THREE STRUCTURAL RESULTS ON THE LASSO PROBLEM

Pingmei Xu, Peter J. Ramadge

Department of Electrical Engineering
Princeton University, Princeton NJ 08540
{pingmeix,ramadge}@princeton.edu

ABSTRACT

The lasso problem, least squares with a ℓ_1 regularization penalty, has been very successful as a tool for obtaining sparse representations of data in terms of given dictionary. It is known, but not widely appreciated, that the lasso problem need not have a unique solution. Sufficient conditions which ensure uniqueness of the solution are known but necessary and sufficient conditions have been elusive. We present three structural results on the lasso problem. First, we show that when the dictionary has more columns than rows, it is always possible to ensure that the dictionary has full row rank. Next we show that the feasible set for the dual lasso problem is bounded if and only if the dictionary has full row rank. Lastly, we give necessary and sufficient conditions for the uniqueness of a lasso solution.

Index Terms— Lasso, Uniqueness, Necessary and Sufficient Conditions, Dual Problem, Bounded

1. INTRODUCTION

The sparse representation of data, sound, images, video, etc., with respect to a dictionary of codewords has provided a successful new nonlinear data representation for machine learning and signal/image processing. At the heart of many such sparse representation methods is the squares problem with ℓ_1 regularization, often called the lasso problem [1]:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad {}^1\!/_2\|\mathbf{x} - B\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1. \tag{1}$$

Here $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^p$ are vectors, $B \in \mathbb{R}^{n\times p}$ is a matrix, $\|\mathbf{w}\|_1 = \sum_{j=1}^p |\mathbf{w}_i|$ denotes the ℓ_1 norm of $\mathbf{w}$ and $\lambda > 0$ is a regularization parameter. Problem (1) seeks a sparse representation of $\mathbf{x} \in \mathbb{R}^n$ as a linear combination of a subset of the columns in B. The ℓ_1 regularization encourages the solution $\tilde{\mathbf{w}}$ of (1) to have many zero components and hence to use relatively few of the columns of B. We will call B the *dictionary*, its columns $\{\mathbf{b}_i\}_{i=1}^p$ *codewords*, and $\mathbf{x}$ the *target* vector.

Problem (1) is convex and hence a local minimum is a global minimum. However, it is known that in general (1) may not have a unique solution [2–7]. This stems from the problem's inherent representational character: it seeks a linear combination of a selected subset of the columns of B. Sufficient conditions for uniqueness are known, e.g., [3, Theorem 1], see also [2, 4–7]. Among other things, these conditions assume that at a solution $\tilde{\mathbf{w}}$, the columns of B with nonzero weights are linearly independent. Recently, [7] examined the uniqueness question in a probabilistic framework and showed that when the entries of the dictionary are drawn from a continuous distribution, the lasso has a unique solution almost surely. This is encouraging from an application viewpoint and probably contributes to the widespread success of the lasso framework. Nevertheless, when one does encounter a situation when the solution is nonunique (or close to being so) one would like to be able to verify this with a necessary and sufficient test. Given one lasso solution, [2] provided a necessary and sufficient condition for the existence of a unique solution by checking the intersection between $\mathcal{N}(B)$ (the null space of B) and a convex cone. In this paper, we show in §3 that this condition can actually be expressed by the intersection between $\mathcal{N}(B)$ and a tighter set which is determined by the given lasso solution and the active set of codewords which can be uniquely defined from the dual solution.

A second problem of interest is the Lagrangian dual of (1) [2,8–10]. This problem can be parameterized as follows [10]:

$$\begin{aligned} \max_{\boldsymbol{\theta}\in\mathbb{R}^n} \quad & {}^1\!/_2\|\mathbf{x}\|_2^2 - {}^{\lambda^2}\!/_2\|\boldsymbol{\theta} - \mathbf{x}/\lambda\|_2^2 \\ \text{s.t.} \quad & |\mathbf{b}_i^T\boldsymbol{\theta}| \le 1 \quad \forall i = 1, 2, \ldots, p, \end{aligned} \tag{2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^n$. A point $\boldsymbol{\theta}$ is said to be *dual feasible* if it satisfies the constraints in (2). Problem (2) seeks to maximize the loss function in (2), by finding the dual feasible point $\boldsymbol{\theta}$ that is closest to $\mathbf{x}/\lambda$. By standard results, the dual problem has a unique solution for every $\lambda > 0$ [11].

The main contribution of the paper provides necessary and sufficient conditions for the uniqueness of the lasso solution. We also show two other structural results for the lasso. First, that without loss of generality when $n \le p$, we can always assume that B has full row rank. Second, that the feasible set of the dual lasso problem is bounded if and only of $\operatorname{rank}(B) = n$. Hence when $n \le p$, without loss of generality we can always assume that the dual feasible set is bounded and hence that the set of dual solutions over $\lambda > 0$ is bounded.

We present two of the structural results in §2, the uniqueness result in §3, and conclude in §4.

2. TWO STRUCTURAL RESULTS

The first structural result deals with the rank of B when $n \leq p$, i.e., when B has more codewords than the data dimension. In this case, B is full rank if and only if $\operatorname{rank}(B) = n$. If $\operatorname{rank}(B) < n \leq p$, then (1) and (2) can easily be transformed into equivalent problems with a full rank dictionary. This is our first result.

Theorem 1. *Suppose that B in* (1) *has rank $r < n \leq p$ and let the columns of $U \in \mathbb{R}^{n\times r}$ be an orthonormal basis for the range of B. Then* (1) *is equivalent to the problem:*

$$\min_{\mathbf{w}\in\mathbb{R}^p} {}^1\!/_2\|(\bar{\mathbf{x}} - \bar{B}\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1, \tag{3}$$

with $\bar{\mathbf{x}} = U^T\mathbf{x}$, $\bar{B} = U^T B$ and $\operatorname{rank}(\bar{B}) = r$. Moreover, corresponding columns of the dictionaries B and $\bar{B}$ have the same norm.

Proof. Let $\hat{\mathbf{x}} = UU^T\mathbf{x}$ be the orthogonal projection of $\mathbf{x}$ onto $\mathcal{R}(B)$. Then

$$\begin{aligned} &{}^1\!/_2\|\mathbf{x} - B\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ &= {}^1\!/_2\|(\mathbf{x} - \hat{\mathbf{x}}) + (\hat{\mathbf{x}} - B\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ &= {}^1\!/_2\|(\mathbf{x} - \hat{\mathbf{x}})\|_2^2 + {}^1\!/_2\|(\hat{\mathbf{x}} - B\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1, \end{aligned}$$

where we used the fact that $\mathbf{x} - \hat{\mathbf{x}}$ is orthogonal to $\hat{\mathbf{x}} - B\mathbf{w}$. Since the first term in the last equation is a constant, problem (1) is equivalent to $\min_{\mathbf{w}\in\mathbb{R}^n} {}^1\!/_2\|(\hat{\mathbf{x}} - B\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1$. Now use the fact that $\|U\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ and that $\mathbf{y} \in \mathcal{R}(B)$ implies $UU^T\mathbf{y} = \mathbf{y}$, to write:

$$\begin{aligned} &{}^1\!/_2\|\hat{\mathbf{x}} - B\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ &= {}^1\!/_2\|(U^T\mathbf{x}) - (U^T B)\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ &= {}^1\!/_2\|(\bar{\mathbf{x}} - \bar{B}\mathbf{w})\|_2^2 + \lambda\|\mathbf{w}\|_1, \end{aligned}$$

where $\bar{x} = U^T\mathbf{x} \in \mathbb{R}^r$ and $\bar{B} = U^T B \in \mathbb{R}^{r\times p}$. For each column $\bar{\mathbf{b}}_i$ of $\bar{B}$, $\|\bar{\mathbf{b}}_i\|_2 = \|U^T\mathbf{b}_i\|_2 = \|UU^T\mathbf{b}_i\|_2 = \|\mathbf{b}_i\|_2$. □

Our second result deals with the feasible set of the dual problem (2). We first recall how dual problem (2) is obtained. Setting $\mathbf{z} = \mathbf{x} - B\mathbf{w}$ in (1) gives the constrained problem: $\min_{\mathbf{z},\mathbf{w}} {}^1\!/_2\, \mathbf{z}^T\mathbf{z} + \lambda\|\mathbf{w}\|_1$, subject to $\mathbf{z} = \mathbf{x} - B\mathbf{w}$. Minimization of the Lagrangian $L(\mathbf{z}, \mathbf{w}, \boldsymbol{\mu}) = {}^1\!/_2\, \mathbf{z}^T\mathbf{z} + \lambda\|\mathbf{w}\|_1 + \boldsymbol{\mu}^T(\mathbf{x} - B\mathbf{w} - \mathbf{z})$ with respect to $\mathbf{z}$ and $\mathbf{w}$ yields $\boldsymbol{\mu} = \tilde{\mathbf{z}}$ and the constraints $|\boldsymbol{\mu}^T\mathbf{b}_i| \leq \lambda$, $i = 1, \ldots, p$. This leads to the dual problem: $\max_{\boldsymbol{\mu}} {}^1\!/_2\, \|\mathbf{x}\|_2^2 - {}^1\!/_2\|\boldsymbol{\mu} - \mathbf{x}\|_2^2$, subject to $|\boldsymbol{\mu}^T\mathbf{b}_i| \leq \lambda$, $i = 1, \ldots, p$, which always has a unique solution. The change of variable $\boldsymbol{\theta} = \boldsymbol{\mu}/\lambda$ then gives (2). Finally, by this construction the solutions $\tilde{\mathbf{w}}$ and $\tilde{\boldsymbol{\theta}}$ of (1) and (2) are related through:

$$\mathbf{x} = \mathbf{B}\tilde{\mathbf{w}} + \lambda\tilde{\boldsymbol{\theta}}, \tag{4}$$

$$\mathbf{b}_i^T\tilde{\boldsymbol{\theta}} = \begin{cases} \operatorname{sign} \tilde{w}_i & \text{if } \tilde{w}_i \neq 0; \\ \sigma \in [-1, 1] & \text{if } \tilde{w}_i = 0. \end{cases} \tag{5}$$

Letting $\mathfrak{B} = \{\pm\mathbf{b}_i\}_{i=1}^p$ allows the constraints in (2) to be stated as $\forall\mathbf{b} \in \mathfrak{B} : \mathbf{b}^T\boldsymbol{\theta} \leq 1$.

Let $H(\mathbf{y}) = \{\mathbf{z} : \mathbf{z}^T\mathbf{y} \leq 1\}$ denote the corresponding closed half space containing the origin. So a constraint of the form $\mathbf{b}^T\boldsymbol{\theta} \leq 1$ requires that $\boldsymbol{\theta}$ lies in the closed half space $H(\mathbf{b})$. It follows that the set of dual feasible points $\mathcal{F}$ is a nonempty, closed, convex polyhedron formed by the intersection of a finite set of closed half spaces. And it is symmetric with respect to the origin, i.e. $\mathcal{F} = -\mathcal{F}$.

The boundedness of $\mathcal{F}$ is determined by the rank of the dictionary. This is our second result.

Theorem 2. *The set $\mathcal{F}$ of dual feasible points is bounded if and only if* $\operatorname{rank}(B) = n$.

When $n \leq p$, it follows from Theorems 1 and 2 that without loss of generality we can always assume that the set of dual feasible points is bounded. On the other hand, when $p < n$, $\operatorname{rank}(B) \leq p < n$ and the set $\mathcal{F}$ cannot be bounded. There are simply not enough codewords to form a bounded region in $\mathbb{R}^n$.

Proof. ($\Leftarrow$) Suppose $\mathcal{F}$ is unbounded. Then there exists $\mathbf{h}$ with $\|\mathbf{h}\|_2 = 1$ such that $\alpha\mathbf{h} \in \mathcal{F}$ for all $\alpha > 0$. Since $\operatorname{rank}(B) = n$, there exists $\mathbf{u} \in \mathbb{R}^p$ with $\mathbf{h} = B\mathbf{u}$. Then $1 = \mathbf{h}^T\mathbf{h} = (\mathbf{h}^T B)\mathbf{u}$. It follows that for some i, $\mathbf{h}^T\mathbf{b}_i u_i > 0$. If $u_i > 0$, then $\mathbf{h}^T\mathbf{b}_i > 0$; and if $u_i < 0$, then $\mathbf{h}^T(-\mathbf{b}_i) > 0$. Hence there exists $\mathbf{b} \in \mathfrak{B}$ such that $\mathbf{h}^T\mathbf{b} = c > 0$. But then $\mathbf{b}^T(\alpha\mathbf{h}) = \alpha c$ is greater than 1 for α sufficiently large. This contradicts $\alpha\mathbf{h} \in \mathcal{F}$ for all $\alpha > 0$. Hence $\mathcal{F}$ must be bounded. ($\Rightarrow$) Suppose $\operatorname{rank}(B) < n$. Then there exists $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\|_2 = 1$ such that $\boldsymbol{\theta}^T B = 0$. Hence $\mathbf{b}^T\boldsymbol{\theta} = 0$ for each $\mathbf{b} \in \mathfrak{B}$. It follows that $\boldsymbol{\theta} \in \mathcal{F}$. Moreover, for each $\alpha > 0$, $\alpha\boldsymbol{\theta}$ is also in $\mathcal{F}$. But then $\mathcal{F}$ is unbounded; a contradiction. Thus $\operatorname{rank}(B) = n$. □

We end this section with a few additional observations about $\mathcal{F}$. We note that the dual problem seeks a point within $\mathcal{F}$ that is closest to a given point in $\mathbb{R}^n$. This is a well studied problem. For any closed convex set $C \subset \mathbb{R}^n$ and any point $\mathbf{z} \in \mathbb{R}^n$, there is a unique point $\hat{\mathbf{z}} \in C$ that is closest to $\mathbf{z}$ [11, §3.1]. Hence for any $\mathbf{x}$ and any $\lambda > 0$, the dual problem has a unique solution $\tilde{\boldsymbol{\theta}}$.

For λ sufficiently small the dual solution $\tilde{\boldsymbol{\theta}}(\lambda)$ will lie on the boundary of $\mathcal{F}$. The boundary is the surface of a polyhedron in $\mathbb{R}^n$. For any $\boldsymbol{\theta} \in \mathcal{F}$, define

$$A^+(\boldsymbol{\theta}) = \{i \colon \mathbf{b}_i^T\boldsymbol{\theta} = 1\}\,,\ A^-(\boldsymbol{\theta}) = \{i \colon \mathbf{b}_i^T\boldsymbol{\theta} = -1\} \tag{6}$$

and set $A(\boldsymbol{\theta}) = A^+(\boldsymbol{\theta}) \cup A^-(\boldsymbol{\theta})$. The codewords indexed in $A(\boldsymbol{\theta})$ are the *active constraints* at $\boldsymbol{\theta}$. If $A(\boldsymbol{\theta}) = \emptyset$, $\boldsymbol{\theta}$ is in the interior of $\mathcal{F}$. Each cell on the boundary of the polyhedron $\mathcal{F}$ has a nonempty configuration of active sets A^+, A^- and $A = A^+ \cup A^-$ for all points $\boldsymbol{\theta}$ within that cell.

3. UNIQUENESS OF THE LASSO SOLUTION

Now consider the uniqueness of the solution of the lasso problem (1). For any $\lambda > 0$, let $\tilde{\boldsymbol{\theta}}(\lambda)$ denote the unique dual solution, $\tilde{A}^+(\lambda)$, $\tilde{A}^-(\lambda)$ denote the corresponding set of active constraints at $\tilde{\boldsymbol{\theta}}(\lambda)$ and let $\tilde{A}(\lambda) = \tilde{A}^+(\lambda) \cup \tilde{A}^-(\lambda)$. We first establish the following simple lemma, which was also partially stated in [7].

Lemma 1. *Let* $\tilde{\mathbf{x}}(\lambda) = \mathbf{x} - \lambda\tilde{\boldsymbol{\theta}}(\lambda)$. *Then*

$$i \in \tilde{A}^+(\lambda) \Leftrightarrow \mathbf{b}_i^T(\mathbf{x} - \tilde{\mathbf{x}}(\lambda)) = \lambda, \tag{7}$$

$$i \in \tilde{A}^-(\lambda) \Leftrightarrow \mathbf{b}_i^T(\mathbf{x} - \tilde{\mathbf{x}}(\lambda)) = -\lambda, \tag{8}$$

$$i \notin \tilde{A}(\lambda) \Leftrightarrow \mathbf{b}_i^T(\mathbf{x} - \tilde{\mathbf{x}}(\lambda)) \in (-\lambda, \lambda). \tag{9}$$

If $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$ *are solutions of* (1)*, then* $\tilde{\mathbf{x}}(\lambda) = B\tilde{\mathbf{w}}_1 = B\tilde{\mathbf{w}}_2$ *and* $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$ *have the same least squares error* $\|\mathbf{x} - B\tilde{\mathbf{w}}_1\|_2^2 = \|\mathbf{x} - B\tilde{\mathbf{w}}_2\|_2^2$*, the same* ℓ_1 *norm* $\|\tilde{\mathbf{w}}_1\|_1 = \|\tilde{\mathbf{w}}_2\|_1$ *and the same active sets* $A^+(\lambda), A^-(\lambda)$.

Proof. By definition, $i \in \tilde{A}^+(\lambda) \Leftrightarrow \mathbf{b}_i^T\tilde{\boldsymbol{\theta}}(\lambda) = 1$. Noting that $\tilde{\boldsymbol{\theta}}(\lambda) = (\mathbf{x} - \tilde{\mathbf{x}}(\lambda))/\lambda$ yields $i \in \tilde{A}^+(\lambda) \Leftrightarrow \mathbf{b}_i^T(\mathbf{x} - \tilde{\mathbf{x}}(\lambda)) = \lambda$. The proof for $\tilde{A}^-(\lambda)$ is similar. For every $\mathbf{b}_i$, $\mathbf{b}_i^T\tilde{\boldsymbol{\theta}}(\lambda) \in [-1, 1]$. By definition, $i \notin \tilde{A}(\lambda) \Leftrightarrow \mathbf{b}_i^T\tilde{\boldsymbol{\theta}}(\lambda) \in (-1, 1)$. Since $\tilde{\boldsymbol{\theta}}(\lambda) = (\mathbf{x} - \tilde{\mathbf{x}}(\lambda))/\lambda$, this is equivalent to (9). Now let $\tilde{\mathbf{w}}$ be a solution of (1) with the regularization parameter set to λ. Then by (4), $\tilde{\mathbf{x}}(\lambda) = B\tilde{\mathbf{w}}$. Hence if $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$ are solutions of (1), then $\tilde{\mathbf{x}}(\lambda) = B\tilde{\mathbf{w}}_1 = B\tilde{\mathbf{w}}_2$. It follows that $\|\mathbf{x} - B\tilde{\mathbf{w}}_1\|_2^2 = \|\mathbf{x} - B\tilde{\mathbf{w}}_2\|_2^2$. Since $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$ both minimize the lasso objective and have the same least squares cost, it must also hold that $\|\tilde{\mathbf{w}}_1\|_1 = \|\tilde{\mathbf{w}}_2\|_1$. By (7), (8), the active sets are determined by $\tilde{\mathbf{x}} = B\tilde{\mathbf{w}}$. Hence $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2$ have the same active constraint sets. □

The vector $\tilde{\mathbf{x}}(\lambda)$ is the approximation of $\mathbf{x}$, and $\tilde{\mathbf{e}}(\lambda) = \mathbf{x} - \tilde{\mathbf{x}}(\lambda)$ is the corresponding error or residual, produced by solving the lasso problem. Since $\tilde{\boldsymbol{\theta}}(\lambda)$ is unique, these points are uniquely defined. Moreover, $\tilde{A}^+(\lambda)$ and $\tilde{A}^-(\lambda)$, defined by $\tilde{\boldsymbol{\theta}}(\lambda)$, are also uniquely determined by $\tilde{\mathbf{x}}(\lambda)$. Note that $\tilde{A}^+$ and $\tilde{A}^-$ are based on the active constraints of the dual solution, not on the support set of a particular solution $\tilde{\mathbf{w}}$ of the primal problem. A solution $\tilde{\mathbf{w}}$ of (1) simply gives a representation of $\tilde{\mathbf{x}}$ as $B\tilde{\mathbf{w}}$. It is well known, however, that this representation may not be unique. But any two solutions will have the same least squares error, the same ℓ_1 norm and the same active constraint sets $\tilde{A}^+, \tilde{A}^-$.

For the next step it will be useful to recall the following basic optimality result, see e.g., [3].

Lemma 2. *A vector* $\mathbf{w}$ *is a solution of* (1) *if and only if for each codeword* b_i*:*

$$b_i^T(\mathbf{x} - B\mathbf{w}) = \begin{cases} \lambda, & \text{if } \mathbf{w}_i > 0 \\ -\lambda, & \text{if } \mathbf{w}_i < 0 \\ \sigma \in [-\lambda, \lambda], & \text{if } \mathbf{w}_i = 0. \end{cases} \tag{10}$$

The known sufficient condition for a solution $\tilde{\mathbf{w}}$ of (1) to be unique is that it satisfies (10) with the third line satisfied with inclusion in $(-\lambda, \lambda)$ and the matrix of consisting of the columns of B corresponding to the nonzero entries of $\tilde{\mathbf{w}}$ has full rank [3, Theorem 1]. Here we focus on developing a necessary and sufficient condition.

For any matrix M, let $\mathcal{N}(M)$ denote the null space of M. For a subspace $\mathcal{U}$ of $\mathbb{R}^k$, let $\mathcal{U}^\perp$ denote the subspace of all vectors orthogonal to every vector in $\mathcal{U}$. If $\mathbf{u} \in \mathbb{R}^k$, $\operatorname{span}\{\mathbf{u}\}$ is the subspace spanned by $\mathbf{u}$. In this case, $\operatorname{span}\{\mathbf{u}\}^\perp$ is the $k-1$ dimensional hyperplane with normal $\mathbf{u}$. For brevity, we abbreviate $\operatorname{span}\{\mathbf{u}\}^\perp$ to simply $\mathbf{u}^\perp$.

In what follows, we abbreviate $\tilde{A}^+(\lambda)$ to simply $\tilde{A}^+$. Similarly, for $\tilde{A}^-(\lambda)$ and $\tilde{A}(\lambda)$. It will be understood that these sets depend on λ. Note that for fixed values of $\mathbf{x}$ and λ, all solutions $\tilde{\mathbf{w}}$ of (1) share the same active set $\tilde{A}$. What will be important is the projection of $\tilde{\mathbf{w}}$ onto its coordinates indexed by this set. Let $\tilde{\mathbf{w}}_{\downarrow\tilde{A}} \in \mathbb{R}^{|\tilde{A}|}$ be formed by retaining only the entries of $\tilde{\mathbf{w}}$ indexed by $\tilde{A}$. For simplicity, we also assume that the the vector $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ has the entries corresponding to $\tilde{A}^+$ at the top followed by the entries corresponding to $\tilde{A}^-$. This can always be achieved by a permutation of the columns of B. This defines a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^{|\tilde{A}|}$ satisfying $\|\tilde{\mathbf{w}}\|_1 = \|\tilde{\mathbf{w}}_{\downarrow\tilde{A}}\|_1$ for every solution $\tilde{\mathbf{w}}$. Correspondingly, let $B_{\downarrow\tilde{A}}$ denote the submatrix of B formed by selecting only the columns with indices in $\tilde{A}$. From these definitions and the definition of $\tilde{A}$, it follows that $B\tilde{\mathbf{w}} = B_{\downarrow\tilde{A}}\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$.

For given $\mathbf{x}$ and λ, let $\tilde{\mathbf{a}}^+ \in \mathbb{R}^p$ be the indicator vector of $\tilde{A}^+$: $\tilde{\mathbf{a}}_i^+ = 1$ if $i \in \tilde{A}^+$, and zero otherwise. Similarly, define the indicator vector of $\tilde{A}^-$ as $\tilde{\mathbf{a}}^-$. Then set $\tilde{\mathbf{a}} = \tilde{\mathbf{a}}^+ - \tilde{\mathbf{a}}^-$. You can think of $\tilde{\mathbf{a}}$ as a signed indicator vector of the active set of constraints. For a solution $\tilde{\mathbf{w}}$, this indicator vector $\tilde{\mathbf{a}}$ ensures $\tilde{\mathbf{a}}^T\tilde{\mathbf{w}} = \|\tilde{\mathbf{w}}\|_1$.

Since each entry of $\tilde{\mathbf{a}}_{\downarrow\tilde{A}}$ is in $\{\pm 1\}$, it specifies a face (including its edges, vertices etc) of the unit ℓ_1-ball in $\mathbb{R}^{|\tilde{A}|}$ and for every solution $\tilde{\mathbf{w}}$, the projection $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ must lie on this face after scaling by $\|\tilde{\mathbf{w}}\|_1$. The $(|\tilde{A}| - 1)$-dimensional subspace $(\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp$is parallel to the selected face of the unit ℓ_1-ball. Moreover, translation of a solution $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ in this subspace perserves the ℓ_1 norm. Similarly, translation of a solution $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ in $\mathcal{N}(B_{\downarrow\tilde{A}})$ perserves the least squares error term. Hence a sufficient condition for a unique solution is that $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp = \{\mathbf{0}\}$. Note that this is only sufficient because there is one more constraint to take into account. For a solution $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ on the selected face of the ℓ_1-ball, one can only move in a set of feasible directions, i.e., $\tilde{w}_{\downarrow\tilde{A},i} \geq 0$ for $i = 1, 2, ...|\tilde{A}^+|$, and

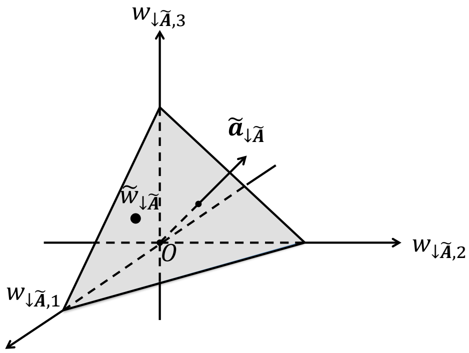

Fig. 1. All solutions must lie on the same face of the ℓ_1-ball.

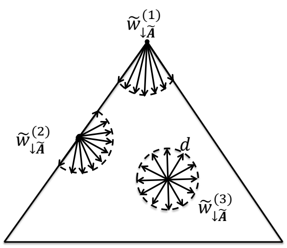

Fig. 2. $D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$ depends on where $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ lies on the face.

$\tilde{w}_{\downarrow\tilde{A},i} \leq 0$ for $i = |\tilde{A}^+| + 1, |\tilde{A}^+| + 2, ...|\tilde{A}|$, or stated equivalently, $\tilde{\mathbf{w}}_{\downarrow\tilde{A}} \in \mathbb{R}_+^{|\tilde{A}^+|} \times \mathbb{R}_-^{|\tilde{A}-|}$. Now we define a mobility set of $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ as $D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{v} \in \mathbb{R}^{|\tilde{A}|} : \mathbf{v} + \tilde{\mathbf{w}}_{\downarrow\tilde{A}} \in \mathbb{R}_+^{|\tilde{A}^+|} \times \mathbb{R}_-^{|\tilde{A}-|}\} = \mathbb{R}_+^{|\tilde{A}^+|} \times \mathbb{R}_-^{|\tilde{A}-|} - \tilde{\mathbf{w}}_{\downarrow\tilde{A}}$. Then $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$ remains a solution when shifted by any vector in $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$. Hence we have a tighter sufficent condition for uniqueness: $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$.

This condition is also necessary. To see this, we need to restate Lemma 2. Suppose we have already found the dual solution $\tilde{\boldsymbol{\theta}}(\lambda)$ and known $\tilde{A}^+(\lambda), \tilde{A}^-(\lambda)$, sufficient and necessary conditions stated in (4) and (5) are equivalent to the following Lemma.

Lemma 3. *A vector* $\mathbf{w}$ *with active sets* $\tilde{A}^+$ *and* $\tilde{A}^-$ *is a solution of* (1) *if and only if:*

$$B_{\downarrow\tilde{A}}\mathbf{w}_{\downarrow\tilde{A}} = \mathbf{x} - \lambda\tilde{\boldsymbol{\theta}} \tag{11}$$

$$w_{\downarrow\tilde{A},i} \begin{cases} \geq 0 & \text{if } i \in \tilde{A}^+ \\ \leq 0 & \text{if } i \in \tilde{A}^-. \end{cases} \tag{12}$$

Once we fix the active sets and determine one solution, the set of all solutions is determined by the intersection of a linear constraint and a set of simple linear inequalities. This is a convex set. The nature of the lasso problem ensures at least one solution, denoted as $\tilde{\mathbf{w}}_{\downarrow\tilde{A}}$. For any $\mathbf{h} = \mathbf{w}_{\downarrow\tilde{A}} - \tilde{\mathbf{w}}_{\downarrow\tilde{A}} \in \mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$, $\mathbf{w}_{\downarrow\tilde{A}} = \tilde{\mathbf{w}}_{\downarrow\tilde{A}} + \mathbf{h}$ satisfies (11) and (12), and therefore is also a solution, vice versa. We now state our third result.

Theorem 3. *Let* $S(\mathbf{x}, \lambda)$ *denote the solution set of* (1), $\tilde{A}^+, \tilde{A}^-$ *denote its active constraint sets and* $\tilde{\mathbf{a}}$ *denote its corresponding signed indicator vector. Then:*
1) $S(\mathbf{x}, \lambda)$ *is a closed, bounded convex set of constant* ℓ_1 *norm.*
2) The lasso solution $\tilde{\mathbf{w}}$ *is unique if and only if*

$$\mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}.$$

3) If $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp = \{\mathbf{0}\}$*, then the solution is unique.*
4) If $\mathcal{N}(B_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$*, then the solution is unique.*

Note that $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$. The sufficient condition given in 4) has been reported in the previous literature [3, 7].

Proof. 1) By Lemma 1, all solutions $\tilde{\mathbf{w}} \in S(\mathbf{x}, \lambda)$ have the same ℓ_1-norm and the same active sets $\tilde{A}^+$, $\tilde{A}-$ and $\tilde{A}$. The invariance of the ℓ_1-norm ensures $S(\mathbf{x}, \lambda)$ is bounded. The continuity of the lasso ojective in $\tilde{\mathbf{w}}$ ensures that $S(\mathbf{x}, \lambda)$ is closed. By Lemma 3, $S(\mathbf{x}, \lambda)$ is convex.

2) ($\Leftarrow$) Assume that $\tilde{\mathbf{w}}^{(1)}, \tilde{\mathbf{w}}^{(2)} \in S(\mathbf{x}, \lambda)$. Let $\mathbf{h} = \tilde{\mathbf{w}}^{(1)} - \tilde{\mathbf{w}}^{(2)}$. Then by Lemma 3, any feasible shift vector satisfies $\mathbf{h} \in \mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$; a contradiction unless $\mathbf{h} = \mathbf{0}$. So the solution is unique.
($\Rightarrow$) Let $S(\mathbf{x}, \lambda) = \{\tilde{\mathbf{w}}\}$, i.e. $\tilde{\mathbf{w}}$ is the unique solution to (1). If there exists a nonzero $\mathbf{h} \in \mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$, by Lemma 3, $\tilde{\mathbf{w}} + \mathbf{h}$ satisfies (11) because $\mathbf{h} \in \mathcal{N}(B_{\downarrow\tilde{A}})$, and (12) because $\mathbf{h} \in D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}})$. As a result, $\tilde{\mathbf{w}} + \mathbf{h}$ is another solution to (1). Contradiction. So $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$.

In addition, $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$ is also a necessary and sufficient condition. To see its sufficiency, we only have to additionally prove $\mathbf{h} = \tilde{\mathbf{w}}^{(1)} - \tilde{\mathbf{w}}^{(2)}$ lies in $(\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp$, which is readily seen from $\tilde{\mathbf{a}}^T\mathbf{h} = \tilde{\mathbf{a}}^T(\tilde{\mathbf{w}}^{(1)} - \tilde{\mathbf{w}}^{(2)}) = \|\tilde{\mathbf{w}}^{(1)}\|_1 - \|\tilde{\mathbf{w}}^{(2)}\|_1 = 0$. To see its necessity, from 2) we know $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$ and clearly $\mathbf{0}$ is in set, so $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp \cap D(\tilde{\mathbf{w}}_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$.

3), 4) If $\mathcal{N}(B_{\downarrow\tilde{A}}) \cap (\tilde{\mathbf{a}}_{\downarrow\tilde{A}})^\perp = \{\mathbf{0}\}$ or if $\mathcal{N}(B_{\downarrow\tilde{A}}) = \{\mathbf{0}\}$, then the second condition in part 2) is satisfied. ☐

4. CONCLUSION

In this paper we present three structural results on the lasso problem, covering full-rank represenation, boundedness of dual feasible set and uniqueness of the lasso solution. An important contribution of the paper is to point out the importance of working on the active set $\tilde{A}$ determined from the dual solution, rather than the support set of the primal solution. The support set of a solution $\tilde{\mathbf{w}}$, i.e., the indices i such that $\tilde{w}_i \neq 0$, is always a subset of the active set of the corresponding unique dual solution $\tilde{\boldsymbol{\theta}} = (\mathbf{x} - B\tilde{\mathbf{w}})/\lambda$, but it can be a proper subset. With help of this active set $\tilde{A}$, we are able to express the primal solution set explicitly, and close the gap between sufficient and necessary conditions for the uniqueness of lasso solution.

5. REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[2] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, June 2000.

[3] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *Information Theory, IEEE Transactions on*, vol. 50, no. 6, pp. 1341–1344, june 2004.

[4] J.J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *Information Theory, IEEE Transactions on*, vol. 51, no. 10, pp. 3601 –3608, oct. 2005.

[5] M.J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[6] E.J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.

[7] R.J. Tibshirani, "The lasso problem and uniqueness," *arXiv:1206.0313 [math.ST]*, 4th Nov. 2012.

[8] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large scale ℓ_1-regularized least squares," *IEEE Selected Topics in Signal Processing*, vol. 1, pp. 606–617, 2007.

[9] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," Tech. Rep. UCB/EECS-2010-126, EECS Department, University of California, Berkeley, Sep 2010.

[10] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems*, 2011.

[11] J. B. Hiriart-Urruty and C. Lemarechal, *Fundamentals of Convex Analysis*, Springer, 2001.