Advances in Applied Mathematics 93 (2018) 121–141

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Phase retrieval from the magnitudes of affine linear measurements

Bing Gao[a], Qiyu Sun[b,1], Yang Wang[c,2], Zhiqiang Xu[a,d,*,3]

[a] LSEC, Inst. Comp. Math., Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
[c] Department of Mathematics, the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[d] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 27 August 2016
Received in revised form 21 September 2017
Accepted 25 September 2017
Available online 5 October 2017

MSC:
primary 42C15

Keywords:
Phase retrieval
Frame
Sparse signals
Algebraic variety

ABSTRACT

In this paper, we consider the affine phase retrieval problem in which one aims to recover a signal from the magnitudes of affine measurements. Let $\{\mathbf{a}_j\}_{j=1}^m \subset \mathbb{H}^d$ and $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{H}^m$, where $\mathbb{H} = \mathbb{R}$ or $\mathbb{C}$. We say $\{\mathbf{a}_j\}_{j=1}^m$ and $\mathbf{b}$ are affine phase retrievable for $\mathbb{H}^d$ if any $\mathbf{x} \in \mathbb{H}^d$ can be recovered from the magnitudes of the affine measurements $\{|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j|, 1 \le j \le m\}$. We develop general framework for affine phase retrieval and prove necessary and sufficient conditions for $\{\mathbf{a}_j\}_{j=1}^m$ and $\mathbf{b}$ to be affine phase retrievable. We establish results on minimal measurements and generic measurements for affine phase retrieval as well as on sparse affine phase retrieval. In particular, we also highlight some notable differences between affine phase retrieval and the standard phase retrieval in which one aims to recover a signal $\mathbf{x}$ from the magnitudes of its linear measurements. In standard phase retrieval, one can only recover $\mathbf{x}$ up to a unimodular constant,

* Corresponding author.
E-mail addresses: gaobing@lsec.cc.ac.cn (B. Gao), qiyu.sun@ucf.edu (Q. Sun), yangwang@ust.hk (Y. Wang), xuzq@lsec.cc.ac.cn (Z. Xu).
[1] Qiyu Sun is partially supported by National Science Foundation (DMS-1412413).
[2] Yang Wang was supported in part by the Hong Kong Research Grant Council grant 16306415 and 16317416 as well as the AFOSR grant FA9550-12-1-0455.
[3] Zhiqiang Xu was supported by NSFC grant (11422113, 91630203, 11331012) and by National Basic Research Program of China (973 Program 2015CB856000).

https://doi.org/10.1016/j.aam.2017.09.004
0196-8858/

while affine phase retrieval removes this ambiguity. We prove that unlike standard phase retrieval, the affine phase retrievable measurements $\{\mathbf{a}_j\}_{j=1}^m$ and $\mathbf{b}$ do not form an open set in $\mathbb{H}^{m\times d} \times \mathbb{H}^m$. Also in the complex setting, the standard phase retrieval requires $4d - O(\log_2 d)$ measurements, while the affine phase retrieval only needs $m = 3d$ measurements.

1. Introduction

1.1. Phase retrieval

Phase retrieval is an active topic of research in recent years as it arises in many different areas of studies (see [2,5,6,9–12,15] and the references therein). For a vector (signal) $\mathbf{x} \in \mathbb{H}^d$, where $\mathbb{H} = \mathbb{R}$ or $\mathbb{C}$, the aim of phase retrieval is to recover $\mathbf{x}$ from $|\langle \mathbf{a}_j, \mathbf{x}\rangle|$, $j = 1, \ldots, m$, where $\mathbf{a}_j \in \mathbb{H}^d$ and we usually refer to $\{\mathbf{a}_j\}_{j=1}^m$ as the *measurement vectors*. Since for any unimodular $c \in \mathbb{H}$, we have $|\langle \mathbf{a}_j, \mathbf{x}\rangle| = |\langle \mathbf{a}_j, c\mathbf{x}\rangle|$, the best outcome phase retrieval can achieve is to recover $\mathbf{x}$ up to a unimodular constant.

We briefly overview some of the results in phase retrieval and introduce some notations. For the set of measurement vectors $\{\mathbf{a}_j\}_{j=1}^m$, we set $\mathbf{A} := (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{H}^{m\times d}$ which we shall refer to as the *measurement matrix*. We shall in general identify the set of measurement vectors $\{\mathbf{a}_j\}_{j=1}^m$ with the corresponding measurement matrix $\mathbf{A}$, and often use the two terms interchangeably whenever there is no confusion. Define the map $\mathbf{M}_\mathbf{A} : \mathbb{H}^d \to \mathbb{R}^m_{\geq 0}$ by

$$\mathbf{M}_\mathbf{A}(\mathbf{x}) \;=\; (|\langle \mathbf{a}_1, \mathbf{x}\rangle|, \ldots, |\langle \mathbf{a}_m, \mathbf{x}\rangle|).$$

We say $\mathbf{A}$ is *phase retrievable* for $\mathbb{H}^d$ if $\mathbf{M}_\mathbf{A}(\mathbf{x}) = \mathbf{M}_\mathbf{A}(\mathbf{y})$ implies $\mathbf{x} \in \{c\mathbf{y} : c \in \mathbb{H}, |c| = 1\}$. There have been extensive studies of phase retrieval from various different angles. For example many efficient algorithms to recover $\mathbf{x}$ from $\mathbf{M}_\mathbf{A}(\mathbf{x})$ have been developed, see [7–9,17] and their references. One of the fundamental problems on the theoretical side of phase retrieval is the following question: *How many vectors in the measurement matrix* $\mathbf{A}$ *are needed so that* $\mathbf{A}$ *is phase retrievable?* It is shown in [2] that for $\mathbf{A}$ to be phase retrievable for $\mathbb{R}^d$, it is necessary and sufficient that $m \geq 2d - 1$.

In the complex case $\mathbb{H} = \mathbb{C}$, the same question becomes much more challenging, however. The minimality question remains open today. Balan, Casazza and Edidin [2] first show that $\mathbf{A}$ is phase retrievable if it contains $m \geq 4d - 2$ generic vectors in $\mathbb{C}^d$. Bodmann and Hammen [5] show that $m = 4d - 4$ measurement vectors are possible for phase retrieval through construction (see also Fickus, Mixon, Nelson and Wang [12]). Bandeira, Cahill, Mixon and Nelson [4] conjecture that (a) $m \geq 4d - 4$ is necessary for $\mathbf{A}$ to be phase retrievable and, (b) $\mathbf{A}$ with $m \geq 4d - 4$ generic measurement vectors is phase retrievable. Part (b) of the conjecture is proved by Conca, Edidin, Hering and Vinzant [11].

They also confirm part (a) for the case where d is in the form of 2^k+1, $k \in \mathbb{Z}_+$. However, Vinzant in [18] presents a phase retrievable $\mathbf{A}$ for $\mathbb{C}^4$ with $m = 11 = 4d-5 < 4d-4$ measurement vectors, thus disproving the conjecture. The measurement vectors in the counterexample are obtained using Gröbner basis and algebraic computation.

1.2. Phase retrieval from magnitudes of affine linear measurements

Here we consider the affine phase retrieval problem, where instead of being given the magnitudes of linear measurements, we are given the magnitudes of affine linear measurements that include shifts. More precisely, instead of recovering $\mathbf{x}$ from $\{|\langle \mathbf{a}_j, \mathbf{x}\rangle|\}_{j=1}^m$, we consider recovering $\mathbf{x}$ from the absolute values of the affine linear measurements

$$|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j|, \quad j = 1, \ldots, m,$$

where $\mathbf{a}_j \in \mathbb{H}^d$, $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{H}^m$. Unlike in the classical phase retrieval, where $\mathbf{x}$ can only be recovered up to a unimodular constant, we will show that one can recover $\mathbf{x}$ *exactly* from $(|\langle \mathbf{a}_1, \mathbf{x}\rangle + b_1|, \ldots, |\langle \mathbf{a}_m, \mathbf{x}\rangle + b_m|)$ if the vectors $\mathbf{a}_j$ and shifts b_j are properly chosen.

Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{H}^{m\times d}$ and $\mathbf{b} \in \mathbb{H}^m$. Define the map $\mathbf{M}_{\mathbf{A},\mathbf{b}} : \mathbb{H}^d \to \mathbb{R}^m_{\geq 0}$ by

$$\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = (|\langle \mathbf{a}_1, \mathbf{x}\rangle + b_1|, \ldots, |\langle \mathbf{a}_m, \mathbf{x}\rangle + b_m|)\,. \tag{1.1}$$

We say the pair $(\mathbf{A}, \mathbf{b})$ (which can also be viewed as a matrix in $\mathbb{H}^{m\times(d+1)}$) is *affine phase retrievable* for $\mathbb{H}^d$, or simply *phase retrievable* whenever there is no confusion, if $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ is injective on $\mathbb{H}^d$. Note that sometimes it is more convenient to consider the map

$$\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) := (|\langle \mathbf{a}_1, \mathbf{x}\rangle + b_1|^2, \ldots, |\langle \mathbf{a}_m, \mathbf{x}\rangle + b_m|^2). \tag{1.2}$$

Clearly $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable if and only if $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ is injective on $\mathbb{H}^d$. The goal of this paper is to develop a framework of affine phase retrieval.

There are several motivations for studying affine phase retrieval. It arises naturally in holography, see [16]. It could also arise in other phase retrieval applications, such as reconstruction of signals in a shift-invariant space from their phaseless samples [10], where some entries of $\mathbf{x}$ might be known in advance. In such scenarios, assume that the object signal is $\mathbf{y} \in \mathbb{H}^{d+k}$ and the first k entries of $\mathbf{y}$ are known. We can write $\mathbf{y} = (y_1, \ldots, y_k, \mathbf{x})$, where $y_1, \ldots, y_k$ are known and $\mathbf{x} \in \mathbb{H}^d$. Suppose that $\tilde{\mathbf{a}}_j = (a_{j1}, \ldots, a_{jk}, \mathbf{a}_j) \in \mathbb{H}^{d+k}$, $j = 1, \ldots, m$ are the measurement vectors. Then

$$|\langle \tilde{\mathbf{a}}_j, \mathbf{y}\rangle| \;=\; |\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j|,$$

where $b_j := a_{j1}y_1 + \cdots + a_{jk}y_k$. So if $(y_1, \ldots, y_k)$ is a nonzero vector, we can take advantage of knowing the first k entries and reduce the standard phase retrieval in $\mathbb{H}^{d+k}$ to affine phase retrieval in $\mathbb{H}^d$.

1.3. Our contribution

This paper considers affine phase retrieval for both real and complex signals. In Section 2, we consider the real case $\mathbb{H} = \mathbb{R}$ and prove several necessary and sufficient conditions under which $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ is injective on $\mathbb{R}^d$. For an index set $T \subset \{1, \ldots, m\}$, we use $\mathbf{A}_T$ to denote the sub-matrix $\mathbf{A}_T := (\mathbf{a}_j : j \in T)^\top$ of $\mathbf{A}$. Let $\#T$ denote the cardinality of T, $\mathrm{span}(\mathbf{A}_T) \subset \mathbb{R}^{\#T}$ denote the subspace spanned by the column vectors of $\mathbf{A}_T$. In particularly, we show that $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable for $\mathbb{R}^d$ if and only if $\mathrm{span}\{\mathbf{a}_j : j \in S^c\} = \mathbb{R}^d$ for any index set $S \subset \{1, \ldots, m\}$ satisfying $\mathbf{b}_S \in \mathrm{span}(\mathbf{A}_S)$. Based on this result, we prove that the measurement vectors set $\mathbf{A}$ must have at least $m \geq 2d$ elements for $(\mathbf{A}, \mathbf{b})$ to be affine phase retrievable. Furthermore, we prove any generic $\mathbf{A} \in \mathbb{R}^{m\times d}$ and $\mathbf{b} \in \mathbb{R}^m$, where $m \geq 2d$ will be affine phase retrievable. The recovery of sparse signals from phaseless measurements also attracts much attention recently [13,19]. In this section, we consider the real affine phase retrieval for sparse vectors.

We turn to the complex case $\mathbb{H} = \mathbb{C}$ in Section 3. First we establish equivalent necessary and sufficient conditions for $(\mathbf{A}, \mathbf{b})$ to be affine phase retrievable for $\mathbb{C}^d$. Using these conditions, we show that $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is *not* affine phase retrievable for $\mathbb{C}^d$ if $m < 3d$. The result is sharp as we also construct an affine phase retrievable $(\mathbf{A}, \mathbf{b})$ for $\mathbb{C}^d$ with $m = 3d$. This result shows that the nature of affine phase retrieval can be quite different from that of the standard phase retrieval in the complex setting, where it is known that $4d - O(\log_2 d)$ measurements are needed for phase retrieval [15,20].

Note that for $j = 1, \ldots, m$ we have

$$|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j| \;=\; |\langle \tilde{\mathbf{a}}_j, \tilde{\mathbf{x}}\rangle|, \quad \text{where} \;\; \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}, \tilde{\mathbf{a}}_j = \begin{pmatrix} \mathbf{a}_j \\ b_j \end{pmatrix}. \tag{1.3}$$

It shows that affine phase retrieval for $\mathbf{x}$ can be reduced to the classical phase retrieval for $\tilde{\mathbf{x}} \in \mathbb{C}^{d+1}$ from $|\langle \tilde{\mathbf{a}}_j, \tilde{\mathbf{x}}\rangle|$, $j = 1, \ldots, m$. Because the last entry of $\tilde{\mathbf{x}}$ is 1, it allows us to recover $\mathbf{x}$ without the unimodular constant ambiguity. Observe also from [11] that $4(d + 1) - 4 = 4d$ generic measurements are enough to recover $\tilde{\mathbf{x}}$ up to a unimodular constant, and hence they are also enough to recover $\mathbf{x}$. In Section 3, we prove the stronger result that a generic $(\mathbf{A}, \mathbf{b})$ in $\mathbb{C}^{m\times(d+1)}$ with $m \geq 4d - 1$ is affine phase retrievable. We furthermore consider the complex affine phase retrieval for sparse signals in this section.

The classical phase retrieval has the property that the set of phase retrievable $\mathbf{A} \in \mathbb{H}^{m\times d}$ is an open set, and hence the phase retrievable property is stable under small perturbations [1,3]. Surprisingly, viewing $(\mathbf{A}, \mathbf{b})$ as an element in $\mathbb{H}^{m\times(d+1)}$, we prove that the set of affine phase retrievable $(\mathbf{A}, \mathbf{b})$ is *not* an open set.

As far as stability of affine phase retrieval is concerned, we prove several new results in Section 4. For the standard phase retrieval, one uses $\min_{|\alpha|=1} \|\mathbf{x} - \alpha\mathbf{y}\|$ to measure the distance between $\mathbf{x}$ and $\mathbf{y}$. The robustness of phase retrieval is established via the lower bound of the following bi-Lipschitz type inequalities for any phase retrievable $\mathbf{A}$,

$$c \min_{\alpha\in\mathbb{C},|\alpha|=1} \|\mathbf{x} - \alpha\mathbf{y}\| \;\leq\; \|\mathbf{M}_{\mathbf{A}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A}}(\mathbf{y})\| \;\leq\; C \min_{\alpha\in\mathbb{C},|\alpha|=1} \|\mathbf{x} - \alpha\mathbf{y}\|, \tag{1.4}$$

where $c, C > 0$ depend only on $\mathbf{A}$ [6]. Throughout this paper, we use $\|\cdot\|$ to denote the ℓ^2-norm. More explicit estimate of the constant c was given in [3]. For the affine phase retrieval, we use $\|\mathbf{x}-\mathbf{y}\|$ to measure the distance between $\mathbf{x}$ and $\mathbf{y}$ because it is possible to recover $\mathbf{x}$ exactly in the affine phase retrieval. For the affine phase retrieval, we show that both $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ and $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ are bi-Lipschitz continuous on any compact sets, but are not bi-Lipschitz on $\mathbb{H}^d$.

2. Affine phase retrieval for real signals

We consider affine phase retrieval of real signals in this section. Several equivalent conditions for affine phase retrieval are established. We also study affine phase retrieval for sparse signals. In particular we answer the minimality question, namely what is the smallest number of measurements needed for affine phase retrievability for $\mathbb{R}^d$.

2.1. Real affine phase retrieval

Let $T \subset \{1, 2, \ldots, m\}$. We first recall that for the measurement matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{R}^{m\times d}$, we use $\mathbf{A}_T$ to denote the submatrix of $\mathbf{A}$ consisting only those rows indexed in T, i.e. $\mathbf{A}_T := (\mathbf{a}_j : j \in T)^\top$. Similarly we use $\mathbf{b}_T$ to denote the sub-vector of $\mathbf{b}$ consisting only entries indexed in T. For any matrix $\mathbf{B}$, we use $\operatorname{span}(\mathbf{B})$ to denote the subspace spanned by the *columns* of $\mathbf{B}$. Thus for any index subset T, the notation $\operatorname{span}(\mathbf{A}_T)$ denotes the subspace of $\mathbb{R}^{\#T}$ spanned by the columns of $\mathbf{A}_T$.

Theorem 2.1. *Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{R}^{m\times d}$ and $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$. Then the followings are equivalent:*

(A) *$(\mathbf{A}, \mathbf{b})$ is affine phase retrievable for $\mathbb{R}^d$.*
(B) *The map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ is injective on $\mathbb{R}^d$, where $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ is defined in (1.2).*
(C) *For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and $\mathbf{u} \neq 0$, there exists a k with $1 \le k \le m$ such that*

$$\langle \mathbf{a}_k, \mathbf{u}\rangle \big(\langle \mathbf{a}_k, \mathbf{v}\rangle + b_k\big) \neq 0.$$

(D) *For any $S \subset \{1, 2, \ldots, m\}$, if $\mathbf{b}_S \in \operatorname{span}(\mathbf{A}_S)$ then $\operatorname{span}(\mathbf{A}^\top_{S^c}) = \operatorname{span}\{\mathbf{a}_j : j \in S^c\} = \mathbb{R}^d$.*
(E) *The Jacobian $J(\mathbf{x})$ of the map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ has rank d for all $\mathbf{x} \in \mathbb{R}^d$.*

Proof. The equivalence of (A) and (B) has already been discussed earlier. We focus on the other conditions.

(A) $\Leftrightarrow$ (C). Assume that $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ for some $\mathbf{x} \neq \mathbf{y}$ in $\mathbb{R}^d$. For any j, we have

$$|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j|^2 - |\langle \mathbf{a}_j, \mathbf{y}\rangle + b_j|^2 = \langle \mathbf{a}_j, \mathbf{x}-\mathbf{y}\rangle(\langle \mathbf{a}_j, \mathbf{x}+\mathbf{y}\rangle + 2b_j). \tag{2.1}$$

Set $2\mathbf{u} = \mathbf{x} - \mathbf{y}$ and $2\mathbf{v} = \mathbf{x} + \mathbf{y}$. Then $\mathbf{u} \neq 0$ and for all j,

$$\langle \mathbf{a}_j, \mathbf{u} \rangle \big(\langle \mathbf{a}_j, \mathbf{v} \rangle + b_j \big) = 0. \tag{2.2}$$

Conversely, assume that (2.2) holds for all j. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ be given by $\mathbf{x} - \mathbf{y} = 2\mathbf{u}$ and $\mathbf{x} + \mathbf{y} = 2\mathbf{v}$. Then $\mathbf{x} \neq \mathbf{y}$. However, we would have $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ and hence $(\mathbf{A}, \mathbf{b})$ cannot be affine phase retrievable.

(C) $\Leftrightarrow$ (D). Assume that (C) holds. If for some $S \subset \{1, 2, \ldots, m\}$ with $\mathbf{b}_S \in \operatorname{span}(\mathbf{A}_S)$, we have $\operatorname{span}\{\mathbf{a}_j : \ j \in S^c\} \neq \mathbb{R}^d$, then we can find $\mathbf{u} \neq 0$ such that $\langle \mathbf{a}_j, \mathbf{u} \rangle = 0$ for all $j \in S^c$. Moreover, since $\mathbf{b}_S \in \operatorname{span}(\mathbf{A}_S)$, we can find $\mathbf{v} \in \mathbb{R}^d$ such that $-b_j = \langle \mathbf{a}_j, \mathbf{v} \rangle$ for all $j \in S$. Thus for all $1 \leq j \leq m$, we have

$$\langle \mathbf{a}_j, \mathbf{u} \rangle \big(\langle \mathbf{a}_j, \mathbf{v} \rangle + b_j \big) = 0.$$

This is a contradiction. The converse clearly also holds.

(C) $\Leftrightarrow$ (E). Note that the Jacobian $J(\mathbf{v})$ of the map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ at the point $\mathbf{v} \in \mathbb{R}^d$ is precisely

$$J(\mathbf{v}) = \Big((\langle \mathbf{a}_1, \mathbf{v} \rangle + b_1)\mathbf{a}_1, (\langle \mathbf{a}_2, \mathbf{v} \rangle + b_2)\mathbf{a}_2, \ldots, (\langle \mathbf{a}_m, \mathbf{v} \rangle + b_m)\mathbf{a}_m \Big),$$

i.e. the j-th column of $J(\mathbf{v})$ is $(\langle \mathbf{a}_j, \mathbf{v} \rangle + b_j)\mathbf{a}_j$. Thus $\operatorname{rank}(J(\mathbf{v})) \neq d$ if and only if there exists a nonzero $\mathbf{u} \in \mathbb{R}^d$ such that

$$\mathbf{u}^\top J(\mathbf{v}) = \Big(\langle \mathbf{a}_1, \mathbf{u} \rangle \big(\langle \mathbf{a}_1, \mathbf{v} \rangle + b_1 \big), \ldots, \langle \mathbf{a}_m, \mathbf{u} \rangle \big(\langle \mathbf{a}_m, \mathbf{v} \rangle + b_m \big) \Big) = 0.$$

The equivalence of (C) and (E) now follows. $\square$

As an application of Theorem 2.1, we show that the minimal number of affine measurements to recover all d-dimensional real signals is $2d$.

Theorem 2.2. *Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$. If $m \leq 2d - 1$, then $(\mathbf{A}, \mathbf{b})$ is not affine phase retrievable for $\mathbb{R}^d$.*

Proof. We divide the proof into two cases.

Case 1: $\operatorname{rank}(\mathbf{A}) \leq d - 1$.

In this case, there exists a nonzero vector $\mathbf{u} \in \mathbb{R}^d$ such that $\langle \mathbf{a}_j, \mathbf{u} \rangle = 0$, $1 \leq j \leq m$. Thus for any $\mathbf{x} \in \mathbb{R}^d$,

$$|\langle \mathbf{a}_j, \mathbf{x} \rangle + b_j|^2 = |\langle \mathbf{a}_j, \mathbf{x} + \mathbf{u} \rangle + b_j|^2, \quad 1 \leq j \leq m.$$

This means that $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ is not injective.

Case 2: $\mathrm{rank}(\mathbf{A}) = d$.

In this case, there exists an index set $S_0 \subset \{1, \dots, m\}$ with cardinality d so that the square matrix $\mathbf{A}_{S_0}$ has full rank d, which implies

$$\mathbf{b}_{S_0} \in \mathrm{span}(\mathbf{A}_{S_0}). \tag{2.3}$$

In other words, there exists $\mathbf{v} \in \mathbb{R}^d$ such that $\langle \mathbf{a}_j, \mathbf{v} \rangle + b_j = 0$ for all $j \in S_0$. Now since $m \leq 2d - 1$ and $\#S_0 = d$, we have $\#S_0^c = m - d \leq d - 1$ where $S_0^c := \{1, \dots, m\} \setminus S_0$. Hence there exists a nonzero $\mathbf{u} \in \mathbb{R}^d$ such that $\mathbf{u} \perp \{\mathbf{a}_j : \ j \in S_0^c\}$. The non-injectivity follows immediately from Theorem 2.1 (C). $\square$

We next consider generic measurements. There are various ways one can define the meaning of being generic. A rigorous definition involves the use of Zariski topology. In this paper, we adopt a simpler definition. We say that a generic $\mathbf{u}$ in $\mathbb{H}^N$ has a certain property if there is an open dense set $X \subset \mathbb{H}^N$ so that every $\mathbf{u}$ in X has that property. Sometimes in actual proofs, we obtain the stronger result where $X^c := \mathbb{H}^N \setminus X$ is an algebraic variety. The following theorem on generic measurements also shows that the lower bound given in Theorem 2.2 is optimal.

Theorem 2.3. *Let $m \geq 2d$. Then a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m \times (d+1)}$ is affine phase retrievable.*

Proof. The theorem follows readily from Theorem 2.1 (D). Note that for a generic $\mathbf{A} \in \mathbb{R}^{m \times d}$, any d rows are linearly independent, so that $\mathrm{span}(\mathbf{A}_{S^c}^\top) = \mathbb{R}^d$ as long as $\#S^c \geq d$. On the other hand, $\mathrm{span}(\mathbf{A}_S)$ is a d dimensional subspace in $\mathbb{R}^{\#S}$ and so $\mathbf{b}_S \notin \mathrm{span}(\mathbf{A}_S)$ if $\#S > d$. Thus if $\mathbf{b}_S \in \mathrm{span}(\mathbf{A}_S)$, then $\#S \leq d$, which implies $\#S^c \geq d$. Consequently $\mathrm{span}\{\mathbf{a}_j : \ j \in S^c\} = \mathrm{span}(\mathbf{A}_{S^c}^\top) = \mathbb{R}^d$. Hence $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable. $\square$

The following theorem highlights a difference between the classical linear phase retrieval and the affine phase retrieval.

Theorem 2.4. *Let $m \geq 2d$. Then the set of affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m \times (d+1)}$ is not an open set.*

Proof. We only need to find an affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m \times (d+1)}$ such that for each $\epsilon > 0$, there is a small perturbation $(\mathbf{A}', \mathbf{b}) \in \mathbb{R}^{m \times (d+1)}$ with $\|\mathbf{A} - \mathbf{A}'\|_F < \epsilon$ such that $(\mathbf{A}', \mathbf{b})$ is not affine phase retrievable, where $\| \cdot \|_F$ denotes the l^2-norm (Frobenius norm). We first do so for $m = 2d$. Set

$$\mathbf{A} = (I_d, I_d)^\top, \qquad \mathbf{b} = (b_{11}, \dots, b_{d1}, b_{12}, \dots, b_{d2})^\top.$$

Here we require that $b_{j1} \neq b_{j2}$ for all j and specially suppose $b_{12} = 0$. Then $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable. To see this, assume that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})$.

Then for each j, we must have $|x_j + b_{jk}| = |y_j + b_{jk}|$ for $k = 1, 2$. Since $b_{j1} \neq b_{j2}$, we must have $x_j = y_j$. Thus $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ is injective and hence $(\mathbf{A}, \mathbf{b})$ is phase retrievable.

Now let $\delta > 0$ be sufficiently small. We perturb $\mathbf{A}$ to

$$\mathbf{A}' = (I_d + b_{11}\delta E_{21}, I_d)^\top, \tag{2.4}$$

where E_{ij} denotes the matrix with the (i, j)-th entry being 1 and all other entries being 0. Now set $\mathbf{x} = (b_{11}, -1/\delta, 0, \ldots, 0)^\top$ and $\mathbf{y} = (-b_{11}, -1/\delta, 0, \ldots, 0)^\top$. It is easy to see that

$$|\mathbf{A}'\mathbf{x} + \mathbf{b}| = |\mathbf{A}'\mathbf{y} + \mathbf{b}|.$$

Hence $(\mathbf{A}', \mathbf{b})$ is not affine phase retrievable. By taking δ sufficiently small, we will have $\|\mathbf{A}' - \mathbf{A}\|_F \leq \epsilon$. It follows that for $m = 2d$, the set of affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ is not an open set.

In general for $m > 2d$, we can simply take the above construction $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{2d\times(d+1)}$ and augment it to a matrix $(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}) \in \mathbb{R}^{m\times(d+1)}$ by appending $m - 2d$ rows of zero vectors to form its last $m - 2d$ rows. The $(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})$ is clearly affine phase retrievable, and the same perturbation above applied to the first $2d$ rows of $\mathbf{A}$ now breaks the affine phase retrievability. Thus for any $m \geq 2d$, the set of affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ is not an open set. □

2.2. Real sparse affine phase retrieval

Set

$$\Sigma_s(\mathbb{H}^d) := \{\mathbf{x} \in \mathbb{H}^d : \|\mathbf{x}\|_0 \leq s\}.$$

We say that $(\mathbf{A}, \mathbf{b}) \in \mathbb{H}^{m\times(d+1)}$ is *s-sparse affine phase retrievable for* $\mathbb{H}^d$ if $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ is injective on $\Sigma_s(\mathbb{H}^d)$. In this subsection, we show that the minimal number of affine measurements to recover all s-sparse real signals is $2s + 1$.

Theorem 2.5.

(i) *Let $1 \leq s \leq d-1$ and $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ be s-sparse affine phase retrievable for $\mathbb{R}^d$. Then $m \geq 2s + 1$.*
(ii) *Let $m \geq 2s+1$ and $(\mathbf{A}, \mathbf{b})$ be a generic element in $\mathbb{R}^{m\times(d+1)}$. Then $(\mathbf{A}, \mathbf{b})$ is s-sparse affine phase retrievable for $\mathbb{R}^d$.*

Proof. (i) We first show that if $(\mathbf{A}, \mathbf{b})$ is s-sparse affine phase retrievable, then $m \geq 2s+1$. First we claim that the rank of $\mathbf{A}$ is at least $r = \min(d, 2s)$. Indeed, suppose that the claim is false. Then there exists a nonzero vector $\mathbf{x} \in \Sigma_r(\mathbb{R}^d)$, such that $\mathbf{A}\mathbf{x} = \mathbf{0}$. Write $\mathbf{x} = \mathbf{u} - \mathbf{v}$ with $\mathbf{u}, \mathbf{v} \in \Sigma_s(\mathbb{R}^d)$. Then $\mathbf{u} \neq \mathbf{v}$ and $\mathbf{A}\mathbf{u} = \mathbf{A}\mathbf{v}$. Hence for all $1 \leq j \leq m$, we have

$$|\langle \mathbf{a}_j, \mathbf{u}\rangle + b_j| = |\langle \mathbf{a}_j, \mathbf{v}\rangle + b_j|,$$

which is a contradiction. Thus $\operatorname{rank}(\mathbf{A}) \ge r = \min(d, 2s)$.

Assume that $m \le 2s$. We derive a contradiction. Since $s < d$, it follows that $r \ge s+1$. Thus there exists an index set $T \subset \{1, 2, \dots, m\}$ with $\#T = s+1$, such that $\operatorname{rank}(\mathbf{A}_T) = s+1$. Without of loss of generality we may assume that $T = \{1, 2, \dots, s+1\}$. Moreover, we may also without of loss of generality assume that the first $s+1$ columns of $\mathbf{A}_T$ are linearly independent. In other words, the $(s+1) \times (s+1)$ submatrix of $\mathbf{A}$ restricted to the first $s+1$ rows and columns is nonsingular. Call this matrix $\mathbf{B}$. It follows that there exists a $\mathbf{y} \in \mathbb{R}^{s+1}$ such that $\mathbf{B}\mathbf{y} = -\mathbf{b}_T$. Write $\mathbf{y} = (y_1, \dots, y_{s+1})^\top$ and set

$$\mathbf{v}_0 = (y_1, \dots, y_{s+1}, 0, \dots, 0)^\top \in \mathbb{R}^d.$$

Then $\mathbf{A}_T\mathbf{v}_0 = -\mathbf{b}_T$.

If $y_j = 0$ for some $1 \le j \le s+1$, say $y_{s+1} = 0$, we let $\mathbf{u} = (u_1, \dots, u_s, 0, \dots, 0)^\top$. Since $\#T^c = m - (s+1) \le s-1$, there exists such a $\mathbf{u}_0 \ne 0$ such that $\langle \mathbf{a}_j, \mathbf{u}_0\rangle = 0$ for all $j \in T^c$. Now for $\mathbf{x} = \mathbf{v}_0 + \mathbf{u}_0$ and $\mathbf{y} = \mathbf{v}_0 - \mathbf{u}_0$, we have $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ and $\mathbf{x} \ne \mathbf{y}$. Furthermore, $\mathbf{x}, \mathbf{y} \in \Sigma_s(\mathbb{R}^d)$. This is a contradiction. Hence $y_j \ne 0$ for all $1 \le j \le s$.

Now for any $1 \le j_1 < j_2 \le s+1$ consider

$$\mathbf{u}_{j_1,j_2} = (u_1, \dots, u_{s+1}, 0, \dots, 0)^\top \in \mathbb{R}^d, \qquad u_{j_1} = ty_{j_1},\ u_{j_2} = -ty_{j_2}. \tag{2.5}$$

We view the other u_j's and t as unconstrained variables, so there are s variables. Since $\#T^c = m - (s+1) \le s-1$, it follows that there exists a $\tilde{\mathbf{u}}_{j_1,j_2} \ne 0$ satisfying (2.5) such that $\langle \mathbf{a}_j, \tilde{\mathbf{u}}_{j_1,j_2}\rangle = 0$ for all $j \in T^c$. If $t \ne 0$, then we may normalize $\tilde{\mathbf{u}}_{j_1,j_2}$ so that $t = 1$. Set $\mathbf{x} = \mathbf{v}_0 + \tilde{\mathbf{u}}_{j_1,j_2}$ and $\mathbf{y} = \mathbf{v}_0 - \tilde{\mathbf{u}}_{j_1,j_2}$. It follows that $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ and

$$\operatorname{supp}(\mathbf{x}) \subset \{1, 2, \dots, s+1\} \setminus \{j_2\}, \quad \operatorname{supp}(\mathbf{y}) \subset \{1, 2, \dots, s+1\} \setminus \{j_1\}.$$

This is a contradiction.

To complete the proof of $m \ge 2s+1$, we finally need to consider the case that $t = 0$ in $\tilde{\mathbf{u}}_{j_1,j_2} \ne 0$ for every pair of indices $1 \le j_1 < j_2 \le s+1$. But if so, it implies that any $s-1$ columns among the first $s+1$ columns of $\mathbf{A}_{T^c}$ are linearly dependent. In particular, it means the $(m-s-1) \times (s+1)$ submatrix of $\mathbf{A}_{T^c}$ restricted to the first $s+1$ columns has rank at most $s-2$. Now because the $(s+1) \times (s+1)$ submatrix of $\mathbf{A}$ restricted to the first $s+1$ rows and columns is nonsingular, we may without loss of generality assume that $s \times s$ submatrix of $\mathbf{A}$ restricted to the first s rows and columns is nonsingular, for otherwise we can make a simple permutation of the indices. The key now is to observe that $(\mathbf{A}, \mathbf{b})$ is not s-sparse affine phase retrievable, because $(\mathbf{A}, \mathbf{b})$ restricted to the first s columns is not affine phase retrievable for $\mathbb{R}^s$. To see this, let $\mathbf{A}'$ be the submatrix of $\mathbf{A}$ consisting of only the first s columns of $\mathbf{A}$. We show $(\mathbf{A}', \mathbf{b})$ is not affine phase retrievable for $\mathbb{R}^s$. Note that for $S = \{1, 2, \dots, s\}$, we have $\mathbf{b}_S \in \operatorname{span}(\mathbf{A}'_S)$ because by assumption $\mathbf{A}'_S$ is nonsingular. But we also know that the rows of $\mathbf{A}_{S^c}$ do not span $\mathbb{R}^s$ because it has $\mathrm{rank}(\mathbf{A}_{S^c}) \le s-1$. Hence $(\mathbf{A}', \mathbf{b})$ is not affine phase retrievable by Theorem 2.1 (D). This completes the proof of $m \ge 2s+1$.

(ii) Next we prove for $m \ge 2s+1$, a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ is s-sparse affine phase retrievable. The set of all $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ has real dimension $m(d+1)$. The goal is to show that the set of $(\mathbf{A}, \mathbf{b})$ that are not s-sparse affine phase retrievable lies in a finite union of subsets of dimension strictly less than $m(d+1)$. Our result then follows.

For any subset of indices $I, J \subset \{1, \dots, m\}$ with $\#I, \#J \le s$, we say $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ is not (I, J)-sparse affine phase retrievable if there exist $\mathbf{x} \ne \mathbf{y}$ in $\mathbb{R}^d$ such that

$$\mathrm{supp}(\mathbf{x}) \subset I, \quad \mathrm{supp}(\mathbf{y}) \subset J, \quad \text{and} \quad \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y}). \tag{2.6}$$

Let $\mathcal{A}_{I,J}$ denote the set of all 4-tuples $(\mathbf{A}, \mathbf{b}, \mathbf{x}, \mathbf{y})$ satisfying (2.6) and $\mathbf{x} \ne \mathbf{y}$. Then

$$\mathcal{A}_{I,J} \subset \mathbb{R}^{m\times(d+1)} \times \mathbb{R}^d \times \mathbb{R}^d.$$

Then $\mathcal{A}_{I,J}$ is a well-defined real quasi-projective variety ([14, Page 18]). Write $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m)^\top$ and $\mathbf{b} = (b_1, \dots, b_m)^\top$. Then by (2.1), $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ is equivalent to

$$\langle \mathbf{a}_j, \mathbf{x} - \mathbf{y}\rangle(\langle \mathbf{a}_j, \mathbf{x} + \mathbf{y}\rangle + 2b_j) = 0, \qquad j = 1, 2, \dots, m. \tag{2.7}$$

Fix any j, the above equation holds if and only if

$$\langle \mathbf{a}_j, \mathbf{x} - \mathbf{y}\rangle = 0 \qquad \text{or} \qquad \langle \mathbf{a}_j, \mathbf{x} + \mathbf{y}\rangle + 2b_j = 0.$$

Thus for any $\mathbf{x} \ne \mathbf{y}$, the first condition requires $\mathbf{a}_j$ to lie on a hyperplane, which has co-dimension 1, while the second condition fixes b_j to be $-\langle \mathbf{a}_j, \mathbf{x} + \mathbf{y}\rangle/2$. Overall, for any given $\mathbf{x} \ne \mathbf{y}$, these two conditions constraint the j-th row of $(\mathbf{A}, \mathbf{b})$ to lie on a real projective variety of codimension 1. We shall use $X_j(\mathbf{x}, \mathbf{y})$ to denote this variety which is a subvariety of $\mathbb{R}^{d+1}$. Now let $\pi_2 : \mathcal{A}_{I,J} \longrightarrow \mathbb{R}^d \times \mathbb{R}^d$ be the projection $(\mathbf{A}, \mathbf{b}, \mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}, \mathbf{y})$ onto the last two coordinates. Then for any $\mathbf{x}_0 \ne \mathbf{y}_0$ in $\mathbb{R}^d$ with $\mathrm{supp}(\mathbf{x}_0) \subset I$ and $\mathrm{supp}(\mathbf{y}_0) \subset J$, we have

$$\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\} = X_1(\mathbf{x}_0, \mathbf{y}_0) \times X_2(\mathbf{x}_0, \mathbf{y}_0) \times \ldots \times X_m(\mathbf{x}_0, \mathbf{y}_0) \times \{\mathbf{x}_0\} \times \{\mathbf{y}_0\}.$$

Hence the dimension of $\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\}$ is

$$\dim\left(\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\}\right) = m(d+1) - m = md.$$

It follows that $\dim(\mathcal{A}_{I,J}) \le md + \#I + \#J \le md + 2s$.

We now let $\pi_1 : \mathcal{A}_{I,J} \longrightarrow \mathbb{R}^{m\times(d+1)}$ be the projection $(\mathbf{A}, \mathbf{b}, \mathbf{x}, \mathbf{y}) \mapsto (\mathbf{A}, \mathbf{b})$. Since projections cannot increase the dimension of a variety, we know that

$$\dim(\pi_1(\mathcal{A}_{I,J})) \le md + 2s = m(d+1) + 2s - m < m(d+1).$$

However, $\pi_1(\mathcal{A}_{I,J})$ contains precisely those $(\mathbf{A}, \mathbf{b})$ in $\mathbb{R}^{m\times(d+1)}$ that are not (I, J)-sparse affine phase retrievable. Thus a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ is (I, J)-sparse affine phase retrievable.

Finally, there are only finitely many indices subsets I, J. Hence a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{m\times(d+1)}$ $(m \ge 2s + 1)$ is (I, J)-sparse affine phase retrievable for any I, J with $\#I, \#J \le s$. The theorem is proved. $\square$

3. Affine phase retrieval for complex signals

In this section, we consider affine phase retrieval for complex signals. Affine phase retrieval for complex signals, like in the case of the classical phase retrieval, poses additional challenges.

3.1. Complex affine phase retrieval

We first establish the analogue of Theorem 2.1 for complex signals. Throughout this paper, $\langle \mathbf{u}, \mathbf{v}\rangle := \sum_{j=1}^{d} u_j \overline{v_j}$ for $\mathbf{u} = (u_1, \ldots, u_d) \in \mathbb{C}^d$, $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{C}^d$.

Theorem 3.1. *Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{C}^{m\times d}$ and $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{C}^m$. Then the followings are equivalent:*

(A) *$(\mathbf{A}, \mathbf{b})$ is affine phase retrievable for $\mathbb{C}^d$.*
(B) *The map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ is injective on $\mathbb{C}^d$.*
(C) *For any $\mathbf{u}, \mathbf{v} \in \mathbb{C}^d$ and $\mathbf{u} \neq 0$, there exists a $1 \le k \le m$ such that*

$$\Re\Big(\langle \mathbf{u}, \mathbf{a}_k\rangle \big(\langle \mathbf{a}_k, \mathbf{v}\rangle + b_k\big)\Big) \neq 0.$$

(D) *Viewing $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ as a map $\mathbb{R}^{2d} \longrightarrow \mathbb{R}^m$, its (real) Jacobian $J(\mathbf{x})$ has rank $2d$ for all $\mathbf{x} \in \mathbb{R}^{2d}$.*

Proof. The equivalence of (A) and (B) has already been discussed earlier. We focus on the other conditions.

(A) $\Leftrightarrow$ (C). Assume that $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ for some $\mathbf{x} \neq \mathbf{y}$ in $\mathbb{C}^d$. Observe that for any $a, b \in \mathbb{C}$, we have $|a|^2 - |b|^2 = \Re((\bar{a} - \bar{b})(a + b))$. Thus for any j, we have

$$|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_j|^2 - |\langle \mathbf{a}_j, \mathbf{y}\rangle + b_j|^2 = \Re\Big(\langle \mathbf{x} - \mathbf{y}, \mathbf{a}_j\rangle (\langle \mathbf{a}_j, \mathbf{x} + \mathbf{y}\rangle + 2b_j)\Big). \tag{3.1}$$

Set $2\mathbf{u} = \mathbf{x} - \mathbf{y}$ and $2\mathbf{v} = \mathbf{x} + \mathbf{y}$. Then $\mathbf{u} \neq 0$ and for all j,

$$\Re\Big(\langle \mathbf{u}, \mathbf{a}_j\rangle (\langle \mathbf{a}_j, \mathbf{v}\rangle + b_j)\Big) = 0. \tag{3.2}$$

Conversely, assume that (3.2) holds for all j. Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^d$ be given by $\mathbf{x} - \mathbf{y} = 2\mathbf{u}$ and $\mathbf{x} + \mathbf{y} = 2\mathbf{v}$. Then $\mathbf{x} \neq \mathbf{y}$. However, we would have $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ and hence $(\mathbf{A}, \mathbf{b})$ cannot be affine phase retrievable.

(C) $\Leftrightarrow$ (D). The k-th entry of $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is $|\langle \mathbf{a}_k, \mathbf{x}\rangle + b_k|^2$. Since all variables here are complex, we shall separate them into the real and imaginary parts by adopting the notation $\mathbf{x} = \mathbf{x}_R + i\mathbf{x}_I$, $\mathbf{a}_k = \mathbf{a}_{k,R} + i\mathbf{a}_{k,I}$ and $b_k = b_{k,R} + ib_{k,I}$. The k-th entry of $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is now

$$|\langle \mathbf{a}_k, \mathbf{x}\rangle + b_k|^2 = \left(\langle \mathbf{a}_{k,R}, \mathbf{x}_R\rangle + \langle \mathbf{a}_{k,I}, \mathbf{x}_I\rangle + b_{k,R}\right)^2 + \left(\langle \mathbf{a}_{k,R}, \mathbf{x}_I\rangle - \langle \mathbf{a}_{k,I}, \mathbf{x}_R\rangle - b_{k,I}\right)^2.$$

It follows that the (real) Jacobian of $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}_R, \mathbf{x}_I)$ is

$$J(\mathbf{x}) := J(\mathbf{x}_R, \mathbf{x}_I) = 2\begin{pmatrix} \mathbf{a}_{1,R}^\top \cdot \alpha_1(\mathbf{x}) - \mathbf{a}_{1,I}^\top \cdot \beta_1(\mathbf{x}) & \mathbf{a}_{1,I}^\top \cdot \alpha_1(\mathbf{x}) + \mathbf{a}_{1,R}^\top \cdot \beta_1(\mathbf{x}) \\ \mathbf{a}_{2,R}^\top \cdot \alpha_2(\mathbf{x}) - \mathbf{a}_{2,I}^\top \cdot \beta_2(\mathbf{x}) & \mathbf{a}_{2,I}^\top \cdot \alpha_2(\mathbf{x}) + \mathbf{a}_{2,R}^\top \cdot \beta_2(\mathbf{x}) \\ \vdots & \vdots \\ \mathbf{a}_{m,R}^\top \cdot \alpha_m(\mathbf{x}) - \mathbf{a}_{m,I}^\top \cdot \beta_m(\mathbf{x}) & \mathbf{a}_{m,I}^\top \cdot \alpha_m(\mathbf{x}) + \mathbf{a}_{m,R}^\top \cdot \beta_m(\mathbf{x}) \end{pmatrix},$$

where $\alpha_j(\mathbf{x}) := \langle \mathbf{a}_{j,R}, \mathbf{x}_R\rangle + \langle \mathbf{a}_{j,I}, \mathbf{x}_I\rangle + b_{j,R}$ and $\beta_j(\mathbf{x}) := \langle \mathbf{a}_{j,R}, \mathbf{x}_I\rangle - \langle \mathbf{a}_{j,I}, \mathbf{x}_R\rangle - b_{j,I}$ for all $0 \le j \le m$.

Now assume that $\operatorname{rank}(J(\mathbf{x}))$ is not $2d$ everywhere. Then there exist $\mathbf{v} = \mathbf{v}_R + i\mathbf{v}_I$ and $\mathbf{u} = \mathbf{u}_R + i\mathbf{u}_I \neq 0$, such that $\mathbf{u}$ as a vector in $\mathbb{R}^{2d}$ is in the null space of $J(\mathbf{v})$, i.e.,

$$J(\mathbf{v})\begin{pmatrix} \mathbf{u}_R \\ \mathbf{u}_I \end{pmatrix} = 0.$$

It follows that for all $1 \le k \le m$, we have

$$C_k := \langle \mathbf{a}_{k,R}, \mathbf{u}_R\rangle\alpha_k(\mathbf{v}) - \langle \mathbf{a}_{k,I}, \mathbf{u}_R\rangle\beta_k(\mathbf{v}) + \langle \mathbf{a}_{k,I}, \mathbf{u}_I\rangle\alpha_k(\mathbf{v}) + \langle \mathbf{a}_{k,R}, \mathbf{u}_I\rangle\beta_k(\mathbf{v}) = 0. \tag{3.3}$$

But one can readily check that C_k is precisely

$$C_k = \Re\Big(\langle \mathbf{u}, \mathbf{a}_k\rangle(\langle \mathbf{a}_k, \mathbf{v}\rangle + b_k)\Big).$$

Thus $(\mathbf{A}, \mathbf{b})$ cannot be affine phase retrievable by (C).

The converse clearly also holds. Assume that (C) is false. Then there exist $\mathbf{v}, \mathbf{u} \in \mathbb{C}^d$ and $\mathbf{u} \neq 0$ such that

$$\Re\Big(\langle \mathbf{u}, \mathbf{a}_k\rangle(\langle \mathbf{a}_k, \mathbf{v}\rangle + b_k)\Big) = 0$$

for all $1 \le k \le m$. It follows that (3.3) holds for all k and hence

$$J(\mathbf{v})\begin{pmatrix}\mathbf{u}_R\\ \mathbf{u}_I\end{pmatrix}=0.$$

Thus $\mathrm{rank}(J(\mathbf{v}))<2d$. □

3.2. Minimal measurement number

We now show that the minimal number of measurements needed to be affine phase retrievable is $3d$. This is surprising compared to the classical affine phase retrieval, where the minimal number is $4d - O(\log_2 d)$.

Lemma 3.1. *Let $z_1, z_2 \in \mathbb{C}$. Suppose that $b_1, b_2, b_3 \in \mathbb{C}$ are not collinear on the complex plane. Then $z_1 = z_2$ if and only if $|z_1 + b_j| = |z_2 + b_j|$, $j = 1, 2, 3$.*

Proof. We use $z_{j,R}$ and $z_{j,I}$ to denote the real and imaginary part of z_j, and similarly for $b_{j,R}$ and $b_{j,I}$. Assume the lemma is false, and that there exist z_1, z_2 with $z_1 \neq z_2$ so that $|z_1 + b_j|^2 = |z_2 + b_j|^2$, $j = 1, 2, 3$. Note that $|z_1 + b_j|^2 = |z_2 + b_j|^2$ implies that

$$(z_{2,R} - z_{1,R}) \cdot b_{j,R} + (z_{2,I} - z_{1,I}) \cdot b_{j,I} = \frac{|z_1|^2 - |z_2|^2}{2}, \quad j = 1, 2, 3. \tag{3.4}$$

The (3.4) together with $z_1 \neq z_2$ implies that b_1, b_2, b_3 are collinear. This is a contradiction. □

Theorem 3.2.

(i) *Suppose that $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is affine phase retrievable in $\mathbb{C}^d$. Then $m \geq 3d$.*
(ii) *Let $B := (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \mathbb{C}^{d\times d}$ be nonsingular. Set $\mathbf{A} = (B, B, B)^\top \in \mathbb{C}^{3d\times d}$. Let*

$$\mathbf{b} = (b_{11}, \ldots, b_{d1}, b_{12}, \ldots, b_{d2}, b_{13}, \ldots, b_{d3})^\top \in \mathbb{C}^{3d}$$

such that b_{j1}, b_{j2}, b_{j3} are not collinear in $\mathbb{C}$ for any $1 \leq j \leq d$. Then $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable in $\mathbb{C}^d$.

Proof. (i) Write $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^\top \in \mathbb{C}^{m\times d}$. Assume that $m < 3d$. Clearly $\mathrm{rank}(\mathbf{A}) = d$, for otherwise we will have $\mathbf{A}\overline{\mathbf{x}} = 0$ for some $\mathbf{x} \neq 0$ and hence $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}_{\mathbf{A},\mathbf{b}}(0)$. Hence there exists a $T \subset \{1, \ldots, m\}$ with $\#T = d$ such that $\mathrm{rank}(A_T) = d$, which means we can find $\mathbf{v} \in \mathbb{C}^d$ such that

$$\langle \mathbf{a}_k, \mathbf{v}\rangle + b_k = 0, \qquad k \in T.$$

Now, because $\#T^c = m - d < 2d$, and the system of homogeneous linear equations for the variable $\mathbf{u}$ with $\mathbf{v}$ fixed,

$$\Re\Big(\langle \mathbf{u}, \mathbf{a}_j\rangle(\langle \mathbf{a}_j, \mathbf{v}\rangle + b_j)\Big) = 0, \qquad j \in T^c$$

has $2d$ real variables $\mathbf{u}_R$, $\mathbf{u}_I$, it must have a nontrivial solution. The two vectors $\mathbf{u} \neq 0$, $\mathbf{v}$ combine to yield

$$\Re\Big(\langle \mathbf{u}, \mathbf{a}_j\rangle(\langle \mathbf{a}_j, \mathbf{v}\rangle + b_j)\Big) = 0$$

for all $1 \le j \le m$. This contradicts with (C) in Theorem 3.1.

(ii) To prove $(\mathbf{A}, \mathbf{b})$ is affine phase retrievable, we prove that $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ implies $\mathbf{x} = \mathbf{y}$ in $\mathbb{C}^d$. The property $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ implies that

$$|\langle \mathbf{a}_j, \mathbf{x}\rangle + b_{jk}| \;=\; |\langle \mathbf{a}_j, \mathbf{y}\rangle + b_{jk}|, \qquad j = 1, \ldots, d, \quad k = 1, 2, 3. \tag{3.5}$$

Thus by Lemma 3.1, for each fixed j we have

$$\langle \mathbf{a}_j, \mathbf{x}\rangle = \langle \mathbf{a}_j, \mathbf{y}\rangle.$$

This implies $\mathbf{x} = \mathbf{y}$ since the matrix $B = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ is nonsingular. □

It is well known that in the classical phase retrieval, the set of all phase retrievable $\mathbf{A} \in \mathbb{C}^{m\times d}$ is an open set in $\mathbb{C}^{m\times d}$. But for affine phase retrieval, as with the real affine phase retrieval case, this property no longer holds. The following theorem shows that this property also doesn't hold in the complex case when $m \ge 3d$.

Theorem 3.3. *Let $m \ge 3d$. Then the set of affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is not an open set in $\mathbb{C}^{m\times(d+1)}$. In fact, there exists an affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{3d\times(d+1)}$, which satisfies the conditions in Theorem 3.2 (ii). Given any $\epsilon > 0$, there exists $(\mathbf{A}', \mathbf{b}) \in \mathbb{C}^{3d\times(d+1)}$ which does not have affine phase retrievable property such that*

$$\|\mathbf{A}' - \mathbf{A}\|_F \le \epsilon,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. Following the construction given in Theorem 3.2 (ii), we set $\mathbf{A} = (B, B, B)^\top$, where B is nonsingular and

$$\mathbf{b} := (\underbrace{\mathrm{i}, \ldots, \mathrm{i}}_{d}, \underbrace{0, \ldots, 0}_{d}, \underbrace{1, \ldots, 1}_{d})^\top \in \mathbb{C}^{3d}.$$

We will show that there exists an arbitrarily small perturbation $\mathbf{A}'$ such that $(\mathbf{A}', \mathbf{b})$ is no longer affine phase retrievable. Making a simple linear transformation $\mathbf{x} = B^{-1}\mathbf{y}$, we see that all we need is to show that this property holds for $\mathbf{A} = (I_d, I_d, I_d)^\top$, where I_d is the $d \times d$ identity matrix. Let $\delta > 0$ be sufficiently small. We perturb $\mathbf{A}$ to

$$\mathbf{A}' = (I_d + \mathrm{i}\delta E_{21}, I_d, I_d)^\top, \tag{3.6}$$

where E_{21} denotes the matrix with the $(2,1)$-th entry being 1 and all other entries being 0. Now set $\mathbf{x} = (\mathrm{i}, -1/\delta, 0, \ldots, 0)^\top$ and $\mathbf{y} = (-\mathrm{i}, -1/\delta, 0, \ldots, 0)^\top$. It is easy to see that

$$|\mathbf{A}'\mathbf{x} + \mathbf{b}| = |\mathbf{A}'\mathbf{y} + \mathbf{b}|$$

which implies that

$$\mathbf{M}_{(\mathbf{A}',\mathbf{b})}(\overline{\mathbf{x}}) = \mathbf{M}_{(\mathbf{A}',\mathbf{b})}(\overline{\mathbf{y}}).$$

Thus $(\mathbf{A}', \mathbf{b})$ is not affine phase retrievable. By taking δ sufficiently small we will have $\|\mathbf{A}' - \mathbf{A}\|_F \leq \epsilon$.

In general for $m > 3d$, like the real case, we can simply take the above construction $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{3d\times(d+1)}$ and augment it to a matrix $(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}) \in \mathbb{C}^{m\times(d+1)}$ by appending $m-3d$ rows of zero vectors to form its last $m-3d$ rows. $(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})$ is clearly affine phase retrievable, and the same perturbation above applied to the first $3d$ rows of $\tilde{\mathbf{A}}$ now breaks the affine phase retrievability.

Thus for any $m \geq 3d$, the set of affine phase retrievable $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is not an open set. $\square$

We next consider complex affine phase retrieval for generic measurements. We have the following theorem:

Theorem 3.4. *Suppose that $m \geq 4d - 1$. Then a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is affine phase retrievable in $\mathbb{C}^d$.*

Proof. Let $N = m+1$. Then $N \geq 4d = 4(d+1)-4$. Hence by [11], there is an open dense set of full measure $X \subset \mathbb{C}^{N\times(d+1)}$, such that any $\mathbf{F} \in X$ is linear phase retrievable in the classical sense. Write $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N)^\top$, where each $\mathbf{f}_j \in \mathbb{C}^{d+1}$. For each $\mathbf{g} \in \mathbb{C}^{d+1}$, denote $X_{\mathbf{g}} := \{\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N)^\top \in X : \ \mathbf{f}_N = \mathbf{g}\}$. Then there exists a $\mathbf{g}_0 \in \mathbb{C}^{d+1}$, such that the projection of $X_{\mathbf{g}_0}$ onto $\mathbb{C}^{(N-1)\times(d+1)}$ with the last row removed is a dense open set with full measure. Thus $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_{N-1}, \mathbf{g}_0)^\top$ is phase retrievable in $\mathbb{C}^{d+1}$ in the classical sense for a generic $(\mathbf{f}_1, \ldots, \mathbf{f}_{N-1})^\top \in \mathbb{C}^{(N-1)\times(d+1)}$.

Now let $P_0 \in \mathbb{C}^{(d+1)\times(d+1)}$ be nonsingular such that $P_0\mathbf{g}_0 = \mathbf{e}_{d+1}$. Then for any $\mathbf{F} \in X_{\mathbf{g}_0}$, we have

$$\mathbf{G} := \mathbf{F}P_0^\top = (P_0\mathbf{f}_1, \ldots, P_0\mathbf{f}_{N-1}, \mathbf{e}_{\mathbf{d+1}})^\top =: (\mathbf{g}_1, \ldots, \mathbf{g}_{N-1}, \mathbf{e}_{d+1})^\top.$$

It is linear phase retrievable in the classical sense for generic $\mathbf{g}_1, \ldots, \mathbf{g}_{N-1}$. In particular, any vector $\mathbf{y} = (x_1, \ldots, x_d, 1)^\top$ can be recovered by $|\mathbf{G}\mathbf{y}|$, where $|\cdot|$ means the entry-wise absolute value. However, note that the last entry of $\mathbf{y}$ is 1, and the last row of $\mathbf{G}$ is $\mathbf{e}_{d+1}^\top$.

So the measurement from last row provides no information. In other words, the above $\mathbf{y}$ can be recovered exactly from the measurements provided by the first $N-1=m$ rows of $\mathbf{G}$. This means precisely that the first $N-1=m$ rows of $\mathbf{G}$ are affine phase retrievable. Let $(\mathbf{A},\mathbf{b})$ denote the first m rows of $\mathbf{G}$. It follows that $(\mathbf{A},\mathbf{b})\in\mathbb{C}^{m\times(d+1)}$ is affine phase retrievable. Therefore a generic $(\mathbf{A},\mathbf{b})\in\mathbb{C}^{m\times(d+1)}$ $(m\geq 4d-1)$ is affine phase retrievable. $\square$

3.3. Complex sparse affine phase retrieval

We now focus on sparse affine phase retrieval by proving that generic $(\mathbf{A},\mathbf{b})$ is s-sparse affine phase retrievable if $m\geq 4s+1$.

Theorem 3.5. *Let $m\geq 4s+1$. Then a generic $(\mathbf{A},\mathbf{b})\in\mathbb{C}^{m\times(d+1)}$ is s-sparse affine phase retrievable.*

Proof. The proof here is very similar to the proof in the real case. The set of all $(\mathbf{A},\mathbf{b})$ has real dimension $\dim_{\mathbb{R}}(\mathbb{C}^{m\times(d+1)})=2m(d+1)$. The goal is to show that the set of $(\mathbf{A},\mathbf{b})$ that are not s-sparse affine phase retrievable lies in a finite union of subsets, each of which is a projection of real hypersurfaces of dimension strictly less than $2m(d+1)$. This would yield our result.

For any subset of indices $I,J\subset\{1,\ldots,m\}$ with $\#I,\#J\leq s$, we say $(\mathbf{A},\mathbf{b})\in\mathbb{C}^{m\times(d+1)}$ is not (I,J)-sparse affine phase retrievable if there exist $\mathbf{x}\neq\mathbf{y}$ in $\mathbb{C}^d$ such that

$$\operatorname{supp}(\mathbf{x})\subset I,\quad \operatorname{supp}(\mathbf{y})\subset J,\quad \text{and}\quad \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})=\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y}). \tag{3.7}$$

Let $\mathcal{A}_{I,J}$ denote the set of all 4-tuples $(\mathbf{A},\mathbf{b},\mathbf{x},\mathbf{y})$ satisfying (3.7) and $\mathbf{x}\neq\mathbf{y}$. Then

$$\mathcal{A}_{I,J}\subset\mathbb{C}^{m\times(d+1)}\times\mathbb{C}^d\times\mathbb{C}^d,$$

where we view $(\mathbf{A},\mathbf{b})$ as an element of $\mathbb{C}^{m\times(d+1)}$. For our proof we shall identify $\mathbb{C}^{m\times(d+1)}\times\mathbb{C}^d\times\mathbb{C}^d$ with $\mathbb{R}^{m\times 2(d+1)}\times\mathbb{R}^{2d}\times\mathbb{R}^{2d}$. In this case $\mathcal{A}_{I,J}$ is a well-defined real quasi-projective variety ([14, Page 18]). Note that $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})=\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})$ yields $|\langle\mathbf{a}_j,\mathbf{x}\rangle+b_j|^2=|\langle\mathbf{a}_j,\mathbf{y}\rangle+b_j|^2$ for all $1\leq j\leq m$, where $\mathbf{A}=(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_m)^\top$ and $\mathbf{b}=(b_1,\ldots,b_m)^\top$. By (3.1), this is equivalent to

$$\Re\Big(\langle\mathbf{x}-\mathbf{y},\mathbf{a}_j\rangle(\langle\mathbf{a}_j,\mathbf{x}+\mathbf{y}\rangle+2b_j)\Big)=0,\qquad j=1,2,\ldots,m. \tag{3.8}$$

Fix any j, the above equation holds if and only if

- $\langle\mathbf{x}-\mathbf{y},\mathbf{a}_j\rangle=0$; or
- $\langle\mathbf{x}-\mathbf{y},\mathbf{a}_j\rangle\neq 0$ but (3.8) holds.

Thus for any $\mathbf{x} \neq \mathbf{y}$, the first condition requires $\mathbf{a}_j$ to lie on a hyperplane, which has real co-dimension 2, while the second condition requires b_j to be on a line in $\mathbb{C}$ (depending on $\mathbf{x}$, $\mathbf{y}$, $\mathbf{a}_j$). Overall, for any given $\mathbf{x} \neq \mathbf{y}$, these two conditions constraint the j-th row of $(\mathbf{A}, \mathbf{b})$ to lie on a real projective variety of codimension 1. We shall use $X_j(\mathbf{x}, \mathbf{y})$ to denote this variety which is a subvariety of $\mathbb{R}^{2d+2}$. Now let $\pi_2 : \mathcal{A}_{I,J} \longrightarrow \mathbb{C}^d \times \mathbb{C}^d$ be the projection $(\mathbf{A}, \mathbf{b}, \mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}, \mathbf{y})$ onto the last two coordinates. Then for any $\mathbf{x}_0 \neq \mathbf{y}_0$ in $\mathbb{C}^d$ with $\mathrm{supp}(\mathbf{x}_0) \subset I$ and $\mathrm{supp}(\mathbf{y}_0) \subset J$, we have

$$\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\} = X_1(\mathbf{x}_0, \mathbf{y}_0) \times X_2(\mathbf{x}_0, \mathbf{y}_0) \times \ldots \times X_m(\mathbf{x}_0, \mathbf{y}_0) \times \{\mathbf{x}_0\} \times \{\mathbf{y}_0\}.$$

Hence the real dimension of $\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\}$ is

$$\dim_{\mathbb{R}} \left(\pi_2^{-1}\{(\mathbf{x}_0, \mathbf{y}_0)\}\right) = 2m(d+1) - m = 2md + m.$$

It follows that $\dim_{\mathbb{R}}(\mathcal{A}_{I,J}) \leq 2md + m + 2\#I + 2\#J \leq 2md + m + 4s$.

We now let $\pi_1 : \mathcal{A}_{I,J} \longrightarrow \mathbb{C}^{m\times(d+1)}$ be the projection $(\mathbf{A}, \mathbf{b}, \mathbf{x}, \mathbf{y}) \mapsto (\mathbf{A}, \mathbf{b})$. Since projections cannot increase the dimension of a variety, we know that

$$\dim_{\mathbb{R}} \left(\pi_1(\mathcal{A}_{I,J})\right) \leq 2md + m + 4s = 2m(d+1) + 4s - m < 2m(d+1).$$

However, $\pi_1(\mathcal{A}_{I,J})$ contains precisely those $(\mathbf{A}, \mathbf{b})$ in $\mathbb{C}^{m\times(d+1)}$ that are not (I, J)-sparse affine phase retrievable. Thus a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ is (I, J)-sparse affine phase retrievable.

Finally, there are only finitely many indices subsets I, J. Hence a generic $(\mathbf{A}, \mathbf{b}) \in \mathbb{C}^{m\times(d+1)}$ ($m \geq 4s + 1$) is (I, J)-sparse affine phase retrievable for any I, J with $\#I, \#J \leq s$. The theorem is proved. □

4. Stability and robustness of affine phase retrieval

Stability and robustness are important properties for affine phase retrieval. For the standard phase retrieval, stability and robustness have been studied in several papers, see [3,4,6,13]. In this section, we establish stability and robustness results for both maps $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ and $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$.

Theorem 4.1. *Assume that $(\mathbf{A}, \mathbf{b}) \in \mathbb{H}^{m\times(d+1)}$ is affine phase retrievable. Assume that $\Omega \subset \mathbb{H}^d$ is a compact set. Then there exist positive constants C_1, C_2, c_1, c_2 depending on $(\mathbf{A}, \mathbf{b})$ and Ω such that for any $\mathbf{x}, \mathbf{y} \in \Omega$, we have*

$$\frac{c_1}{1 + \|\mathbf{x}\| + \|\mathbf{y}\|}\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\| \leq C_1\|\mathbf{x} - \mathbf{y}\|, \tag{4.1}$$

$$c_2\|\mathbf{x} - \mathbf{y}\| \leq \left\|\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})\right\| \leq C_2(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\|\mathbf{x} - \mathbf{y}\|. \tag{4.2}$$

Proof. Write $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)^\top$ and $\mathbf{b} = (b_1, \dots, b_m)^\top$. We first establish the inequality for the map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})$, where we recall

$$\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \left(|\langle \mathbf{a}_1, \mathbf{x}\rangle + b_1|^2, \dots, |\langle \mathbf{a}_m, \mathbf{x}\rangle + b_m|^2\right).$$

Denote the matrix $(\mathbf{A}, \mathbf{b}) \in \mathbb{H}^{m\times(d+1)}$ by $(\mathbf{A}, \mathbf{b}) = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_m)^\top$, where $\tilde{\mathbf{a}}_j := \begin{pmatrix} \mathbf{a}_j \\ b_j \end{pmatrix}$, $j = 1, \dots, m$. Similarly we augment $\mathbf{x}, \mathbf{y} \in \mathbb{H}^d$ into $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathbb{H}^{d+1}$ by appending 1 to the $(d+1)$-th entry. Now we have

$$\begin{aligned} \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) &= (|\langle \mathbf{a}_1, \mathbf{x}\rangle + b_1|^2, \dots, |\langle \mathbf{a}_m, \mathbf{x}\rangle + b_m|^2) \\ &= (\mathrm{tr}(\tilde{\mathbf{a}}_1\tilde{\mathbf{a}}_1^*\tilde{\mathbf{x}}\tilde{\mathbf{x}}^*), \dots, \mathrm{tr}(\tilde{\mathbf{a}}_m\tilde{\mathbf{a}}_m^*\tilde{\mathbf{x}}\tilde{\mathbf{x}}^*)) =: \mathbf{T}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^*), \end{aligned}$$

where $\mathbf{T}$ is a linear transformation from $\mathbb{H}^{(d+1)\times(d+1)}$ to $\mathbb{R}^m$.

Let $X_\Omega = \{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* \in \mathbb{H}^{(d+1)\times(d+1)} : \ \mathbf{x} \in \Omega\}$,

$$\Theta_\Omega = \{\mathbf{S} \in \mathbb{H}^{(d+1)\times(d+1)} : \ \|\mathbf{S}\|_F = 1, \ t\mathbf{S} \in X_\Omega - X_\Omega \text{ for some } t > 0\}$$

and

$$\tilde{\Theta}_\Omega = \left\{ \mathbf{S} := \begin{pmatrix} \mathbf{z}\mathbf{w}^* + \mathbf{w}\mathbf{z}^* & \mathbf{z} \\ \mathbf{z}^* & 0 \end{pmatrix} : \ \mathbf{z} \in \mathbb{H}^d, \mathbf{w} \in (\Omega + \Omega)/2 \text{ and } \|\mathbf{S}\|_F = 1 \right\},$$

where $\|\cdot\|_F$ denotes the l^2-norm (Frobenius norm) of a matrix. Then

$$\Theta_\Omega \subset \tilde{\Theta}_\Omega \tag{4.3}$$

because

$$\mathbf{S} = t^{-1}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*) = \begin{pmatrix} \mathbf{z}\mathbf{w}^* + \mathbf{w}\mathbf{z}^* & \mathbf{z} \\ \mathbf{z}^* & 0 \end{pmatrix} \in \tilde{\Theta}_\Omega \quad \text{for all } \mathbf{S} \in \Theta_\Omega,$$

where the existence of $t > 0$, $\mathbf{x}, \mathbf{y} \in \Omega$ in the first equality follows from the definition of Θ_Ω and the second equality holds for $\mathbf{z} = (\mathbf{x} - \mathbf{y})/t$ and $\mathbf{w} = (\mathbf{x} + \mathbf{y})/2$.

For any $\mathbf{S} = \begin{pmatrix} \mathbf{z}\mathbf{w}^* + \mathbf{w}\mathbf{z}^* & \mathbf{z} \\ \mathbf{z}^* & 0 \end{pmatrix} \in \tilde{\Theta}_\Omega$, we have

$$\mathbf{T}(\mathbf{S}) = \mathbf{T}\begin{pmatrix} \mathbf{x}\mathbf{x}^* - \mathbf{y}\mathbf{y}^* & \mathbf{x} - \mathbf{y} \\ \mathbf{x}^* - \mathbf{y}^* & 0 \end{pmatrix} = \mathbf{T}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*) = \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y}) \neq 0 \tag{4.4}$$

by the affine phase retrievability of $(\mathbf{A}, \mathbf{b})$, where $\mathbf{x} = \mathbf{w} + \mathbf{z}/2$ and $\mathbf{y} = \mathbf{w} - \mathbf{z}/2$. Clearly $\tilde{\Theta}_\Omega$ is a compact set. This together with (4.4) implies that

$$c_2 := \inf_{\mathbf{S}\in\tilde{\Theta}_\Omega} \|\mathbf{T}(\mathbf{S})\| > 0. \tag{4.5}$$

Therefore

$$\begin{aligned}
&\left\|\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})\right\| = \|\mathbf{T}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*)\| \\
&\geq \Big(\inf_{\mathbf{S}\in\Theta_\Omega} \|\mathbf{T}(\mathbf{S})\|\Big)\|\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*\|_F \geq c_2\|\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*\|_F,
\end{aligned} \tag{4.6}$$

where the first equality follows from (4.4) and the last inequality holds by (4.3).

Now for the unit vector $\mathbf{e}_{d+1}$, we have

$$\|\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*\|_F \geq \|(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*)\mathbf{e}_{d+1}\| = \|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\| = \|\mathbf{x} - \mathbf{y}\|.$$

This, together with (4.5) and (4.6), establishes the lower bound in (4.2).

Because $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is linear in $X = \tilde{\mathbf{x}}\tilde{\mathbf{x}}^*$, we must also have

$$\left\|\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})\right\| \leq C_2'\|\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*\|_F.$$

However using the standard estimate, we have

$$\|\tilde{\mathbf{x}}\tilde{\mathbf{x}}^* - \tilde{\mathbf{y}}\tilde{\mathbf{y}}^*\|_F \leq \|\tilde{\mathbf{x}}\|\,\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\| + \|\tilde{\mathbf{y}}\|\,\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\| \leq 2(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\,\|\mathbf{x} - \mathbf{y}\|.$$

Here we have used the facts that $\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\| = \|\mathbf{x} - \mathbf{y}\|$ and $\|\tilde{\mathbf{x}}\| \leq 1 + \|\mathbf{x}\|$. Taking $C_2 = 2C_2'$ yields the upper bound in (4.2).

We now prove the inequalities for $\mathbf{M}_{\mathbf{A},\mathbf{b}}$. The upper bound in (4.1) is straightforward. Note that

$$\Big||\langle\mathbf{a}_j, \mathbf{x}\rangle + b_j| - |\langle\mathbf{a}_j, \mathbf{y}\rangle + b_j|\Big| \leq |\langle\mathbf{a}_j, \mathbf{x} - \mathbf{y}\rangle| \leq \|\mathbf{a}_j\|\,\|\mathbf{x} - \mathbf{y}\|.$$

It follows that

$$\|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\| \leq \Big(\sum_{j=1}^m \|\mathbf{a}_j\|\Big)\,\|\mathbf{x} - \mathbf{y}\|.$$

The upper bound in (4.1) thus follows by letting $C_1 = \sum_{j=1}^m \|\mathbf{a}_j\|$.

To prove the lower bound, we observe that

$$\begin{aligned}
&\Big||\langle\mathbf{a}_j, \mathbf{x}\rangle + b_j|^2 - |\langle\mathbf{a}_j, \mathbf{y}\rangle + b_j|^2\Big| \\
&\quad = \Big||\langle\mathbf{a}_j, \mathbf{x}\rangle + b_j| - |\langle\mathbf{a}_j, \mathbf{y}\rangle + b_j|\Big|(|\langle\mathbf{a}_j, \mathbf{x}\rangle + b_j| + |\langle\mathbf{a}_j, \mathbf{y}\rangle + b_j|) \\
&\quad \leq L(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\Big||\langle\mathbf{a}_j, \mathbf{x}\rangle + b_j| - |\langle\mathbf{a}_j, \mathbf{y}\rangle + b_j|\Big|,
\end{aligned}$$

where $L > 0$ is a constant depending only on $(\mathbf{A}, \mathbf{b})$. Hence

$$\left\|\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})\right\| \le L(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\, \|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\|\,.$$

It now follows from the lower bound $\|\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{y})\| \ge c_2\|\mathbf{x} - \mathbf{y}\|$ and setting $c_2 = c_1/L$ that

$$\|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\| \ge \frac{c_1}{1 + \|\mathbf{x}\| + \|\mathbf{y}\|}\, \|\mathbf{x} - \mathbf{y}\|.$$

The theorem is proved. $\square$

Proposition 4.1. *Neither* $\mathbf{M}_{\mathbf{A},\mathbf{b}}$ *nor* $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}$ *is bi-Lipschitz on* $\mathbb{H}^d$.

Proof. The map $\mathbf{M}^2_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is not bi-Lipschitz follows from the simple observation that it is quadratic in $\mathbf{x}$ (more precisely, in $\Re(\mathbf{x})$ and $\Im(\mathbf{x})$). No quadratic function can be bi-Lipschitz on the whole Euclidean space.

To see $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is not bi-Lipschitz, we fix a nonzero $\mathbf{x}_0 \in \mathbb{H}^d$. Take $\mathbf{x} = r\mathbf{x}_0$ and $\mathbf{y} = -r\mathbf{x}_0$, where $r > 0$. Note that

$$\|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\| = \Big(\sum_{j=1}^m (|r\langle \mathbf{a}_j, \mathbf{x}_0\rangle + b_j| - |r\langle \mathbf{a}_j, \mathbf{x}_0\rangle - b_j|)^2\Big)^{1/2}$$

and

$$\|\mathbf{x} - \mathbf{y}\| \;=\; 2r\|\mathbf{x}_0\|.$$

Then

$$\frac{\|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} = \frac{1}{2\|\mathbf{x}_0\|}\Big(\sum_{j=1}^m (|\langle \mathbf{a}_j, \mathbf{x}_0\rangle + b_j/r| - |\langle \mathbf{a}_j, \mathbf{x}_0\rangle - b_j/r|)^2\Big)^{1/2}. \tag{4.7}$$

A simple observation is that the right side of (4.7) tending to 0 as $r \to \infty$. Hence for any $\delta > 0$, we can choose r large enough so that

$$\frac{\|\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x}) - \mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \le \delta.$$

Thus $\mathbf{M}_{\mathbf{A},\mathbf{b}}(\mathbf{x})$ is not bi-Lipschitz. $\square$

References

[1] R. Balan, Stability of phase retrievable frames, Proc. SPIE 8858 (2013), http://dx.doi.org/10.1117/12.2026135, SPIE, Los Angeles.

[2] R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, Appl. Comput. Harmon. Anal. 20 (2006) 345–356.

[3] R. Balan, Y. Wang, Invertibility and robustness of phaseless reconstruction, Appl. Comput. Harmon. Anal. 38 (3) (2015) 469–488.
[4] A.S. Bandeira, J. Cahill, D.G. Mixon, A.A. Nelson, Saving phase: injectivity and stability for phase retrieval, Appl. Comput. Harmon. Anal. 37 (1) (2014) 106–125.
[5] B.G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames, Adv. Comput. Math. 41 (2) (2015) 317–331.
[6] J. Cahill, P.G. Casazza, I. Daubechies, Phase retrieval in infinite-dimensional Hilbert spaces, Trans. Am. Math. Soc. Ser. B 3 (2016) 63–76.
[7] E.J. Candès, Y. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion, SIAM J. Imaging Sci. 6 (1) (2013) 199–225.
[8] E.J. Candès, X.D. Li, M. Soltanolkotabi, Phase retrieval via Wirtinger flow: theory and algorithm, IEEE Trans. Inform. Theory 61 (4) (2015) 1985–2007.
[9] E.J. Candès, T. Strohmer, V. Voroninski, PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming, Comm. Pure Appl. Math. 66 (8) (2013) 1241–1274.
[10] Y. Chen, C. Cheng, Q. Sun, H.C. Wang, Phase retrieval of real signals in a principal shift-invariant space, available at http://arxiv.org/abs/1603.01592, 2016.
[11] A. Conca, D. Edidin, M. Hering, C. Vinzant, Algebraic characterization of injectivity in phase retrieval, Appl. Comput. Harmon. Anal. 38 (2) (2015) 346–356.
[12] M. Fickus, D.G. Mixon, A.A. Nelson, Y. Wang, Phase retrieval from very few measurements, Linear Algebra Appl. 449 (2014) 475–499.
[13] B. Gao, Y. Wang, Z.Q. Xu, Stable signal recovery from phaseless measurements, J. Fourier Anal. Appl. 22 (4) (2016) 787–808.
[14] J. Harris, Algebraic Geometry, first ed., Springer-Verlag, New York, 1992.
[15] T. Heinosaari, L. Mazzarella, M.M. Wolf, Quantum tomography under prior information, Comm. Math. Phys. 318 (2) (2013) 355–374.
[16] M. Liebling, T. Blu, E. Cuche, P. Marquet, C.D. Depeursinge, M. Unser, Local amplitude and phase retrieval method for digital holography applied to microscopy, Proc. SPIE 5143 (2003) 210–214.
[17] P. Netrapalli, P. Jain, S. Sanghavi, Phase retrieval using alternating minimization, IEEE Trans. Signal Process. 63 (18) (2015) 4814–4826.
[18] C. Vinzant, A small frame and a certificate of its injectivity, in: Sampling Theory and Applications (SampTA) Conference Proceedings, 2015, pp. 197–200.
[19] Y. Wang, Z.Q. Xu, Phase retrieval for sparse signals, Appl. Comput. Harmon. Anal. 37 (3) (2014) 531–544.
[20] Y. Wang, Z.Q. Xu, Generalized phase retrieval: measurement number, matrix recovery and beyond, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.09.003.