High-Dimensional Regression: Lasso

Advanced Topics in Statistical Learning, Spring 2024

Ryan Tibshirani

1 Introduction

In this lecture, we'll move on from low-dimensional nonparametric to high-dimensional parametric regression. Though this might seems like very different problems, as we'll see, they do share some similarities.

Below, we provide a quick recap of what we know about least squares and motivations for regularization (as also covered in the review lecture), laying the groundwork for the main estimators we'll study in this and the next lecture on high-dimensional regression: lasso and ridge.

1.1 Recap: least squares regression

Suppose we are given n observations of the form (x_i, y_i), $i = 1, \ldots, n$ where each $x_i \in \mathbb{R}^d$ denotes a feature vector and $y_i \in \mathbb{R}$ an associated response value. Let $X \in \mathbb{R}^{n \times d}$ denote the predictor matrix (whose i^{th} row is x_i) and $Y \in \mathbb{R}^n$ denote the response vector. Recall that the least squares regression coefficients of Y on X are given by solving

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2. \tag{1}$$

When $d \leq n$ and $\text{rank}(X) = d$, this produces the unique solution

$$\hat{\beta} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y.$$

The fitted values (i.e., in-sample predictions) are

$$X\hat{\beta} = X(X^\mathsf{T} X)^{-1} X^\mathsf{T} Y = P_X Y,$$

where $P_X = X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$ denotes the projection onto the column space of X.

Risk properties. Now let's recall the risk properties of least squares. Assume (x_i, y_i), $i = 1, \ldots, n$ are i.i.d. such that $X^\mathsf{T} X$ is almost surely invertible, and

$$y_i = x_i^\mathsf{T} \beta_0 + \epsilon_i, \quad i = 1, \ldots, n,$$

where ϵ_i has mean zero and variance σ^2, and $\epsilon_i \perp\!\!\!\perp x_i$. Of course, we can equivalently write this as

$$Y = X\beta_0 + \epsilon. \tag{2}$$

The in-sample prediction risk of least squares is

$$\frac{1}{n}\mathbb{E}\big[\|X\hat{\beta} - X\beta_0\|_2^2 \,|\, X\big] = \sigma^2 \frac{d}{n}. \tag{3}$$

Meanwhile, the out-of-sample prediction risk is

$$\mathbb{E}\big[(x_0^\mathsf{T}\hat{\beta} - x_0^\mathsf{T}\beta_0)^2\big] = \sigma^2 \,\text{tr}\left(\mathbb{E}[x_0 x_0^\mathsf{T}]\,\mathbb{E}[(X^\mathsf{T} X)^{-1}]\right) \approx \sigma^2 \frac{d}{n-d}, \tag{4}$$

where the expectation is taken over the training data (x_i, y_i), $i = 1, \ldots, n$ and an independent draw x_0 from the predictor distribution. The middle expression above is an equality in general. The rightmost expression is an approximation that holds as n, d grow large in a random matrix theory model, which we'll learn the details of when we study ridge regression. (Further, the rightmost expression is exact for Gaussian features.)

1.2 Recap: trouble in high dimensions

As we just saw in (3), (4), the risk of least squares regression degrades as d grows close to n—and looking at the rightmost expression in (4), the out-of-sample risk actually diverges at $d = n$.

Meanwhile, the least squares estimator itself is not even well-defined when $d > n$, in that the optimization problem (1) does not have a unique solution. In this case, any vector of the form

$$\hat{\beta} = (X^\mathsf{T} X)^+ X^\mathsf{T} Y + \eta, \quad \text{where } \eta \in \text{null}(X), \tag{5}$$

solves (1), where we write A^+ to denote the generalized inverse of a matrix A, and $\text{null}(A)$ to denote its null space.

If all we care about is out-of-sample prediction, then this is not the end of the story for least squares—it turns out that taking $\eta = 0$ in (5), which yields the *minimum ℓ_2 norm least squares solution*, can still have interesting predictive properties when $d > n$. We'll study this later, in the overparametrization lecture.

But if we additionally care about the estimated coefficients themselves, then it really is the end of the road for least squares. This is because, for any $\hat{\beta}$ of the form (5) with $\hat{\beta}_j > 0$ for some component j, we can always find[1] another $\tilde{\beta}$ of the form (5) with $\tilde{\beta}_j < 0$. So we cannot even consistently interpret the sign of any estimated coefficient (let alone its magnitude).

1.3 Recap: regularization to the rescue

Regularization will finesse the problems described above. At a high level, it gives us a way to produce nontrivial coefficient estimates, and it may well give us better predictions as well. (It typically does, and most people would have traditionally said that it pretty much *always* does. However, recent developments in overparametrization have shown us that there is more nuance to the prediction story, and whether or not explicit regularization helps depends very strongly on the operating characteristics of the prediction problem.)

In the least squares setting, traditional approaches for regularization take two forms:

$$\begin{aligned} &\text{Constrained form:} \quad \underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 \ \text{ subject to } \ \beta \in C \\ &\text{Penalized form:} \quad \underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + h(\beta). \end{aligned}$$

Here C is some (typically convex) set, and h is some (typically convex) penalty function. Typically $C = \{\beta : \|\beta\| \leq t\}$ is the sublevel set of a norm $\|\cdot\|$, and $h(\beta) = \lambda\|\beta\|$ is a nonnegative multiple of a norm. In this case, the constrained and penalized forms can be seen to equivalent, via convex duality: that is, for any $t \geq 0$ and solution $\hat{\beta}$ in the constrained problem, there is a value of $\lambda \geq 0$ such that $\hat{\beta}$ also solves the penalized problem, and vice versa.

We'll mostly focus on the penalized form in this lecture, since it is generally more commonly studied, but we'll also cover the constrained form when we work out prediction and estimation theory.

Canonical regularizers: ℓ_0, ℓ_1, and ℓ_2. In regression, arguably the three canonical choices for regularizers are the ℓ_0, ℓ_1, and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^d 1\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^d |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^d \beta_j^2\right)^{1/2}.$$

This gives rise to what we call best subset selection, the lasso, and ridge regression. In the current lecture we'll focus on the lasso, and in the next we'll focus on ridge regression.

[1] Technically, this is only true if $\text{null}(X) \not\perp e_j$, where e_j is the j^{th} standard basis vector. Note that the latter condition must hold for at least one j, as $\text{null}(X) = \{0\}$. And for random features, under very weak conditions, it will be true almost surely for any j.

Calling $\|\cdot\|_0$ the "ℓ_0 norm" is a misnomer, as it is not actually a norm: it does not satisfy positive homogeneity, i.e., $\|a\beta\|_0 = a\|\beta\|_0$ for all $a > 0$. (It would be more accurate to call it the "ℓ_0 pseudonorm", but in keeping with common convention, we'll simply use "norm".)

Critically, $\|\cdot\|_0$ is *not convex*, while $\|\cdot\|_1$ and $\|\cdot\|_2$ are convex (note any norm is a convex function). This makes best subset selection a nonconvex problem, and one that is generally very hard to solve in practice except for very small d. On the other hand, the lasso and ridge regression problems are convex, and many efficient algorithms exist for them.

2 Lasso basics

This section covers some basic properties of the *least absolute selection and shrinkage operator* or lasso, defined by

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{6}$$

for a tuning parameter $\lambda \geq 0$. (The factor of $\frac{1}{2}$ is just for convenience.) We start off by recalling a central property that the lasso has in common with best subset selection, which replaces $\|\cdot\|_1$ in (6) by $\|\cdot\|_0$: they produce *sparse* solutions, that is, at their solutions $\hat{\beta}$, we will have

$$\hat{\beta}_j = 0, \text{ for many } j.$$

Larger values of the tuning parameter λ typically means sparser solutions. Why care about sparsity? This is often desirable, for two reasons: (i) it corresponds to performing variable selection in the fitted linear model (providing a level of interpretability of what features may be important), and (ii) it can often predict better (in situations where the underlying regression function is well-approximated by a sparse linear model).

In constrast, the *ridge regression* estimator,

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{7}$$

has a generically dense solution (all nonzero components), for any $\lambda > 0$. The fact that sparsity arises in (6) but not (7) is often explained using the "classic" picture, in Figure 1 (which illustrates the problems in constrained form). We can also see a clear difference in how they behave in the case of an orthogonal predictor matrix X (meaning $X^\mathsf{T}X = I$); in this case, the solutions in (6), (7) are

$$\begin{aligned}\hat{\beta}^{\text{lasso}} &= S_\lambda(X^\mathsf{T}Y),\\ \hat{\beta}^{\text{ridge}} &= (X^\mathsf{T}Y)/(1+\lambda),\end{aligned}$$

respectively, where S_λ is soft-thresholding operator at the level λ applied componentwise, which is defined as $S_\lambda(a) = \text{sign}(a)(|a| - \lambda)_+$.

Another difference between the lasso (6) and ridge (7) problems is that the latter is always strictly convex, whereas the former is not strictly convex when $d > n$. This means that we are always guaranteed a unique ridge solution, but not necessarily a unique lasso solution.

Some basic convex analysis, as developed over the next few subsections, will help us understand this better (and reveal that there is often not much to worry about here). Our treament follows Tibshirani (2013).

2.1 Sign interconsistency

First, we make a basic observation: although the lasso solution is not always unique, the lasso fit $X\hat{\beta}$ is always unique. This is true because the least squares loss $f(u) = \|Y - u\|_2^2$ is strictly convex in u. (Note that the same is true of the least squares fit: it is always unique, even when the solution is not.)

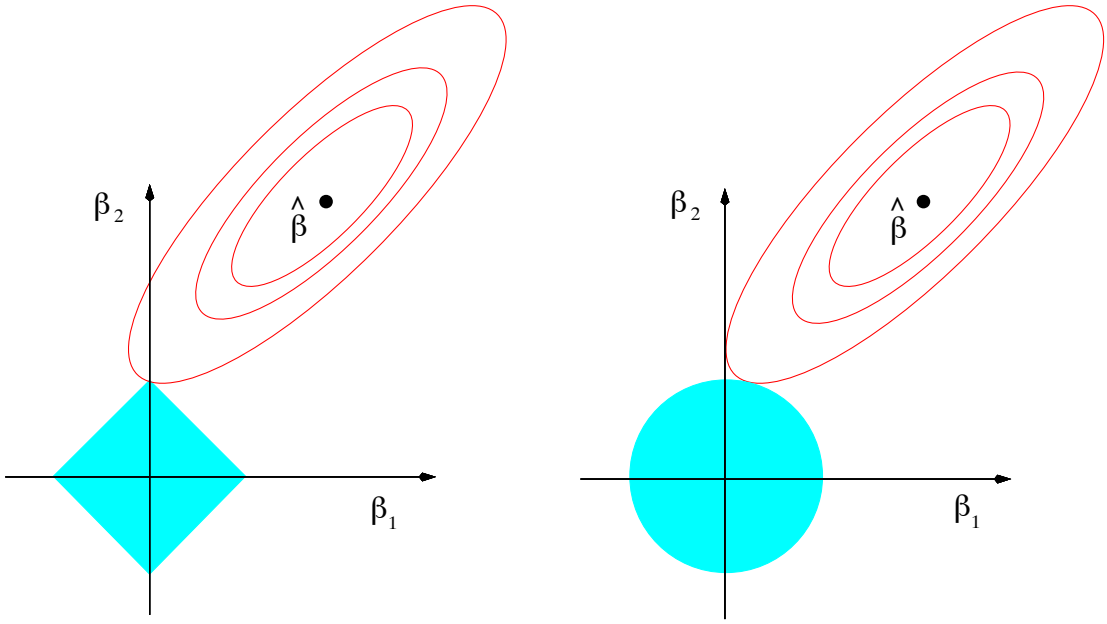

Figure 1: *The "classic" illustration comparing lasso and ridge constraints. Credit: Chapter 3.4 of Hastie et al. (2009).*

Next, consider the subgradient optimality condition (sometimes called the KKT condition) for the lasso problem (6), which is

$$X^T(Y - X\hat{\beta}) = \lambda s, \tag{8}$$

where $s \in \partial\|\hat{\beta}\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$. Precisely,

$$s_j \in \begin{cases} \{+1\} & \hat{\beta}_j > 0 \\ \{-1\} & \hat{\beta}_j < 0 \\ [-1, 1] & \hat{\beta}_j = 0, \end{cases}, \quad j = 1, \ldots, d. \tag{9}$$

From (8), (9) we can read off a straightforward but important fact: the optimal subgradient s is always unique, for any $\lambda > 0$. This is because it is determined by the unique fitted value $X\hat{\beta}$.

Why is this important? It tells us that even in the case when lasso solutions are nonunique, *any two solutions must always agree on the signs of common nonzero coefficients*. That is, we cannot have two solutions $\hat{\beta}, \tilde{\beta}$ such that $\hat{\beta}_j > 0$ but $\tilde{\beta}_j < 0$. In this sense, the lasso is already much better behaved than least squares when $d > n$. The next subsection shows that a much stronger statement can be made about the lasso solution itself, when $d > n$.

2.2 Structure of solutions

It is not hard to prove that the lasso solution is "often" unique even when $d > n$. But first, we'll need to do a little bit of work to to learn about the structure of lasso solutions (which should be interesting in its own right). Define the *equicorrelation set*

$$E = \big\{j \in \{1, \ldots, d\} \,:\, |X_j^T(Y - X\hat{\beta})| = \lambda\big\}.$$

This is the set of variables that achieves the maximum absolute inner product (correlation, for standardized predictors) with the lasso residual vector $Y - X\hat{\beta}$. Assuming $\lambda > 0$, this is the same as

$$E = \big\{j \in \{1, \ldots, d\} \,:\, |s_j| = 1\big\}.$$

The equicorrelation set E is always unique (as $X\hat{\beta}, s$ are unique). Note that the set E contains the support set—also called the *active set*, and denoted $A = \text{supp}(\hat{\beta})$ of any lasso solution $\hat{\beta}$, because for $j \notin E$, we have $|s_j| < 1$, which implies that $\hat{\beta}_j = 0$.

Thus we can write $X\hat{\beta} = X_E\hat{\beta}_E$ for any lasso solution $\hat{\beta}$, where $\hat{\beta}_E$ denotes the components of $\hat{\beta}$ indexed by E, and X_E denotes the columns of X indexed by E. The subgradient condition (8) implies

$$X_E^\mathsf{T}(Y - X_E\hat{\beta}_E) = \lambda s_E,$$

and solving this for $\hat{\beta}_E$ gives

$$\hat{\beta}_E = (X_E^\mathsf{T}X_E)^+(X_E^\mathsf{T}Y - \lambda s_E) + \eta, \quad \text{where } \eta \in \text{null}(X_E).$$

From this we learn the following sufficient condition for uniqueness: *if the equicorrelated predictors are linearly independent,* $\text{rank}(X_E) = |E|$*, then the lasso solution is unique and given by*

$$\begin{aligned} \hat{\beta}_E &= (X_E^\mathsf{T}X_E)^{-1}(X_E^\mathsf{T}Y - \lambda s_E), \\ \hat{\beta}_{-E} &= 0. \end{aligned} \tag{10}$$

(Here $\hat{\beta}_{-E}$ denotes the components of $\hat{\beta}$ indexed by $E^c = \{1, \ldots, d\} \setminus E$.) Interestingly, we can see above that this is a certain *shrunken* least squares estimator on the active set E.

2.3 Uniqueness and saturation

The previous subsection established that when $\text{rank}(X_E) = |E|$, then the lasso solution is unique. A short calculation will show us that when the columns of X are in what is known as *general position*, a very weak condition, then it must be the case that $\text{rank}(X_E) = |E|$, and so the lasso solution is unique. We state and prove this next.

Proposition 1. *Assume that X has columns $X_1, \ldots, X_d \in \mathbb{R}^n$ that are in general position. This means that for any $k < \min\{n, d\}$, indices $i_1, \ldots, i_{k+1} \in \{1, \ldots, d\}$, and signs $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, +1\}$, the affine span of $\sigma_1 X_{i_1}, \ldots, \sigma_{k+1} X_{i_{k+1}}$ does not contain any element of $\{\pm X_i : i \neq i_1, \ldots, i_{k+1}\}$. Then for any $Y \in \mathbb{R}^n$ and $\lambda > 0$, the lasso problem (6) has a unique solution.*

Proof. We will prove the contrapositive. We assume the lasso solution is not unique, thus $\text{rank}(X_E) < |E|$, and will prove that this means the columns of X are not in general position. Since $\text{rank}(X_E) < |E|$, there exists some $i \in E$ such that

$$X_i = \sum_{j \in E \setminus \{i\}} c_j X_j.$$

Hence

$$s_i X_i = \sum_{j \in E \setminus \{i\}} (s_i s_j c_j) \cdot (s_j X_j).$$

By definition of the equicorrelation set, $X_j^\mathsf{T} r = s_j \lambda$ for any $j \in E$, where $r = Y - X\hat{\beta}$ is the lasso residual vector. Taking the inner product of both sides above with r, we get

$$\lambda = \sum_{j \in E \setminus \{i\}} (s_i s_j c_j)\lambda,$$

and since $\lambda > 0$,

$$\sum_{j \in E \setminus \{i\}} (s_i s_j c_j) = 1.$$

Therefore, denoting $a_j = s_i s_j c_j$ for $j \in E \setminus \{i\}$, we have shown that

$$s_i X_i = \sum_{j \in E \setminus \{i\}} c_j X_j, a_j \cdot s_j X_j, \quad \text{and} \quad \sum_{j \in E \setminus \{i\}} a_j = 1,$$

which means that $s_i X_i$ lies in the affine span of $s_j X_j$, $j \in E \setminus \{i\}$. Note that we can assume without a loss of generality that $E \setminus \{i\}$ has at most n elements, since otherwise we can simply repeat the above arguments replacing E by any one of its subsets with $n+1$ elements. This means that the columns of X cannot be in general position, which completes the proof. □

The lasso problem is an intriguing example where we get a unique solution without strict convexity. To emphasize just how weak the general position assumption, we note that if the entries of X have a density on $\mathbb{R}^{nd}$, then it is not hard to show that X has columns in general position almost surely. To emphasize, this gives us the following corollary.

Corollary 1. *Assume that the entries of X are drawn from some continuous distribution on $\mathbb{R}^{nd}$. Then almost surely, for any $Y \in \mathbb{R}^n$ and $\lambda > 0$, the lasso problem (6) has a unique solution.*

Here is another intriguing property of the lasso.

Proposition 2. *For any $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$, and $\lambda > 0$, there exists a solution in the lasso problem (6) whose active set A has at most $\min\{n, d\}$ elements. In particular, when the lasso is unique, this means it must have at most $\min\{n, d\}$ nonzero components.*

This property is often called *saturation* of the lasso solution. It can be shown using Carathéodory's theorem. Note that it is not necessarily a good thing—when we have, say, $d = 10000$ variables and $n = 10$ observations, then any lasso solution, when unique, can only have at most 10 nonzero coefficients. This is often used as a point of motivation for the *elastic net*, which is an estimator that is based on combining the lasso and ridge penalties into one criterion.

There is a lot more that can be said about the lasso, including some interesting geometry that underlies it, which reveals important properties about its local stability and effective degrees of freedom. There are also a number of interesting recent developments surrounding lasso inference, in high-dimensional, post-selection, and other settings. Unfortunately we can't cover all of this (without budgeting more time), and in the rest of the lecture we'll instead cover some of the most foundational estimation theory for the lasso. We refer to Hastie et al. (2015) for a nice treatment of some topics we skipped (including "close cousins" of the lasso, which are crafted to have similar properties in related but often different problem settings). It is also a good reference for the theory we cover below, as is Buhlmann and van de Geer (2011).

3 "Slow" rates

In this section, we develop what are sometimes called "slow" rates for the lasso. You'll see a strong parallel to the way we analyzed nonparametric estimators in the empirical process theory lecture, and developing a basic inequality will play a leading role. That said, in terms of the probabilistic tools that are needed, the analysis here will be much simpler.

We assume the linear model (2), where the noise vector $\epsilon \in \mathbb{R}^n$ has i.i.d. sub-Gaussian entries with mean zero and variance proxy σ^2. We will take $X \in \mathbb{R}^{n \times d}$ to be fixed (or equivalently, we condition on it), and assume that each $\max_{j=1,\ldots,d} \|X_j\|_2 \leq \sqrt{n}$. (Note that we can always rescale to make this true.)

3.1 Penalized form

First, we analyze the lasso in penalized form (6). As in the empirical process theory lecture, we start by deriving a basic inequality. Let $\hat{\beta}$ denote any solution in (6). For any coefficient vector $\beta \in \mathbb{R}^d$,

$$\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Rearranging,

$$\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 - \frac{1}{2}\|Y - X\beta\|_2^2 \leq \lambda(\|\beta\|_1 - \|\hat{\beta}\|_1).$$

Adding and subtracting $X\beta$ in the leftmost term, and expanding, we get

$$\|X\hat{\beta} - X\beta\|_2^2 \leq 2\langle Y - X\beta, X\hat{\beta} - X\beta\rangle + \lambda(\|\beta\|_1 - \|\hat{\beta}\|_1).$$

where we have moved the inner product term to the right-hand side. This is true for any vector β. Taking $\beta = \beta_0$ in particular, the true coefficient vector from (2), and recognizing $Y - X\beta_0 = \epsilon$, the noise vector, we get from the last display our *basic inequality* for $\hat{\beta}$,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X\hat{\beta} - X\beta_0 \rangle + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1). \tag{11}$$

A result on the in-sample prediction risk for the lasso is only a few lines away. Observe that

$$\begin{aligned} \langle \epsilon, X\hat{\beta} - X\beta_0 \rangle &= \langle X^\mathsf{T}\epsilon, \hat{\beta} - \beta_0 \rangle \\ &\leq \|X^\mathsf{T}\epsilon\|_\infty \|\hat{\beta} - \beta_0\|_1 \end{aligned}$$

by Hölder's inequality. Thus from (11), we learn that

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\|X^\mathsf{T}\epsilon\|_\infty \|\hat{\beta} - \beta_0\|_1 + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1), \tag{12}$$

and using the triangle inequality,

$$\begin{aligned} \|X\hat{\beta} - X\beta_0\|_2^2 &\leq 2\|X^\mathsf{T}\epsilon\|_\infty (\|\hat{\beta}\|_1 + \|\beta_0\|_1) + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\ &\leq 2\lambda\|\beta_0\|_1. \end{aligned} \tag{13}$$

where the second line holds if we take $\lambda \geq 2\|X^\mathsf{T}\epsilon\|_\infty$. So far this has all been deterministic. Next comes the probabilistic argument. Note that $X^\mathsf{T}\epsilon$ has sub-Gaussian entries with mean zero and variance proxy $\max_{j=1,\ldots,d} \|X_j\|_2^2 \sigma^2 \leq n\sigma^2$. By a result on the maximum of sub-Gaussian random variables (stated in the empirical process theory, and you'll prove it on the homework),

$$\mathbb{P}\Big(\|X^\mathsf{T}\epsilon\|_\infty \geq \sigma\sqrt{2n(\log(2d) + u)}\Big) \leq e^{-u},$$

for any $u > 0$. Therefore, from (13), taking $\lambda = 2\sigma\sqrt{2n(\log(2d) + u)}$, we get

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1 \sqrt{\frac{2(\log(2d) + u)}{n}}, \tag{14}$$

with probability at least $1 - e^{-u}$. This bound yields what is called the "slow" rate for the penalized lasso estimator: the in-sample prediction risk scales as $\|\beta_0\|_1\sqrt{(\log d)/n}$.

3.2 Constrained form

To analyze the lasso in constrained form,

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 \ \ \text{subject to} \ \ \|\beta\|_1 \leq t, \tag{15}$$

where now $t \geq 0$ is our tuning parameter, the arguments are even easier. Take $t = \|\beta_0\|_1$, so that the true coefficient vector is feasible for (15); then following the same steps as in the previous subsection,

$$\begin{aligned} \frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 &\leq \frac{2}{n}\langle \epsilon, X\hat{\beta} - X\beta_0 \rangle \\ &\leq \frac{2}{n}\|X^\mathsf{T}\epsilon\|_\infty (\|\hat{\beta}\|_1 + \|\beta_0\|_1) \\ &\leq 4\sigma\|\beta_0\|_1 \sqrt{\frac{2(\log(2d) + u)}{n}}, \end{aligned} \tag{16}$$

where the last line holds with probability at least $1 - e^{-u}$. Note that the bound in (16) for the constrained estimator is exactly the same as that in (14) for the penalized one.

3.3 Oracle inequality

If we don't want to assume a linear model (2) for the mean, then we can still derive an interesting bound on the in-sample prediction risk, that characterizes its risk in excess of the best linear predictor. We will demonstrate this in the constrained case because the argument is simplest. Assume that

$$Y = f_0(X) + \epsilon,$$

for some function $f_0 : \mathbb{R}^d \to \mathbb{R}$, where we abbreviate $f_0(X) = (f_0(x_1), \ldots, f_0(x_n)) \in \mathbb{R}^n$ for the rowwise application of f_0 to X. Then as in the last subsection, for any solution $\hat{\beta}$ in the constrained lasso problem (15), and any coefficient vector $\bar{\beta}$ with $\|\bar{\beta}\|_1 \leq t$,

$$\begin{aligned}\|X\hat{\beta} - X\bar{\beta}\|_2^2 &\leq 2\langle Y - X\bar{\beta}, X\hat{\beta} - X\bar{\beta}\rangle \\ &= 2\langle f_0(X) - X\bar{\beta}, X\hat{\beta} - X\bar{\beta}\rangle + 2\langle \epsilon, X\hat{\beta} - X\bar{\beta}\rangle.\end{aligned}$$

Now we use the polarization identity $\|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2 = 2\langle a, b\rangle$ on the first term on the right-hand side above, yielding

$$\|X\hat{\beta} - X\bar{\beta}\|_2^2 \leq \|X\bar{\beta} - f_0(X)\|_2^2 + \|X\hat{\beta} - X\bar{\beta}\|_2^2 - \|X\hat{\beta} - f_0(X)\|_2^2 + 2\langle X^\mathsf{T}\epsilon, \hat{\beta} - \bar{\beta}\rangle.$$

Cancelling the common term of $\|X\hat{\beta} - X\bar{\beta}\|_2^2$ from each side, then rearranging, and following familiar arguments,

$$\begin{aligned}\|X\hat{\beta} - f_0(X)\|_2^2 &\leq \|X\bar{\beta} - f_0(X)\|_2^2 + 2\langle X^\mathsf{T}\epsilon, \hat{\beta} - \bar{\beta}\rangle \\ &\leq \|X\bar{\beta} - f_0(X)\|_2^2 + 4t\|X^\mathsf{T}\epsilon\|_\infty \\ &\leq \|X\bar{\beta} - f_0(X)\|_2^2 + 4\sigma t\sqrt{\frac{2(\log(2d) + u)}{n}},\end{aligned}$$

where the last line holds with probability at least $1 - e^{-u}$. To be clear, the above statement holds simultaneously over all $\bar{\beta}$ such that $\|\bar{\beta}\|_1 \leq t$, with probability at least $1 - e^{-u}$. Thus we can take an infimum over all such $\bar{\beta}$ on the right-hand side, yielding

$$\frac{1}{n}\|X\hat{\beta} - f_0(X)\|_2^2 \leq \inf_{\|\bar{\beta}\|_1 \leq t} \left(\frac{1}{n}\|X\bar{\beta} - f_0(X)\|_2^2\right) + 4\sigma t\sqrt{\frac{2(\log(2d) + u)}{n}}, \tag{17}$$

with probability at least $1 - e^{-u}$. The bound in (17) is called an *oracle inequality* for the constrained lasso estimator. It says, in terms of in-sample risk, that we can expect the lasso to perform close as well as the best ℓ_1-sparse linear predictor to f_0.

Further, let $\bar{\beta}^{\text{best}}$ achieve the infimum of $\frac{1}{n}\|X\bar{\beta} - f_0(X)\|_2^2$ over $\|\bar{\beta}\|_1 \leq t$ (we are minimizing a continuous function over a compact set, hence its infimum is achieved). A slight variation on the above argument (actually it comes from a sharpening of the basic inequality itself, where we remove the factor 2 that multiplies the sub-Gaussian process term) can be used to show that

$$\frac{1}{n}\|X\hat{\beta} - X\bar{\beta}^{\text{best}}\|_2^2 \leq 4\sigma t\sqrt{\frac{2(\log(2d) + u)}{n}}, \tag{18}$$

with probability at least $1 - e^{-u}$. The bound in (18) says something interesting, above and beyond the oracle inequality in (17): it says the in-sample predictions from the lasso are close to those from the best ℓ_1-sparse linear predictor, *even when this best ℓ_1-sparse linear predictor is far from f_0*.

4 "Fast" rates

Next, we develop what are sometimes called "fast" rates for the lasso. The "slow" rates for the lasso in the last section assume nothing about the predictors and gave us rates (as measured by in-sample prediction risk) that scale as $\sqrt{(\log d)/n}$. In the current section, by assuming (admittedly fairly strong) conditions on X, we'll be able to get rates that scale as $(\log d)/n$. Before developing this, we'll pause to describe why achieving such a rate is notable.

4.1 Interlude: theory for subset selection

Suppose that in the linear model (2), the true coefficient vector β_0 is ℓ_0-sparse, with $s_0 = \|\beta_0\|_0$. In other words, β_0 has s_0 nonzero components. Denote by $S = \mathrm{supp}(\beta_0)$ the true active set. Then we could formulate an oracle estimator—with knowledge of S—by simply performing least squares regression on the active predictors,

$$\hat{\beta}^{\mathrm{oracle}}_S = (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} Y,$$
$$\hat{\beta}^{\mathrm{oracle}}_{-S} = 0.$$

From our previous calculations, we know that this has in-sample prediction risk

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta}^{\mathrm{oracle}} - X\beta_0\|_2^2 = \sigma^2 \frac{s_0}{n}.$$

(This is just as in (3), recalling that X here is fixed.)

How would subset selection do by comparison? Then Foster and George (1994) consider this question, and study the in-sample prediction risk of a solution $\hat{\beta}$ of

$$\underset{\beta}{\mathrm{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_0. \tag{19}$$

They show that if we choose $\lambda \asymp \sigma^2 \log d$, then the best subset selection estimator satisfies

$$\frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \leq 4\log d + 2 + o(1), \quad \text{as } n, d \to \infty. \tag{20}$$

This holds without any conditions on the predictor matrix X. Moreover, they prove the lower bound

$$\inf_{\hat{\beta}} \sup_{X,\beta_0} \frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \geq 2\log d - o(\log d),$$

where the infimum is over all estimators $\hat{\beta}$, and the supremum is over all predictor matrices X and underlying coefficients β_0 that have ℓ_0-sparsity equal at most s_0. Hence, in terms of rate, best subset selection is optimal: precisely, it achieves the optimal *risk inflation* over the oracle risk of $\sigma^2 s_0/n$.

Comparison: subset selection versus lasso, when the true model is ℓ_0-sparse. It is informative to compare rates from ℓ_0 and ℓ_1 penalization in the sparse linear model setting, where $s_0 = \|\beta_0\|_0$.

- From the result in (20), we see that best subset selection (19) leads to an-sample prediction risk on the order of $s_0(\log d)/n$. Some authors like to say that the factor of $\log d$ is the "price it pays" for searching over which of the d variables is relevant for prediction (which is, in a sense, a remarkably small price). And notably, best subset selection achieves this with no assumptions on X whatsoever.
- From the result in (14), we see that the lasso (6) leads to an in-sample prediction risk on the order of $\|\beta_0\|_1\sqrt{(\log d)/n}$, again with no assumptions on X. If each nonzero entry of β_0 is of constant order, then this will be $s_0\sqrt{(\log d)/n}$ which is a still "full square root factor slower" than the subset selection rate.

This observation motivates us to find a way to get sharper rates for in-sample prediction risk with the lasso. We will see next that we can do so with conditions on X that limit correlations between features.

Interlude on an interlude: is subset selection the gold standard? (You can skip this if you want and just head over to the analysis in the next subsection ... but this point is too important to not mention at all.) Many authors seem to treat best subset selection as the gold standard. That is, if we could compute it, then we would always want to use it over a method like the lasso. However, the story is actually much more nuanced than that. Yes, the last subsection showed that in an idealized setting where the true model linear and ℓ_0-sparse, using ℓ_0 penalization results in sharper guarantees than ℓ_1 penalization, for a

generic feature matrix X. But this does not mean that it will perform better in practice for every (or even a typical) high-dimensional regression problem that we might want to solve.

Best subset selection tends to have much higher variance than the lasso, because there is shrinkage inherent in the latter's coefficient estimates (recall that we can see this directly from (10)). As a result, which estimator performs better in practice really depends on a lot of factors like the SNR (signal-to-noise ratio); whether the true model is ℓ_0-sparse, ℓ_1-sparse, approximately sparse, etc.; feature correlations; and so on. See Hastie et al. (2020) for a discussion of this, and extensive empirical comparisons.

4.2 Compatibility condition

Returning to the lasso, we'll now show how certain assumptions on X can give us sharper rates. As in the last subsection, we assume that β_0 in (2) has support $S = \mathrm{supp}(\beta_0)$ and we denote $s_0 = |S|$. We note that there are many flavors of "fast" rates, and the conditions required are all fairly closely related. We'll limit our discussion to only two such conditions, for simplicity.

The first condition, which we study in this subsection, is called the *compatibility condition* on X. This is defined with respect to the true support set S, and says that for some compatibility constant $\phi_0 > 0$,

$$\frac{1}{n}\|Xv\|_2^2 \geq \frac{\phi_0^2}{s_0}\|v_S\|_1^2, \quad \text{for all } v \in \mathbb{R}^d \text{ such that } \|v_{-S}\|_1 \leq 3\|v_S\|_1. \tag{21}$$

While this may look like an odd condition, we will see it being useful in the analysis below, and we will also have some help interpreting it when we discuss the restricted eigenvalue condition shortly. Roughly, it means the true active predictors can't be too correlated.

Recall our previous analysis for the lasso estimator in penalized form (6). It turns out to be helpful to peel back to the step right before we used the triangle inequality, i.e., the line above (13). Applying the result on the maximum of sub-Gaussians, we get

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\sigma\sqrt{2n(\log(2d) + u)}\|\hat{\beta} - \beta_0\|_1 + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1),$$

on an event Ω of probability at least $1 - e^{-u}$. The remainder of the analysis will be performed on Ω, implicitly. Choosing $\lambda \geq 4\sigma\sqrt{2n(\log(2d) + u)}$ (note this is a factor of 2 larger than before) we have

$$\begin{aligned}
\|X\hat{\beta} - X\beta_0\|_2^2 &\leq \frac{\lambda}{2}\|\hat{\beta} - \beta_0\|_1 + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\
&\leq \frac{\lambda}{2}\|\hat{\beta}_S - \beta_{0,S}\|_1 + \lambda\|\hat{\beta}_{-S}\|_1 + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\
&\leq \frac{\lambda}{2}\|\hat{\beta}_S - \beta_{0,S}\|_1 + \lambda\|\hat{\beta}_{-S}\|_1 + \lambda(\|\beta_{0,S} - \hat{\beta}_S\|_1 - \|\hat{\beta}_{-S}\|_1) \\
&= \frac{3\lambda}{2}\|\hat{\beta}_S - \beta_{0,S}\|_1 - \frac{\lambda}{2}\|\hat{\beta}_{-S}\|_1,
\end{aligned} \tag{22}$$

where the two inequalities each follow from the triangle inequality. As the left-hand side is nonnegative, $\|X\hat{\beta} - X\beta_0\|_2^2 \geq 0$, we have shown

$$\|\hat{\beta}_{-S} - \hat{\beta}_{0,-S}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1,$$

and thus we may apply the compatibility condition (21) to the vector $v = \hat{\beta} - \beta_0$. This gives us two bounds: one on the fitted values, and the other on the coefficients. Both use as a jumping off point the following key inequality, which is a consequence of (22),

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\|\hat{\beta}_S - \beta_{0,S}\|_1. \tag{23}$$

In-sample prediction risk. To bound the in-sample prediction risk, we upper bound the right-hand side in (23) using (21),

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\sqrt{\frac{s_0}{\phi_0^2 n}}\|X\hat{\beta} - X\beta_0\|_2.$$

Dividing through both sides by $\|X\hat{\beta} - X\beta_0\|_2$, then squaring both sides, and dividing by n,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{9s_0\lambda^2}{\phi_0^2 n^2}.$$

Plugging in $\lambda = 4\sigma\sqrt{2n(\log(2d) + u)}$, we have shown that

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{288\sigma^2 s_0(\log(2d) + u)}{\phi_0^2 n}, \tag{24}$$

with probability at least $1 - e^{-u}$. As desired, we have achieved a "fast" rate for the lasso estimator: the in-sample prediction risk scales as $s_0(\log d)/n$.

Coefficient risk. To bound the coefficient risk, we can essentially just perform the reverse argument: we lower bound the left-hand side in the key inequality (23), giving

$$\frac{n\phi_0^2}{s_0}\|\hat{\beta}_S - \beta_{0,S}\|_1^2 \leq 3\lambda\|\hat{\beta}_S - \beta_{0,S}\|_1.$$

Dviding through both sides by $\|\hat{\beta}_S - \beta_{0,S}\|_1$, and recalling $\|\hat{\beta}_{-S}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1$, which implies by the triangle inequality that $\|\hat{\beta} - \beta_0\|_1 \leq 4\|\hat{\beta}_S - \beta_{0,S}\|_1$, we get

$$\|\hat{\beta} - \beta_0\|_1 \leq \frac{12s_0\lambda}{\phi_0^2 n}.$$

Plugging in $\lambda = 4\sigma\sqrt{2n(\log(2d) + u)}$, we have shown that

$$\|\hat{\beta} - \beta_0\|_1 \leq \frac{48\sigma s_0}{\phi_0^2}\sqrt{\frac{2(\log(2d) + u)}{n}}, \tag{25}$$

with probability at least $1 - e^{-u}$. We see that the coefficient risk in the ℓ_1 norm scales as $s_0\sqrt{(\log d)/n}$.

4.3 Restricted eigenvalue condition

Instead of compatibility, we may assume that X satisfies what is called the *restricted eigenvalue condition* with constant $\phi_0 > 0$,

$$\begin{aligned}\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2, \quad &\text{for all } v \in \mathbb{R}^d \text{ such that } \|v_{-J}\|_1 \leq 3\|v_J\|_1, \\ &\text{and all subsets } J \subseteq \{1, \ldots, d\} \text{ such that } |J| = s_0.\end{aligned} \tag{26}$$

This produces similar results as in (24), (25), but instead of the latter we get a coefficient bound in the ℓ_2 norm, of the form

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0 \log d}{\phi_0^4 n}.$$

with high probability. Notice the similarity between (26) and (21). The restricted eigenvalue condition is actually stronger, i.e., it implies the compatibility condition, as we always have $\|\beta\|_2^2 \geq \|\beta_J\|_2^2 \geq \|\beta_J\|_1^2/s_0$. We may interpret the restricted eigenvalue condition roughly as follows. The requirement

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2, \quad \text{for all } v \in \mathbb{R}^d,$$

would be a lower bound of ϕ_0^2 on the smallest eigenvalue of $X^\mathsf{T}X/n$. We don't assume this (and this would of course mean that X needs to be full column rank, which couldn't happen when $d > n$), but instead in (26) we assume that this inequality holds for all vectors v that are "mostly" supported on small subsets J of variables, with $|J| = s_0$.

References

Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22(4):1947–1975, 1994.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009. Second edition.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall, 2015.

Trevor Hastie, Robert Tibshirani, and Ryan J. Tibshirani. Best subset, forward stepwise, or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.