 / (2\sigma^2 L^2)]$ for the maximum exceeding t ; independence between coordinates is not required. Insert $t = \sigma L \sqrt{2 \log(2p/\alpha)/n}$.

□

Mid priority *Basis pursuit as a linear program*

Exercise 8 (Basis pursuit as a linear program).

Express BP using variables $u, v \geq 0$ with $\theta = u - v$, and justify equivalence.

Solution.

Minimize $\sum_j (u_j + v_j)$ subject to $X(u - v) = y$ and nonnegative u, v . Every θ has a feasible representation by its positive and negative parts with objective $\|\theta\|_1$. Conversely every such representation has objective at least $\|u - v\|_1$. Hence the two optimal values and recovered θ minimizers agree. This explains convex computational tractability without implying uniqueness.

□

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 16, 2026: Clarified the basis-pursuit limit of LASSO, used the enlarged error set consistently, and separated computational hardness from nonconvexity alone. Added optional-part guidance, a backward recap and concise reading summaries. October 17, 2026: Published in optional Part VI with lecture links, student downloads and issue reporting; this material is not examined.

References

- [1] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [2] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.