Sparse Linear Models in High Dimensions

Guo Zerui

Sun Yat-sen University

May 30, 2022

Contents

Introduction

Suppose that we observe a vector $y \in \mathbb{R}^n$ and a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ that are linked via the standard linear model:

$$y = \mathbf{X}\theta^* + w \tag{1}$$

where $w \in \mathbb{R}^n$ is a vector of noise variables.

If $d > n$, it is necessary to impose additional structure on the unknown regression vector $\theta^* \in \mathbb{R}^d$ for obtaining consistent estimators, and this chapter focuses on different types of sparse models.

Introduction

Example (Basis function and regularization) We know that the so-called linear model is linear in the parameters not in the input variables, so we can augment the simplest linear model with nonlinear basis functions as:

$$\begin{aligned} h(\mathbf{x}, \mathbf{w}) &= w_0 \phi_0(\mathbf{x}) + \cdots + w_{M-1} \phi_{M-1}(\mathbf{x}) \\ &= \sum_{m=0}^{M-1} w_m \phi_m(\mathbf{x}) \\ &= \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) \end{aligned}$$

By using nonlinear basis functions it is possible for h to adapt to nonlinear relationships of x, we call these models linear basis function models.

Introduction

We can see that increasing the number of basis functions makes a better model, until we start overfitting.

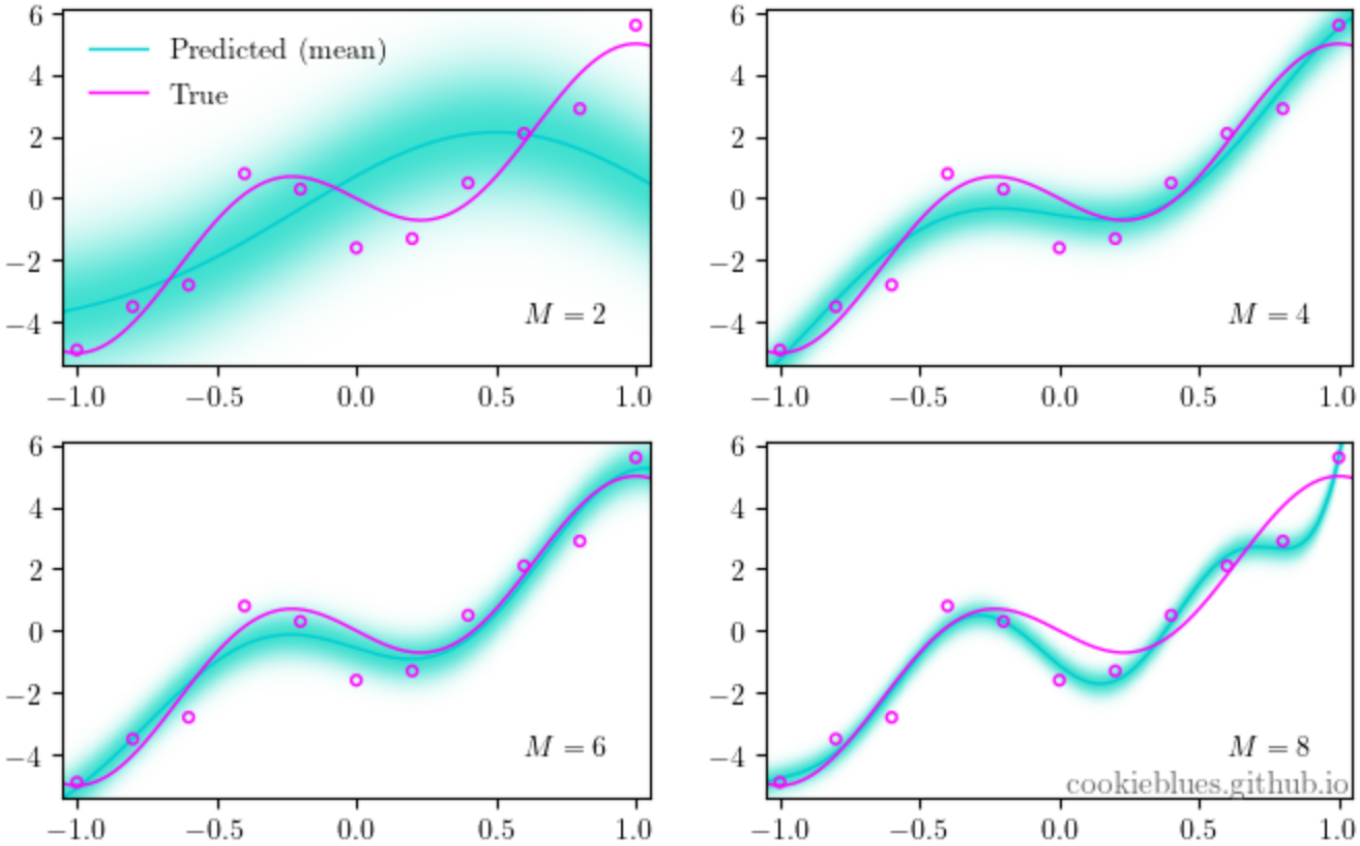


Figure: Illustration of the effect of using Gaussian basis functions

Introduction

One of the simplest kinds of structure in a linear model is a hard sparsity assumption, meaning that the set:

$$S(\theta^*) := \{j \in \{1,2,\ldots,d\} \mid \theta_j^* \neq 0\} \tag{2}$$

known as the support set of θ^*, has cardinality $s := |S(\theta^*)|$ substantially smaller than d.

Assuming that the model is exactly supported on s coefficients may be overly restrictive, in which case it is also useful to consider various relaxations of hard sparsity. There are different ways in which to formalize such an idea, and we will focus on ℓ_1-*norm*.

Recovery in the Noiseless Setting

We begin by focusing on the simplest case in which the observations are noiseless. More concretely, we wish to find a solution θ to the linear system $y = \mathbf{X}\theta$, where $y \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ are given. When $d > n$, this is an underdetermined set of linear equations, and we would like to find a sparse solution with $s \ll d$ non-zero entries. This noiseless problem has applications in signal representation and compression.

Recovery in the Noiseless Setting

This problem can be cast as a non-convex optimization problem involving the ℓ_0–*norm*:

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad \text{such that } \mathbf{X}\theta = y \tag{3}$$

If we could solve this problem, then we would obtain a solution to the linear equations that has the fewest number of non-zero entries.

Given the computational difficulties associated with ℓ_0-minimization, a natural strategy is to replace the troublesome ℓ_0-objective by the nearest convex member of the ℓ_q-family, namely the ℓ_1-norm:

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that } \mathbf{X}\theta = y \tag{4}$$

Recovery in the Noiseless Setting

We now turn to an interesting theoretical question: when is solving the ℓ_1– problem (4) equivalent to solving the original ℓ_0– problem (3)?

We would like to introduce a property first, where
$\mathbb{C}(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$ and $\text{null}(\mathbf{X}) := \{\Delta \in \mathbb{R}^d \mid \mathbf{X}\Delta = 0\}$.

Definition 1

The matrix $\mathbf{X}$ *satisfies the restricted nullspace property with respect to* S *if* $\mathbb{C}(S) \cap \text{null}(\mathbf{X}) = \{0\}$

Recovery in the Noiseless Setting

Now we can prove that the restricted nullspace property is equivalent to the equivalence of the problem (4) and problem (3):

Theorem 1

1. *For any vector* $\theta^* \in \mathbb{R}^d$ *with support* S*, the problem* (4) *applied with* $y = \mathbf{X}\theta^*$ *has unique solution* $\widehat{\theta} = \theta^*$*;*
2. *The matrix* $\mathbf{X}$ *satisfies the restricted nullspace property with respect to* S*.*

(Proof)

Recovery in the Noisy Setting

Let us now turn to the noisy setting, a natural extension of the problem (4) is based on minimizing $\|y - \mathbf{X}\theta\|_2^2$ with the ℓ_1-norm penalty, say of the form

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\} \tag{5}$$

where $\lambda_n > 0$ is a regularization parameter to be chosen by the user, and we refer to it as the **Lasso** following Tibshirani (1996).

Recovery in the Noisy Setting

Alternatively, one can consider different constrained forms of the Lasso, that is either:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \right\} \quad \text{such that } \|\theta\|_1 \leq R \tag{6}$$

for some radius $R > 0$, or

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that } \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 \leq b^2 \tag{7}$$

for some noise tolerance $b > 0$. By Lagrangian duality theory, all three families of convex programs are equivalent.

Recovery in the Noisy Setting

In the noisy setting, we can no longer expect to achieve perfect recovery. Instead, we focus on bounding the ℓ_2-error $\left\|\widehat{\theta} - \theta^*\right\|_2$ between a Lasso solution $\widehat{\theta}$ and the unknown regression vector θ^*.
Before proving the consistency theorem, we would like to introduce a lemma and a condition for the following proof, which are the basic inequality and the restricted eigenvalue (RE) condition.

Recovery in the Noisy Setting

In the general linear model, we know that the prediction error:

$$\left\|\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right\|_2^2/\sigma^2 \tag{8}$$

is χ_p^2-distributed, where $\hat{\theta} := \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. And there is a closed form of $\mathbb{E}\left\|\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right\|_2^2/n$:

$$\frac{\mathbb{E}\left\|\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right\|_2^2}{n} = \frac{\sigma^2}{n}p \tag{9}$$

Recovery in the Noisy Setting

In the context of Lasso, we find an inequality for $\mathbb{E}\left\|\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right\|_2^2/n$ instead, which is called the basic inequality:

$$\left\|\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right\|_2^2/n+\lambda\|\hat{\theta}\|_1 \leq 2\varepsilon^T\mathbf{X}\left(\hat{\theta}-\theta_0\right)/n+\lambda\left\|\theta_0\right\|_1 \tag{10}$$

and the first part of right hand side can be bounded in terms of the ℓ_1-norm of the parameters involved:

$$2\left|\varepsilon^T\mathbf{X}\left(\hat{\theta}-\theta_0\right)\right| \leq \left(\max_{1\leq j\leq p} 2\left|\varepsilon^T\mathbf{X}^{(j)}\right|\right)\left\|\hat{\theta}-\theta_0\right\|_1 \tag{11}$$

Recovery in the Noisy Setting

Refer to the previous section, we require a condition that is closely related to but slightly stronger than the restricted nullspace property:

Definition 2

The matrix $\mathbf{X}$ *satisfies the restricted eigenvalue (RE) condition over* S *with parameters* (κ, α) *if*

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \quad \textit{for all } \Delta \in \mathbb{C}_\alpha(S).$$

where $\mathbb{C}_\alpha(S) := \left\{\Delta \in \mathbb{R}^d \middle| \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\right\}$, and we will prove that under the RE condition, the error $\left\|\widehat{\theta} - \theta^*\right\|_2$ in the Lasso solution is well controlled.

Recovery in the Noisy Setting

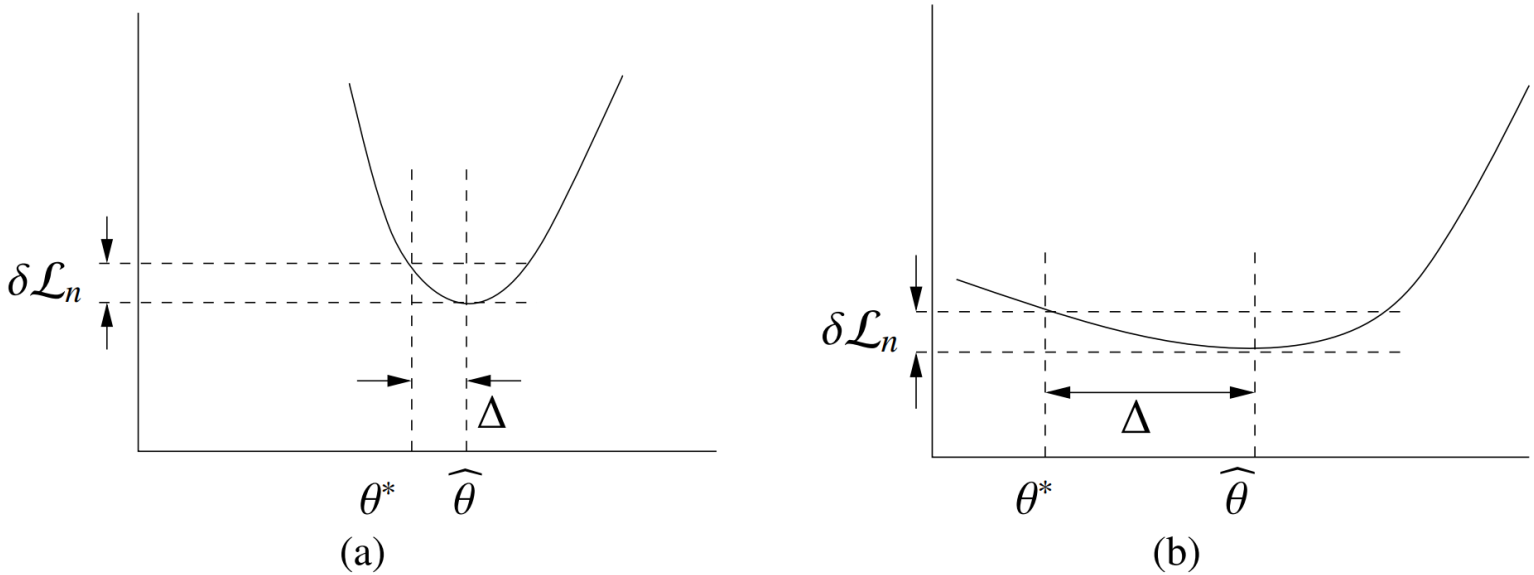


Figure 7.5 Illustration of the connection between curvature (strong convexity) of the cost function, and estimation error. (a) In a favorable setting, the cost function is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ 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.

Recovery in the Noisy Setting

For a function in d dimensions, the curvature of a cost function is captured by the structure of its Hessian matrix $\nabla^2 \mathscr{L}_n(\theta)$, which can be calculated as

$$\nabla^2 \mathscr{L}_n(\theta) = \frac{1}{n} \mathbf{X}^{\mathrm{T}} \mathbf{X} \tag{12}$$

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

$$\frac{\|\mathbf{X}\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 > 0 \quad \text{for all } \Delta \in \mathbb{R}^d \backslash \{0\} \tag{13}$$

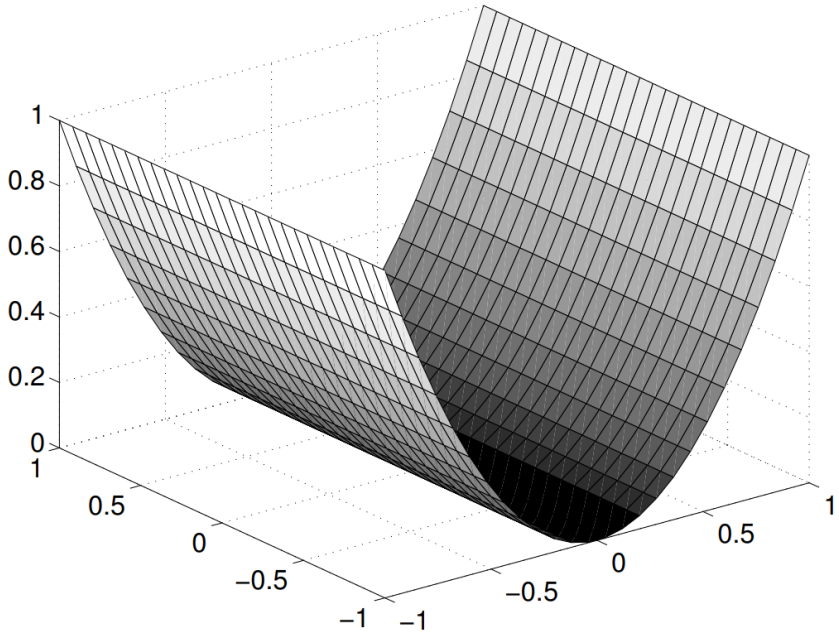
then we would be assured of having curvature in all directions.

Recovery in the Noisy Setting

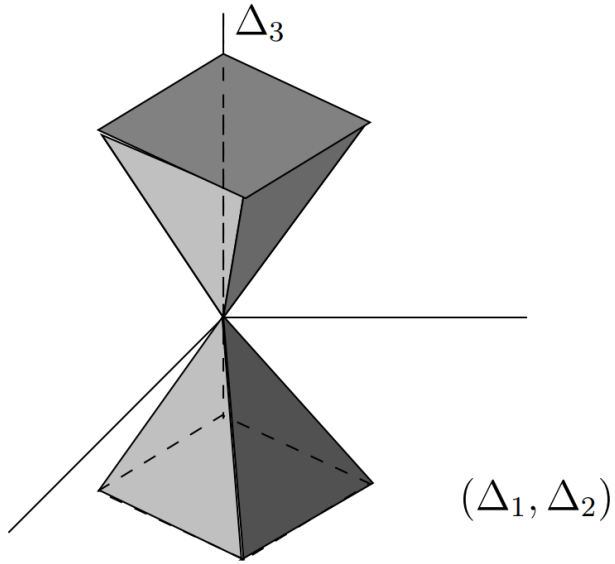
But in the high-dimensional setting, this Hessian is impossible to guarantee that it has a positive curvature in all directions, and the quadratic cost function may be curved in some directions, there is always a $(d-n)$-dimensional subspace of directions in which it is completely flat.

Consequently, the uniform lower bound (13) is never satisfied, and we need to relax the stringency of the uniform curvature condition, and require that it holds only for a subset $\mathbb{C}_\alpha(S)$ of vectors.

Recovery in the Noisy Setting



(a)



(b)

Figure 7.6 (a) A convex cost function in high-dimensional settings (with $d \gg n$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. (b) The Lasso error $\widehat{\Delta}$ must lie in the restricted subset $\mathbb{C}_\alpha(S)$ of $\mathbb{R}^d$. For this reason, it is only necessary that the cost function be curved in certain directions of space.

Recovery in the Noisy Setting

We now state a result that provides a bound on the error $\left\|\widehat{\theta} - \theta^*\right\|_2$. In particular, let us impose the following conditions:

1. The vector θ^* is supported on a subset $S \subseteq \{1, 2, \ldots, d\}$ with $|S| = s$;
2. The design matrix satisfies the restricted eigenvalue condition over S with parameters $(\kappa, 3)$.

Recovery in the Noisy Setting

Under assumptions above, we can prove the following theorem:

Theorem 2

1. *Any solution of the Lagrangian Lasso* (5) *with regularization parameter lower bounded as* $\lambda_n \geq 2\left\|\frac{\mathbf{X}^{\mathrm{T}}w}{n}\right\|_\infty$ *satisfies the bound:*

$$\left\|\widehat{\theta}-\theta^*\right\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n \tag{14}$$

2. *Any solution of the constrained Lasso* (7) *with* $R = \|\theta^*\|_1$ *satisfies the bound*

$$\left\|\widehat{\theta}-\theta^*\right\|_2 \leq \frac{4}{\kappa}\sqrt{s}\left\|\frac{\mathbf{X}^{\mathrm{T}}w}{n}\right\|_\infty \tag{15}$$

Recovery in the Noisy Setting

Theorem 3

3. *Any solution of the relaxed basis pursuit program* (8) *with* $b^2 \geq \frac{\|w\|_2^2}{2n}$ *satisfies the bound*

$$\left\|\widehat{\theta} - \theta^*\right\|_2 \leq \frac{4}{\kappa}\sqrt{s}\left\|\frac{\mathbf{X}^{\mathrm{T}} w}{n}\right\|_\infty + \frac{2}{\sqrt{\kappa}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}} \tag{16}$$

(Proof and Example)

Recovery in the Noisy Setting

If we assume the design matrix is orthonormal and $n > p$, then we have $X^T X = D$, which is a full rank and diagonal matrix, and the OLS solution $\hat{\beta}$ is uniquely defined. In this setting, we have a uniform closed form of lasso as:

$$\left(\hat{\beta}_j\right)_{\text{lasso}} = \text{sign}\left(\hat{\beta}_j\right)\left(\left|\hat{\beta}_j\right| - \frac{\lambda}{\sigma_j^2}\right)_+ \tag{17}$$

(Proof)