Performance Analysis of LASSO-Based Signal Parameter Estimation

Ashkan Panahi, Student Member, IEEE, and Mats Viberg, Fellow, IEEE,

Abstract

The Least Absolute Shrinkage and Selection Operator (LASSO) has gained attention in a wide class of continuous parametric estimation problems with promising results. In these applications the desired information is given by the unknown support of a sparse vector represented by some continuous parameters. The objective of this work is to provide a theoretical analysis of such a LASSO-based estimator in terms of the classical statistical measures, i.e. variance and bias. Employing LASSO, which only admits a discrete set of candidate regressors, to a continuous valued problem complicates the analysis significantly. We respond to this dilemma by introducing a new approach considering an intermediate sparse estimator over the continuum, which we show to be asymptotically equivalent to LASSO but easier to analyze. This provides us theoretical expressions for the LASSO-based estimation error in the asymptotic case of high SNR and dense grids. We specifically show that beyond the RIP-based results, such an asymptotic case may be consistent in many individual cases of interest. Without loss of generality, we present the comparative numerical results in the context of Direction of Arrival (DOA) estimation using a sensor array.

Index Terms

Compressed Sensing, performance analysis, sparse estimation, sparse regression, continuous regression

I. INTRODUCTION

The last decade witnessed the rapid emergence of the sparse models and their corresponding techniques in many traditional signal processing areas [1]–[5]. Although the basic principles of sparsity are easily recognized in many conventional studies, more exotic approaches such as the $\ell_1$ penalized least squares, well known as LASSO [6] (Least Absolute Shrinkage and Selection Operator), basis pursuit [7] or global matched filter [8], and its variants [9]–[12] have been unknown to the majority of the researchers
until recently. The LASSO method succeeded to fascinate scholars in different experimental instances, yet it is viewed skeptically for its ambiguous theoretical properties. One reason is that the previous brilliant theoretical studies such as [13] can not be simply compared to that of the other well know parametric estimation methods in certain contexts. Note that, many applications such as the sensor array problems mainly concern the continuous support parameter estimates, which need to be discretized (finitely sampled) in order to apply LASSO [14]. The corresponding error analysis is significantly more complicated since the continuous interpolation from the discretized estimates is not straightforward. Relative to the subject of continuous inversion/estimation in partly linear models, the current study offers an approach to fill this theoretical gap.

LASSO belongs to a rich family of techniques for estimating a large set of real or complex variables, only a few of which are assumed to linearly contribute to the given observation. The group of contributors, also known as the support, is not priorly known. This is generally known as a sparse linear regression problem. In such a sparse regression, finding a proper representation of the observed data by trying all possibilities of a small support implies a heavy computational burden, especially when the number of unknown variables grows. Accordingly, many suboptimal regression techniques such as matching pursuit [15] and orthogonal matching pursuit [16] have been proposed, to which LASSO strongly connects. It is also worth to mention some interesting later developments such as Least Angle Regression (LARS) [17], and Compressive Sampling Matching Pursuit (CoSaMP) [18].

Recently, LASSO has been applied to a class of parametric inversion (estimation) problems. To avoid abstraction, we will take the illustrative and typical example of estimating multiple Directions of Arrival (DOAs), where the bases, also known as steering vectors, correspond to directions and the observed vector is modeled as a superposition of these bases scaled by the complex-valued source amplitudes. This clearly results in a sparse regression. To be more specific, in the cases of interest herein the regressors (e.g. steering vectors) depend on a continuous "position parameter" (e.g. the direction in the DOA estimation case) in a non-linear fashion and are superimposed after scaling by the corresponding "amplitude parameters" to create the "observation vector". We refer to this type of problems as the "continuous regression". Other familiar continuous regression problems are frequency and spectrum estimation [19], [20], sensor array analysis [14], image processing [21], [22], tomography [23], [24], seismology [5], etc. The purpose of this work is to provide the analysis of LASSO applied to such continuous regression problems.

We mainly focus the error analysis of LASSO on estimating the position parameters. From a practical perspective, an interesting and popular case is that of the high Signal to Noise Ratio (SNR), when the
noise level in observed data is low. This is well studied by extensive Monte Carlo (MC) methods in both cases of single and multiple snapshots, e.g. in [25] for the DOA example case. However, MC results are specific and provide a limited insight. Thus, it is still beneficial to provide a theoretical argument. Note that the theoretical results mainly dealing with the question of Compressed Sensing (CS) [13], [21], [26], [27] may not be applicable, since they are generally in terms of the \( \ell_2 \) distance of the sparse vectors. Note that, to be comparable to the conventional continuous techniques, such as spectral methods, subspace techniques, Maximum Likelihood (ML), etc as well as the continuous bounds such as the Cramer-Rao Lower Bound (CRLB) [28], the error analysis should be presented in terms of the statistical bias and mean squared error of the position parameter estimates. This is simply seen to be out of the scope of the \( \ell_2 \)-type error analysis. A similar concern is observed in [29] where the analysis of the support estimates is separated by the so-called oracle property. However, it includes only an asymptotic consistency analysis and may not provide the desirable error bounds.

As already pointed out, we may also note that the direct application of LASSO to continuous regression involves discretizing (finitely sampling) the space of position parameters, which further complicates the analysis. The sampling also seems inevitable and highly restrictive by the previously mentioned CS-based result, known as the Restricted Isometry Property (RIP). When the true position parameters are off-grid, this also brings confusion as it is not clear how the resulting on-grid estimates, usually consisting of the neighbor samples should be combined to provide continuous ones. Many suboptimal methods [30]–[32] have been recently proposed. However, we show that this is not theoretically necessary by putting the analysis in a more practical framework, in which there is no theoretical restriction on the level of discretization. This leads to consider a theoretical continuous estimation scheme, which asymptotically coincides with LASSO for a dense grid. This continuous estimator, denoted by ”Continuous LASSO” (CLASS) is more straightforward to analyze. Although difficult to implement, CLASS plays the role of a bridge to the asymptotic analysis of LASSO. We then show that the more practical and numerically reliable fine-grid LASSO is also strongly connected to that of the asymptotically dense one, so that the analysis of the latter (which is the same as the analysis of CLASS) applies to the former.

Following the above mentioned method, our current results show that beyond the prediction of the RIP-based argument, LASSO can resolve well separated bases with an infinite dynamic range of amplitudes. However, there exists a resolution limit comparable to that of the spectral techniques. Note that the latter have a finite consistent dynamic range compared to LASSO, i.e. it may not resolve sources with too different amplitudes due to the sidelobe effect. The LASSO technique also has support estimates biased proportionally to the Regularization Parameter (RP), which affects its performance in terms of Mean
Square Error (MSE). However, numerical computations here show that LASSO is nearly optimal in the high SNR case in terms of error variance.

A. Notation

In this paper we refer to the vector and the matrix entities by small and capital bold letters, respectively. The vector \( \mathbf{\theta} \) may also especially denote the (unordered) set of its elements, whenever there is no risk of confusion. The transpose and conjugate transpose are shown by \((\cdot)^T\) and \((\cdot)^H\). The real part is denoted by \(\text{Re}(\cdot)\) and \(\odot\) denotes the element-wise product. The Landau notations \(f = \omega(g)\) and \(f = O(g)\) for the two arbitrary functions \(f\) and \(g\) imply that \(f/g\) tends to zero and \(f/g\) is bounded, respectively in an asymptotic case inferred from the context.

II. Mathematical Modeling

In this section, we formalize what we previously introduced as continuous regression and its conventional solution by (discretized) LASSO.

A. Continuous Sparse Representation

Consider a closed index set \( \Theta \subset \mathbb{R} \) and a collection of complex basis vectors \( \mathbf{a}(\theta) \in \mathbb{C}^m \) indexed by the elements \( \theta \in \Theta \). For our purpose, it suffices to assume that \( \mathbf{a}(\theta) \) is a smooth function of \( \theta \), where it is referred to as a manifold. In most applications of interest, the dependence of \( \mathbf{a}(\theta) \) on \( \theta \) is non-linear. Consider a set of \( n \) indexes \( \mathbf{\theta} = [\theta_1, \theta_2, \ldots, \theta_n] \), its corresponding discrete-time amplitudes \( \mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)] \in \mathbb{C}^n \) and the noise vectors \( \mathbf{n}(t) \in \mathbb{C}^m \) for \( t = 1, 2, \ldots, T \). For an arbitrary sequence of \( T \) observed vectors \( \mathbf{x}(t) \in \mathbb{C}^m \) for \( t = 1, 2, \ldots, T \), the three sequences of indexes, components and noise provide a static noisy representation, if

\[
\mathbf{x}(t) = \mathbf{A}(\mathbf{\theta})\mathbf{s}(t) + \mathbf{n}(t),
\]

where \( \mathbf{A}(\mathbf{\theta}) = [\mathbf{a}(\theta_1) \ \mathbf{a}(\theta_2) \ldots \mathbf{a}(\theta_n)] \) is the collection of basis vectors corresponding to \( \mathbf{\theta} \). As seen, (1) provides a noisy linear representation of \( \mathbf{x}(t) \) in the linear space spanned by the bases corresponding to the elements of \( \mathbf{\theta} \). The model in (1) is called static, since the index parameters \( \mathbf{\theta} \) may not change in different times \( t \). The parameter \( n \) is referred to as the order or the cardinality of the representation. Finally, the expression in (1) can be rewritten as

\[
\mathbf{X} = \mathbf{A}(\mathbf{\theta})\mathbf{S} + \mathbf{N},
\]
where \( X = [x(1) \, x(2) \ldots x(T)] \), \( S = [s(1) \, s(2) \ldots s(T)] \) and \( N = [n(1) \, n(2) \ldots n(T)] \) are the matrix collection of the observations, the amplitudes and the noise respectively. More formally we may also refer to the index parameters \( \theta \) as the position parameters.

According to (2), one can roughly state that the element \( s_k(t) \) contains the contribution of basis \( a(\theta_k) \) in representing \( x(t) \). If \( s_k(t) = 0 \), the corresponding parameter \( \theta_k \) can be removed from the representation at time \( t \). Such an amplitude and the corresponding index are referred to as inactive at time \( t \). If an index \( \theta_k \) is inactive at every time, it is totally inactive or simply an inactive index, which means that the \( k \)th row in \( S \) corresponding to \( s_k(t) \) for \( t = 1, 2, \ldots, T \) is completely zero. We refer to a representation by \( \theta \) and \( S \) as reducible if a row, say the \( k \)th row in \( S \) is zero, i.e. the corresponding index \( \theta_k \) is totally inactive. In this case, \( \theta \) can be reduced to \( \theta' = \theta - \{\theta_k\} \) by removing the corresponding zero row in \( S \). If a representation is repeatedly reduced until an irreducible one is obtained, we refer to the latter as the irreducible root of the former.

To obtain a desirable representation without any further assumption, the model in (1) is strongly underdetermined. Then, a general consideration is that the noise level should be as small as possible. However, the problem remains ambiguous even if the noise is assumed to vanish (\( n(t) = 0 \)), which is referred to as the noiseless case. Thus, the stronger condition of sparsity is considered, which essentially bounds the order of the representation. In general, there is a trade-off between the noise level and the order, and different sparse techniques attempt to obtain the best of this compromise.

**B. The Sensor Array Example**

As an example, we consider the planar Direction Of Arrival (DOA) estimation problem, in which a set of \( m \) sensors listen to \( n \) far and narrow band sources and decide on their directions. The received data is modeled by (2), where the basis manifold is given by (3)

\[
a(\theta) = \left[ e^{j \frac{2\pi}{d} r_1 \cos(\theta - \rho_1)} \, e^{j \frac{2\pi}{d} r_2 \cos(\theta - \rho_2)} \ldots \right.
\]

\[
\left. e^{j \frac{2\pi}{d} r_m \cos(\theta - \rho_m)} \right]^T,
\]

in which \((r_i, \rho_i)\) is the polar coordinate pair of the \( i \)th sensor \((i = 1, 2, \ldots, m)\) and \( d \) is the wavelength at the central frequency. Then, the goal is to estimate \( \theta \) which represents the directions given \( X \). Obviously, the problem is defined in a complex-valued space of variables. For simplicity and without loss of generality, we focus on the half-wavelength \( (r_i = \frac{(i-1)d}{2}) \) Uniform Linear Array (ULA), which means that \( \rho_i = 0 \). In this case, it is more convenient to write (3) in terms of the electrical angle \( \phi = \pi \cos \theta \). The ULA manifold resembles the classical Fourier basis, when represented in terms of the electrical angle.
Thus, the sensor array example essentially includes other applications such as frequency estimation and sampling.

C. Conventional LASSO-based Solution

Attaining a desirable level of noise and order in (2) is challenging. A recently proposed approach is to first discretize the parameter space and next reformulate the model in the sparse framework [14]. Note that discretization will naturally introduce an additional quantization noise to the parameter estimates.

Consider a big set of \( N \) candidate points \( \theta_G = [\theta_1^g, \theta_2^g, \ldots, \theta_N^g] \subset \Theta \). Let us denote \( A^g = A(\theta_G) \). Then, the canonical LASSO representation is given by

\[
\hat{S}^g(\lambda) = \arg \min_{S^g} \frac{1}{2} \|X - A^g S^g\|_F^2 + \lambda \|S^g\|_{1,2}, \tag{4}
\]

where, for every \( S \) with \( S_{ij} \) as the element in the \( i \)th row and \( j \)th column we define

\[
\|S\|_{1,2} = \sum_i \gamma_i, s,
\]

\[
\gamma_i, s = \sqrt{\sum_j |S_{ij}|^2}. \tag{5b}
\]

We further introduce the support \( \hat{\theta}^g(\lambda) \subset \theta_G \) as the set of all indexes \( \theta_i^g \) corresponding to the nonzero rows in \( \hat{S}(\lambda) \). We also refer to (4) as the \( P_g \) optimization. As explained in [6], the estimate can be calculated in the noiseless case by solving

\[
\hat{S}^g_{nl} = \arg \min_{S^g} \|S^g\|_{1,2} \text{ subject to } A^g S^g = X, \tag{6}
\]

with the index estimates \( \hat{\theta}^g_{nl} \) corresponding to the non-zero amplitudes. We will refer to this optimization as \( P_{g,nl} \). An interesting observation is that the solution \( \hat{S}^g(\lambda) \) is a continuous matrix function of \( \lambda \) and tends to \( \hat{S}^g_{nl} \) as \( \lambda \to 0 \). This is referred to as the homotopy rule. Thus, the noiseless case may be defined as an analytical extension of \( P_g \) to \( \lambda = 0 \), which is not possible by direct evaluation. Note that the desired estimate \( \hat{\theta} \) is always chosen from the fixed grid \( \theta^g \).

The LASSO optimization in (4) is convex. Thus, its global optimality is guaranteed by the local Karush-Kuhn-Tucker (KKT) condition which can also be stated as follows.

Lemma 1. Suppose \( \hat{S} \) is an \( N \times T \) matrix with \( \hat{S}_i \) as the \( i \)th row and \( \hat{\theta}^g \) as its support (active indexes). By introducing \( \hat{N} = X - A^g \hat{S} \), the matrix \( \hat{S} \) is a global minimum in (4) if and only if first,

\[
a^H(\hat{\theta}^g_i) \hat{N} = \lambda \gamma_{i,\hat{S}}^{-1} \hat{S}_i, \tag{7}
\]
for every $\hat{\theta}_i \in \hat{\theta}^g$ corresponding to an active row in $\hat{S}$ and second,

$$\|a^H(\hat{\theta}_i)\hat{N}\|_2 \leq \lambda,$$

for $i = 1, 2, \ldots, N$. (See [33, Theorem 1] for a very similar discussion and proof and [34] for a proof in the real-valued case.)

III. CONTINUOUS LASSO

The discretization step in the conventional application of LASSO remarkably complicates the analysis of LASSO in estimating position parameters. In simple words, the fixed grid method is not satisfactory, as the true index parameters may possess off-grid values. To overcome this, we seek an intermediate continuous estimator, closely related to LASSO which is easier to analyze. One natural approach is to directly estimate the continuous index parameters by letting the grid $\theta_G$ in (4) vary so that $A^g$ covers all possible index combinations. For example, if the grid size increases, one may assume that the summations tend to proper integrals. This path is pursued in [35] for a slightly different purpose of developing a super-resolution theory. However, due to the technical difficulties of redefining the index estimates as an irregularly thin support of a function, we avoid such a method, and instead introduce the following extension of $P_g$. Later in this section, we show the equivalence of this method to the former integral extension of [35] and its deep relation to the Global Matched Filters (GMF) [8] which also does not admit discretization in principle.

A. Constructing CLASS

In [36] the authors look back to the optimality condition in Lemma [1] and discuss the possibility to consider it for all the index continuum $\Theta$ rather than $\theta^g$. To attain a point satisfying the generalized conditions, it also provides a reliable program generalizing the homotopy-based LASSO [33], [37], [38]. It finally empirically observes that the resulting algorithm provides satisfactory estimates. It is simple to see that our above quest for an intermediate LASSO-connected continuous estimator may be fulfilled by an optimization equivalently satisfying the generalized conditions in [36]. We fist introduce this so-called CLASS optimization in the sequel and later show its close relation to the conventional LASSO by a dense discretization.
The CLASS method is formally defined as the optimization

$$\left( \hat{\theta}(\lambda), \hat{S}(\lambda) \right) = \underset{S, \theta}{\text{argmin}} \frac{1}{2} \|X - A(\theta)S\|_F^2 + \lambda \|S\|_{1,2}$$

subject to

$$(S, \theta) \text{ is irreducible}$$

in the noisy case, where $\theta$ can be any finite subset of $\Theta$. Comparing (9) to (4), one may regard CLASS as a discretized LASSO with an adaptive grid $\theta$ substituting $\theta_G$. The irreducibility condition in (9), explained in Section II-A and later with more details, is crucial due to the following observation. For now, assume that a minimal pair $(\theta_0, S_0)$ of (9) exists. Since $\theta_0$ is a finite set, it is possible to add an arbitrary new index and a corresponding all-zero row in $S$ (we remind the discussion in Section II-A). This will not change the value of the cost in (9), which means that a new minimum point is reached. As seen, the set of unconstrained solutions of (9) is large and connected by the actions of adding and removing an inactive index, thus sharing a common irreducible root. The constraint in (9) introduces this irreducible root as the sole solution of CLASS. Finally, note that the irreducibility constraint does not exclude the possibility of multiple solutions to (9). However, the multiple solutions are of different nature as they are not connected by the reduction action.

B. Existence of the CLASS Solution

The existence of the CLASS solution is not obvious, since the dimension of $\theta$ is allowed to freely increase, which provides the opportunity to gradually decrease the cost in (9) without attaining any minimum. The fact that the minimal point exists may imply that increasing the dimension of $\theta$ is not beneficial for a higher order than that of the optimal point. For convenience, we can separate the analysis of CLASS into two steps. First, we fix $\theta$ to any arbitrary finite subset of $\Theta$, which results in the same type of optimization as in $P_g$. At this step, we denote

$$P(X, \theta, \lambda) = \min_S \Psi(X, \theta, S, \lambda),$$

where

$$\Psi(X, \theta, S, \lambda) = \frac{1}{2} \|X - A(\theta)S\|_F^2 + \lambda \|S\|_{1,2}.$$  (11)

Next, we minimize the function $P$ in (10) over all possible $\theta$ vectors. Note that from (9) the minimization should only be taken over the $\theta$ vectors, for which there does not exist a minimal point $S$ in (10) with a
zero row. We refer to such an index vector \( \theta \) as an irreducible vector in (9). Then, the second optimization can be written as

\[
\hat{\theta}(X, \lambda) = \arg\min\limits_{\theta} P(X, \theta, \lambda)
\]

subject to \( \theta \) is irreducible

(12)

To assure that the above two-stage optimization is equivalent to (9), take an arbitrary pair \((S, \theta)\). We show in the following that \( \Psi(X, \theta, S, \lambda) \geq \Psi(X, \hat{\theta}(X, \lambda), \hat{S}(X, \lambda), \lambda) \) where \( \hat{\theta}(X, \lambda) \) is given by (12) and \( \hat{S}(X, \lambda) \) is its corresponding minimum point in (10). This implies the minimality of the resulting representation from the two-stage procedure in (9). To see this note that

\[
\Psi(X, \theta, S, \lambda) \geq P(X, \theta, \lambda) = \Psi(X, \hat{\theta}(\theta), \hat{S}(\theta), \lambda),
\]

where \( \hat{S}(\theta) \) is the minimum point in (10). Note that if \((\theta, \hat{S}(\theta))\) is reducible, denoting its irreducible root by \((\theta_r, \hat{S}_r(\theta))\), we have that \( \hat{S}_r(\theta) \) also minimizes (10) for \( \theta_r \), and consequently \( P(X, \theta_r, \lambda) = P(X, \hat{\theta}_r, \lambda) \). To see this take an arbitrary \( S' \) compatible with \( \theta_r \) and add enough zero rows to obtain \( S'_c \) compatible with \( \theta \). Then, \( \Psi(X, \theta_r, S', \lambda) = \Psi(X, \theta, S'_c, \lambda) \geq \Psi(X, \theta, \hat{S}(\theta), \lambda) = \Psi(X, \theta_r, \hat{S}_r(\theta), \lambda) \) from the fact that adding and removing zero rows does not change the LASSO cost \( \Psi \). Finally, from (12), \( \Psi(X, \theta, S, \lambda) \geq P(X, \theta, \lambda) = P(X, \theta_r, \lambda) \geq P(X, \hat{\theta}, \lambda) = \Psi(X, \hat{\theta}(X, \lambda), \hat{S}(X, \lambda), \lambda) \).

From (10), it is also simple to see that the first optimization step in (10) always attains a minimum point. Finally, if the dimension \( n \) of \( \theta \) is greater than \( 2m \) it is reducible, i.e. there exists a minimum point \( S_0 \) in (10), which has an all-zero row. This can be immediately seen from [39, Chapter 2.6] and the structure of the extreme points of the \( \ell_1 \) ball as a polyhedral set. As the order of an irreducible solution of CLASS is bounded by \( 2m \), it is also simple to see that the minimization over \( P \) in (12) for such a bounded order of \( \theta \) always has a minimum point.

We refer to (12) which is equivalent to CLASS in (9) as the \( P_t \) optimization. We also define the noiseless continuous optimization \( P_{t, nl} \) by

\[
(\hat{\theta}_{nl}, \hat{S}_{nl}) = \arg\min\|S\|_1 \text{ subject to } A(\theta)S = X
\]

and \( \theta \) is irreducible

(13)

C. Relation to the Integral Formalism

The idea of generalizing LASSO to a continuous regression case is also discussed in [35], where a straightforward extension of (4) is considered, in which the grid size increases so that the sums tend to integrals. As the main concern in [35] is to construct a theory of super-resolution, the analysis of the
resulting estimates is not considered. Note that any attempt to give results similar to that of the current work with the integral formalism in [35] requires a complicated parameterization of the support of the amplitudes. On the other hand, it is simple to show that this form is equivalent to the current CLASS definition. Thus, the analysis of CLASS simplifies that of the equivalent integral extension.

Formally, as the grid grows, the vector $S^\theta$ tends to a vector measure $s : B_\Theta \to \mathbb{C}^T$, where $B_\Theta$ is the restriction of the real Borel $\sigma$-algebra to $\Theta$. For a certain realization of $X$, assume that the $P_t$ solution is given by $\hat{\theta}(X, \lambda)$. Then,

$$P(X, \hat{\theta}(X, \lambda), \lambda) = \min \frac{1}{2} \left\| X - \int_{\Theta} a(\theta) ds^H \right\|_F^2 + \lambda \|s\|,$$

(14)

where the minimum is taken over all $T$-dimensional complex vector measures $s$ on $B_\Theta$ and $\|s\|$ denotes the total variation of $s$ over $\Theta$. See Appendix A for more details. This shows that although unlike the CLASS formalism, the integral formalism takes an asymptotically dense grid, the fact that the solution is sparse for any grid, enables us to reduce the solution to a low dimensional one which is exactly given by CLASS for an asymptotically dense grid.

D. Optimality Condition

Now, we give a characterization of the global minimum of $P_t$ generalizing Lemma 1 to the gridless conditions given in [36]. This plays a central role in the later analysis of CLASS. Based on Lemma 1, we draw two conclusions. The first in Theorem 1 is the CLASS optimality condition. The second in Proposition 1 is crucial in applying LASSO to a multiple snapshot case and shows that LASSO can be equivalently solved over a reduced dimensional set of pre-processed data.

Following the same logical pattern as in Section III-B by separating the two optimization steps in (10) and (12), it is seen that the first step in (10) is LASSO in (4) when $\theta^g = \theta$. Thus the result of Lemma 1 applies to this step and can be combined to the second one to result in the following theorem, which reminds that roughly speaking, CLASS is LASSO when the finite set $\theta^g$ is replaced by the continuum $\Theta$.

Theorem 1. (CLASS optimality condition) The matrix $\hat{S}$ and the DOA vector $\hat{\theta}$ in Lemma 1 form a

\footnote{In fact, this can be equivalently expressed by a simpler machinery of functions. However, as the minimum point always happens at an atomic function, represented by the Dirac impulse function, it is convenient to avoid the confusions by a measure based analysis.}
global minimum for (9) if and only if \( \hat{S} \) is irreducible and the following conditions are satisfied:

\[
A^H(\hat{\theta})\hat{N} = \lambda \Gamma_S^{-1} \hat{S}
\]

\[
\max_{\theta \in \Theta} \|a^H(\theta)\hat{N}\|_2 \leq \lambda, \quad (15b)
\]

where

\[
\Gamma_S = \begin{bmatrix}
    \gamma_{1,S} & 0 & \cdots & 0 \\
    0 & \gamma_{2,S} & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & \gamma_{n,S}
\end{bmatrix},
\]

(16)

**Proof:** First, assume that \( \hat{S} \) and \( \hat{\theta} \) form the irreducible global minimum of CLASS. Note that an optimal point of CLASS also satisfies the optimality of the two optimization steps in (10) and (12). Since adding a zero row may not change the unconstrained optimality, for an arbitrary choice of \( \theta \in \Theta \), the vector \( \hat{S}_1 = [\hat{S}^H \ 0]^H \) is a global minimum of (10) with \( \hat{\theta}_1 = [\hat{\theta} \ \theta] \), although it is reducible. Using Lemma 1 for the \( P_g \) optimization with \( \theta_G = \hat{\theta}_1 \) and after straightforward manipulations, we conclude that \( \|a^H(\theta)\hat{N}\| \leq \lambda \) as well as (15a), which are identical to (15) when considered for every \( \theta \).

Conversely, suppose that \( \hat{S} \) and \( \hat{\theta} \) satisfy (15a). For each different DOA set \( \theta_1 \), introducing \( \theta_2 = [\theta_1 \ \hat{\theta}] \) we observe that \( \hat{S}_2 = [\hat{S}^H \ 0]^H \) satisfies the conditions in Lemma 1. Thus, \( \hat{S}_2 \) is an optimal solution of (10) for \( \theta_2 \). Hence, \( \theta_2 \) is reducible to \( \theta \), which results in \( P(X, \theta_2, \lambda) = P(X, \theta, \lambda) \). On the other hand, because \( \theta_1 \subset \theta_2 \) we have \( P(X, \theta_2, \lambda) \leq P(X, \theta_1, \lambda) \). Thus, \( P(X, \hat{\theta}, \lambda) \leq P(X, \hat{\theta}, \lambda) \). As the choice of \( \theta_1 \) is arbitrary, we conclude that \( \hat{\theta} \) is the minimum of \( \mathcal{P}_t \) and CLASS.

**Proposition 1.** (Second order sufficiency) The \( \mathcal{P}_t \) solution and the solution to (10) are only functions of the observed sample correlation matrix \( R_x = \frac{1}{T} XX^H \).

**Proof:** We only need to show that the solution to \( \hat{S} \) in (7) is a function of \( R_x \). This can easily be shown by substituting the definition of \( \hat{N} \) in Theorem 1 into (7) to obtain

\[
\hat{S} = (A^H A + \lambda \Gamma_S)^{-1} A^H X,
\]

(17)

which solves \( \hat{S} \) with respect to \( \Gamma_S \). According to the fact that \( \Gamma_S \) consists of the diagonal elements of \( \hat{S} \hat{S}^H \), and substituting (17), it is obvious that \( \Gamma_S \) and, consequently, \( \hat{S} \) are only functions of \( R_x \).

As a consequence of Proposition 1, we may assume that \( X \) is of full column rank. Otherwise, there exists a full rank matrix \( X' \) with fewer columns and the same sample correlation matrix. However, to maintain the statistical properties of \( X \), we avoid such a restriction in this stage.
E. Uniqueness of the CLASS Solution

As we have already stated in Section III, LASSO and CLASS may have multiple global optima. We provide a uniqueness condition, which we claim to be met in many practical situations. We first define a proper basis manifold as follows.

Definition 1. We call a smooth basis manifold \( \mathbf{a}(\theta) \in \mathbb{C}^m \) proper if

1) It is non-ambiguous, i.e. for every collection \( \theta \) of \( m \) parameters, the matrix \( \mathbf{A}(\theta) \) has full column rank.
2) For any matrix \( \mathbf{Z} \in \mathbb{C}^{m \times T} \) with an arbitrary \( T \), the function \( f(\theta) = \| \mathbf{a}^H(\theta) \mathbf{Z} \|_2 \) may have at most \( m - 1 \) global maxima.

As an example, we show in Appendix B, part A that the ULA manifold is proper. However, we claim that many other array manifolds are proper without giving a proof. Then, the following theorem gives the uniqueness condition.

Theorem 2. (CLASS solution uniqueness) For a proper basis manifold, the solutions to \( \mathcal{P}_{t,\text{nl}} \) and \( \mathcal{P}_t \), for every \( \lambda > 0 \), are unique and, at most, of order \( n = m - 1 \).

Proof: The proof is given in Appendix B, part B.

As a remark, note that the global sparsity upper bound of \( n = m - 1 \) is not obvious in a complex-valued scenario. The reader may verify that the resulting sparse vector from LASSO by a random complex-valued regression matrix can be nonzero in possibly \( 2m - 1 \) elements. This tighter bound reflects the unique characteristics of a basis manifold.

F. Relation of CLASS and LASSO with a Fine Grid

In this part, we clarify the relation of CLASS and LASSO in a mathematical context. The most important relation is seen to be the fact that the CLASS optimality conditions in Theorem \( \square \) implies the condition in Lemma \( \square \). This means that: "For given \( \mathbf{X} \) and \( \lambda \), if the CLASS estimates are included in a grid \( \theta_G \), the corresponding solution of LASSO coincides with the former CLASS estimates". Otherwise, it is impossible for the LASSO to achieve them. However, the following result show that the latter ones lie very close to the former ones.

More precisely, we show that with a sufficiently dense grid, CLASS is indistinguishable from LASSO under any arbitrarily finite precision limit. To see this, we first show that if the grid in \( \mathcal{P}_g \) is sufficiently
close to every element of the $P_t$ estimate, then the $P_g$ solution is arbitrarily close to that of $P_t$. Thus, if the grid is dense enough, so that it is sufficiently close to any index combination including the CLASS estimate, then the solutions of $P_t$ and $P_g$ are always arbitrarily close.

We first need a distance measure between two unordered sets of estimates which may also be of different cardinalities. We define $\Delta(\theta_1, \theta_2) = \max_{\theta_1 \in \theta_1} \min_{\theta_2 \in \theta_2} |\theta_1 - \theta_2|$ as such a distance between two index sets $\theta_1$ and $\theta_2$. This denotes the uniformly largest distance from an element of $\theta_1$ to all elements of $\theta_2$. In other words, assigning to each element of $\theta_1$ its minimum distance to the elements of $\theta_2$, the $\Delta$ distance provides the largest of such assignments. Note that $\Delta(\theta_1, \theta_2) = 0$ only implies that $\theta_1 \subseteq \theta_2$, as it does not consider the worst uniform distance to $\theta_1$ for the elements of $\theta_2$. For each grid $\theta$, we also define $\zeta(\theta) = \max_{\theta \in \Theta} \min_{\theta' \in \theta} |\theta - \theta'|$ as the measure of fineness, since it computes the worst minimum distance of an arbitrary point to the grid. Then, it can be seen that assuming the $P_t$ minimum point uniquely given by $\hat{\theta}$, taking a grid $\theta_G$ and its corresponding $P_g$ solution at $\hat{\theta}^g$, for each $\epsilon > 0$ there exists a $\delta > 0$ so that if $\Delta(\hat{\theta}, \theta_G) \leq \delta$ then $\Delta(\hat{\theta}^g, \hat{\theta}) \leq \epsilon$. This shows that if the grid $\theta_G$ in $P_g$ satisfies $\zeta(\theta_G) < \delta$ with a proper $\delta$ then the active basis of the $P_g$ solution are $\epsilon$ close to the solution of $P_t$ for any possible $\hat{\theta}$, as $\Delta(\hat{\theta}, \theta_G) \leq \zeta(\theta_G) < \delta$. In other words, CLASS is equivalent to LASSO with a sufficiently fine grid under an arbitrary precision level. Thus, the analysis of CLASS applies to the grid-based LASSO and is comparable to the previous studies, e.g. [25].

IV. Performance Analysis

In the previous section, we built an optimization concept, which we showed to be equivalent to LASSO, but it is simpler to analyze by continuous linearization techniques. We are interested in providing the performance of LASSO (or more precisely CLASS) in the Mean Square Error sense when the data $X$ is actually provided by a sparse model in (2) with an asymptotically vanishing level of noise $N$. As the reader may recall from the classical performance analysis literature, this is performed by providing two steps of study. In the first step, the ideal cases in which the estimates are identical to the true parameters are verified. This is referred to as the consistency analysis. In the second step, a first order noise effect is analyzed by linearizing the optimality condition in the noiseless case, which is known as the error analysis. However, the application of this method to the current problem is difficult due to the following reasons:

First, CLASS (or LASSO) does not provide a single, but a group of estimates $\hat{\theta}(\lambda)$, indexed by the RP. We refer to this group as the solution path. It is not straightforward to decide which estimate is the best in the solution path. However, it is obvious that as the noise level tends to zero, the best estimate in the path
is recognized by an RP value \( \lambda \), which itself tends to zero. Accordingly, we consider the consistent cases and the error analysis in the vicinity of \( \lambda = 0 \) and \( N = 0 \). Second, the CLASS optimality conditions in (15a) are based on inequalities, due to the singularity of the \( \ell_1 \) norm, which makes linearization difficult. Third, the estimate of \( \hat{\theta} \) belongs to a space with a variable order, and therefore the standard linear algebra of vectors as well as conventional continuity and differentiability concepts do not apply directly.

We will address these difficulties in the following and provide solutions eventually leading to our final results. As the analysis concerns a nearly noiseless case, we refer to the observed data by \( X + \Delta X \), where \( X \) always denotes the noiseless data, obtained in (2) only from a set of true parameters \( \theta \) and components \( S \), and \( \Delta X = N \) is a noisy first order perturbation. We finally remind that in the following the uniqueness is always assumed.

A. Continuity of the solution path

The immediate consequence of Theorem 2 is that the unique solutions of \( P_t \) and \( P_{t, nl} \) are continuous with respect to all their variables (\( X \) and \( \lambda \)). Furthermore, the \( P_t \) continuum may also extend to \( \lambda = 0 \), in which case it approaches the \( P_{t, nl} \) solution. This provides the corresponding homotopy rule in the context of CLASS. However, due to variable dimensionality, the definitions should be clarified. Note that infinitesimally changing \( X \) and \( \lambda \) in CLASS not only changes the estimates but it may also result in creation or annihilation of new indexes with small corresponding component. Then, the solutions to \( P_t \) and \( P_{t, nl} \) are continuous, in the sense that for every \( \lambda > 0 \), \( X \), and \( \epsilon > 0 \), there exists a \( \delta > 0 \) so that if \( \max\{\|\Delta X\|, |\Delta \lambda|\} < \delta \), then for every elements \( \hat{\theta}_k \) in \( \hat{\theta} (X + \Delta X, \lambda + \Delta \lambda) = [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n] \) either there exists an element \( \hat{\theta}'_l \) in \( \hat{\theta}' (X, \lambda) = [\hat{\theta}'_1, \hat{\theta}'_2, \ldots, \hat{\theta}'_n] \) so that \( \max\{||\theta_k - \theta'_l||, ||\hat{S}_k - \hat{S}'_l||\} < \epsilon \) or \( ||\hat{S}'_l|| < \epsilon \), where \( \hat{S}_k \) and \( \hat{S}'_l \) denote the \( k \)th and \( l \)th row in the component estimates corresponding to \( \hat{\theta} \) and \( \hat{\theta}' \) respectively. Furthermore, as \( \lambda \) tends to zero, the solution to \( P_t \) tends to the solution of \( P_{t, nl} \) in the above sense. Although the proof follows a standard procedure, the reader may simply accept the fact by intuition. Thus, we omit the formal proof in the interest of saving space and clarity.

B. The near-zero behavior of the LASSO path

In this section, we compute the linearized perturbation of the CLASS solution for very small values of \( \lambda \) and a small deviation \( \Delta X \) from the ideal measurement \( X \). Note that the argument in Section [IV-A] does not prove that the cardinalities of \( \hat{\theta}_{nl} \) and \( \hat{\theta}(\lambda) \) are equal for sufficiently small values of \( \lambda \). However as \( \lambda \) decreases, some additional rows of the component estimate vanish so that the estimate almost reduces to the noiseless one. For simplicity, we confine the analysis to a case in which, fixing \( X \), the dimension
of the solution is equal to the dimension of the noiseless solution in a sufficiently small neighborhood of 
\( \lambda = 0 \). We may call such a case pure. We also denote 
\( d(\theta) = \frac{d\alpha(\theta)}{d\theta} \) and 
\( D(\theta) = [d(\theta_1) \ d(\theta_2) \ldots d(\theta_n)] \).

In simple words, we linearize the conditions in Theorem 1. Note that from Theorem 1, for any solution 
\((\hat{\theta}, \hat{S})\) of CLASS with \( \lambda > 0 \), the estimates \( \hat{\theta}_i \in \hat{\theta} \) are the local optima of the function 
\( f(\theta) = \|a^H(\theta)\hat{N}\|_2 \), where 
\( \hat{N} = X - A(\hat{\theta})\hat{S} \). Thus the derivative function of \( f(\theta) \) equals zero at these points. Thus,

\[
\text{Re} \left( d^H(\hat{\theta}_i)\hat{N} \right) = 0 \tag{18}
\]

Furthermore, as we have already shown by continuity, for sufficiently small values of \( \lambda \) and \( \Delta X \), the solution is close to the noiseless one. Let us say that the solution pair at such a point is given by 
\((\hat{\theta} + \Delta \hat{\theta}, \hat{S} + \Delta \hat{S})\), where \((\hat{\theta}, \hat{S}) = (\hat{\theta}_{nl}, \hat{S}_{nl}) \) simply denotes the solution of the noiseless CLASS \((\mathcal{P}_{nl})\).

We recall that we neglect the possibility of the existence of additional vanishing parameters. Expanding the condition \((15a)\) in Theorem 1 up to the first order, we get

\[
a^H(\hat{\theta}_i) \left( \Delta X - A(\hat{\theta})\Delta \hat{S} - D(\hat{\theta})\Delta Q U \right) = \lambda U_i, \tag{19}
\]

where we introduced 
\( U = \Gamma^{-1}_S \), with \( U_i \) as the \( i \)th row of \( U \), and \( \Delta Q = \Delta Q' \Gamma_\hat{S} \) with \( \Delta Q' \) as the diagonal matrix whose diagonal elements are identical to \( \Delta \hat{\theta} \), and where \( \Gamma_\hat{S} \) is given in \((5)\). Note that 
\( \|U_i\|_2 = 1 \) by definition. Moreover, expanding \((18)\), we get

\[
\text{Re} \left( d^H(\hat{\theta}_i) \left( \Delta X - A(\hat{\theta})\Delta \hat{S} - D(\hat{\theta})\Delta Q U \right) U_i^H \right) = 0. \tag{20}
\]

We can solve \((20)\) and \((19)\) to obtain the first order expansions. Let us define 
\( P_A^\perp = I - A(A^H A)^{-1}A^H \) as the orthogonal projection matrix onto the range space of \( A \), and

\[
\Xi = UU^H = [\xi_1 \ \xi_2 \ldots \ \xi_n]
\]

\[
R = \text{Re} \left[ D^H P_A^\perp D \otimes \Xi^T \right] \tag{21}
\]

Then, we have

\[
\Delta \hat{\theta} = \delta + \lambda \beta
\]

\[
\Delta \hat{S} = (A^H A)^{-1} \left[ A^H(\Delta X - D\Delta Q U) - \lambda U \right] \tag{22}
\]
where

$$\beta = \Gamma_S^{-1} R^{-1} \text{Re} \begin{bmatrix} d_1^H A (A^H A)^{-1} \xi_1 \\ \vdots \\ d_n^H A (A^H A)^{-1} \xi_n \end{bmatrix}$$

(23a)

and

$$\delta = \Gamma_S^{-1} R^{-1} \text{Re} \begin{bmatrix} d_1^H P_A \Delta X U_1^H \\ \vdots \\ d_n^H P_A \Delta X U_n^H \end{bmatrix}$$

(23b)

where we dropped the arguments of $A(\hat{\theta})$ and $D(\hat{\theta})$ for simplicity. We recognize (23b) as the first order expansion of the standard Non-Linear Least Square (NLLS) $\theta$ estimation error [40], whereas (23a) represents the contribution due to the $\ell_1$ regularization. The $\Delta S$ can similarly be separated with a different biasing contribution term from $\lambda$, which is unimportant for the current purpose.

C. Consistency Analysis

In this section, we discuss consistency, i.e. the correctness of the solution in the asymptotic noiseless case. It turns out that the LASSO based estimate is not generally consistent, and the ULA example case in Section IV-D shows that this is satisfied only when the parameters (i.e. elements of $\hat{\theta}$) are well separated. This means LASSO suffers from a resolution limit. We provide a condition in Proposition 3 under which the LASSO based estimator gives correct result (i.e. $\hat{\theta}_{nl} = \theta$) irrespective of the value of $S$. As a comparison, note that there is no such a consistent case for the classical spectral techniques such as beamforming [41], which cannot resolve a representation with too different level of components (rows of $S$). Although we only consider the noiseless data and its CLASS estimates, the homotopy rule in IV-A provides that a sufficiently small level of noise $N (= \Delta X)$ results in an arbitrarily close estimate to the true one for a noiseless consistent case.

We use the condition (15b) with the results in Section IV-B to obtain a tight upper bound. If the noiseless solution is identical to the true solution, (15b) holds with the true parameters $(\theta, S)$ perturbed by the corresponding terms given by (22). Since $\Delta X = 0$, we have $\delta = 0$. Then, $\Delta \hat{\theta} = \lambda \beta$ and $\Delta \hat{S} = -\lambda (A^H A)^{-1} [A^H D \Delta Q_0 + I] U$, where $\Delta Q_0$ is the diagonal matrix of the elements of $\Gamma_S \beta$. Substituting in (15b), dividing by $\lambda$ and letting $\lambda$ tend to zero, we arrive at the following result.

Proposition 2. For any CLASS-consistent true parameters $(\theta, S)$, we have

$$\forall \theta, \| a^H(\theta) \left( P_A D \Delta Q_0 - A (A^H A)^{-1} \right) \Xi_2 \|_2 \leq 1,$$

(24)
where $\Xi$ is given in (20) and $\Xi^{\frac{1}{2}}$ is a square root of $\Xi$.

For a certain $m$, at a DOA set $\theta$, if the condition (24) is true for every $\theta$ and a certain choice of $\Xi$, this DOA set is said to be consistent over $\Xi$. Note that $\Delta Q_0$ is also a function of $\Xi$. We can further simplify the analysis by focusing on the asymptotic case of very large number of sensors $m$. A more general terminology is given as follows.

**Definition 2.** For a fixed $m$, a DOA set $\theta$ is said to be **absolutely consistent** if (24) is true for every choice of $\Xi$. A DOA set is called **almost consistent** if (24) is true for almost all matrices $\Xi$. Furthermore, a sequence $\theta_m$ is called **asymptotically consistent** if for every $\Xi$ there exists a large number $M$ such that for $m > M$, the set $\theta_m$ is consistent over $\Xi$. Finally, a sequence is **almost asymptotically consistent** if the latter condition is true for almost all matrices $\Xi$.

**D. Example: Consistency for DOA estimation by a ULA**

In this part, we discuss a more detailed analysis for the example case of the ULA manifold introduced in Section II-B. For simplicity, we focus on a case, in which the order $n$ equals to 2 in the true representation $(\theta, S)$. Note that in the case of DOA estimation by a ULA, there exists a simple symmetry in linear arrays. It is simple to see that fixing the true $S$, a rigid shift in true DOA electrical angles results in a rigid shift in the CLASS solution. Thus, we may take two DOAs $\theta = \frac{\pi}{2}, \frac{\pi}{2} + \Delta \theta$ corresponding to $\phi = 0, \Delta \phi$ respectively without loss of generality. We also define $\Delta \phi = \frac{\delta}{m}$ and show that in terms of $\delta$ the consistency condition in (24) can be simplified for a large value of $m$. Note that fixing $\delta$, the separation $\Delta \phi$ varies by $m$.

We first introduce the functions $F_a(\delta) = \frac{e^{j\delta} - 1}{j\delta}$, $G_a(\delta) = \frac{dF_a}{d\delta}$, and $H_a(\delta) = -\frac{dG_a}{d\delta}$. It is simple to show that $\frac{1}{m}A^H A(\frac{\delta_1}{m})a(\frac{\delta_2}{m}) \rightarrow F_a(\delta_2 - \delta_1)$ as $m$ increases to infinity, so that we also have that $\frac{1}{m}A^H A = \begin{bmatrix} 1 & F_a(\delta) \\ F_a(-\delta) & 1 \end{bmatrix}$, $\frac{1}{m}A^H D = G_a(\delta) = \begin{bmatrix} \frac{1}{2}j & G_a(\delta) \\ G_a(-\delta) & \frac{1}{2}j \end{bmatrix} = [g_{a1} \ g_{a2}]$, $\frac{1}{m}D^H D = H_a(\delta) = \begin{bmatrix} 1 \frac{1}{3} \\ H_a(-\delta) \frac{1}{3} \end{bmatrix}$ for asymptotically large $m$. We also define the $1 \times 2$ vectors $g(\delta', \delta) = [G_a(\delta) \ G_a(\delta + \delta')]$ and $f(\delta', \delta) = [F_a(\delta) \ F_a(\delta + \delta')]$.

Then, substituting the above asymptotic results into Proposition 2 when $\Delta \phi = \frac{\delta}{m}$ for a large value of
\[ m, \text{ defining} \]
\[
R_a = \text{Re} \left[ (H_a - G_a^H F_a^{-1} G_a) \odot \Xi^T \right] \tag{25a}
\]
\[
\beta_a = R_a^{-1} \text{Re} \begin{bmatrix}
g_{a1}^H F_a^{-1} \xi_1 \\
g_{a2}^H F_a^{-1} \xi_2
\end{bmatrix} \tag{25b}
\]
and after some straightforward manipulations, (24) leads to the following result.

**Proposition 3.** Assuming a sequence \( \phi_m = [0 \Delta \phi_m] \), then if \( \lim_{m \to \infty} m \Delta \phi_m = \delta \), the sequence is almost asymptotically consistent if for every \( \delta' \in \mathbb{R} \)
\[
\|(g_a(\delta', \delta) - f_a(\delta', \delta) F_a^{-1} G_a) \Delta Q_a - f_a(\delta', \delta) F_a^{-1}\|_2 \leq 1, \tag{26}
\]
where \( \Delta Q_a \) is the diagonal matrix of elements of \( \beta_a \) in (25b). In particular,

- \( \phi_m \) is almost asymptotically consistent if \( m \Delta \phi_m \to \infty \).
- It is not almost asymptotically consistent if \( m \Delta \phi_m \to 0 \).

Proposition 3 shows that in the above asymptotic case, the inconsistency is reflected by a resolution limit of the form \( \frac{\delta}{m} \), where \( \delta \) is the smallest number satisfying (26). It is interesting to see that this is similar to the classical Rayleigh resolution for a ULA. In Section V we use (26) and (24) to find out a threshold rate in the asymptotic case and the resolution of small arrays respectively.

**E. Asymptotic Estimator Variance**

In this section, we complete the LASSO based estimation performance analysis by giving the asymptotic performance of the CLASS estimator in the low-noise case. In this case, the noiseless solution will not give the right number of parameters and one may either consider the \( \mathcal{P}_t \) optimization with \( \lambda > 0 \) or to remove some outliers in the estimate. We only consider the former case as follows, which sacrifices a small portion of the LASSO performance in terms of bias. Then, it is important to select a proper \( \lambda \) value.

1) Choice of \( \lambda \): We follow the strategy of choosing \( \lambda \) such that CLASS estimate has a correct cardinality, which is possible only for pure cases in high SNR. Since a greater value of \( \lambda \) introduces a higher bias, the best choice in terms of the minimum error is the smallest such \( \lambda \). Assume that the true parameter vector \( \theta \) is pure and consistent. The noiseless estimates are then equal to the true ones. Thus, we may use the true parameters in the expansion (23a). Then, the parameter \( \lambda \) can be computed as the smallest value for which the computed perturbations satisfy (15b). Recall that the consistency guarantees
that for sufficiently small $\Delta X = N$ such a $\lambda$ value exist. Using (15b) and the linearized terms in (22), we get

$$\lambda = \min \lambda'$$

such that

$$\forall \theta, \lambda' \geq \|a^H(\theta)\{P^\perp_A(N - D\Delta QU) + \lambda' A(A^HA)^{-1}U\}\|.$$  (27)

Let us denote the set of all non-negative values of $\lambda$ satisfying the inequality constraint in (27) for a certain $\theta$ by $S_\theta$. Then $\lambda = \min_\theta S_\theta$. Note that for each active parameter $\theta \in \Theta$, $S_\theta = [0, \infty)$, while for other values $S_\theta = [\Lambda(\theta), \infty)$, where $\Lambda(\theta)$ is the optimal value of a modified version of (27), in which the constraint is only for a fixed given $\theta$. Now, obviously, $\lambda = \max_\theta \Lambda(\theta)$ (see a similar argument in [33] for more illustration).

2) Statistical Assumptions and Results: We assume a white, Gaussian, centered, circularly symmetric noise vector process $n_i \sim N(0, C)$ and $N = [n_1 n_2 \ldots n_T]$. According to the results in Section IV-B we have $\Delta \hat{\theta} = \delta + \lambda \beta$ where $\delta$ and $\beta$ are given in (23b) and (23a) respectively. Recall that $\delta$ is actually the asymptotic Maximum Likelihood (ML) error in a high SNR scenario [40].

**Proposition 4.** For the choice of $\lambda$ in (27), the asymptotic statistical DOA bias is given by

$$\mathcal{E}(\Delta \hat{\theta}) = \mathcal{E}(\lambda)\beta,$$  (28)

where we used the fact that $\mathcal{E}(\delta) = 0$. Moreover, the asymptotic error covariance is

$$\text{Cov}(\Delta \hat{\theta}) = \frac{1}{2} \Gamma^{-1} R^{-1} \text{Re}[D_p^H C D_p \otimes \Xi] R^{-1} \Gamma^{-1} + \text{Var}(\lambda) \beta \beta^T,$$  (29)

where more details are presented in Appendix C.

3) Simplification for a Large ULA: For a ULA with a large enough number of sensors $m$, the relation (27) can be simplified. We assume that $\Lambda(\theta)$ for different values of $\theta$ is almost at the same level so that $\theta_1 = \arg \max \Lambda(\theta)$ is well-distributed and $\Pr(\min_{\theta \in \theta_0} |\theta_1 - \theta| = O(\frac{1}{m})) = 0$, where $\theta_0$ is the true DOA electrical angle set. This means that $a^H(\theta_1)A = o(m)$, $a^H(\theta_1)D = o(m^2)$ and, $a^H(\theta_1)P_A^\perp$ is identical.
to $a^H(\theta_1)$ almost surely. Following this line of reasoning, after some manipulations and keeping the dominant terms, we get

$$\lambda \approx \max_\theta \|a^H(\theta)N\|_2.$$  \hspace{1cm} (30)

Even with such a simplification, finding the two first statistical moments of $\lambda$ is a difficult task. However, for the case of uncorrelated noise, $C = \sigma^2 I$, and large enough $m$ in a ULA, we can proceed further by focusing on the extreme value, $\lambda_s^2$ of the sampled set $\{z_k = \|a^H(\phi_m = \frac{2k\pi}{m})N\|_2\}_{k=1}^m$. In this case,

$$\frac{\lambda_s^2}{\sigma^2 m} = \ln m + (T - 1) \ln \ln m + \gamma - \ln(T - 1)!, \hspace{1cm} (31)$$

where $\gamma$ converges weakly to a Gumble $[42]$ random variable. It should be also reminded that this result is valid when $T = o\left(\frac{\ln m}{\ln \ln m}\right)$. See Appendix [C] for more details. Now, it is expected that the true regularization $\lambda$ has the same expansion with $\gamma$ converging to a different variable. The result in (31) leads to

**Proposition 5.** For a ULA with large $m$ and $T = o\left(\frac{\ln m}{\ln \ln m}\right)$, the unknown expectations in Proposition 4 can be computed by

$$\mathcal{E}(\lambda^2) = \sigma^2 m (\ln m + (T - 1) \ln \ln m + \mathcal{E}(\gamma) - \ln(T - 1)!), \hspace{1cm} (32)$$

and

$$\mathcal{E}(\lambda) = \sigma \sqrt{m \ln m \left\{1 + \frac{(T - 1) \ln \ln m + \mathcal{E}(\gamma) - \ln(T - 1)!}{2 \ln m}\right\}}, \hspace{1cm} (33)$$

where $\gamma$ is a proper random variable and its expectation is a proper positive number.

The observations in Section [V] will show the accuracy of this approximation, using both a theoretical and an empirical value of $\mathcal{E}(\gamma)$.

**V. Numerical Results**

In this section, we present some results supporting and illustrating the analysis in Sections [IV-E] and [IV-C] for the case of estimating 2 DOAs by a ULA.

**A. Consistency**

In Section [IV-C] we analyzed the consistency of CLASS and identified the so called pure cases for which CLASS behaves consistently under the assumption of infinite estimation dynamic range. Then,
(24) gives a condition for a DOA to be consistent. We also showed that there asymptotically exists a resolution threshold $\Delta \theta = \frac{\delta}{m}$ that provides consistency. The exact value of $\delta$ can be found by (26). A simple numerical search shows that $\delta = 2.26\pi$ is a good approximation of the threshold. We remind that in practice, with a finite dynamic range and a discretized space, this upper bound can be reduced.

**B. Asymptotic Performance**

In Section IV-E, we computed the performance of CLASS in pure cases in many steps. Equations (29) and (28) connect the error covariance and the statistical bias to the two first moments of the optimal regularization parameter $\lambda$ given in (27). The $\lambda$ value is then approximated by (30) for the asymptotic case of large $m$. Finally, (30) was further simplified to (31), assuming uncorrelated noise and almost constant number of snapshots. Here, we compare these results to the experimental ones for the single-snapshot case. To find out the best estimate, the best $\lambda$ value in necessary. Note that while theoretically sufficient, (27) is not applicable, since it is actually an approximation to the exact best such value in the high SNR case. Accordingly, we use the CLASS implementation for a single snapshot in [36], which gives a continuous estimate of $\theta$ under an arbitrary precision with a tractable method.

As explained in Section IV-E, the first guess for $E(\gamma)$ in (33) and (32) is the expected value of a normalized Gumble random variable, the Euler-Mascheroni constant [43]. In Figure 1, the simple dashed line shows the theoretical value in (32) assuming the Gumble distribution. A better approximation can be found by fitting the experimental estimates, which gives $E(\gamma) = 1.3$. Note that this is a fundamental constant for every DOA estimate by a half-wavelength ULA. In Figure 1, then, the dashed line with triangles shows the same result by the empirical value.

Fixing $E(\gamma) = 1.3$, and combining (33) and (32) with (28) and (29) respectively, we evaluate the theoretical bias and covariance. Figures 2 and 3 show the performance of the estimator for different number of sensors. Note that the source separation $\Delta \phi$ (as well as $\Delta \theta$) changes by the number of sources as $\Delta \phi = \frac{4\pi}{m}$, which is twice the Rayleigh resolution limit. We obtain the experimental results by using fixed values of the sources $s_1 = s_2 = 1$ and 1000 different realizations of the noise vector for a single snapshot case. The noise variance is fixed to $\sigma^2 = 10^{-6}$. However, since the performance is proportional to the noise level, the normalized results are shown. The plots show good agreement between the theory and the experimental results.

We finally compared the LASSO performance to that of the ML with NLLS implementation [40] and Conventional BeamForming (CBF) [41]. Figures 4 and 5 compare the estimate Mean Squared Errors and variances of three different estimators; CLASS, ML with NLLS implementation and CBF, respectively.
Fig. 1. The experimental value of $E(\lambda^2)$ for different number of sensors compared to the theoretical results. The simple dashed line shows evaluation of (33) with $E(\gamma) = 0.58$ while the dashed one with triangle shows the same evaluation with $E(\gamma) = 1.3$.

Fig. 2. The DOA MSE versus different number of sensors. The estimation is based on one snapshot measurement of two sources separated by $\Delta \theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$, with the noise standard deviation $\sigma = 0.001$.

Fig. 3. The statistical bias normalized by true noise standard deviation for different number of sensors in the same setup as in Figure 2.
The setup is similar to the one in figures 2 and 3 while the number of sensors $m$ is fixed to 15. The results are the average of the outcomes of 100 trials at each noise level. We see that while the asymptotic variances of CLASS and ML methods coincide, the CLASS estimator has a higher asymptotic MSE. We conclude that CLASS modifies the solution of ML mostly by adding a bias term in the very high SNR regime. However, as SNR decreases, the MSE of CLASS reaches the one for the ML estimator in the SNR regime between -2 and 5 dBs. The two methods reach the threshold region at almost the same SNR.
VI. Conclusion

In this work, we studied the behavior of LASSO-based signal parameter estimation in terms of bias and variance. We remind that the aim of this work is to analyze a previously introduced method and not to devise a new technique. However, the analysis needs extra elaboration due to the improper definition of LASSO for statistical analysis. Accordingly, the concept of LASSO-based estimation was first clarified by introducing CLASS, a generalization of LASSO, which acts over a continuous set of indexes $\Theta$ rather than the finite matrix indexes. Then, we introduced conditions and definitions guaranteeing a good performance. Finally, we gave the theoretical performance under such consistency conditions. We finally tested our results by introducing an implementation of CLASS in a high SNR scenario. The results were confined to the single-snapshot scenario as the worst case.

From the theoretical and experimental results, one may conclude that the LASSO technique sacrifices performance only in terms of statistical bias with the same threshold SNR to attain a better implementation as we explained in Section V. Note that our implementation of ML by searching for the best initial point is of exponential complexity if the number of sources increase, while the computational cost of implementing LASSO grows by a polynomial speed. This is of great interest due to the flexibility of LASSO with the choice of the manifold, which may vary due to the specific problem. Note also that the complex-valued nature of the problem limits the implementation techniques, in which case the LASSO method should be considered as an effective general solver.

Last but not least, we address the RP selection issue. As shown in Section C, the performance of CLASS is improved by choosing a smaller value of $\lambda$. Accordingly, one may propose implementing the noiseless LASSO, which theoretically guarantees the performance of the exact ML rule. However, by introducing noise, some outliers will be introduced to the noiseless CLASS solution, which may be removed by a threshold. While unproven, we expect this strategy to have a higher SNR threshold, which puts the trade-off between the SNR threshold and bias in focus. We consider this an interesting extension of our work, and similarly a study of the case of large samples but finite SNR is a natural continuation.

Appendix A

Integral Form Equivalence

Assume $\hat{S} = [s_1 \ldots s_n]$ and $\hat{\theta} = [\theta_1 \ldots \theta_n]$ as the CLASS solution for $X$ and $\lambda$. Obviously, $s = \sum_{i=1}^{n} s_i \delta_{\theta_i}$ with $\delta$ as the Dirac measure achieves $P(X, \hat{\theta}(X, \lambda), \lambda)$ in (14). To show that there is no other measure with smaller value, first note that as the cost of the integral form for $s = 0$ equals to $1/2\|X\|_F$, the search can be restricted to the measures $s$ with $\|s(\Theta)\| \leq 1/2\|X\|_F$. Take an arbitrary such vector measure $s$
On the other hand, for every $k \in K$ one may take an element $\theta_k$ in $E_k^n \cap V_n$, for which we have $\|f^n_k - a(\theta_k)\| \leq \epsilon$. Combining this with (34) results in

$$\frac{1}{2} \left\| \int_{\Theta} a(\theta) ds^H - \sum_{k \in K} f^n_k s^H(E_k^n \cap V_n) \right\|_F^2 + C \epsilon \geq \frac{1}{2} \left\| \sum_{k \in K} a(\theta_k) s^H(E_k^n \cap V_n) - X \right\|_F^2,$$

(35)

where $C = 1/2 + \|X\|_F$ is a positive constant. We also have $\|s\|(\Theta) \geq \|s\|(V_n) = \sum_{k \in K} \|s\|(E_k^n \cap V_n) \geq \sum_{k \in K} \|s(E_k^n \cap V_n)\|$ which in combination with (35) gives

$$\frac{1}{2} \left\| \int_{\Theta} a(\theta) ds^H - X \right\|_F^2 + \lambda \|s\|(\Theta) + C \epsilon \geq \frac{1}{2} \left\| \sum_{k \in K} a(\theta_k) s^H(E_k^n \cap V_n) - X \right\|_F^2 + \lambda \sum_{k \in K} \|s(E_k^n \cap V_n)\|$$

(36)

The RHS in (36) is of the form in (10) and can be bounded below by $P(X, \hat{\theta}(X, \lambda), \lambda)$. Then, letting $\epsilon$ tend to zero, the proof is complete.
APPENDIX B

Uniqueness

A. The ULA Manifold is Proper

To see this, note that for any collection of \( m \) indexes \( \theta_1, \theta_2, \ldots, \theta_m \), their respective corresponding electrical angles \( \phi_1, \phi_2, \ldots, \phi_m \) and \( z_i = e^{j\phi_i} \) for \( i = 1, 2, \ldots, m \), the matrix

\[
A(\theta) = \begin{bmatrix}
1 & 1 & \ldots & 1 \\
z_1 & z_2 & \ldots & z_m \\
z_1^2 & z_2^2 & \ldots & z_m^2 \\
\vdots & \vdots & \ddots & \vdots \\
z_1^{m-1} & z_2^{m-1} & \ldots & z_m^{m-1}
\end{bmatrix}
\]  

is a Vandermonde type matrix, and thus full rank. To see the second condition, note that for a ULA, in terms of the electrical angle, we have \( a(\phi) = [1 e^{j\phi} e^{j2\phi} \ldots e^{j(m-1)\phi}]^T \). We may also extend this vector field to the whole complex plane as \( a(z) = [1 z z^2 \ldots z^{m-1}]^T \), where \( z \in \mathbb{C} \). Obviously, \( a(\phi) = a(z = e^{j\phi}) = a(z = e^{j\pi \cos \theta}) \). Take a matrix \( Z \) and define \( F(\phi) = \|a^H(\phi)Z\|_2^2 \), \( \lambda = \max F(\phi) \), \( T = ZZ^H \) and \( T_\alpha = \sum \{i,j|i-i=\alpha-m+1\} T_{ij} \) for \( 0 \leq \alpha \leq 2(m-1) \). Furthermore, define the polynomial

\[
S(z) = \sum_{\alpha=0}^{2(m-1)} T_\alpha z^\alpha - \lambda z^{m-1}
\]

It is straightforward to show that \( F(\phi) = \frac{S(z)}{z^{m-1}} + \lambda \) for \( z = e^{-j\phi} \). Note that since \( F(z) = \frac{S(z)}{z^{m-1}} + \lambda \) is a complex differentiable function, from the Cauchy-Riemann equations we get \( \frac{dF}{dz} = -\frac{1}{z^{m-2}} \frac{\partial F}{\partial \phi} \) if \( z = e^{-j\phi} \). Suppose now that \( \phi = [\phi_1 \phi_2 \ldots \phi_n] \) is the vector of all global maxima of \( F(\phi) \). Then, \( F(\phi_i) = \lambda \) and \( F(\phi) < \lambda \) for any \( \phi \notin \phi \). Thus, \( F(z_i) = \lambda \) and \( \frac{dF}{dz}(z_i) = -\frac{1}{z_i} \frac{\partial F}{\partial \phi}(\phi) = 0 \), where \( z_i = e^{-j\phi_i} \). It can simply be concluded that \( S(z_i) = S'(z_i) = 0 \), which means that all numbers \( z_i \) are multiple roots of \( S(z) \). However, \( S(z) \) is of degree \( 2(m-1) \) and the number of multiple zeros cannot be more than \( \frac{2(m-1)}{2} = m - 1 \).

B. CLASS Uniqueness for a Proper Manifold

Here, we give a proof of uniqueness of CLASS for a proper manifold. We use the following result to prove the uniqueness in the noiseless case.

Lemma 2. For any solution \((\hat{\Theta}, \hat{S})\) of \( \mathcal{P}_{t, nl} \) or \( \mathcal{P}_t \) with \( T \) snapshots and order \( n \), there exists a vector \( Z \in \mathbb{C}^{m \times T} \), such that

\[
\forall \theta \in \Theta \quad \|a(\theta)Z\| \leq 1
\]

and \( \|a(\hat{\theta}_i)Z\| = 1 \) for all \( \hat{\theta}_i \in \hat{\Theta} \).
Proof: For any $\mathcal{P}_t$ solution with $\lambda > 0$ choose $Z = \frac{X - A(\theta) \hat{S}}{\lambda}$. Then, from the optimality condition, the result is provided. For the noiseless case, take an arbitrary sequence of vector parameters $\theta_n$ each with order $n$ such that $\zeta(\theta_n) \to 0$ as $n \to \infty$ and define $\theta_0^n = [\hat{\theta}_n \theta_n]$. Take $S_0^n = [S^T \theta_0^n]^T$ as the corresponding components. Note that $S_0^n$ is a minimum point for optimization in (43) only over components with the parameters fixed to $\theta_0^n$. From the constraint, such an optimization is over the components of the form $S = S_0^n + N$, where $N = [\nu_1 \nu_2 \ldots \nu_T]$, in which $\nu_i$ are arbitrary vectors in $\mathcal{N}_{\mathcal{A}(\theta_0^n)}$. From [39, Theorem 3.4.3], we conclude that there exists a matrix $\Sigma \in \mathbb{C}^{(n+1) \times T}$ such that $||\Sigma_{n+1}|| \leq 1$ and $\Sigma_k = S_{n+1,k}/||S_{n+1,k}||_2$, (i.e. a subgradient of $||\cdot||_2$ at $S_0$) such that for any $N$ corresponding to an arbitrary collection $\nu_i \in \mathcal{N}_{\mathcal{A}(\theta_0^n)}$ for $i = 1, 2, \ldots, n + 1$, we have $\text{Tr}(\Sigma^T N) \leq 0$. But, this is only possible if $\sigma_i \in \mathcal{N}_{\mathcal{A}(\theta_0^n)} = \mathcal{R}_{\mathcal{A}(\theta_0^n)}$ for each $i = 1, 2, \ldots, T$. Thus, there exists a matrix $Z_n$ such that $\Sigma = A(\theta_0^n)^T Z_n$. As $Z_n$ is bounded, a subsequence of it has a limit $Z$ which provides the result as $\zeta(\theta_n) \to 0$.

assume there exists two irreducible global minima $(\theta_1, S_1)$ and $(\theta_2, S_2)$.

Define $\theta = \theta_1 \cup \theta_2$ and extend $S_1$ and $S_2$ on $\theta$ to $S^1$ and $S^2$ as

$$S^i_k = \begin{cases} S_{i,l} & \theta_k = \theta_{i,l} \\ 0 & \theta_k \notin \theta_i \end{cases}$$

(39)

Note that the pairs $(\theta, S^i)$ are reducible global minima.

For $\lambda > 0$, Since $\Psi(X, \theta, S)$ is a convex function of $S$, fixing $\lambda$, we have

$$\Psi(X, \theta, (1 - \mu)S^1 + \mu S^2) \leq (1 - \mu)\Psi(X, \theta, S^1) + \mu \Psi(X, \theta, S^2) = \Psi(X, \theta, S^1)$$

(40)

for every $0 \leq \mu \leq 1$. For such $\mu$, this means that the point $(\theta, (1 - \mu)S^1 + \mu S^2)$ is also a global minimum. Note also that the two terms $\frac{1}{2} ||\cdot||_F^2$ and $||\cdot||_{1,2}$ are convex, thus

$$||(1 - \mu)S^1 + \mu S^2||_{1,2} = (1 - \mu)||S^1||_{1,2} + \mu||S^2||_{1,2}$$

(41)

and

$$\frac{1}{2} \left\| X - A(\theta) \left((1 - \mu)S^1 + \mu S^2\right) \right\|_F^2 = (1 - \mu)\frac{1}{2} \left\| X - A(\theta)S^1 \right\|_F^2 + \mu\frac{1}{2} \left\| X - A(\theta)S^2 \right\|_F^2$$

(42)

since, otherwise, (42) and (41) with strict inequality results in strict inequality in (40). Moreover, $\frac{1}{2} \left\| \cdot \right\|_F^2$ is a strictly convex function and from (42) we conclude that $A(\theta)(S^1 - S^2) = 0$. For the noiseless case, we arrive at $A(\theta)(S^1 - S^2) = 0$ directly, due to the constraint.
In either noisy or noiseless cases, \( A(\theta)(S^1 - S^2) = 0 \) can be possible only if the dimension of \( \theta \) is more than \( m \), since for \( n = m \) the matrix \( A(\theta) \) is full-rank. Choose \( \mu \) such that \( \frac{\mu}{1-\mu} < \frac{\|S^1\|_2}{\|S^2\|_2} \) for every \( i \). Then the pair \((\theta, (1-\mu)S^1 + \mu S^2)\) is irreducible with \( n > m \). This is a contradiction, since due to Lemma 2 and the properness property, the dimension \( n \) of every irreducible global minimum \( \theta \) of CLASS, is less than the dimension of the received vector \( m \).

**APPENDIX C**

**A. Computing the error covariance**

We start by \( \Delta \theta = \lambda \beta + \delta \). Note that since \( \delta \) is the first order ML error, its covariance can be found, e.g. in [40]. We get \( \text{Cov}(\Delta \theta) = \frac{1}{2} \Gamma^{-1} R^{-1} \text{Re} \left[ \sum D_H C \right] + \text{Var}(\lambda) \beta \beta^T + k \) where \( k \) is proportional to \( E(\lambda \Delta X) \).

We can write the best \( \lambda \) value in (27) as

\[
\lambda = \min \left\{ a | a \geq \max_{\theta} ||a^H(\theta)[Y + a\Lambda]||_2 \right\}
\]  

(A.43)  

where \( Y \) is a linear (but not analytic) matrix function of \( \Delta X \) and \( \Lambda \) is a constant one. Thus, \( Y \) is centered Gaussian but not circularly symmetric. We may show \( \lambda = \lambda(Y) \) or \( \lambda(\Delta X) = \lambda(Y(\Delta X)) \).

The first observation is that \( \lambda(\alpha Y) = |\alpha| \lambda(Y) \) for every complex \( \alpha \), which implies that \( \lambda(\alpha \Delta X) = |\alpha| \lambda(\Delta X) \). In other words, for every \( a > 0 \), \( \{ \Delta X | \lambda(\Delta X) = a \} = \{ a \Delta X | \lambda(\Delta X) = 1 \} \) which means that \( \mathcal{E}(\Delta X | \lambda = a) = a \mathcal{E}(\Delta X | \lambda = 1) \). We conclude that \( \mathcal{E}(\lambda)\mathcal{E}(\Delta X | \lambda = 1) = \mathcal{E}(\lambda \Delta X | \lambda) = \mathcal{E}(\Delta X) = 0 \). Thus, \( \mathcal{E}(\Delta X | \lambda = 1) = 0 \) and \( \mathcal{E}(\Delta X \lambda) = \mathcal{E}(\lambda \Delta X | \lambda) = \mathcal{E}(\lambda^2) \mathcal{E}(\Delta X | \lambda = 1) = 0 \). This shows that \( k \propto \mathcal{E}(\lambda \Delta X) = 0 \).

**B. The asymptotic extreme value expansion**

The variables \( z_i \) introduced in Section [IV-E] are independent with chi squared distribution. Thus,

\[
\Pr(z_i > \alpha) = \frac{1}{(T-1)!} \int_{\frac{\alpha}{\sigma}}^{\infty} z^{T-1} e^{-z} \, dz
\]  

which is in the order of \( \frac{\alpha^{T-1} e^{-\alpha}}{(T-1)! (\sigma^2 m)^{\frac{T-1}{2}}} \). This can be seen, for example, by L’Hopital’s rule. Since \( z_i \)s are independent, it is simple to see that

\[
\Pr(\lambda^2 < \alpha) = (1 - \Pr(z_i > \alpha))^m = \beta
\]  

(A.45)
Then, for large enough \( m \), we get \( \Pr(z_i > \alpha) = \frac{-\ln \beta}{m} \) which using (44), can be written as \( (T - 1) \ln \frac{\alpha}{\sigma^2 m} - \ln(T - 1)! = \ln(-\ln(\beta)) - m \). It can then be simplified to

\[
\frac{\alpha}{\sigma^2 m} = \ln m + (T - 1) \ln \ln m + \gamma - \ln(T - 1)! + o(1)
\]  

(46)

where \( \gamma = -\ln(-\ln(\beta)) \). We conclude (31).

REFERENCES

[1] T. T.Wu et al., “Genome-wide association analysis by lasso penalized logistic regression,” Bioinformatics, vol. 25, pp. 714–721, Mar. 2009.
[2] F.Parvaresh et al., “Recovering sparse signals using sparse measurement matrices in compressed dna microarrays,” IEEE, J. Select. Topics Signal Processing, vol. 2, pp. 275–285, June 2008.
[3] W.Tu and S.Sun, “Spatial filter selection with lasso for eeg classification,” in Advanced Data Mining and Applications, Chongqing, China, 2010, pp. 142–149.
[4] H.Konno and H.Yamazaki, “Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market,” Manage. Sci., vol. 37, pp. 519–531, May 1991.
[5] H.Yao et al., “Compressive sensing of the tohoku-oki mw 9.0 earthquake: Frequency-dependent rupture modes,” Geophys. Res. Lett., vol. 38, Oct. 2011.
[6] R.Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Stat. Soc., Series B (Methodological), vol. 58, pp. 267–288, Jan. 1996.
[7] S. S.Chen et al., “Atomic Decomposition by Basis Pursuit,” SIAM J. Sci. Comput., vol. 20, no. 1, pp. 33–61, Dec. 1998.
[8] J. J.Fuchs, “On the application of the global matched filter to doa estimation with uniform circular arrays,” IEEE Trans. Signal Processing, vol. 49, no. 4, pp. 702–709, Apr. 2001.
[9] M.Figueiredo et al., “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” IEEE Select. Topic. Signal Processing, vol. 1, pp. 586–597, Dec. 2007.
[10] H.Zou, “The adaptive lasso and its oracle properties,” J. Amer. Stat. Assoc., vol. 101, no. 476, pp. 418–429, 2006.
[11] T.Park and G.Casella, “The bayesian lasso,” J. Amer. Stat. Assoc., vol. 103, pp. 681–686, 2008.
[12] H.Zayyani et al., “Bayesian pursuit algorithm for sparse representation,” in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Taipei, Taiwan, Apr. 2009, pp. 1549–1552.
[13] D.Donoho, “Compressed sensing,” IEEE Trans. Inform. Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[14] D.Malioutov et al., “Source localization by enforcing sparsity through a laplacian prior: an svd-based approach,” IEEE Workshop Stat. Signal Processing, pp. 573–576, Sept. 2003.
[15] S. G.Mallat and Z.Zhang, “Matching pursuits with time-frequency dictionaries,” IEEE Trans. Signal Processing, vol. 41, pp. 3397–3415, Dec. 1993.
[16] J. A.Tropp and A. C.Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” IEEE Trans. Inform. Theory, vol. 53, pp. 4655–4666, Dec. 2007.
[17] B.Efron et al., “Least angle regression,” Ann. Stat., vol. 32, pp. 407–499, Apr. 2004.
[18] D.Needell and J. A.Tropp, “Cosamp: Iterative signal recovery from incomplete and inaccurate samples,” Elsevier, Appl. Comput. Harmon. Anal., vol. 26, pp. 301–321, May 2009.
[19] M. Mishali and Y. C. Eldar, “From theory to practice: Sub-nyquist sampling of sparse wideband analog signals,” IEEE J. Select. Topics Signal Processing, vol. 4, pp. 375–391, Apr. 2010.

[20] J. A. Tropp et al., “Beyond nyquist: Efficient sampling of sparse bandlimited signals,” IEEE Trans. Inform. Theory, vol. 56, pp. 520–544, Jan. 2010.

[21] R. G. Baraniuk, “Compressive sensing [lecture notes],” IEEE Signal Processing Mag., vol. 24, pp. 118–121, July 2007.

[22] M. Arigovindan et al., “Variational image reconstruction from arbitrarily spaced samples: A fast multiresolution spline solution,” IEEE Trans. Image Processing, vol. 14, pp. 450–460, Apr. 2005.

[23] M. Lustig et al., “Sparse mri: The application of compressed sensing for rapid mr imaging,” Resonance Med. Mag., vol. 58, pp. 1182–1195, Dec. 2007.

[24] P. Milanfar et al., “A moment-based variational approach to tomographic reconstruction,” IEEE Trans. Image Processing, vol. 5, pp. 459–470, Mar. 1996.

[25] D. Malioutov, “A sparse signal reconstruction perspective for source localization with sensor arrays,” Master’s thesis, MIT, July 2003.

[26] E. J. Candès and Y. Plan, “Near-ideal model selection by $\ell_1$ minimization,” Ann. Stat., vol. 37, pp. 2145–2177, Oct. 2009.

[27] E. J. Candès and Y. Plan, “A Probabilistic and RIPless Theory of Compressed Sensing,” IEEE Trans. Inform. Theory, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[28] P. Stoica and A. Nehorai, “Performance study of conditional and unconditional direction-of-arrival estimation,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, no. 10, pp. 1783–1795, 1990.

[29] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” J. Amer. Stat. Assoc., vol. 96, no. 456, pp. 1348–1360, 2001.

[30] Z. Tan et al., “Joint sparse recovery method for compressed sensing with structured dictionary mismatch,” arXiv preprint arXiv:1309.0858, 2013.

[31] C. Ekanadham et al., “Recovery of sparse translation-invariant signals with continuous basis pursuit,” IEEE Trans. Signal Processing, 2011.

[32] Y. Chi et al., “Sensitivity to basis mismatch in compressed sensing,” IEEE Trans. Signal Processing, 2011.

[33] A. Panahi and M. Viberg, “Fast candidate points selection in the lasso path,” IEEE Signal Processing Lett., vol. 19, no. 2, pp. 79–82, Feb. 2012.

[34] E. J. Candès and T. Tao, “Decoding by linear programming,” IEEE Trans. Inform. Theory, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[35] E. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” arXiv preprint arXiv:1203.5871, 2012.

[36] A. Panahi and M. Viberg, “A robust $\ell_1$ penalized doa estimator,” in Asilomar Conference on Signals, Systems, and Computation, Nov. 2012.

[37] D. Donoho and Y. Tsaig, “Fast solution of 1-norm minimization problems when the solution may be sparse,” IEEE Trans. Inform. Theory, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[38] M. R. Osborne et al., “A new approach to variable selection in least squares problems,” 1999.

[39] M. S. Bazaraa et al., Nonlinear Programming: Theory and Algorithms, 3rd ed. Wiley, 2006.

[40] B. Ottersten et al., Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing. Springer, 1993.
[41] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric approach,” *IEEE Signal Processing Mag.*, vol. 13, pp. 67–94, July 1996.

[42] J. Beirlant *et al.*, *Statistics of Extremes: Theory and Applications*, 1st ed. Wiley, Oct. 2004.

[43] D. E. Knuth, “Euler constant to 1271 places,” *Math. Comput.*, vol. 16, no. 79, pp. 275–281, 1962.