# Sparse phase retrieval via group-sparse optimization

Fabien Lauer, Henrik Ohlsson

# Sparse phase retrieval via group-sparse optimization

F. Lauer[1] and H. Ohlsson[2,3]

[1] LORIA, Université de Lorraine, CNRS, Inria, France

[2] Dept. of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA

[3] Dept. of Electrical Engineering, Linköping University, Sweden

February 24, 2014

**Abstract**

This paper deals with sparse phase retrieval, i.e., the problem of estimating a vector from quadratic measurements under the assumption that few components are nonzero. In particular, we consider the problem of finding the sparsest vector consistent with the measurements and reformulate it as a group-sparse optimization problem with linear constraints. Then, we analyze the convex relaxation of the latter based on the minimization of a block $\ell_1$-norm and show various exact recovery and stability results in the real and complex cases. Invariance to circular shifts and reflections are also discussed for real vectors measured via complex matrices.

## 1 Introduction

The problem of recovering a signal from quadratic measurements is known as phase retrieval. A typical case with many applications, for instance in optics [23] or crystallography [11], is when the measurements correspond to the (squared) magnitude of the Fourier transform of the signal. Here, we consider the more general setting of arbitrary quadratic measurements, $y_i = \boldsymbol{x}^H \boldsymbol{Q}_i \boldsymbol{x}$, $i = 1, \dots, N$, while focusing on the case where the signal $\boldsymbol{x} \in \mathbb{C}^n$ is assumed to be sparse, i.e., with few nonzero entries in $\boldsymbol{x}$. As in compressive sensing [5, 1], which deals with the recovery of sparse signals from linear measurements, the sparsity prior reduces the number of measurements required to recover the signal.

**Related work.** Seminal works on phase retrieval [13, 9, 10, 8] did not consider the sparsity prior. Though these methods were able to incorporate prior information on the support, they were not designed to estimate the support with limited information on its size. More recently, matrix lifting techniques were developed in [3, 2, 22] for phase retrieval and more particularly for the case of sparse signals in [21, 12, 18]. The basic idea is to apply a change of variable resulting in linearized measurement equations with a rank-1 constraint on the new matrix variable $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^H$. Then, the rank-1 constraint is relaxed to the problem of minimizing the rank of the matrix, which is further relaxed to the minimization of the nuclear norm. In these methods, the sparsity prior is typically incorporated as the minimization of an $\ell_1$-norm, which induces a trade-off between the satisfaction of the rank-1 constraint and the sparsity of the solution. Other methods focusing on sparsity are typically iterative, like Fienup-type methods [15] using alternate projections or the GESPAR method [20] implementing a local search strategy with a bi-directional greedy algorithm. These iterative methods usually come without recovery guarantees.

Finally, note that sparse phase retrieval also enters the more general framework of nonlinear compressed sensing, as investigated in [17] for analytic functions computing the measurements, in [7] for quasi-linear functions and in [14] for polynomials.

**Contribution.** We propose a convex approach to the sparse phase retrieval problem. This approach relies on two main steps: the linearization of the constraints inducing a group-sparse structure on the variables and a convex relaxation of the group-sparse optimization problem enforcing this structure. More precisely, the linearization is based on the Veronese map lifting the signal to a higher dimensional space. This map is invariant to a global sign change and preserves sparsity in the sense that the lifted signal has a group-sparsity similar to the sparsity of the original signal. Thus, the proposed method amounts to estimating a group-sparse signal satisfying linear constraints, from which the original signal can be recovered. This estimation relies on a convex relaxation of the group-sparse problem based on a sum of norms, or block $\ell_1$-norm. Thus, while methods based on matrix lifting lead to semi-definite programming problems, the proposed approach yields a more amenable second-order cone programming formulation. This formulation is also easily extended to deal with noisy measurements. In addition to these algorithmic benefits, we derive exact recovery conditions in the noiseless case and stable recovery guarantees in the presence of noise.

Note that the approach taken here is similar in spirit to the one derived in [14] for the more general problem of finding sparse solutions of polynomial systems of equations. However, the analysis in [14] is limited to the real case and does not apply to polynomials without linear terms as the ones found in phase retrieval.

**Paper organization.** For the sake of clarity, we first detail in Sect. 2 the proposed method in the real case before extending it in Sect. 3 to the complex case. The effect of noise and stability results are discussed in Sect. 4. The case of real signals measured via complex vectors is considered in Sect. 5 which also deals with the invariance of the measurements to circular shifts and reflections. Finally, Section 6 tests the proposed methods in numerical experiments.

**Notations.** Matrices are written with bold uppercase letters and vectors in bold lowercase letters, except for the $i$th column $\boldsymbol{A}_i$ of a matrix $\boldsymbol{A}$. The notation $(\boldsymbol{A})_{i,j}$ denotes the element at the $i$th row and $j$th column of a matrix $\boldsymbol{A}$. $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of a complex number, vector or matrix, and i the imaginary unit. The superscripts $^T$ and $^H$ denote the transpose and conjugate transpose, respectively, i.e., $\boldsymbol{z}^H = \overline{\boldsymbol{z}}^T$. $\|\cdot\|_p$ denotes the $\ell_p$-norm in $\mathbb{R}^n$, while $\|\cdot\|$ denotes the norm in $\mathbb{C}^n$ induced by the inner product as $\|\boldsymbol{z}\| = \sqrt{\boldsymbol{z}^H \boldsymbol{z}}$. The $\ell_0$-pseudo-norm of a real or complex vector $\boldsymbol{x}$ of dimension $n$ is defined as $\|\boldsymbol{x}\|_0 = |\{j \in \{1,\dots,n\} : x_j \neq 0\}|$ and denotes the number of nonzero components $x_j$. We also define the $\ell_0$-pseudo-norm of a vector-valued sequence $\{\boldsymbol{u}_i\}_{i=1}^N$ as $\|\{\boldsymbol{u}_i\}_{i=1}^N\|_0 = |\{i \in \{1,\dots,N\} : \boldsymbol{u}_i \neq \boldsymbol{0}\}|$.

# 2 The real case

We write the sparse phase retrieval problem as

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \ & \|\boldsymbol{x}\|_0 \\ \text{s.t. } & y_i = (\boldsymbol{q}_i^T \boldsymbol{x})^2 = \boldsymbol{x}^T \boldsymbol{Q}_i \boldsymbol{x}, \quad i = 1,\dots,N, \end{aligned} \tag{1}$$

where $y_i \in \mathbb{R}$ are the measurements and $\boldsymbol{Q}_i = \boldsymbol{q}_i \boldsymbol{q}_i^T \in \mathbb{R}^{n\times n}$. Due to symmetry, solutions to (1) are defined up to their sign and the goal is to obtain an estimate $\hat{\boldsymbol{x}} = \pm\boldsymbol{x}_0$, for $\boldsymbol{x}_0$ in the solution set of (1). In particular, we are interested in the case where (1) has a unique pair of solutions $\{\boldsymbol{x}_0, -\boldsymbol{x}_0\}$, while conditions ensuring such a uniqueness are discussed in [16, 19].

The proposed method relies on two subsequent relaxations. While the first one linearizes the constraints, the second one convexifies the objective function.

## 2.1 First level of relaxation

Let the Veronese map of degree 2, $\nu : \mathbb{R}^n \to \mathbb{R}^M$, be defined by

$$\nu(\boldsymbol{x}) = [x_1^2,\ x_1x_2,\ \dots,\ x_2^2,\ x_2x_3,\ \dots,\ x_{n-1}^2,\ x_{n-1}x_n,\ x_n^2]^T,$$

and the subscript $_{ij}$ denote the index of its component equal to $x_i x_j$, i.e.,

$$_{ij} = \sum_{k=1}^{\min\{i,j\}-1} (n-k+1) + |j-i| + 1 = \sum_{k=1}^{\min\{i,j\}-1} (n-k) + \min\{i,j\} + |j-i|. \tag{2}$$

This notation is symmetric, i.e., $_{ij}$ and $_{ji}$ denote the same index, and will be used throughtout the paper to index the components of vectors of $\mathbb{R}^M$ or $\mathbb{C}^M$.

The constraints of the phase retrieval problem (1) can be rewritten as

$$\boldsymbol{A}\nu(\boldsymbol{x}) = \boldsymbol{y}$$

with $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $M = \binom{n+1}{2} = n(n+1)/2$.

Let $\boldsymbol{W}_j$ be an $n \times M$-binary matrix such that[1] $\boldsymbol{W}_j\nu(\boldsymbol{x}) = x_j\boldsymbol{x}$, i.e., $\boldsymbol{W}_j\nu(\boldsymbol{x})$ is the vector of $n$ entries corresponding to the monomials in $\nu(\boldsymbol{x})$ including $x_j$. Then, we have

$$x_j = 0 \;\Leftrightarrow\; \boldsymbol{W}_j\nu(\boldsymbol{x}) = \boldsymbol{0}$$

and the objective function in (1) can be written as

$$\|\boldsymbol{x}\|_0 = \|\{\boldsymbol{W}_j\nu(\boldsymbol{x})\}_{j=1}^n\|_0. \tag{3}$$

Thus, (1) can be reformulated as the nonlinear group-sparse optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \; & \|\{\boldsymbol{W}_j\nu(\boldsymbol{x})\}_{j=1}^n\|_0 \\ \text{s.t. } & \boldsymbol{A}\nu(\boldsymbol{x}) = \boldsymbol{y}. \end{aligned} \tag{4}$$

Note that $\nu(\boldsymbol{x}) = \nu(-\boldsymbol{x})$, but that for $\boldsymbol{x} \neq \pm\boldsymbol{x}_0$, $\nu(\boldsymbol{x}) \neq \nu(\boldsymbol{x}_0)$. Thus, the problem can be posed as the one of recovering the value of $\nu(\boldsymbol{x}_0)$, from which $\boldsymbol{x}_0$ can be inferred up to its sign.

To estimate $\nu(\boldsymbol{x}_0)$, we relax (4) to

$$\begin{aligned} \min_{\boldsymbol{v}\in\mathbb{R}^M} \; & \|\{\boldsymbol{W}_j\boldsymbol{v}\}_{j=1}^n\|_0 \\ \text{s.t. } & \boldsymbol{A}\boldsymbol{v} = \boldsymbol{y} \\ & v_{jj} \geq 0,\; j = 1,\dots,n, \end{aligned} \tag{5}$$

where the variables in $\boldsymbol{v}$ estimating the components of $\nu(\boldsymbol{x})$ are not constrained to be interdepent monomials of $n$ base variables, but the last constraints in (5) nonetheless ensure that the $v_{jj}$'s estimating the $x_j^2$'s are positive.

## 2.2 Convex relaxation

Problem (5) is a (linear) group-sparse optimization problem with highly overlapping groups. While groupwise-greedy algorithms, such as the one proposed in [14], can be applied, their analysis is not available for the case of overlapping groups. Therefore, here, we consider the convex relaxation approach which aims at solving (5) via the following surrogate formulation:

$$\begin{aligned} \hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^M} \; & \sum_{j=1}^n \|\boldsymbol{W}_j\boldsymbol{W}\boldsymbol{v}\|_2 \\ \text{s.t. } & \boldsymbol{A}\boldsymbol{v} = \boldsymbol{y} \\ & v_{jj} \geq 0,\; j = 1,\dots,n, \end{aligned} \tag{6}$$

where we introduced the diagonal matrix $\boldsymbol{W}$ of precompensating weights $(\boldsymbol{W})_{i,i} = w_i = \|\boldsymbol{A}_i\|_2$, and which can be solved efficiently by off-the-shelf Second-Order Cone Programming (SOCP) solvers.

[1] More precisely, the entries of $\boldsymbol{W}_j$ are given by $(\boldsymbol{W}_j)_{k,l} = \delta_{l,jk}$, $l = 1,\dots,M$, $k = 1,\dots,n$, where $\delta$ is the Kronecker delta and $_{jk}$ is an index as in (2).

Then, we easily obtain an estimate of $\boldsymbol{x}_0$ from the estimate $\hat{\boldsymbol{v}}$ of $\nu(\boldsymbol{x})$ as $\hat{\boldsymbol{x}} = \nu^{-1}(\hat{\boldsymbol{v}})$, where the inverse mapping $\nu^{-1}$ is defined as

$$\nu^{-1}(\boldsymbol{v}) = \begin{cases} \dfrac{1}{\sqrt{v_{ii}}} \left[v_{1i},\ v_{2i},\ \ldots,\ v_{ni}\right]^T, & \text{if } i > 0 \text{ and } \dfrac{v_{ji}^2}{v_{ii}} = v_{jj},\ \forall j \in \{1, \ldots, n\} \\ \mathbf{0},\ \text{otherwise,} \end{cases}$$

where

$$i = \begin{cases} \min\limits_{j \in \{1,\ldots,n\}} j,\ \text{s.t. } v_{jj} > 0,\ \text{if } \exists j \text{ such that } \ v_{jj} > 0 \\ 0, \quad \text{otherwise.} \end{cases}$$

In the above, the first nonzero entry (of index $i$) is assumed to be positive to fix the signs and make $\nu^{-1}$ injective. This definition also ensures that

$$\nu^{-1}(\nu(\boldsymbol{x})) = \pm \boldsymbol{x}$$

and that

$$\|\nu^{-1}(\boldsymbol{v})\|_0 \leq \|\{\boldsymbol{W}_j \boldsymbol{v}\}_{j=1}^n\|_0 \tag{7}$$

since, for $\nu^{-1}(\boldsymbol{v}) \neq \mathbf{0}$, $\boldsymbol{W}_j \boldsymbol{v} = 0 \Rightarrow v_{ji} = 0 \Rightarrow \left(\nu^{-1}(\boldsymbol{v})\right)_j = 0$.

## 2.3 Analysis

We now turn to theoretical guarantees offered by the proposed approach. First, the following theorem provides the rationale for tackling the sparse phase retrieval problem via the group-sparse optimization formulation (5).

**Theorem 1.** *If the solution $\boldsymbol{v}^*$ to (5) is unique and yields $\boldsymbol{x}^* = \nu^{-1}(\boldsymbol{v}^*) \neq \mathbf{0}$ such that $y_i = (\boldsymbol{x}^*)^T \boldsymbol{Q}_i \boldsymbol{x}^*$, $i = 1, \ldots, N$, then $\{\boldsymbol{x}^*, -\boldsymbol{x}^*\}$ is the unique pair of solutions of (1).*

*Proof.* Assume there is an $\boldsymbol{x}_0 \neq \pm \boldsymbol{x}^*$ satisfying the constraints of (1) and at least as sparse as $\boldsymbol{x}^*$. Then, $\boldsymbol{A}\nu(\boldsymbol{x}_0) = \boldsymbol{y}$ and, by using (3) and (7),

$$\|\{\boldsymbol{W}_j \nu(\boldsymbol{x}_0)\}_{j=1}^n\|_0 = \|\boldsymbol{x}_0\|_0 \leq \|\boldsymbol{x}^*\|_0 \leq \|\{\boldsymbol{W}_j \boldsymbol{v}^*\}_{j=1}^n\|_0,$$

which contradicts the fact that $\boldsymbol{v}^*$ is the unique solution to (4) unless $\nu(\boldsymbol{x}_0) = \boldsymbol{v}^*$. But since $\boldsymbol{x}_0 \neq \pm \boldsymbol{x}^*$, we have $x_{0j}^2 \neq (x_j^*)^2$ for some $j \in \{1, \ldots, n\}$, which implies $(\nu(\boldsymbol{x}_0))_{jj} \neq (\nu(\boldsymbol{x}^*))_{jj} = \left(\nu(\nu^{-1}(\boldsymbol{v}^*))\right)_{jj}$. Therefore, by using Lemma 1 in Appendix A with the assumption $\boldsymbol{x}^* = \nu^{-1}(\boldsymbol{v}^*) \neq \mathbf{0}$, there cannot be such an $\boldsymbol{x}_0$. □

The following results regarding the convex formulation (6) are based on the notion of mutual coherence.

**Definition 1.** *The* mutual coherence *of a matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M] \in \mathbb{R}^{N \times M}$ is*

$$\mu(\boldsymbol{A}) = \max_{1 \leq i < j \leq M} \frac{|\boldsymbol{A}_i^T \boldsymbol{A}_j|}{\|\boldsymbol{A}_i\|_2 \|\boldsymbol{A}_j\|_2}.$$

With this definition, we can state an exact recovery result (proof given in Appendix B.1).

**Theorem 2.** *Let $\boldsymbol{x}_0$ be such that $y_i = \boldsymbol{x}_0^T \boldsymbol{Q}_i \boldsymbol{x}_0$, $i = 1, \ldots, N$, and $\boldsymbol{v}_0 = \nu(\boldsymbol{x}_0)$. If the condition*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2\sqrt{n}} \sqrt{1 + \frac{1}{\mu^2(\boldsymbol{A})}}$$

*holds, then $\boldsymbol{v}_0$ is the unique solution to (6).*

**Corollary 1.** *Let $\boldsymbol{x}_0$ be a feasible point of* (1)*. If the condition*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2\sqrt{n}}\sqrt{1+\frac{1}{\mu^2(\boldsymbol{A})}}$$

*holds, then $\{\boldsymbol{x}_0, -\boldsymbol{x}_0\}$ is the unique pair of solutions to the minimization problem* (1) *and they can be computed as $\boldsymbol{x}_0 = \pm\nu^{-1}(\hat{\boldsymbol{v}})$ with $\hat{\boldsymbol{v}}$ the solution to* (6)*.*

*Proof.* Assume there exists another solution $\boldsymbol{x}_1 \neq \pm\boldsymbol{x}_0$ to (1), and thus with $\|\boldsymbol{x}_1\|_0 \leq \|\boldsymbol{x}_0\|_0$. Then, Theorem 2 implies that both $\nu(\boldsymbol{x}_1)$ and $\nu(\boldsymbol{x}_0)$ are *unique* solutions to (6) and thus that $\nu(\boldsymbol{x}_1) = \nu(\boldsymbol{x}_0) = \hat{\boldsymbol{v}}$. But this contradicts the definition of the mapping $\nu$ implying $\nu(\boldsymbol{x}_1) \neq \nu(\boldsymbol{x}_0)$ whenever $\boldsymbol{x}_1 \neq \pm\boldsymbol{x}_0$. Therefore the assumption $\boldsymbol{x}_1 \neq \pm\boldsymbol{x}_0$ cannot hold and $\{\boldsymbol{x}_0, -\boldsymbol{x}_0\}$ is the unique pair of solutions to (1), while $\nu^{-1}(\hat{\boldsymbol{v}}) = \nu^{-1}(\nu(\boldsymbol{x}_0)) = \pm\boldsymbol{x}_0$. □

# 3 The complex case

Consider now the problem in complex domain:

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{x}\|_0 \\ &\text{s.t. } y_i = |\boldsymbol{q}_i^H\boldsymbol{x}|^2, \quad i = 1,\ldots,N, \end{aligned} \tag{8}$$

where $y_i \in \mathbb{R}$ and $\boldsymbol{q}_i \in \mathbb{C}^n$.

The equations in the problem above are invariant to multiplication by a unit complex scalar $z$ with $|z| = 1$. Thus, there are sets of solutions of the form $T(\boldsymbol{x}_0) = \{\boldsymbol{x} \in \mathbb{C}^n : \boldsymbol{x} = z\boldsymbol{x}_0,\ z \in \mathbb{C},\ |z| = 1\}$, and the goal is to obtain an estimate $\hat{\boldsymbol{x}} \in T(\boldsymbol{x}_0)$, from which $T(\boldsymbol{x}_0) = T(\hat{\boldsymbol{x}})$ can be inferred due to the property of the invariance set:

$$\forall \boldsymbol{x} \in T(\boldsymbol{x}_0), \quad T(\boldsymbol{x}) = T(\boldsymbol{x}_0), \tag{9}$$

which can be proved as follows. Let $\boldsymbol{x} = z\boldsymbol{x}_0$ with $|z| = 1$, then $T(\boldsymbol{x}) = \{\boldsymbol{a} \in \mathbb{C}^n : \boldsymbol{a} = b\boldsymbol{x},\ b \in \mathbb{C},\ |b| = 1\} = \{\boldsymbol{a} \in \mathbb{C}^n : \boldsymbol{a} = bz\boldsymbol{x}_0,\ b \in \mathbb{C},\ |b| = 1\} = \{\boldsymbol{a} \in \mathbb{C}^n : \boldsymbol{a} = c\boldsymbol{x}_0,\ c \in \mathbb{C},\ |c| = 1\} = T(\boldsymbol{x}_0)$.

## 3.1 First level of relaxation

As for the real case, the linearization of the equations will use the Veronese map, which we redefine for complex vectors as follows.

**Definition 2** (Complex Veronese map)**.** *The complex Veronese map $\nu : \mathbb{C}^n \to \mathbb{C}^M$ is defined by*

$$\nu(\boldsymbol{x}) = [x_1\overline{x_1},\ x_1\overline{x_2},\ \ldots,\ x_2\overline{x_2},\ x_2\overline{x_3},\ \ldots,\ x_{n-1}\overline{x_{n-1}},\ x_{n-1}\overline{x_n},\ x_n\overline{x_n}]^T,$$

*for which the subscript $_{ij}$, defined as in* (2)*, denotes the component index such that $(\nu(\boldsymbol{x}))_{ij}$ equals either $x_i\overline{x_j}$ or $x_j\overline{x_i}$ (note that, for all pairs $(i,j)$, there is exactly one such component).*

Note that $\nu(\boldsymbol{x}) = \nu(\boldsymbol{x}')$ for all $\boldsymbol{x}' \in T(\boldsymbol{x})$, since for $1 \leq i \leq j \leq n$, $x_i'\overline{x_j'} = zx_j\overline{zx_j} = |z|^2x_i\overline{x_j} = x_i\overline{x_j}$, but that $\boldsymbol{x}' \notin T(\boldsymbol{x})$ implies $\nu(\boldsymbol{x}) \neq \nu(\boldsymbol{x}')$, since $\boldsymbol{x}' \notin T(\boldsymbol{x}) \Rightarrow |x_j'| \neq |x_j| \Rightarrow x_j'\overline{x_j'} \neq x_j\overline{x_j}$.

Then, we define the inverse mapping as follows.

**Definition 3** (Inverse complex Veronese map)**.** *The inverse complex Veronese map, $\nu^{-1} : \mathbb{C}^M \to \mathbb{C}^n$, is defined by*

$$\nu^{-1}(\boldsymbol{v}) = \begin{cases} \dfrac{1}{\sqrt{v_{ii}}}\,[\overline{v_{1i}},\ \overline{v_{2i}},\ \ldots,\ \overline{v_{ni}}]^T, & \text{if } i > 0 \text{ and } \dfrac{|v_{ji}|^2}{v_{ii}} = v_{jj},\ \forall j \in \{1,\ldots,n\} \\ \mathbf{0}, \text{ otherwise,} \end{cases}$$

*where*

$$i = \begin{cases} \min\limits_{j\in\{1,\ldots,n\}} j, \text{ s.t. } \Re(v_{jj}) > 0,\ \Im(v_{jj}) = 0, \text{ if } \exists j \text{ such that } \Re(v_{jj}) > 0, \Im(v_{jj}) = 0 \\ 0, \quad \text{otherwise.} \end{cases}$$

*In particular, we have $\nu^{-1}(\nu(\boldsymbol{x})) \in T(\boldsymbol{x})$.*

Note that the square root acts on a real and positive number $v_{ii}$ and is thus also a real positive number. This implies $x_i = \sqrt{v_{ii}} \in \mathbb{R}^+$ and that we arbitrarily set $\Im(x_i) = 0$ to fix the value of $z$ in the equation $\nu^{-1}(\nu(\boldsymbol{x})) = z\boldsymbol{x}$, for a complex number $z$ with $|z| = 1$, and thus make $\nu^{-1}$ injective.

With these definitions at hand, the equations in (8) are reformulated via Lemma 3 (all Lemmas are given in Appendix A) as follows:

$$y_i = \boldsymbol{x}^H \boldsymbol{q}_i \boldsymbol{q}_i^H \boldsymbol{x} = 2\Re\left(\nu(\boldsymbol{q}_i)^H \nu(\boldsymbol{x})\right) - \sum_{j=1}^{n} \left(\nu(\boldsymbol{q}_i)\right)_{jj} \left(\nu(\boldsymbol{x})\right)_{jj}, \quad i = 1, \ldots, N.$$

Define the vectors $\boldsymbol{a}_i \in \mathbb{C}^M$, $i = 1, \ldots, N$, with components given by $(\boldsymbol{a}_i)_{jk} = 2\left(\nu(\boldsymbol{q}_i)\right)_{jk}$ for $1 \leq j < k \leq n$ and $(\boldsymbol{a}_i)_{jj} = \left(\nu(\boldsymbol{q}_i)\right)_{jj}$, $j = 1 \ldots, n$. Then, the equations above, linear wrt. to $\nu(\boldsymbol{x})$, can be rewritten as $y_i = \Re\left(\boldsymbol{a}_i^H \nu(\boldsymbol{x})\right)$, $i = 1, \ldots, N$.

Additionally define the binary matrices $\boldsymbol{W}_j$ such that the vector $\boldsymbol{W}_j \nu(\boldsymbol{x})$ contains all the monomials of $\nu(\boldsymbol{x})$ including either $x_j$ or $\overline{x_j}$:

$$\boldsymbol{W}_j \nu(\boldsymbol{x}) = [x_1\overline{x_j},\ x_2\overline{x_j},\ \ldots,\ x_{j-1}\overline{x_j},\ x_j\overline{x_j},\ x_j\overline{x_{j+1}},\ \ldots,\ x_j\overline{x_n}]^T \in \mathbb{C}^n.$$

Then, we have

$$\|\boldsymbol{x}\|_0 = \|\{\boldsymbol{W}_j \nu(\boldsymbol{x})\}_{j=1}^n\|_0 \tag{10}$$

and problem (8) can be rewritten as the nonlinear group-sparse optimization program

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{C}^n} \ & \|\{\boldsymbol{W}_j \nu(\boldsymbol{x})\}_{j=1}^n\|_0 \\ \text{s.t. } & \boldsymbol{y} = \Re\left(\boldsymbol{A}\nu(\boldsymbol{x})\right), \end{aligned} \tag{11}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]^H$.

Next, we relax this formulation by substituting $\boldsymbol{v} \in \mathbb{C}^M$ for $\nu(\boldsymbol{x})$, which yields

$$\begin{aligned} \min_{\boldsymbol{v} \in \mathbb{C}^M} \ & \|\{\boldsymbol{W}_j \boldsymbol{v}\}_{j=1}^n\|_0 \\ \text{s.t. } & \boldsymbol{y} = \Re\left(\boldsymbol{A}\boldsymbol{v}\right) \\ & v_{jj} \in \mathbb{R}^+,\ j = 1, \ldots, n, \end{aligned} \tag{12}$$

where the last constraints ensure that the $v_{jj}$'s estimating the modulus of the base variables are positive real numbers.

The following theorem shows that this relaxation can be used as a proxy to solve the original problem.

**Theorem 3.** *If the solution $\boldsymbol{v}^*$ to (12) is unique and yields $\boldsymbol{x}^* = \nu^{-1}(\boldsymbol{v}^*) \neq \boldsymbol{0}$ such that $y_i = |\boldsymbol{q}_i^H \boldsymbol{x}^*|^2$, $i = 1, \ldots, N$, then $T(\boldsymbol{x}^*)$ is the unique set of solutions of (8).*

*Proof.* Assume there is an $\boldsymbol{x}_0 \neq T(\boldsymbol{x}^*)$ satisfying the constraints of (8) and at least as sparse as $\boldsymbol{x}^*$. Then $\boldsymbol{y} = \Re\left(\boldsymbol{A}\nu(\boldsymbol{x}_0)\right)$, and, by using (10) and Lemma 4,

$$\|\{\boldsymbol{W}_j \nu(\boldsymbol{x}_0)\}_{j=1}^n\|_0 = \|\boldsymbol{x}_0\|_0 \leq \|\boldsymbol{x}^*\|_0 \leq \|\{\boldsymbol{W}_j \boldsymbol{v}^*\}_{j=1}^n\|_0,$$

which contradicts the fact that $\boldsymbol{v}^*$ is the unique solution to (12) unless $\nu(\boldsymbol{x}_0) = \boldsymbol{v}^*$. But since $\boldsymbol{x}_0 \notin T(\boldsymbol{x}^*)$, we have $|x_{0j}| \neq |x_j^*|$ and thus $x_{0j}\overline{x_{0j}} \neq x_j^*\overline{x_j^*}$ for some $j \in \{1, \ldots, n\}$, which implies $\left(\nu(\boldsymbol{x}_0)\right)_{jj} \neq \left(\nu(\boldsymbol{x}^*)\right)_{jj} = \left(\nu(\nu^{-1}(\boldsymbol{v}^*))\right)_{jj}$. Therefore, by using Lemma 5 with the assumption $\boldsymbol{x}^* = \nu^{-1}(\boldsymbol{v}^*) \neq \boldsymbol{0}$, there cannot be such an $\boldsymbol{x}_0$. □

## 3.2 Convex relaxation

We now introduce a second level of relaxation by replacing the $\ell_0$-pseudo norm by a block-$\ell_1$ norm. This leads to a convex relaxation in the form of a SOCP:

$$\begin{aligned} \min_{\boldsymbol{v} \in \mathbb{C}^M} \ & \sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{v}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v})\| \\ \text{s.t. } & \boldsymbol{y} = \Re\left(\boldsymbol{A}\boldsymbol{v}\right) \\ & v_{jj} \in \mathbb{R}^+,\ j = 1, \ldots, n, \end{aligned} \tag{13}$$

where $\boldsymbol{W}_j^R = \boldsymbol{W}_j \boldsymbol{W}^R$ and $\boldsymbol{W}_j^I = \boldsymbol{W}_j \boldsymbol{W}^I$ with $\tilde{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W}^R & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}^I \end{pmatrix}$ a diagonal matrix of precompensating weights given by $(\boldsymbol{W}^R)_{i,i} = \|\Re(\boldsymbol{A}_i)\|_2$ and $(\boldsymbol{W}^I)_{i,i} = \|\Im(\boldsymbol{A}_i)\|_2$.

**Theorem 4.** *Let $\boldsymbol{x}_0$ be such that $y_i = \boldsymbol{x}_0^H \boldsymbol{q}_i \boldsymbol{q}_i^H \boldsymbol{x}_0$, $i = 1, \ldots, N$, $\boldsymbol{v}_0 = \nu(\boldsymbol{x}_0)$ and $\tilde{\boldsymbol{A}} = [\Re(\boldsymbol{A}),\ -\Im(\boldsymbol{A})]$. If the condition*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2\sqrt{2n}} \sqrt{1 + \frac{1}{\mu^2(\tilde{\boldsymbol{A}})}}$$

*holds, then $\boldsymbol{v}_0$ is the unique solution to* (13).

*Proof.* The vector $\boldsymbol{v}_0$ is the unique solution to (13) if the inequality

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{v}_0 + \boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v}_0 + \boldsymbol{\delta})\| > \sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{v}_0) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v}_0)\|$$

holds for all $\boldsymbol{\delta} \in \mathbb{C}^M$ such that $\Re(\boldsymbol{A}(\boldsymbol{v}_0 + \boldsymbol{\delta})) = \boldsymbol{y}$, which implies the constraint $\Re(\boldsymbol{A}\boldsymbol{\delta}) = \boldsymbol{0}$ on $\boldsymbol{\delta}$. The inequality above can be rewritten as

$$\sum_{j \in I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| + \sum_{j \notin I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{v}_0 + \boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v}_0 + \boldsymbol{\delta})\| - \|\boldsymbol{W}_j^R \Re(\boldsymbol{v}_0) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v}_0)\| > 0,$$

where $I_0 = \{j \in \{1, \ldots, n\} : \boldsymbol{W}_j^R \Re(\boldsymbol{v}_0) = \boldsymbol{0} \ \wedge \ \boldsymbol{W}_j^I \Im(\boldsymbol{v}_0) = \boldsymbol{0}\}$. By the triangle inequality, $\|\boldsymbol{a} + \boldsymbol{b}\| - \|\boldsymbol{a}\| \geq -\|\boldsymbol{b}\|$ with $\boldsymbol{a} = \boldsymbol{W}_j^R \Re(\boldsymbol{v}_0) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{v}_0)$, this condition is met if

$$\sum_{j \in I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| - \sum_{j \notin I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| > 0$$

or

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| - 2 \sum_{j \notin I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| > 0. \tag{14}$$

By defining $G_j$ as the set of indexes corresponding to nonzero columns of $\boldsymbol{W}_j$, Lemma 6 yields

$$\begin{aligned} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|^2 &= \sum_{i \in G_j} (w_i^R)^2 \Re(\delta_i)^2 + (w_i^I)^2 \Im(\delta_i)^2 \\ &\leq 2n \frac{\mu^2(\tilde{\boldsymbol{A}})}{1 + \mu^2(\tilde{\boldsymbol{A}})} \|\boldsymbol{W}^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I \Im(\boldsymbol{\delta})\|^2. \end{aligned}$$

Due to the fact that $\bigcup_{k \in \{1, \ldots, n\}} G_k = \{1, \ldots, M\}$, we also have

$$\begin{aligned} \|\boldsymbol{W}^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I \Im(\boldsymbol{\delta})\|^2 &= \sum_{i=1}^{M} (w_i^R)^2 \Re(\delta_i)^2 + (w_i^I)^2 \Im(\delta_i)^2 \\ &\leq \sum_{k=1}^{n} \sum_{i \in G_k} (w_i^R)^2 \Re(\delta_i)^2 + (w_i^I)^2 \Im(\delta_i)^2 \\ &= \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_k^I \Im(\boldsymbol{\delta})\|^2 \\ &\leq \left( \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_k^I \Im(\boldsymbol{\delta})\| \right)^2, \end{aligned}$$

which then leads to

$$\|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|^2 \leq 2n \frac{\mu^2(\tilde{\boldsymbol{A}})}{1 + \mu^2(\tilde{\boldsymbol{A}})} \left( \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_k^I \Im(\boldsymbol{\delta})\| \right)^2.$$

Introducing this result in (14) gives the condition

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| - 2(n - |I_0|)\frac{\mu(\tilde{\boldsymbol{A}})\sqrt{2n}}{\sqrt{1+\mu^2(\tilde{\boldsymbol{A}})}} \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_k^I \Im(\boldsymbol{\delta})\| > 0.$$

Finally, given that $|I_0| = n - \|\boldsymbol{x}_0\|_0$, this yields

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\| - 2\|\boldsymbol{x}_0\|_0\frac{\mu(\tilde{\boldsymbol{A}})\sqrt{2n}}{\sqrt{1+\mu^2(\tilde{\boldsymbol{A}})}} \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_k^I \Im(\boldsymbol{\delta})\| > 0. \tag{15}$$

or, for $\boldsymbol{\delta} \neq \boldsymbol{0}$,

$$\|\boldsymbol{x}_0\|_0 < \frac{\sqrt{1+\mu^2(\tilde{\boldsymbol{A}})}}{2\mu(\tilde{\boldsymbol{A}})\sqrt{2n}},$$

which can be rewritten as in the statement of the Theorem. □

**Corollary 2.** *Let $\boldsymbol{x}_0$ be a feasible point of* (8)*. If the condition*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2\sqrt{2n}}\sqrt{1+\frac{1}{\mu^2(\tilde{\boldsymbol{A}})}}$$

*holds, then $T(\boldsymbol{x}_0)$ is the unique set of solutions to the minimization problem* (8) *and it can be computed as $T(\boldsymbol{x}_0) = T(\nu^{-1}(\hat{\boldsymbol{v}}))$ with $\hat{\boldsymbol{v}}$ the solution to* (13)*.*

*Proof.* Assume there exists another solution $\boldsymbol{x}_1 \notin T(\boldsymbol{x}_0)$ to (8), and thus with $\|\boldsymbol{x}_1\|_0 \leq \|\boldsymbol{x}_0\|_0$. Then, Theorem 4 implies that both $\nu(\boldsymbol{x}_1)$ and $\nu(\boldsymbol{x}_0)$ are *unique* solutions to (13) and thus that $\nu(\boldsymbol{x}_1) = \nu(\boldsymbol{x}_0) = \hat{\boldsymbol{v}}$. But this contradicts Definition 2 implying $\nu(\boldsymbol{x}_1) \neq \nu(\boldsymbol{x}_0)$ whenever $\boldsymbol{x}_1 \notin T(\boldsymbol{x}_0)$. Therefore the assumption $\boldsymbol{x}_1 \notin T(\boldsymbol{x}_0)$ cannot hold and $T(\boldsymbol{x}_0)$ is the unique set of solutions to (8), while $\nu^{-1}(\hat{\boldsymbol{v}}) = \nu^{-1}(\nu(\boldsymbol{x}_0)) \in T(\boldsymbol{x}_0)$, which, by using (9), implies $T(\boldsymbol{x}_0) = T(\nu^{-1}(\hat{\boldsymbol{v}}))$. □

# 4 Stable recovery in the presence of noise

Consider now the case where the measurements $\boldsymbol{y}$ are perturbed by an additive noise $\boldsymbol{e} \in \mathbb{R}^N$ of bounded $\ell_2$-norm, $\|\boldsymbol{e}\|_2 \leq \varepsilon$. Then, the equations in (1) and (8) are of the form $y_i = |\boldsymbol{q}_i^H \boldsymbol{x}|^2 + e_i$ with the noise terms $e_i$ to be estimated together with the sparse signal. Note that in this context, multiple solutions with different noise vectors can be valid. Thus, we aim at stability results bounding the error on the estimates by a function of $\varepsilon$ rather exact recovery ones. Details on the proposed method to achieve these goals are given below, first for real signals and then for complex ones.

## 4.1 Stability in the real case

In the noisy case with real data, the problem that we need to solve becomes

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{e}\in\mathbb{R}^N} \ & \|\boldsymbol{x}\|_0 \\ \text{s.t. } & y_i = |\boldsymbol{q}_i^T \boldsymbol{x}|^2 + e_i, \quad i = 1, \ldots, N, \\ & \|\boldsymbol{e}\|_2 \leq \varepsilon, \end{aligned} \tag{16}$$

where $y_i \in \mathbb{R}$ and $\boldsymbol{q}_i \in \mathbb{R}^n$. Following the approach of Sect. 2 leads to a convex relaxation in the form of the SOCP

$$\begin{aligned} \min_{\boldsymbol{v}\in\mathbb{R}^M} \ & \sum_{j=1}^{n} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{v}\|_2 \\ \text{s.t. } & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{v}\|_2 \leq \varepsilon \\ & v_{jj} \geq 0, \ j = 1, \ldots, n, \end{aligned} \tag{17}$$

for which we have the following stability result.

**Theorem 5.** *Let* $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ *denote a solution to* (16)*. If the inequality*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2n^2(n+1)}\left(1+\frac{1}{\mu(\boldsymbol{A})}\right) \tag{18}$$

*holds, then the solution* $\hat{\boldsymbol{v}}$ *to* (17) *must obey*

$$\|\boldsymbol{W}(\hat{\boldsymbol{v}}-\nu(\boldsymbol{x}_0))\|_2^2 \leq \frac{4n\varepsilon^2}{1-\mu(\boldsymbol{A})[2n^2(n+1)\|\boldsymbol{x}_0\|_0-1]}. \tag{19}$$

*If, in addition,* $\varepsilon = 0$*, then* $\hat{\boldsymbol{x}} = \pm\boldsymbol{x}_0$*.*

We omit the proof which is similar to the one of Theorem 6 in [14], except for the last statement concluding on the stability of $\hat{\boldsymbol{x}}$, and which closely follows the one for the complex case of Theorem 6 to be detailed below.

## 4.2 Stability in the complex case

Consider now the complex variant of the problem perturbed by noise:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{C}^n, \boldsymbol{e}\in\mathbb{R}^N} \ & \|\boldsymbol{x}\|_0 \\ \text{s.t. } & y_i = |\boldsymbol{q}_i^H\boldsymbol{x}|^2 + e_i, \quad i = 1,\dots,N, \\ & \|\boldsymbol{e}\|_2 \leq \varepsilon, \end{aligned} \tag{20}$$

where $y_i \in \mathbb{R}$ and $\boldsymbol{q}_i \in \mathbb{C}^n$. The solution to this problem can be approached via the convex relaxation

$$\begin{aligned} \min_{\boldsymbol{v}\in\mathbb{C}^M} \ & \sum_{j=1}^{n} \|\boldsymbol{W}_j^R\Re(\boldsymbol{v}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{v})\| \\ \text{s.t. } & \|\boldsymbol{y} - \Re\left(\boldsymbol{A}\boldsymbol{v}\right)\|_2 \leq \varepsilon \\ & v_{jj} \in \mathbb{R}^+, \ j = 1,\dots,n. \end{aligned} \tag{21}$$

As for the real case, we have a stability result for the estimation of $\nu(\boldsymbol{x})$.

**Theorem 6.** *Let* $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ *denote a solution to* (20)*. If the inequality*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2n^2(n+1)}\left(1+\frac{1}{\mu(\tilde{\boldsymbol{A}})}\right) \tag{22}$$

*holds, then the solution* $\hat{\boldsymbol{v}}$ *to* (21) *must obey*

$$\|\boldsymbol{W}^R\Re(\hat{\boldsymbol{v}}-\nu(\boldsymbol{x}_0)) + \mathrm{i}\boldsymbol{W}^I\Im(\hat{\boldsymbol{v}}-\nu(\boldsymbol{x}_0))\|^2 \leq \frac{4n\varepsilon^2}{1-\mu(\tilde{\boldsymbol{A}})[2n^2(n+1)\|\boldsymbol{x}_0\|_0-1]}. \tag{23}$$

*If, in addition,* $\varepsilon = 0$*, then* $\hat{\boldsymbol{x}} = \nu^{-1}(\hat{\boldsymbol{v}}) \in T(\boldsymbol{x}_0)$*.*

*Proof.* Assume (20) has a solution $(\boldsymbol{x}_0, \boldsymbol{e}_0)$. Let define $\boldsymbol{v}_0 = \nu(\boldsymbol{x}_0)$ and $\boldsymbol{\delta} = \hat{\boldsymbol{v}} - \boldsymbol{v}_0$. The proof follows a path similar to that of Theorem 3.1 in [6], which was adapted in [14] to the group-sparse setting and which is here further extended to the complex case.

Due to the definition of $\hat{\boldsymbol{v}}$ as a minimizer of (21), $\boldsymbol{\delta}$ must satisfy either

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R\Re(\boldsymbol{v}_0+\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{v}_0+\boldsymbol{\delta})\| < \sum_{j=1}^{n} \|\boldsymbol{W}_j^R\Re(\boldsymbol{v}_0) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{v}_0)\|$$

or $\boldsymbol{\delta} = \boldsymbol{0}$, in which case the statement is obvious. The inequality above can be rewritten as

$$\sum_{j\in I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\| + \sum_{j\notin I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{v}_0+\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{v}_0+\boldsymbol{\delta})\| - \|\boldsymbol{W}_j^R\Re(\boldsymbol{v}_0) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{v}_0)\| < 0,$$

where $I_0 = \{j \in \{1,\dots,n\} : \boldsymbol{W}_j^R\Re(\boldsymbol{v}_0) = \boldsymbol{0} \ \wedge \ \boldsymbol{W}_j^I\Im(\boldsymbol{v}_0) = \boldsymbol{0}\}$. By the triangle inequality, $\|\boldsymbol{a}+\boldsymbol{b}\| - \|\boldsymbol{a}\| \geq -\|\boldsymbol{b}\|$, this implies

$$\sum_{j\in I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\| - \sum_{j\notin I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\| < 0. \tag{24}$$

In addition, $\boldsymbol{\delta}$ must satisfy the constraints in (21) as

$$\|\Re(\boldsymbol{A}(\boldsymbol{v}_0+\boldsymbol{\delta})) - \boldsymbol{y}\|_2 \leq \varepsilon,$$

in which $\boldsymbol{y}$ can be replaced by $\Re(\boldsymbol{A}\boldsymbol{v}_0) + \boldsymbol{e}_0$, leading to

$$\|\Re(\boldsymbol{A}\boldsymbol{\delta}) - \boldsymbol{e}_0\|_2 \leq \varepsilon.$$

Using $\|\boldsymbol{a}\|_2 \leq \|\boldsymbol{a}-\boldsymbol{b}\|_2 + \|\boldsymbol{b}\|_2$, this implies $\|\Re(\boldsymbol{A}\boldsymbol{\delta})\|_2 \leq 2\varepsilon$, which further gives

$$\begin{aligned}
4\varepsilon^2 &\geq \|\Re(\boldsymbol{A}\boldsymbol{\delta})\|_2^2 = \|\tilde{\boldsymbol{A}}\tilde{\boldsymbol{\delta}}\|_2^2 = \|\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 = (\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}})^T\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1}(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}) \\
&= \|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 + (\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}})^T(\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1} - \boldsymbol{I})(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}) \\
&\geq \|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 - \left|(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}})^T(\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1} - \boldsymbol{I})(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}})\right| \\
&\geq \|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 - |\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}|^T|\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1} - \boldsymbol{I}||\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}| \\
&\geq \|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 - \mu(\tilde{\boldsymbol{A}})(\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_1^2 - \|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2) \\
&= (1+\mu(\tilde{\boldsymbol{A}}))\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 - \mu(\tilde{\boldsymbol{A}})\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_1^2
\end{aligned} \tag{25}$$

where we used $\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{W}} = \boldsymbol{I}$ and the fact that the diagonal entries of $|\tilde{\boldsymbol{W}}^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\tilde{\boldsymbol{W}}^{-1} - \boldsymbol{I}|$ are zeros while off-diagonal entries are bounded from above by $\mu(\tilde{\boldsymbol{A}})$.

Due to $\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})$ being a vector with a subset of entries from $\boldsymbol{W}^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I\Im(\boldsymbol{\delta})$, we have $\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 = \|\boldsymbol{W}^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I\Im(\boldsymbol{\delta})\|^2 \geq \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2$, $j = 1,\dots,n$, and thus

$$\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 \geq \frac{1}{n}\sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2. \tag{26}$$

Since the groups defined by the $\boldsymbol{W}_j$'s overlap, $\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2 = \|\boldsymbol{W}^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I\Im(\boldsymbol{\delta})\| \leq \sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|$, and the squared $\ell_1$-norm in (25) can be bounded by

$$\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_1^2 \leq M\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 \leq M\left(\sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|\right)^2. \tag{27}$$

Introducing the bounds (26)–(27) in (25) yields

$$\frac{1+\mu(\tilde{\boldsymbol{A}})}{n}\sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2 - \mu(\tilde{\boldsymbol{A}})M\left(\sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|\right)^2 \leq 4\varepsilon^2. \tag{28}$$

We will now use this inequality to derive an upper bound on $\sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2$, which will also apply to $\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 \leq \sum_{j=1}^n \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2 = \sum_{j=1}^n(\|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta})\|_2^2 + \|\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|_2^2)$, since the groups overlap and the squared components of $\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}$ are summed multiple times in the right-hand side. To derive the upper bound, we first introduce a few notations:

$$a = \sum_{j\in I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|, \quad b = \sum_{j\notin I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|,$$

and

$$c_0 = \frac{\sum_{j\in I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|^2}{\left(\sum_{j\in I_0} \|\boldsymbol{W}_j^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I\Im(\boldsymbol{\delta})\|\right)^2} \in \left[\frac{1}{|I_0|}, 1\right],$$

$$c_1 = \frac{\sum_{j\notin I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|^2}{\left(\sum_{j\notin I_0} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|\right)^2} \in \left[\frac{1}{n-|I_0|}, 1\right],$$

where the box bounds are obtained by classical relations between the $\ell_1$ and $\ell_2$ norms[2]. With these notations, the term to bound is rewritten as

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|^2 = c_0 a^2 + c_1 b^2,$$

while the inequality (28) becomes

$$\frac{1+\mu(\tilde{\boldsymbol{A}})}{n}(c_0 a^2 + c_1 b^2) - \mu(\tilde{\boldsymbol{A}}) M (a+b)^2 \leq 4\varepsilon^2.$$

We further reformulate this constraint by letting $a = \rho b$:

$$\frac{1+\mu(\tilde{\boldsymbol{A}})}{n}(c_0\rho^2 + c_1) b^2 - \mu(\tilde{\boldsymbol{A}}) M (1+\rho)^2 b^2 \leq 4\varepsilon^2. \tag{29}$$

Let $\gamma = (1+\rho)^2/(c_0\rho^2 + c_1)$. Due to (24), we have $a < b$ and thus $\rho \in [0, 1)$, which, together with the bounds on $c_0$ and $c_1$, gives the constraints $1 \leq \gamma \leq 4(n - |I_0|)$. By setting $V = (c_0\rho^2 + c_1)b^2$, (29) is rewritten as

$$\frac{1+\mu(\tilde{\boldsymbol{A}})}{n} V - \mu(\tilde{\boldsymbol{A}}) M \gamma V \leq 4\varepsilon^2,$$

where

$$\frac{1+\mu(\tilde{\boldsymbol{A}})}{n} - \mu(\tilde{\boldsymbol{A}}) M \gamma \geq \frac{1+\mu(\tilde{\boldsymbol{A}})}{n} - 4(n-|I_0|)\mu(\tilde{\boldsymbol{A}}) M > 0,$$

since $\gamma \leq 4(n-|I_0|)$ and the positivity is ensured by the condition (22) and the fact that $\|\boldsymbol{x}_0\|_0 = n - |I_0|$. Thus,

$$\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 \leq \sum_{j=1}^{n} \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|^2 = V \leq \frac{4n\varepsilon^2}{1+\mu(\tilde{\boldsymbol{A}}) - 4\mu(\tilde{\boldsymbol{A}}) n M \|\boldsymbol{x}_0\|_0},$$

which proves the stability result in (23) since $\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\|_2^2 = \|\boldsymbol{W}^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I \Im(\boldsymbol{\delta})\|^2$.

To conclude in the case $\varepsilon = 0$, it remains to see that (23) implies $\hat{\boldsymbol{v}} = \nu(\boldsymbol{x}_0)$ and that Definition 3 ensures $\nu^{-1}(\nu(\boldsymbol{x}_0)) \in T(\boldsymbol{x}_0)$. □

# 5 Complex data, but real solutions

Consider now the following problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \ & \|\boldsymbol{x}\|_0 \\ \text{s.t. } & y_i = |\boldsymbol{q}_i^H \boldsymbol{x}|^2, \quad i = 1, \ldots, N, \end{aligned} \tag{30}$$

where the signal $\boldsymbol{x} \in \mathbb{R}^n$ and measurements $y_i \in \mathbb{R}$ are assumed to be real while the vectors $\boldsymbol{q}_i \in \mathbb{C}^n$ can be complex. In this case, $\boldsymbol{v} = \nu(\boldsymbol{x})$ is also a real vector and solutions can be approximated via a dedicated version of (13):

$$\begin{aligned} \min_{\boldsymbol{v}\in\mathbb{R}^M} \ & \sum_{j=1}^{n} \|\boldsymbol{W}_j^R \boldsymbol{v}\|_2 \\ \text{s.t. } & \boldsymbol{y} = \Re(\boldsymbol{A}\boldsymbol{v}) = \Re(\boldsymbol{A})\,\boldsymbol{v} \\ & v_{jj} \geq 0, \ j = 1, \ldots, n, \end{aligned} \tag{31}$$

[2]Let $\boldsymbol{u} \in \mathbb{R}^{|I_0|}$ with $u_j = \|\boldsymbol{W}_j^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}_j^I \Im(\boldsymbol{\delta})\|$. Then, $c_0 = (\|\boldsymbol{u}\|_2/\|\boldsymbol{u}\|_1)^2$ and the bounds are obtained by the classical relation $\forall \boldsymbol{u} \in \mathbb{R}^k$, $\|\boldsymbol{u}\|_2 \leq \|\boldsymbol{u}\|_1 \leq \sqrt{k}\|\boldsymbol{u}\|_2$.

where $\boldsymbol{A} \in \mathbb{C}^{N\times M}$ is defined as in Sect. 3.1 and $\boldsymbol{W}_j^R = \boldsymbol{W}_j \boldsymbol{W}^R$ with precompensating weights given by $(\boldsymbol{W}^R)_{i,i} = \|\Re(\boldsymbol{A}_i)\|_2$.

For this particular case, Theorem 7 below is similar in spirit to Theorem 4, but allows us to gain a $\sqrt{2}$ factor by using Lemma 7 instead of Lemma 6 in order to take into account that $\boldsymbol{\delta}$ belongs to $\mathbb{R}^M$ (detailed proof given in Appendix B.2). This results in a bound on $\|\boldsymbol{x}_0\|_0$ similar to the one in Theorem 2 for the real case and based on the mutual coherence of the real part of $\boldsymbol{A}$. Since $\mu(\Re(\boldsymbol{A})) \leq \mu(\tilde{\boldsymbol{A}})$, this also improves (relaxes) the bound compared with the one of Theorem 4.

**Theorem 7.** *Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ be such that $y_i = \boldsymbol{x}_0^T \boldsymbol{q}_i \boldsymbol{q}_i^H \boldsymbol{x}_0$, $i = 1, \ldots, N$, and $\boldsymbol{v}_0 = \nu(\boldsymbol{x}_0)$. If the condition*

$$\|\boldsymbol{x}_0\|_0 < \frac{1}{2\sqrt{n}} \sqrt{1 + \frac{1}{\mu^2(\Re(\boldsymbol{A}))}}$$

*holds, then $\boldsymbol{v}_0$ is the unique solution to* (31).

In a typical instance of Problem (30), the measurements $\boldsymbol{y}$ correspond to the power spectrum of a real signal. However, in this case, Theorem 7 does not apply since the solution is known not to be unique due to invariances of the measurements. The following subsections first describe a practical technique to deal with such invariances and then focus on the power spectrum case.

## 5.1 Invariances

Consider the case where the measurements are invariant to some transformation of the signal. A typical example is when $\boldsymbol{x} \in \mathbb{R}^n$ and $\{\boldsymbol{q}_i \in \mathbb{C}^n\}$ forms a Fourier basis. Then, the equations $y_i = |\boldsymbol{q}_i^H \boldsymbol{x}|^2$ are invariant to circular shifts of the components of $\boldsymbol{x}$. This is problematic since shifted versions of $\boldsymbol{x}$ lead to shuffled[3] versions of $\nu(\boldsymbol{x})$. Thus, multiple shuffled $\boldsymbol{v}$'s satisfy the linearized constraints, $y_i = \Re(\boldsymbol{A})\boldsymbol{v}$, and so does any convex combination of them. These convex combinations need not be sparse as they combine vectors with different sparsity patterns, but lead to lower or equal values of the convex cost function of (31).

To circumvent this issue, we need to linearize the constraints by a shift-invariant transformation $\phi$, such that $\phi(\boldsymbol{x}_0) = \phi(\boldsymbol{x}_1)$ for $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ two shifted versions of the same vector. To lead to an effective estimation method, the transformation $\phi$ must also retain sparsity in a sense similar to $\nu$.

Consider the transformation $\phi : \mathbb{R}^n \to \mathbb{R}^M$ defined by[4]

$$\phi(\boldsymbol{x}) = \nu(\text{shift}(\boldsymbol{x}, 1-k)), \quad \text{with } k = \arg \max_{i\in\{1,\ldots,n\}} |x_i|, \tag{32}$$

where $\text{shift}(\cdot, k)$ stands for the $k$-steps circular shift operator. The transformation $\phi$ in (32) first shifts the vector $\boldsymbol{x}$ such that the first component is the one with maximal magnitude. This results in a shift-invariant transformation which retains sparsity and the linearization ability via the mapping $\nu$. The linearized constraints remain the same, i.e., $y_i = \Re(\boldsymbol{A})\boldsymbol{\phi}$, but another constraint must be added to (31) in order to account for the shift-invariance.

More precisely, the definition of $\phi$ ensures that

$$(\phi(\boldsymbol{x}))_{11} \geq (\phi(\boldsymbol{x}))_{jj}, \quad j = 2, \ldots, n.$$

Thus, a valid vector $\hat{\boldsymbol{\phi}}$ estimating $\phi(\boldsymbol{x}_0)$ can be found by solving

$$\begin{aligned} \hat{\boldsymbol{\phi}} = \arg \min_{\boldsymbol{\phi}\in\mathbb{R}^M} & \sum_{j=2}^{n} \|\boldsymbol{W}_j^R \boldsymbol{\phi}\|_2 \\ \text{s.t. } & \boldsymbol{y} = \Re(\boldsymbol{A})\,\boldsymbol{\phi} \\ & \phi_{11} \geq \phi_{jj} \geq 0,\ j = 2, \ldots, n. \end{aligned} \tag{33}$$

[3] Circular shifts of $\boldsymbol{x}$ lead to rearrangements of the components of $\nu(\boldsymbol{x})$ which are not exactly circular shifts. For example, with $\boldsymbol{x}_0 = [1,1,0,0]^T$ and $\boldsymbol{x}_1 = [0,1,1,0]^T$, we have $\nu(\boldsymbol{x}_0) = [1,1,0,0,1,0,0,0,0,0]^T$ and $\nu(\boldsymbol{x}_1) = [0,0,0,0,1,1,0,1,0,0]^T$.

[4] In the case where the argmax in (32) is not a singleton, $k$ is arbitrarily set to the minimum of the indexes in the argmax and $\phi$ cannot be shift-invariant. Assumptions regarding this issue will be made clear in Proposition 1 and Theorem 8 below.

Note that the the cost function does not involve the first group of variables since $x_1$ is assumed to be nonzero. In comparison with (31), this formulation is still convex but cannot have multiple solutions that are shifted/shuffled versions of one another.

Finally, the set of shifted solutions $\{\boldsymbol{x}_k\}_{k=1}^n$ to (30) is approximated by the set

$$\phi^{-1}(\hat{\boldsymbol{\phi}}) = \left\{\text{shift}(\nu^{-1}(\hat{\boldsymbol{\phi}}), k)\right\}_{k=1}^{n}.$$

Following a similar approach, we can additionally take into account reflections by defining

$$\begin{cases} \phi(\boldsymbol{x}) = \nu(\boldsymbol{x}_2) \\ \boldsymbol{x}_2 = \varphi(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}_1, \text{ if } \sum_{i=2}^{n/2} |x_{1i}|^2 \geq \sum_{i=2+\frac{n}{2}}^{n} |x_{1i}|^2 \\ \text{shift}(\text{reflection}(\boldsymbol{x}_1), 1), \text{ otherwise} \end{cases} \\ \boldsymbol{x}_1 = \text{shift}(\boldsymbol{x}, 1-k) \\ k = \arg\max_{i\in\{1,\dots,n\}} |x_i|, \end{cases} \tag{34}$$

where we assume $n$ to be even and reflection$(\cdot)$ is the reflection operator defined by reflection$(\boldsymbol{x}) = [x_n, x_{n-1}, \dots, x_2, x_1]^T$. In plain words, the transformation $\phi$ in (34) first shifts the vector $\boldsymbol{x}$ such that the first component is the one with maximal magnitude. Then, it applies a centered reflection to the shifted $\boldsymbol{x}$, named $\boldsymbol{x}_1$, only if the sum of squares over the first entries of $\boldsymbol{x}_1$ (without the first one) is smaller than the one over the last ones. If this is the case, the result is shifted again to recover the first component of $\boldsymbol{x}_2$ with maximal magnitude. Finally, the Veronese map is applied to $\boldsymbol{x}_2$ to give $\phi(\boldsymbol{x})$.

Invariance of $\phi = \nu \circ \varphi$ to circular shifts and reflections is implied by the invariance of $\varphi$ given in the Proposition below (proof in Appendix B.3).

**Proposition 1.** *For all $\boldsymbol{x} \in \mathbb{C}^n$ such that $\left|\arg\max_{i\in\{1,\dots,n\}} |x_i|\right| = 1$, the following statements hold for the transformation $\varphi$ defined in* (34)*:*

1. *$\varphi$ is idempotent, i.e., $\varphi \circ \varphi(\boldsymbol{x}) = \varphi(\boldsymbol{x})$;*

*and, for all $\boldsymbol{x}$ additionally satisfying $\sum_{i=2}^{n/2} |x_{1i}|^2 \neq \sum_{i=2+\frac{n}{2}}^{n} |x_{1i}|^2$ with $\boldsymbol{x}_1 = \textit{shift}(\boldsymbol{x}, \arg\max_{i\in\{1,\dots,n\}} |x_i|)$,*

2. *$\varphi$ is shift-invariant, i.e., $\forall s \in \mathbb{Z},\ \varphi(\boldsymbol{x}) = \varphi(\textit{shift}(\boldsymbol{x}, s))$;*

3. *$\varphi$ is reflection-invariant, i.e., $\varphi(\boldsymbol{x}) = \varphi(\textit{reflection}(\boldsymbol{x}))$.*

*Note that statements 2 and 3 imply that $\varphi$ is invariant to any combination of shifts and reflections.*

Proposition 1 also shows the idempotence of $\varphi$, which of course does not transfer to $\phi$ directly but which is however very useful. Indeed, this allows us to test if a candidate $\boldsymbol{x}_2$, supposed to be the result of $\varphi$ applied to some vector, is consistent with the definition of $\varphi$ as $\boldsymbol{x}_2 \stackrel{?}{=} \varphi(\boldsymbol{x}_2)$, which can be checked via simple inequalities. Since these inequalities only involve the (squared) magnitude of the entries in the vector, they can also be easily embedded as linear constraints in (33) to compute an estimate $\hat{\boldsymbol{\phi}}$ that is consistent with the transformation (34). This yields the convex program

$$\begin{aligned} \hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathbb{R}^M} \ & \sum_{j=2}^{n} \|\boldsymbol{W}_j^R \boldsymbol{\phi}\|_2 && (35) \\ \text{s.t. } & \boldsymbol{y} = \Re(\boldsymbol{A})\,\boldsymbol{\phi} && \text{(data fitting)} \\ & \phi_{11} \geq \phi_{jj} \geq 0,\ j = 2, \dots, n && \text{(shift-invariance)} \\ & \sum_{i=2}^{n/2} \phi_{ii} \geq \sum_{i=2+\frac{n}{2}}^{n} \phi_{ii} && \text{(reflection-invariance),} \end{aligned}$$

where $\phi_{jj}$ estimates $|x_j|^2$. The advantage of using (35) is that it has a single solution independently of the number of shifted/reflected solutions to (30), in the sense of the next theorem.

**Theorem 8.** *If Problem* (30) *has a unique set of shifted/reflected solutions* $S(\boldsymbol{x}_0) = \{\textit{shift}(\boldsymbol{x}_0, k) : k \in \{1, \dots, n\}\} \cup \{\textit{shift}(\textit{reflection}(\boldsymbol{x}_0), k) : k \in \{1, \dots, n\}\}$*, and if* $\left|\arg\max_{i\in\{1,\dots,n\}} |x_{0i}|\right| = 1$*, then there is exactly one vector* $\nu(\boldsymbol{x})$ *with* $\boldsymbol{x} \in S(\boldsymbol{x}_0)$ *in the feasible set of* (35)*.*

*Proof.* Since all $\boldsymbol{x} \in S(\boldsymbol{x}_0)$ are solutions to (30), they satisfy the first constraint in (35) as $\boldsymbol{y} = \Re(\boldsymbol{A})\,\nu(\boldsymbol{x})$. By Proposition 1, $\varphi$ is constant over the set $S(\boldsymbol{x}_0)$ and equal to one of the vectors from this set, say $\boldsymbol{x}_0$, for which we have in particular $\boldsymbol{x}_0 = \varphi(\boldsymbol{x}_0)$. Then, $\nu(\boldsymbol{x}_0) = \nu \circ \varphi(\boldsymbol{x}_0) = \phi(\boldsymbol{x}_0)$ and thus $\nu(\boldsymbol{x}_0)$ satisfies all the constraints of (35) and is a feasible point.

Now take an $\boldsymbol{x} \in S(\boldsymbol{x}_0)$ with $\boldsymbol{x} \neq \boldsymbol{x}_0$. Then, either $\boldsymbol{x} = \text{shift}(\boldsymbol{x}_0, k)$ or $\boldsymbol{x} = \text{shift}(\text{reflection}(\boldsymbol{x}_0), k)$ for some $k \in \{1, \dots, n\}$. This implies $\arg\max_i(\nu(\boldsymbol{x}))_{ii} = \arg\max_i |x_i| \neq \arg\max_i |x_{0i}| = \arg\max_i(\nu(\boldsymbol{x}_0))_{ii}$. Since $\boldsymbol{x}_0$ satisfies the constraints in (35), we have $\arg\max_i(\nu(\boldsymbol{x}_0))_{ii} = 1$ and thus $\arg\max_i(\nu(\boldsymbol{x}))_{ii} \neq 1$, which shows that $\nu(\boldsymbol{x})$ violates the second constraint and is not a feasible point of (35). □

**Illustrative example.** Consider the vectors ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not related to the notations of (34))

$$\boldsymbol{x}_1 = [1, 2, 3, 4, 0, 0]^T, \quad \boldsymbol{x}_2 = [4, 0, 0, 1, 2, 3]^T, \quad \boldsymbol{x}_3 = [0, 0, 4, 3, 2, 1]^T,$$

$$\boldsymbol{x}_4 = [2, 1, 0, 0, 4, 3]^T, \quad \boldsymbol{x}_5 = [4, 3, 2, 1, 0, 0]^T,$$

which are all obtained by shifts and reflections of the same vector. They all lead to the same $\phi(\boldsymbol{x}_i) = \nu(\boldsymbol{x}_5)$. Indeed, the transformation $\phi$ first applies the required shift and reflect operations to map $\boldsymbol{x}_i$ to $\boldsymbol{x}_5$ which has a first component with the largest magnitude and the largest half-sum over its first entries. Then, $\phi$ computes the Veronese map of $\boldsymbol{x}_5$. Also note that all the $\nu(\boldsymbol{x}_i)$ are feasible with respect to $\boldsymbol{y} = \Re(\boldsymbol{A})\,\nu(\boldsymbol{x}_i)$, but only $\nu(\boldsymbol{x}_5)$ satisfies the additional constraints implementing shift and reflection invariance in (35). Therefore, we can obtain $\hat{\boldsymbol{\phi}} = \nu(\boldsymbol{x}_5)$ as the unique solution to (35).

## 5.2 Support recovery from the power spectrum

When the measurements correspond to the squared magnitude of the Fourier transform of the signal, i.e., its (squared) power spectrum, there is another issue beside shift/reflection-invariances. Even with the correct support $\mathcal{S} = \textit{supp}(\boldsymbol{x}_0)$, the linear system $\boldsymbol{y} = \Re(\boldsymbol{A}_{\mathcal{S}})\boldsymbol{\phi}_{\mathcal{S}}$, where $\boldsymbol{A}_{\mathcal{S}}$ is the submatrix of $\boldsymbol{A}$ with the corresponding columns, is under-determined. More precisely, for all $i$ and all $j$, $|q_{ij}| = 1$, which implies $(\boldsymbol{a}_i)_{jj} = (\nu(\boldsymbol{q}_i))_{jj} = 1$ and $\boldsymbol{A}_{jj} = \boldsymbol{1}$. Thus, even when limited to the support $\mathcal{S}$, $\boldsymbol{A}_{\mathcal{S}}$ has $|\mathcal{S}|$ similar columns and is rank-defficient.

Therefore, in this case the proposed approach cannot exactly recover $\boldsymbol{x}_0$ and (35) is used to estimate the support $\mathcal{S}$. Indeed, knowing the correct support can be useful for other methods dedicated to the classical phase retrieval problem [8].

Though this case seems unfavorable, if the number of measurements satisfies $N \geq 2n - 1$, the autocorreleation of $\boldsymbol{x}_0$ (with zero padding),

$$r_k = \sum_{i=1}^{n} x_i x_{i+k}, \quad k = -n+1, \dots, n-1,$$

can be computed via the inverse Fourier transform of $\boldsymbol{y}$ and thus is available. This information can be included as linear constraints on the components $\phi_{i(i+k)}$ of $\boldsymbol{\phi}$ estimating the cross products $x_i x_{i+k}$ to drive the solution towards a satisfactory one.

In addition, following [12, 20], this can be used to restrict the support of the solution as follows. First, note that with $N \geq 2n - 1$, the measurements are not invariant to circular shifts of $\boldsymbol{x}$ but of $\boldsymbol{x}$ with zero-padding. Thus, we cannot assume $|x_1| \geq |x_j|$, $j = 2, \dots, n$. However, we can fix $x_1 \neq 0$. Then, for all base variable index $j \in \{1, \dots, n\}$,

$$r_k = 0,\ k = j-1, \dots, n-1 \quad \Rightarrow \quad x_j = 0. \tag{36}$$

Thus, the corresponding variables in $\boldsymbol{\phi}$ can be set to zero, i.e., $\boldsymbol{W}_j\boldsymbol{\phi} = \boldsymbol{0}$. Moreover, let $j_m$ be the minimum of the $j$'s satisfying the condition in (36), then $x_{j_m-1} \neq 0$, which can be favored

by removing the corresponding term in the cost function. Combining all the information on the solution in a convex program leads to the final formulation

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathbb{R}^M} \sum_{j=2}^{j_m-2} \|\boldsymbol{W}_j^R \boldsymbol{\phi}\|_2 \tag{37}$$

$$\begin{aligned}
\text{s.t. } & \boldsymbol{y} = \Re(\boldsymbol{A})\,\boldsymbol{\phi} && \text{(data fitting)}\\
& \phi_{jj} \geq 0,\ j = 1,\ldots,j_m-1 && \text{(structural knowledge)}\\
& \sum_{i=2}^{n/2} \phi_{ii} \geq \sum_{i=2+\frac{n}{2}}^{n} \phi_{ii} && \text{(reflection-invariance)}\\
& \sum_{i=1}^{n-k} \phi_{i(i+k)} = r_k,\ k = 0,\ldots,n-1 && \text{(autocorrelation)}\\
& \boldsymbol{W}_j\boldsymbol{\phi} = \boldsymbol{0},\ j = j_m,\ldots,n && \text{(restricted support).}
\end{aligned}$$

A noise-tolerant version of (37) is obtained by replacing the constraint $\boldsymbol{y} = \Re(\boldsymbol{A})\,\boldsymbol{\phi}$ by $\|\boldsymbol{y} - \Re(\boldsymbol{A})\,\boldsymbol{\phi}\|_2 \leq \varepsilon$. This modification also applies to (31) and (35).

# 6 Experiments

For the experiments, we consider two variants of the proposed approach: the convex method described in details in the previous sections and the greedy method for solving the group-sparse optimization problems (5) and (12). Implementation details for these two methods are as described in [14], with slight modifications to handle complex variables. In particular, the convex method uses the iterative reweighting of [4] adapted to the group-sparse setting to enhance the sparsity of the solution. The greedy method starts with an empty support and, at each iteration, adds the group of variables in $\boldsymbol{v}$ corresponding to the base variable $x_j$ that results in the best approximation of $\boldsymbol{y}$.

For complex signals, $\boldsymbol{x}_0$, we measure the relative error corresponding to the normalized distance between the estimate $\hat{\boldsymbol{x}}$ and the set $T(\boldsymbol{x}_0)$, i.e., $\min_{\boldsymbol{x}\in T(\boldsymbol{x}_0)} \|\hat{\boldsymbol{x}}-\boldsymbol{x}\|/\|\boldsymbol{x}\|$. Exact recovery is detected when this error is smaller than $10^{-6}\|\boldsymbol{x}_0\|$.

**Exact recovery in the noiseless case.** We estimate the probability of exact recovery at various sparsity levels, $\|\boldsymbol{x}_0\|_0$, in the complex case where $n = 20$ and $N = 50$. For each sparsity level, the probability is estimated as the percentage of successful trials over a Monte Carlo experiment with 100 trials. In each trial, $N$ complex vectors $\boldsymbol{q}_i \in \mathbb{C}^n$ are drawn from a zero-mean Gaussian distribution of unit variance to measure a random signal $\boldsymbol{x}_0 \in \mathbb{C}^n$ with $\|\boldsymbol{x}_0\|_0$ nonzero entries at random locations whose values are drawn from a zero-mean Gaussian distribution of unit variance. Results shown in Fig. 1 indicate that the proposed approach can exactly recover sufficiently sparse signals with high probability.

Similar experiments are performed to evaluate the influence of the number of measurements $N$ on the probability of exact recovery. Results reported in the right plot of Fig. 1 show that, with a sparsity level of $\|\boldsymbol{x}_0\|_0 = 4$, the convex method requires more measurements for perfect recovery than the greedy strategy. However, $N \approx \|\boldsymbol{x}_0\|_0^3$ measurements are already sufficient to exactly recover $\boldsymbol{x}_0$ in all trials.

**Stable recovery in the noisy case.** We now consider the noisy case where $y_i = |\boldsymbol{q}_i^H \boldsymbol{x}_0|^2 + e_i$, $i = 1,\ldots,N$, and $\|\boldsymbol{e}\|_2 \leq \varepsilon$. Over 100 trials, Figure 2 reports the mean relative error and the rate of successful support recovery for $N = 50$ and $\varepsilon = 3$. These results show that sufficiently sparse signals can be accurately estimated from noisy quadratic measurements and that the probability of support recovery in this case follows a curve similar to the one of the probability of exact recovery in the noiseless case. This means that the methods are robust to noise regarding the recovery of the correct support.

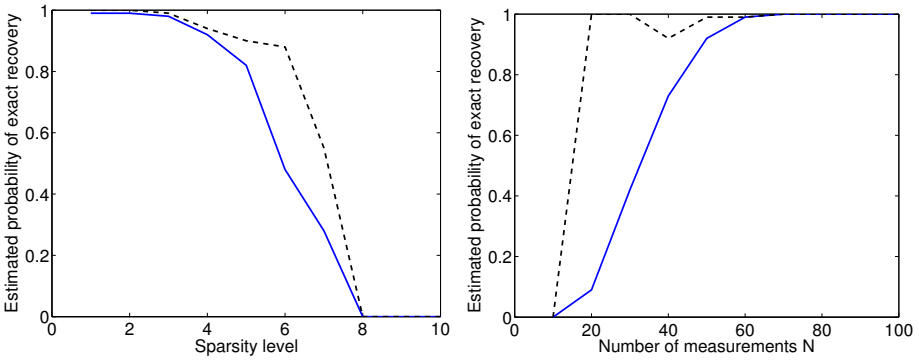


Figure 1: Estimated probability of exact recovery in the noiseless case for the convex relaxation (13) (plain line) and the greedy method applied to (12) (dashed line) versus the number of nonzeros (left) and the number of measurements (right).

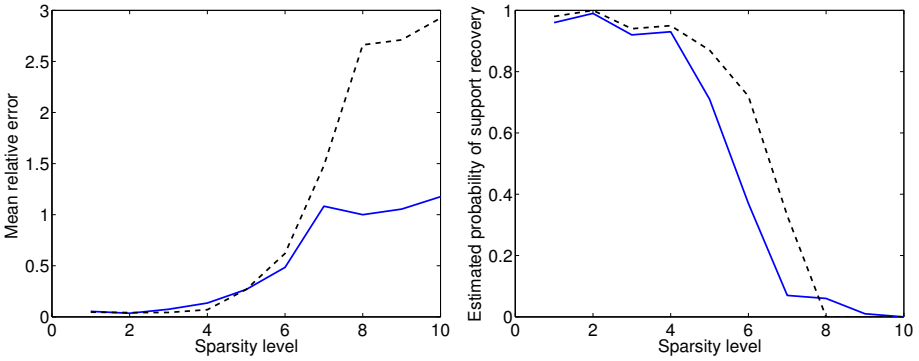


Figure 2: Mean relative error (left) and estimated probability of support recovery (right) in the noisy case for the convex relaxation (13) (plain line) and the greedy method applied to (12) (dashed line).

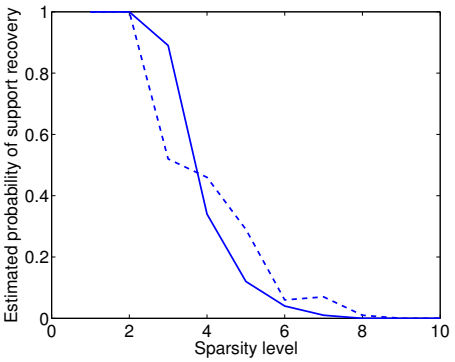


Figure 3: Estimated probability of support recovery from the magnitude of the Fourier transform when using (35) for $N = n$ (plain line) and when using (37) for $N = 2n$ (dashed line) versus the sparsity level.

**Support estimation from the power spectrum.** We now test if the convex method is robust to invariances, such as the ones discussed in Sect. 5.2, when estimating the support. In particular, we start with a setting in which $\boldsymbol{x} \in \mathbb{R}^n$, $N = n = 20$ and $\boldsymbol{q}_i^H$, $i = 1, \ldots, N$, are the rows of the $n$-point Fourier matrix. In this case, the autocorrelation cannot be computed and the convex formulation (35) is used to estimate the support. Then, we perform similar experiments but with oversampling ($N = 2n$), thus allowing for the computation of the autocorrelation and the use of (37). Results shown in Fig. 3 indicate that by using (35) we can recover the correct support for sufficiently sparse signals without oversampling, i.e., with as many measurements as unknowns, while using more information extracted from the autocorrelation of the signal only slightly helps to recover larger supports.

# 7 Conclusions

The paper proposed a new approach to phase retrieval of sparse signals. This approach is based on a group-sparse optimization formulation of the problem with linearized constraints. Exact and stable recovery results were shown for a convex relaxation of this formulation both in the real and complex case. Invariances to circular shifts and reflections that are common in phase retrieval problems were also discussed and a practical technique was given to prevent these from breaking the sparsity of the solution.

Future work will focus on deriving theoretical guarantees for the invariance issue. Another direction of research concerns the analysis of greedy algorithms for the group-sparse optimization problem, which proved as valuable as the convex relaxations in experiments for measurements without invariance. How to deal with invariances in these methods will also be investigated.

# References

[1] E. J Candès. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: invited lectures*, pages 1433–1452, 2006.

[2] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[3] E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[4] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted $\ell_1$ minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[5] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[7] M. Ehler, M. Fornasier, and J. Sigl. Quasi-linear compressed sensing. Technical report, 2013. http://www-m15.ma.tum.de/foswiki/pub/M15/Allgemeines/PublicationsEN/greedy_21.pdf.

[8] J. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[9] R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[10] R. Gonsalves. Phase retrieval from modulus data. *Journal of Optical Society of America*, 66(9):961–964, 1976.

[11] R. W. Harrison. Phase problem in crystallography. *Journal of the Optical Society of America A*, 10(5):1046–1055, 1993.

[12] K. Jaganathan, S. Oymak, and B. Hassibi. Recovery of sparse 1-d signals from the magnitudes of their Fourier transform. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1473–1477, 2012.

[13] D. Kohler and L. Mandel. Source reconstruction from the modulus of the correlation function: a practical approach to the phase problem of optical coherence theory. *Journal of the Optical Society of America*, 63(2):126–134, 1973.

[14] F. Lauer and H. Ohlsson. Finding sparse solutions of systems of polynomial equations via group-sparsity optimization. *arXiv preprint*, arXiv:1311.5871, 2013.

[15] S. Mukherjee and C. S. Seelamantula. An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 553–556, 2012.

[16] H. Ohlsson and Y. C. Eldar. On conditions for uniqueness in sparse phase retrieval. *CoRR*, abs/1308.5447, 2013.

[17] H. Ohlsson, A. Y. Yang, R. Dong, and S. Sastry. Nonlinear basis pursuit. *arXiv preprint arXiv:1304.5802*, 2013.

[18] H. Ohlsson, A. Y. Yang, R. Dong, M. Verhaegen, and S. Sastry. Quadratic basis pursuit. *arXiv preprint arXiv:1301.7002*, 2013.

[19] J. Ranieri, A. Chebira, Y. M. Lu, and M. Vetterli. Phase retrieval for sparse signals: Uniqueness conditions. *CoRR*, abs/1308.3058, 2013.

[20] Y. Shechtman, A. Beck, and Y. C. Eldar. GESPAR: Efficient phase retrieval of sparse signals. *arXiv preprint arXiv:1301.1018*, 2013.

[21] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics Express*, 19(16):14807–14822, 2011.

[22] I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *arXiv preprint arXiv:1206.0102*, 2012.

[23] A. Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963.

# A Lemmas

**Lemma 1.** *Let $\nu$ and $\nu^{-1}$ be defined as in Section 2. For all $\boldsymbol{v} \in \mathbb{R}^M$ such that $\nu^{-1}(\boldsymbol{v}) \neq \boldsymbol{0}$ and $\boldsymbol{v}' \in \mathbb{R}^M$, if $\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} \neq v'_{jj}$ for some $j \in \{1, \dots, n\}$, then $\boldsymbol{v} \neq \boldsymbol{v}'$.*

*Proof.* Assume that $\nu^{-1}(\boldsymbol{v}) \neq \boldsymbol{0}$ and let $i = \min_j j$, s.t. $v_{jj} > 0$. Then, $\forall j \in \{1, \dots, n\}$, we have

$$\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} = \left(\nu^{-1}(\boldsymbol{v})\right)_j^2 = \frac{v_{ji}^2}{v_{ii}} = v_{jj},$$

where the last equality follows from the definition of $\nu^{-1}$. Therefore, if $\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} = v_{jj} \neq v'_{jj}$ for some $j \in \{1, \dots, n\}$, we obtain $\boldsymbol{v} \neq \boldsymbol{v}'$. □

**Lemma 2.** *Let $\boldsymbol{A} = [\boldsymbol{A}_1, \dots, \boldsymbol{A}_M]$ be an $N \times M$ real matrix with mutual coherence $\mu(\boldsymbol{A})$ as in Definition 1. Let $\boldsymbol{W}$ be the $M \times M$-diagonal matrix of entries $w_i = \|\boldsymbol{A}_i\|_2$. Then, for all $\boldsymbol{\delta} \in Ker(\boldsymbol{A})$ and $i \in \{1, \dots, M\}$, the bound*

$$w_i^2 \delta_i^2 \leq \frac{\mu^2(\boldsymbol{A})}{1 + \mu^2(\boldsymbol{A})} \|\boldsymbol{W}\boldsymbol{\delta}\|_2^2 \tag{38}$$

*holds.*

*Proof.* See Lemma 2 in [14]. □

**Lemma 3.** *Let the complex Veronese map $\nu$ be as in Definition 2. Then, for all $\boldsymbol{q} \in \mathbb{C}^n$ and $\boldsymbol{x} \in \mathbb{C}^n$, the following equality holds:*

$$\boldsymbol{x}^H \boldsymbol{q}\boldsymbol{q}^H \boldsymbol{x} = 2\Re\left(\nu(\boldsymbol{q})^H \nu(\boldsymbol{x})\right) - \sum_{j=1}^{n} \left(\nu(\boldsymbol{q})\right)_{jj} \left(\nu(\boldsymbol{x})\right)_{jj}.$$

*Proof.*

$$\begin{aligned}
\boldsymbol{x}^H \boldsymbol{q}\boldsymbol{q}^H \boldsymbol{x} &= \left(\sum_{k=1}^{n} \overline{x_k} q_k\right) \left(\sum_{j=1}^{n} \overline{q_j} x_j\right) \\
&= \sum_{j=1}^{n} \overline{q_j} x_j \sum_{k=1}^{n} q_k \overline{x_k} = \sum_{j=1}^{n} \sum_{k=1}^{n} \overline{q_j} q_k x_j \overline{x_k} \\
&= \sum_{j=1}^{n} q_j \overline{q_j} x_j \overline{x_j} + \sum_{j=1}^{n} \sum_{k \neq j} \overline{q_j} q_k x_j \overline{x_k} \\
&= \sum_{j=1}^{n} q_j \overline{q_j} x_j \overline{x_j} + \sum_{j=1}^{n} \left(\sum_{k<j} \overline{q_j} q_k x_j \overline{x_k} + \sum_{k>j} \overline{q_j} q_k x_j \overline{x_k}\right) \\
&= \sum_{j=1}^{n} q_j \overline{q_j} x_j \overline{x_j} + \sum_{j=1}^{n} \sum_{k>j} \left(\overline{q_k} q_j x_k \overline{x_j} + \overline{q_j} q_k x_j \overline{x_k}\right) \\
&= \sum_{j=1}^{n} q_j \overline{q_j} x_j \overline{x_j} + \sum_{j=1}^{n} \sum_{k>j} \left(\overline{\overline{q_j} q_k x_j \overline{x_k}} + \overline{q_j} q_k x_j \overline{x_k}\right)
\end{aligned}$$

At this point, we use the fact that $z + \overline{z} = 2\Re(z)$, which yields

$$\begin{aligned}
\boldsymbol{x}^H \boldsymbol{q}\boldsymbol{q}^H \boldsymbol{x} &= \sum_{j=1}^{n} q_j \overline{q_j} x_j \overline{x_j} + 2 \sum_{j=1}^{n} \sum_{k>j} \Re\left(\overline{q_j} q_k x_j \overline{x_k}\right) \\
&= \sum_{j=1}^{n} \left(\nu(\boldsymbol{q})\right)_{jj} \left(\nu(\boldsymbol{x})\right)_{jj} + 2 \sum_{j=1}^{n} \sum_{k>j} \Re\left(\left(\overline{\nu(\boldsymbol{q})}\right)_{jk} \left(\nu(\boldsymbol{x})\right)_{jk}\right)
\end{aligned}$$

Since $(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj} = q_j\overline{q_j}x_j\overline{x_j} = |q_jx_j|^2$ is a real number, it can be introduced in the second sum as

$$\begin{aligned}
\boldsymbol{x}^H\boldsymbol{q}\boldsymbol{q}^H\boldsymbol{x} &= -\sum_{j=1}^{n}(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj} + 2\sum_{j=1}^{n}\sum_{k\geq j}\Re\left(\left(\overline{\nu(\boldsymbol{q})}\right)_{jk}(\nu(\boldsymbol{x}))_{jk}\right)\\
&= -\sum_{j=1}^{n}(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj} + 2\sum_{j=1}^{n}\sum_{k\geq j}\left[\Re\left(\overline{\nu(\boldsymbol{q})}\right)_{jk}\Re\left(\nu(\boldsymbol{x})\right)_{jk} - \Im\left(\overline{\nu(\boldsymbol{q})}\right)_{jk}\Im\left(\nu(\boldsymbol{x})\right)_{jk}\right]\\
&= -\sum_{j=1}^{n}(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj} + 2\left[\Re\left(\overline{\nu(\boldsymbol{q})}\right)^T\Re\left(\nu(\boldsymbol{x})\right) - \Im\left(\overline{\nu(\boldsymbol{q})}\right)^T\Im\left(\nu(\boldsymbol{x})\right)\right]\\
&= -\sum_{j=1}^{n}(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj} + 2\left[\Re\left(\nu(\boldsymbol{q})\right)^T\Re\left(\nu(\boldsymbol{x})\right) + \Im\left(\nu(\boldsymbol{q})\right)^T\Im\left(\nu(\boldsymbol{x})\right)\right]\\
&= 2\Re\left(\nu(\boldsymbol{q})^H\nu(\boldsymbol{x})\right) - \sum_{j=1}^{n}(\nu(\boldsymbol{q}))_{jj}\,(\nu(\boldsymbol{x}))_{jj}\,,
\end{aligned}$$

where the last equality is due to $\Re(\boldsymbol{a}^H\boldsymbol{b}) = \Re(\boldsymbol{a})^T\Re(\boldsymbol{b}) + \Im(\boldsymbol{a})^T\Im(\boldsymbol{b})$. □

**Lemma 4.** *Let $\nu^{-1}$ be as in Definition 3 and the matrices $\boldsymbol{W}_j$ as in Sect. 3.1. Then, for all $\boldsymbol{v}\in\mathbb{C}^M$,*

$$\|\nu^{-1}(\boldsymbol{v})\|_0 \leq \|\{\boldsymbol{W}_j\boldsymbol{v}\}_{j=1}^n\|_0. \tag{39}$$

*Proof.* According to Definition 3, $|\left(\nu^{-1}(\boldsymbol{v})\right)_j|^2 = \frac{|v_{ji}|^2}{v_{ii}} = v_{jj}$. On the other hand, $v_{jj}$ belongs to a single group of variables generated by the $\boldsymbol{W}_j$, i.e., $(\boldsymbol{W}_k)_{(jj)} \neq \boldsymbol{0} \Leftrightarrow k = j$. Thus, $\left(\nu^{-1}(\boldsymbol{v})\right)_j \neq 0 \Rightarrow v_{jj} > 0 \Rightarrow \boldsymbol{W}_j\boldsymbol{v} \neq \boldsymbol{0} \Rightarrow \|\nu^{-1}(\boldsymbol{v})\|_0 \leq \|\{\boldsymbol{W}_j\boldsymbol{v}\}_{j=1}^n\|_0$. □

However, note that the converse is not true: we can have $\|\nu^{-1}(\boldsymbol{v})\|_0 \neq \|\{\boldsymbol{W}_j\boldsymbol{v}\}_{j=1}^n\|_0$, e.g., when $v_{jj} = 0$, $j = 1,\ldots,n$, and $v_{ij} \neq 0$ for some $i$ and $j$.

**Lemma 5.** *Let $\nu$ and $\nu^{-1}$ be defined as in Definitions 2 and 3. For all $\boldsymbol{v}\in\mathbb{C}^M$ such that $\nu^{-1}(\boldsymbol{v}) \neq \boldsymbol{0}$ and $\boldsymbol{v}'\in\mathbb{C}^M$, if $\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} \neq v'_{jj}$ for some $j\in\{1,\ldots,n\}$, then $\boldsymbol{v}\neq\boldsymbol{v}'$.*

*Proof.* Assume that $\nu^{-1}(\boldsymbol{v}) \neq \boldsymbol{0}$ and let $i = \min_j j$, s.t. $\Re(v_{jj}) > 0$, $\Im(v_{jj}) = 0$. Then, $\forall j \in \{1,\ldots,n\}$, we have

$$\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} = \left(\nu^{-1}(\boldsymbol{v})\right)_j\left(\overline{\nu^{-1}(\boldsymbol{v})}\right)_j = \frac{\overline{v_{ij}}}{\sqrt{v_{ii}}}\frac{v_{ij}}{\sqrt{v_{ii}}} = \frac{|v_{ij}|^2}{v_{ii}} = v_{jj},$$

where the last equality follows from Definition 3. Therefore, if $\left(\nu(\nu^{-1}(\boldsymbol{v}))\right)_{jj} = v_{jj} \neq v'_{jj}$ for some $j\in\{1,\ldots,n\}$, we obtain $\boldsymbol{v}\neq\boldsymbol{v}'$. □

**Lemma 6.** *For all $\boldsymbol{\delta}\in\mathbb{C}^M$ such that $\Re(\boldsymbol{A}\boldsymbol{\delta}) = \boldsymbol{0}$, the inequality*

$$(w_i^R)^2\Re(\delta_i)^2 + (w_i^I)^2\Im(\delta_i)^2 \leq \frac{2\mu^2(\tilde{\boldsymbol{A}})}{1+\mu^2(\tilde{\boldsymbol{A}})}\|\boldsymbol{W}^R\Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I\Im(\boldsymbol{\delta})\|^2$$

*holds with $\tilde{\boldsymbol{A}} = [\Re(\boldsymbol{A}),\ -\Im(\boldsymbol{A})] \in \mathbb{R}^{N\times 2M}$.*

*Proof.* We rewrite the assumption as

$$\boldsymbol{0} = \Re(\boldsymbol{A}\boldsymbol{\delta}) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\delta}},$$

where $\tilde{\boldsymbol{\delta}} = [\Re(\boldsymbol{\delta})^T,\ \Im(\boldsymbol{\delta})^T]^T \in \mathbb{R}^{2M}$. Then, we can use Lemma 2 to bound the entries in $\tilde{\boldsymbol{\delta}}$ as

$$(w_i^R)^2\Re(\delta_i)^2 \leq \frac{\mu^2(\tilde{\boldsymbol{A}})}{1+\mu^2(\tilde{\boldsymbol{A}})}\left\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\delta}}\right\|_2^2,$$

with $\tilde{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W}^R & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}^I \end{pmatrix}$ a diagonal matrix of entries $(\tilde{\boldsymbol{W}})_{i,i} = \|\tilde{\boldsymbol{A}}_i\|_2$, and

$$(w_i^I)^2 \Im(\delta_i)^2 \leq \frac{\mu^2(\tilde{\boldsymbol{A}})}{1+\mu^2(\tilde{\boldsymbol{A}})} \left\| \tilde{\boldsymbol{W}} \tilde{\boldsymbol{\delta}} \right\|_2^2.$$

These inequalities lead to

$$(w_i^R)^2 \Re(\delta_i)^2 + (w_i^I)^2 \Im(\delta_i)^2 \leq \frac{2\mu^2(\tilde{\boldsymbol{A}})}{1+\mu^2(\tilde{\boldsymbol{A}})} \|\tilde{\boldsymbol{W}} \tilde{\boldsymbol{\delta}}\|_2^2,$$

where

$$\|\tilde{\boldsymbol{W}} \tilde{\boldsymbol{\delta}}\|_2^2 = \|\boldsymbol{W}^R \Re(\boldsymbol{\delta})\|_2^2 + \|\boldsymbol{W}^I \Im(\boldsymbol{\delta})\|_2^2 = \|\boldsymbol{W}^R \Re(\boldsymbol{\delta}) + \mathrm{i}\boldsymbol{W}^I \Im(\boldsymbol{\delta})\|^2.$$

□

**Lemma 7.** *For all $\boldsymbol{\delta} \in \mathbb{R}^M$ such that $\Re(\boldsymbol{A})\boldsymbol{\delta} = \boldsymbol{0}$, the inequality*

$$(w_i^R)^2 \delta_i^2 \leq \frac{\mu^2\left(\Re(\boldsymbol{A})\right)}{1+\mu^2\left(\Re(\boldsymbol{A})\right)} \|\boldsymbol{W}^R \boldsymbol{\delta}\|_2^2,$$

*where $\boldsymbol{W}^R$ is a diagonal matrix of entries $(\boldsymbol{W}^R)_{i,i} = w_i^R = \|\Re(\boldsymbol{A}_i)\|_2$.*

*Proof.* This is a direct consequence of Lemma 2 applied to $\Re(\boldsymbol{A})$ and $\boldsymbol{\delta} \in Ker(\Re(\boldsymbol{A}))$. □

# B Proofs

## B.1 Proof of Theorem 2

This proof is similar to the one of Theorem 3 in [14].

*Proof.* The vector $\boldsymbol{v}_0$ is the unique solution to (6) if the inequality

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j \boldsymbol{W}(\boldsymbol{v}_0 + \boldsymbol{\delta})\|_2 > \sum_{j=1}^{n} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{v}_0\|_2$$

holds for all $\boldsymbol{\delta} \neq \boldsymbol{0}$ satisfying $\boldsymbol{A}\boldsymbol{\delta} = 0$. The inequality above can be rewritten as

$$\sum_{j \in I_0} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2 + \sum_{j \notin I_0} \|\boldsymbol{W}_j \boldsymbol{W}(\boldsymbol{v}_0 + \boldsymbol{\delta})\|_2 - \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{v}_0\|_2 > 0,$$

where $I_0 = \{j \in \{1, \ldots, n\} : \boldsymbol{W}_j \boldsymbol{W} \boldsymbol{v}_0 = \boldsymbol{0}\}$. By the triangle inequality, $\|\boldsymbol{a} + \boldsymbol{b}\|_2 - \|\boldsymbol{a}\|_2 \geq -\|\boldsymbol{b}\|_2$, this condition is met if

$$\sum_{j \in I_0} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2 - \sum_{j \notin I_0} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2 > 0$$

or

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2 - 2 \sum_{j \notin I_0} \|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2 > 0. \tag{40}$$

By defining $G_j$ as the set of indexes corresponding to nonzero columns of $\boldsymbol{W}_j$, Lemma 2 yields

$$\|\boldsymbol{W}_j \boldsymbol{W} \boldsymbol{\delta}\|_2^2 = \sum_{i \in G_j} w_i^2 \delta_i^2 \leq n \frac{\mu^2(\boldsymbol{A})}{1+\mu^2(\boldsymbol{A})} \|\boldsymbol{W} \boldsymbol{\delta}\|_2^2,$$

Due to the fact that $\bigcup_{k \in \{1,\ldots,n\}} G_k = \{1, \ldots, M\}$, we also have

$$\|\boldsymbol{W} \boldsymbol{\delta}\|_2^2 = \sum_{i=1}^{M} w_i^2 \delta_i^2 \leq \sum_{k=1}^{n} \sum_{i \in G_k} w_i^2 \delta_i^2 = \sum_{k=1}^{n} \|\boldsymbol{W}_k \boldsymbol{W} \boldsymbol{\delta}\|_2^2 \leq \left( \sum_{k=1}^{n} \|\boldsymbol{W}_k \boldsymbol{W} \boldsymbol{\delta}\|_2 \right)^2,$$

which then leads to

$$\|\boldsymbol{W}_j\boldsymbol{W}\boldsymbol{\delta}\|_2^2 \leq n\frac{\mu^2(\boldsymbol{A})}{1+\mu^2(\boldsymbol{A})}\left(\sum_{k=1}^{n}\|\boldsymbol{W}_k\boldsymbol{W}\boldsymbol{\delta}\|_2\right)^2.$$

Introducing this result in (40) gives the condition

$$\sum_{j=1}^{n}\|\boldsymbol{W}_j\boldsymbol{W}\boldsymbol{\delta}\|_2 - 2(n-|I_0|)\frac{\mu(\boldsymbol{A})\sqrt{n}}{\sqrt{1+\mu^2(\boldsymbol{A})}}\sum_{k=1}^{n}\|\boldsymbol{W}_k\boldsymbol{W}\boldsymbol{\delta}\|_2 > 0.$$

Finally, given that $|I_0| = n - \|\{\boldsymbol{W}_j\boldsymbol{v}_0\}_{j=1}^n\|_0 = n - \|\boldsymbol{x}_0\|_0$, this yields

$$\sum_{j=1}^{n}\|\boldsymbol{W}_j\boldsymbol{W}\boldsymbol{\delta}\|_2 - 2\|\boldsymbol{x}_0\|_0\frac{\mu(\boldsymbol{A})\sqrt{n}}{\sqrt{1+\mu^2(\boldsymbol{A})}}\sum_{k=1}^{n}\|\boldsymbol{W}_k\boldsymbol{W}\boldsymbol{\delta}\|_2 > 0.$$

or, after rearranging the terms,

$$\|\boldsymbol{x}_0\|_0 < \frac{\sqrt{1+\mu^2(\boldsymbol{A})}}{2\mu(\boldsymbol{A})\sqrt{n}},$$

which can be rewritten as in the statement of the Theorem. □

## B.2 Proof of Theorem 7

This proof is very similar to the ones of Theorems 2 and 4.

*Proof.* The vector $\boldsymbol{v}_0$ is the unique solution to (31) if the inequality

$$\sum_{j=1}^{n}\|\boldsymbol{W}_j^R(\boldsymbol{v}_0+\boldsymbol{\delta})\|_2 > \sum_{j=1}^{n}\|\boldsymbol{W}_j^R\boldsymbol{v}_0\|_2$$

holds for all $\boldsymbol{\delta}\in\mathbb{R}^M$ such that $\Re(\boldsymbol{A})(\boldsymbol{v}_0+\boldsymbol{\delta}) = \boldsymbol{y}$, which implies the constraint $\Re(\boldsymbol{A})\,\boldsymbol{\delta} = \boldsymbol{0}$ on $\boldsymbol{\delta}$. The inequality above can be rewritten as

$$\sum_{j\in I_0}\|\boldsymbol{W}_j^R\boldsymbol{\delta}\|_2 + \sum_{j\notin I_0}\|\boldsymbol{W}_j^R(\boldsymbol{v}_0+\boldsymbol{\delta})\|_2 - \|\boldsymbol{W}_j^R\boldsymbol{v}_0\|_2 > 0,$$

where $I_0 = \{j\in\{1,\dots,n\} : \boldsymbol{W}_j^R\boldsymbol{v}_0 = \boldsymbol{0}\}$. By the triangle inequality, $\|\boldsymbol{a}+\boldsymbol{b}\| - \|\boldsymbol{a}\| \geq -\|\boldsymbol{b}\|$ with $\boldsymbol{a} = \boldsymbol{W}_j^R\boldsymbol{v}_0$, this condition is met if

$$\sum_{j=1}^{n}\|\boldsymbol{W}_j^R\boldsymbol{\delta}\|_2 - 2\sum_{j\notin I_0}\|\boldsymbol{W}_j^R\boldsymbol{\delta}\|_2 > 0. \tag{41}$$

By defining $G_j$ as the set of indexes corresponding to nonzero columns of $\boldsymbol{W}_j$, Lemma 7 yields

$$\|\boldsymbol{W}_j^R\boldsymbol{\delta}\|_2^2 = \sum_{i\in G_j}(w_i^R)^2\delta_i^2 \leq n\frac{\mu^2(\Re(\boldsymbol{A}))}{1+\mu^2(\Re(\boldsymbol{A}))}\|\boldsymbol{W}^R\boldsymbol{\delta}\|_2^2.$$

Due to the fact that $\bigcup_{k\in\{1,\dots,n\}}G_k = \{1,\dots,M\}$, we also have

$$\|\boldsymbol{W}^R\boldsymbol{\delta}\|_2^2 = \sum_{i=1}^{M}(w_i^R)^2\delta_i^2 \leq \sum_{k=1}^{n}\sum_{i\in G_k}(w_i^R)^2\delta_i^2 = \sum_{k=1}^{n}\|\boldsymbol{W}_k^R\boldsymbol{\delta}\|_2^2 \leq \left(\sum_{k=1}^{n}\|\boldsymbol{W}_k^R\boldsymbol{\delta}\|_2\right)^2,$$

which then leads to

$$\|\boldsymbol{W}_j^R\boldsymbol{\delta}\|_2^2 \leq n\frac{\mu^2(\Re(\boldsymbol{A}))}{1+\mu^2(\Re(\boldsymbol{A}))}\left(\sum_{k=1}^{n}\|\boldsymbol{W}_k^R\boldsymbol{\delta}\|_2\right)^2.$$

Introducing this result in (41) gives the condition

$$\sum_{j=1}^{n} \|\boldsymbol{W}_j^R \boldsymbol{\delta}\|_2 - 2(n - |I_0|) \frac{\mu(\Re(\boldsymbol{A}))\sqrt{n}}{\sqrt{1+\mu^2(\Re(\boldsymbol{A}))}} \sum_{k=1}^{n} \|\boldsymbol{W}_k^R \boldsymbol{\delta}\|_2 > 0.$$

Finally, given that $|I_0| = n - \|\boldsymbol{x}_0\|_0$, this yields , for $\boldsymbol{\delta} \neq \boldsymbol{0}$,

$$\|\boldsymbol{x}_0\|_0 < \frac{\sqrt{1+\mu^2(\Re(\boldsymbol{A}))}}{2\mu(\Re(\boldsymbol{A}))\sqrt{n}},$$

which can be rewritten as in the statement of the Theorem. □

## B.3 Proof of Proposition 1

*Proof.* To prove Statement 1, note that the operation $\boldsymbol{x}_2 = \text{shift}(\text{reflection}(\boldsymbol{x}_1), 1)$ is equivalent to

$$\boldsymbol{x}_2 = \begin{bmatrix} x_{11} \\ \text{reflection}(\tilde{\boldsymbol{x}}_1) \end{bmatrix}, \quad \tilde{\boldsymbol{x}}_1 = [x_{12}, x_{13}, \ldots, x_{1n}]^T,$$

where the reflection is centered on $\tilde{x}_{1(n/2)} = x_{1(1+n/2)}$. Thus,

$$\boldsymbol{x}_2 = [x_{11}, x_{1n}, x_{1(n-1)}, \ldots, x_{12}]^T$$

and

$$\sum_{i=2}^{n/2} |x_{2i}|^2 = \sum_{i=0}^{n/2-2} |x_{1(n-i)}|^2 = \sum_{i=2+\frac{n}{2}}^{n} |x_{1i}|^2 > \sum_{i=2}^{n/2} |x_{1i}|^2 = \sum_{i=0}^{n/2-2} |x_{2(n-i)}|^2 = \sum_{i=2+\frac{n}{2}}^{n} |x_{2i}|^2,$$

where the inequality holds whenever this operation is performed in (34) (i.e., if this is not the case, then $\boldsymbol{x}_2 = \boldsymbol{x}_1$). Therefore, for all $\boldsymbol{x}$, we obtain a vector $\boldsymbol{x}_2 = \varphi(\boldsymbol{x})$ such that $\sum_{i=2}^{n/2} |x_{2i}|^2 \geq \sum_{i=2+\frac{n}{2}}^{n} |x_{2i}|^2$. Since $\arg\max_{i\in\{1,\ldots,n\}} |x_{2i}| = 1$ and $\text{shift}(\boldsymbol{x}_2, 1 - \arg\max_{i\in\{1,\ldots,n\}} |x_{2i}|) = \boldsymbol{x}_2$, this implies that $\varphi(\boldsymbol{x}_2) = \boldsymbol{x}_2$, i.e., $\varphi(\varphi(\boldsymbol{x})) = \varphi(\boldsymbol{x})$.

Statement 2 is easily seen from the fact that $\boldsymbol{x}_1$ does not change when $\boldsymbol{x}$ is shifted.

To prove Statement 3, let $\boldsymbol{x}' = \text{reflection}(\boldsymbol{x}) = [x_n, x_{n-1}, \ldots, x_1]^T$, $k = \arg\max_{i\in\{1,\ldots,n\}} |x_i|$, and $k' = \arg\max_{i\in\{1,\ldots,n\}} |x'_i|$. Then, $k' = n - k + 1$ and

$$\begin{aligned} \boldsymbol{x}'_1 &= \text{shift}(\boldsymbol{x}', 1 - k') = \text{shift}(\boldsymbol{x}', k - n) = \text{shift}(\boldsymbol{x}', k) \\ &= [x'_{n-k+1}, x'_{n-k+2}, \ldots, x'_n, x'_1, x'_2, \ldots, x'_{n-k}]^T \\ &= [x_k, x_{k-1}, \ldots, x_1, x_n, x_{n-1}, \ldots, x_{k+1}]^T. \end{aligned}$$

Define $\boldsymbol{x}_1 = \text{shift}(\boldsymbol{x}, 1 - k) = [x_k, x_{k+1}, \ldots, x_n, x_1, x_2, \ldots, x_{k-1}]^T$.

If $k < n/2$, we have

$$\sum_{i=2}^{n/2} |x'_{1i}|^2 = \sum_{i=1}^{k-1} |x_i|^2 + \sum_{i=1}^{\frac{n}{2}-k} |x_{n-i+1}|^2 = \sum_{i=n+2-k}^{n} |x_{1i}|^2 + \sum_{i=n/2+2}^{n-k+1} |x_{1i}|^2 = \sum_{i=n/2+2}^{n} |x_{1i}|^2,$$

and otherwise, if $k \geq n/2$, we obtain

$$\sum_{i=2}^{n/2} |x'_{1i}|^2 = \sum_{i=k-\frac{n}{2}+1}^{k-1} |x_i|^2 = \sum_{i=n/2+2}^{n} |x_{1i}|^2.$$

On the other hand, if $k \leq n/2$, then

$$\sum_{i=2+\frac{n}{2}}^{n} |x'_{1i}|^2 = \sum_{i=k+1}^{\frac{n}{2}+k-1} |x_i|^2 = \sum_{i=2}^{n/2} |x_{1i}|^2,$$

while if $k > n/2$:

$$\sum_{i=2+\frac{n}{2}}^{n} |x'_{1i}|^2 = \sum_{i=k+1}^{n} |x_i|^2 + \sum_{i=1}^{k-\frac{n}{2}-1} |x_i|^2 = \sum_{i=2}^{n-k+1} |x_{1i}|^2 + \sum_{i=n-k+2}^{n/2} |x_{1i}|^2 = \sum_{i=2}^{n/2} |x_{1i}|^2.$$

Therefore,

$$\sum_{i=2}^{n/2} |x_{1i}|^2 > \sum_{i=2+\frac{n}{2}}^{n} |x_{1i}|^2 \quad \Leftrightarrow \quad \sum_{i=2}^{n/2} |x'_{1i}|^2 < \sum_{i=2+\frac{n}{2}}^{n} |x'_{1i}|^2$$

and $\varphi$ applies a reflection to $\boldsymbol{x}$ if and only if it does not apply one to $\boldsymbol{x}'$ (the inequalities above are strict by assumption). Thus, in the case $\boldsymbol{x}$ is reflected by $\varphi$, we have

$$\varphi(\boldsymbol{x}') = \text{shift}(\text{reflection}(\boldsymbol{x}), k) = \text{reflection}(\text{shift}(\boldsymbol{x}, n-k))$$

and

$$\begin{aligned}
\varphi(\boldsymbol{x}) &= \text{shift}(\text{reflection}(\text{shift}(\boldsymbol{x}, 1-k)), 1) \\
&= \text{reflection}(\text{shift}(\text{shift}(\boldsymbol{x}, 1-k), n-1)) \\
&= \text{reflection}(\text{shift}(\boldsymbol{x}, n-k)) \\
&= \varphi(\boldsymbol{x}'),
\end{aligned}$$

while in the case $\boldsymbol{x}$ is not reflected by $\varphi$, we have

$$\varphi(\boldsymbol{x}) = \text{shift}(\boldsymbol{x}, 1-k)$$

and

$$\begin{aligned}
\varphi(\boldsymbol{x}') &= \text{shift}(\text{reflection}(\text{shift}(\text{reflection}(\boldsymbol{x}), k)), 1) \\
&= \text{reflection}(\text{shift}(\text{shift}(\text{reflection}(\boldsymbol{x}), k), n-1)) \\
&= \text{reflection}(\text{shift}(\text{reflection}(\boldsymbol{x}), n+k-1)) \\
&= \text{reflection}(\text{reflection}(\text{shift}(\boldsymbol{x}, -k+1))) \\
&= \text{shift}(\boldsymbol{x}, 1-k) \\
&= \varphi(\boldsymbol{x}).
\end{aligned}$$

□