Sparsity-based Recovery of Finite Alphabet Solutions of Underdetermined Linear System

Abdeldjalil Aïssa-El-Bey Senior Member, IEEE, Dominique Pastor Member, IEEE, Si Mohamed Aziz Sbaï and Yasser Fadlallah Graduate Student Member, IEEE
Institut Télécom; Télécom Bretagne; UMR CNRS 3192 Lab-STIC, Technopôle Brest-Iroise CS 83818 29238 Brest, France
Université européenne de Bretagne

Abstract—We consider the problem of estimating a deterministic finite alphabet vector \( f \) from underdetermined measurements \( y = Af \) where \( A \) is a given (random) matrix. Two new methods are introduced for the recovery of finite alphabet signals via \( \ell_1 \)-minimization through convex optimization. The first method is based on regularization, whereas in the second approach, the problem is formulated as the recovery of sparse signals after a suitable sparse transform. The regularization-based method is less complex than the transformation-based one. When \( p = 2 \) and \( (n, N) \) grows proportionally, it is proved that the same conditions are necessary and sufficient for each method to perform signal recovery. According to these necessary and sufficient conditions, the two methods yield similar performance results. When \( p > 2 \), the existence of necessary and sufficient conditions for signal recovery remains an open problem. However, a conjecture on the behaviour of the transform-based method can be given and is supported by experimental results. In addition, these experimental results show that the transform-based approach performs better than the regularization-based one.

I. INTRODUCTION

The source separation problem is an important research topic in a variety of fields, including speech and audio processing [1], radar processing [2], medical imaging [3], and communication [4]. As such, it has been intensively investigated in the literature in the recent three decades. Basically, source separation aims to estimate original source signals from their mixtures. Approaches in this area can be classified according to the nature of the mixing process (instantaneous, convolutive) and the ratio between the number of sources and the number of sensors of the problem (determined, underdetermined, overdetermined). The more difficult case is clearly the underdetermined case, where the number of sources is more than the number of observed signals and for which solutions cannot be derived without additional assumptions. For instance, the sources can be separated thanks to their sparse representation in the time-frequency domain [5], a source being said to be sparse in a given signal representation domain if most of its samples are close to zero. Another approach can be based on geometric properties of signals as in [6].

In the present paper, we address separation of finite alphabet signals in the instantaneous case with underdetermined known mixing matrix. This problem is important in data communications for symbol demodulation and in image processing. In this respect, we pose the problem in the noiseless case as in [7]–[9]. However, in contrast to the aforementioned references that consider alphabets with two elements only, we study the case of alphabets with any finite size without assuming that the signal to reconstruct is necessarily sparse. Therefore, we propose criteria different from those introduced in [7]–[9]. These criteria exploit appropriate transforms of the initial problem into sparse recovering ones. We thus follow an approach similar to that proposed in [10] and [11], where an \( \ell_0 \)-optimization problem is transformed into a sparse recovering problem involving \( \ell_1 \) minimization. More specifically, we show that separating finite alphabet signals in the instantaneous case with underdetermined known mixing matrix can be rewritten as a sparse recovering problem that can then be relaxed into an \( \ell_1 \)-minimization. This convex relaxation provides good recovery performance for generic random mixing matrices [7], satisfying appropriate asymptotic properties. This is mathematically proved when the alphabet has two elements. For any alphabet size, we give a conjecture, which is supported by experimental results. In the particular case where the alphabet comprises two elements only, these experimental results show that the criteria we propose lead to better performance measurements in less computational time than the solution given in [7].

II. PROBLEM STATEMENT

Henceforth, bold upper cases denote real-valued matrices. The transpose of a given matrix \( A \) is denoted by \( A^T \). All vectors will be column vectors unless transposed. Throughout the paper, \( 0_m \) and \( 1_m \) are the (column) vectors of \( \mathbb{R}^m \) with respectively zero and one entries only. For a vector \( x \), the notation \( x_i \) will stand for the \( i \)-th component of \( x \). As usual, for any integer \( m \), \([1, m]\) stands for \( \{1, 2, \ldots, m\} \). Notation and terminology introduced below are used throughout with always the same meaning.

We hereafter consider the underdetermined linear system of equations or noise free mixing model

\[
y = Ax,
\]

where \( x = [x_1, x_2, \cdots, x_N]^T \) is the \( N \times 1 \) source vector, \( y = [y_1, y_2, \cdots, y_n]^T \) is the \( n \times 1 \) observed vector and \( A \) is an \( n \times N \) real-valued generic random matrix with \( n < N \). For the sake of readiness, we recall the following definition.

Definition II.1 : Generic random matrices [7] A given matrix \( A \) is an \( n \times N \) generic random matrix if all sets of \( n \) columns
are linearly independent with probability 1 and each column is symmetrically distributed about the origin.

System (2) is underdetermined. To discard the case when it has no solution, we shall assume throughout that $\mathbf{A}$ has full-rank, which implies that the image of $\mathbf{A}$ is $\mathbb{R}^n$ and that the solutions of Eq. (1) form a vector space. In telecommunication systems, most signals are generated via some finite alphabet. For practical applications, it thus makes sense to assume that most signals are generated via some finite alphabet. For readiness and because this notion will be used in the sequel, we recall the definition of a sequence of couples $(n,N_n)$ grows proportionally, the authors in [7] conclude that equality is attained when $\sum_{i=1}^{p} N_i = N$ and, thus, when $\mathbf{x} \in \mathcal{F}^N$. Under assumption III.1, $\mathbf{x}$ is therefore the unique $f \in \mathcal{F}^N$ that satisfies Eq. (2).

Solving a $\ell_0$-norm minimization problem is in general complex and may require exhaustive search strategy, which can be intractable in practice for large values of $N$ or $p$. Therefore, by mimicking literature on sparse reconstruction [14], we propose to replace the $\ell_0$-norm by the $\ell_1$-norm as prior knowledge to incorporate into the penalty function. A basic way to involve this constraint is simply to use the $\ell_1$-norm to count the coordinates of any given $\mathbf{x}$.

\[ y = \mathbf{A}f, \quad f \in \mathcal{F}^N \quad (2) \]

In this respect, after exhibiting a sufficient condition for the problem to be solvable in the next section, we propose two new frameworks for the recovery of finite alphabet signals. These two frameworks are presented in Sections IV. Both are based on reformulations of the finite alphabet constraints as sparsity constraints in incomplete measurements. The resulting sparse problems can then be relaxed as for sparse reconstructions in [10] and [11], by using $\ell_1$ minimization. Simulations results are given in Section V.

III. SOLVABILITY ASSUMPTION

In order to recover the finite alphabet signal from the underdetermined linear model, we make the following assumption, where we use the same terminology as [6]:

**Assumption III.1 [Solvability] Model** (2) is $\mathcal{F}$-well-posed, in that there exists a unique element $f \in \mathcal{F}^N$ such that $y = \mathbf{A}f$.

The authors in [7] give a necessary and sufficient condition for assumption III.1 to be satisfied in the integer case, that is, when $\mathcal{F} = \{-1,1\}$. When $\mathbf{A}$ is generic random, they also show that this necessary and sufficient condition is satisfied with probability

\[ P_{n,N} = 2^{-N+1} \sum_{i=0}^{n-1} \binom{N-1}{i}. \]

This probabilistic result is also a direct consequence of [12, Theorem 1.8 & Lemma 5.2] in the particular case where $k = 0$. By using Hoeffding’s inequality [13] when the pair $(n,N)$ grows proportionally, the authors in [7] conclude that the probability that assumption III.1 is satisfied tends to 1 if $\frac{n}{N} > \frac{1}{2}$ and to 0 if $\frac{n}{N} < \frac{1}{2}$. For readiness and because this notion will be used in the sequel, the definition of proportional growth.

**Definition III.1 : Proportional Growth Setting [12]**

A sequence of couples $(n,N_n)$ will be said to grow proportionally if there is $\delta \in (0,1)$ so that $\frac{n}{N_n} \rightarrow \delta$ when $n \rightarrow +\infty$. To alleviate the notation, subscript $n$ will henceforth be omitted.

IV. SPARSITY-BASED RECOVERY METHODS

A solution to Eq. (2) is given in [7] for the special case $\mathcal{F} = \{-1,1\}$. This solution is obtained by solving the $\ell_\infty$ norm minimization

\[ (P_{\infty}) : \quad \arg \min_{\mathbf{x} \in \{-1,1\}^N} \|\mathbf{x}\|_\infty \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (3) \]

since solutions in $\mathcal{F}^N$ to Eq. (2) are basically vertices of the hypercube $[-1,1]^N$. Readily, $(P_{\infty})$ can be solved by linear programming. Nevertheless, by construction, this solution does not apply to alphabets with cardinality $p > 2$. Indeed, when $\mathcal{F}$ has more than 2 elements, the set of solutions in $\mathcal{F}^N$ to Eq. (2) not only involves vertices of $[-1,1]^N$ but also vertices of nested hypercubes included in $[-1,1]^N$.

To overcome this limitation, we hereafter propose two criteria aimed at solving Eq. (2) by seeking sparse solutions of a possibly transformed linear system of equations. The first method presented in the next section is regularization-based. The second one in Section IV-B involves a suitable sparse transform.

A. Regularization-based method

Our first approach is to consider the finite alphabet constraint as prior knowledge to incorporate into the penalty function. A basic way to involve this constraint is simply to use the $\ell_0$-norm to count the coordinates of any given $\mathbf{x} \in \mathbb{R}^N$ that do not belong to $\mathcal{F}$. Thence, the use of $\sum_{i=1}^{p} \|\mathbf{x} - \alpha_i \mathbf{1}_N\|_0$ as the cost to minimize. Actually, we have the following result.

**Proposition IV.1 Under solvability assumption III.1, the unique element of $\mathcal{F}^N$ satisfying Eq. (2) is the unique solution of the optimization problem**

\[ (P_{\mathcal{F},0}) : \quad \arg \min_{\mathbf{x} \in \mathbb{R}^N} \sum_{i=1}^{p} \|\mathbf{x} - \alpha_i \mathbf{1}_N\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \]

**Proof:** Let $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$. For any $i \in [1,p]$, let

\[ N_i = \text{card} \{j \in [1,N] : x_j = \alpha_i\}. \]

Then,

\[ \sum_{i=1}^{p} \|\mathbf{x} - \alpha_i \mathbf{1}_N\|_0 = Np - \sum_{i=1}^{p} N_i \geq N(p - 1). \]

Equality is attained when $\sum_{i=1}^{p} N_i = N$ and, thus, when $\mathbf{x} \in \mathcal{F}^N$. Under assumption III.1, $\mathbf{x}$ is therefore the unique $f \in \mathcal{F}^N$ that satisfies Eq. (2).

Solving a $\ell_0$-norm minimization problem is in general complex and may require exhaustive search strategy, which can be intractable in practice for large values of $N$ or $p$. Therefore, by mimicking literature on sparse reconstruction [14], we propose to replace the $\ell_0$-norm by the $\ell_1$-norm. Thus, we address the much simpler problem $(P_{\mathcal{F},1})$.

\[ (P_{\mathcal{F},1}) : \quad \arg \min_{\mathbf{x} \in \mathbb{R}^N} \sum_{i=1}^{p} \|\mathbf{x} - \alpha_i \mathbf{1}_N\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \]
On the practical side, unlike the $\ell_0$-norm, the $\ell_1$-norm is convex. Furthermore, $(P_{\mathcal{F},1})$ can be solved by linear programming and, thus, in polynomial time. However, problem $(P_{\mathcal{F},1})$ does not always have the same solution as $(P_{\mathcal{F},0})$. For the special case $p = 2$, the next theorem states the conditions under which $(P_{\mathcal{F},1})$ yields the unique solution to Eq. (2) with strong probability.

**Theorem IV.1** Let $A$ be an $n \times N$ generic random matrix. Under solvability assumption III.1, let $f$ be the unique solution to Eq. (2) with coefficients in $F = \{\alpha_1, \alpha_2\}$ ($p = 2$) where $\alpha_1 < \alpha_2$. When $(n, N)$ grows proportionally, the probability that $f$ be the unique solution of $(P_{\mathcal{F},1})$ tends to 0 when $\frac{n}{N} < \frac{1}{2}$ and tends to 1 when $\frac{n}{N} > \frac{1}{2}$.

**Proof:** By using the triangular inequality, we have that
\[ \forall x \in \mathbb{R}^N, \|x - \alpha_1 1_N\|_1 + \|x - \alpha_2 1_N\| \geq N(\alpha_2 - \alpha_1) \]
with equality if and only if, for every $i \in [1, N]$, \[ |x_i - \alpha_1| + |x_i - \alpha_2| = \alpha_2 - \alpha_1. \]
Since $[\alpha_1, \alpha_2] = \{t \in \mathbb{R} : |t - \alpha_1| + |t - \alpha_2| = \alpha_2 - \alpha_1\}$, it follows that:
\[ \arg \min_{x \in \mathbb{R}^N} \left(\|x - \alpha_1 1_N\|_1 + \|x - \alpha_2 1_N\| \right) = [\alpha_1, \alpha_2]^N \]
Therefore, when $p = 2$, solving $(P_{\mathcal{F},1})$ is equivalent to solving
\[ y = Ax, \quad \text{subject to} \quad x \in [\alpha_1, \alpha_2]^N \]
By setting $x' = \frac{x - \alpha_1 1_N + (x - \alpha_2 1_N)}{\alpha_2 - \alpha_1}$, the linear problem (4) is equivalent to
\[ y' = Ax', \quad \text{subject to} \quad y' \in [-1, 1]^N, \]
with $y' = \frac{1}{\alpha_2 - \alpha_1}(2y - (\alpha_2 + \alpha_1)A 1_N)$. In [7], Mangasarian et al. compute the probability that there exists a unique solution to (5) when $A$ is a generic random matrix. When $(n, N)$ grows proportionally, this probability tends to 0 when $\frac{n}{N} < \frac{1}{2}$ and tends to 1 when $\frac{n}{N} > \frac{1}{2}$. Hence, the result.

**B. Transformation-based method**

In this section, we propose a solution based on a suitable sparse transform. Indeed, we plunge $\mathcal{F}^N$ into $\mathbb{R}^{Np}$, so that any element $f \in \mathcal{F}^N$ is represented by a sparse vector $s(f) \in \mathbb{R}^{Np}$. This sparse vector is composed of $N$ consecutive $p$-uples, such that each $p$-uple contains one 1 only and $p - 1$ zeros. By so proceeding, the initial problem (2) is equivalent to a problem of sparse signal recovery from highly incomplete measurements.

For any given $f \in \mathcal{F}^N$, we define
\[ s(f) = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T \in \mathbb{R}^{Np} \]
with $\epsilon_i = [I(f_i = a_1), I(f_i = a_2), \ldots, I(f_i = a_p)]$ and where $I(f_i = a_j)$ is the indicator function equal to one if $f_i = a_j$ and zero otherwise.

We also introduce the matrices $B_\alpha$ and $B_1$ in $\mathbb{R}^{N \times Np}$ defined by:
\[ B_\alpha = I_N \otimes \alpha^T \quad \text{and} \quad B_1 = I_N \otimes 1_N^T, \quad (7) \]
where $\alpha = [\alpha_1, \ldots, \alpha_p]^T$. By construction, we have:
\[ f = B_\alpha s(f) \quad \text{and} \quad B_1 s(f) = 1_N. \]
Therefore, by setting $\Phi = \left(AB_\alpha \quad B_1\right)$, we have
\[ \Phi s(f) = \left(AB_\alpha \right) s(f) = \left( y \right) = 1_N. \]

For any $i \in [1, N]$, we henceforth put $T_i = \{(i - 1)p + 1, ip\}$. Note that, for any $u = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^{Np}$, $\sum_{j \in T_i} u_j$ is the $i$th coordinate of $B_1 u$. This simple remark will prove helpful in the sequel.

On the basis of the above transform, the following proposition states that the recovering of finite alphabet signal $f$ amounts to recovering the sparse signal $s(f)$ from measurements $b$.

**Proposition IV.2** Under solvability assumption III.1 and with definitions (6), (7) and (9), vector $s(f)$ is the unique solution to the optimization problem:
\[ (TP_{\mathcal{F},0}) : \arg \min_{u \in \mathbb{R}^{Np}} \|u\|_0 \quad \text{subject to} \quad \Phi u = b. \]

has a unique solution and is equal to $s(f)$.

**Proof:** Let $u \in \mathbb{R}^{Np}$ be any solution to $(TP_{\mathcal{F},0})$. We thus have $B_1 u = 1_N$, which induces that $\|u\|_0 \geq N$. According to Eqs. (6) and (9), $s(f)$ is solution to $(TP_{\mathcal{F},0})$ with $\|s(f)\|_0 = N$. Therefore, $\|u\|_0 = N$.

For any $i \in [1, N]$, put $T_i = \{(i - 1)p + 1, ip\}$, for any $u = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^{Np}$, $\sum_{j \in T_i} u_j$ is the $i$th coordinate of $B_1 u$. Since $B_1 u = 1_N$ and $\|u\|_0 = N$, for every $i \in [1, N]$, all the values $u_j$ for $j \in T_i$ are null except one, which equals 1. It follows that $B_\alpha u \in \mathcal{F}^N$. Thence, the result by using solvability assumption III.1.

In order to recover $f$ under assumption III.1, it is thus sufficient to find $s(f)$ by seeking the unique solution to $(TP_{\mathcal{F},0})$ and put $f = B_\alpha s(f)$. For the same reasons as those evoked in the previous section, we aim at reducing the complexity cost of the optimization by relaxing the $\ell_0$-norm into the $\ell_1$-norm. In this respect, consider the optimization problem
\[ (TP_{\mathcal{F},1}) : \arg \min_{u \in \mathbb{R}^{Np}} \|u\|_1 \quad \text{subject to} \quad \Phi u = b. \]

Although the set of solutions to $(TP_{\mathcal{F},1})$ is not guaranteed to involve $s(f)$ only, it is however expected that $f = B_\alpha s$ for any solution $s$ to $(TP_{\mathcal{F},1})$. After establishing that solutions to $(TP_{\mathcal{F},1})$ are in $\mathbb{R}^{Np}_+$, Theorem IV.2 below gives a necessary and sufficient condition for $f$ to be the unique element of the image by $B_\alpha$ of the set of all solutions to $(TP_{\mathcal{F},1})$.

**Theorem IV.2** With the notation introduced above,

a) The optimization problem $(TP_{\mathcal{F},1})$ is equivalent to the optimization problem
\[ (TP_{\mathcal{F},1}^+) : \arg \min_{u \in \mathbb{R}^{Np}} \|u\|_1 \quad \text{subject to} \quad \Phi u = b. \]
b) Let \( S \) stand for the set of solutions in \( \mathbb{R}^{N_p} \) to \((TP_{F,1})\). Let \( f \) be the unique element of \( \mathcal{F}^N \) satisfying Eq. (2) under solvability assumption III.1. Vector \( s(f) \) defined by Eq. (6) belongs to \( S \), \( f \in B_\alpha(S) \) and \( f \) is the unique element of \( B_\alpha(S) \) if and only if
\[
\ker \Phi \cap \mathcal{K}^{N,p}_f \subset \ker B_\alpha, \tag{10}
\]
where
\[
\mathcal{K}^{N,p}_f = \{ h \in \mathbb{R}^{N_p} : \forall i \in [1, N], -1 \leq h_{ni}(f) \leq 0 \text{ and } \forall j \in T_i \setminus \{ n_i(f) \}, h_j \geq 0 \}.
\]
and, for every \( i \in [1, N] \), \( n_i(f) \) is the unique element of \( T_i \) such that the \( n_i(f) \)th coordinate of \( s(f) \) is 1.

Proof: Let \( u \in \mathbb{R}^{N_p} \). By the triangle inequality, we have
\[
\| u \|_1 = \sum_{i=1}^{N_p} |u_i| = \sum_{i=1}^{N_p} \sum_{j \in T_i} |u_i| = \sum_{i=1}^{N} \sum_{j \in T_i} u_i.
\]
Since \( \sum_{j \in T_i} u_i \) is the \( i \)th coordinate of \( B_1 u \),
\[
\sum_{i=1}^{N} \sum_{j \in T_i} u_i = \| B_1 u \|_1.
\]
Therefore:
\[
\| u \|_1 \geq \| B_1 u \|_1. \tag{11}
\]
For all \( u \in \mathbb{R}^{N_p} \) such that \( \Phi u = b \), we have \( B_1 u = 1_N \) and thus, according to (11)
\[
\Phi u = b \implies \| u \|_1 \geq N. \tag{12}
\]
Under solvability assumption III.1, \( \Phi s(f) = b \). By construction of \( s(f) \), \( \| s(f) \|_1 = N \). It thus follows from (12) that \( s(f) \) is a solution to \((TP_{F,1})\). Therefore, any solution \( s \) to \((TP_{F,1})\) is such that \( B_1 s = 1_N \) and \( \| s \|_1 = N \). It follows from Lemma B.2, established in Appendix A, that the coordinates of \( s \) are all non-negative. Hence, the first statement of the theorem.

We now prove the second statement. It directly follows from the foregoing that the set \( S \) of solutions to \((TP_{F,1})\) can be written as
\[
S = \{ s(f) + h : \Phi h = 0 \text{ and } \| s(f) + h \|_1 = N \}
\]
with \( \mathcal{C}^{N,p}_f = \{ h \in \mathbb{R}^{N_p} : \| s(f) + h \|_1 = N \} \) and the convention \( x + D = \{ x + d : d \in D \} \) for any \( x \in \mathbb{R}^{N_p} \) and any \( D \subset \mathbb{R}^{N_p} \). According to appendix A, we have
\[
\ker \Phi \cap \mathcal{C}^{N,p}_f = \ker \Phi \cap \mathcal{K}^{N,p}_f. \tag{14}
\]
From (13) and (14), we derive that
\[
S = s(f) + \left( \ker \Phi \cap \mathcal{K}^{N,p}_f \right).
\]
Thereby, according to Eq. (8),
\[
B_\alpha(S) = B_\alpha s(f) + B_\alpha \left( \ker \Phi \cap \mathcal{K}^{N,p}_f \right)
\]
Then, \( f \) is the unique element of \( B_\alpha(S) \) if and only if \( B_\alpha \left( \ker \Phi \cap \mathcal{K}^{N,p}_f \right) = 0 \), which is equivalent to \( \ker \Phi \cap \mathcal{K}^{N,p}_f \subset \ker B_\alpha \).

We now prove that, under certain conditions on \( n \) and \( N \) (precisely on \( n/N \)), generic random matrices satisfy the exact recovery conditions for the special case \( p = 2 \). A conjecture will hereafter be given for the general case.

Theorem IV.3 Let \( A \) be an \( n \times N \) generic random matrix. Let \( \mathcal{F} = \{ \alpha_1, \alpha_2 \} \) with \( \alpha_1 < \alpha_2 \) (p = 2).

(i) \( \Phi \) satisfies (10) with probability
\[
P_{n,N} = 2^{-N+1} \sum_{i=0}^{n-1} \binom{N-1}{i}.
\]
(ii) When \( (n, N) \) grows proportionally, the probability that \( \Phi \) satisfies (10) tends to 1 if \( \frac{n}{N} > \frac{1}{2} \) and tends to 0 if \( \frac{n}{N} < \frac{1}{2} \).

Proof: Given any \( f \in \mathcal{F}^N \) and any \( i \in [1, N] \), we define:
\[
\varepsilon_i(f) = \begin{cases} 1, & \text{if } n_i(f) = 2i - 1 \\ -1, & \text{if } n_i(f) = 2i \end{cases}
\]
Let \( h \) be any element of \( \mathcal{K}^{N,2}_f \). Since \( T_i = \{ 2i-1, 2i \} \) for any \( i \in [1, N] \), we set
\[
\gamma_i = |h_{2i-1}| = |h_{2i}| \in [0, 1]
\]
and
\[
h_{2i-1} = -\varepsilon_i(f) \gamma_i \quad \text{and} \quad h_{2i} = \varepsilon_i(f) \gamma_i.
\]
Then \( h = (D_{\varepsilon_i(f)}(\gamma)) \otimes [-1 \ 1]^T \) with \( \gamma = (\gamma_1, \ldots, \gamma_N)^T \) and
\[
D_{\varepsilon_i} = \text{diag}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N). \tag{15}
\]
Conversely, any vector \( (D_{\varepsilon_i}(\gamma)) \otimes [-1 \ 1]^T \) where \( \gamma \in [0, 1]^N \) belongs to \( \mathcal{K}^{N,2}_f \). Therefore,
\[
\mathcal{K}^{N,2}_f = \left\{ (D_{\varepsilon_i}(\gamma)) \otimes [-1 \ 1]^T : \gamma \in [0, 1]^N \right\}. \tag{16}
\]
We now prove that (10) when \( p = 2 \) is equivalent to \( \mathcal{K}^{N,2}_f \cap \ker \Phi = \{ 0 \} \). In other words, we want to prove the equivalence:
\[
\mathcal{K}^{N,2}_f \cap \ker \Phi \subset \ker B_\alpha \iff \mathcal{K}^{N,2}_f \cap \ker (AB_\alpha) = \{ 0 \}. \tag{17}
\]
In the sequel, we use the following property. For any \( h \in \mathcal{K}^{N,2}_f \), it follows from (7) and usual properties of the Kronecker product that:
\[
B_\alpha h = (I_N \otimes \alpha^T)(D_{\varepsilon_i}(\gamma)) \otimes [-1 \ 1]^T
\]
\[
= (I_N D_{\varepsilon_i}(\gamma)) \otimes (\alpha^T[-1 \ 1]^T)
\]
\[
= (\alpha_2 - \alpha_1) D_{\varepsilon_i}(\gamma). \tag{18}
\]
We begin by the direct implication in (17). So, we suppose that \( K_f^{N,2} \cap \text{Ker } \Phi \subset \text{Ker } B_0 \). We thus have \( \text{Ker } \Phi = \text{Ker } B_1 \cap \text{Ker } (AB_0) \) and \( K_f^{N,2} \subset \text{Ker } B_1 \). Therefore:

\[
K_f^{N,2} \cap \text{Ker } \Phi = K_f^{N,2} \cap \text{Ker } B_1 \cap \text{Ker } (AB_0) = K_f^{N,2} \cap \text{Ker } (AB_0)
\]

Since \( K_f^{N,2} \cap \text{Ker } (AB_0) \) is not empty because \( 0 \) belongs to it, let \( h \) be one of its elements. We then have \( h \in K_f^{N,2} \) and \( AB_0 h = 0 \). It follows from (10) with \( p = 2 \) that \( B_0 h = 0 \).

According to Eqs. (16) and (18), \( h = (D_{\varepsilon(f)}(\gamma) \otimes \text{[-1 1]}^T) \) with \( \gamma \in [0,1]^N \) and \( B_0 h = (\alpha_2 - \alpha_1) D_{\varepsilon(f)}(\gamma) \). Since \( \alpha_1 \neq \alpha_2 \), \( B_0 h = 0 \) is equivalent to \( D_{\varepsilon(f)}(\gamma) = 0 \) and the unique solution in \( \gamma \) to this equality is \( \gamma = 0 \) since \( |D_{\varepsilon(f)}| \neq 0 \). Thereby, \( h = 0 \) so that \( K_f^{N,2} \cap \text{Ker } (AB_0) = \{0\} \). The converse is straightforward.

To get the probability value for (10) to be satisfied, we compute the probability that \( K_f^{N,2} \cap \text{Ker } (AB_0) = \{0\} \). To this end, we start by describing the complementary event. We have \( K_f^{N,2} \cap \text{Ker } (AB_0) \neq \{0\} \) equivalent to the existence of some \( \gamma \in [0,1]^N \setminus \{0\} \) such that \( AB_0 (D_{\varepsilon(f)}(\gamma) \otimes \text{[-1 1]}^T) = 0 \) and, thus, according to (18), such that \( (\alpha_2 - \alpha_1) A D_{\varepsilon(f)}(\gamma) = 0 \). This condition is itself equivalent to the existence of some non-null \( y \in \mathbb{R}^N \) with non-negative coordinates such that \( A D_{\varepsilon(f)}y = 0 \). Gordan’s theorem of the alternative [15] then implies that \( K_f^{N,2} \cap \text{Ker } (AB_0) = \{0\} \) if and only if there exists \( y \in \mathbb{R}^N \) such that \( y^T A D_{\varepsilon(f)} > 0 \), which is equivalent to the fact that the columns of \( A D_{\varepsilon(f)} \) lie in the same hemisphere. Since \( A \) is an \( n \times N \) generic random matrix, \( A D_{\varepsilon(f)} \) is a generic random matrix as well. Statement (i) then derives from (16).

Statement (ii) results from statement (i) and [7, p.5, Eq. (21)] based on Hoeffding’s inequality [13]. Statement (iii) is a direct consequence of statement (ii) and Theorem IV.3.

The experimental results presented in the next section suggest the following conjecture.

**Conjecture IV.1** Let \( A \) be an \( n \times N \) generic random matrix. Assuming solvability assumption III.1, let \( f \) be the unique solution of Eq. (2) with coefficients \( f_i \in \mathcal{F} = \{\alpha_1, \cdots, \alpha_p\} \). When \( (n, N) \) grows proportionally, the exact recovery probability of minimization problem \((TP_{F,1})\) tends to 0 when \( \frac{n}{N} < \frac{p-1}{p} \) and tends to 1 when \( \frac{n}{N} > \frac{p-1}{p} \).

**V. Experimental results**

The following simulation results are aimed at illustrating and completing the theoretical framework exposed above. In particular, these simulation results will emphasize the relevance of conjecture IV.1 for the transform-based approach. These results make it possible to assess the regularization- and transform-based approaches in terms of complexity and performance for different values alphabet sizes. This assessment will also involve Mangasarian’s approach [7] as a reference in the specific case \( p = 2 \). Thanks to these comparisons, we will conclude on the use and relevance of these different methods.

The experimental set-up, common to all the simulations whose results are given below, is the following one. We have used even values for \( p \) and chosen \( \mathcal{F} = \{\pm(2k - 1) : k = 1, \ldots, p/2\} \) for each simulation, we fix \( N \in \{128, 256, 512\} \) and make \( n \) vary so as to assess a significant number of values for ratio \( n/N \). For each pair \((n, N)\) and each method assessed, 1000 iterations of the experiment are carried out. For each iteration, we generate a realization of the generic random matrix \( A \) with size \( n \times N \) by drawing its entries from the normal distribution. We then generate a vector \( f \) with entries drawn uniformly from \( \mathcal{F} \). Once \( f \) is generated, we compute \( y = Af \) and solve \((P_\infty)\) — when \( p = 2 \), \((P_{\infty})\) and \((TP_{F,1})\). We compute the solutions to these optimization problems by using the Matlab CVX toolbox [17], [18], a package for solving convex problems. Finally, we compare the solution \( f \) returned for a given optimization problem to the true signal \( f \). The recovery is said to be correct if the relative error \( \|f-f\|_2 \) is less than \( 10^{-6} \).

Figures 1, 2 and 3 are the phase diagrams for the case \( p = 2 \). They involve Mangasarian’s approach [7] to \((P_\infty)\). These phase diagrams show that the three methods perform equivalently. They also confirm theorems IV.1 and IV.3. In particular, we observe that the breakpoint is actually \( n/N = 1/2 \), as established theoretically. The computational time to solve \((P_\infty)\) [7] however turns out to be higher than that required by the two methods proposed in this paper. This is illustrated by Figures 4, 5 and 6. These results were obtained by using on a PC with OS Linux Ubuntu 12.04 with processor Intel Xeon 3.2 GHz and 3 GO of RAM memory. The values of ratio \( n/N \) considered in these figures are those for which the algorithms under consideration recover the solution with high probability, in accordance with the phase diagrams and the theoretical results of Theorems IV.1 and IV.3. The regularization-based approach is somewhat faster than the transform-based one. Therefore, for \( p = 2 \), it is recommended to use the regularization-based approach since it provides the best trade-off between performance measurements and computational load.

![Figure 1. Phase diagrams of Mangasarian’s approach [7], regularization- and transform-based methods for the case \( p = 2 \) and \( N = 128 \).](image-url)
ian’s approach is not applicable. Figures 7, 8, 9 and 10 provide the phase diagrams of the regularization-based and transformation-based methods for $p = 4$, $p = 6$ and different values of $N$. The regularization-based approach performs less well than the transform-based one, with a breakpoint much higher. The breakpoint of the transform-based approach is actually the value expected by Conjecture IV.1. Clearly, when $p$ increases, the transform-based approach may have to cope with much higher dimensions than the regularization-based one. Therefore, the computational time required by the former can significantly be larger than that of the latter. To conclude this paragraph, we experimentally observed that the set $S$ of solutions to $(TP_{F,1})$ has elements other than $s(f)$ when $p > 2$. Since Proposition IV.2 states that $s(f)$ is the unique solution to $(TP_{F,0})$, whatever the value of $p$, we conclude that problems $(TP_{F,0})$ and $(TP_{F,1})$ are not equivalent when $p > 2$.

VI. DISCUSSION

In this section, through three independent remarks, we discuss to what extent some notions and results, complementary to those mentioned above, relate to the theoretical framework developed in the present paper.

1) We recall that the Kruskal rank $\text{Krank}(A)$ of any given matrix $A$ is the maximal $L$ such that every $L$ columns are linearly independent. The computation of $\text{Krank}(A)$ is NP-complete [19]. On the other hand, it is known that the system $y = Ax$ of linear equations is well-posed for $k$-sparse vectors $x$ — that is, vectors $x$ with $k$ non-zero coordinates — if and only if $2k \leq \text{Krank}(A)$. For the transform-based model of Eq. (9), we derive from Theorem IV.3 that $\text{Krank}(\Phi) \geq 2N$ since $s(f)$ is $N$-sparse.

2) Since $s(f)$ belongs to $\{0,1\}^{Np}$, $s(f)$ is 0-simple in the sense given in [12, Section 5.2, p. 540]. By applying
Figure 6. Computational time in seconds of Mangasarian’s approach [7], regularization- and transform-based methods for the case \( p = 2 \) and \( N = 512 \).

Figure 7. Phase diagrams of the regularization-based method, for \( p = 4 \) and \( N \in \{128, 256, 512\} \).

Figure 8. Phase diagrams of the transformation-based method, for \( p = 4 \) and \( N \in \{128, 256, 512\} \).

Figure 9. Phase diagrams of the regularization-based method, for \( p = 6 \) and \( N \in \{128, 256, 512\} \).

VII. Conclusion

Two frameworks have been proposed for the underdetermined source separation problem of finite alphabet signals. The first one is based on regularization and the second one relies on a suitable sparse transform. Both frameworks are based on convex relaxation aimed at recovering the ideal finite alphabet signal by solving \( \ell_1 \) optimization problems. Simulation results illustrate the effectiveness of the proposed approaches, although the RIP condition is not satisfied. For \( p = 2 \), the regularization- and transform-based approaches provide the same performance measurements as Mangasarian’s approach [7]. These performance measurements confirm the asymptotic behaviour predicted by the theoretical results of Theorems IV.1 and IV.3. However, the computational time of the regularization-based approach is lesser than that of the transform-based one, which, in turn, is significantly smaller than that of Mangasarian’s method. For \( p > 2 \), the com-
putational cost of the regularization-based approach remains lesser than that of the transform-based one. However, the performance measurements of the latter exceed significantly those of the former.

The behaviour of the methods proposed above must theoretically be studied in the general case $p > 2$. In particular, for the transform-based method, Conjecture IV.1 should be proved. In this respect, it could be wondered whether condition $\frac{N}{p} > \frac{(p-1)}{p}$ put forward by the said conjecture could not be regarded as some uncertainty principle similar to those given in [22]–[24]. Another point of importance is the study and the extension of such approaches to the noisy case.

**APPENDIX A**

**PROOF OF EQUALITY (14)**

**A. Proof of direct inclusion** $\text{Ker } \Phi \cap K^{N,p}_f \subset \text{Ker } \Phi \cap C^{N,p}_f$.

For $j \in \{1, 2, \ldots, N\}$, let $s_j(f)$ be the $j^{th}$ coordinate of $s(f)$. Given any $h \in K^{N,p}_f$, we have

$$\|s(f) + h\|_1 = \sum_{i=1}^{N} \left| \sum_{j \in T_i \setminus \{n_i(f)\}} h_j + h_{n_i(f)} + 1 \right|$$

since $s_j(f) = 0$ if $j \neq n_i(f)$ and $s_j(f) = 1$, otherwise. According to the definition of $K^{N,p}_f$, the absolute values are useless in the last equality above. Therefore,

$$\|s(f) + h\|_1 = \sum_{i=1}^{N} \left( \sum_{j \in T_i \setminus \{n_i(f)\}} h_j + h_{n_i(f)} + 1 \right)$$

by definition of $K^{N,p}_f$. Therefore, $K^{N,p}_f \subset C^{N,p}_f$, which implies that

$$\text{Ker } \Phi \cap K^{N,p}_f \subset \text{Ker } \Phi \cap C^{N,p}_f$$

**B. Proof of converse inclusion** $\text{Ker } \Phi \cap C^{N,p}_f \subset \text{Ker } \Phi \cap K^{N,p}_f$.

Let $h$ be any element of $\text{Ker } B_1 \cap C^{N,p}_f$. We have:

$$\|s(f) + h\|_1 = N \quad \text{and} \quad B_1 h = 0$$

We derive from Eq. (8) that

$$B_1(s(f) + h) = 1_N.$$  \hspace{1cm} (22)

According to Eqs. (21), (22) and Lemma B.2, $h \in K^{N,p}_f$. We thus have $\text{Ker } B_1 \cap C^{N,p}_f \subset K^{N,p}_f$. Since $\text{Ker } \Phi \subset \text{Ker } B_1$, it follows from the foregoing that $\text{Ker } \Phi \cap C^{N,p}_f = \text{Ker } \Phi \cap \text{Ker } B_1 \cap C^{N,p}_f \subset \text{Ker } \Phi \cap K^{N,p}_f$.

**APPENDIX B**

**TWO USEFUL LEMMAS**

**Lemma B.1** Let $(a_i)$ be a sequence of real values with $i \in [1, n]$. \hspace{1cm} (23)

$$\sum_{i=1}^{n} |a_i| = n$$

if and only if every $a_i$ is non-negative.

**Proof:** Suppose that the set $I = \{ i \in [1, n] : a_i < 0 \}$ is not empty. We can write

$$\sum_{i=1}^{n} |a_i| = \sum_{i \in I^c} |a_i| + \sum_{i \in I} |a_i| = - \sum_{i \in I} a_i + \sum_{i \in I} a_i,$$

with $I^c = [1, n] \setminus I$. Therefore, equality (23) implies that $\sum_{i \in I} a_i = 0$, which contradicts that $a_i < 0$ for any $i \in I$.

**Lemma B.2** The coordinates of any $u \in \mathbb{R}^N$ such that $B_1 u = 1_N$ and $\|u\|_1 = N$ are non-negative.

**Proof:** If $u = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^N$ is such that $B_1 u = 1_N$, then $\sum_{j \in T_i} u_j = 1$ for every $i \in \{1, 2, \ldots, N\}$. It follows that $\sum_{i=1}^{N} \sum_{j \in T_i} u_j = N$. If, in addition, $\|u\|_1 = N$, we have $\sum_{i=1}^{N} \sum_{j \in T_i} u_j = \sum_{i=1}^{N} \sum_{j \in T_i} u_j$. Thence the result as a direct consequence of Lemma B.1.

**REFERENCES**


