# A Closer Look at Sparse Regression
Ryan Tibshirani
(ammended by Larry Wasserman)

# 1 Introduction

In these notes we take a closer look at sparse linear regression. Throughout, we make the very strong assumption that $Y_i = \beta^T X_i + \epsilon_i$ where $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\mathrm{Var}(\epsilon_i|X_i) = \sigma^2$. These assumptions are highly unrealistic but they permit a more detailed analysis. There are several books on high-dimensional estimation: Hastie, Tibshirani & Wainwright (2015), Buhlmann & van de Geer (2011), Wainwright (2017).

# 2 Best subset selection, ridge regression, and the lasso

## 2.1 Three norms: $\ell_0$, $\ell_1$, $\ell_2$

In terms of regularization, we typically choose the constraint set $C$ to be a sublevel set of a norm (or seminorm), and equivalently, the penalty function $P(\cdot)$ to be a multiple of a norm (or seminorm)

Let's consider three canonical choices: the $\ell_0$, $\ell_1$, and $\ell_2$ norms:

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

(Truthfully, calling it "the $\ell_0$ norm" is a misnomer, since it is not a norm: it does not satisfy positive homogeneity, i.e., $\|a\beta\|_0 \neq a\|\beta\|_0$ whenever $a \neq 0, 1$.)

In constrained form, this gives rise to the problems:

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_0 \leq k \qquad \text{(Best subset selection)} \quad (1)$$

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq t \qquad \text{(Lasso regression)} \quad (2)$$

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_2^2 \leq t \qquad \text{(Ridge regession)} \quad (3)$$

where $k, t \geq 0$ are tuning parameters. Note that it makes sense to restrict $k$ to be an integer; in best subset selection, we are quite literally finding the best subset of variables of size $k$, in terms of the achieved training error

Though it is likely the case that these ideas were around earlier in other contexts, in statistics we typically subset selection to Beale et al. (1967), Hocking & Leslie (1967), ridge regression to Hoerl & Kennard (1970), and the lasso to Tibshirani (1996), Chen et al. (1998)

In penalized form, the use of $\ell_0, \ell_1, \ell_2$ norms gives rise to the problems:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \qquad \text{(Best subset selection)} \tag{4}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{(Lasso regression)} \tag{5}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad \text{(Ridge regression)} \tag{6}$$

with $\lambda \geq 0$ the tuning parameter. In fact, problems (2), (5) are equivalent. By this, we mean that for any $t \geq 0$ and solution $\widehat{\beta}$ in (2), there is a value of $\lambda \geq 0$ such that $\widehat{\beta}$ also solves (5), and vice versa. The same equivalence holds for (3), (6). (The factors of $1/2$ multiplying the squared loss above are inconsequential, and just for convenience)

It means, roughly speaking, that computing solutions of (2) over a sequence of $t$ values and performing cross-validation (to select an estimate) should be basically the same as computing solutions of (5) over some sequence of $\lambda$ values and performing cross-validation (to select an estimate). Strictly speaking, this isn't quite true, because the precise correspondence between equivalent $t, \lambda$ depends on the data $X, y$

Notably, problems (1), (4) are *not equivalent*. For every value of $\lambda \geq 0$ and solution $\widehat{\beta}$ in (4), there is a value of $t \geq 0$ such that $\widehat{\beta}$ also solves (1), but the converse is not true

## 2.2 A Toy Example

It is helpful to first consider a toy example. Suppose that $Y \sim N(\mu, 1)$. Let's consider the three different estimators we get using the following three different loss functions:

$$\frac{1}{2}(Y-\mu)^2 + \lambda||\mu||_0, \quad \frac{1}{2}(Y-\mu)^2 + \lambda|\mu|, \quad \frac{1}{2}(Y-\mu)^2 + \lambda\mu^2.$$

You should verify that the solutions are

$$\widehat{\mu} = H(Y; \sqrt{2\lambda}), \quad \widehat{\mu} = S(Y; \lambda), \quad \widehat{\mu} = \frac{Y}{1+2\lambda}$$

where $H(y; a) = yI(|y| > a)$ is the hard-thresholding operator, and

$$S(y; a) = \begin{cases} y - a & \text{if } y > a \\ 0 & \text{if } -a \leq y \leq a \\ y + a & \text{if } y < a. \end{cases}$$

Hard thresholding creates a "zone of sparsity" but it is discontinuous. Soft thresholding also creates a "zone of sparsity" but it is scontinuous. The $L_2$ loss creates a nice smooth estimator but it is never sparse. (You can verify the solution to the $L_1$ problem using sub-differentials if you know convex analysis, or by doing three cases separately: $\mu > 0$, $\mu = 0$, $\mu < 0$.)

## 2.3 Sparsity

The best subset selection and the lasso estimators have a special, useful property: their solutions are *sparse*, i.e., at a solution $\widehat{\beta}$ we will have $\widehat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$. In problem (1), this is obviously true, where $k \geq 0$ controls the sparsity level. In problem (2), it is less obviously true, but we get a higher degree of sparsity the smaller the value of $t \geq 0$. In the penalized forms, (4), (5), we get more sparsity the larger the value of $\lambda \geq 0$

This is not true of ridge regression, i.e., the solution of (3) or (6) generically has all nonzero components, no matter the value of $t$ or $\lambda$. Note that sparsity is desirable, for two reasons: (i) it corresponds to performing variable selection in the constructed linear model, and (ii) it provides a level of interpretability (beyond sheer accuracy)

That the $\ell_0$ norm induces sparsity is obvious. But, why does the $\ell_1$ norm induce sparsity and not the $\ell_2$ norm? There are different ways to look at it; let's stick with intuition from the constrained problem forms (2), (5). Figure 1 shows the "classic" picture, contrasting the way the contours of the squared error loss hit the two constraint sets, the $\ell_1$ and $\ell_2$ balls. As the $\ell_1$ ball has sharp corners (aligned with the coordinate axes), we get sparse solutions

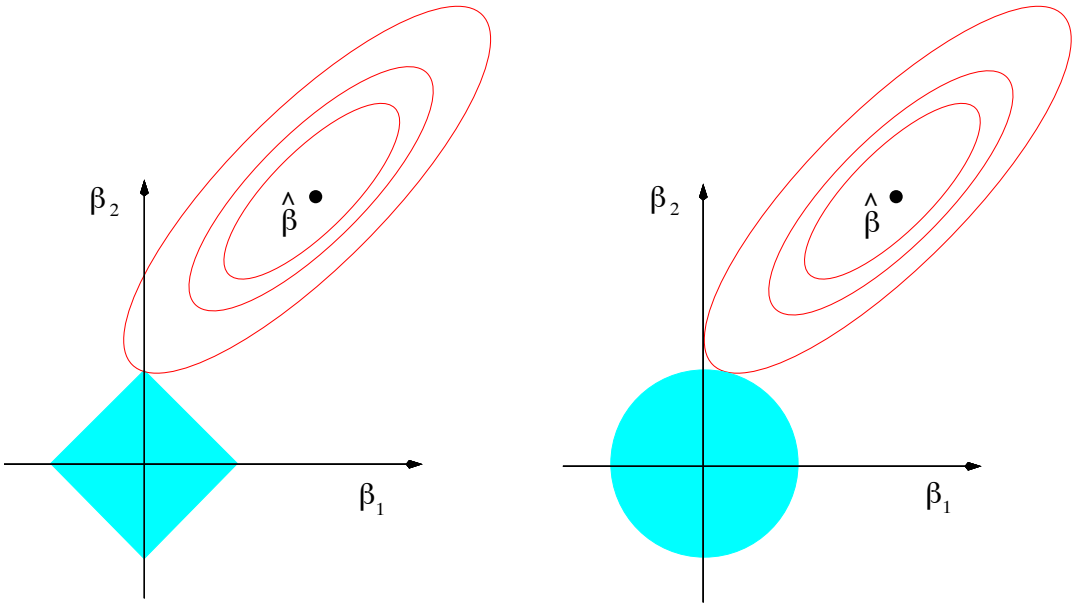


Figure 1: *The "classic" illustration comparing lasso and ridge constraints. From Chapter 3 of Hastie et al. (2009)*

Intuition can also be drawn from the orthogonal case. When $X$ is orthogonal, it is not hard to show that the solutions of the penalized problems (4), (5), (6) are

$$\widehat{\beta}^{\text{subset}} = H_{\sqrt{2\lambda}}(X^T y), \quad \widehat{\beta}^{\text{lasso}} = S_\lambda(X^T y), \quad \widehat{\beta}^{\text{ridge}} = \frac{X^T y}{1 + 2\lambda}$$

respectively, where $H_t(\cdot), S_t(\cdot)$ are the componentwise hard- and soft-thresholding functions at the level $t$. We see several revealing properties: subset selection and lasso solutions exhibit sparsity when the componentwise least squares coefficients (inner products $X^T y$) are small enough; the lasso solution exihibits shrinkage, in that large enough least squares coefficients are shrunken towards zero by $\lambda$; the ridge regression solution is never sparse and compared to the lasso, preferentially shrinkage the larger least squares coefficients even more

## 2.4 Convexity

The lasso and ridge regression problems (2), (3) have another very important property: they are convex optimization problems. Best subset selection (1) is not, in fact it is very far from being convex. Consider using the norm $||\beta||_p$ as a penalty. Sparsity requires $p \leq 1$ and convexity requires $p \geq 1$. The only norm that gives sparsity and convexity is $p = 1$. The appendix has a brief review of convexity.

## 2.5 Theory For Subset Selection

Despite its computational intractability, best subset selection has some attractive risk properties. A classic result is due to Foster & George (1994), on the in-sample risk of best subset selection in penalized form (4), which we will paraphrase here. First, we raise a very simple point: if $A$ denotes the support (also called the active set) of the subset selection solution $\widehat{\beta}$ in (4)—meaning that $\widehat{\beta}_j = 0$ for all $j \notin A$, and denoted $A = \text{supp}(\widehat{\beta})$—then we have

$$\begin{aligned} \widehat{\beta}_A &= (X_A^T X_A)^{-1} X_A^T y, \\ \widehat{\beta}_{-A} &= 0. \end{aligned} \tag{7}$$

Here and throughout we write $X_A$ for the columns of matrix $X$ in a set $A$, and $x_A$ for the components of a vector $x$ in $A$. We will also use $X_{-A}$ and $x_{-A}$ for the columns or components not in $A$. The observation in (7) follows from the fact that, given the support set $A$, the $\ell_0$ penalty term in the subset selection criterion doesn't depend on the actual magnitudes of the coefficients (it contributes a constant factor), so the problem reduces to least squares.

Now, consider a standard linear model as with $X$ fixed, and $\epsilon \sim N(0, \sigma^2 I)$. Suppose that the underlying coefficients have support $S = \text{supp}(\beta_0)$, and $s_0 = |S|$. Then, the estimator given by least squares on $S$, i.e.,

$$\begin{aligned} \widehat{\beta}_S^{\text{oracle}} &= (X_S^T X_S)^{-1} X_S^T y, \\ \widehat{\beta}_{-S}^{\text{oracle}} &= 0. \end{aligned}$$

is is called *oracle estimator*, and as we know from our previous calculations, has in-sample risk

$$\frac{1}{n} \|X\widehat{\beta}^{\text{oracle}} - X\beta_0\|_2^2 = \sigma^2 \frac{s_0}{n}.$$

Foster & George (1994) consider this setup, and compare the risk of the best subset selection estimator $\widehat{\beta}$ in (4) to the oracle risk of $\sigma^2 s_0/n$. They show that, if we choose $\lambda \asymp \sigma^2 \log p$, then the best subset selection estimator satisfies

$$\frac{\mathbb{E}\|X\widehat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \leq 4 \log p + 2 + o(1), \tag{8}$$

as $n, p \to \infty$. This holds without any conditions on the predictor matrix $X$. Moreover, they prove the lower bound

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

where the infimum is over all estimators $\widehat{\beta}$, and the supremum is over all predictor matrices $X$ and underlying coefficients with $\|\beta_0\|_0 = s_0$. Hence, in terms of rate, best subset selection achieves the optimal risk inflation over the oracle risk.

Returning to what was said above, the kicker is that we can't really compute the best subset selection estimator for even moderately-sized problems. As we will in the following, the lasso provides a similar risk inflation guarantee, though under considerably stronger assumptions.

Lastly, it is worth remarking that even if we *could* compute the subset selection estimator at scale, it's not at all clear that we would want to use this in place of the lasso. (Many people assume that we would.) We must remind ourselves that theory provides us an understanding of the performance of various estimators under typically idealized conditions, and it doesn't tell the complete story. It could be the case that the lack of shrinkage in the subset selection coefficients ends up being harmful in practical situations, in a signal-to-noise regime, and yet the lasso could still perform favorably in such settings.

**Update.** Some nice recent work in optimization (Bertsimas et al. 2016) shows that we can cast best subset selection as a mixed integer quadratic program, and proposes to solve it (in general this means approximately, though with a certified bound on the duality gap) with an industry-standard mixed integer optimization package like Gurobi. However, in a recent paper, Hastie, Tibshirani and Tibshirani (arXiv:1707.08692) show that best subset selection does not do well statistically unless there is an extremely high signal to noise ratio.

# 3 Basic properties and geometry of the lasso

## 3.1 Ridge regression and the elastic net

A quick refresher: the ridge regression problem (6) is always strictly convex (assuming $\lambda > 0$), due to the presense of the squared $\ell_2$ penalty $\|\beta\|_2^2$. To be clear, this is true regardless of $X$, and so the ridge regression solution is always well-defined, and is in fact given in closed-form by $\widehat{\beta} = (X^T X + 2\lambda I)^{-1} X^T y$.

## 3.2 Lasso

Now we turn to subgradient optimality (sometimes called the KKT conditions) for the lasso problem in (5). They tell us that any lasso solution $\widehat{\beta}$ must satisfy

$$X^T(y - X\widehat{\beta}) = \lambda s, \tag{9}$$

where $s \in \partial\|\widehat{\beta}\|_1$, a subgradient of the $\ell_1$ norm evaluated at $\widehat{\beta}$. Precisely, this means that

$$s_j \in \begin{cases} \{+1\} & \widehat{\beta}_j > 0 \\ \{-1\} & \widehat{\beta}_j < 0 \\ [-1, 1] & \widehat{\beta}_j = 0, \end{cases} \quad j = 1, \ldots, p. \tag{10}$$

From (9) we can read off a straightforward but important fact: even though the solution $\widehat{\beta}$ may not be uniquely determined, the optimal subgradient $s$ is a function of the unique fitted value $X\widehat{\beta}$ (assuming $\lambda > 0$), and hence is itself unique.

Now from (10), note that the uniqueness of $s$ implies that any two lasso solutions must have the same signs on the overlap of their supports. That is, it cannot happen that we find two different lasso solutions $\widehat{\beta}$ and $\widetilde{\beta}$ with $\widehat{\beta}_j > 0$ but $\widetilde{\beta}_j < 0$ for some $j$, and hence we have no problem interpretating the signs of components of lasso solutions.

Let's assume henceforth that the columns of $X$ are in general position (and we are looking at a nontrivial end of the path, with $\lambda > 0$), so the lasso solution $\widehat{\beta}$ is unique. Let $A = \mathrm{supp}(\widehat{\beta})$ be the lasso active set, and let $s_A = \mathrm{sign}(\widehat{\beta}_A)$ be the signs of active coefficients. From the subgradient conditions (9), (10), we know that

$$X_A^T(y - X_A\widehat{\beta}_A) = \lambda s_A,$$

and solving for $\widehat{\beta}_A$ gives

$$\begin{aligned} \widehat{\beta}_A &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \\ \widehat{\beta}_{-A} &= 0 \end{aligned} \tag{11}$$

(where recall we know that $X_A^T X_A$ is invertible because $X$ has columns in general position). We see that the active coefficients $\widehat{\beta}_A$ are given by taking the least squares coefficients on $X_A$, $(X_A^T X_A)^{-1} X_A^T y$, and shrinking them by an amount $\lambda(X_A^T X_A)^{-1} s_A$. Contrast this to, e.g., the subset selection solution in (7), where there is no such shrinkage.

Now, how about this so-called shrinkage term $(X_A^T X_A)^{-1} X_A^T y$? Does it always act by moving each one of the least squares coefficients $(X_A^T X_A)^{-1} X_A^T y$ towards zero? Indeed, this is not always the case, and one can find empirical examples where a lasso coefficient is actually larger (in magnitude) than the corresponding least squares coefficient on the active set. Of course, we also know that this is due to the correlations

between active variables, because when $X$ is orthogonal, as we've already seen, this never happens.

On the other hand, it is always the case that the lasso solution has a strictly smaller $\ell_1$ norm than the least squares solution on the active set, and in this sense, we are (perhaps) justified in always referring to $(X_A^T X_A)^{-1} X_A^T y$ as a shrinkage term. To see this, note that, for any vector $b$, $||b||_1 = s^T b$ where $s$ is the vector of signs of $b$. So $||\widehat{\beta}||_1 = s^T \widehat{\beta} = s_A^T \widehat{\beta}_A$ and so

$$\|\widehat{\beta}\|_1 = s_A^T (X_A^T X_A)^{-1} X_A^T y - \lambda s_A^T (X_A^T X_A)^{-1} s_A < \|(X_A^T X_A)^{-1} X_A^T y\|_1. \tag{12}$$

The first term is less than or equal to $\|(X_A^T X_A)^{-1} X_A^T y\|_1$, and the term we are subtracting is strictly negative (because $(X_A^T X_A)^{-1}$ is positive definite).

# 4 Theoretical analysis of the lasso

## 4.1 Slow rates

There has been an enormous amount theoretical work analyzing the performance of the lasso. Some references (warning: a highly incomplete list) are Greenshtein & Ritov (2004), Fuchs (2005), Donoho (2006), Candes & Tao (2006), Meinshausen & Buhlmann (2006), Zhao & Yu (2006), Candes & Plan (2009), Wainwright (2009); a helpful text for these kind of results is Buhlmann & van de Geer (2011).

We begin by stating what are called *slow rates* for the lasso estimator. Most of the proofs are simple enough that they are given below. These results don't place any real assumptions on the predictor matrix $X$, but deliver slow(er) rates for the risk of the lasso estimator than what we would get under more assumptions, hence their name.

We will assume the standard linear model with $X$ fixed, and $\epsilon \sim N(0, \sigma^2)$. We will also assume that $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$. That the errors are Gaussian can be easily relaxed to sub-Gaussianity.

The lasso estimator in bound form (2) is particularly easy to analyze. Suppose that we choose $t = \|\beta_0\|_1$ as the tuning parameter. Then, simply by virtue of optimality of the solution $\widehat{\beta}$ in (2), we find that

$$\|y - X\widehat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2,$$

or, expanding and rearranging,

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X\widehat{\beta} - X\beta_0 \rangle.$$

Here we denote $\langle a, b \rangle = a^T b$. The above is sometimes called the *basic inequality* (for the lasso in bound form). Now, rearranging the inner product, using Holder's inequality, and recalling the choice of bound parameter:

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 2\langle X^T\epsilon, \widehat{\beta} - \beta_0 \rangle \leq 4\|\beta_0\|_1 \|X^T\epsilon\|_\infty.$$

Notice that $\|X^T\epsilon\|_\infty = \max_{j=1,\ldots,p} |X_j^T\epsilon|$ is a maximum of $p$ Gaussians, each with mean zero and variance upper bounded by $\sigma^2 n$. By a standard maximal inequality for Gaussians, for any $\delta > 0$,

$$\max_{j=1,\ldots,p} |X_j^T\epsilon| \leq \sigma\sqrt{2n\log(ep/\delta)},$$

with probability at least $1-\delta$. Plugging this to the second-to-last display and dividing by $n$, we get the finite-sample result for the lasso estimator

$$\frac{1}{n}\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{13}$$

with probability at least $1-\delta$.

The high-probability result (13) implies an in-sample risk bound of

$$\frac{1}{n}\mathbb{E}\|X\widehat{\beta} - X\beta_0\|_2^2 \lesssim \|\beta_0\|_1\sqrt{\frac{\log p}{n}}.$$

Compare to this with the risk bound (8) for best subset selection, which is on the (optimal) order of $s_0 \log p/n$ when $\beta_0$ has $s_0$ nonzero components. If each of the nonzero components here has constant magnitude, then above risk bound for the lasso estimator is on the order of $s_0\sqrt{\log p/n}$, which is much slower.

**Predictive risk.** Instead of in-sample risk, we might also be interested in out-of-sample risk, as after all that reflects actual (out-of-sample) predictions. In least squares, recall, we saw that out-of-sample risk was generally higher than in-sample risk. The same is true for the lasso Chatterjee (2013) gives a nice, simple analysis of out-of-sample risk for the lasso. He assumes that $x_0$, $x_i$, $i = 1,\ldots,n$ are i.i.d. from an arbitrary distribution supported on a compact set in $\mathbb{R}^p$, and shows that the lasso estimator in bound form (2) with $t = \|\beta_0\|_1$ has out-of-sample risk satisfying

$$\mathbb{E}(x_0^T\widehat{\beta} - x_0^T\beta)^2 \lesssim \|\beta_0\|_1^2\sqrt{\frac{\log p}{n}}.$$

The proof is not much more complicated than the above, for the in-sample risk, and reduces to a clever application of Hoeffding's inequality, though we omit it for brevity. Note here the dependence on $\|\beta_0\|_1^2$, rather than $\|\beta_0\|_1$ as in the in-sample risk. This agrees with the analysis we did in the previous set of notes where we did not assume the linear model. (Only the interpretation changes.)

**Oracle inequality.** If we don't want to assume linearity of the mean then we can still derive an *oracle inequality* that characterizes the risk of the lasso estimator in excess of the risk of the best linear predictor. For this part only, assume the more general model

$$y = \mu(X) + \epsilon,$$

with an arbitrary mean function $\mu(X)$, and normal errors $\epsilon \sim N(0, \sigma^2)$. We will analyze the bound form lasso estimator (2) for simplicity. By optimality of $\widehat{\beta}$, for any other $\widetilde{\beta}$ feasible for the lasso problem in (2), it holds that[1]

$$\langle X^T(y - X\widehat{\beta}), \widetilde{\beta} - \widehat{\beta} \rangle \leq 0. \tag{14}$$

Rearranging gives

$$\langle \mu(X) - X\widehat{\beta}, X\widetilde{\beta} - X\widehat{\beta} \rangle \leq \langle X^T\epsilon, \widehat{\beta} - \widetilde{\beta} \rangle. \tag{15}$$

Now using the polarization identity $\|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2 = 2\langle a, b \rangle$,

$$\|X\widehat{\beta} - \mu(X)\|_2^2 + \|X\widehat{\beta} - X\widetilde{\beta}\|_2^2 \leq \|X\widetilde{\beta} - \mu(X)\|_2^2 + 2\langle X^T\epsilon, \widehat{\beta} - \widetilde{\beta} \rangle,$$

and from the exact same arguments as before, it holds that

$$\frac{1}{n}\|X\widehat{\beta} - \mu(X)\|_2^2 + \frac{1}{n}\|X\widehat{\beta} - X\widetilde{\beta}\|_2^2 \leq \frac{1}{n}\|X\widetilde{\beta} - \mu(X)\|_2^2 + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$. This holds simultaneously over all $\widetilde{\beta}$ with $\|\widetilde{\beta}\|_1 \leq t$. Thus, we may write, with probability $1 - \delta$,

$$\frac{1}{n}\|X\widehat{\beta} - \mu(X)\|_2^2 \leq \left\{ \inf_{\|\widetilde{\beta}\|_1 \leq t} \frac{1}{n}\|X\widetilde{\beta} - \mu(X)\|_2^2 \right\} + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}}.$$

Also if we write $X\widetilde{\beta}^{\text{best}}$ as the best linear that predictor of $\ell_1$ at most $t$, achieving the infimum on the right-hand side (which we know exists, as we are minimizing a continuous function over a compact set), then

$$\frac{1}{n}\|X\widehat{\beta} - X\widetilde{\beta}^{\text{best}}\|_2^2 \leq 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$

## 4.2 Fast rates

Under **very** strong assumptions we can get faster rates. For example, if we assume that $X$ satisfies the *restricted eigenvalue condition* with constant $\phi_0 > 0$, i.e.,

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2 \quad \text{for all subsets } J \subseteq \{1, \ldots, p\} \text{ such that } |J| = s_0 \\ \text{and all } v \in \mathbb{R}^p \text{ such that } \|v_{J^c}\|_1 \leq 3\|v_J\|_1 \tag{16}$$

[1] To see this, consider minimizing a convex function $f(x)$ over a convex set $C$. Let $\widehat{x}$ be a minimizer. Let $z \in C$ be any other point in $C$. If we move away from the solution $\widehat{x}$ we can only increase $f(\widehat{x})$. In other words, $\langle \nabla f(\widehat{x}), z - \widehat{z} \rangle \geq 0$.

then

$$\|\widehat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0 \log p}{n\phi_0^2} \tag{17}$$

with probability tending to 1. (This condition can be slightly weakened, but not much.) The condition is unlikely to hold in any real problem. Nor is it checkable. The proof is in the appendix.

### 4.3 Support recovery

Here we discuss results on support recovery of the lasso estimator. There are a few versions of support recovery results and again Buhlmann & van de Geer (2011) is a good place to look for a thorough coverage. Here we describe a result due to Wainwright (2009), who introduced a proof technique called the *primal-dual witness method*. The assumptions are even stronger (and less believable) than in the previous section. In addition to the previous assumptions we need:

*Mutual incoherence:* for some $\gamma > 0$, we have

$$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \gamma, \quad \text{for } j \notin S,$$

*Minimum eigenvalue:* for some $C > 0$, we have

$$\Lambda_{\min}\left(\frac{1}{n} X_S^T X_S\right) \geq C,$$

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix $A$

*Minimum signal:*

$$\beta_{0,\min} = \min_{j \in S} |\beta_{0,j}| \geq \lambda \|(X_S^T X_S)^{-1}\|_\infty + \frac{4\gamma\lambda}{\sqrt{C}},$$

where $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^q |A_{ij}|$ denotes the $\ell_\infty$ norm of an $m \times q$ matrix $A$

Under these assumptions, once can show that, if $\lambda$ is chosen just right, then

$$P(\text{support}(\widehat{\beta}) = \text{support}(\beta)) \to 1. \tag{18}$$

The proof is in the appendix.

## References


Beale, E. M. L., Kendall, M. G. & Mann, D. W. (1967), 'The discarding of variables in multivariate analysis', *Biometrika* **54**(3/4), 357–366.

Bertsimas, D., King, A. & Mazumder, R. (2016), 'Best subset selection via a modern optimization lens', *The Annals of Statistics* **44**(2), 813–852.

Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer.

Candes, E. J. & Plan, Y. (2009), 'Near ideal model selection by $\ell_1$ minimization', *Annals of Statistics* **37**(5), 2145–2177.

Candes, E. J. & Tao, T. (2006), 'Near optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406–5425.

Chatterjee, S. (2013), Assumptionless consistency of the lasso. arXiv: 1303.5817.

Chen, S., Donoho, D. L. & Saunders, M. (1998), 'Atomic decomposition for basis pursuit', *SIAM Journal on Scientific Computing* **20**(1), 33–61.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(12), 1289–1306.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499.

Foster, D. & George, E. (1994), 'The risk inflation criterion for multiple regression', *The Annals of Statistics* **22**(4), 1947–1975.

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**(6), 971–988.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer. Second edition.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, Chapman & Hall.

Hocking, R. R. & Leslie, R. N. (1967), 'Selection of the best subset in regression analysis', *Technometrics* **9**(4), 531–540.

Hoerl, A. & Kennard, R. (1970), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Meinshausen, N. & Buhlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Osborne, M., Presnell, B. & Turlach, B. (2000*a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

Osborne, M., Presnell, B. & Turlach, B. (2000*b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

Raskutti, G., Wainwright, M. J. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls', *IEEE Transactions on Information Theory* **57**(10), 6976–6994.

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

van de Geer, S. & Buhlmann, P. (2009), 'On the conditions used to prove oracle results for the lasso', *Electronic Journal of Statistics* **3**, 1360–1392.

Wainwright, M. (2017), *High-Dimensional Statistics: A Non-Asymptotic View*, Cambridge University Press. To appear.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564.

# 5 Appendix: Convexity

It is convexity that allows to equate (2), (5), and (3), (6) (and yes, the penalized forms are convex problems too). It is also convexity that allows us to both efficiently solve, and in some sense, precisely understand the nature of the lasso and ridge regression solutions

Here is a (far too quick) refresher/introduction to basic convex analysis and convex optimization. Recall that a set $C \subseteq \mathbb{R}^n$ is called *convex* if for any $x, y \in C$ and $t \in [0, 1]$, we have

$$tx + (1 - t)y \in C,$$

i.e., the line segment joining $x, y$ lies entirely in $C$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if its domain $\mathrm{dom}(f)$ is convex, and for any $x, y \in \mathrm{dom}(f)$ and $t \in [0, 1]$,

$$f\big(tx + (1 - t)y\big) \leq tf(x) + (1 - t)f(y),$$

i.e., the function lies below the line segment joining its evaluations at $x$ and $y$. A function is called *strictly convex* if this same inequality holds strictly for $x \neq y$ and $t \in (0, 1)$

E.g., lines, rays, line segments, linear spaces, affine spaces, hyperplans, halfspaces, polyhedra, norm balls are all convex sets

E.g., affine functions $a^T x + b$ are convex and concave, quadratic functions $x^T Q x + b^T x + c$ are convex if $Q \succeq 0$ and strictly convex if $Q \succ 0$, norms are convex

Formally, an *optimization problem* is of the form

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & \ell_j(x) = 0, \;\; j = 1, \ldots r \end{aligned}$$

Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(h_i) \cap \bigcap_{j=1}^r \mathrm{dom}(\ell_j)$ is the common domain of all functions. A *convex optimization problem* is an optimization problem in which all functions $f, h_1, \ldots h_m$ are convex, and all functions $\ell_1, \ldots \ell_r$ are affine. (Think: why affine?) Hence, we can express it as

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & Ax = b \end{aligned}$$

Why is a convex optimization problem so special? The short answer: because *any local minimizer is a global minimizer*. To see this, suppose that $x$ is feasible for the convex problem formulation above and there exists some $R > 0$ such that

$$f(x) \leq f(y) \quad \text{for all feasible } y \text{ with } \|x - y\|_2 \leq R.$$

Such a point $x$ is called a local minimizer. For the sake of contradiction, suppose that $x$ was not a global minimizer, i.e., there exists some feasible $z$ such that $f(z) < f(x)$. By convexity of the constraints (and the domain $D$), the point $tz + (1-t)x$ is feasible for any $0 \leq t \leq 1$. Furthermore, by convexity of $f$,

$$f\big(tz + (1-t)x\big) \leq tf(z) + (1-t)f(x) < f(x)$$

for any $0 < t < 1$. Lastly, we can choose $t > 0$ small enough so that $\|x - (tz + (1-t)x)\|_2 = t\|x - z\|_2 \leq R$, and we obtain a contradiction

Algorithmically, this is a very useful property, because it means if we keep "going downhill", i.e., reducing the achieved criterion value, and we stop when we can't do so anymore, then we've hit the global solution

Convex optimization problems are also special because they come with a beautiful theory of beautiful convex duality and optimality, which gives us a way of understanding the solutions. We won't have time to cover any of this, but we'll mention what subgradient optimality looks like for the lasso

Just based on the definitions, it is not hard to see that (2), (3), (5), (6) are convex problems, but (1), (4) are not. In fact, the latter two problems are known to be NP-hard, so they are in a sense even the worst kind of nonconvex problem

# 6 Appendix: Geometry of the solutions

One undesirable feature of the best subset selection solution (7) is the fact that it behaves discontinuously with $y$. As we change $y$, the active set $A$ must change at some point, and the coefficients will jump discontinuously, because we are just doing least squares onto the active set. So, does the same thing happen with the lasso solution (11)? The answer it not immediately clear. Again, as we change $y$, the active set $A$ must change at some point; but if the shrinkage term were defined "just right", then perhaps the coefficients of variables to leave the active set would gracefully and continously drop to zero, and coefficients of variables to enter the active set would continuously move form zero. This would make whole the lasso solution continuous. Fortuitously, this is indeed the case, and the lasso solution $\widehat{\beta}$ is continuous as a function of $y$. It might seem a daunting task to prove this, but a certain perspective using convex geometry provides a very simple proof. The geometric perspective in fact proves that the lasso fit $X\widehat{\beta}$ is nonexpansive in $y$, i.e., 1-Lipschitz continuous, which is a very strong form of continuity. Define the convex polyhedron $C = \{u : \|X^T u\|_\infty \leq \lambda\} \subseteq \mathbb{R}^n$. Some simple manipulations of the KKT conditions show that the lasso fit is given by

$$X\widehat{\beta} = (I - P_C)(y),$$

the residual from projecting $y$ onto $C$. A picture to show this (just look at the left panel for now) is given in Figure 2.

The projection onto any convex set is nonexpansive, i.e., $\|P_C(y) - P_C(y')\|_2 \leq \|y - y'\|_2$ for any $y, y'$. This should be visually clear from the picture. Actually, the same is true with the residual map: $I - P_C$ is also nonexpansive, and hence the lasso fit is 1-Lipschitz continuous. Viewing the lasso fit as the residual from projection onto a convex polyhedron is actually an even more fruitful perspective. Write this polyhedron as

$$C = (X^T)^{-1}\{v : \|v\|_\infty \leq \lambda\},$$

where $(X^T)^{-1}$ denotes the preimage operator under the linear map $X^T$. The set $\{v : \|v\|_\infty \leq \lambda\}$ is a hypercube in $\mathbb{R}^p$. Every face of this cube corresponds to a subset $A \subseteq \{1, \ldots p\}$ of dimensions (that achieve the maximum value $|\lambda|$) and signs $s_A \in \{-1, 1\}^{|A|}$ (that tell which side of the cube the face will lie on, for each dimension). Now, the faces of $C$ are just faces of $\{v : \|v\|_\infty \leq \lambda\}$ run through the (linear) preimage transformation, so each face of $C$ can also indexed by a set $A \subseteq \{1, \ldots p\}$ and signs $s_A \in \{-1, 1\}^{|A|}$. The picture in Figure 2 attempts to convey this relationship with the colored black face in each of the panels.

Now imagine projecting $y$ onto $C$; it will land on some face. We have just argued that this face corresponds to a set $A$ and signs $s_A$. One can show that this set $A$ is exactly the active set of the lasso solution at $y$, and $s_A$ are exactly the active signs. The size of the active set $|A|$ is the co-dimension of the face. Looking at the picture: we can that see that as we wiggle $y$ around, it will project to the same face. From the

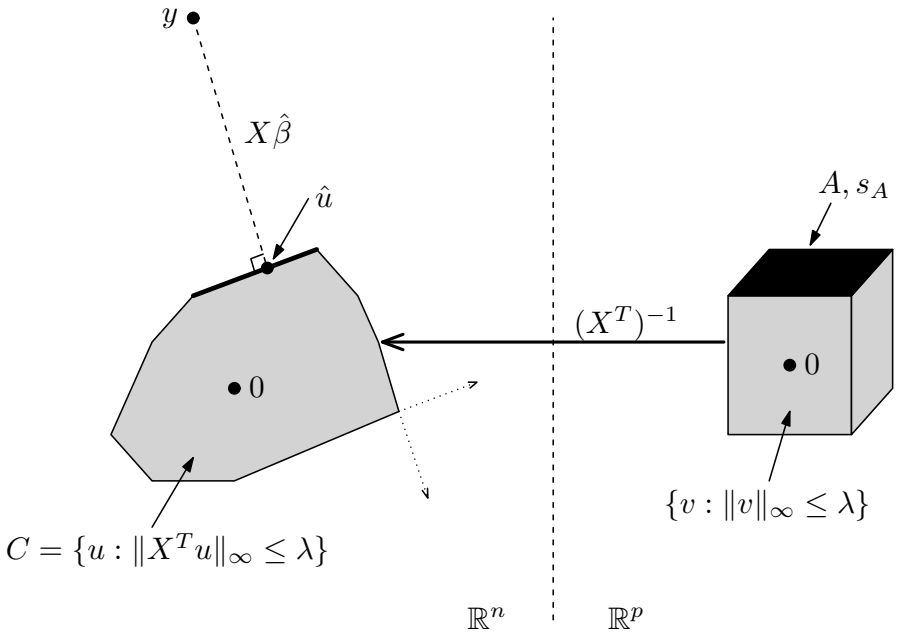


Figure 2: *A geometric picture of the lasso solution. The left panel shows the polyhedron underlying all lasso fits, where each face corresponds to a particular combination of active set $A$ and signs $s$; the right panel displays the "inverse" polyhedron, where the dual solutions live*

correspondence between faces and active set and signs of lasso solutions, this means that $A, s_A$ do not change as we perturb $y$, i.e., they are locally constant. But this isn't true for all points $y$, e.g., if $y$ lies on one of the rays emanating from the lower right corner of the polyhedron in the picture, then we can see that small perturbations of $y$ do actually change the face that it projects to, which invariably changes the active set and signs of the lasso solution. However, this is somewhat of an exceptional case, in that such points can be form a of Lebesgue measure zero, and therefore we can assure ourselves that the active set and signs $A, s_A$ are locally constant for almost every $y$.

From the lasso KKT conditions (9), (10), it is possible to compute the lasso solution in (5) as a function of $\lambda$, which we will write as $\widehat{\beta}(\lambda)$, for all values of the tuning parameter $\lambda \in [0, \infty]$. This is called the *regularization path* or *solution path* of the problem (5). Path algorithms like the one we will describe below are not always possible; the reason that this ends up being feasible for the lasso problem (5) is that the solution path $\widehat{\beta}(\lambda)$, $\lambda \in [0, \infty]$ turns out to be a piecewise linear, continuous function of $\lambda$. Hence, we only need to compute and store the *knots* in this path, which we will denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, and the lasso solution at these knots. From this information, we can then compute the lasso solution at any value

of $\lambda$ by linear interpolation.

The knots $\lambda_1 \geq \ldots \geq \lambda_r$ in the solution path correspond to $\lambda$ values at which the active set $A(\lambda) = \mathrm{supp}(\widehat{\beta}(\lambda))$ changes. As we decrease $\lambda$ from $\infty$ to 0, the knots typically correspond to the point at which a variable enters the active set; this connects the lasso to an incremental variable selection procedure like forward stepwise regression. Interestingly though, as we decrease $\lambda$, a knot in the lasso path can also correspond to the point at which a variables leaves the active set. See Figure 3.

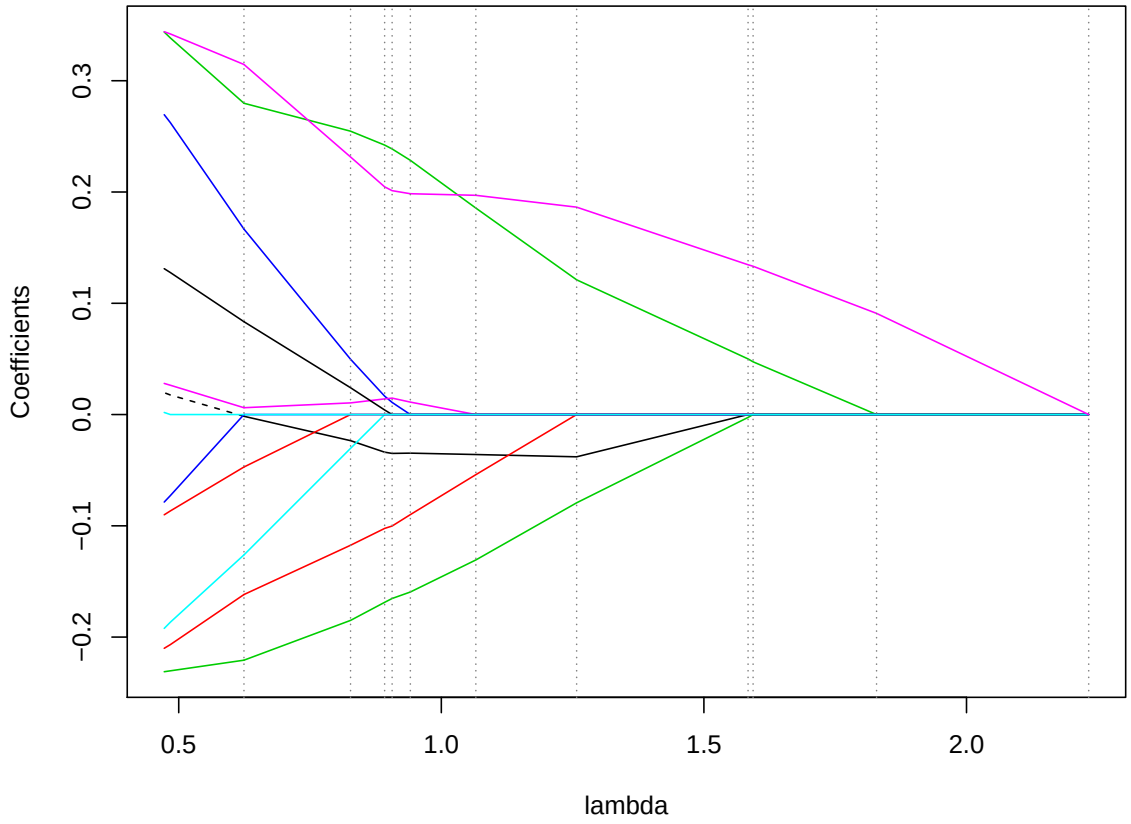


Figure 3: *An example of the lasso path. Each colored line denotes a component of the lasso solution* $\widehat{\beta}_j(\lambda)$, $j = 1, \ldots, p$ *as a function of* $\lambda$. *The gray dotted vertical lines mark the knots* $\lambda_1 \geq \lambda_2 \geq \ldots$

The lasso solution path was described by Osborne et al. (2000*a*,*b*), Efron et al. (2004). Like the construction of all other solution paths that followed these seminal works, the lasso path is essentially given by an iterative or inductive verification of the KKT conditions; if we can maintain that the KKT conditions holds as we decrease $\lambda$, then we know we have a solution. The trick is to start at a value of $\lambda$ at which the solution is trivial; for the lasso, this is $\lambda = \infty$, at which case we know the solution must be $\widehat{\beta}(\infty) = 0$.

Why would the path be piecewise linear? The construction of the path from the

KKT conditions is actually rather technical (not difficult conceptually, but somewhat tedious), and doesn't shed insight onto this matter. But we can actually see it clearly from the projection picture in Figure 2.

As $\lambda$ decreases from $\infty$ to 0, we are shrinking (by a multiplicative factor $\lambda$) the polyhedron onto which $y$ is projected; let's write $C_\lambda = \{u : \|X^T u\|_\infty \leq \lambda\} = \lambda C_1$ to make this clear. Now suppose that $y$ projects onto the relative interior of a certain face $F$ of $C_\lambda$, corresponding to an active set $A$ and signs $s_A$. As $\lambda$ decreases, the point on the boundary of $C_\lambda$ onto which $y$ projects, call it $\widehat{u}(\lambda) = P_{C_\lambda}(y)$, will move along the face $F$, and change linearly in $\lambda$ (because we are equivalently just tracking the projection of $y$ onto an affine space that is being scaled by $\lambda$). Thus, the lasso fit $X\widehat{\beta}(\lambda) = y - \widehat{u}(\lambda)$ will also behave linearly in $\lambda$. Eventually, as we continue to decrease $\lambda$, the projected point $\widehat{u}(\lambda)$ will move to the relative boundary of the face $F$; then, decreasing $\lambda$ further, it will lie on a different, neighboring face $F'$. This face will correspond to an active set $A'$ and signs $s_{A'}$ that (each) differ by only one element to $A$ and $s_A$, respectively. It will then move linearly across $F'$, and so on.

Now we will walk through the technical derivation of the lasso path, starting at $\lambda = \infty$ and $\widehat{\beta}(\infty) = 0$, as indicated above. Consider decreasing $\lambda$ from $\infty$, and continuing to set $\widehat{\beta}(\lambda) = 0$ as the lasso solution. The KKT conditions (9) read

$$X^T y = \lambda s,$$

where $s$ is a subgradient of the $\ell_1$ norm evaluated at 0, i.e., $s_j \in [-1, 1]$ for every $j = 1, \ldots, p$. For large enough values of $\lambda$, this is satisfied, as we can choose $s = X^T y/\lambda$. But this ceases to be a valid subgradient if we decrease $\lambda$ past the point at which $\lambda = |X_j^T y|$ for some variable $j = 1, \ldots, p$. In short, $\widehat{\beta}(\lambda) = 0$ is the lasso solution for all $\lambda \geq \lambda_1$, where

$$\lambda_1 = \max_{j=1,\ldots,p} |X_j^T y|. \tag{19}$$

What happens next? As we decrease $\lambda$ from $\lambda_1$, we know that we're going to have to change $\widehat{\beta}(\lambda)$ from 0 so that the KKT conditions remain satisfied. Let $j_1$ denote the variable that achieves the maximum in (19). Since the subgradient was $|s_{j_1}| = 1$ at $\lambda = \lambda_1$, we see that we are "allowed" to make $\widehat{\beta}_{j_1}(\lambda)$ nonzero. Consider setting

$$\begin{aligned} \widehat{\beta}_{j_1}(\lambda) &= (X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \\ \widehat{\beta}_j(\lambda) &= 0, \quad \text{for all } j \neq j_1, \end{aligned} \tag{20}$$

as $\lambda$ decreases from $\lambda_1$, where $s_{j_1} = \text{sign}(X_{j_1}^T y)$. Note that this makes $\widehat{\beta}(\lambda)$ a piecewise linear and continuous function of $\lambda$, so far. The KKT conditions are then

$$X_{j_1}^T \Big( y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \Big) = \lambda s_{j_1},$$

which can be checked with simple algebra, and

$$\Big| X_j^T \Big( y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \Big) \Big| \leq \lambda,$$

for all $j \neq j_1$. Recall that the above held with strict inequality at $\lambda = \lambda_1$ for all $j \neq j_1$, and by continuity of the constructed solution $\widehat{\beta}(\lambda)$, it should continue to hold as we decrease $\lambda$ for at least a little while. In fact, it will hold until one of the piecewise linear paths

$$X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})), \quad j \neq j_1$$

becomes equal to $\pm\lambda$, at which point we have to modify the solution because otherwise the implicit subgradient

$$s_j = \frac{X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}))}{\lambda}$$

will cease to be in $[-1, 1]$. It helps to draw yourself a picture of this.

Thanks to linearity, we can compute the critical "hitting time" explicitly; a short calculation shows that, the lasso solution continues to be given by (20) for all $\lambda_1 \geq \lambda \geq \lambda_2$, where

$$\lambda_2 = \max_{j \neq j_1,\, s_j \in \{-1,1\}}{}^{+} \frac{X_j^T(I - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}X_{j_1})y}{s_j - X_j^T X_{j_1}(X_{j_1}^T X_{j_1})^{-1}s_{j_1}}, \tag{21}$$

and $\max^+$ denotes the maximum over all of its arguments that are $< \lambda_1$.

To keep going: let $j_2, s_2$ achieve the maximum in (21). Let $A = \{j_1, j_2\}$, $s_A = (s_{j_1}, s_{j_2})$, and consider setting

$$\begin{aligned} \widehat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \widehat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{22}$$

as $\lambda$ decreases from $\lambda_2$. Again, we can verify the KKT conditions for a stretch of decreasing $\lambda$, but will have to stop when one of

$$X_j^T(y - X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \quad j \notin A$$

becomes equal to $\pm\lambda$. By linearity, we can compute this next "hitting time" explicitly, just as before. Furthermore, though, we will have to check whether the active components of the computed solution in (22) are going to cross through zero, because past such a point, $s_A$ will no longer be a proper subgradient over the active components. We can again compute this next "crossing time" explicitly, due to linearity. Therefore, we maintain that (22) is the lasso solution for all $\lambda_2 \geq \lambda \geq \lambda_3$, where $\lambda_3$ is the maximum of the next hitting time and the next crossing time. For convenience, the lasso path algorithm is summarized below.

As we decrease $\lambda$ from a knot $\lambda_k$, we can rewrite the lasso coefficient update in Step 1 as

$$\begin{aligned} \widehat{\beta}_A(\lambda) &= \widehat{\beta}_A(\lambda_k) + (\lambda_k - \lambda)(X_A^T X_A)^{-1} s_A, \\ \widehat{\beta}_{-A}(\lambda) &= 0. \end{aligned} \tag{23}$$

We can see that we are moving the active coefficients in the direction $(\lambda_k - \lambda)(X_A^T X_A)^{-1} s_A$ for decreasing $\lambda$. In other words, the lasso fitted values proceed as

$$X\widehat{\beta}(\lambda) = X\widehat{\beta}(\lambda_k) + (\lambda_k - \lambda) X_A (X_A^T X_A)^{-1} s_A,$$

for decreasing $\lambda$. Efron et al. (2004) call $X_A(X_A^T X_A)^{-1} s_A$ the *equiangular direction*, because this direction, in $\mathbb{R}^n$, takes an equal angle with all $X_j \in \mathbb{R}^n$, $j \in A$.

For this reason, the lasso path algorithm in Algorithm **??** is also often referred to as the *least angle regression* path algorithm in "lasso mode", though we have not mentioned this yet to avoid confusion. Least angle regression is considered as another algorithm by itself, where we skip Step 3 altogether. In words, Step 3 disallows any component path to cross through zero. The left side of the plot in Figure 3 visualizes the distinction between least angle regression and lasso estimates: the dotted black line displays the least angle regression component path, crossing through zero, while the lasso component path remains at zero.

Lastly, an alternative expression for the coefficient update in (23) (the update in Step 1) is

$$\begin{aligned} \widehat{\beta}_A(\lambda) &= \widehat{\beta}_A(\lambda_k) + \frac{\lambda_k - \lambda}{\lambda_k} (X_A^T X_A)^{-1} X_A^T r(\lambda_k), \\ \widehat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{24}$$

where $r(\lambda_k) = y - X_A \widehat{\beta}_A(\lambda_k)$ is the residual (from the fitted lasso model) at $\lambda_k$. This follows because, recall, $\lambda_k s_A$ are simply the inner products of the active variables with the residual at $\lambda_k$, i.e., $\lambda_k s_A = X_A^T (y - X_A \widehat{\beta}_A(\lambda_k))$. In words, we can see that the update for the active lasso coefficients in (24) is in the direction of the least squares coefficients of the residual $r(\lambda_k)$ on the active variables $X_A$.

# 7 Appendix: Fast Rates

Here is a proof of (17). There are many flavors of fast rates, and the conditions required are all very closely related. van de Geer & Buhlmann (2009) provides a nice review and discussion. Here we just discuss two such results, for simplicity.

**Compatibility result.** Assume that $X$ satisfies the *compatibility condition* with respect to the true support set $S$, i.e., for some compatibility constant $\phi_0 > 0$,

$$\frac{1}{n} \|Xv\|_2^2 \geq \frac{\phi_0^2}{s_0} \|v_S\|_1^2 \quad \text{for all } v \in \mathbb{R}^p \text{ such that } \|v_{-S}\|_1 \leq 3\|v_S\|_1. \tag{25}$$

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

Recall from our previous analysis for the lasso estimator in penalized form (5), we showed on an event $E_\delta$ of probability at least $1-\delta$,

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 2\sigma\sqrt{2n\log(ep/\delta)}\|\widehat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\widehat{\beta}\|_1).$$

Choosing $\lambda$ large enough and applying the triangle inequality then gave us the slow rate we derived before. Now we choose $\lambda$ just slightly larger (by a factor of 2): $\lambda \geq 2\sigma\sqrt{2n\log(ep/\delta)}$. The remainder of the analysis will be performed on the event $E_\delta$ and we will no longer make this explicit until the very end. Then

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

where the two inequalities both followed from the triangle inequality, one application for each of the two terms, and we have used that $\widehat{\beta}_{0,-S} = 0$. As $\|X\widehat{\beta} - X\beta_0\|_2^2 \geq 0$, we have shown

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

and thus we may apply the compatibility condition (25) to the vector $v = \widehat{\beta} - \beta_0$. This gives us two bounds: one on the fitted values, and the other on the coefficients. Both start with the key inequality (from the second-to-last display)

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\|\widehat{\beta}_S - \beta_{0,S}\|_1. \tag{26}$$

For the fitted values, we upper bound the right-hand side of the key inequality (26),

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

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

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

Plugging in $\lambda = 2\sigma\sqrt{2n\log(ep/\delta)}$, we have shown that

$$\frac{1}{n}\|X\widehat{\beta} - X\beta_0\|_2^2 \leq \frac{72\sigma^2 s_0\log(ep/\delta)}{n\phi_0^2}, \tag{27}$$

with probability at least $1-\delta$. Notice the similarity between (27) and (8): both provide us in-sample risk bounds on the order of $s_0\log p/n$, but the bound for the lasso requires a strong compability assumption on the predictor matrix $X$, which roughly means the predictors can't be too correlated

For the coefficients, we lower bound the left-hand side of the key inequality (26),

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

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

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

Plugging in $\lambda = 2\sigma\sqrt{2n\log(ep/\delta)}$, we have shown that

$$\|\widehat{\beta} - \beta_0\|_1 \leq \frac{24\sigma s_0}{\phi_0^2}\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{28}$$

with probability at least $1-\delta$. This is a error bound on the order of $s_0\sqrt{\log p/n}$ for the lasso coefficients (in $\ell_1$ norm)

**Restricted eigenvalue result.** Instead of compatibility, we may assume that $X$ satisfies the *restricted eigenvalue condition* with constant $\phi_0 > 0$, i.e.,

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2 \quad \text{for all subsets } J \subseteq \{1,\ldots,p\} \text{ such that } |J| = s_0 \\ \text{and all } v \in \mathbb{R}^p \text{ such that } \|v_{J^c}\|_1 \leq 3\|v_J\|_1. \tag{29}$$

This produces essentially the same results as in (27), (28), but additionally, in the $\ell_2$ norm,

$$\|\widehat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0\log p}{n\phi_0^2}$$

with probability tending to 1

Note the similarity between (29) and the compatibility condition (25). The former is actually stronger, i.e., it implies the latter, because $\|\beta\|_2^2 \geq \|\beta_J\|_2^2 \geq \|\beta_J\|_1^2/s_0$. We may interpret the restricted eigenvalue condition roughly as follows: the requirement $(1/n)\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2$ for all $v \in \mathbb{R}^n$ would be a lower bound of $\phi_0^2$ on the smallest eigenvalue of $(1/n)X^TX$; we don't require this (as this would of course mean that $X$ was full column rank, and couldn't happen when $p > n$), but instead that require that the same inequality hold for $v$ that are "mostly" supported on small subsets $J$ of variables, with $|J| = s_0$

# 8 Appendix: Support Recovery

Again we assume a standard linear model (**??**), with $X$ fixed, subject to the scaling $\|X_j\|_2^2 \leq n$, for $j = 1,\ldots,p$, and $\epsilon \sim N(0,\sigma^2)$. Denote by $S = \text{supp}(\beta_0)$ the true support set, and $s_0 = |S|$. Assume that $X_S$ has full column rank

We aim to show that, at some value of $\lambda$, the lasso solution $\widehat{\beta}$ in (5) has an active set that exactly equals the true support set,

$$A = \text{supp}(\widehat{\beta}) = S,$$

with high probability. We actually aim to show that the signs also match,

$$\text{sign}(\widehat{\beta}_S) = \text{sign}(\beta_{0,S}),$$

with high probability. The primal-dual witness method basically plugs in the true support $S$ into the KKT conditions for the lasso (9), (10), and checks when they can be verified

We start by breaking up (9) into two blocks, over $S$ and $S^c$. Suppose that $\text{supp}(\widehat{\beta}) = S$ at a solution $\widehat{\beta}$. Then the KKT conditions become

$$X_S^T(y - X_S\widehat{\beta}_S) = \lambda s_S \tag{30}$$

$$X_{-S}^T(y - X_S\widehat{\beta}_S) = \lambda s_{-S}. \tag{31}$$

Hence, if we can satisfy the two conditions (30), (31) with a proper subgradient $s$, such that

$$s_S = \text{sign}(\beta_{0,S}) \quad \text{and} \quad \|s_{-S}\|_\infty = \max_{j \notin S} |s_j| < 1,$$

then we have met our goal: we have recovered a (unique) lasso solution whose active set is $S$, and whose active signs are $\text{sign}(\beta_{0,S})$

So, let's solve for $\widehat{\beta}_S$ in the first block (30). Just as we did in the work on basic properties of the lasso estimator, this yields

$$\widehat{\beta}_S = (X_S^T X_S)^{-1}\big(X_S^T y - \lambda \text{sign}(\beta_{0,S})\big), \tag{32}$$

where we have substituted $s_S = \text{sign}(\beta_{0,S})$. From (31), this implies that $s_{-S}$ must satisfy

$$s_{-S} = \frac{1}{\lambda} X_{-S}^T\big(I - X_S(X_S^T X_S)^{-1}X_S^T\big)y + X_{-S}^T X_S(X_S^T X_S)^{-1}\text{sign}(\beta_{0,S}). \tag{33}$$

To lay it out, for concreteness, the primal-dual witness method proceeds as follows:

1. Solve for the lasso solution over the $S$ components, $\widehat{\beta}_S$, as in (32), and set $\widehat{\beta}_{-S} = 0$
2. Solve for the subgradient over the $S^c$ components, $s_{-S}$, as in (33)
3. Check that $\text{sign}(\widehat{\beta}_S) = \text{sign}(\beta_{0,S})$, and that $\|s_{-S}\|_\infty < 1$. If these two checks pass, then we have certified there is a (unique) lasso solution that exactly recovers the true support and signs

The success of the primal-dual witness method hinges on Step 3. We can plug in $y = X\beta_0 + \epsilon$, and rewrite the required conditions, $\text{sign}(\widehat{\beta}_S) = \text{sign}(\beta_{0,S})$ and $\|s_{-S}\|_\infty < 1$, as

$$\text{sign}(\beta_{0,j} + \Delta_j) = \text{sign}(\beta_{0,j}), \text{ where}$$
$$\Delta_j = e_j^T (X_S^T X_S)^{-1} \big( X_S^T \epsilon - \lambda \text{sign}(\beta_{0,S}) \big), \text{ for all } j \in S, \quad (34)$$

and

$$\Big\| \frac{1}{\lambda} X_{-S}^T \big( I - X_S (X_S^T X_S)^{-1} X_S^T \big) \epsilon + X_{-S}^T X_S (X_S^T X_S)^{-1} \text{sign}(\beta_{0,S}) \Big\|_\infty < 1. \quad (35)$$

As $\epsilon \sim N(0, \sigma^2 I)$, we see that the two required conditions have been reduced to statements about Gaussian random variables. The arguments we need to check these conditions actually are quite simply, but we will need to make assumptions on $X$ and $\beta_0$. These are:

With these assumptions in place on $X$ and $\beta_0$, let's first consider verifying (34), and examine $\Delta_S$, whose components $\Delta_j$, $j \in S$ are as defined in (34). We have

$$\|\Delta_S\|_\infty \leq \|(X_S^T X_S)^{-1} X_S^T \epsilon\|_\infty + \lambda \|(X_S^T X_S)^{-1}\|_\infty.$$

Note that $w = (X_S^T X_S)^{-1} X_S^T \epsilon$ is Gaussian with mean zero and covariance $\sigma^2 (X_S^T X_S)^{-1}$, so the variances of components of $w$ are bounded by

$$\sigma^2 \Lambda_{\max} \Big( (X_S^T X_S)^{-1} \Big) \leq \frac{\sigma^2 n}{C},$$

where we have used the minimum eigenvalue assumption. By a standard result on the maximum of Gaussians, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\begin{aligned} \|\Delta_S\|_\infty &\leq \frac{\sigma}{\sqrt{C}} \sqrt{2n \log(es_0/\delta)} + \lambda \|(X_S^T X_S)^{-1}\|_\infty \\ &\leq \beta_{0,\min} + \frac{\gamma}{\sqrt{C}} \underbrace{\left( \frac{\sigma}{\gamma} \sqrt{2n \log(es_0/\delta)} - 4\lambda \right)}_{a}. \end{aligned}$$

where in the second line we used the minimum signal condition. As long as $a < 0$, we can see that the sign condition (34) is verified

Now, let's consider verifying (35). Using the mutual incoherence condition, we have

$$\Big\| \frac{1}{\lambda} X_{-S}^T \big( I - X_S (X_S^T X_S)^{-1} X_S^T \big) \epsilon + X_{-S}^T X_S (X_S^T X_S)^{-1} \text{sign}(\beta_{0,S}) \Big\|_\infty \leq \|z\|_\infty + (1 - \gamma),$$

where $z = (1/\lambda) X_{-S}^T (I - X_S (X_S^T X_S)^{-1} X_S^T) \epsilon = (1/\lambda) X_{-S}^T P_{X_S} \epsilon$, with $P_{X_S}$ the projection matrix onto the column space of $X_S$. Notice that $z$ is Gaussian with mean zero

and covariance $(\sigma^2/\lambda^2)X_{-S}^T P_{X_S} X_{-S}$, so the components of $z$ have variances bounded by

$$\frac{\sigma^2 n}{\lambda^2}\Lambda_{\max}(P_{X_S}) \leq \frac{\sigma^2 n}{\lambda^2}.$$

Therefore, again by the maximal Gaussian inequality, for any $\delta > 0$, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\Big\|\frac{1}{\lambda}X_{-S}^T\big(I - X_S(X_S^TX_S)^{-1}X_S^T\big)\epsilon + X_{-S}^TX_S(X_S^TX_S)^{-1}\mathrm{sign}(\beta_{0,S})\Big\|_\infty & \\ \leq \frac{\sigma}{\lambda}\sqrt{2n\log\left(e(p-s_0)/\delta\right)} + (1-\gamma) & \\ = 1 + \underbrace{\left(\frac{\sigma}{\lambda}\sqrt{2n\log\left(e(p-s_0)/\delta\right)} - \gamma\right)}_{b}, &\end{aligned}$$

Thus as long as $b < 0$, we can see that the subgradient condition (35) is verified

So it remains to choose $\lambda$ so that $a, b < 0$. For $\lambda \geq (2\sigma/\gamma)\sqrt{2n\log(ep/\delta)}$, we can see that

$$a \leq 2\lambda - 4\lambda < 0, \quad b \leq \gamma/2 - \gamma < 0,$$

so (34), (35) are verified—and hence lasso estimator recovers the correct support and signs—with probability at least $1-2\delta$

## 8.1 A note on the conditions

As we moved from the slow rates, to fast rates, to support recovery, the assumptions we used just got stronger and stronger. For the slow rates, we essentially assumed nothing about the predictor matrix $X$ except for column normalization. For the fast rates, we had to additionally assume a compatibility or restricted eigenvalue condition, which roughly speaking, limited the correlations of the predictor variables (particularly concentrated over the underlying support $S$). For support recovery, we still needed whole lot more. The minimum eigenvalue condition on $(1/n)(X_S^TX_S)^{-1}$ is somewhat like the restricted eigenvalue condition on $X$. But the mutual incoherence condition is even stronger; it requires the regression coefficients

$$\eta_j(S) = (X_S^TX_S)^{-1}X_S^TX_j,$$

given by regressing each $X_j$ on the truly active variables $X_S$, to be small (in $\ell_1$ norm) for all $j \notin S$. In other words, no truly inactive variables can be highly correlated (or well-explained, in a linear projection sense) by any of the truly active variables. Finally, this minimum signal condition ensures that the nonzero entries of the true coefficient vector $\beta_0$ are big enough to detect. This is quite restrictive and is not needed for risk bounds, but it is crucial to support recovery.

## 8.2 Minimax bounds

Under the data model (**??**) with $X$ fixed, subject to the scaling $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$, and $\epsilon \sim N(0, \sigma^2)$, Raskutti et al. (2011) derive upper and lower bounds on the minimax prediction error

$$M(s_0, n, p) = \inf_{\widehat{\beta}} \sup_{\|\beta_0\|_0 \leq s_0} \frac{1}{n} \|X\widehat{\beta} - X\beta_0\|_2^2.$$

(Their analysis is acutally considerably more broad than this and covers the coefficient error $\|\widehat{\beta} - \beta_0\|_2$, as well $\ell_q$ constraints on $\beta_0$, for $q \in [0, 1]$.) They prove that, under no additional assumptions on $X$,

$$M(s_0, n, p) \lesssim \frac{s_0 \log(p/s_0)}{n},$$

with probability tending to 1

They also prove that, under a type of restricted eigenvalue condition in which

$$c_0 \leq \frac{(1/n)\|Xv\|_2^2}{\|v\|_2^2} \leq c_1 \text{for all } v \in \mathbb{R}^p \text{ such that } \|v\|_0 \leq 2s_0,$$

for some constants $c_0 > 0$ and $c_1 < \infty$, it holds that

$$M(s_0, n, p) \gtrsim \frac{s_0 \log(p/s_0)}{n},$$

with probability at least $1/2$

The implication is that, for some $X$, minimax optimal prediction may be able to be performed at a faster rate than $s_0 \log(p/s_0)/n$; but for low correlations, this is the rate we should expect. (This is consistent with the worst-case-$X$ analysis of Foster & George (1994), who actually show the worst-case behavior is attained in the orthogonal $X$ case)