Bounds on l_2-error for hard sparse models

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

1. Basis pursuit problem: $\hat{\theta} = \arg\min \|\theta\|_1$, s.t. $y = X\theta^*$, where $X \in \mathbb{R}^{n \times d}$ and θ^* is s-sparse.
2. Restricted Nullspace Property: $\{\text{null}(X) \cap \{\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}\} = 0$
3. Pairwise Incoherence (PI): $\left\| \frac{X'X}{n} - I \right\|_2 \leq \delta$
4. Restricted Isometry Property (RIP): $\left\| \left\| \frac{X_S' X_S}{n} - I_S \right\| \right\|_2 \leq \delta_s(X)$, where $\delta_s(X) > 0$ and $|S| \leq s$
5. If PI holds with $\delta_s(X) < \frac{1}{3s} \rightarrow$ RN holds
6. If RIP holds with $\delta_{2s}(X) < \frac{1}{3} \rightarrow$ RN holds

Our general aim is to solve

$$\hat{\theta} = \arg\min \|\theta\|_0 \quad s.t. \quad y = X\theta$$

which is hard to compute, and hence why we replace the troublesome L_0-norm by L_1-norm to obtain $\hat{\theta}$. For $\|\hat{\theta}\|_0 \leq \|\theta^*\|_0 \leq s$, $\|\Delta\|_0 \leq 2s$, where $\Delta := \hat{\theta} - \theta^*$. In order to argue that θ^* is unique , it suffices to show that $\|X\Delta\|_2^2 > 0$ for any at most $2s$-sparse vector Δ. We will now show the RIP in fact implies this condition. Observe that

$$\left\| \frac{X\Delta}{\sqrt{n}} \right\|_2^2 = \frac{\Delta' X' X \Delta}{n} = \Delta' \left(\frac{X'X}{n} - I \right) \Delta + \|\Delta\|_2^2 \geq -\delta \|\Delta\|_2^2 + \|\Delta\|_2^2 > 0$$

hold when

$$\left\| \frac{X_S' X_S}{n} - I_S \right\|_2 < 1, \quad \forall |S| \leq 2s.$$

This shows that an RIP condition is sufficient for the success of the L_0 minimization program.

Let us now turn our attention to the noisy setting, in which we observe the vector matrix pair $(y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$ linked by the linear model

$$y = X\theta^* + \varepsilon$$

where $\varepsilon \in \mathbb{R}^n$ is the noise vector. The aim is to answer the following questions:

(a) Can we ensure $\|\hat{\theta} - \theta^*\|_2^2$ is small?

(b) What conditions on X is needed to ensure (a)?

(c) What choice of (n, d, s) do we need to ensure (a)?

A natural extension of the basis pursuit program is based on minimizing a weighted combination of the data-fidelity term $\|y - X\theta\|_2^2$ with the L_1-norm penalty, say of the form

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}$$

where $\lambda_n > 0$ is a regularization parameter to be chosen by the user. What we just described, following Tibshirani (1996) is referred to as the LASSO program.
The LASSO problem can be formulated in constrained form as follows

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \quad s.t. \quad \|\theta\|_1 \leq R$$

for some radius $R > 0$, or

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad s.t. \quad \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2$$

for some tolerance $b > 0$, where the latter form of the constrained problem is referred to as the relaxed basis pursuit by .
Looking at the regularized version of the LASSO, by the basic inequality:

$$\frac{1}{2n}\|y - X\hat{\theta}\|_2^2 + \lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_n\|\theta^*\|_1.$$

Setting $\Delta := \hat{\theta} - \theta^*$, and noting that $Y - X\hat{\theta} = \varepsilon - X\Delta$, we can expand the first term in the above inequality to obtain

$$\frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle \varepsilon, X\Delta\rangle + \lambda_n\|\hat{\theta}\|_1 \leq \lambda_n\|\theta^*\|_1.$$

The goal is then to bound $\|\Delta\|_2^2$. In low dimension, a bound on $\|X\Delta\|_2^2$ would provide guarantees for $\|\Delta\|_2^2$. This is however not true in high dimension, as X has a non-trivial null space.

- Perfect recovery is no longer feasible in noisy settings.
- We thus focus on bounding the L_2-error $\|\hat{\theta} - \theta^*\|_2$.
- In noisy setting, the required condition is slightly stronger than the Restricted Nullspace Property - namely that the restricted eigenvalues of the matrix $\frac{X'X}{n}$ are lower bounded over a cone. In particular, for a constant $\alpha \geq 1$, let us define the set

$$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}$$

- Notice that the above set is a special case of our definition in the RNS property, which corresponds to the special case of $\alpha = 1$.

Definition (Restricted Eigenvalue condition)

The matrix X satisfies the Restricted Eigenvalue condition over S with parameters (κ, α), if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2, \quad \forall \Delta \in \mathbb{C}_\alpha(S)$$

The Restricted Eigenvalue condition (RE hereafter) strengthens the Restricted Nullspace Property. In particular, if the RE condition holds with parameters $(\kappa, 1)$ for any $\kappa > 0$, then the RNS property holds. In what follows, we will prove that the error $\|\hat{\theta} - \theta^*\|_2$ in the LASSO condition is well controlled.

In the optimisation problem

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \quad \text{s.t. } \|\theta\|_1 \leq R$$

let us consider the radius $R = \|\theta^*\|_1$. With this setting, the true parameter vector θ^* is feasible for the problem. By definition, the Lasso estimate $\hat{\theta}$ minimises the quadratic cost function

$$L_n(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2$$

over the l_1-ball of radius R. As $n \to \infty$, it is expected that θ^* becomes a near-minimiser of the same cost function, so that

$$L_n(\hat{\theta}) \approx L_n(\theta).$$

But when does closeness in cost imply that the error vector $\Delta := \hat{\theta} - \theta^*$ is small? Let us first take a look at the following image

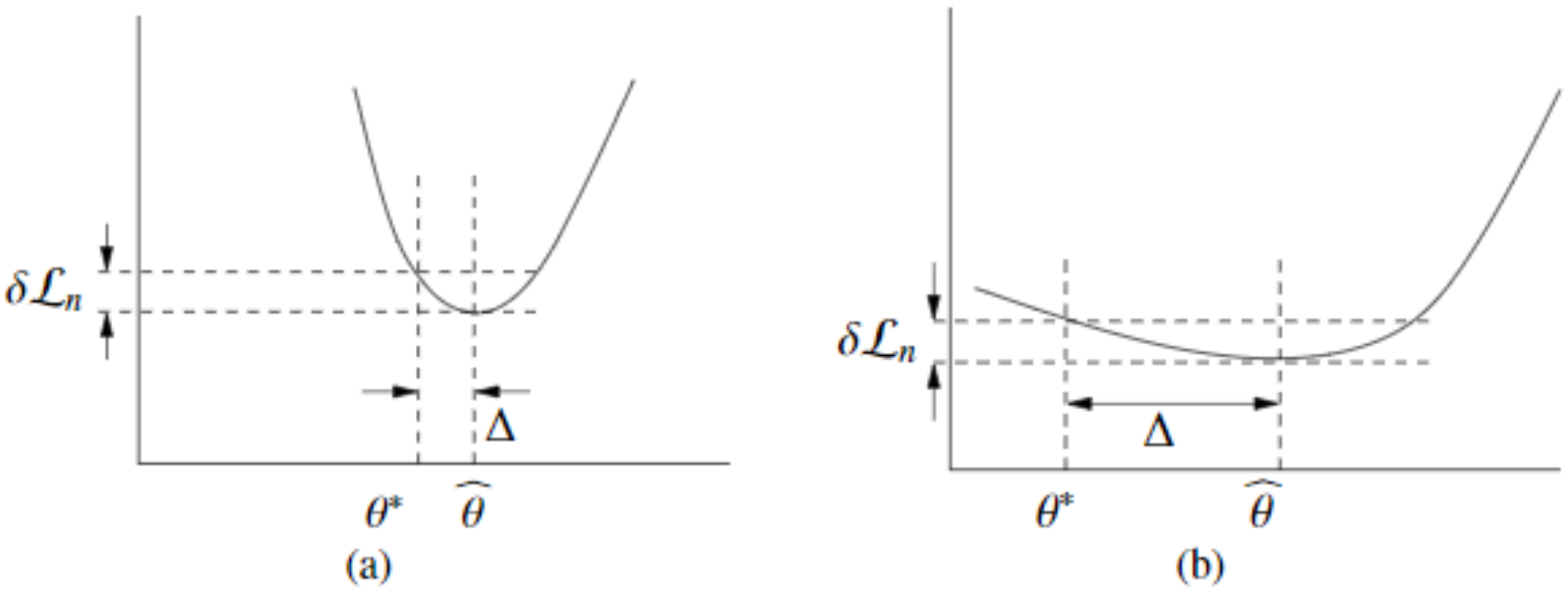


Figure 7.5 Illustration of the connection between curvature (strong convexity) of the cost function, and estimation error. (a) In a favorable setting, the cost function is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ in the cost implies that the error vector $\Delta = \widehat{\theta} - \theta^*$ is not too large. (b) In an unfavorable setting, the cost is very flat, so that a small cost difference $\delta\mathcal{L}_n$ need not imply small error.

- There is a link between the cost difference of the one-dimensional function $\delta L_n := L_n(\theta^*) - L_n(\hat{\theta})$ and the error $\Delta = \hat{\theta} - \theta^*$ is controlled by the curvature of the cost function.
- For a function in d dimensions, the curvature of a cost function is captured by the structure of its Hessian matrix $\nabla^2 L_n(\theta)$, which is symmetric positive semidefinite matrix.
- Notice that in the special case of the quadratic cost function that underlies the LASSO

$$\nabla L_n(\theta) = \frac{1}{n}(\theta X'X - X'y)$$

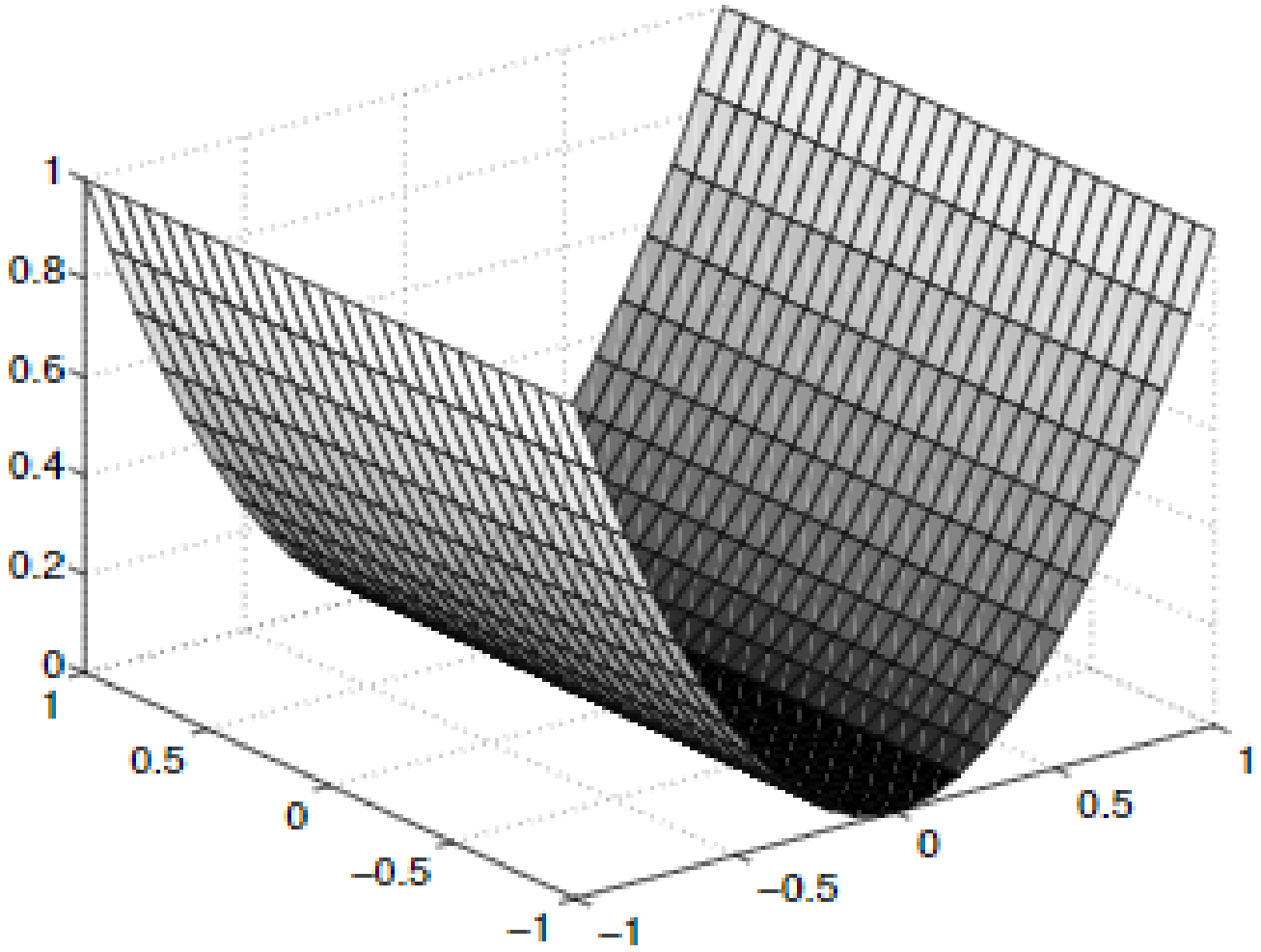
and the Hessian is calculated as

$$\nabla^2 L_n(\theta) = \frac{1}{n}X'X$$

- If we could guarantee that the eigenvalues of this matrix were uniformly bounded away from zero, say that

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 > 0, \quad \forall \Delta \in \mathbb{R}^d \backslash \{0\}$$

  then we would be assured of having curvature in all directions.
- When $d > n$ and the Hessian is a $d \times d$ with rank at most n, it is impossible to guarantee that it has positive curvature in all directions.
- The quadratic cost function always has the below form


1
0.8
0.6
0.4
0.2
0
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

- Although, it is curved in some directions, there is always $(d - n)$-dimensional subspace of directions in which it is completely flat, and consequently the uniform lower bound above is never satisfied.
- Thus, it is necessary to relax the stringency of the uniform curvature condition and require that it holds only for a subset $\mathbb{C}_\alpha(S)$ of vectors.
- If we can be assured that the subset $\mathbb{C}_\alpha(S)$ is well aligned with the curved direction of the Hessian, then a small difference in the cost function will translate into bounds on the difference between $\hat{\theta} and \theta^*$

References

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.