On Robustness of $\ell_1$-Regularization Methods for Spectral Estimation

Johan Karlsson, Member, IEEE, and Lipeng Ning, Member, IEEE,

Abstract—The use of $\ell_1$-regularization in sparse estimation methods has received huge attention during the last decade, and applications in virtually all fields of applied mathematics have benefited greatly. This interest was sparked by the recovery results of Candès, Donoho, Tao, Tropp, et al. and has resulted in a framework for solving a set of combinatorial problems in polynomial time by using convex relaxation techniques.

In this work we study the use of $\ell_1$-regularization methods for high-resolution spectral estimation. In this problem, the dictionary is typically coherent and existing theory for robust/exact recovery does not apply. In fact, the robustness cannot be guaranteed in the usual strong sense. Instead, we consider metrics inspired by the Monge-Kantorovich transportation problem and show that the magnitude can be robustly recovered if the original signal is sufficiently sparse and separated. We derive both worst case error bounds as well as error bounds based on assumptions on the noise distribution.

Index Terms—Spectral estimation, sparse recovery, robustness, error bounds, coherent dictionaries.

I. INTRODUCTION

Consider a discrete-time signal $y_n \in \mathbb{C}, n \in \mathbb{Z}$, consisting of sinusoids in noise,

$$y_n = \sum_{\ell=1}^{L} x_{\ell} e^{i n \lambda_{\ell}} + w_n, \quad (1)$$

where $w_n$ is a stochastic or deterministic but unknown error, and $x_{\ell} \in \mathbb{C}$ determine the magnitude and phase of the complex sinusoids. The spectral estimation problem, which is a key element for modelling and estimation of signals, is to determine the energy distribution over frequency of the time signal based on a finite sample $y_1, \ldots, y_N$. There is a large number of methods such as the periodogram, MUSIC, ESPRIT, matching pursuit, maximum likelihood, and maximum entropy, for this spectral estimation problem [1], [2]. In this work, we will consider a recently popular approach for addressing this spectral estimation problems, namely, the use of sparse approaches based on $\ell_1$ regularization [3], [4], [5].

For these sparse methods, typically the frequency grid is discretized resulting in the linear system

$$y = Ax + w \quad (2)$$

where $x \in \mathbb{C}^K$, $w \in \mathbb{C}^N$, $N < K$, and the columns of $A \in \mathbb{C}^{N \times K}$ form a normalized overcomplete Fourier basis

$$a(\theta_k) = \frac{1}{\sqrt{N}} (1, e^{i \theta_k}, \ldots, e^{i(N-1)\theta_k})^T, \quad (3)$$

for a set of frequencies $\Omega = \{\theta_1, \theta_2, \ldots, \theta_K\}$. The system (2) is underdetermined and hence has infinitely many solutions; and to single out a unique solution, sparse methods use $\ell_1$ regularization as a convex surrogate for the cardinality (see, e.g., the review articles [6], [7]; and [8], [9], [10], [11] for applications to signals and systems). This sparse recovery problem has been intensely studied during the last decade for various classes of underdetermined systems and significant advances have been made in understanding when these approaches recover a sparse signal. In fact, it can be proved that if $A$ is sufficiently incoherent, then these recovery methods can be used to recover a sparse solution robustly [6], [7], [12], [13]. However, these recovery results can not be used for typical spectral estimation problems since $A$ is highly coherent.

Nevertheless, empirical results suggest that these sparse methods have significantly higher performance compared to, e.g., the periodogram. In this paper we will study sparse recovery methods based on $\ell_1$-regularization in order to determine what can be guaranteed in terms of error bounds and robustness. As we will see, no useful bounds can be obtained in terms of the $\ell_2$ norm (see also [14]). Instead we consider distances based on optimal mass transport and show that the magnitude of the true solution can be recovered robustly under certain conditions. In Section II we discuss sparse recovery methods and state known results on robustness. In Section III we consider a motivating example and show error bounds that apply here. In Section IV we define the Monge-Kantorovich transportation distance. In Section V we derive worst case bounds for a sparse recovery method expressed in terms of the transportation distance. Finally in Section VI we show how to use the results from V so that they apply also for cases when the error is large, given that the number of samples is sufficiently large.

We use the following notation. Let the columns of $A$ be $a(\theta)$, for $\theta \in \Omega$ and where the columns are indexed by the set $\theta \in \Omega$. For $\Lambda \subset \Omega$, let $A_\Lambda$ denote the matrix with columns $\{a(\theta)\}_{\theta \in \Lambda}$. Let $e(\lambda)$ be the $K \times 1$ unit vector with a 1 at the position $\lambda \in \Omega$. For a vector $x \in \mathbb{C}^K$, let $|x|_e \in \mathbb{R}^K$ be the vector where the absolute value is applied element-wise, i.e., $(|x|_e)_\theta = |x(\theta)|$. Also, let $x|_{\Lambda}$ be the restriction of $x$ to the indices $\theta \in \Lambda$. 
II. METHODS BASED ON $\ell_1$ REGULARIZATION

A common objective in estimation is to find the most sparse solution consistent with a linear system (2), e.g., the optimum of

$$\arg\min_{x\in\mathbb{C}^K} \|x\|_0 \quad \text{subject to} \quad \|y - Ax\|_2 \leq \epsilon.$$  (4)

This is a combinatorial problem, but using the $\ell_1$-norm as a surrogate for the cardinality:

$$\arg\min_{x\in\mathbb{C}^K} \|x\|_1 \quad \text{subject to} \quad \|y - Ax\|_2 \leq \epsilon,$$  (5)

leads to a problem that is convex (and hence computationally feasible) and in many cases achieves its minimum close to the minimum of (4). This recovery method is often called LASSO [15] or basis pursuit denoising (BPDN) [7]. In fact, this may be used for robust recovery of a sparse vector.

Theorem 1 ([6]): Assume that for $\delta_{2s} < \sqrt{2} - 1$, the inequality

$$(1 - \delta_{2s})\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_{2s})\|x\|_2^2$$  (6)

holds for all $2s$-sparse $x$. Let $y = Ax_{\text{true}} + w$ where $x_{\text{true}}$ is $s$-sparse and $\|w\|_2 \leq \epsilon$, then the minimizer $x_{\text{est}}$ of (5) satisfy

$$\|x_{\text{true}} - x_{\text{est}}\|_2 \leq C(\delta_{2s})\epsilon.$$  

The condition (6) on $A$ is known as the restricted isometry property, and also restricts the coherency of $A$:

$$\mu = \max_{\theta_i \neq \theta_j} |a(\theta_i)^* a(\theta_j)|.$$  

In fact, $\mu \leq \delta_2$, hence the result is only applicable if $\mu < \sqrt{2} - 1$. For spectral estimation, the dictionary typically consists of the Fourier vectors (3) corresponding to the set of frequencies $\Omega = \{\theta_1, \theta_2, \ldots, \theta_K\}$ where $\theta_k = 2\pi k/K$. In these applications $K$ is typically between $5N$ and $10N$ [5], which corresponds to $\mu \approx 0.94$ and $\mu \approx 0.98$, respectively. Since $A$ is highly coherent, there exists 2-sparse vectors $x$ so that $\|Ax\|_2^2 \leq \|x\|_2^2$, and robust recovery in the sense of Theorem 1 is not possible.

III. MOTIVATING EXAMPLE

In this section we will consider the case when the signal consists of a single frequency. We will first show an example where a small noise contribution results in a large $\ell_2$ error (see Figure 1). Then we will derive an error bound and relate this bound to optimal mass transport.

Let the signal consist of a single sinusoid in noise

$$y = Ax_{\text{true}} + w = a(0.1\pi) + w,$$

of length $N = 100$. Here, the true $x_{\text{true}}$ has a unique non-zero element at the point $0.1\pi$, i.e., $x_{\text{true}} = e(0.1\pi)$, and $A \in \mathbb{C}^{N \times K}$ consists of an overcomplete Fourier basis (3) with $K = 1000$. Figure 1 shows the recovered solution $x_{\text{est}}$ using (5) for an example where the noise level is $10\% = \|w\|_2 / \|a(0.1\pi)\|_2$ and $\epsilon = \|w\|$. This results in a relative $\ell_2$ error of $130\%$. This shows that robust recovery is not possible in $\ell_2$-norm sense even if $x_{\text{true}}$ only has one non-zero element. However, there is only a small error in the estimated frequency and this precision is typically sufficient for most applications. Thus a more reasonable quantitative guarantee is needed for using (5) in spectral estimation.

![Figure 1](image_url)

Fig. 1. Estimation example where the true signal is a sinusoid, and where a small noise contribution results in a relative $\ell_2$ error of $130\%$. Here SNR:= $\|Ax\|_2 / \|w\|_2 = 10$, $N = 100$, and $K = 1000$.

To show the approach taken in this paper, we consider the basic example $x_{\text{true}} = e(\lambda)$, where $x_{\text{true}}$ has support in only one point $\lambda$. Here we will study $x_{\text{est}}$ obtained via (2) and (5) and show that the energy of $x_{\text{est}}$ must be located close to $\lambda$ if the noise term $w$ is small. The signal is given by

$$y = Ax_{\text{true}} + w = a(\lambda) + w, \quad \text{where} \quad \|w\|_2 \leq \epsilon,$$

and let $x_{\text{est}}$ be the recovered solution from (5):

$$x_{\text{est}} = \arg\min \|x\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 \leq \epsilon.$$  

Note that

$$\epsilon \geq \|Ax_{\text{est}} - y\|_2 \geq |a(\lambda)^* (Ax_{\text{est}} - y)| = \left| \sum_{\theta \in \Omega} x_{\text{est}}(\theta) a(\lambda)^* a(\theta) - a(\lambda)^* a(\lambda) - a(\lambda)^* w \right| \geq 1 - \sum_{\theta \in \Omega} |x_{\text{est}}(\theta)||a(\lambda)^* a(\theta)| - \epsilon,$$

by using first the definition of the $2$-norm and then the triangle inequality; hence we deduce

$$2\epsilon \geq \sum_{\theta \in \Omega} |x_{\text{est}}(\theta)|(1 - |a(\lambda)^* a(\theta)|) + \|x_{\text{true}}\|_1 - \|x_{\text{est}}\|_1.$$  (7)

This gives a bound on two terms; where the first term quantifies how much the mass deviates from the location $\theta = \lambda$, and the second term penalizes deviation in the total mass ($\ell_1$ norm). Note that since $x_{\text{true}}$ is feasible for (5), the second term, $\|x_{\text{true}}\|_1 - \|x_{\text{est}}\|_1$, representing the mass deviation is non-negative. Therefore, if $\epsilon$ is small, then equation (7) ensures both a lower bound on $\|x_{\text{est}}\|_1$ as well as a bound limiting the energy spread of $x_{\text{est}}$.

The first term of (7), quantifying the spread of $x_{\text{est}}$ from the spectral line, can be seen as a transportation cost where
the cost of transporting a “unit” spectral line from \( \theta \) to \( \lambda \) is
\[
1 - |a(\lambda)^* a(\theta)| = 1 - \frac{\sin(N(\lambda - \theta)/2)}{\sin((\lambda - \theta)/2)}.
\]
depicted in Figure 2. By multiplying the weight of the spectral line with this cost function, we obtain the “mass” transport cost.

This simple example is to illustrate that a more meaningful robustness result of spectral line estimation problem is to obtain bounds for the deviation of spectral lines and the “mass” deviation. As we will see, metrics inspired by the Monge-Kantorovich transportation problem are naturally used to derive the bounds.

This is known as the Monge-Kantorovich distance [17]. Monge-Kantorovich distances are not metrics, in general, but they readily give rise to a class of the so-called Wasserstein metrics
\[
W_p(\rho_0, \rho_1) = T(\rho_0, \rho_1)^{\min(1, \frac{1}{p})}
\]
where the cost function is of the form \( c(\theta_0, \theta_1) = d(\theta_0, \theta_1)^p \) where \( d \) is a metric and \( p \in (0, \infty) \) [16].

The Monge-Kantorovich theory deals with mass distributions of equal mass, however, in [18] we develop distances based on similar principles, that applies to distributions of possibly unequal mass. Given the two mass distributions \( \rho_0 \) and \( \rho_1 \), we postulate that these are perturbations of two other mass distributions \( \sigma_0, \sigma_1 \in \mathbb{R}^K \), that have equal mass. Then, the cost of transporting \( \rho_0 \) and \( \rho_1 \) to one another can be thought of as the cost of transporting \( \sigma_0 \) and \( \sigma_1 \) to one another plus the size of the respective perturbations:
\[
\hat{T}(\rho_0, \rho_1) := \inf_{\|\sigma_0\|_1 = \|\sigma_1\|_1} T(\sigma_0, \sigma_1) + \sum_{j=0}^{1} \|\rho_j - \sigma_j\|_1. \tag{9}
\]

IV. TRANSPORTATION DISTANCE

The Monge-Kantorovich distance quantifies the optimal transportation cost of transferring one “mass” distribution to another with specified cost of moving one unit amount of mass from one location to another [16].

Consider two \( K \)-dimensional elementwise non-negative vectors \( \rho_0 \) and \( \rho_1 \) that each represent a distribution of “mass” at the \( K \) different locations \( \Omega \). Let \( m(\theta_0, \theta_1) \) denote the amount of mass transported from location \( \theta_0 \) to location \( \theta_1 \), and we say that \( M = (m(\theta_0, \theta_1))_{\theta_0, \theta_1 \in \Omega} = \mathbb{R}^{K \times K} \) is a feasible transportation plan from \( \rho_0 \) to \( \rho_1 \) if the respective marginals are equal to \( \rho_0 \) and \( \rho_1 \), respectively, i.e., if \( M \) is in
\[
\Pi(\rho_0, \rho_1) := \left\{ M = (m(\theta_0, \theta_1))_{\theta_0, \theta_1} : m(\theta_0, \theta_1) \geq 0, \sum_{\theta_1 \in \Omega} m(\theta_0, \theta_1) = \rho_0(\theta_0), \ \theta_0 \in \Omega \right\}.
\]

Let \( C := (c(\theta_0, \theta_1))_{\theta_0, \theta_1 \in \Omega} \in \mathbb{R}^{K \times K} \) be the matrix whose elements \( c(\theta_0, \theta_1) \) represents the cost of transferring one unit amount of mass from the location \( \theta_0 \) to location \( \theta_1 \). Then the minimum cost of transporting mass with distribution \( \rho_0 \) to a distribution \( \rho_1 \) is
\[
T(\rho_0, \rho_1) = \inf_{M \in \Pi(\rho_0, \rho_1)} \sum_{\theta_0, \theta_1 \in \Omega} m(\theta_0, \theta_1)c(\theta_0, \theta_1). \tag{8}
\]

For the problem of comparing sparse recovery vector \( x_{\text{est}} \) and \( x_{\text{true}} \), we consider the absolute values \( \rho_0 = |x_{\text{est}}|_c \) and \( \rho_1 = |x_{\text{true}}|_c \) as two discrete mass distributions on the interval \([0, 2\pi]\). The cost of transferring one unit of mass from \( \theta_0 \) to \( \theta_1 \) is given by the function \( c(\theta_0, \theta_1) = 1 - |a(\theta_0)^* a(\theta_1)| \) as shown in Figure 2. From the definition of the transportation distance (8), it is clear that the first term in (7) represents the cost of transferring all the mass in \( |x_{\text{est}}(\theta)| \) to location \( \lambda \) while the second term in (7) represents a mass perturbation which penalize the difference in \( \ell_1 \) norm between \( x_{\text{est}} \) and \( x_{\text{true}} \).

V. WORST CASE ERROR BOUNDS FOR BASIS PURSUIT DENOISING

In this section we will generalize the error bound in the motivating example (in Section III) to hold for the case
\[
x_{\text{true}} = \sum_{\lambda \in \Lambda} \alpha_{\lambda} e(\lambda), \tag{10}
\]
when \( x_{\text{true}} \) contains more than 1 non-zero element. We will derive the bound based on the following properties
\[
11a) \quad x_{\text{true}} \text{ has support } \Lambda,
11b) \quad \|x_{\text{est}}\|_1 \leq \|x_{\text{true}}\|_1, \tag{11}
11c) \quad \|Ax_{\text{est}} - Ax_{\text{true}}\|_2 \leq 2\epsilon,
\]
which clearly holds for \( x_{\text{est}} \) obtained from (5). If \( \Lambda \) contains two frequencies that are closely located, the signals may cancel with each other, i.e., there are \( x_{\text{true}} \) such that \( \|Ax_{\text{true}}\|_2 \ll \|x_{\text{true}}\|_2 \), and there is no hope to reconstruct \( x_{\text{true}} \). Therefore we can only hope to achieve such a bound if the support \( \Lambda \) is sufficiently separated (c.f., page 17 in [19]). To quantify this, define the cumulative intercoherence with respect to \( \Lambda \).
Definition 2: Let $A$ be a dictionary with index set $\Omega$ and let $\Lambda \subset \Omega$. Then we define
\[
\mu_\Lambda := \max_{\theta \in \Omega} \left( \sum_{\lambda \in \Lambda \setminus \phi} |a(\lambda)^*a(\theta)| - \max_{\lambda \in \Lambda} |a(\lambda)^*a(\theta)| \right)
= \max_{\theta \in \Omega} \min_{\phi \in \Lambda \setminus \phi} \sum_{\lambda \in \Lambda \setminus \phi} |a(\lambda)^*a(\theta)|,
\]
which we denote by the cumulative intercoherence.

The cumulative coherence $\mu_\Lambda$ is a quantitative measure on how closely spaced a set of spectral lines are. If all frequencies of $\Lambda$ are sufficiently separated, then for any $\theta$ we will have $|a(\lambda)^*a(\theta)| \ll 1$ for all $\lambda \in \Lambda$ except one, hence $\mu_\Lambda \ll 1$. Figure 3 depicts $\mu_\Lambda$ where $\Lambda = \{0, \theta, 2\theta, \ldots, (L-1)\theta\}$ is equispaced and $N = 100$, for the three cases $L \in \{2, 3, 10\}$.

![Diagram](image.png)

Fig 3. Intercorherence for $L \in \{2, 3, 10\}$ and $\Lambda$ consisting of $L$ equispaced frequencies with distance $\theta$. Here $N = 100$.

This definition allows us to state the following lemma, which is proved in the Appendix.

Lemma 3: Let $L = |\Lambda|$. Then following inequalities hold:
1) $\|A_\Lambda^*A_\Lambda\|_{1,1} \leq \|A_\Lambda^*A\|_{1,1} \leq 1 + \mu_\Lambda$,
2) $\left| (A_\Lambda^*A_\Lambda)^{-1} \right|_{1,1} \leq \frac{1}{1 - \mu_\Lambda}$,
3) $\|A_\Lambda^*\|_{2,1} \leq \sqrt{L(1 + \mu_\Lambda)}$.

A. Bounding the transportation cost

In this subsection we will bound the transportation distance $\tilde{T}(x_{est}, x_{true})$ for any two vectors $x_{true}$ and $x_{est}$ satisfying (11). The proof idea is to show that there exists a vector $\tilde{x} \in \mathbb{C}^K$ such that
1) $\|\tilde{x}\|_1 = \|x_{est}\|_1$,
2) $T(\tilde{x}, x_{true})$ is small,
3) $\text{supp}(\tilde{x}) \subseteq \Lambda$,
4) $\|\tilde{x}\|_e - \|x_{true}\|_e$ is small.

The existence of such an $\tilde{x}$ is established by the following lemma.

Lemma 4 (Main Lemma): Let $x_{true}, x_{est}$ be vectors satisfying (11). Then there exists $\tilde{x} \in \mathbb{C}^{K \times 1}$ with $\|\tilde{x}\|_1 = \|x_{est}\|_1$ and where the support of $\tilde{x}$ is a subset of $\Lambda$, such that
\[
a(\theta)^*w \leq \delta \|w\|_2 \quad \text{for all} \quad \theta \in \Omega.
\]

In this case, assuming that $\|w\|_2 = \epsilon$ and letting $y = Ax_{true} + w$, we have that the optimizer $x_{est}$ of (5) satisfies
\[
\|Ax_{est} - Ax_{true} - w\|_2 \leq \epsilon.
\]

By expanding the left hand side, it follows that
\[
\|Ax_{est} - Ax_{true}\|_2 \leq 2w^*(Ax_{est} - Ax_{true})
\]
\[
\leq 2\delta \|Ax_{est}\|_1 + \|Ax_{true}\|_1
\]
\[
\leq 4\delta \|x_{true}\|_1,
\]

and where the support of $\tilde{x}$ is a subset of $\Lambda$, such that

a) $T(x_{est}, \tilde{x}) \leq 2 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right)$,

b) $\|\tilde{x}\|_e - \|x_{true}\|_e \leq 4 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right)$.

Proof: See the Appendix.

By the definition of $T$ we have that
\[
\tilde{T}(x_{est}, x_{true}) \leq T(x_{est}, \tilde{x}) + \|\tilde{x}\|_e - \|x_{true}\|_e,
\]

which gives a bound on the transportation distance between $x_{est}$ and $x_{true}$.

Proposition 5: Let $x_{true}, x_{est}$ be vectors satisfying (11), then
\[
\tilde{T}(x_{est}, x_{true}) \leq 6 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right).
\]

Proof: This follows directly from Lemma 4 and (9).

Thus, Proposition 5 can be used to quantify the worst case error in the recovery (5) in terms of the transportation distance. This is our first main result.

Theorem 6: Let $x_{true} \in \mathbb{C}^{K \times 1}$ be a vector with support $\Lambda$, and let $w \in \mathbb{C}^{K \times 1}$ with $\|w\|_2 \leq \epsilon$. Then the optimal solution $x_{est}$ of (5) with $y = Ax_{true} + w$ satisfies
\[
\tilde{T}(x_{est}, x_{true}) \leq 6 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right),
\]

where $\mu_\Lambda$ is given by Definition 2 and $L = |\Lambda|$.

Proof: The assumptions (11) hold in the given scenario, hence the statement follows from Proposition 5.

Since this is a worst case error bound, it is rather conservative and useful only when the noise level is small. In the next section we use assumptions on the noise distributions to derive bounds with a given confidence level.

VI. ERROR BOUNDS FOR BASIS PURSUIT DECONOISING WITH A CONFIDENCE LEVEL

When the noise is large compared to the signal, any worst case error bound is too large to be useful. However, if the noise is white, it will typically be nearly orthogonal to the columns of $A$ when the number of samples is large. In this section we will use this to give bounds that also applies when the signal to noise ratio is low.

First consider the situation when the noise is nearly orthogonal to all the columns of $A$, i.e., for some $\delta \ll 1$ we have that
\[
\|a(\theta)^*w\| \leq \delta \|w\|_2 \quad \text{for all} \quad \theta \in \Omega.
\]

In this case, assuming that $\|w\|_2 = \epsilon$ and letting $y = Ax_{true} + w$, we have that the optimizer $x_{est}$ of (5) satisfies
\[
\|Ax_{est} - Ax_{true} - w\|_2 \leq \epsilon.
\]

By expanding the left hand side, it follows that
\[
\|Ax_{est} - Ax_{true}\|_2 \leq 2|w^*(Ax_{est} - Ax_{true})|
\]
\[
\leq 2\delta \|Ax_{est}\|_1 + \|Ax_{true}\|_1
\]
\[
\leq 4\delta \|x_{true}\|_1,
\]

and where the support of $\tilde{x}$ is a subset of $\Lambda$, such that

a) $T(x_{est}, \tilde{x}) \leq 2 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right)$,

b) $\|\tilde{x}\|_e - \|x_{true}\|_e \leq 4 \left( \epsilon \sqrt{L(1 + \mu_\Lambda)} \right)$.
where we use that $\|A^*w\|_\infty \leq \delta \|w\|_2$, which follows from (12). Now we may apply Proposition 5 using (13) instead of (11c). This leads to the following result.

**Theorem 7:** Let $x_{\text{true}} \in \mathbb{C}^{K \times 1}$ be a vector with support $\Lambda$, and let $w \in \mathbb{C}^{K \times 1}$ with $\|w\|_2 = \epsilon$ and which satisfies (12). Then the optimal solution $x_{\text{est}}$ of (5) with $y = Ax_{\text{true}} + w$ satisfies

$$
\tilde{T}(x_{\text{est}}, x_{\text{true}}) \leq 6 \left( \sqrt{28\epsilon \|x_{\text{true}}\|_1 L(1 + \mu_{\Lambda}) + \mu_{\Lambda} \|x_{\text{true}}\|_1} \right),
$$

where $\mu_{\Lambda}$ is given by Definition 2 and $L = |\Lambda|$.

Next question is to determine when the near orthogonality assumption (12) is justified. If the noise is white, then for any given $\delta > 0$, the probability that a random noise satisfies (12) goes to 1 as the number of samples $N$ go to infinity. This can be seen from the following proposition.

**Proposition 8:** Let $\delta \in (0, 1)$ be given and let $w \in \mathbb{C}^N$ be a random vector that is complex gaussian with zero mean and unit variance. Then

$$
\text{Prob} \left( \max_{0 \leq \theta \leq 2\pi} |a(\theta)^*w| \geq \delta \|w\|_2 \right) \leq (1 - \delta)^{N-3/2} \left( 1 + \frac{\sqrt{\epsilon^2}}{\sqrt{6}} N^{3/2} \right).
$$

**Proof:** See the Appendix.

By combining Theorem 7 and Proposition 8 it is possible to construct an error bound for $x_{\text{est}}$ that holds with a given confidence level. We will conclude with a few remarks.

**Remark 9:** In general there is a tuning parameter of these sparse estimation methods which has to be determined (in our case the noise level $\epsilon$). In this paper we have assumed that this is given, hence, the results give bounds for (5) when the noise parameter is correctly tuned. See, e.g., [15], [5], [20] for how to find the tuning parameter in practice.

**Remark 10:** The absolute value of the representation vector $|x|_\epsilon$ is considered as the representative of “mass” in this paper. However, we may consider $x$ itself as a complex-valued “mass” distribution where the “mass” also has a direction on the complex plane. The distance developed in [21] may provide a suitable transportation distance between complex-valued mass distributions.

**Remark 11:** Assume that $y$ is given by (1) and we consider the estimates from (5) as the number of samples $N$ go to infinity. Then $\delta \to 0$ with probability 1, $\mu_{\Lambda} \to 0$, and the noise level $\epsilon$ is constant, as $N \to \infty$. In a forthcoming paper we will investigate this in detail, and we conjecture that $x_{\text{est}} \to x_{\text{true}}$ with probability 1 in the weak* topology, i.e., the topology of the Wasserstein metrics. This conjecture should be compared with the results of [22] where the uncertainty of spectral estimation based on covariances is studied. Here it is shown that the uncertainty set converge to a singelton in the weak* topology as the number of covariances tend to infinity.

**Remark 12:** In [23] it is shown that the noiseless version of (5) and (2) gives exact recovery provided the non-zero components of $x_{\text{true}}$ are separated by at least $4K/N$. For the noisy case, they give an error-bound in terms of the $\ell_1$-norm. However, considering our example in Section III where a small noise results in a large $\ell_1$ error, it seems that robust bounds in terms of the $\ell_1$-norm are only useful when the noise level is very small. Moreover, their bound depends on $K/N$. In our work no assumptions are made on $K/N$. In fact, our theory applies for arbitrary grids $\Omega$ and can be generalized to hold for the continuous version of the problem.

**Remark 13:** The theory provided here can be applied to any application where the dictionary is parameterized as a low dimensional manifold. One example is estimation problems in radar, the atoms are often parametrized by parameters such a frequency, location, or Doppler shift. For these example, the transportation cost will depend on the waveform, and can be obtained from, e.g., the autocorrelation function or the ambiguity function (c.f., [10]).

VII. CONCLUSION

The spectral line estimation problem is fundamental in many applications such as antenna array and image processing. In this work we analyse a recently popular sparse method that use $\ell_1$-regularization for this problem. This method have been shown to provide high resolution estimates, and our main contribution in this work is to provide error bounds for this method by using a distance inspired by the theory of optimal mass transport.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank Professor Tryphon Georgiou for interesting discussions and useful suggestions regarding this paper.

IX. APPENDIX

A. Proof of Lemma 3

1) Follows from the definition:

$$
\|A_A^*A_{\Lambda}\|_{1,1} = \max_{\lambda \in \Lambda} \|A_A^*a(\Lambda)\|_1 \\
\leq \max_{\omega \in \Omega} \|A_A^*a(\omega)\|_1 = 1 + \mu_{\Lambda}.
$$

2) This is analogous to (12) in the proof of Theorem 3.5 in [24]. Note that $\|A_A^*A_{\Lambda} - IA\|_{1,1} \leq \mu_{\Lambda}$.

3) The inequality follows from

$$
\|A_A\|_{2,1} = \|A_A\|_{\infty,2} \\
= \max_{\|v\|_2 \leq 1} \|A_A v\|_2 \\
= \max_{\|v\|_2 \leq 1} \sqrt{v^* A_A^* A_A v} \\
\leq \sqrt{L(1 + \mu_{\Lambda})}.
$$
B. Proof of Lemma 4 (Main lemma)

To prove our main lemma, first introduce a partitioning of \(\{\Omega(\lambda), \lambda \in \Lambda\}\) of \(\Omega\). The partition is selected such that for each \(\theta \in \Omega(\lambda)\) it holds that

\[
|a(\theta) \ast a(\lambda)| \geq |a(\theta) \ast a(\phi)| \quad \text{for all } \phi \in \Lambda
\]
or equivalently \(\theta \in \Omega(\lambda)\) for \(\lambda\) maximizing \(\{|a(\theta) \ast a(\phi)| : \phi \in \Lambda\}\). Next, define

\[
\tilde{x} = \sum_{\lambda \in \Lambda} \|x_{\Omega(\lambda)}\|_1 e_{\Lambda}.
\] (14)

Note that the \(\|\tilde{x}\|_1 = \|x_{\text{est}}\|_1\) and that the transportation distance \(T(x, \tilde{x})\) can be bounded as follows.

**Lemma 14:** Let \(\Lambda \subseteq \Omega\), let \(x_{\text{est}} \in \mathbb{C}^{K \times 1}\), and define \(\tilde{x}\) by (14). Then

\[
T(x, \tilde{x}) \leq (1 + \mu_\Lambda)\|x\|_1 - \|A_\Lambda^* A x\|_1.
\]

**Proof:** [Proof of Lemma 14]

The following sequence of inequalities give the result:

\[
\|A_\Lambda^* A x\|_1 = \sum_{\lambda \in \Lambda} |a(\lambda) \ast A x| \leq \sum_{\lambda \in \Lambda} \sum_{\theta \in \Omega(\lambda)} |a(\lambda) \ast a(\theta) x(\theta)| \leq \sum_{\lambda \in \Lambda} \sum_{\theta \in \Omega(\lambda)} |a(\lambda) \ast a(\theta) x(\theta)| + \sum_{\lambda \in \Lambda} \sum_{\theta \notin \Omega(\lambda)} |a(\lambda) \ast a(\theta) x(\theta)| \leq \left(\|x\|_1 - T(x, \tilde{x}) + \mu_\Lambda \|\tilde{x}\|_1\right).
\]

The first two inequalities follow from the triangle inequality. In the last inequality we use that \(\tilde{x}\) has support on \(\Lambda\) and the transportation distance between \(\tilde{x}\) and \(x\) is given by

\[
T(x, \tilde{x}) = \sum_{\theta \in \Omega(\lambda)} \min_{\lambda \in \Lambda} |x(\theta)|(1 - |a_\Lambda^*(\theta)|) = \|x\|_1 - \sum_{\lambda \in \Lambda} \sum_{\theta \in \Omega(\lambda)} |a(\lambda) \ast a(\theta) x(\theta)|.
\]

The next step is the proof of Lemma 4 is the bound in a). Using first 3) in Lemma 3, and then the definition of \(\mu_\Lambda\) together with Lemma 14 we get

\[
2\sqrt{L(1 + \mu_\Lambda)} \geq \|A_\Lambda^* (A x_{\text{true}} - A x_{\text{est}})\|_1 \geq \|A_\Lambda^* A x_{\text{true}}\|_1 - \|A_\Lambda^* A x_{\text{est}}\|_1 \geq (1 - \mu_\Lambda)\|x_{\text{true}}\|_1 + T(x_{\text{est}}, \tilde{x}) - (1 + \mu_\Lambda)\|x_{\text{est}}\|_1
\]

This leads to the desired result a):

\[
T(x_{\text{est}}, \tilde{x}) \leq 2\sqrt{L(1 + \mu_\Lambda)} - (1 - \mu_\Lambda)\|x_{\text{true}}\|_1 + (1 + \mu_\Lambda)\|x_{\text{est}}\|_1 \leq 2\sqrt{L(1 + \mu_\Lambda)} + 2\mu_\Lambda\|x_{\text{true}}\|_1.
\]

Next we will prove b), i.e., the bound on \(\|x_{\text{true}} - |\tilde{x}|\|_1\). Define the two subsets of \(\Lambda\):

\[
\Lambda_+ = \{\lambda \in \Lambda : |\tilde{x}(\lambda)| \leq |x_\lambda|\},
\]

\[
\Lambda_- = \{\lambda \in \Lambda : |\tilde{x}(\lambda)| > |x_\lambda|\},
\]

and following previous notation, denote by \(A_{\Lambda_+}\) the matrix with columns \(a(\lambda)\), where \(\lambda \in \Lambda_+\). Also define the set \(\Omega(\Lambda_+) := \cup_{\lambda \in \Lambda_+} \Omega(\lambda)\). We define the matrix \(A_{\Lambda_-}\), and the set \(\Omega(\Lambda_-)\) analogously. Note that

\[
\|x_{\text{true}} - |\tilde{x}|\|_1 \leq \|x_{\text{true}} - |\tilde{x}|\|_1 + \|x_{\text{true}}\|_1 - |\tilde{x}|_1 = 2\left(\|x_{\text{true}}\|_1 - \|\tilde{x}\|_1\right)
\]

where \(\|\tilde{x}\|_1 = \|x_{\text{est}}\|_1 \leq \|x_{\text{true}}\|_1\) is used in the inequality. We also have that

\[
2\sqrt{L(1 + \mu_\Lambda)} \geq \|A_{\Lambda_+}^* (A x_{\text{true}} - A x_{\text{est}})\|_1 \geq \|A_{\Lambda_+}^* A x_{\text{true}}\|_1 - \|A_{\Lambda_+}^* A x_{\text{est}}\|_1 - \|A_{\Lambda_-}^* A x_{\text{est}}\|_1 \geq (1 - \mu_\Lambda)\|x_{\text{true}}\|_1 - \mu_\Lambda\|x_{\text{true}}\|_1 = \|x_{\text{true}}\|_1 - \|x_{\text{est}}\|_1 \geq \|x_{\text{true}}\|_1 + \|x_{\text{est}}\|_1
\]

Using both these inequalities and noting that \(\|x_{\text{est}}\|_1 = \|\tilde{x}\|_1\) holds (from the definition of \(\tilde{x}\)), we get

\[
\|x_{\text{true}}\|_1 - \|\tilde{x}\|_1 \leq 4\left(\sqrt{L(1 + \mu_\Lambda)} + \mu_\Lambda\|x_{\text{true}}\|_1\right).
\]

C. Proof of Proposition 8

We denote \(a(\theta) = a_\mathbb{R}(\theta) + i a_\mathbb{T}(\theta)\) and similarly, we denote \(w = w_\mathbb{R} + i w_\mathbb{T}\), where \(a_\mathbb{R}(\theta), a_\mathbb{T}(\theta), w_\mathbb{R}, w_\mathbb{T} \in \mathbb{R}^N\), for \(\theta \in [0, 2\pi]\). Thus

\[
|a(\theta)w| = \left((\gamma_1(\theta)^T w_{\text{aug}})^2 + (\gamma_2(\theta)^T w_{\text{aug}})^2\right)^{1/2}
\]

where

\[
\gamma_1(\theta) = \begin{bmatrix} a_{\mathbb{R}}(\theta) \\ a_{\mathbb{T}}(\theta) \end{bmatrix}, \quad \gamma_2(\theta) = \begin{bmatrix} -a_{\mathbb{T}}(\theta) \\ a_{\mathbb{R}}(\theta) \end{bmatrix}, \quad w_{\text{aug}} = \begin{bmatrix} w_\mathbb{R} \\ w_\mathbb{T} \end{bmatrix}.
\]

Note that \(\gamma_1(\theta)\) and \(\gamma_2(\theta)\) are both closed curves on the \(S^{2N-1} = \{z \in \mathbb{R}^{2N} : \|z\|_2 = 1\}\) and \(\gamma_1^* \gamma_2 = 0\). The vector \(w_{\text{aug}}\) is a sample from a uniform distribution on \(S^{2N-1}\), hence it follows from Theorem 4.1 in [25] that

\[
\text{Prob}\left(\max_{0 \leq \theta < 2\pi} |a(\theta)w| \geq \delta\right) \leq \left(1 - \delta\right)^{N-1} \frac{2 \pi \Gamma(N)}{\pi^2} \frac{\Gamma(N - \frac{1}{2})}{\Gamma(N - 1/2)} \times \int_0^{2\pi} \int_0^{2\pi} \left\|\gamma_1(\theta) \cos(\omega) + \gamma_2(\theta) \sin(\omega)\right\|^2 - (\gamma_1^T \gamma_2)^2 \frac{1}{2} d\omega d\theta
\]

where \(\Gamma(\cdot)\) denotes the Gamma function. By Sterlings formula, the following inequality holds

\[
\frac{\Gamma(N)}{\Gamma(N - 1/2)} \leq \frac{\pi^{N/2}}{2^{N/2} \sqrt{N}}
\]

and by straightforward calculations we get

\[
\left\|\gamma_1(\theta) \cos(\omega) + \gamma_2(\theta) \sin(\omega)\right\|^2 - (\gamma_1^T \gamma_2)^2 \leq \frac{N^2}{12}.
\]

By substituting these two inequalities into (15) and using that \(\delta \in (0, 1)\), we get the desired result.
REFERENCES