On the LASSO and Dantzig Selector Equivalence

M. Salman Asif and Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332, USA
Email: {sasif, jrom}@gatech.edu

***Abstract*—Recovery of sparse signals from noisy observations is a problem that arises in many information processing contexts. LASSO and the Dantzig selector (DS) are two well-known schemes used to recover high-dimensional sparse signals from linear observations. This paper presents some results on the equivalence between LASSO and DS. We discuss a set of conditions under which the solutions of LASSO and DS are same. With these conditions in place, we formulate a shrinkage procedure for which LASSO and DS follow the same solution path. Furthermore, we show that under these shrinkage conditions the solution to LASSO and DS can be attained in at most S homotopy steps, where S is the number of nonzero elements in the final solution. Thus the computational cost for finding complete homotopy path for an $M \times N$ system is merely $O(SMN)$.**

I. Introduction

In recent years, there has been tremendous progress in the area of sparse signal approximation and reconstruction, with applications in wide ranging disciplines. Compressive Sensing (CS) theory has played a significant role in this regard by providing general conditions under which it is possible to recover sparse signals from a limited number of observations [1]–[3].

The general sparse recovery problem can be formulated as follows. Suppose we are given a data vector $y \in \mathbb{R}^M$ which obeys the linear model

$$y = Ax_0 + e, \tag{1}$$

where $x_0 \in \mathbb{R}^N$ is the unknown S-sparse vector (i.e., it has at most S nonzero components), A is an $M \times N$ matrix with $M < N$, and e is a noise vector. The goal is to reliably estimate x_0 from y. This problem has received a good deal of attention recently in two different signal processing contexts. In the CS framework, we view x_0 as an unknown signal of interest which we wish to recover from indirect observations through A. In the sparse approximation framework, we view y as an observed signal, A as an overcomplete dictionary, and x_0 as a sparse decomposition of y in terms of the columns of matrix A [4].

In this paper we discuss two well-known convex programs commonly used to recover sparse signals from noisy measurements, namely LASSO [5] and the Dantzig Selector (DS) [6]. Given the observations y, LASSO solves[1]

$$\text{minimize } \tau\|x\|_1 + \frac{1}{2}\|Ax - y\|_2^2, \tag{2}$$

[1]This problem formulation is actually Lagrangian form of the LASSO in [5], also known as Basis Pursuit DeNoising (BPDN) [7].

and DS solves

$$\text{minimize } \|x\|_1 \text{ subject to } \|A^T(Ax - y)\|_\infty \le \tau, \tag{3}$$

where $\tau > 0$ is some suitably chosen threshold parameter, which essentially controls the trade-off between the sparsity of the solution to these problems and its fidelity to the measurements y. There are number of performance guarantees associated with both (2) and (3) [6], [8]–[12]. In a nutshell, if the underlying signal x_0 is sufficiently sparse and the matrix A obeys some *incoherence* or *restricted isometry* conditions, then the reconstruction error for the estimates of both (2) and (3) comes within a small factor of the minimum achievable error.

These sparse recovery results are usually given in terms of two properties of the matrix A: the coherence or the restricted isometry constants. The coherence of a matrix A, with normalized columns a_j, is defined as

$$\mu(A) = \max_{i \neq j} |\langle a_i, a_j \rangle|. \tag{4}$$

The matrix A is called incoherent if $\mu(A)$ is small; a typical case of interest is $\mu \sim 1/\sqrt{M}$. Low coherence means that the columns of A are very different from one another, which helps disambiguate short linear combinations when finding the sparse vector x_0 in (1) [8], [13]. A matrix A is said to obey the restricted isometry property (RIP) of order $2S$ if there exists a constant $\delta_{2S} < 1$ such that for all $2S$-sparse signals

$$(1 - \delta_{2S})\|x\|_2 \le \|Ax\|_2 \le (1 + \delta_{2S})\|x\|_2. \tag{5}$$

If the RIP constant δ_{2S} is sufficiently small, then any S-sparse signal can be reliably estimated using some suitable recovery scheme, such as (2) and (3) [14], [15].

The main contribution of this paper is to outline a set of general conditions, in terms of the sparsity of the underlying vector x_0 and the coherence parameter μ, under which LASSO and DS have the same solutions. Further, we show that under similar conditions LASSO and DS follow the same homotopy paths as the regularization parameter τ changes, and that their final solutions can be achieved in at most S homotopy steps, where S is the number of nonzero elements in the final solution. Lastly, we discuss the types of incoherent matrices that obey the conditions required for the equivalence of LASSO and DS.

II. Preliminaries on LASSO and Dantzig Selector

A. Optimality conditions

We start with a discussion of the *optimality conditions* for LASSO and DS which must be satisfied by their respective solutions at any given value of τ.

LASSO: To be a solution to (2), a vector $x^\star$ must obey the following condition:

$$\|A^T(Ax^\star - y)\|_\infty \leq \tau. \tag{L}$$

We can view (L) as a set of n different constraints, one on each entry of the vector of residual correlations $A^T(Ax^\star - y)$, which can be derived by taking subgradient of the objective in (2) [16]. In addition, a sufficient condition for the optimality of $x^\star$ is that the set of locations for which the constraints in (L) are active (i.e., equal to τ) will be the same as the support of $x^\star$ (the set of locations on which $x^\star$ is non-zero) [16]. Denoting this set by Γ, we can write the optimality conditions for any given value of τ as

L1. $A_\Gamma^T(Ax^\star - y) = -\tau z$

L2. $\|A_{\Gamma^c}^T(Ax^\star - y)\|_\infty < \tau$,

where A_Γ is the $m \times |\Gamma|$ matrix formed from the columns of A indexed by elements in Γ, and z is a $|\Gamma|$-vector containing the signs of $x^\star$ on Γ. From this we see that $x^\star$ can be calculated directly from the support Γ and sign sequence z as

$$x^\star = \begin{cases} (A_\Gamma^T A_\Gamma)^{-1}(A_\Gamma^T y - \tau z) & \text{on } \Gamma \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Dantzig Selector: On the other hand, the solution to (3) obeys a similar set of optimality conditions, with the addition of a dual variable. The dual problem to (3) can be written as

$$\begin{aligned} &\text{maximize} \;\; -(\tau\|\lambda\|_1 + \langle \lambda, A^T y\rangle) \\ &\text{subject to} \;\; \|A^T A \lambda\|_\infty \leq 1, \end{aligned} \tag{7}$$

where $\lambda \in \mathbb{R}^n$ is the dual optimization variable. We can derive the optimality conditions by recognizing that at the solution, the objectives in (3) and (7) will be equal, due to strong duality [17]. This fact, along with the complementary slackness property, tell us that $x^\star, \lambda^\star$ are solutions to (3) and (7) for a given value of τ if they satisfy the following optimality conditions [18]:

DS1. $A_{\Gamma_\lambda}^T(Ax^\star - y) = \tau z_\lambda$

DS2. $A_{\Gamma_x}^T A\lambda^\star = -z_x$

DS3. $\|A_{\Gamma_\lambda^c}^T(Ax^\star - y)\|_\infty < \tau$

DS4. $\|A_{\Gamma_x^c}^T A\lambda^\star\|_\infty < 1$,

where Γ_x and Γ_λ are the supports of $x^\star$ and $\lambda^\star$ respectively, z_x and z_λ are the sign sequences of $x^\star$ and $\lambda^\star$ on their respective supports. We call (DS1,DS3) the *primal constraints*, and (DS2,DS4) the *dual constraints*. It can be shown using standard convex optimization that these four conditions are necessary and sufficient for $(x^\star, \lambda^\star)$ to be the unique primal-dual solution pair. Furthermore, the active primal constraints correspond to the support of dual variable and the active dual constraints correspond to the support of primal variable. Therefore, from these optimality conditions we can see that the primal and dual solutions can be calculated using the knowledge of primal-dual supports and sign sequences $(\Gamma_x, \Gamma_\lambda, z_x, z_\lambda)$ as

$$x^\star = \begin{cases} (A_{\Gamma_\lambda}^T A_{\Gamma_x})^{-1}(A_{\Gamma_\lambda}^T y + \tau z_\lambda) & \text{on } \Gamma_x \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\lambda^\star = \begin{cases} -(A_{\Gamma_x}^T A_{\Gamma_\lambda})^{-1} z_x & \text{on } \Gamma_\lambda \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Note that the essential difference between the LASSO and DS optimality conditions (L1-L2) and (DS1-DS4), and their respective solutions (6) and (8), comes from the fact that in DS the primal and dual support $(\Gamma_x, \Gamma_\lambda)$ and sign sequence (z_x, z_λ) can be very different from each other. However, if at the given value of τ the solution of DS has same primal and dual supports, i.e., $\Gamma_\lambda = \Gamma_x$, and the primal and dual sign sequences are opposite to each other, i.e., $z_\lambda = -z_x$, then for that particular value of τ the solutions of LASSO and DS will be identical. We pursue this line of reasoning further in the following sections.

B. Homotopy paths

Some further insight into the working of LASSO and DS can be gained by looking at the solutions of (2) and (3) for different values of τ. Homotopy methods provide an efficient way to trace the solution paths of these problem for a range of values of τ [18]–[23].

LASSO: The complete solution path for LASSO can be traced by starting from a large value of τ where the solution of (2) is known (e.g., $\tau > \|A^T y\|_\infty$, where solution is a zero vector), and reduce it towards zero while updating the intermediate solutions. Suppose we are at an intermediate solution along the homotopy path, given as $x^\star$ in (6). As we reduce τ, the solution moves along a line with direction

$$\partial x = \begin{cases} (A_\Gamma^T A_\Gamma)^{-1} z & \text{on } \Gamma \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

until one of two things happens: an element in $x^\star$ shrinks to zero, removing it from the support of the new solution $x^\star$, or another constraint in (L) becomes active, adding a new element to the support of $x^\star$. At these so-called *critical points*, both the support of $x^\star$ and the direction of the solution path change. Also, at any point on the solution path it is straightforward to calculate how much we need to vary τ to take us to a critical point in either direction. Therefore, we start with a solution $x^\star$ at τ with support Γ and sign z, and reduce τ to the desired value while hopping from one critical point to the next. At each critical point along this path, a single element is either added to or removed from Γ, and the new direction can be computed from the old using a rank-1 update. Thus multiple solutions over a range of τ can be calculated at very little marginal cost.

Dantzig Selector: The homotopy path for DS can be traced in a similar way, with the additional requirement of keeping track of dual variable, and its support and signs. We start from a large value of τ and reduce τ gradually by updating the primal-dual supports and sign sequences at every critical point. As we change τ, the solution moves along a line in the direction

$$\partial x = \begin{cases} -(A_{\Gamma_\lambda}^T A_{\Gamma_x})^{-1} z_\lambda & \text{on } \Gamma_x \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

until one of the two things happens at a new critical point: an element in $x^\star$ shrinks to zero (removing an element from the support of $x^\star$) or an inactive primal constraint becomes active (adding an element to the support of $\lambda^\star$). We call this first phase the *primal update*. This gives us the value of $x^\star$ at the new critical point but the value of $\lambda^\star$ is still unknown. We use the information about the change in the support from the primal update phase to find the new value for the dual solution $\lambda^\star$ at this critical point, during which either an existing element in $\lambda^\star$ shrinks to zero (removing an element from the support of $\lambda^\star$) or an inactive dual constraint becomes active (adding an element to the support of $x^\star$). We call this second phase the *dual update*. For further details on the DS homotopy see [18], [23].

Note that the solutions of both LASSO and DS follow piecewise linear paths, in the respective directions (10) and (11) which depend only on the supports and signs at critical points. Therefore, the intuition tells us that if the solution of DS has same primal and dual support and opposite primal and dual signs at every value of τ, then the entire solution paths for LASSO and DS will be identical.

III. SHRINKABILITY AND S-STEP RECOVERY

In the previous section we observed that the solutions of DS and LASSO will be identical whenever the supports for primal-dual solution pair of DS are same and the sign sequences are opposite to each other. In this section we derive a set of conditions, under which LASSO and DS are guaranteed to have identical solution paths and the number of homotopy steps required for the recovery is same as the number of nonzero components in the sparse signal, i.e., S-steps for S-sparse signal recovery. We start with the case of noiseless measurements.

A. Noiseless measurements

Suppose we have noiseless measurements $y = Ax_0$ of the signal supported on the set Γ with sign sequence z on Γ. The following lemma gives us a set of conditions which ensure that (2) and (3) have same solution for small values of τ.

Lemma 1. *Let $x_0 \in \mathbb{R}^N$ be supported on a set Γ with $z := \text{sign}[x_\Gamma]$. Set $y = Ax_0$. Suppose that A obeys following three conditions with Γ and z:*

H1. A_Γ *is full rank,*
H2. $\|A_{\Gamma^c}^T A_\Gamma (A_\Gamma^T A_\Gamma)^{-1} z\|_\infty < 1$,
H3. $\text{sign}[(A_\Gamma^T A_\Gamma)^{-1} z] = z$.

Take

$$\lambda^\star = \begin{cases} -(A_\Gamma^T A_\Gamma)^{-1} z & \text{on } \Gamma \\ 0 & \text{on } \Gamma^c \end{cases},$$

and

$$x_\tau^\star = x_0 + \tau\lambda^\star. \quad (12)$$

Then for all τ in the range

$$0 \le \tau \le \tau_{crit} = \min_{\gamma\in\Gamma} \left(\frac{x_0(\gamma)}{-\lambda^\star(\gamma)} \right),$$

$x_\tau^\star$ is the unique solution to (2) *and* (3)*, and $\lambda^\star$ is the unique solution to* (7).

Proof: If A_Γ is full rank (H1), then the proposed $\lambda^\star$ is well-defined. We need to show that the proposed pair $(x^\star, \lambda^\star)$ obeys optimality conditions in L1-L2 and DS1-DS4. First note that $\Gamma_x = \Gamma_\lambda = \Gamma$. If H3 holds, then also $-z_\lambda = z_x = z$, and DS2 is satisfied. In addition, this makes H2 same as DS4. Finally, with $x_\tau^\star$ as in (12),

$$A^T(Ax_\tau^\star - y) = -\tau A^T A_\Gamma (A_\Gamma^T A_\Gamma)^{-1} z,$$

and so H2 implies L1, L2, DS1, and DS3. ■

Under the conditions H1-H3, we can interpret the solutions of (2) and (3) as "shrinkage" of x_0. As τ increases, the magnitude of all the nonzero entries in $x_\tau^\star$ decrease. Although, unlike soft-thresholding, the rate of decrease is not the same, the decrease at component γ is proportional to $\lambda^\star(\gamma)$. τ_{crit} is the value of τ for which one of the components in x_0 shrinks to zero.

It is natural, then, to ask if this "shrinkage" property holds for $x_{\tau_{crit}}^\star$ supported on $\Gamma_1 \subset \Gamma$, and if so, can we continue the shrinkage process until $x^\star = 0$. Note that since x_0 has only S elements, it will take exactly S steps to shrink it to zero. Conversely, it will take S steps to move from zero to x_0. To make this precise, we call x_0 *δ-shrinkable* with respect to A, for some $0 < \delta \le 1$ if the following shrinkage procedure terminates in `Success`:

1) Set $k = 0$, $\Gamma_0 = \Gamma$, and $z_0 = z$. Check that A_{Γ_0} is full rank; if so continue to 2, otherwise return `Failure`.
2) If $x_k = 0$, return `Success`.
3) Check that

$$\|A_{\Gamma_k^c}^T A_{\Gamma_k} (A_{\Gamma_k}^T A_{\Gamma_k})^{-1} z\|_\infty < \delta \quad \text{(H2')}$$

$$\text{sign}[(A_{\Gamma_k}^T A_{\Gamma_k})^{-1} z] = z \quad \text{(H3')}$$

If either condition fails, break and return `Failure`. (Note that if A_{Γ_0} is full rank, then A_{Γ_k} will be full rank for any $\Gamma_k \subset \Gamma_0$.)

4) Set

$$\lambda_k = \begin{cases} -(A_{\Gamma_k}^T A_{\Gamma_k})^{-1} z_k & \text{on} \quad \Gamma_k \\ 0 & \text{on} \quad \Gamma_k^c \end{cases},$$

$$\epsilon_{k+1} = \min_{\gamma \in \Gamma_k} \left(\frac{x_k(\gamma)}{-\lambda_k(\gamma)} \right),$$

$$x_{k+1} = x_k + \epsilon_{k+1}\lambda_k,$$

$$\gamma'_{k+1} = \arg\min_{\gamma \in \Gamma_k} \left(\frac{x_k(\gamma)}{-\lambda_k(\gamma)} \right),$$

$$\Gamma_{k+1} = \Gamma_k \backslash \gamma'_{k+1},$$

$$z_{k+1} = z_k \text{ restricted to } \Gamma_{k+1}.$$

5) Set $k \leftarrow k+1$, and return to step 2.

As x_0 shrinks to zero in the procedure above, $\tau = \sum_k \epsilon_{k+1}$ increases and the solution x_k follows the path of solutions to (2) and (3), as shown in the next lemma.

Lemma 2. *Suppose x_0 is δ-shrinkable with respect to A, and define $x_k, \lambda_k, \epsilon_k$ as above. For any given $\tau \in [0, \infty)$, let K be the largest integer such that*

$$\tau \geq \sum_{k=0}^{K-1} \epsilon_{k+1} =: E_K.$$

Then the solution to (2) *and* (3) *will be exactly*

$$x_\tau^\star = x_K + (\tau - E_K)\lambda_K.$$

Proof: Set $\theta = \tau - E_K$. Since x_0 is δ-shrinkable, we can write

$$x_\tau^\star = x_0 + \sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K$$

and

$$A^T(Ax_\tau^\star - y) = A^T(Ax_0 - y) + A^T A \left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K \right).$$

Since $\Gamma_K \subset \Gamma_k$ for all $k = 0, \ldots, K-1$, let z_K denote sign of x_K on Γ_K,

$$\begin{aligned} A_{\Gamma_K}^T(Ax_\tau^\star - y) &= -(\sum_{k=0}^{K-1} \epsilon_{k+1} + \theta) z_K \\ &= -\tau z_K. \end{aligned}$$

Similarly, since

$$|a_\gamma^T A\lambda_k| \leq 1 \quad \text{for } \gamma \in \Gamma_K^c \text{ and for all} \quad k = 1, \ldots, K-1,$$

and

$$|a_\gamma^T A\lambda_K| < 1 \quad \text{for all} \quad \gamma \in \Gamma_K^c,$$

thus for all $\gamma \in \Gamma_K^c$

$$\left| a_\gamma^T(Ax_0 - y) + a_\gamma^T A \left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \delta\lambda_K \right) \right| < (E_K + \theta) = \tau.$$

Hence L1, L2, and DS1,DS3 are satisfied. DS2 and DS4 are automatically satisfied with our choice of λ_K. ■

Remark. *An immediate consequence of Lemma 2 is that since x_0 is shrinkable and it takes S steps to shrink to zero, and the path it takes coincides with the homotopy paths of LASSO and DS. Thus a LASSO or DS homotopy algorithm starting from zero will terminate in exactly S steps, recovering the original signal x_0 with total complexity $O(SMN)$.*

The shrinkability conditions (H2,H3) have a distinct effect in the working of the homotopy algorithm. The correlation condition (H2) ensures that only correct elements enter the support, and once an element has entered the support the sign condition (H3) ensures that it never leaves the support, hence providing the solution of S-sparse signal in S homotopy steps. Also note that, when the S-step solution property holds, the homotopy algorithms for both LASSO and DS reduce to LARS [20]. Conditions similar to H3 have also been studied in [23], [24], to establish the equivalence between LASSO and DS solutions.

B. Noisy measurements

The analysis for the shrinkability in the presence of noise is very similar to the noiseless case. The main difference is that we cannot recover the original signal exactly. Therefore, we start with some other shrinkable signal as the starting point of the shrinkage procedure, while the remaining steps are almost identical to the noiseless case.

Consider the noisy measurements in (1), where x_0 is supported on the set Γ_0. We assume that

$$\|A_{\Gamma_0^c}^T(I - P[\Gamma_0])e\|_\infty \leq \rho, \tag{13}$$

where $P[\Gamma_0]$ denotes projection onto the space spanned by columns in A_{Γ_0}. The choice of final threshold parameter τ in (2) and (3) depends on the noise level. The following lemma shows that if the threshold τ in (2) and (3) is chosen safely above the noise level, the LASSO and DS have same solution.

The following is a variation of a lemma from [25].

Lemma 3. *Given noisy measurements $y = Ax_0 + e$ of the signal x_0 supported on Γ_0, where e satisfies* (13)*. Let x_{orc} be the oracle estimate*

$$x_{\text{orc}} = \begin{cases} (A_{\Gamma_0}^T A_{\Gamma_0})^{-1} A_{\Gamma_0}^T y & \textit{on } \Gamma \\ 0 & \textit{otherwise.} \end{cases} \tag{14}$$

If x_{orc} is δ-shrinkable with respect to A, then for $\tau > \frac{\rho}{1-\delta}$ the solution $x_\tau^\star$ obtained by running (2) *and* (3) *will be identical.*

Proof: The proof is quite similar to the argument in Lemma 2, where we replace x_0 with x_{orc} in the shrinkage procedure. Since $\lambda^\star$ is a dual solution to (2) by construction, therefore it obeys DS2,DS4. We only need to show that the solution $x^\star$ obeys L1,L2, and DS1,DS3. Since x_{orc} is δ-shrinkable, let us first write $x_\tau^\star$ with support Γ at some $\tau > \frac{\rho}{1-\delta}$ as

$$x_\tau^\star = x_{\text{orc}} + \sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K,$$

where $\tau = \sum_{k=0}^{K-1} \epsilon_{k+1} + \theta$. This gives us

$$Ax^\star_\tau - y = Ax_{\text{orc}} + A\left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K\right) - y$$
$$= A\left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K\right) - (I - P[\Gamma_0])e.$$

Recall that $\Gamma \subset \Gamma_0$. Suppose first that $\gamma \in \Gamma_0$ but $\gamma \notin \Gamma$. Then $\langle a_\gamma, (I - P[\Gamma_0])e\rangle = 0$ and

$$|\langle a_\gamma, A\left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K\right)\rangle| < \delta\tau.$$

Now suppose that $\gamma \in \Gamma_0^c$. Then we are guaranteed that

$$|\langle a_\gamma, A\left(\sum_{k=0}^{K-1} \epsilon_{k+1}\lambda_k + \theta\lambda_K\right)\rangle| < \delta\tau.$$

and

$$|\langle a_\gamma, Ax^\star_\tau - y\rangle| < \delta\tau + \rho < \tau,$$

where the last inequality follows from (13). Similar procedure can be used to show that in at most S steps x_{orc} shrinks to zero. ■

Although in Lemma 3 we only discuss the equivalence between LASSO and DS solutions and the S-step property, but similar conditions can be used to prove various near-optimality results for both LASSO and DS solutions, where we need some conditions (very similar to H1-H3) only for the final support of the solution [12], [25].

IV. Shrinkability conditions and incoherence

In this section we examine the types of signals and matrices which satisfy the shrinkability conditions. In particular, we show that an incoherent matrix satisfies δ-shrinkability conditions for any S-sparse signal.

In order to show that a vector x with support Γ and sign sequence z on Γ is δ-shrinkable for some given value of $\delta \leq 1$ with respect to the matrix A, we need to show that conditions H1-H3 are satisfied on every $\Gamma_k \subset \Gamma$ and respective sign sequence. If we assume that A_Γ is full rank[2] then A_{Γ_k} will also be full rank for any $\Gamma_k \subset \Gamma$ [26]. Let us define $G = I - A_\Gamma^T A_\Gamma$. We claim that condition (H2) and (H3) will be satisfied if $\|G\| < 1$ and

$$\max_{\gamma \in \{1,\dots,n\}} |\langle (A_\Gamma^T A_\Gamma)^{-1} Y_\gamma, z\rangle| < \delta, \tag{15}$$

with

$$Y_\gamma = \begin{cases} A_\Gamma^T a_\gamma & \gamma \in \Gamma^c \\ A_\Gamma^T a_\gamma - \mathbf{1}_\gamma & \gamma \in \Gamma \end{cases}, \tag{16}$$

where a_γ is the column of A indexed by γ, and $\mathbf{1}_\gamma$ is a vector which is equal to 1 at γ and zero elsewhere.

To see why this is true, first note that (H2) is same as

$$\max_{\gamma \in \Gamma^c} |\langle a_\gamma, A_\Gamma (A_\Gamma^T A_\Gamma)^{-1} z\rangle| < \delta.$$

[2] Geršgorin's disc theorem [26] can be used to show that $\|I - A_\Gamma^T A_\Gamma\| < 1$ if $S < 1 + \frac{1}{\mu}$, which implies that A_Γ has full rank.

Whenever $\|G\| < 1$, the Neumann's series $\sum_{\ell=0}^{\infty} G^l$ converges to the inverse $(I - G)^{-1}$. Thus we can write $(A_\Gamma^T A_\Gamma)^{-1} z$ as

$$(A_\Gamma^T A_\Gamma)^{-1} z = (I - G)^{-1} z = \sum_{\ell=0}^{\infty} G^\ell z = \left(z + \sum_{\ell=1}^{\infty} G^\ell z\right),$$

and condition (H3) will be satisfied for any $\delta \leq 1$ if

$$\|(A_\Gamma^T A_\Gamma)^{-1} z - z\|_\infty = \left\|\sum_{\ell=1}^{\infty} G^\ell z\right\|_\infty < \delta. \tag{17}$$

We can rewrite (17) as

$$\max_{\gamma \in \Gamma} \left|\langle \mathbf{1}_\gamma, \sum_{\ell=1}^{\infty} G^\ell z\rangle\right| = \max_{\gamma \in \Gamma} \left|\langle \sum_{\ell=1}^{\infty} G^\ell \mathbf{1}_\gamma, z\rangle\right|$$
$$= \max_{\gamma \in \Gamma} \left|\langle \sum_{\ell=1}^{\infty} G^{\ell-1} g_\gamma, z\rangle\right|$$
$$= \max_{\gamma \in \Gamma} |\langle (A_\Gamma^T A_\Gamma)^{-1} g_\gamma, z\rangle|$$

where g_γ is the column of G indexed by γ, $g_\gamma = \mathbf{1}_\gamma - A_\Gamma^T a_\gamma$. The first equality above comes from the self-adjointness of G, the second comes from simple fact that $g_\gamma = G\mathbf{1}_\gamma$, and the third because $\sum_{\ell \geq 1} G^{\ell-1} = \sum_{\ell \geq 0} G^\ell = (A_\Gamma^T A_\Gamma)^{-1}$.

The following theorem gives conditions on coherence of matrix A and sparsity S of signal x under which (15) is satisfied for Γ. The argument can be easily extended to any support $\Gamma_k \subset \Gamma$; a crucial step needed to satisfy the shrinkability conditions H1-H3.

Theorem 1 ([16]). *Let A be the incoherent matrix with coherence μ. Let x be an S-sparse signal supported on Γ with sign sequence z, and Y_γ be as defined in* (16)*.*

$$S \leq \frac{\delta}{1+\delta}\left(1 + \frac{1}{\mu}\right) \tag{18}$$

then

$$\max_{\gamma \in \{1,\dots,n\}} |\langle (A_\Gamma^T A_\Gamma)^{-1} Y_\gamma, z\rangle| < \delta. \tag{15}$$

Proof: This result can be proved easily by following the arguments in [16, Theorem 3], where the left side in (15) can be bounded by the absolute values of the individual components as

$$\langle Y_\gamma, (A_\Gamma^T A_\Gamma)^{-1} z\rangle \leq \langle |Y_\gamma^T|, |(A_\Gamma^T A_\Gamma)^{-1}||z|\rangle.$$

The only difference is that our definition of Y_γ is slightly different. For the case when $\gamma \notin \Gamma$, the argument is exactly same. For the case where $\gamma \in \Gamma$, we can use similar argument because γth component of Y_γ will become zero (since $\|a_\gamma\|_2 = 1$). ■

V. Discussion

In the last section we showed that if $S \leq \frac{\delta}{1+\delta}(1 + \frac{1}{\mu})$ then both LASSO and DS have identical solution paths and that the final solution can be recovered in at most S homotopy steps. The result, although aimed at the incoherent matrices, is not yet optimal for the random matrices usually used in CS, e.g., Gaussian or Bernoulli. The CS results for random matrices suggest that for an S-sparse signal typically $M = O(S \cdot \log N)$ measurements are sufficient to reliably recover the signal. Whereas, the coherence based results for similar matrices would unfortunately require $M = O(S^2 \cdot \log N)$ measurements. Although this comparison is not fair, as we not only guarantee signal recovery but recovery in S-steps, but it would still be desirable to improve the results for S-step recovery.

Under the incoherence conditions in Theorem 1, shrinkage conditions H1-H3 hold for all possible supports of size at most S. Whereas, to prove the S-step solution property for a signal supported on Γ_0, we only need to show that (15) holds for all subsets $\Gamma_k \subset \Gamma_0$. Note that for a random matrix A, we can show that for a *fixed* support Γ_0 of size S and random sign sequence z, if $M = O(S \cdot \log N)$ then (15) holds with very high probability [27]. The main challenge we face while trying to prove the so-called S-step property with $O(S \cdot \log N)$ measurements for random ensembles, is the independence of elements as the signal shrinks. We want to show that (15) holds for all $\Gamma_k \subset \Gamma_0$ after every shrinkage step, but the independence between random sign sequence z and $(A_\Gamma^T A_\Gamma)^{-1} Y_\gamma$, which is a crucial part in proving $O(S \cdot \log N)$ result for the Γ_0, does not exist after the first shrinkage step. On the other hand, if we use a union bound on all possible supports (a set with 2^S elements), the result would be again on the order of $S^2 \cdot \log N$. Although we know that there are S possible supports involved in the shrinkage procedure, but they are not fixed a priori.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[3] D. Donoho, “Compressed sensing,” *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[4] D. Donoho and M. Elad, “Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization,” *Proc. Nat. Aca. Sci*, vol. 100, no. 5, pp. 2197–2202, 2003.

[5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] E. Candès and T. Tao, “The Dantzig selector: Statistical estimation when p is much larger than n,” *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[8] D. Donoho, M. Elad, and V. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, 2006.

[9] E. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Comm. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, 2006.

[10] N. Meinshausen and B. Yu, “Lasso-type recovery of sparse representations for high-dimensional data,” *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2008.

[11] P. Bickel and A. T. Ya’acov Ritov, “Simultaneous analysis of Lasso and Dantzig selector,” *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

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

[13] D. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *Information Theory, IEEE Transactions on*, vol. 47, no. 7, pp. 2845–2862, Nov 2001.

[14] E. Candès and T. Tao, “Decoding by linear programming,” *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[15] E. Candès, “Compressive sampling,” *Proceedings of the International Congress of Mathematicians, Madrid, Spain*, vol. 3, pp. 1433–1452, 2006.

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

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[18] M. S. Asif, “Primal Dual Pursuit: A homotopy based algorithm for the Dantzig selector,” Master’s thesis, Georgia Institute of Technology, August 2008.

[19] M. Osborne, B. Presnell, and B. Turlach, “A new approach to variable selection in least squares problems,” *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, 2000.

[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[21] D. Malioutov, M. Cetin, and A. Willsky, “Homotopy continuation for sparse signal representation,” *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, vol. 5, pp. v/733–v/736, March 2005.

[22] D. L. Donoho and Y. Tsaig, “Fast solution of ℓ_1-norm minimization problems when the solution may be sparse,” *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, 2008.

[23] G. James, P. Radchenko, and J. Lv, “The DASSO algorithm for fitting the Dantzig selector and the Lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 71, pp. 127–142, 2009.

[24] N. Meinshausen, G. Rocha, and B. Yu, “Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig,” *Annals of Statistics*, vol. 35, no. 6, pp. 2373–2384, 2007.

[25] J. Romberg, “The Dantzig selector and generalized thresholding,” *Information Sciences and Systems, (CISS). 42nd Annual Conference on*, pp. 22–25, March 2008.

[26] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[27] E. Candès and J. Romberg, “Sparsity and incoherence in compressive sampling,” *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.