Electronic Journal of Statistics
Vol. 7 (2013) 1456–1490
ISSN: 1935-7524
DOI: 10.1214/13-EJS815

The lasso problem and uniqueness

Ryan J. Tibshirani

Department of Statistics
Baker Hall, Carnegie Mellon University
Pittsburgh, PA 15217
e-mail: ryantibs@cmu.edu

Abstract: The lasso is a popular tool for sparse linear regression, especially for problems in which the number of variables p exceeds the number of observations n. But when $p > n$, the lasso criterion is not strictly convex, and hence it may not have a unique minimizer. An important question is: when is the lasso solution well-defined (unique)? We review results from the literature, which show that if the predictor variables are drawn from a continuous probability distribution, then there is a unique lasso solution with probability one, regardless of the sizes of n and p. We also show that this result extends easily to ℓ_1 penalized minimization problems over a wide range of loss functions.

A second important question is: how can we manage the case of non-uniqueness in lasso solutions? In light of the aforementioned result, this case really only arises when some of the predictor variables are discrete, or when some post-processing has been performed on continuous predictor measurements. Though we certainly cannot claim to provide a complete answer to such a broad question, we do present progress towards understanding some aspects of non-uniqueness. First, we extend the LARS algorithm for computing the lasso solution path to cover the non-unique case, so that this path algorithm works for any predictor matrix. Next, we derive a simple method for computing the component-wise uncertainty in lasso solutions of any given problem instance, based on linear programming. Finally, we review results from the literature on some of the unifying properties of lasso solutions, and also point out particular forms of solutions that have distinctive properties.

AMS 2000 subject classifications: Primary 62J07; secondary 90C46.
Keywords and phrases: Lasso, high-dimensional, uniqueness, LARS.

Received August 2012.

Contents

1. Introduction

We consider ℓ_1 penalized linear regression, also known as the lasso problem (Chen et al., 1998; Tibshirani, 1996). Given a outcome vector $y \in \mathbb{R}^n$, a matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables, and a tuning parameter $\lambda \geq 0$, the lasso estimate can be defined as

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1}$$

The lasso solution is unique when $\operatorname{rank}(X) = p$, because the criterion is strictly convex. But the criterion is not strictly convex when $\operatorname{rank}(X) < p$, and so there can be multiple minimizers of the lasso criterion (emphasized by the element notation in (1)). Note that when the number of variables exceeds the number of observations, $p > n$, we must have $\operatorname{rank}(X) < p$.

The lasso is quite a popular tool for estimating the coefficients in a linear model, especially in the high-dimensional setting, $p > n$. Depending on the value of the tuning parameter λ, solutions of the lasso problem will have many coefficients set exactly to zero, due to the nature of the ℓ_1 penalty. We tend to think of the support set of a lasso solution $\hat{\beta}$, written $\mathcal{A} = \operatorname{supp}(\hat{\beta}) \subseteq \{1, \ldots p\}$ and often referred to as the active set, as describing a particular subset of important variables for the linear model of y on X. Recently, there has been a lot of interesting work legitimizing this claim by proving desirable properties of $\hat{\beta}$ or its active set $\mathcal{A}$, in terms of estimation error or model recovery. Most of this work falls into the setting $p > n$. But such properties are not the focus of the current paper. Instead, our focus somewhat simpler, and at somewhat more of a basic level: we investigate issues concerning the uniqueness or non-uniqueness of lasso solutions.

Let us first take a step back, and consider the usual linear regression estimate (given by $\lambda = 0$ in (1)), as a motivating example. Students of statistics are taught to distrust the coefficients given by linear regression when $p > n$. We may ask:

why? Arguably, the main reason is that the linear regression solution is not unique when $p > n$ (or more precisely, when $\text{rank}(X) < p$), and further, this non-uniqueness occurs in such a way that we can always find a variable $i \in \{1, \ldots p\}$ whose coefficient is positive at one solution and negative at another. (Adding any element of the null space of X to one least squares solution produces another solution.) This makes it generally impossible to interpret the linear regression estimate when $p > n$.

Meanwhile, the lasso estimate is also not unique when $p > n$ (or when $\text{rank}(X) < p$), but it is commonly used in this case, and in practice little attention is paid to uniqueness. Upon reflection, this seems somewhat surprising, because non-uniqueness of solutions can cause major problems in terms of interpretation (as demonstrated by the linear regression case). Two basic questions are:

- Do lasso estimates suffer from the same sign inconsistencies as do linear regression estimates? That is, for a fixed λ, can one lasso solution have a positive ith coefficient, and another have a negative ith coefficient?
- Must any two lasso solutions, at the same value of λ, necessarily share the same support, and differ only in their estimates of the nonzero coefficient values? Or can different lasso solutions exhibit different active sets?

Consider the following example, concerning the second question. Here we let $n = 5$ and $p = 10$. For a particular outcome $y \in \mathbb{R}^5$ and predictor matrix $X \in \mathbb{R}^{5\times 10}$, and $\lambda = 1$, we found two solutions of the lasso problem (1), using two different algorithms. These are

$$\hat{\beta}^{(1)} = (-0.893, 0.620, 0.375, 0.497, \ldots, 0)^T \text{ and}$$
$$\hat{\beta}^{(2)} = (-0.893, 0.869, 0.624, 0, \ldots, 0)^T,$$

where we use ellipses to denote all zeros. In other words, the first solution has support set $\{1, 2, 3, 4\}$, and the second has support set $\{1, 2, 3\}$. This is not at all ideal for the purposes of interpretation, because depending on which algorithm we used to minimize the lasso criterion, we may have considered the 4th variable to be important or not. Moreover, who knows which variables may have zero coefficients at other solutions?

In Section 2, we show that if the entries of the predictor matrix X are drawn from a continuous probability distribution, then we essentially never have to worry about the latter problem—along with the problem of sign inconsistencies, and any other issues relating to non-uniqueness—because the lasso solution is unique with probability one. We emphasize that here uniqueness is ensured with probability one (over the distribution of X) regardless of the sizes of n and p. This result has basically appeared in various forms in the literature, but is perhaps not as well-known as it should be. Section 2 gives a detailed review of why this fact is true.

Therefore, the two questions raised above only need to be addressed in the case that X contains discrete predictors, or contains some kind of post-processed versions of continuously drawn predictor measurements. To put it bluntly (and save any dramatic tension), the answer to the first question is "no". In other

words, no two lasso solutions can attach opposite signed coefficients to the same variable. We show this using a very simple argument in Section 4. As for the second question, the example above already shows that the answer is unfortunately "yes". However, the multiplicity of active sets can be dealt with in a principled manner, as we argue in Section 4. Here we show how to compute lower and upper bounds on the coefficients of lasso solutions of any particular problem instance—this reveals exactly which variables are assigned zero coefficients at some lasso solutions, and which variables have nonzero coefficients at all lasso solutions.

Apart from addressing these two questions, we also attempt to better understand the non-unique case through other means. In Section 3, we extend the well-known LARS algorithm for computing the lasso solution path (over the tuning parameter λ) to cover the non-unique case. Therefore the (newly proposed) LARS algorithm can compute a lasso solution path for any predictor matrix X. (The existing LARS algorithm cannot, because it assumes that for any λ the active variables form a linearly independent set, which is not true in general.) The special lasso solution computed by the LARS algorithm, also called the LARS lasso solution, possesses several interesting properties in the non-unique case. We explore these mainly in Section 3, and to a lesser extent in Section 5. Section 5 contains a few final miscellaneous properties relating to non-uniqueness, and the work of the previous three sections.

In this paper, we both review existing results from the literature, and establish new ones, on the topic of uniqueness of lasso solutions. We do our best to acknowledge existing works in the literature, with citations either immediately preceeding or succeeding the statements of lemmas. The contents of this paper were already discussed above, but this was presented out of order, and hence we give a proper outline here. We begin in Section 2 by examining the KKT optimality conditions for the lasso problem, and we use these to derive sufficient conditions for the uniqueness of the lasso solution. This culminates in a result that says that if the entries of X are continuously distributed, then the lasso solution is unique with probability one. We also show that this same result holds for ℓ_1 penalized minimization problems over a broad class of loss functions. Essentially, the rest of the paper focuses on the case of a non-unique lasso solution. Section 3 presents an extension of the LARS algorithm for the lasso solution path that works for any predictor matrix X (the original LARS algorithm really only applies to the case of a unique solution). We then discuss some special properties of the LARS lasso solution. Section 4 develops a method for computing component-wise lower and upper bounds on lasso coefficients for any given problem instance. In Section 5, we finish with some related properties, concerning the different active sets of lasso solutions, and a necessary condition for uniqueness. Section 6 contains some discussion.

Finally, our notation in the paper is as follows. For a matrix A, we write $\mathrm{col}(A)$, $\mathrm{row}(A)$, and $\mathrm{null}(A)$ to denote its column space, row space, and null space, respectively. We use $\mathrm{rank}(A)$ for the rank of A. We use A^+ to denote the Moore-Penrose pseudoinverse of A, and when A is rectangular, this means $A^+ = (A^T A)^+ A^T$. For a linear subspace L, we write P_L for the projection

map onto L. Suppose that $A \in \mathbb{R}^{n\times p}$ has columns $A_1, \ldots A_p \in \mathbb{R}^n$, written $A = [A_1, \ldots A_p]$. Then for an index set $S = \{i_1, \ldots i_k\} \subseteq \{1, \ldots p\}$, we let $A_S = [A_{i_1}, \ldots A_{i_k}]$; in other words, A_S extracts the columns of A in S. Similarly, for a vector $b \in \mathbb{R}^p$, we let $b_S = (b_{i_1}, \ldots b_{i_r})^T$, or in other words, b_S extracts the components of b in S. We write A_{-S} or b_{-S} to extract the columns or components not in S.

2. When is the lasso solution unique?

In this section, we review the question: when is the lasso solution unique? In truth, we only give a partial answer, because we provide sufficient conditions for a unique minimizer of the lasso criterion. Later, in Section 5, we study the other direction (a necessary condition for uniqueness).

2.1. Basic facts and the KKT conditions

We begin by recalling a few basic facts about lasso solutions.

Lemma 1. *For any y, X, and $\lambda \geq 0$, the lasso problem (1) has the following properties:*

(i) There is either a unique lasso solution or an (uncountably) infinite number of solutions.
(ii) Every lasso solution $\hat{\beta}$ gives the same fitted value $X\hat{\beta}$.
(iii) If $\lambda > 0$, then every lasso solution $\hat{\beta}$ has the same ℓ_1 norm, $\|\hat{\beta}\|_1$.

Proof. (i) The lasso criterion is convex and has no directions of recession (strictly speaking, when $\lambda = 0$ the criterion can have directions of recession, but these are directions in which the criterion is constant). Therefore it attains its minimum over $\mathbb{R}^p$ (see, for example, Theorem 27.1 of Rockafellar (1970)), that is, the lasso problem has at least one solution. Suppose now that there are two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$, $\hat{\beta}^{(1)} \neq \hat{\beta}^{(2)}$. Because the solution set of a convex minimization problem is convex, we know that $\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}$ is also a solution for any $0 < \alpha < 1$, which gives uncountably many lasso solutions as α varies over $(0, 1)$.

(ii) Suppose that we have two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ with $X\hat{\beta}^{(1)} \neq X\hat{\beta}^{(2)}$. Let c^* denote the minimum value of the lasso criterion obtained by $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$. For any $0 < \alpha < 1$, we have

$$\frac{1}{2}\|y - X(\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)})\|_2^2 + \lambda\|\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}\|_1 < \alpha c^* + (1-\alpha)c^* = c^*,$$

where the strict inequality is due to the strict convexity of the function $f(x) = \|y - x\|_2^2$ along with the convexity of $f(x) = \|x\|_1$. This means that $\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}$ attains a lower criterion value than c^*, a contradiction.

(iii) By (ii), any two solutions must have the same fitted value, and hence the same squared error loss. But the solutions also attain the same value of the lasso criterion, and if $\lambda > 0$, then they must have the same ℓ_1 norm. □

To go beyond the basics, we turn to the Karush-Kuhn-Tucker (KKT) optimality conditions for the lasso problem (1). These conditions can be written as

$$X^T(y - X\hat{\beta}) = \lambda\gamma, \tag{2}$$

$$\gamma_i \in \begin{cases} \{\text{sign}(\hat{\beta}_i)\} & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}, \quad \text{for } i = 1, \dots p. \tag{3}$$

Here $\gamma \in \mathbb{R}^p$ is called a subgradient of the function $f(x) = \|x\|_1$ evaluated at $x = \hat{\beta}$. Therefore $\hat{\beta}$ is a solution in (1) if and only if $\hat{\beta}$ satisfies (2) and (3) for some γ.

We now use the KKT conditions to write the lasso fit and solutions in a more explicit form. In what follows, we assume that $\lambda > 0$ for the sake of simplicity (dealing with the case $\lambda = 0$ is not difficult, but some of the definitions and statements need to be modified, avoided here in order to preserve readibility). First we define the equicorrelation set $\mathcal{E}$ by

$$\mathcal{E} = \big\{i \in \{1, \dots p\} : |X_i^T(y - X\hat{\beta})| = \lambda\big\}. \tag{4}$$

The equicorrelation set $\mathcal{E}$ is named as such because when y, X have been standardized, $\mathcal{E}$ contains the variables that have equal (and maximal) absolute correlation with the residual. We define the equicorrelation signs s by

$$s = \text{sign}\big(X_{\mathcal{E}}^T(y - X\hat{\beta})\big). \tag{5}$$

Recalling (2), we note that the optimal subgradient γ is unique (by the uniqueness of the fit $X\hat{\beta}$), and we can equivalently define $\mathcal{E}, s$ in terms of γ, as in $\mathcal{E} = \{i \in \{1, \dots p\} : |\gamma_i| = 1\}$ and $s = \gamma_{\mathcal{E}}$. The uniqueness of $X\hat{\beta}$ (or the uniqueness of γ) implies the uniqueness of $\mathcal{E}, s$.

By definition of the subgradient γ in (3), we know that $\hat{\beta}_{-\mathcal{E}} = 0$ for any lasso solution $\hat{\beta}$. Hence the $\mathcal{E}$ block of (2) can be written as

$$X_{\mathcal{E}}^T(y - X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}) = \lambda s. \tag{6}$$

This means that $\lambda s \in \text{row}(X_{\mathcal{E}})$, so $\lambda s = X_{\mathcal{E}}^T(X_{\mathcal{E}}^T)^+\lambda s$. Using this fact, and rearranging (6), we get

$$X_{\mathcal{E}}^T X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}} = X_{\mathcal{E}}^T\big(y - (X_{\mathcal{E}}^T)^+\lambda s\big).$$

Therefore the (unique) lasso fit $X\hat{\beta} = X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}$ is

$$X\hat{\beta} = X_{\mathcal{E}}(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda s\big), \tag{7}$$

and any lasso solution $\hat{\beta}$ is of the form

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_{\mathcal{E}} = (X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda s\big) + b, \tag{8}$$

where $b \in \text{null}(X_\mathcal{E})$. In particular, any $b \in \text{null}(X_\mathcal{E})$ produces a lasso solution $\hat{\beta}$ in (8) provided that $\hat{\beta}$ has the correct signs over its nonzero coefficients, that is, $\text{sign}(\hat{\beta}_i) = s_i$ for all $\hat{\beta}_i \neq 0$. We can write these conditions together as

$$b \in \text{null}(X_\mathcal{E}) \quad \text{and} \quad s_i \cdot \Big(\big[(X_\mathcal{E})^+ \big(y - (X_\mathcal{E}^T)^+ \lambda s \big) \big]_i + b_i \Big) \geq 0 \quad \text{for } i \in \mathcal{E}, \tag{9}$$

and hence any b satisfying (9) gives a lasso solution $\hat{\beta}$ in (8). In the next section, using a sequence of straightforward arguments, we prove that the lasso solution is unique under somewhat general conditions.

2.2. *Sufficient conditions for uniqueness*

From our work in the previous section, we can see that if $\text{null}(X_\mathcal{E}) = \{0\}$, then the lasso solution is unique and is given by (8) with $b = 0$. (We note that $b = 0$ necessarily satisfies the sign condition in (9), because a lasso solution is guaranteed to exist by Lemma 1.) Then by rearranging (8), done to emphasize the rank of $X_\mathcal{E}$, we have the following result.

Lemma 2. *For any y, X, and $\lambda > 0$, if $\text{null}(X_\mathcal{E}) = \{0\}$, or equivalently if $\text{rank}(X_\mathcal{E}) = |\mathcal{E}|$, then the lasso solution is unique, and is given by*

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_\mathcal{E} = (X_\mathcal{E}^T X_\mathcal{E})^{-1} (X_\mathcal{E}^T y - \lambda s), \tag{10}$$

where $\mathcal{E}$ and s are the equicorrelation set and signs as defined in (4) and (5). Note that this solution has at most $\min\{n, p\}$ nonzero components.

This sufficient condition for uniqueness has appeared many times in the literature. For example, see Osborne et al. (2000*b*), Fuchs (2005), Wainwright (2009), Candes and Plan (2009). We will show later in Section 5 that the same condition is actually also necessary, for all almost every $y \in \mathbb{R}^n$.

Note that $\mathcal{E}$ depends on the lasso solution at y, X, λ, and hence the condition $\text{null}(X_\mathcal{E}) = \{0\}$ is somewhat circular. There are more natural conditions, depending on X alone, that imply $\text{null}(X_\mathcal{E}) = \{0\}$. To see this, suppose that $\text{null}(X_\mathcal{E}) \neq \{0\}$; then for some $i \in \mathcal{E}$, we can write

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

where $c_j \in \mathbb{R}$, $j \in \mathcal{E} \setminus \{i\}$. Hence,

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

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

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

or

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

assuming that $\lambda > 0$. Therefore, we have shown that if $\text{null}(X_\mathcal{E}) \neq \{0\}$, then for some $i \in \mathcal{E}$,

$$s_i X_i = \sum_{j\in\mathcal{E}\setminus\{i\}} a_j \cdot s_j X_j,$$

with $\sum_{j\in\mathcal{E}\setminus\{i\}} a_j = 1$, which means that $s_i X_i$ lies in the affine span of $s_j X_j$, $j \in \mathcal{E}\setminus\{i\}$. Note that we can assume without a loss of generality that $\mathcal{E}\setminus\{i\}$ has at most n elements, since otherwise we can simply repeat the above arguments replacing $\mathcal{E}$ by any one of its subsets with $n+1$ elements; hence the affine span of $s_j X_j$, $j \in \mathcal{E}\setminus\{i\}$ is at most $n-1$ dimensional.

We say that the matrix $X \in \mathbb{R}^{n\times p}$ has columns in *general position* if no k-dimensional subspace $L \subseteq \mathbb{R}^n$, for $k < \min\{n, p\}$, contains more than $k+1$ elements of the set $\{\pm X_1, \ldots \pm X_p\}$, excluding antipodal pairs. Another way of saying this: the affine span of any $k+1$ points $\sigma_1 X_{i_1}, \ldots \sigma_{k+1} X_{i_{k+1}}$, for arbitrary signs $\sigma_1, \ldots \sigma_{k+1} \in \{-1, 1\}$, does not contain any element of $\{\pm X_i : i \neq i_1, \ldots i_{k+1}\}$. From what we have just shown, the predictor matrix X having columns in general position is enough to ensure uniqueness.

Lemma 3. *If the columns of X are in general position, then for any y and $\lambda > 0$, the lasso solution is unique and is given by* (10).

This result has also essentially appeared in the literature, taking various forms when stated for various related problems. For example, Rosset et al. (2004) give a similar result for general convex loss functions. Dossal (2012) gives a related result for the noiseless lasso problem (also called basis pursuit). Donoho (2006) gives results tying togther the uniqueness (and equality) of solutions of the noiseless lasso problem and the corresponding ℓ_0 minimization problem.

Although the definition of general position may seem somewhat technical, this condition is naturally satisfied when the entries of the predictor matrix X are drawn from a continuous probability distribution. More precisely, if the entries of X follow a joint distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{np}$, then the columns of X are in general position with probability one. To see this, first consider the probability $\mathrm{P}(X_{k+2} \in \text{aff}\{X_1, \ldots X_{k+1}\})$, where $\text{aff}\{X_1, \ldots X_{k+1}\}$ denotes the affine span of $X_1, \ldots X_{k+1}$. Note that, by continuity,

$$\mathrm{P}(X_{k+2} \in \text{aff}\{X_1, \ldots X_{k+1}\} \,|\, X_1, \ldots X_{k+1}) = 0,$$

because (for fixed $X_1, \ldots X_{k+1}$) the set $\text{aff}\{X_1, \ldots X_{k+1}\} \subseteq \mathbb{R}^n$ has Lebesgue measure zero. Therefore, integrating over $X_1, \ldots X_{k+1}$, we get that $\mathrm{P}(X_{k+2} \in \text{aff}\{X_1, \ldots X_{k+1}\}) = 0$. Taking a union over all subsets of $k+2$ columns, all combinations of $k+2$ signs, and all $k < n$, we conclude that with probability zero the columns are not in general position. This leads us to our final sufficient condition for uniqueness of the lasso solution.

Lemma 4. *If the entries of $X \in \mathbb{R}^{n\times p}$ are drawn from a continuous probability distribution on $\mathbb{R}^{np}$, then for any y and $\lambda > 0$, the lasso solution is unique and is given by (10) with probability one.*

According to this result, we essentially never have to worry about uniqueness when the predictor variables come from a continuous distribution, regardless of the sizes of n and p. Actually, there is nothing really special about ℓ_1 penalized linear regression in particular—we show next that the same uniqueness result holds for ℓ_1 penalized minimization with any differentiable, strictly convex loss function.

2.3. General convex loss functions

We consider the more general minimization problem

$$\hat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} f(X\beta) + \lambda\|\beta\|_1, \tag{11}$$

where the loss function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and strictly convex. To be clear, we mean that f is strictly convex in its argument, so for example the function $f(u) = \|y - u\|_2^2$ is strictly convex, even though $f(X\beta) = \|y - X\beta\|_2^2$ may not be strictly convex in β.

The main ideas from Section 2.1 carry over to this more general problem. The arguments given in the proof of Lemma 1 can be applied (relying on the strict convexity of f) to show that the same set of basic results hold for problem (11): (i) there is either a unique solution or uncountably many solutions;[1] (ii) every solution $\hat{\beta}$ gives the same fit $X\hat{\beta}$; (iii) if $\lambda > 0$, then every solution $\hat{\beta}$ has the same ℓ_1 norm. The KKT conditions for (11) can be expressed as

$$X^T(-\nabla f)(X\hat{\beta}) = \lambda\gamma, \tag{12}$$

$$\gamma_i \in \begin{cases} \{\operatorname{sign}(\hat{\beta}_i)\} & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}, \quad \text{for } i = 1, \ldots p, \tag{13}$$

where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is the gradient of f, and we can define the equicorrelation set and signs in the same way as before,

$$\mathcal{E} = \big\{i \in \{1, \ldots p\} : |X_i^T(-\nabla f)(X\hat{\beta})| = \lambda\big\},$$

and

$$s = \operatorname{sign}\big(X_{\mathcal{E}}^T(-\nabla f)(X\hat{\beta})\big).$$

[1] To be precise, if $\lambda = 0$ then problem (11) may not have a solution for an arbitrary differentiable, strictly convex function f. This is because f may have directions of recession that are not directions in which f is constant, and hence it may not attain its minimal value. For example, the function $f(u) = e^{-u}$ is differentiable and strictly convex on $\mathbb{R}$, but does not attain its minimum. Therefore, for $\lambda = 0$, the statements in this section should all be interpreted as conditional on the existence of a solution in the first place. For $\lambda > 0$, the ℓ_1 penalty gets rid of this issue, as the criterion in (11) has no directions of recession, implying the existence of a solution.

The subgradient condition (13) implies that $\hat{\beta}_{-\mathcal{E}} = 0$ for any solution $\hat{\beta}$ in (11). For squared error loss, recall that we then explicitly solved for $\hat{\beta}_{\mathcal{E}}$ as a function of $\mathcal{E}$ and s. This is not possible for a general loss function f; but given $\mathcal{E}$ and s, we can rewrite the minimization problem (11) over the coordinates in $\mathcal{E}$ as

$$\hat{\beta}_{\mathcal{E}} \in \operatorname*{argmin}_{\beta_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}} f(X_{\mathcal{E}}\beta_{\mathcal{E}}) + \lambda \|\beta_{\mathcal{E}}\|_1. \tag{14}$$

Now, if $\text{null}(X_{\mathcal{E}}) = \{0\}$ (equivalently $\text{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$), then the criterion in (14) is strictly convex, as f itself is strictly convex. This implies that there is a unique solution $\hat{\beta}_{\mathcal{E}}$ in (14), and therefore a unique solution $\hat{\beta}$ in (11). Hence, we arrive at the same conclusions as those made in Section 2.2, that there is a unique solution in (11) if the columns of X are in general position, and ultimately, the following result.

Lemma 5. *If $X \in \mathbb{R}^{n\times p}$ has entries drawn from a continuous probability distribution on $\mathbb{R}^{np}$, then for any differentiable, strictly convex function f, and for any $\lambda > 0$, the minimization problem (11) has a unique solution with probability one. This solution has at most $\min\{n, p\}$ nonzero components.*

This general result applies to any differentiable, strictly convex loss function f, which is quite a broad class. For example, it applies to logistic regression loss,

$$f(u) = \sum_{i=1}^{n} \big[-y_i u_i + \log\big(1 + \exp(u_i)\big)\big],$$

where typically (but not necessarily) each $y_i \in \{0, 1\}$, and Poisson regression loss,

$$f(u) = \sum_{i=1}^{n} \big[-y_i u_i + \exp(u_i)\big],$$

where typically (but again, not necessarily) each $y_i \in \mathbb{N} = \{0, 1, 2, \ldots\}$.

We shift our focus in the next section, and without assuming any conditions for uniqueness, we show how to compute a solution path for the lasso problem (over the regularization parameter λ).

3. The LARS algorithm for the lasso path

The LARS algorithm is a great tool for understanding the behavior of lasso solutions. (To be clear, here and throughout we use the term "LARS algorithm" to refer to the version of the algorithm that computes the lasso solution path, and not the version that performs a special kind of forward variable selection.) The algorithm begins at $\lambda = \infty$, where the lasso solution is trivially $0 \in \mathbb{R}^p$. Then, as the parameter λ decreases, it computes a solution path $\hat{\beta}^{\text{LARS}}(\lambda)$ that is piecewise linear and continuous as a function of λ. Each knot in this path corresponds to an iteration of the algorithm, in which the path's linear trajectory is altered in order to satisfy the KKT optimality conditions. The LARS algorithm

was proposed (and named) by Efron et al. (2004), though essentially the same idea appeared earlier in the works of Osborne et al. (2000*a*) and Osborne et al. (2000*b*). It is worth noting that the LARS algorithm (as proposed in any of these works) assumes that $\operatorname{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$ throughout the lasso path. This is not necessarily correct when $\operatorname{rank}(X) < p$, and can lead to errors in computing lasso solutions. (However, from what we showed in Section 2, this "naive" assumption is indeed correct with probability one when the predictors are drawn from a continuous distribution, and this is likely the reason why such a small oversight has gone unnoticed since the time of the original publications.)

In this section, we extend the LARS algorithm to cover a generic predictor matrix X.[2] Though the lasso solution is not necessarily unique in this general case, and we may have $\operatorname{rank}(X_{\mathcal{E}}) < |\mathcal{E}|$ at some points along path, we show that a piecewise linear and continuous path of solutions still exists, and computing this path requires only a simple modification to the previously proposed LARS algorithm. We describe the algorithm and its steps in detail, but delay the proof of its correctness until Appendix A.1. We also present a few properties of this algorithm and the solutions along its path.

3.1. Description of the LARS algorithm

We start with an overview of the LARS algorithm to compute the lasso path (extended to cover an arbitrary predictor matrix X), and then we describe its steps in detail at a general iteration k. The algorithm presented here is of course very similar to the original LARS algorithm of Efron et al. (2004). The key difference is the following: if $X_{\mathcal{E}}^T X_{\mathcal{E}}$ is singular, then the KKT conditions over the variables in $\mathcal{E}$ no longer have a unique solution, and the current algorithm uses the solution with the minimum ℓ_2 norm, as in (15) and (16). This seemingly minor detail is the basis for the algorithm's correctness in the general X case.

Algorithm 1 (**The LARS algorithm for the lasso path**).

Given y and X.

- *Start with the iteration counter $k = 0$, regularization parameter $\lambda_0 = \infty$, equicorrelation set $\mathcal{E} = \emptyset$, and equicorrelation signs $s = \emptyset$.*
- *While $\lambda_k > 0$:*
  1. *Compute the LARS lasso solution at λ_k by least squares, as in (15) (or (16)). Continue in a linear direction from the solution for $\lambda \le \lambda_k$.*
  2. *Compute the next joining time $\lambda_{k+1}^{\text{join}}$, when a variable outside the equicorrelation set achieves the maximal absolute inner product with the residual, as in (17) and (18).*
  3. *Compute the next crossing time $\lambda_{k+1}^{\text{cross}}$, when the coefficient path of an equicorrelation variable crosses through zero, as in (19) and (20).*

[2]The description of this algorithm and its proof of correctness previously appeared in Appendix B of the author's doctoral dissertation (Tibshirani, 2011).

4. *Set* $\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{join}}, \lambda_{k+1}^{\text{cross}}\}$. *If* $\lambda_{k+1}^{\text{join}} > \lambda_{k+1}^{\text{cross}}$, *then add the joining variable to* $\mathcal{E}$ *and its sign to* s; *otherwise, remove the crossing variable from* $\mathcal{E}$ *and its sign from* s. *Update* $k = k + 1$.

At the start of the kth iteration, the regularization parameter is $\lambda = \lambda_k$. For the path's solution at λ_k, we set the non-equicorrelation coefficients equal to zero, $\hat{\beta}_{-\mathcal{E}}^{\text{LARS}}(\lambda_k) = 0$, and we compute the equicorrelation coefficients as

$$\hat{\beta}_{\mathcal{E}}^{\text{LARS}}(\lambda_k) = (X_{\mathcal{E}})^{+}\big(y - (X_{\mathcal{E}}^T)^{+}\lambda_k s\big) = c - \lambda_k d, \tag{15}$$

where $c = (X_{\mathcal{E}})^{+}y$ and $d = (X_{\mathcal{E}})^{+}(X_{\mathcal{E}}^T)^{+}s = (X_{\mathcal{E}}^T X_{\mathcal{E}})^{+}s$ are defined to help emphasize that this is a linear function of the regularization parameter. This estimate can be viewed as the minimum ℓ_2 norm solution of a least squares problem on the variables in $\mathcal{E}$ (in which we consider $\mathcal{E}, s$ as fixed):

$$\hat{\beta}_{\mathcal{E}}^{\text{LARS}}(\lambda_k) = \operatorname{argmin}\left\{\|\hat{\beta}_{\mathcal{E}}\|_2 : \hat{\beta}_{\mathcal{E}} \in \operatorname*{argmin}_{\beta_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}} \|y - (X_{\mathcal{E}}^T)^{+}\lambda_k s - X_{\mathcal{E}}\beta_{\mathcal{E}}\|_2^2\right\}. \tag{16}$$

Now we decrease λ, keeping $\hat{\beta}_{-\mathcal{E}}^{\text{LARS}}(\lambda) = 0$, and letting

$$\hat{\beta}_{\mathcal{E}}^{\text{LARS}}(\lambda) = c - \lambda d,$$

that is, moving in the linear direction suggested by (15). As λ decreases, we make two important checks. First, we check when (that is, we compute the value of λ at which) a variable outside the equicorrelation set $\mathcal{E}$ should join $\mathcal{E}$ because it attains the maximal absolute inner product with the residual—we call this the next joining time $\lambda_{k+1}^{\text{join}}$. Second, we check when a variable in $\mathcal{E}$ will have a coefficient path crossing through zero—we call this the next crossing time $\lambda_{k+1}^{\text{cross}}$.

For the first check, for each $i \notin \mathcal{E}$, we solve the equation

$$X_i^T\big(y - X_{\mathcal{E}}(c - \lambda d)\big) = \pm\lambda.$$

A simple calculation shows that the solution is

$$t_i^{\text{join}} = \frac{X_i^T(y - X_{\mathcal{E}}c)}{\pm 1 - X_i^T X_{\mathcal{E}}d} = \frac{X_i^T\big(I - X_{\mathcal{E}}(X_{\mathcal{E}})^{+}\big)y}{\pm 1 - X_i^T(X_{\mathcal{E}}^T)^{+}s}, \tag{17}$$

called the joining time of the ith variable. (Although the notation is ambiguous, the quantity t_i^{join} is uniquely defined, as only one of $+1$ or -1 above will yield a value in the interval $[0, \lambda_k]$.[3]) Hence the next joining time is

$$\lambda_{k+1}^{\text{join}} = \max_{i \notin \mathcal{E}} t_i^{\text{join}}, \tag{18}$$

and the joining coordinate and its sign are

$$i_{k+1}^{\text{join}} = \operatorname*{argmax}_{i \notin \mathcal{E}} t_i^{\text{join}} \quad \text{and} \quad s_{k+1}^{\text{join}} = \operatorname{sign}\Big(X_{i_{k+1}^{\text{join}}}^T\big\{y - X\hat{\beta}^{\text{LARS}}(\lambda_{k+1}^{\text{join}})\big\}\Big).$$

[3] If i corresponds to the variable that left the equicorrelation set in the last iteration, then the value of ± 1 here is determined here by the sign opposite to that of its own coefficient.

As for the second check, note that a variable $i \in \mathcal{E}$ will have a zero coefficient when $\lambda = c_i/d_i = [(X_\mathcal{E})^+ y]_i / [(X_\mathcal{E}^T X_\mathcal{E})^+ s]_i$. Because we are only considering $\lambda \leq \lambda_k$, we define the crossing time of the ith variable as

$$t_i^{\text{cross}} = \frac{[(X_\mathcal{E})^+ y]_i}{[(X_\mathcal{E}^T X_\mathcal{E})^+ s]_i} \cdot 1\left\{ \frac{[(X_\mathcal{E})^+ y]_i}{[(X_\mathcal{E}^T X_\mathcal{E})^+ s]_i} \leq \lambda_k \right\}. \tag{19}$$

The next crossing time is therefore

$$\lambda_{k+1}^{\text{cross}} = \max_{i \in \mathcal{E}} t_i^{\text{cross}}, \tag{20}$$

and the crossing coordinate and its sign are

$$i_{k+1}^{\text{cross}} = \operatorname*{argmax}_{i \in \mathcal{E}} t_i^{\text{cross}} \quad \text{and} \quad s_{k+1}^{\text{cross}} = s_{i_{k+1}^{\text{cross}}}.$$

Finally, we decrease λ until the next joining time or crossing time—whichever happens first—by setting $\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{join}}, \lambda_{k+1}^{\text{cross}}\}$. If $\lambda_{k+1}^{\text{join}} > \lambda_{k+1}^{\text{cross}}$, then we add the joining coordinate i_{k+1}^{join} to $\mathcal{E}$ and its sign s_{k+1}^{join} to s. Otherwise, we delete the crossing coordinate i_{k+1}^{cross} from $\mathcal{E}$ and its sign s_{k+1}^{cross} from s.

The proof of correctness for this algorithm shows that computed path $\hat{\beta}^{\text{LARS}}(\lambda)$ satisfies the KKT conditions (2) and (3) at each λ, and is hence indeed a lasso solution path. It also shows that the computed path is continuous at each knot in the path λ_k, and hence is globally continuous in λ. The fact that $X_\mathcal{E}^T X_\mathcal{E}$ can be singular makes the proof somewhat complicated (at least more so than it is for the case $\operatorname{rank}(X) = p$), and hence we delay its presentation until Appendix A.1. Appendix A.2 contains more details on the joining times and crossing times.

3.2. Properties of the LARS algorithm and its solutions

Two basic properties of the LARS lasso path, as mentioned in the previous section, are piecewise linearity and continuity with respect to λ. The algorithm and the solutions along its computed path possess a few other nice properties, most of them discussed in this section, and some others later in Section 5. We begin with a property of the LARS algorithm itself.

Lemma 6. *For any y, X, the LARS algorithm for the lasso path performs at most*

$$\sum_{k=0}^{p} \binom{p}{k} 2^k = 3^p$$

iterations before termination.

Proof. The idea behind the proof is quite simple, and was first noticed by Osborne et al. (2000*a*) for their homotopy algorithm: any given pair of equicorrelation set $\mathcal{E}$ and sign vector s that appear in one iteration of the algorithm cannot be revisited in a future iteration, due to the linear nature of the solution

path. To elaborate, suppose that $\mathcal{E}, s$ were the equicorrelation set and signs at iteration k and also at iteration k', with $k' > k$. Then this would imply that the constraints

$$\left|X_i^T\left(y - X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}^{\mathrm{LARS}}(\lambda)\right)\right| < \lambda \quad \text{for each } i \notin \mathcal{E}, \tag{21}$$

$$s_i \cdot \hat{\beta}_{\mathcal{E}}^{\mathrm{LARS}}(\lambda) > 0 \quad \text{for each } i \in \mathcal{E}, \tag{22}$$

hold at both $\lambda = \lambda_k$ and $\lambda = \lambda_{k'}$. But $\hat{\beta}_{\mathcal{E}}^{\mathrm{LARS}}(\lambda) = c - \lambda d$ is a linear function of λ, and this implies that (21) and (22) also hold at every $\lambda \in [\lambda'_k, \lambda_k]$, contradicting the fact that k' and k are distinct iterations. Therefore the total number of iterations performed by the LARS algorithm is bounded by the number of distinct pairs of subsets $\mathcal{E} \subseteq \{1, \ldots p\}$ and sign vectors $s \in \{-1, 1\}^{|\mathcal{E}|}$. □

Remark. Mairal and Yu (2012) showed recently that the upper bound for the number of steps taken by the original LARS algorithm, which assumes that $\mathrm{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$ throughout the path, can actually be improved to $(3^p + 1)/2$. Their proof is based on the following observation: if $\mathcal{E}, s$ are the equicorrelation set and signs at one iteration of the algorithm, then $\mathcal{E}, -s$ cannot appear as the equicorrelation set and signs in a future iteration. Indeed, this same observation is true for the extended version of LARS presented here, by essentially the same arguments. Hence the upper bound in Lemma 6 can also be improved to $(3^p + 1)/2$. Interestingly, Mairal and Yu (2012) further show that this upper bound is tight: they construct, for any p, a problem instance (y and X) for which the LARS algorithm takes exactly $(3^p + 1)/2$ steps.

Next, we show that the end of the LARS lasso solution path ($\lambda = 0$) is itself an interesting least squares solution.

Lemma 7. *For any y, X, the LARS lasso solution converges to a minimum ℓ_1 norm least squares solution as $\lambda \to 0^+$, that is,*

$$\lim_{\lambda \to 0^+} \hat{\beta}^{\mathrm{LARS}}(\lambda) = \hat{\beta}^{\mathrm{LS},\ell_1},$$

where $\hat{\beta}^{\mathrm{LS},\ell_1} \in \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2$ and achieves the minimum ℓ_1 norm over all such solutions.

Proof. First note that by Lemma 6, the algorithm always takes a finite number of iterations before terminating, so the limit here is always attained by the algorithm (at its last iteration). Therefore we can write $\hat{\beta}^{\mathrm{LARS}}(0) = \lim_{\lambda \to 0^+} \hat{\beta}^{\mathrm{LARS}}(\lambda)$. Now, by construction, the LARS lasso solution satisfies

$$\left|X_i^T\left(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda)\right)\right| \leq \lambda \quad \text{for each } i = 1, \ldots p,$$

at each $\lambda \in [0, \infty]$. Hence at $\lambda = 0$ we have

$$X_i^T\left(y - X\hat{\beta}^{\mathrm{LARS}}(0)\right) = 0 \quad \text{for each } i = 1, \ldots p,$$

implying that $\hat{\beta}^{\mathrm{LARS}}(0)$ is a least squares solution. Suppose that there exists another least squares solution $\hat{\beta}^{\mathrm{LS}}$ with $\|\hat{\beta}^{\mathrm{LS}}\|_1 < \|\hat{\beta}^{\mathrm{LARS}}(0)\|_1$. Then by continuity of the LARS lasso solution path, there exists some $\lambda > 0$ such that still $\|\hat{\beta}^{\mathrm{LS}}\|_1 < \|\hat{\beta}^{\mathrm{LARS}}(\lambda)\|_1$, so that

$$\frac{1}{2}\|y - X\hat{\beta}^{\mathrm{LS}}\|_2^2 + \lambda\|\hat{\beta}^{\mathrm{LS}}\|_1 < \frac{1}{2}\|y - X\hat{\beta}^{\mathrm{LARS}}(\lambda)\|_2^2 + \lambda\|\hat{\beta}^{\mathrm{LARS}}(\lambda)\|_1.$$

This contradicts the fact that $\hat{\beta}^{\mathrm{LARS}}(\lambda)$ is a lasso solution at λ, and therefore $\hat{\beta}^{\mathrm{LARS}}(0)$ achieves the minimum ℓ_1 norm over all least squares solutions. □

We showed in Section 3.1 that the LARS algorithm constructs the lasso solution

$$\hat{\beta}^{\mathrm{LARS}}_{-\mathcal{E}}(\lambda) = 0 \quad \text{and} \quad \hat{\beta}^{\mathrm{LARS}}_{\mathcal{E}}(\lambda) = (X_{\mathcal{E}})^{+}\big(y - (X_{\mathcal{E}}^T)^{+}\lambda s\big),$$

by decreasing λ from ∞, and continually checking whether it needs to include or exclude variables from the equicorrelation set $\mathcal{E}$. Recall our previous description (8) of the set of lasso solutions at any given λ. In (8), different lasso solutions are formed by choosing different vectors b that satisfy the two conditions given in (9): a null space condition, $b \in \mathrm{null}(X_{\mathcal{E}})$, and a sign condition,

$$s_i \cdot \Big(\big[(X_{\mathcal{E}})^{+}\big(y - (X_{\mathcal{E}}^T)^{+}\lambda s\big)\big]_i + b_i\Big) \geq 0 \quad \text{for } i \in \mathcal{E}.$$

We see that the LARS lasso solution corresponds to the choice $b = 0$. When $\mathrm{rank}(X) = |\mathcal{E}|$, $b = 0$ is the only vector in $\mathrm{null}(X_{\mathcal{E}})$, so it satisfies the above sign condition by necessity (as we know that a lasso solution must exist Lemma 1). On the other hand, when $\mathrm{rank}(X) < |\mathcal{E}|$, it is certainly true that $0 \in \mathrm{null}(X_{\mathcal{E}})$, but it is not at all obvious that the sign condition is satisfied by $b = 0$. The LARS algorithm establishes this fact by constructing an entire lasso solution path with exactly this property ($b = 0$) over $\lambda \in [0, \infty]$. At the risk of sounding repetitious, we state this result next in the form of a lemma.

Lemma 8. *For any y, X, and $\lambda > 0$, a lasso solution is given by*

$$\hat{\beta}^{\mathrm{LARS}}_{-\mathcal{E}} = 0 \quad \textit{and} \quad \hat{\beta}^{\mathrm{LARS}}_{\mathcal{E}} = (X_{\mathcal{E}})^{+}\big(y - (X_{\mathcal{E}}^T)^{+}\lambda s\big), \tag{23}$$

and this is the solution computed by the LARS lasso path algorithm.

For one, this lemma is perhaps interesting from a computational point of view: it says that for any y, X, and $\lambda > 0$, a lasso solution (indeed, the LARS lasso solution) can be computed directly from $\mathcal{E}$ and s, which themselves can be computed from the unique lasso fit. Further, for any y, X, we can start with a lasso solution at $\lambda > 0$ and compute a local solution path using the same LARS steps; see Appendix A.3 for more details. Aside from computational interests, the explicit form of a lasso solution given by Lemma 8 may be helpful for the purposes of mathematical analysis; for example, this form is used by Tibshirani and Taylor (2012) to give a simpler proof of the degrees of freedom of the lasso

fit, for a general X, in terms of the equicorrelation set. As another example, it is also used in Section 5 to prove a necessary condition for the uniqueness of the lasso solution (holding almost everywhere in y).

We show in Section 5 that, for almost every $y \in \mathbb{R}^n$, the LARS lasso solution is supported on all of $\mathcal{E}$ and hence has the largest support of any lasso solution (at the same y, X, λ). As lasso solutions all have the same ℓ_1 norm, by Lemma 1, this means that the LARS lasso solution spreads out the common ℓ_1 norm over the largest number of coefficients. It may not be surprising, then, that the LARS lasso solution has the smallest ℓ_2 norm among lasso solutions, shown next.

Lemma 9. *For any y, X, and $\lambda > 0$, the LARS lasso solution $\hat{\beta}^{\mathrm{LARS}}$ has the minimum ℓ_2 norm over all lasso solutions.*

Proof. From (8), we can see that any lasso solution has squared ℓ_2 norm

$$\|\hat{\beta}\|_2^2 = \big\|(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda s\big)\big\|_2^2 + \|b\|_2^2,$$

since $b \in \mathrm{null}(X_{\mathcal{E}})$. Hence $\|\hat{\beta}\|_2^2 \geq \|\hat{\beta}^{\mathrm{LARS}}\|_2^2$, with equality if and only if $b = 0$. □

Mixing together the ℓ_1 and ℓ_2 norms brings to mind the elastic net (Zou and Hastie, 2005), which penalizes both the ℓ_1 norm and the squared ℓ_2 norm of the coefficient vector. The elastic net utilizes two tuning parameters $\lambda_1, \lambda_2 \geq 0$ (this notation should not to be confused with the knots in the LARS lasso path), and solves the criterion[4]

$$\hat{\beta}^{\mathrm{EN}} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|_2^2. \tag{24}$$

For any $\lambda_2 > 0$, the elastic net solution $\hat{\beta}^{\mathrm{EN}} = \hat{\beta}^{\mathrm{EN}}(\lambda_1, \lambda_2)$ is unique, since the criterion is strictly convex.

Note that if $\lambda_2 = 0$, then (24) is just the lasso problem. On the other hand, if $\lambda_1 = 0$, then (24) reduces to ridge regression. It is well-known that the ridge regression solution $\hat{\beta}^{\mathrm{ridge}}(\lambda_2) = \hat{\beta}^{\mathrm{EN}}(0, \lambda_2)$ converges to the minimum ℓ_2 norm least squares solution as $\lambda_2 \to 0^+$. Our next result is analogous to this fact: it says that for any fixed $\lambda_1 > 0$, the elastic net solution converges to the minimum ℓ_2 norm lasso solution—that is, the LARS lasso solution—as $\lambda_2 \to 0^+$,

Lemma 10. *Fix any X and $\lambda_1 > 0$. For almost every $y \in \mathbb{R}^n$, the elastic net solution converges to the LARS lasso solution as $\lambda_2 \to 0^+$, that is,*

$$\lim_{\lambda_2 \to 0^+} \hat{\beta}^{\mathrm{EN}}(\lambda_1, \lambda_2) = \hat{\beta}^{\mathrm{LARS}}(\lambda_1).$$

[4]This is actually what Zou and Hastie (2005) call the "naive" elastic net solution, and the modification $(1 + \lambda_2)\hat{\beta}^{\mathrm{EN}}$ is what the authors refer to as the elastic net estimate. But in the limit as $\lambda_2 \to 0^+$, these two estimates are equivalent, so our result in Lemma 10 holds for this modified estimate as well.

Proof. By Lemma 13, we know that for any $y \notin \mathcal{N}$, where $\mathcal{N} \subseteq \mathbb{R}^n$ is a set of measure zero, the LARS lasso at λ_1 satisfies $\hat{\beta}^{\mathrm{LARS}}(\lambda_1)_i \neq 0$ for all $i \in \mathcal{E}$. Hence fix $y \notin \mathcal{N}$. First note that we can rewrite the LARS lasso solution as

$$\hat{\beta}^{\mathrm{LARS}}_{-\mathcal{E}}(\lambda_1) = 0 \quad \text{and} \quad \hat{\beta}^{\mathrm{LARS}}_{\mathcal{E}}(\lambda_1) = (X_{\mathcal{E}}^T X_{\mathcal{E}})^+ (X_{\mathcal{E}}^T y - \lambda_1 s).$$

Define the function

$$\begin{aligned} f(\lambda_2) &= (X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)^{-1}(X_{\mathcal{E}}^T y - \lambda_1 s) \quad \text{for } \lambda_2 > 0, \\ f(0) &= (X_{\mathcal{E}}^T X_{\mathcal{E}})^+ (X_{\mathcal{E}}^T y - \lambda_1 s). \end{aligned}$$

For fixed $\mathcal{E}, s$, the function f is continuous on $[0, \infty)$ (continuity at 0 can be verified, for example, by looking at the singular value decomposition of $(X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)^{-1}$.) Hence it suffices to show that for small enough $\lambda_2 > 0$, the elastic net solution at λ_1, λ_2 is given by

$$\hat{\beta}^{\mathrm{EN}}_{-\mathcal{E}}(\lambda_1, \lambda_2) = 0 \quad \text{and} \quad \hat{\beta}^{\mathrm{EN}}_{\mathcal{E}}(\lambda_1, \lambda_2) = f(\lambda_2).$$

To this end, we show that the above proposed solution satisfies the KKT conditions for small enough λ_2. The KKT conditions for the elastic net problem are

$$X^T(y - X\hat{\beta}^{\mathrm{EN}}) - \lambda_2 \hat{\beta}^{\mathrm{EN}} = \lambda_1 \gamma, \tag{25}$$

$$\gamma_i \in \begin{cases} \{\mathrm{sign}(\hat{\beta}_i^{\mathrm{EN}})\} & \text{if } \hat{\beta}_i^{\mathrm{EN}} \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i^{\mathrm{EN}} = 0 \end{cases}, \quad \text{for } i = 1, \ldots p. \tag{26}$$

Recall that $f(0) = \hat{\beta}^{\mathrm{LARS}}_{\mathcal{E}}(\lambda_1)$ are the equicorrelation coefficients of the LARS lasso solution at λ_1. As $y \notin \mathcal{N}$, we have $f(0)_i \neq 0$ for each $i \in \mathcal{E}$, and further, $\mathrm{sign}(f(0)_i) = s_i$ for all $i \in \mathcal{E}$. Therefore the continuity of f implies that for small enough λ_2, $f(\lambda_2)_i \neq 0$ and $\mathrm{sign}(f(\lambda_2)_i) = s_i$ for all $i \in \mathcal{E}$. Also, we know that $\|X_{-\mathcal{E}}^T(y - X_{\mathcal{E}} f(0))\|_\infty < \lambda_1$ by definition of the equicorrelation set $\mathcal{E}$, and again, the continuity of f implies that for small enough λ_2, $\|X_{-\mathcal{E}}^T(y - X_{\mathcal{E}} f(\lambda_2))\|_\infty < \lambda_1$. Finally, direct calculcation shows that

$$\begin{aligned} X_{\mathcal{E}}^T\big(y - X_{\mathcal{E}} f(\lambda_2)\big) - \lambda_2 f(\lambda_2) &= X_{\mathcal{E}}^T y - (X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)(X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)^{-1} X_{\mathcal{E}}^T y \\ &\quad + (X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)(X_{\mathcal{E}}^T X_{\mathcal{E}} + \lambda_2 I)^{-1} \lambda_1 s \\ &= \lambda_1 s. \end{aligned}$$

This verifies the KKT conditions for small enough λ_2, and completes the proof. □

In Section 5, we discuss a few more properties of LARS lasso solutions, in the context of studying the various support sets of lasso solutions. In the next section, we present a simple method for computing lower and upper bounds on the coefficients of lasso solutions, useful when the solution is not unique.

4. Lasso coefficient bounds

Here we again consider a general predictor matrix X (not necessarily having columns in general position), so that the lasso solution is not necessarily unique. We show that it is possible to compute lower and upper bounds on the coefficients of lasso solutions, for any given problem instance, using linear programming. We begin by revisiting the KKT conditions.

4.1. Back to the KKT conditions

The KKT conditions for the lasso problem were given in (2) and (3). Recall that the lasso fit $X\hat{\beta}$ is always unique, by Lemma 1. Note that when $\lambda > 0$, we can rewrite (2) as

$$\gamma = \frac{1}{\lambda} X^T (y - X\hat{\beta}),$$

implying that the optimal subgradient γ is itself unique. According to its definition (3), the components of γ give the signs of nonzero coefficients of any lasso solution, and therefore the uniqueness of γ immediately implies the following result.

Lemma 11. *For any y, X, and $\lambda > 0$, any two lasso solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ must satisfy $\hat{\beta}_i^{(1)} \cdot \hat{\beta}_i^{(2)} \geq 0$ for $i = 1, \ldots p$. In other words, any two lasso solutions must have the same signs over their common support.*

In a sense, this result is reassuring—it says that even when the lasso solution is not necessarily unique, lasso coefficients must maintain consistent signs. Note that the same is certainly not true of least squares solutions (corresponding to $\lambda = 0$), which causes problems for interpretation, as mentioned in the introduction. Lemma 11 will be helpful when we derive lasso coefficient bounds shortly.

We also saw in the introduction that different lasso solutions (at the same y, X, λ) can have different supports, or active sets. The previously derived characterization of lasso solutions, given in (8) and (9), provides an understanding of how this is possible. It helps to rewrite (8) and (9) as

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_{\mathcal{E}} = \hat{\beta}_{\mathcal{E}}^{\text{LARS}} + b, \tag{27}$$

where b is subject to

$$b \in \text{null}(X_{\mathcal{E}}) \quad \text{and} \quad s_i \cdot (\hat{\beta}_i^{\text{LARS}} + b_i) \geq 0, \;\; i \in \mathcal{E}, \tag{28}$$

and $\hat{\beta}^{\text{LARS}}$ is the fundamental solution traced by the LARS algorithm, as given in (23). Hence for for a lasso solution $\hat{\beta}$ to have an active set $\mathcal{A} = \text{supp}(\hat{\beta})$, we can see that we must have $\mathcal{A} \subseteq \mathcal{E}$ and $\hat{\beta}_{\mathcal{E}} = \hat{\beta}_{\mathcal{E}}^{\text{LARS}} + b$, where b satisfies (28) and also

$$\begin{aligned} b_i &= -\hat{\beta}_i^{\text{LARS}} \quad \text{for } i \notin \mathcal{E} \setminus \mathcal{A}, \\ b_i &\neq -\hat{\beta}_i^{\text{LARS}} \quad \text{for } i \in \mathcal{E} \setminus \mathcal{A}. \end{aligned}$$

As we discussed in the introduction, the fact that there may be different active sets corresponding to different lasso solutions (at the same y, X, λ) is perhaps concerning, because different active sets provide different "stories" regarding which predictor variables are important. One might ask: given a specific variable of interest $i \in \mathcal{E}$ (recalling that all variables outside of $\mathcal{E}$ necessarily have zero coefficients), is it possible for the ith coefficient to be nonzero at one lasso solution but zero at another? The answer to this question depends on the interplay between the constraints in (28), and as we show next, it is achieved by solving a simple linear program.

4.2. The polytope of solutions and lasso coefficient bounds

The key observation here is that the set of lasso solutions defined by (27) and (28) forms a convex polytope. Consider writing the set of lasso solutions as

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_{\mathcal{E}} \in K = \{x \in \mathbb{R}^{|\mathcal{E}|} : Px = \hat{\beta}_{\mathcal{E}}^{\text{LARS}},\, Sx \geq 0\}, \tag{29}$$

where $P = P_{\text{row}(X_{\mathcal{E}})}$ and $S = \text{diag}(s)$. That (29) is equivalent to (27) and (28) follows from the fact that $\hat{\beta}_{\mathcal{E}}^{\text{LARS}} \in \text{row}(X_{\mathcal{E}})$, hence $Px = \hat{\beta}_{\mathcal{E}}^{\text{LARS}}$ if and only if $x = \hat{\beta}_{\mathcal{E}}^{\text{LARS}} + b$ for some $b \in \text{null}(X_{\mathcal{E}})$.

The set $K \subseteq \mathbb{R}^{|\mathcal{E}|}$ is a polyhedron, since it is defined by linear equalities and inequalities, and furthermore it is bounded, as all lasso solutions have the same ℓ_1 norm by Lemma 1, making it a polytope. The component-wise extrema of K can be easily computed via linear programming. In other words, for $i \in \mathcal{E}$, we can solve the following two linear programs:

$$\hat{\beta}_i^{\text{lower}} = \min_{x \in \mathbb{R}^{|\mathcal{E}|}} x_i \;\text{ subject to }\; Px = \hat{\beta}_{\mathcal{E}}^{\text{LARS}},\, Sx \geq 0, \tag{30}$$

$$\hat{\beta}_i^{\text{upper}} = \max_{x \in \mathbb{R}^{|\mathcal{E}|}} x_i \;\text{ subject to }\; Px = \hat{\beta}_{\mathcal{E}}^{\text{LARS}},\, Sx \geq 0, \tag{31}$$

and then we know that the ith component of any lasso solution satisfies $\hat{\beta}_i \in [\hat{\beta}_i^{\text{lower}}, \hat{\beta}_i^{\text{upper}}]$. These bounds are tight, in the sense that each is achieved by the ith component of some lasso solution (in fact, this solution is just the minimizer of (30), or the maximizer of (31)). By the convexity of K, every value between $\hat{\beta}_i^{\text{lower}}$ and $\hat{\beta}_i^{\text{upper}}$ is also achieved by the ith component of some lasso solution. Most importantly, the linear programs (30) and (31) can actually be solved in practice. Aside from the obvious dependence on y, X, and λ, the relevant quantities P, S, and $\hat{\beta}_{\mathcal{E}}^{\text{LARS}}$ only depend on the equicorrelation set $\mathcal{E}$ and signs s, which in turn only depend on the unique lasso fit. Therefore, one could compute any lasso solution (at y, X, λ) in order to define $\mathcal{E}, s$, and subsequently P, S and $\hat{\beta}_{\mathcal{E}}^{\text{LARS}}$, all that is needed in order to solve (30) and (31). We summarize this idea below.

Algorithm 2 (**Lasso coefficient bounds**).

Given y, X, and $\lambda > 0$.

1. *Compute any solution $\hat{\beta}$ of the lasso problem (at y, X, λ), to obtain the unique lasso fit $X\hat{\beta}$.*

2. *Define the equicorrelation set* $\mathcal{E}$ *and signs* s*, as in* (4) *and* (5)*, respectively.*
3. *Define* $P = P_{\mathrm{row}(X_\mathcal{E})}$*,* $S = \mathrm{diag}(s)$*, and* $\hat{\beta}_\mathcal{E}^{\mathrm{LARS}}$ *according to* (23)*.*
4. *For each* $i \in \mathcal{E}$*, compute the coefficient bounds* $\hat{\beta}_i^{\mathrm{lower}}$ *and* $\hat{\beta}_i^{\mathrm{upper}}$ *by solving the linear programs* (30) *and* (31)*, respectively.*

Lemma 11 implies a valuable property of the bounding interval $[\hat{\beta}_i^{\mathrm{lower}}, \hat{\beta}_i^{\mathrm{upper}}]$, namely, that this interval cannot contain zero in its interior. Otherwise, there would be a pair of lasso solutions with opposite signs over the ith component, contradicting the lemma. Also, we know from Lemma 1 that all lasso solutions have the same ℓ_1 norm L, and this means that $|\hat{\beta}_i^{\mathrm{lower}}|, |\hat{\beta}_i^{\mathrm{upper}}| \leq L$. Combining these two properties gives the next lemma.

Lemma 12. *Fix any* y, X*, and* $\lambda > 0$*. Let* L *be the common* ℓ_1 *norm of lasso solutions at* y, X, λ*. Then for any* $i \in \mathcal{E}$*, the coefficient bounds* $\hat{\beta}_i^{\mathrm{lower}}$ *and* $\hat{\beta}_i^{\mathrm{upper}}$ *defined in* (30) *and* (31) *satisfy*

$$[\hat{\beta}_i^{\mathrm{lower}}, \hat{\beta}_i^{\mathrm{upper}}] \subseteq [0, L] \quad \textit{if } s_i > 0,$$
$$[\hat{\beta}_i^{\mathrm{lower}}, \hat{\beta}_i^{\mathrm{upper}}] \subseteq [-L, 0] \quad \textit{if } s_i < 0.$$

Using Algorithm 2, we can identify all variables $i \in \mathcal{E}$ with one of two categories, based on their bounding intervals:

(i) If $0 \in [\hat{\beta}_i^{\mathrm{lower}}, \hat{\beta}_i^{\mathrm{upper}}]$, then variable i is called *dispensable* (to the lasso model at y, X, λ), because there is a solution that does not include this variable in its active set. By Lemma 12, this can only happen if $\hat{\beta}_i^{\mathrm{lower}} = 0$ or $\hat{\beta}_i^{\mathrm{upper}} = 0$.

(ii) If $0 \notin [\hat{\beta}_i^{\mathrm{lower}}, \hat{\beta}_i^{\mathrm{upper}}]$, then variable i is called *indispensable* (to the lasso model at y, X, λ), because every solution includes this variable in its active set. By Lemma 12, this can only happen if $\hat{\beta}_i^{\mathrm{lower}} > 0$ or $\hat{\beta}_i^{\mathrm{upper}} < 0$.

It is helpful to return to the example discussed in the introduction. Recall that in this example we took $n = 5$ and $p = 10$, and for a given y, X, and $\lambda = 1$, we found two lasso solutions: one supported on variables $\{1, 2, 3, 4\}$, and another supported on variables $\{1, 2, 3\}$. In the introduction, we purposely did not reveal the structure of the predictor matrix X; given what we showed in Section 2 (that X having columns in general position implies a unique lasso solution), it should not be surprising to find out that here we have $X_4 = (X_2 + X_3)/2$. A complete description of our construction of X and y is as follows: we first drew the components of the columns X_1, X_2, X_3 independently from a standard normal distribution, and then defined $X_4 = (X_2 + X_3)/2$. We also drew the components of $X_5, \ldots X_{10}$ independently from a standard normal distribution, and then orthogonalized $X_5, \ldots X_{10}$ with respect to the linear subspace spanned by $X_1, \ldots, X_4$. Finally, we defined $y = -X_1 + X_2 + X_3$. The purpose of this construction was to make it easy to detect the relevant variables $X_1, \ldots X_4$ for the linear model of y on X.

According to the terminology defined above, variable 4 is dispensable to the lasso model when $\lambda = 1$, because it has a nonzero coefficient at one solution

TABLE 1
The results of Algorithm 2 for the small example from the introduction, with $n = 5$, $p = 8$. Shown are the lasso coefficient bounds over the equicorrelation set $\mathcal{E} = \{1, 2, 3, 4\}$

i	$\hat{\beta}_i^{\text{lower}}$	$\hat{\beta}_i^{\text{LARS}}$	$\hat{\beta}_i^{\text{upper}}$
1	−0.8928	−0.8928	−0.8928
2	0.2455	0.6201	0.8687
3	0	0.3746	0.6232
4	0	0.4973	1.2465

but a zero coefficient at another. This is perhaps not surprising, as X_2, X_3, X_4 are linearly dependent. How about the other variables? We ran Algorithm 2 to answer this question. The results are displayed in Table 1.

For the given y, X, and $\lambda = 1$, the equicorrelation set is $\mathcal{E} = \{1, 2, 3, 4\}$, and the sign vector is $s = (-1, 1, 1, 1)^T$ (these are given by running Steps 1 and 2 of Algorithm 2). Therefore we know that any lasso solution has zero coefficients for variables $5, \ldots 10$, has a nonpositive first coefficient, and has nonnegative coefficients for variables $2, 3, 4$. The third column of Table 1 shows the LARS lasso solution over the equicorrelation variables. The second and fourth columns show the component-wise coefficient bounds $\hat{\beta}_i^{\text{lower}}$ and $\hat{\beta}_i^{\text{upper}}$, respectively, for $i \in \mathcal{E}$. We see that variable 3 is dispensable, because it has a lower bound of zero, meaning that there exists a lasso solution that excludes the third variable from its active set (and this solution is actually computed by Algorithm 2, as it is the minimizer of the linear program (30) with $i = 3$). The same conclusion holds for variable 4. On the other hand, variables 1 and 2 are indispensable, because their bounding intervals do not contain zero.

Like variables 3 and 4, variable 2 is linearly dependent on the other variables (in the equicorrelation set), but unlike variables 3 and 4, it is indispensable and hence assigned a nonzero coefficient in every lasso solution. This is the first of a few interesting points about dispensability and indispensability, which we discuss below.

- *Linear dependence does not imply dispensability.* In the example, variable 2 is indispensable, as its coefficient has a lower bound of $0.2455 > 0$, even though variable 2 is a linear function of variables 3 and 4. Note that in order for the 2nd variable to be dispensable, we need to be able to use the others (variables 1, 3, and 4) to achieve both the same fit and the same ℓ_1 norm of the coefficient vector. The fact that variable 2 can be written as a linear function of variables 3 and 4 implies that we can preserve the fit, but not necessarily the ℓ_1 norm, with zero weight on variable 2. Table 1 says that we can make the weight on variable 2 as small as 0.2455 while keeping the fit and the ℓ_1 norm unchanged, but that moving it below 0.2455 (and maintaining the same fit) inflates the ℓ_1 norm.
- *Linear independence implies indispensability (almost everywhere).* In the next section we show that, given any X and λ, and almost every $y \in \mathbb{R}^n$,

the quantity $\mathrm{col}(X_{\mathcal{A}})$ is invariant over all active sets coming from lasso solutions at y, X, λ. Therefore, almost everywhere in y, if variable $i \in \mathcal{E}$ is linearly independent of all $j \notin \mathcal{E}$ (meaning that X_i cannot be expressed as a linear function of X_j, $j \notin \mathcal{E}$), then variable i must be indispensable—otherwise the span of the active variables would be different for different active sets.

- *Individual dispensability does not imply pairwise dispensability.* Back to the above example, variables 3 and 4 are both dispensable, but this does not necessarily mean that there exists a lasso solution that exludes both 3 and 4 simultaneously from the active set. Note that the computed solution that achieves a value of zero for its 3rd coefficient (the minimizer of (30) for $i = 3$) has a nonzero 4th coefficient, and the computed solution that achieves zero for its 4th coefficient (the minimizer of (30) for $i = 4$) has a nonzero 3rd coefficient. While this suggests that variables 3 and 4 cannot simultaneously be zero for the current problem, it does not serve as definitive proof of such a claim. However, we can check this claim by solving (30), with $i = 4$, subject to the additional constraint that $x_3 = 0$. This does in fact yield a positive lower bound, proving that variables 3 and 4 cannot both be zero at a solution. Furthermore, moving beyond pairwise interactions, we can actually enumerate all possible active sets of lasso solutions, by recognizing that there is a one-to-one correspondence between active sets and faces of the polytope K; see Appendix A.4.

Next, we cover some properties of lasso solutions that relate to our work in this section and in the previous two sections, on uniqueness and non-uniqueness.

5. Related properties

We present more properties of lasso solutions, relating to issues of uniqueness and non-uniqueness. The first three sections examine the active sets generated by lasso solutions of a given problem instance, when X is a general predictor matrix. The results in these three sections are reviewed from the literature. In the last section, we give a necessary condition for the uniqueness of the lasso solution.

5.1. The largest active set

For an arbitrary X, recall from Section 4 that the active set $\mathcal{A}$ of any lasso solution is necessarily contained in the equicorrelation set $\mathcal{E}$. We show that the LARS lasso solution has support on all of $\mathcal{E}$, making it the lasso solution with the largest support, for almost every $y \in \mathbb{R}^n$. This result appeared in Tibshirani and Taylor (2012).

Lemma 13. *Fix any X and $\lambda > 0$. For almost every $y \in \mathbb{R}^n$, the LARS lasso solution $\hat{\beta}^{\mathrm{LARS}}$ has an active set $\mathcal{A}$ equal to the equicorrelation set $\mathcal{E}$, and therefore achieves the largest active set of any lasso solution.*

Proof. For a matrix A, let $A_{[i]}$ denote its ith row. Define the set

$$\mathcal{N} = \bigcup_{\mathcal{E},s} \bigcup_{i \in \mathcal{E}} \Big\{ z \in \mathbb{R}^n : \big((X_\mathcal{E})^+\big)_{[i]} \big(z - (X_\mathcal{E}^T)^+ \lambda s\big) = 0 \Big\}. \tag{32}$$

The first union above is taken over all subsets $\mathcal{E} \subseteq \{1, \ldots p\}$ and sign vectors $s \in \{-1, 1\}^{|\mathcal{E}|}$, but implicitly we exclude sets $\mathcal{E}$ such that $(X_\mathcal{E})^+$ has a row that is entirely zero. Then $\mathcal{N}$ has measure zero, because it is a finite union of affine subspaces of dimension $n - 1$.

Now let $y \notin \mathcal{N}$. We know that no row of $(X_\mathcal{E})^+$ can be entirely zero (otherwise, this means that $X_\mathcal{E}$ has a zero column, implying that $\lambda = 0$ by definition of the equicorrelation set, contradicting the assumption in the lemma). Then by construction we have that $\hat{\beta}_i^{\text{LARS}} \neq 0$ for all $i \in \mathcal{E}$. □

Remark 1. In the case that the lasso solution is unique, this result says that the active set is equal to the equicorrelation set, almost everywhere.

Remark 2. Note that the equicorrelation set $\mathcal{E}$ (and hence the active set of a lasso solution, almost everywhere) can have size $|\mathcal{E}| = p$ in the worst case, even when $p > n$. As a trivial example, consider the case when $X \in \mathbb{R}^{n \times p}$ has p duplicate columns, with $p > n$.

5.2. The smallest active set

We have shown that the LARS lasso solution attains the largest possible active set, and so a natural question is: what is the smallest possible active set? The next result is from Osborne et al. (2000*b*) and Rosset et al. (2004).

Lemma 14. *For any y, X, and $\lambda > 0$, there exists a lasso solution whose set of active variables is linearly independent. In particular, this means that there exists a solution whose active set $\mathcal{A}$ has size $|\mathcal{A}| \leq \min\{n, p\}$.*

Proof. We follow the proof of Rosset et al. (2004) closely. Let $\hat{\beta}$ be a lasso solution, let $\mathcal{A} = \text{supp}(\hat{\beta})$ be its active set, and suppose that $\text{rank}(X_\mathcal{A}) < |\mathcal{A}|$. Then by the same arguments as those given in Section 2, we can write, for some $i \in \mathcal{A}$,

$$s_i X_i = \sum_{j \in \mathcal{A} \setminus \{i\}} a_j s_j X_j, \quad \text{where} \quad \sum_{j \in \mathcal{A} \setminus \{i\}} a_j = 1. \tag{33}$$

Now define

$$\theta_i = -s_i \quad \text{and} \quad \theta_j = a_j s_j \quad \text{for } j \in \mathcal{A} \setminus \{i\}.$$

Starting at $\hat{\beta}$, we move in the direction of θ until a coefficient hits zero; that is, we define

$$\tilde{\beta}_{-\mathcal{A}} = 0 \quad \text{and} \quad \tilde{\beta}_\mathcal{A} = \hat{\beta}_\mathcal{A} + \delta\theta,$$

where

$$\delta = \min\{\rho \geq 0 : \hat{\beta}_j + \rho\theta_j = 0 \text{ for some } j \in \mathcal{A}\}.$$

Notice that δ is guaranteed to be finite, as $\delta \leq |\hat{\beta}_i|$. Furthermore, we have $X\tilde{\beta} = X\hat{\beta}$ because $\theta \in \text{null}(X_{\mathcal{A}})$, and also

$$\begin{aligned}\|\tilde{\beta}\|_1 &= |\tilde{\beta}_i| + \sum_{j \in \mathcal{A}\setminus\{i\}} |\tilde{\beta}_j| \\ &= |\hat{\beta}_i| - \delta + \sum_{j \in \mathcal{A}\setminus\{i\}} (|\hat{\beta}_j| + \delta a_j) \\ &= \|\hat{\beta}\|_1.\end{aligned}$$

Hence we have shown that $\tilde{\beta}$ achieves the same fit and the same ℓ_1 norm as $\hat{\beta}$, so it is indeed also lasso solution, and it has one fewer nonzero coefficient than $\hat{\beta}$. We can now repeat this procedure until we obtain a lasso solution whose active set $\mathcal{A}$ satisfies $\text{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$. □

Remark 1. This result shows that, for any problem instance, there exists a lasso solution supported on $\leq \min\{n, p\}$ variables; some works in the literature have misquoted this result by claiming that every lasso solution is supported on $\leq \min\{n, p\}$ variables, which is clearly incorrect. When the lasso solution is unique, however, Lemma 14 implies that its active set has size $\leq \min\{n, p\}$.

Remark 2. In principle, one could start with any lasso solution, and follow the proof of Lemma 14 to construct a solution whose active set $\mathcal{A}$ is such that $\text{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$. But from a practical perspective, this could be computationally quite difficult, as computing the constants a_j in (33) requires finding a nonzero vector in $\text{null}(X_{\mathcal{A}})$—a nontrivial task that would need to be repeated each time a variable is eliminated from the active set. To the best of our knowledge, the standard optimization algorithms for the lasso problem (such as coordinate descent, first-order methods, quadratic programming approaches) do not consistently produce lasso solutions with the property that $\text{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$ over the active set $\mathcal{A}$. This is in contrast to the solution with largest active set, which is computed by the LARS algorithm.

Remark 3. The proof of Lemma 14 does not actually depend on the lasso problem in particular, and the arguments can be extended to cover the general ℓ_1 penalized minimization problem (11), with f differentiable and strictly convex. (This is in the same spirit as our extension of lasso uniqueness results to this general problem in Section 2.) Hence, to put it explicitly, for any differentiable, strictly convex f, any X, and $\lambda > 0$, there exists a solution of (11) whose active set $\mathcal{A}$ is such that $\text{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$.

The title "smallest" active set is justified, because in the next section we show that the subspace $\text{col}(X_{\mathcal{A}})$ is invariant under all choices of active sets $\mathcal{A}$, for almost every $y \in \mathbb{R}^n$. Therefore, for such y, if $\mathcal{A}$ is an active set satisfying $\text{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$, then one cannot possibly find a solution whose active set has size $< |\mathcal{A}|$, as this would necessarily change the span of the active variables.

5.3. Equivalence of active subspaces

With the multiplicity of active sets (corresponding to lasso solutions of a given problem instance), there may be difficulty in identifying and interpreting important variables, as discussed in the introduction and in Section 4. Fortunately, it turns out that for almost every y, the span of the active variables does not depend on the choice of lasso solution, as shown in Tibshirani and Taylor (2012). Therefore, even though the linear models (given by lasso solutions) may report differences in individual variables, they are more or less equivalent in terms of their scope, almost everywhere in y.

Lemma 15. *Fix any X and $\lambda > 0$. For almost every $y \in \mathbb{R}^n$, the linear subspace $\mathrm{col}(X_\mathcal{A})$ is exactly the same for any active set $\mathcal{A}$ coming from a lasso solution.*

Due to the length and technical nature of the proof, we only give a sketch here, and refer the reader to Tibshirani and Taylor (2012) for full details. First, we define a set $\mathcal{N} \subseteq \mathbb{R}^n$—somewhat like the set defined in (32) in the proof of Lemma 13—to be a union of affine subspaces of dimension $\leq n-1$, and hence $\mathcal{N}$ has measure zero. Then, for any y except in this exceptional set $\mathcal{N}$, we consider any lasso solution at y and examine its active set $\mathcal{A}$. Based on the careful construction of $\mathcal{N}$, we can prove the existence of an open set U containing y such that any $y' \in U$ admits a lasso solution that has an active set $\mathcal{A}$. In other words, this is a result on the local stability of lasso active sets. Next, over U, the lasso fit can be expressed in terms of the projection map onto $\mathrm{col}(X_\mathcal{A})$. The uniqueness of the lasso fit finally implies that $\mathrm{col}(X_\mathcal{A})$ is the same for any choice of active set $\mathcal{A}$ coming from a lasso solution at y.

5.4. A necessary condition for uniqueness (almost everywhere)

We now give a necessary condition for uniqueness of the lasso solution, that holds for almost every $y \in \mathbb{R}^n$ (considering X and λ fixed but arbitrary). This is in fact the same as the sufficient condition given in Lemma 2, and hence, for almost every y, we have characterized uniqueness completely.

Lemma 16. *Fix any X and $\lambda > 0$. For almost every $y \in \mathbb{R}^n$, if the lasso solution is unique, then* $\mathrm{null}(X_\mathcal{E}) = \{0\}$.

Proof. Let $\mathcal{N}$ be as defined in (32). Then for $y \notin \mathcal{N}$, the LARS lasso solution $\hat{\beta}^{\mathrm{LARS}}$ has active set equal to $\mathcal{E}$. If the lasso solution is unique, then it must be the LARS lasso solution. Now suppose that $\mathrm{null}(X_\mathcal{E}) \neq \{0\}$, and take any $b \in \mathrm{null}(X_\mathcal{E})$, $b \neq 0$. As the LARS lasso solution is supported on all of $\mathcal{E}$, we know that

$$s_i \cdot \hat{\beta}_i^{\mathrm{LARS}} > 0 \quad \text{for all } i \in \mathcal{E}.$$

For $\delta > 0$, define

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_\mathcal{E} = \hat{\beta}_\mathcal{E}^{\mathrm{LARS}} + \delta b.$$

Then we know that

$$\delta b \in \mathrm{null}(X_\mathcal{E}) \quad \text{and} \quad s_i \cdot (\hat{\beta}_i^{\mathrm{LARS}} + \delta b_i) > 0, \ \ i \in \mathcal{E},$$

the above inequality holding for small enough $\delta > 0$, by continuity. Therefore $\hat{\beta} \neq \hat{\beta}^{\mathrm{LARS}}$ is also a solution, contradicting uniqueness, which means that $\mathrm{null}(X_{\mathcal{E}}) = \{0\}$. □

6. Discussion

We studied the lasso problem, covering conditions for uniqueness, as well as results aimed at better understanding the behavior of lasso solutions in the non-unique case. Some of the results presented in this paper were already known in the literature, and others were novel. We give a summary here.

Section 2 showed that any one of the following three conditions is sufficient for uniqueness of the lasso solution: (i) $\mathrm{null}(X_{\mathcal{E}}) = \{0\}$, where $\mathcal{E}$ is the unique equicorrelation set; (ii) X has columns in general position; (iii) X has entries drawn from a continuous probability distribution (the implication now being uniqueness with probability one). These results can all be found in the literature, in one form or another. They also apply to a more general ℓ_1 penalized minimization problem, provided that the loss function is differentiable and strictly convex when considered a function of $X\beta$ (this covers, for example, ℓ_1 penalized logistic regression and ℓ_1 penalized Poisson regression). Section 5 showed that for the lasso problem, the condition $\mathrm{null}(X_{\mathcal{E}}) = \{0\}$ is also necessary for uniqueness of the solution, almost everywhere in y. To the best of our knowledge, this is a new result.

Sections 3 and 4 contained novel work on extending the LARS path algorithm to the non-unique case, and on bounding the coefficients of lasso solutions in the non-unique case, respectively. The newly proposed LARS algorithm works for any predictor matrix X, whereas the original LARS algorithm only works when the lasso solution path is unique. Although our extension may superficially appear to be quite minor, its proof of correctness is somewhat more involved. In Section 3 we also discussed some interesting properties of LARS lasso solutions in the non-unique case. Section 4 derived a simple method for computing marginal lower and upper bounds for the coefficients of lasso solutions of any given problem instance. It is also in this section that we showed that no two lasso solutions can exhibit different signs for a common active variable, implying that the bounding intervals cannot contain zero in their interiors. These intervals allowed us to categorize each equicorrelation variable as either "dispensable"—meaning that some lasso solution excludes this variable from active set, or "indispensable"—meaning that every lasso solution includes this variable in its active set. We hope that this represents progress towards interpretation in the non-unique case.

The remainder of Section 5 reviewed existing results from the literature on the active sets of lasso solutions in the non-unique case. The first was the fact that the LARS lasso solution is supported on $\mathcal{E}$, and hence attains the largest active set, almost everywhere in y. Next, there always exists a lasso solution whose active set $\mathcal{A}$ satisfies $\mathrm{rank}(X_{\mathcal{A}}) = |\mathcal{A}|$, and therefore has size $|\mathcal{A}| \leq \min\{n, p\}$. The last result gave an equivalence between all active sets of lasso solutions of

a given problem instance: for almost every y, the subspace $\mathrm{col}(X_{\mathcal{A}})$ is the same for any active set $\mathcal{A}$ of a lasso solution.

While this paper was under revision, the referees raised an interesting question: how does current lasso theory deal with issues of non-uniqueness? Before addressing this question, it is worth pointing out that such theoretical results typically assume a linear generative model for the outcome y as a function of the predictors X, with true coefficients β^*, whereas the current paper considers the issues of uniqueness and computation of lasso solutions and makes no assumptions about the true underlying model. Having mentioned this, we can now address the above question in parts, based on the type of theoretical result sought.

First, for results on bounding the difference in the lasso fit and the true mean, $\|X\hat{\beta} - X\beta^*\|_2$, note that it does not matter whether or not the lasso solution is unique, because the fitted value itself is always unique (recall Lemma 1).

Second, for results on bounding $\|\hat{\beta} - \beta^*\|_2$, issues of uniqueness of $\hat{\beta}$ must clearly be considered. A common assumption used to derive sharp bounds for this quantity is the *restricted eigenvalue condition* on X (see, for example, Bickel et al. (2009), Koltchinskii (2009*a*), Koltchinskii (2009*b*); see also van de Geer and Buhlmann (2009) for an extension discussion of this condition and its relation to other common conditions in the literature). With this assumption in place (and some others), one can prove a fast convergence rate for $\|\hat{\beta} - \beta^*\|_2$, when $\hat{\beta}$ is any solution of the lasso problem at a specific value of λ—in other words, in the non-unique case, any one of the infinite number of lasso solutions will do (for example, Negahban et al. (2012) are careful about stating this explicitly). Generally speaking, there is no known prescription for building deterministic (and high-dimensional) matrices X that satisfy the restricted eigenvalue condition; hence, to make this convergence result more concrete, many authors study random matrices X that satisfy the restricted eigenvalue condition with high probability. It is worth mentioning that typical examples of random matrices X with this property use continuous probability distributions (the most common example being X with i.i.d. Gaussian entries), in which case the lasso solution is unique almost surely (recall Lemma 4), and so the derived bounds on $\|\hat{\beta} - \beta^*\|_2$ really only apply to the single unique solution $\hat{\beta}$. Furthermore, we suspect that even those random matrices X known to satisfy the restricted eigenvalue condition with high probability, but are not continuously distributed (for example, X with i.i.d. Bernoulli entries), can still be shown to have columns in general position with high probability, guaranteeing the uniqueness of the lasso solution (recall Lemma 3) with high probability.

Third, and lastly, for results on recovering the true underlying support set, $\mathrm{supp}(\hat{\beta}) = \mathrm{supp}(\beta^*)$, with high probability, one requires a stronger assumption than the restricted eigenvalue condition, namely, that of *mutual incoherence* or *irrepresentability* (see, for example, Wainwright (2009), Zhao and Yu (2006)). Assuming this condition, a common approach to proving exact support recovery is to use the primal-dual witness method, which (with strict dual feasibility) implies the existence of a lasso solution whose equicorrelation set is $\mathcal{E} = \mathcal{A}^*$,

where $\mathcal{A}^* = \mathrm{supp}(\beta^*)$ is the true active set (see Wainwright (2009)). But mutual incoherence (trivially) implies that $\mathrm{rank}(X_{\mathcal{A}^*}) = |\mathcal{A}^*|$, so the constructed lasso solution has an equicorrelation set with $\mathrm{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$, which implies uniqueness (recall Lemma 2).

Acknowledgements

The idea for computing bounds on the coefficients of lasso solutions was inspired by a similar idea of Max Jacob Grazier G'Sell, for bounding the uncertainty in maximum likelihood estimates. We would like to thank Jacob Bien, Trevor Hastie, Jonathan Taylor, and Robert Tibshirani for helpful comments.

Appendix A: Appendix

A.1. Proof of correctness of the LARS algorithm

We prove that for a general X, the LARS algorithm (Algorithm 1) computes a lasso solution path, by induction on k, the iteration counter. The key result is Lemma 17, which shows that the LARS lasso solution is continuous at each knot λ_k in the path, as we change the equicorrelation set and signs from one iteration to the next. We delay the presentation and proof of Lemma 17 until we discuss the proof of correctness, for the sake of clarity.

The base case $k = 0$ is straightforward, hence assume that the computed path is a solution path through iteration $k-1$, that is, for all $\lambda \geq \lambda_k$. Consider the kth iteration, and let $\mathcal{E}$ and s denote the current equicorrelation set and signs. First we note that the LARS lasso solution, as defined in terms of the current $\mathcal{E}, s$, satisfies the KKT conditions at λ_k. This is implied by Lemma 17, and the fact that the KKT conditions were satisfied at λ_k with the old equicorrelation set and signs. To be more explicit, Lemma 17 and the inductive hypothesis together imply that

$$\big\|X_{-\mathcal{E}}^T\big(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda_k)\big)\big\|_\infty < \lambda_k, \quad X_{\mathcal{E}}^T\big(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda_k)\big) = \lambda_k s,$$

and $s = \mathrm{sign}(\hat{\beta}_{\mathcal{E}}^{\mathrm{LARS}}(\lambda_k))$, which verifies the KKT conditions at λ_k. Now note that for any $\lambda \leq \lambda_k$ (recalling the definition of $\hat{\beta}^{\mathrm{LARS}}(\lambda)$), we have

$$\begin{aligned} X_{\mathcal{E}}^T\big(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda)\big) &= X_{\mathcal{E}}^T y - X_{\mathcal{E}}^T X_{\mathcal{E}}(X_{\mathcal{E}})^+ y + X_{\mathcal{E}}^T (X_{\mathcal{E}}^T)^+ \lambda s \\ &= X_{\mathcal{E}}^T (X_{\mathcal{E}}^T)^+ \lambda s \\ &= \lambda s, \end{aligned}$$

where the last equality holds as $s \in \mathrm{row}(X_{\mathcal{E}})$. Therefore, as λ decreases, only one of the following two conditions can break: $\|X_{-\mathcal{E}}^T(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda)\|_\infty < \lambda$, or $s = \mathrm{sign}(\hat{\beta}_{\mathcal{E}}^{\mathrm{LARS}}(\lambda))$. The first breaks at the next joining time $\lambda_{k+1}^{\mathrm{join}}$, and the second breaks at the next crossing time $\lambda_{k+1}^{\mathrm{cross}}$. Since we only decrease λ

to $\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{join}}, \lambda_{k+1}^{\text{cross}}\}$, we have hence verified the KKT conditions for $\lambda \geq \lambda_{k+1}$, completing the proof.

Now we present Lemma 17, which shows that $\hat{\beta}^{\text{LARS}}(\lambda)$ is continuous (considered as a function of λ) at every knot λ_k. This means that the constructed solution path is also globally continuous, as it is simply a linear function between knots. We note that Tibshirani and Taylor (2011) proved a parallel lemma (of the same name) for their dual path algorithm for the generalized lasso.

Lemma 17 (**The insertion-deletion lemma**). *At the kth iteration of the LARS algorithm, let $\mathcal{E}$ and s denote the equicorrelation set and signs, and let $\mathcal{E}^*$ and s^* denote the same quantities at the beginning of the next iteration. The two possibilities are:*

1. *(Insertion) If a variable joins the equicorrelation set at λ_{k+1}, that is, $\mathcal{E}^*$ and s^* are formed by adding elements to $\mathcal{E}$ and s, then:*

$$\begin{bmatrix} (X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big) \\ 0 \end{bmatrix} = \begin{bmatrix} \Big[(X_{\mathcal{E}^*})^+\big(y - (X_{\mathcal{E}^*}^T)^+\lambda_{k+1}s^*\big)\Big]_{-i_{k+1}^{\text{join}}} \\ \Big[(X_{\mathcal{E}^*})^+\big(y - (X_{\mathcal{E}^*}^T)^+\lambda_{k+1}s^*\big)\Big]_{i_{k+1}^{\text{join}}} \end{bmatrix}. \tag{34}$$

2. *(Deletion) If a variable leaves the equicorrelation set at λ_{k+1}, that is, $\mathcal{E}^*$ and s^* are formed by deleting elements from $\mathcal{E}$ and s, then:*

$$\begin{bmatrix} \Big[(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big)\Big]_{-i_{k+1}^{\text{cross}}} \\ \Big[(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big)\Big]_{i_{k+1}^{\text{cross}}} \end{bmatrix} = \begin{bmatrix} (X_{\mathcal{E}^*})^+\big(y - (X_{\mathcal{E}^*}^T)^+\lambda_{k+1}s\big) \\ 0 \end{bmatrix}. \tag{35}$$

Proof. We prove each case separately. The deletion case is actually easier so we start with this first.

Case 2: Deletion. Let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \Big[(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big)\Big]_{-i_{k+1}^{\text{cross}}} \\ \Big[(X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big)\Big]_{i_{k+1}^{\text{cross}}} \end{bmatrix},$$

the left-hand side of (35). By definition, we have $x_2 = 0$ because variable i_{k+1}^{cross} crosses through zero at λ_{k+1}. Now we consider x_1. Assume without a loss of generality that i_{k+1}^{cross} is the last of the equicorrelation variables, so that we can write

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (X_{\mathcal{E}})^+\big(y - (X_{\mathcal{E}}^T)^+\lambda_{k+1}s\big).$$

The point $(x_1, x_2)^T$ is the minimum ℓ_2 norm solution of the linear equation:

$$X_{\mathcal{E}}^T X_{\mathcal{E}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = X_{\mathcal{E}}^T y - \lambda_{k+1}s.$$

Decomposing this into blocks,

$$\begin{bmatrix} X_{\mathcal{E}^*}^T X_{\mathcal{E}^*} & X_{\mathcal{E}^*}^T X_{i_{k+1}^{\text{cross}}} \\ X_{i_{k+1}^{\text{cross}}}^T X_{\mathcal{E}^*} & X_{i_{k+1}^{\text{cross}}}^T X_{i_{k+1}^{\text{cross}}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} X_{\mathcal{E}^*}^T \\ X_{i_{k+1}^{\text{cross}}}^T \end{bmatrix} y - \lambda_{k+1} \begin{bmatrix} s^* \\ s_{k+1}^{\text{cross}} \end{bmatrix}.$$

Solving this for x_1 gives

$$\begin{aligned} x_1 &= (X_{\mathcal{E}^*}^T X_{\mathcal{E}^*})^+ \left[X_{\mathcal{E}^*}^T y - \lambda_{k+1} s^* - X_{\mathcal{E}^*}^T X_{i_{k+1}^{\text{cross}}} x_2 \right] + b \\ &= (X_{\mathcal{E}^*})^+ \left(y - (X_{\mathcal{E}^*}^T)^+ \lambda_{k+1} s^* \right) + b, \end{aligned}$$

where $b \in \text{null}(X_{\mathcal{E}^*})$. Recalling that x_1 must have minimal ℓ_2 norm, we compute

$$\|x_1\|_2^2 = \left\| (X_{\mathcal{E}^*})^+ \left(y - (X_{\mathcal{E}^*}^T)^+ \lambda_{k+1} s^* \right) \right\|_2^2 + \|b\|_2^2,$$

which is smallest when $b = 0$. This completes the proof.

Case 1: Insertion. This proof is similar, but only a little more complicated. Now we let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \left[(X_{\mathcal{E}^*})^+ \left(y - (X_{\mathcal{E}^*}^T)^+ \lambda_{k+1} s^* \right) \right]_{-i_{k+1}^{\text{join}}} \\ \left[(X_{\mathcal{E}^*})^+ \left(y - (X_{\mathcal{E}^*}^T)^+ \lambda_{k+1} s^* \right) \right]_{i_{k+1}^{\text{join}}} \end{bmatrix},$$

the right-hand side of (34). Assuming without a loss of generality that i_{k+1}^{join} is the largest of the equicorrelation variables, the point $(x_1, x_2)^T$ is the minimum ℓ_2 norm solution to the linear equation:

$$X_{\mathcal{E}^*}^T X_{\mathcal{E}^*} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = X_{\mathcal{E}^*}^T y - \lambda_{k+1} s^*.$$

If we now decompose this into blocks, we get

$$\begin{bmatrix} X_{\mathcal{E}}^T X_{\mathcal{E}} & X_{\mathcal{E}}^T X_{i_{k+1}^{\text{join}}} \\ X_{i_{k+1}^{\text{join}}}^T X_{\mathcal{E}} & X_{i_{k+1}^{\text{join}}}^T X_{i_{k+1}^{\text{join}}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} X_{\mathcal{E}}^T \\ X_{i_{k+1}^{\text{join}}}^T \end{bmatrix} y - \lambda_{k+1} \begin{bmatrix} s \\ s_{k+1}^{\text{join}} \end{bmatrix}.$$

Solving this system for x_1 in terms of x_2 gives

$$\begin{aligned} x_1 &= (X_{\mathcal{E}}^T X_{\mathcal{E}})^+ \left[X_{\mathcal{E}}^T y - \lambda_{k+1} s - X_{\mathcal{E}}^T X_{i_{k+1}^{\text{join}}} x_2 \right] + b \\ &= (X_{\mathcal{E}})^+ \left[y - (X_{\mathcal{E}}^T)^+ \lambda_{k+1} s - X_{\mathcal{E}}^T X_{i_{k+1}^{\text{join}}} x_2 \right] + b, \end{aligned}$$

where $b \in \text{null}(X_{\mathcal{E}})$, and as we argued in the deletion case, we know that $b = 0$ in order for x_1 to have minimal ℓ_2 norm. Therefore we only need to show that $x_2 = 0$. To do this, we solve for x_2 in the above block system, plug in what we know about x_1, and after a bit of calculation we get

$$x_2 = \left[X_{i_{k+1}^{\text{join}}}^T (I - P) X_{i_{k+1}^{\text{join}}} \right]^{-1} \left(X_{i_{k+1}^{\text{join}}}^T \left[(I - P) y + (X_{\mathcal{E}}^T)^+ \lambda_{k+1} s \right] - \lambda s_{k+1}^{\text{join}} \right),$$

where we have abbreviated $P = P_{\mathrm{col}(X_\mathcal{E})}$. But the expression inside the parentheses above is exactly

$$X^T_{i^{\mathrm{join}}_{k+1}}\big(y - X\hat{\beta}^{\mathrm{LARS}}(\lambda_{k+1})\big) - \lambda s^{\mathrm{join}}_{k+1} = 0,$$

by definition of the joining time. Hence we conclude that $x_2 = 0$, as desired, and this completes the proof. □

A.2. Alternate expressions for the joining and crossing times

As remarked in Section 3.1, the joining times in (17) are well-defined in that, for each variable i, only one of $+1$ or -1 gives rise to a joining time in the interval $[0, \lambda_k]$. It may be helpful to see these defined these more precisely; we now give two alternative representations for joining times, as well as a related representation for crossing times.

- *Max form of the joining times.* We can express the ith joining time as the maximum over the possibilites for the sign s_i (of the inner product of the ith variable with the current residual). Define $S_i = \{-1, 1\}$, with the exception that $S_i = \{-\mathrm{sign}(\hat{\beta}^{\mathrm{LARS}}_i(\lambda_k))\}$ if i corresponds to the variable that left the equicorrelation set in the last iteration; then

$$t^{\mathrm{join}}_i = \max_{s_i \in S_i} \frac{X_i^T\big(I - X_\mathcal{E}(X_\mathcal{E})^+\big)y}{s_i - X_i^T(X_\mathcal{E}^T)^+ s} \cdot 1\bigg\{\frac{X_i^T\big(I - X_\mathcal{E}(X_\mathcal{E})^+\big)y}{s_i - X_i^T(X_\mathcal{E}^T)^+ s} \le \lambda_k\bigg\}.$$

- *Intercept form of the joining times.* The ith joining time is defined as the value of λ that solves the equation $a_i - b_i\lambda = \pm\lambda$, subject to this value lying in $[0, \lambda_k]$. By construction, we know that $|a_i - b_i\lambda_k| < \lambda_k$. It is not hard to see, then, that λ in fact solves $a_i - b_i\lambda = s_i\lambda$, where $s_i = \mathrm{sign}(a_i)$, the sign of the intercept of the line $a_i - b_i\lambda$. Hence we can write the ith joining time as

$$t^{\mathrm{join}}_i = \frac{X_i^T\big(I - X_\mathcal{E}(X_\mathcal{E})^+\big)y}{s_i - X_i^T(X_\mathcal{E}^T)^+ s} \quad \text{where } s_i = \mathrm{sign}\Big(X_i^T\big(I - X_\mathcal{E}(X_\mathcal{E})^+\big)y\Big).$$

- *Parallel form of the crossing times.* For the case $\mathrm{rank}(X_\mathcal{E}) = |\mathcal{E}|$, Jonathan Taylor pointed out an interesting parallel between the crossing times in (19) and the joining times. In particular, the crossing time of the ith variable can be written as

$$t^{\mathrm{cross}}_i = \frac{X_i^T\big(I - X_{\mathcal{E}\setminus\{i\}}(X_{\mathcal{E}\setminus\{i\}})^+\big)y}{s_i - X_i^T(X^T_{\mathcal{E}\setminus\{i\}})^+ s_{-i}} \cdot 1\bigg\{\frac{X_i^T\big(I - X_{\mathcal{E}\setminus\{i\}}(X_{\mathcal{E}\setminus\{i\}})^+\big)y}{s_i - X_i^T(X^T_{\mathcal{E}\setminus\{i\}})^+ s_{-i}} \le \lambda_k\bigg\}, \tag{36}$$

where s_i is the ith component of the sign vector s (the sign of the ith coefficient), and s_{-i} is the sign vector s with ith component removed. This has the form of the joining time of the ith variable, had the equicorrelation $\mathcal{E}$

set not included i. To see the equivalence between (19) and (36), first note that by the well-known formula for the ith partial regression coefficient,

$$[(X_{\mathcal{E}})^+ y]_i = \frac{X_i^T (I - P_{\mathcal{E}\setminus\{i\}})y}{\|(I - P_{\mathcal{E}\setminus\{i\}})X_i\|_2^2},$$

where $P_{\mathcal{E}\setminus\{i\}} = X_{\mathcal{E}\setminus\{i\}}(X_{\mathcal{E}\setminus\{i\}})^+$ is the projection matrix onto the column space of $X_{\mathcal{E}\setminus\{i\}}$. Furthermore, writing $e_i \in \mathbb{R}^n$ as the ith basis vector,

$$[(X_{\mathcal{E}}^T X_{\mathcal{E}})^{-1} s]_i = e_i^T (X_{\mathcal{E}}^T X_{\mathcal{E}})^{-1} s = \frac{X_i^T (I - P_{\mathcal{E}\setminus\{i\}})X_{\mathcal{E}}}{\|(I - P_{\mathcal{E}\setminus\{i\}})X_i\|_2^2} (X_{\mathcal{E}}^T X_{\mathcal{E}})^{-1} s.$$

The term $X_i^T (I - P_{\mathcal{E}\setminus\{i\}})X_{\mathcal{E}}(X_{\mathcal{E}}^T X_{\mathcal{E}})^{-1} s$ above can be rewritten as

$$s^T (X_{\mathcal{E}})^+ (I - P_{\mathcal{E}\setminus\{i\}})X_i = s_i - s_{-i}^T (X_{\mathcal{E}\setminus\{i\}})^+ X_i,$$

and therefore we have shown that

$$\frac{[(X_{\mathcal{E}})^+ y]_i}{[(X_{\mathcal{E}}^T X_{\mathcal{E}})^+ s]_i} = \frac{X_i^T (I - P_{\mathcal{E}\setminus\{i\}})y}{s_i - X_i^T (X_{\mathcal{E}\setminus\{i\}}^T)^+ s_{-i}},$$

completing the equivalence proof.

A.3. Local LARS algorithm for the lasso path

We argue that there is nothing special about starting the LARS path algorithm at $\lambda = \infty$. Given any solution the lasso problem at y, X, and $\lambda^* > 0$, we can define the unique equicorrelation set $\mathcal{E}$ and signs s, as in (4) and (5). The LARS lasso solution at λ^* can then be explicitly constructed as in (23), and by following the same steps as those outlined in Section 3.1, we can compute the LARS lasso solution path beginning at λ^*, for decreasing values of the tuning parameter; that is, over $\lambda \in [0, \lambda^*]$.

In fact, the LARS lasso path can also be computed in the reverse direction, for increasing values of the tuning parameter. Beginning with the LARS lasso solution at λ^*, it is not hard to see that in this direction (increasing λ) a variable enters the equicorrelation set at the next crossing time—the minimal crossing time larger than λ^*, and a variable leaves the equicorrelation set at the next joining time—the minimal joining time larger than λ^*. This is of course the opposite of the behavior of joining and crossing times in the usual direction (decreasing λ). Hence, in this manner, we can compute the LARS lasso path over $\lambda \in [\lambda^*, \infty]$.

This could be useful in studying a large lasso problem: if we knew a tuning parameter value λ^* of interest (even approximate interest), then we could compute a lasso solution at λ^* using one of the many efficient techniques from convex optimization (such as coordinate descent, or accelerated first-order methods), and subsequently compute a local solution path around λ^* to investigate the

behavior of nearby lasso solutions. This can be achieved by finding the knots to the left and right of λ^* (performing one LARS iteration in the usual direction and one iteration in the reverse direction), and repeating this, until a desired range $\lambda \in [\lambda^* - \delta_L, \lambda^* + \delta_R]$ is achieved.

A.4. Enumerating all active sets of lasso solutions

We show that the facial structure of the polytope K in (29) describes the collection of active sets of lasso solutions, almost everywhere in y.

Lemma 18. *Fix any X and $\lambda > 0$. For almost every $y \in \mathbb{R}^n$, there is a one-to-one correspondence between active sets of lasso solutions and nonempty faces of the polyhedron K defined in* (29).

Proof. Nonempty faces of K are sets F of the form $F = K \cap H \neq \emptyset$, where H is a supporting hyperplane to K. If $\mathcal{A}$ is an active set of a lasso solution, then there exists an $x \in K$ such that $x_{\mathcal{E}\setminus\mathcal{A}} = 0$. Hence, recalling the sign condition in (28), the hyperplane $H_{\mathcal{E}\setminus\mathcal{A}} = \{x \in \mathbb{R}^{|\mathcal{E}|} : u^T x = 0\}$, where

$$u_i = \begin{cases} s_i & \text{if } i \in \mathcal{E} \setminus \mathcal{A} \\ 0 & \text{if } i \in \mathcal{A}, \end{cases}$$

supports K. Furthermore, we have $F = K \cap H = \{x \in K : \sum_{i\in\mathcal{E}\setminus\mathcal{A}} s_i x_i = 0\} = \{x \in K : x_{\mathcal{E}\setminus\mathcal{A}} = 0\}$. Therefore every active set $\mathcal{A}$ corresponds to a nonempty face F of K.

Now we show the converse statement holds, for almost every y. Well, the facets of K are sets of the form $F_i = K \cap \{x \in \mathbb{R}^{|\mathcal{E}|} : x_i = 0\}$ for some $i \in \mathcal{E}$.[5] Each nonempty proper face F can be written as an intersection of facets: $F = \cap_{i\in\mathcal{I}} F_i = \{x \in K : x_{\mathcal{I}} = 0\}$, and hence F corresponds to the active set $\mathcal{A} = \mathcal{E} \setminus \mathcal{I}$. The face $F = K$ corresponds to the equicorrelation set $\mathcal{E}$, which itself is an active set for almost every $y \in \mathbb{R}^n$ by Lemma 13. □

Note that this means that we can enumerate all possible active sets of lasso solutions, at a given y, X, λ, by enumerating the faces of the polytope K. This is a well-studied problem in computational geometry; see, for example, Fukuda et al. (1997) and the references therein. It is worth mentioning that this could be computationally intensive, as the number of faces can grow very large, even for a polytope of moderate dimensions.

References

Bickel, P., Ritov, Y. and Tsybakov, A. (2009), 'Simultaneous analysis of lasso and Dantzig selector', *Annals of Statistics* **37**(4), 1705–1732. MR2533469

[5] This is slightly abusing the notion of a facet, but the argument here can be made rigorous by reparametrizing the coordinates in terms of the affine subspace $\{x \in \mathbb{R}^{|\mathcal{E}|} : Px = \hat{\beta}^{\text{LARS}}_{\mathcal{E}}\}$.

Candes, E. J. and Plan, Y. (2009), 'Near ideal model selection by ℓ_1 minimization', *Annals of Statistics* **37**(5), 2145–2177. MR2543688

Chen, S., Donoho, D. L. and Saunders, M. (1998), 'Atomic decomposition for basis pursuit', *SIAM Journal on Scientific Computing* **20**(1), 33–61. MR1639094

Donoho, D. L. (2006), 'For most large underdetermined systems of linear equations, the minimal ℓ_1 solution is also the sparsest solution', *Communications on Pure and Applied Mathematics* **59**(6), 797–829. MR2217606

Dossal, C. (2012), 'A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization', *Comptes Rendus Mathematique* **350**(1–2), 117–120. MR2887848

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499. MR2060166

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608. MR2237526

Fukuda, K., Liebling, T. M. and Margot, F. (1997), 'Analysis of backtrack algorithms for listing all vertices and all faces of a convex polyhedron', *Computational Geometry: Theory and Applications* **8**(1), 1–12. MR1452921

Koltchinskii, V. (2009*a*), 'The Dantzig selector and sparsity oracle inequalities', *Bernoulli* **15**(3), 799–828. MR2555200

Koltchinskii, V. (2009*b*), 'Sparsity in penalized empirical risk minimization', *Annales de l'Institut Henri Poincare, Probabilites et Statistiques* **45**(1), 7–57. MR2500227

Mairal, J. and Yu, B. (2012), 'Complexity analysis of the lasso regularization path', *Proceedings of the International Conference on Machine Learning* **29**.

Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. To appear in Statistical Science.

Osborne, M., Presnell, B. and Turlach, B. (2000 *a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404. MR1773265

Osborne, M., Presnell, B. and Turlach, B. (2000 *b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337. MR1822089

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press, Princeton. MR0274683

Rosset, S., Zhu, J. and Hastie, T. (2004), 'Boosting as a regularized path to a maximum margin classifier', *Journal of Machine Learning Research* **5**, 941–973. MR2248005

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288. MR1379242

Tibshirani, R. J. (2011), The Solution Path of the Generalized Lasso, PhD thesis, Department of Statistics, Stanford University.

Tibshirani, R. J. and Taylor, J. (2011), Proofs and technical details for "The solution path of the generalized lasso". http://www.stat.cmu.edu/~ryantibs/papers/genlasso-supp.pdf

Tibshirani, R. J. and Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232. MR2985948

van de Geer, S. and Buhlmann, P. (2009), 'On the conditions used to prove oracle results for the lasso', *Electronic Journal of Statistics* **3**, 1360–1392. MR2576316

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202. MR2729873

Zhao, P. and Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564. MR2274449

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B* **67**(2), 301–320. MR2137327