

Lecture 20: Basis pursuit, compressed sensing, and LASSO

Lecturer: Yiqiao Zhong

Last update: 18:46:23, 2024-11-26

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

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

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 with ℓ_0 norm) 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 is unique (why?), and the feasible set (i.e., parameters satisfying linear constraint) is

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

which is a hyperplane.

1.1 Restricted null space property

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

Let us consider the level set $\mathcal{S}_r = \{\theta : \|\theta\|_1 = r\}$ for a given r . The level set $\mathcal{S}_t$ is a collection of simplices. It is easy to see (Figure 1)

$$\begin{cases} \mathcal{S}_r \cap (\theta^* + \text{Ker}(X)) = \emptyset & \implies & \|\theta^*\|_1 > r, \\ \mathcal{S}_r \cap (\theta^* + \text{Ker}(X)) \text{ contains one point} & \implies & \|\theta^*\|_1 = r \\ \mathcal{S}_r \cap (\theta^* + \text{Ker}(X)) \text{ contains more than one points} & \implies & \|\theta^*\|_1 < r \end{cases}$$

The third case is due to the strict convexity of ℓ_1 norm: if two points θ_1, θ_2 are in the intersection, then $(\theta_1 + \theta_2)/2$ is in the feasible set with a smaller ℓ_1 norm.

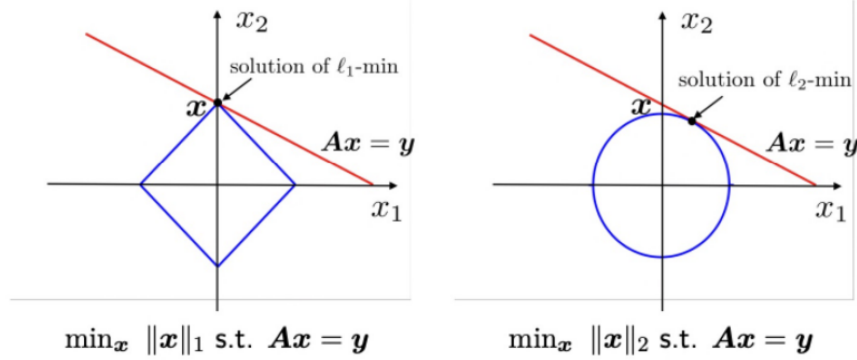
Blessing of high dimensions. The ℓ_1 ball in p dimensions, namely

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

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 hyperplane the ℓ_1 ball touches exactly at θ^* , i.e.,

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

What is the failure probability for a uniformly random hyperplane of dimension $p - 1$ crossing θ^* ?



- Level sets of $\|\cdot\|_1$ are pointed, enabling it to promote sparsity
- Level sets of $\|\cdot\|_2$ are smooth, often leading to dense solutions

Figure 1: Solution to BP is the intersection point between the feasible set and the ℓ_1 ball. Note that the ℓ_1 ball encourages a sparse solution whereas the ℓ_2 ball does not.

- In 2D, basis pursuit successfully recovers a given sparse signal θ^* if the hyperplane $\{\theta \in \mathbb{R}^p : X\theta = y\}$ intersects with the blue square only at θ^* (Figure 2).
- In 3D, the successful directions form a smaller region. Because a corner of ℓ^1 -ball in $\mathbb{R}^3$ is adjacent to 8 cubes, the failure recovery probability is $\frac{1}{2^2} = \frac{1}{4}$.
- In $\mathbb{R}^p$, the failure probability is $\frac{1}{2^{p-1}}$, which very small if p large!

RNS: formalizing geometric intuitions. Denote the support set of the true signal as $S = \{j \leq p : \theta_j^* \neq 0\}$. Define a cone as

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

where Δ_S means $(\Delta_i \mathbb{1}(j \in S))_{j \leq p}$. 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 the converse is true, i.e., we can find another θ in the feasible set with a smaller ℓ_1 norm. Then, there exists $\Delta \neq 0$, such that

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

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, $C(S)$ is the set of “descent directions” where the objective function $\|\theta\|_1$ further decreases. Thus, we can say that BP recovers θ^* iff no possible (nonzero) descent direction is in the feasible set.

Probability of success of ℓ_1 minimization

Suppose A is randomly generated

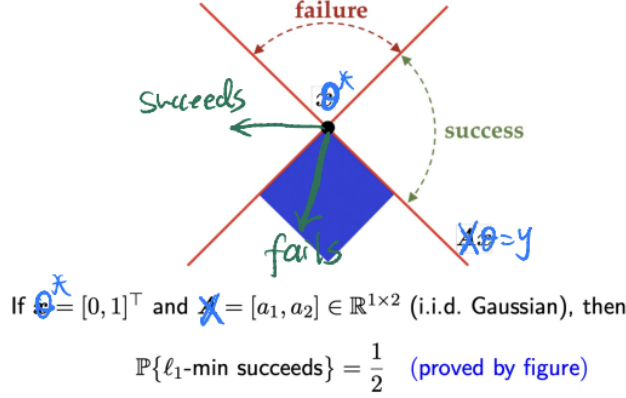


Figure 2: Illustration of when BP is successful at recovering a sparse θ^* in 2D. The blue region indicates the cone $C(S)$ where the objective value decreases.

1.2 Pairwise incoherence: a key condition on 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

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

If $n = p$, then $\delta_{pw}(X)$ simply measures the difference between the empirical covariance matrix to the identity matrix. A remark fact is that we can have a small $\delta_{pw}(X)$ even when $p \gg n$.

Proposition 2 (Exponentially many random vectors have small PI). Suppose $X = (X_{kj})_{k \leq n, j \leq p}$ where $X_{kj} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\delta_{pw}(X) \leq C \frac{\sqrt{\log p} + \sqrt{\log(1/\delta)}}{\sqrt{n}}.$$

where C is an absolute constant.

Remark 1. This result also implies that given exponentially many random normal vectors, all pairs of vectors are nearly orthogonal (*exercise*).

Proof. By Lecture 4, the inner product $\langle X_k, X_j \rangle$ is a sum of independent subexponential random variables. So by Bernstein's inequality,

$$\mathbb{P}\left(\frac{1}{n} |\langle X_k, X_j \rangle - 1| \geq t\right) \leq 2e^{-Cn \min(t, t^2)}, \quad \text{for any } t > 0.$$

By the union bound (over the choice of k, j),

$$\delta_{pw}(X) \leq 2 \exp(-Cn \min(t, t^2) + 2 \log p), \quad \text{for any } t > 0.$$

We can choose a sufficiently large constant C and

$$t = \min \left\{ 1, C \frac{\sqrt{\log p} + \sqrt{\log(1/\delta)}}{\sqrt{n}} \right\}$$

to ensure $\delta_{\text{pw}}(X) \leq \delta$. □

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

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

Corollary 1. *Assume $X_{kj} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and θ^* is a k -sparse. Then, BP recovers θ^* with high probability 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 under-determined 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)!

1.3 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 ,

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

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

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

$$X = \begin{bmatrix} X_S & X_{S^c} \\ n \times |S| & n \times (p - |S|) \end{bmatrix}, \quad X\Delta = X_S\Delta_S + X_{S^c}\Delta_{S^c},$$

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

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

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

since

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

Thus, we have

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

So,

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

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

2 Compressed Sensing

Definition 2. (*Compressed Sensing*) *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 $\delta_k = 0$, $\|Xu\| = \|u\|$, which implies that the singular values of X are all 1. In this case, X as a linear operator preserves Euclidean distances (isometry). The above definition focuses on all k -sparse vectors and relaxes isometry to allow a tolerance δ_k .

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

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

Remark 2 (Convex vx. nonconvex). (1) *Recall $\|\theta\|_0 = \sum_{j=1}^p \mathbb{1}(\theta_j \neq 0)$. The ℓ_0 norm $\|\theta\|_0$ is nonconvex in θ , so this problem is computationally hard in high dimensions.*

(2) *Basis pursuit is a convex problem, which is easy to solve in general. It is in fact a linear programming.*

Theorem 3. *If $X_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then w.h.p. $\frac{1}{\sqrt{n}}X$ satisfies $\delta_k < \delta$ as long as*

$$n \geq C \frac{1}{\delta^2} k \log \frac{p}{k}.$$

where C is a constant.

Finally, we have a tight sample complexity bound in terms of both sparsity k and dimension k .

2.1 Algorithmic perspective*

In general, a convex optimization problem is considered easy to solve. Usually, we expect there exists a polynomial algorithm (e.g., interior point method) to find a solution.

In contrast, a nonconvex problem is usually hard to solve without additional structures. In the problem

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

in the worst case, we need a brute-force search of 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 the complexity theory.

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

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

- (i) Orthogonal design: LASSO = coordinate-wise shrinkage.
- (ii) Noiseless data: LASSO limit = basis pursuit.

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

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

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, the OLS estimator has poor accuracy when $p \gg n$ because of noise in every coordinate.
- Instead of the linear constraint in BP, we use a quadratic loss $L_n(\theta)$ to accommodate noisy measurements.

3.1 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 problem (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, worse Cramér-Rao lower bound, etc.

Key insights. Our strategy is to show that the Hessian is non-degenerate when restricted to approximate descent directions $C(S) \subset \mathbb{R}^p$ instead of the entire space $\mathbb{R}^p$. We have seen earlier that $C(S)$ occupies a small fraction of region 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 $\ell_n(\theta^* + \Delta) \leq \ell_n(\theta^*)$? We hope to show that whenever $\|\theta^* + \Delta\|_1 < \|\theta^*\|_1$ (thus decreases the objective function ℓ_n), the quadratic function $L_n(\theta)$ behaves like a strongly convex function.

3.2 Estimation Error of LASSO

We need to slightly modify the cone of descent directions from BP. Denote

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

Definition 3 (7.12 in [1]). *We say X satisfies restricted eigenvalue (RE) condition over S with parameters κ if*

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

We state a deterministic result assuming the RE condition.

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

$$\lambda \geq 2 \left\| \frac{X^T \epsilon}{n} \right\|_\infty.$$

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

- (i) $\Delta \in C_3(S)$.
- (ii) $\|\Delta\|_2 \leq \frac{3\lambda\sqrt{|S|}}{\kappa}$.

If X is isotropic Gaussian design, i.e. i.i.d. $\mathcal{N}(0, 1)$ entries, then by the subgaussian concentration inequality,

$$\left\| \frac{X^T \epsilon}{n} \right\|_\infty \leq C \sqrt{\frac{\log p}{n}} \quad \text{w.h.p.}$$

Further, more technical analysis can show that κ is bounded below by a constant. So,

$$\|\Delta\|_2 \leq C \sqrt{\frac{k \log p}{n}} \quad \text{w.h.p.}$$

for k -sparse signal θ^* . This estimation error bound is similar to soft-thresholding since both have the factor $\sqrt{k \log p}$.

3.3 Proof of Theorem 4 *

Proof of Theorem 4.

$$\begin{aligned} \ell_n(\hat{\theta}) \leq \ell_n(\theta^*) &\Leftrightarrow \tilde{\ell}_n(\Delta) \leq \tilde{\ell}_n(\mathbf{0}) \\ &\Leftrightarrow \frac{1}{2n}\|X\Delta - \epsilon\|_2^2 + \lambda\|\theta^* + \Delta\|_1 \leq \frac{1}{2n}\|\epsilon\|_2^2 + \lambda\|\theta^*\|_1 \\ &\Leftrightarrow \frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{1}{n}\langle X\Delta, \epsilon \rangle + \lambda\|\theta^*\|_1 - \lambda\|\theta^* + \Delta\|_1 \end{aligned}$$

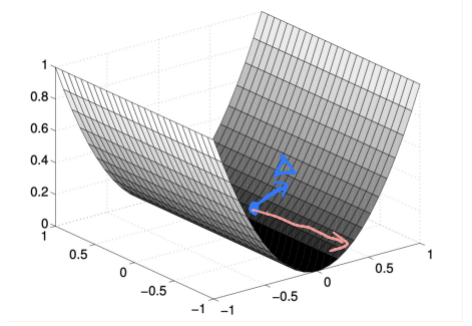


Figure 3: The function $L_n(\theta)$ in high dimensions. There are degenerate directions (red). If we can show that the important directions (blue) are not aligned with the degenerate directions, then we can control the error.

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

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

we have

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

(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

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

by K -restricted eigenvalues,

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

□

4 Further reading

The main reference of sparse linear regression is the book [1].

References

- [1] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.