Phase retrieval from power spectra of masked signals

Afonso S. Bandeira and Yutong Chen†

Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, USA

†Corresponding author: yutong@math.princeton.edu ajsb@math.princeton.edu

and

Dustin G. Mixon

Department of Mathematics and Statistics, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433, USA
dustin.mixon@afit.edu

[Received on 25 June 2013; revised on 31 March 2014; accepted on 1 April 2014]

In diffraction imaging, one is tasked with reconstructing a signal from its power spectrum. To resolve the ambiguity in this inverse problem, one might invoke prior knowledge about the signal, but phase retrieval algorithms in this vein have found limited success. One alternative is to create redundancy in the measurement process by illuminating the signal multiple times, distorting the signal each time with a different mask. Despite several recent advances in phase retrieval, the community has yet to construct an ensemble of masks which uniquely determines all signals and admits an efficient reconstruction algorithm. In this paper, we leverage the recently proposed polarization method to construct such an ensemble. First, we construct four explicit masks which enable polarization recovery of any signal with non-vanishing Fourier transform, and then we construct $\Theta(\log M)$ random masks which, with high probability, simultaneously enable polarization recovery of any signal whatsoever. We also present numerical simulations to illustrate the stability of the polarization method in this setting. In comparison to a state-of-the-art phase retrieval algorithm known as PhaseLift, we find that polarization is much faster with comparable stability.

Keywords: phase retrieval, diffraction imaging, polarization, angular synchronization.

1. Introduction

In many applications, one wishes to reconstruct a signal from the magnitudes of its Fourier coefficients. This problem is known as phase retrieval, and it has been instrumental to many important scientific advances, including the Nobel Prize–winning work that leveraged X-ray diffraction to establish the double helix structure of DNA [29]. Of course, given only the magnitudes of a signal’s Fourier coefficients, one does not have enough information to recover the signal—while the Fourier transform is injective, its point-wise absolute value is not. As such, one is inclined to use a priori knowledge of the signal, and hope it is then uniquely determined by the Fourier magnitudes. For example, to deduce the structure of DNA, Watson and Crick [29] applied certain chemical assumptions, along with a knowledge of van der Waals interactions between atoms.

In order to image more exotic molecules, such assumptions are difficult to apply, and so there has been quite a bit of work attempting to exploit more general assumptions (e.g. positivity or support constraints). To account for this sort of prior information, the most popular phase retrieval algorithms are modifications of Gerchberg and Saxton’s original approach [19], which alternates between the time
and frequency domains, iteratively correcting the current guess by imposing time-domain assumptions or scaling Fourier coefficients to match the measured data. Marchesini \cite{24} surveys and compares the various modifications, but they all have a tendency to stall in local minima.

To address this issue, Candès et al. \cite{10} proposed an alternative methodology whereby non-uniqueness is overcome not by prior information, but by additional illuminations. For each illumination, a different mask (or grating) is used to distort the appearance of the object in question; in mathematical parlance, each mask acts as a multiplication operator on the desired signal before the Fourier transform. To be explicit, let $F^*: \ell(\mathbb{Z}_M) \to \ell(\mathbb{Z}_M)$ denote the discrete Fourier transform (DFT) defined by

$$(F^*x)(m) := \frac{1}{M} \sum_{m' \in \mathbb{Z}_M} x(m') e^{-2\pi imm'/M} \quad \forall m \in \mathbb{Z}_M.$$ 

In this paper, we follow the lead of \cite{10} to find masks $\{D_k\}_{k=0}^{K-1}$ (i.e. multiplication operators on $\ell(\mathbb{Z}_M) = \mathbb{C}^M$, also viewed as diagonal matrices) such that every signal $x \in \ell(\mathbb{Z}_M)$ can be efficiently reconstructed, up to a global phase factor, from measurements of the form

$$\{|F^*D_k^*x|^2\}_{k=0}^{K-1} = \{|(D_kF)^*x|^2\}_{k=0}^{K-1} = \{|\langle x, D_k f_m \rangle|^2\}_{k=0}^{K-1},$$

where $f_m$ denotes the complex sinusoid $(1/M) e^{2\pi i mm'/M}_{m' \in \mathbb{Z}_M}$. Specifically, we will design the masks in such a way that allows for a new reconstruction technique to work, namely the polarization technique introduced in \cite{1}. As established in \cite{1}, every signal $x$ can be stably reconstructed from $|\langle x, \varphi_i \rangle|^2$ if the measurement vectors $\varphi_i$ are constructed using a particular random process. In this paper, we show how to fulfill the design criteria that measurement vectors be of the form $D_k f_m$ while still allowing for polarization recovery. If anything, this demonstrates that polarization is flexible enough to provably accommodate certain measurement design requirements, which are bound to arise from various applications.

In the next section, we briefly review the main ideas behind the polarization technique, and we show how to apply them with Fourier masks. There, we present our two main results:

(i) Theorem 2.1 constructs four explicit masks which enable polarization recovery of any signal with non-vanishing Fourier transform.

(ii) Theorem 2.2 constructs $\Theta(\log M)$ random masks which, with high probability, simultaneously enable polarization recovery of any signal whatsoever.

We compare these results to guarantees for a state-of-the-art phase retrieval algorithm known as PhaseLift \cite{10–13}; specifically, we compare with \cite[Theorem 3.1]{10} and to \cite[Theorem 1.1]{12}. Section 2 also reduces the proof of Theorem 2.2 to a problem in additive combinatorics: Find a small subset of $\mathbb{Z}_M$ with small Fourier bias. In Section 3, we use concentration-of-measure arguments to solve this problem. We also demonstrate that $\Omega(\log M)$ masks are necessary to simultaneously enable polarization recovery of all signals, indicating that our analysis is tight. Next in Section 4, we compare the performance of polarization to PhaseLift; numerical simulations illustrate that polarization provides a comparable amount of stability with Fourier masks, while being orders of magnitude faster. We conclude in Section 5 with some remarks.
2. How to polarize Fourier masks

The main contribution of this paper is that the ideas in [1] can be leveraged to uniquely determine signals with masked Fourier transforms. In this section, we summarize these ideas and explain how we apply them. The following subsection introduces the main ideas of polarization and constructs four explicit masks which yield almost injective intensity measurements (that is, almost every signal is uniquely determined up to global phase [17]). Section 2.2 then dispels the non-generic case by leveraging ideas from spectral graph theory and additive combinatorics, ultimately increasing the number of exposures to \( \Theta(\log M) \).

2.1 Phase retrieval of almost every signal

In [1], signal \( x \) is measured with two ensembles of vectors \( \Phi_V, \Phi_E \subseteq \mathbb{C}^M \). Here, we require that \( \Phi_V \) be a spanning set, and the vectors in \( \Phi_E \) will be polarized combinations of vectors in \( \Phi_V \). More explicitly, define a graph \( G = (V, E) \) whose vertices \( i \in V \) index the measurement vectors \( \varphi_i \in \Phi_V \). Then each edge \((i, j) \in E\) corresponds to three members of \( \Phi_E \), namely \( \{\varphi_i + \omega^r \varphi_j\}_{r=0}^2 \), where \( \omega = e^{2\pi i/3} \). As a big-picture summary, polarization combines the intensity measurements from \( \Phi_E \) using the polarization identity to determine relative phases between measurements from \( \Phi_V \), from which the inner products with \( \Phi_V \) can be determined up to a global phase factor, and then the signal can be recovered (up to global phase) by solving a linear system. Indeed, a version of the polarization identity gives that

\[
\frac{\langle x, \varphi_i \rangle \langle x, \varphi_j \rangle}{\langle x, \varphi_e \rangle} = \frac{1}{3} \sum_{r=0}^{2} \omega^r \langle x, \varphi_i \rangle + \omega^{-r} \langle x, \varphi_j \rangle \right|^2 = \frac{1}{3} \sum_{r=0}^{2} \omega^r \langle x, \varphi_i + \omega^r \varphi_j \rangle \right|^2, \tag{2.1}
\]

and so the relative phase between \( \langle x, \varphi_i \rangle \) and \( \langle x, \varphi_j \rangle \) is calculated by normalizing this quantity (provided it is non-zero). In this subsection, we are only concerned with reconstructing almost every signal, and so we will focus on reconstructing the class of signals \( x \) which are not orthogonal to any vector in \( \Phi_V \) (this guarantees that the relative phases above are well defined). In order to reconstruct these vectors, it suffices for the graph \( G \) to be connected, as this will enable the propagation of relative phases to determine the vertex phases up to a global phase factor.

At this point, we describe how to design \( \Phi_V \) and \( \Phi_E \) using masked Fourier transforms. Here, we will take \( \Phi_V := \{f_m\}_{m \in \mathbb{Z}_M} \), which is certainly a spanning set for \( \mathbb{C}^M \). Note that we can implement measurements with \( \Phi_V \) by simply applying a DFT (i.e. without a mask). Next, let \( G \) be the cyclic graph over \( \mathbb{Z}_M \) such that \( i \leftrightarrow j \) precisely when \( j - i = \pm 1 \). Then, for each edge \((i, i + 1) \) in \( G \), \( \Phi_E \) will receive three measurement vectors \( \{f_i + \omega^r f_{i+1}\}_{r=0}^2 \). Using the modulation operator \( E = \text{diag}(e^{2\pi im/M})_{m=0}^{M-1} \), we can express these members of \( \Phi_E \) in terms of masks:

\[
f_i + \omega^r f_{i+1} = f_i + \omega^r Ef_i = (I + \omega^r E)f_i. \tag{2.2}
\]

As such, measurements with \( \Phi_E \) can be implemented by applying masked DFTs with masks \( \{I + \omega^r E\}_{r=0}^2 \). To summarize, we have the following theorem.

**Theorem 2.1** Take \( \omega = e^{2\pi i/3} \) and let \( E = \text{diag}(e^{2\pi im/M})_{m=0}^{M-1} \) denote the modulation operator. Then the four masks \( \{I\} \cup \{I + \omega^r E\}_{r=0}^2 \) yield almost injective intensity measurements. More explicitly, every \( x \in \mathbb{C}^M \) with non-vanishing Fourier transform can be recovered up to global phase from \( \{|F^* x|^2\} \cup \{|F^*(I + \omega^r E)^* x|^2\}_{r=0}^2 \) using polarization.
At this point, it is appropriate to compare this result to \cite[Theorem 3.1]{10}, which gives that every \( x \in \mathbb{C}^M \) with non-vanishing Fourier transform is uniquely determined (up to global phase) by intensity measurements with three particular masked DFTs. Specifically, the masks they use are very similar to the ones we define above: \( \{I, I + E^s, I + iE^s\} \) for some \( s \) which is coprime with \( M \) (e.g. \( s = 1 \)). The result also guarantees that PhaseLift will recover \( x \) up to global phase whenever it returns a rank-1 solution. In comparison, Theorem 2.1 guarantees the efficient recovery of any \( x \in \mathbb{C}^M \) with non-vanishing Fourier transform (caveat-free) at the price of a fourth masked exposure.

\subsection{Phase retrieval of every signal}

The previous subsection briefly outlined the main ideas behind polarization, culminating in the construction of four masks which yield almost injective intensity measurements. This subsection will describe how to construct masks which yield injectivity; as we will see, \( \Theta(\log M) \) masks are necessary and sufficient for such masks to enable recovery with polarization. To achieve recovery in the previous subsection, we assumed that the signal \( x \) was not orthogonal to any of the vertex measurement vectors \( \Phi^V \), which is true for almost every \( x \in \mathbb{C}^M \). The purpose of this assumption was to ensure that the relative phases, calculated from (2.1), were well defined. In the case where \( x \) is orthogonal to some \( \psi_i \), note that all relative phases corresponding to edges incident to vertex \( i \) are not well defined. As such, relative phase information cannot propagate through vertex \( i \), indicating that orthogonality with \( \psi_i \) has the effect of deleting the vertex \( i \) (and its edges) from the graph \( G \).

To limit this damage to the graph, the authors in \cite{1} design \( \Phi^V \) to have the property that every subcollection of size \( M \) spans \( \mathbb{C}^M \); such ensembles are called \textit{full spark frames}, of which there are several known constructions \cite{2,25}. Indeed, this property ensures that \( x \) is orthogonal to at most \( M - 1 \) vertex vectors, meaning at most \( M - 1 \) vertices can be deleted from the graph. Furthermore, for any remaining connected component \( S \subseteq V \) of size \( \geq M \), we can propagate the relative phases to determine \( \{(x, \psi_i)\}_{i \in S} \) up to a global phase factor, and since \( \{\psi_i\}_{i \in S} \) necessarily span \( \mathbb{C}^M \), we can determine \( x \) from \( \{(x, \psi_i)\}_{i \in S} \) by solving the linear system. Interestingly, there will always be a connected component of size \( \geq M \) after the removal of any \( M - 1 \) vertices, provided the graph \( G \) is regular with a sufficiently large \textit{spectral gap}. To be clear, the spectral gap of a \( d \)-regular graph \( G \) is determined by the eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \) of its adjacency matrix:

\[
\text{spg}(G) = \frac{1}{d} \left( \lambda_1 - \max_{i \neq 1} |\lambda_i| \right),
\]

and the existence of the requisite connected component is given by the following classical result in spectral graph theory (e.g. see \cite[Lemma 5.2]{21}).

\textbf{Lemma 2.1} Consider a \( d \)-regular graph \( G \) of \( n \) vertices. For all \( \varepsilon \leq \text{spg}(G)/6 \), removing any \( \varepsilon dn \) edges from \( G \) results in a connected component of size \( \geq (1 - 2\varepsilon / \text{spg}(G))n \).

In summary, in order to apply the polarization trick in the setting of masked DFTs, it suffices to (i) construct a full spark ensemble \( \Phi^V \) with masked DFTs, and (ii) construct a graph \( G \) with a sufficiently large spectral gap and whose corresponding edge vectors \( \Phi^E \) can also be implemented with masked DFTs. We quickly address (i) with the following result:

\textbf{Lemma 2.2} For each \( k \in \{0, \ldots, K - 1\} \), choose some non-zero \( \alpha_k \in \mathbb{C} \) and define \( D_k := \text{diag}[\alpha_k^m]_{m=0}^{M-1} \). Let \( f_m \) denote the complex sinusoid \( \{(1/M)e^{2\pi i m m'}/M\}_{m' \in \mathbb{Z}_M} \). Then \( \{D_k f_m\}_{k=0, m=0}^{K-1,M-1} \) is full spark if and only if \( \alpha_k \alpha_k^{-1} \) is not an \( m \)th root of unity for any pair of distinct \( k, k' \in \{0, \ldots, K - 1\} \).
Proof. The $M \times KM$ matrix whose columns are $D_kf_m$ has Vandermonde form, as does each of its $M \times M$ submatrices. These submatrices are all invertible precisely when their determinants are non-zero, that is, when \{\alpha_k e^{2\pi im/M} \}_{k=0, m=0}^{K-1, M-1} \text{ are distinct} (by the Vandermonde determinant formula). The lemma immediately follows. 

The condition above is satisfied with probability 1 if the $\alpha_k$’s are drawn uniformly at random from the complex unit circle. For a deterministic alternative, it suffices to take $\alpha_k = e^{2\pi i k/M}$. Now that we have established how to construct masks $D_k$ such that $\Phi = \{D_kf_m\}_{k=0, m=0}^{K-1, M-1}$ is full spark, we turn to solving (ii). Before describing our construction of $G$, we first motivate it by illustrating how polarized combinations of our vertex vectors can be expressed using masked DFTs (this is a generalization of (2.2)):

**Lemma 2.3** Take $\omega = e^{2\pi i/3}$, let $E = \text{diag}(e^{2\pi i m/M})_{m=0}^{M-1}$ denote the modulation operator and consider $\Phi = \{D_kf_m\}_{k=0, m=0}^{K-1, M-1}$, as defined in Lemma 2.2. Then

$$D_kf_m + \omega^r D_kf_{m'} = (D_k + \omega^r E^{m'-m} D_k)f_m$$

for all $k, k' \in \{0, \ldots, K-1\}$ and $m, m' \in \mathbb{Z}_M$.

**Proof.** Consider the $\ell$th entry of $D_kf_m'$:

$$(D_kf_m')(\ell) = \alpha^\ell \cdot \frac{1}{M} e^{2\pi i \ell/M} = e^{2\pi i (m'-m)\ell/M} \cdot \alpha^\ell \cdot \frac{1}{M} e^{2\pi im\ell/M} = (E^{m'-m} D_kf_m)(\ell).$$

Then $D_kf_m + \omega^r D_kf_{m'} = D_kf_m + \omega^r E^{m'-m} D_kf_m = (D_k + \omega^r E^{m'-m} D_k)f_m$. 

When implementing the modulation trick of Lemma 2.3 using DFTs, we will have to fix $m' - m$ (so as to apply a fixed mask) and let $m$ vary (since we will mask an entire DFT). As such, we intend to use auxiliary masks of the form $\{D_k + \omega^r E^m D_k\}_{m=0}^{M-1}$, thereby drawing an edge between every $(k, m)$ and $(k', m')$ such that $m' - m = a \text{ mod } M$. This informs our decision of how to construct the graph $G$, since it enables a masked-DFT implementation.

**Definition 2.1** We construct $G$ as follows: Pick $A \subseteq \mathbb{Z}_M$ such that $0 \not\in A$ and $A = -A$. Then $(k, m)$ and $(k', m')$ are adjacent precisely when $m' - m \in A$.

Note that, for each $a \in A$, the corresponding edges between $\{(k, m)\}_{m \in \mathbb{Z}_M}$ and $\{(k', m')\}_{m' \in \mathbb{Z}_M}$ are implemented with the masks $\{D_k + \omega^r E^a D_k\}_{r=0}^{K-1}$. As such, $A$ is the set of modulations used to combine the $K$ original masks $D_k$ into $3(K+1)/2 |A|$ auxiliary masks. Also note that $G$ has no loops since $0 \not\in A$, $G$ is not directed since $A = -A$ and, furthermore, every vertex has degree $K|A|$, so $G$ is regular. It remains to show that $G$ has a sufficiently large spectral gap. To simplify this analysis, we first relate the spectral gap to a fundamental concept in additive combinatorics called the **Fourier bias**. As in [27], take $A \subseteq \mathbb{Z}_M$ and let $1_A$ denote the characteristic function of $A$. Then the Fourier bias of $A$ is defined as follows:

$$\|A\|_u := \max_{m \equiv 0} |(F^* 1_A)(m)|.$$

Fourier bias is used in additive combinatorics to measure pseudorandomness; here, the main idea is that correlation with any complex sinusoid indicates regularity, which is not typically exhibited by random sets. As indicated earlier, the spectral gap of $G$ is intimately related to the Fourier bias of $A$. 


Lemma 2.4 Consider the graph $G$ defined in Definition 2.1. The spectral gap of $G$ is
\[
\text{spg}(G) = 1 - \frac{M}{|A|} \|A\|_u,
\]
where $\| \cdot \|_u$ denotes Fourier bias.

Proof. To determine the spectral gap, we first determine the adjacency matrix $W$ of $G$. For any pair $k,k' \in \{0, \ldots, K - 1\}$, the adjacency rule for $(k,m)$ and $(k',m')$ is whether $m' - m \in A$. As such, the $(k,k')$th block of $W$ is circulant, with each row being a translation of $\mathbf{1}_A$. This gives the expression $W = J \otimes \text{circ}(\mathbf{1}_A)$, where $J$ is the $K \times K$ all-ones matrix and $\otimes$ denotes the Kronecker product. One useful property of the Kronecker product is that its eigenvalues are products of eigenvalues:
\[
\lambda_{i,j}(W) = \lambda_i(J)\lambda_j(\text{circ}(\mathbf{1}_A)).
\]

Here, we note that the only non-trivial eigenvalue of $J$ is $K$, and the eigenvalues of $\text{circ}(\mathbf{1}_A)$ are the entries of the Fourier transform $MF^*\mathbf{1}_A$. As such, the non-trivial eigenvalues of $W$ are given by
\[
\lambda_{(1,m)}(W) = \lambda_1(J)\lambda_m(\text{circ}(\mathbf{1}_A)) = KM(F^*\mathbf{1}_A)(m) = K \sum_{m' \in \mathbb{Z}_M} \mathbf{1}_A(m') e^{-2\pi i mn'/M}.
\]

By the triangle inequality, we then have
\[
|\lambda_{(1,m)}(W)| = K \sum_{m' \in \mathbb{Z}_M} |\mathbf{1}_A(m') e^{-2\pi i mn'/M}| \leq K \sum_{m' \in \mathbb{Z}_M} |\mathbf{1}_A(m')| = K|A|
\]
with equality when $m = 0$. This means the largest eigenvalue of $W$ is $\lambda_1 = K|A|$, which corresponds to the all-ones eigenvector, and so the spectral gap of $G$ is
\[
\text{spg}(G) = \frac{1}{d} \left( \lambda_1 - \max_{i \neq 1} |\lambda_i| \right) = \frac{1}{K|A|} \left( K|A| - \max_{m \neq 0} |KM(F^*\mathbf{1}_A)(m)| \right) = 1 - \frac{M}{|A|} \|A\|_u,
\]
as claimed. \qed

At this point, we have reduced the problem of finding injective Fourier masks to finding a small set $A \subseteq \mathbb{Z}_M$ of sufficiently small Fourier bias (i.e. a problem in additive combinatorics). In the next section, we use the following random process to construct this set: Draw $B \subseteq \mathbb{Z}_M$ at random and take $A := B \cup (-B) \setminus \{0\}$. With the appropriate choice of distribution, we construct $A$ of size $\Theta(\log M)$ and Fourier bias $O(\log M / M)$, thereby producing a graph $G$ with a spectral gap that is bounded away from zero (see Theorem 3.1). We also show that the spectral gap is bounded away from zero only if $A$ has size $\Omega(\log M)$, thereby demonstrating the optimality of our analysis (see Theorem 3.2). Overall, letting $K$ be a sufficiently large constant, the techniques of this paper use a total of $K + 3(K+1)|A| = \Theta(\log M)$ Fourier masks to uniquely determine any signal.

Theorem 2.2 Take $K = 12$, $c = 144$, suppose the entries of $\mathbf{1}_B$ are independent, identical Bernoulli random variables with mean $c \log M / M$ and take $A := B \cup (-B) \setminus \{0\}$. Then with high probability, the
as described in Lemmas 2.2 and 2.3, yield injective intensity measurements; in particular, if $M \geq 7 \cdot 10^7$, then the success probability is $\geq 1 - 5M^{-1/8}$. Furthermore, in this event, every $x \in \mathbb{C}^M$ can be recovered up to global phase from $\{ |F^* D^* x|^2 \}_{D \in \mathcal{D}}$ using polarization.

In the time since this result was originally proved, Candès et al. [12] proved a competing result which guarantees PhaseLift recovery from $\Omega(\log^4 M)$ masked exposures. However, unlike Theorem 2.2, the main result (Theorem 1.1) in [12] does not guarantee injectivity. Instead, Candès et al. ensure that, for any fixed signal $x \in \mathbb{C}^M$, PhaseLift will recover $x$ from $\Omega(\log^4 M)$ randomly masked exposures with high probability. This distinction between uniform and non-uniform recovery, coupled with the discrepancy in log factors, establishes the theoretical advantages of polarization (at least for now). Note that, however, Phaselift has the advantage that it can be used on any set of masks, albeit without much theoretical understanding of its performance for general masks. On the other hand, the polarization method studied here strongly exploits specific algebraic structure in the particular choice of masks.

3. Small sets with small Fourier bias

In the previous section, we reduced the main problem of this paper to one in additive combinatorics: Find a small set $A \subset \mathbb{Z}_M$ with small Fourier bias $\|A\|_u = \max_{m \neq 0} |(F^* 1_A)(m)|$. In our case, we want $A$ to satisfy two additional properties: $0 \not\in A$ and $A = -A$. To account for these, we will approach this problem by first drawing $B \subset \mathbb{Z}_M$ at random, and then taking $A := B \cup (-B) \setminus \{0\}$. We start with the following lemma.

**Lemma 3.1** Suppose the entries of $1_B$ are independent, identical Bernoulli random variables with mean $c \log M / M$. Then the following simultaneously hold with high probability:

(i) $B \cap (-B) = \emptyset$;

(ii) $0 \not\in B$;

(iii) $\frac{1}{2} c \log M \leq |B| \leq \frac{3}{2} c \log M$.

**Proof.** First, a union bound gives

$$\Pr(B \cap (-B) = \emptyset) = \Pr(\exists m \in \mathbb{Z}_M \text{ s.t. } m, -m \in B) \leq M \Pr(m \in B) \Pr(-m \in B) = \frac{c^2 \log^2 M}{M}.$$ 

Next, we have $\Pr(0 \in B) = c \log M / M$. For (iii), we apply a multiplicative form of the Chernoff bound (see (6) and (7) in [20]): Let $X$ be a sum of independent 0-1 random variables. Then

$$\Pr(|X - \mathbb{E}[X]| \geq \delta \mathbb{E}[X]) \leq 2 e^{-\delta^2 \mathbb{E}[X]/3}$$

where $\delta$ is a constant.
for any choice $0 < \delta < 1$. Here, we let $X := \sum_{m \in \mathbb{Z}_M} 1_B(m) = |B|$ and take $\delta = \frac{1}{2}$. Then
\[
\Pr(|B| - c \log M| \geq \frac{1}{2} c \log M) \leq 2 e^{-(c \log M)/12} = 2M^{-c/12}.
\]
The result then follows from a union bound. \hfill \Box

Note that, in the event of Lemma 3.1, we have $1_A = 1_B + 1_{-B}$, and so
\[
\|A\|_u = \max_{m \neq 0} |(F^* 1_A)(m)| \leq \max_{m \neq 0} (|(F^* 1_B)(m)| + |(F^* 1_{-B})(m)|)
\leq \max_{m \neq 0} |(F^* 1_B)(m)| + \max_{m \neq 0} |(F^* 1_{-B})(m)| = 2\|B\|_u,
\]
where the last step applied a complex conjugate to $(F^* 1_{-B})(m)$. As such, it suffices to show that random sets $B$ have small Fourier bias.

**Lemma 3.2** Pick $c \geq 4$ and suppose the entries of $1_B$ are independent, identical Bernoulli random variables with mean $c \log M/M$. Then $\|B\|_u \leq 3\sqrt{c} \cdot (\log M/M)$ with high probability.

To prove this lemma, we will apply the following version of the Chernoff bound.

**Lemma 3.3** Assume that $\{Z_i\}_{i=1}^n$ are jointly independent complex random variables where $|Z_i - \mathbb{E}[Z_i]| \leq 1$ for all $i$. Take $Z := \sum_{i=1}^n Z_i$ and define $\sigma := \sqrt{\text{Var}(Z)}$. Then
\[
\Pr(|Z - \mathbb{E}[Z]| \geq t\sigma) \leq 4 \max\{e^{-t^2/8}, e^{-t^2/2}\sqrt{3}\}
\]
for any $t > 0$.

**Proof.** Let $X$ and $Y$ denote the real and imaginary parts of $Z$. Then a union bound gives
\[
\Pr(|Z - \mathbb{E}[Z]| \geq t\sigma) = \Pr(|X - \mathbb{E}[X]|^2 + |Y - \mathbb{E}[Y]|^2 \geq t^2 \sigma^2)
\leq \Pr \left( |X - \mathbb{E}[X]| \geq \frac{t\sigma}{\sqrt{2}} \right) + \Pr \left( |Y - \mathbb{E}[Y]| \geq \frac{t\sigma}{\sqrt{2}} \right).
\]
Define $\sigma_X := \sqrt{\text{Var}(X)}$ and $\sigma_Y := \sqrt{\text{Var}(Y)}$. Since $\text{Var}(X) + \text{Var}(Y) = \text{Var}(Z)$, we then have $\sigma_X \leq \sigma$ and $\sigma_Y \leq \sigma$, and so
\[
\Pr(|Z - \mathbb{E}[Z]| \geq t\sigma) \leq \Pr \left( |X - \mathbb{E}[X]| \geq \frac{t\sigma_X}{\sqrt{2}} \right) + \Pr \left( |Y - \mathbb{E}[Y]| \geq \frac{t\sigma_Y}{\sqrt{2}} \right).
\]
The result then follows from applying [27, Theorem 1.8] to both terms of the right-hand side. \hfill \Box

**Proof of Lemma 3.2.** Fix $m \in \mathbb{Z}_M$, $m \neq 0$, and take $Z_{m'} := \frac{1}{2} 1_B(m') e^{-2\pi i mm'/M}$ for each $m' \in \mathbb{Z}_M$. Then $\{Z_{m'}\}_{m' \in \mathbb{Z}_M}$ are jointly independent with
\[
|Z_{m'} - \mathbb{E}[Z_{m'}]| \leq |Z_{m'}| + |\mathbb{E}[Z_{m'}]| \leq 1.
\]
As such, $\{Z_{m'}\}_{m' \in \mathbb{Z}_M}$ satisfy the conditions of Lemma 3.3, and so we now seek the expected value and variance of
\[
Z := \sum_{m' \in \mathbb{Z}_M} Z_{m'} = \frac{1}{2} \sum_{m' \in \mathbb{Z}_M} 1_B(m') e^{-2\pi i mm'/M}.
\]
Take \( p := c \log M / M \) to be the probability that \( 1_B(m') \) is 1. Then by linearity of expectation, we have
\[
\mathbb{E}[Z] = \frac{1}{2} \sum_{m' \in \mathbb{Z}_M} \mathbb{E}[1_B(m')] e^{-2\pi i m' / M} = \frac{p}{2} \sum_{m' \in \mathbb{Z}_M} e^{-2\pi i m' / M}
\]
and so \( \mathbb{E}[Z] = 0 \) by the geometric sum formula (recall that \( m \neq 0 \) by choice). Next, we compute \( \text{Var}(Z) \):
\[
\text{Var}(Z) = \text{Var} \left( \frac{1}{2} \sum_{m' \in \mathbb{Z}_M} 1_B(m') e^{-2\pi i m' / M} \right) = \frac{1}{4} \sum_{m' \in \mathbb{Z}_M} \text{Var}(1_B(m')) = \frac{1}{4} M p (1 - p).
\]

Having calculated the expected value and variance of \( Z \), we are now ready to use Lemma 3.3 (and the fact that \( Z = (M/2) (F \ast 1_B)(m) \)). A union bound gives
\[
\Pr \left( \|B\|_u \geq \frac{2t \sigma}{M} \right) \leq (M - 1) \Pr \left( \| (F \ast 1_B)(m) \| \geq \frac{2t \sigma}{M} \right) = (M - 1) \Pr (|Z - \mathbb{E}[Z]| \geq t \sigma) \leq 4(M - 1) \max \{ e^{-t^2/8}, e^{-t \sigma / \sqrt{2}} \}.
\]

We select \( t = (\sqrt{2}c(1 + \varepsilon) \log M) / \sigma \) and simplify the exponents in our probability bound:
\[
\frac{t^2}{8} = \frac{c(1 + \varepsilon) \log^2 M}{M p (1 - p)} \geq \frac{c(1 + \varepsilon) \log^2 M}{M p} = (1 + \varepsilon) \log M, \quad \frac{t \sigma}{2 \sqrt{2}} = \frac{\sqrt{c(1 + \varepsilon)}}{2} \log M.
\]

With this choice of \( t \), we continue as follows:
\[
\Pr \left( \|B\|_u \geq \sqrt{8c(1 + \varepsilon) \cdot \frac{\log M}{M}} \right) \leq 4(M - 1) \max \{ e^{-(1 + \varepsilon) \log M}, e^{-(\sqrt{c(1 + \varepsilon)/2}) \log M} \} \leq 4 \max \{ M^{-\varepsilon}, M^{1 - \sqrt{c(1 + \varepsilon)/2}} \}.
\]

Taking \( \varepsilon = \frac{1}{8} \), then since \( c \geq 4 \), we therefore have that \( \|B\|_u \leq 3 \sqrt{c} \cdot (\log M / M) \) with high probability. \( \square \)

We can now combine Lemmas 3.1 and 3.2 to produce a graph \( G \) of the form in Definition 2.1 with a large spectral gap.

**Theorem 3.1** Pick \( c \geq 4 \) and suppose the entries of \( 1_B \) are independent, identical Bernoulli random variables with mean \( c \log M / M \). Take \( A := B \cup (-B) \setminus \{0\} \) and define \( G \) according to Definition 2.1. Then
\[
\text{spg}(G) \geq 1 - \frac{6}{\sqrt{c}}
\]
in the events of Lemmas 3.1 and 3.2.
Proof. By parts (i) and (ii) of Lemma 3.1, we have $|A| = 2|B|$, and furthermore by (3.1), $\|A\|_u \leq 2\|B\|_u$. Starting with Lemma 2.4, we then have

$$\text{spg}(G) = 1 - \frac{M}{|A|}\|A\|_u \geq 1 - \frac{M}{|B|}\|B\|_u \geq 1 - \frac{M}{(1/2)c\log M} \cdot 3\sqrt{c} \cdot \log M \cdot M = 1 - \frac{6}{\sqrt{c}},$$

where the second inequality applies Lemmas 3.1(iii) and 3.2. □

We now apply Lemma 2.1 to show that $O(\log M)$ Fourier masks suffice for injectivity.

**Proof of Theorem 2.2.** We will assume that the events of Lemmas 3.1 and 3.2 hold simultaneously, as they will with probability $\geq 1 - 5M^{-1/8}$ when $c = 144$ and $M \geq 7 \cdot 10^7$. This assumption will enable us to apply Theorem 3.1 later.

Note that $G$ has $n = KM$ vertices and is $d$-regular with $d = K|A|$. We need to be robust to the removal of any $M - 1$ vertices, which in turn removes $d(M - 1)$ edges, and so it suffices to be robust to the removal of any $dM$ edges. As such, we observe Lemma 2.1 and take $\varepsilon$ to satisfy

$$\frac{1}{K} = \frac{dM}{dn} = \varepsilon \leq \frac{\text{spg}(G)}{6}. \quad (3.2)$$

We also want a connected component of size $M$ after the removal of these edges, and so, by Lemma 2.1, it suffices to have

$$M \leq \left(1 - \frac{2\varepsilon}{\text{spg}(G)}\right)n = \left(1 - \frac{2}{K\text{spg}(G)}\right)KM. \quad (3.3)$$

Rearranging (3.2) and (3.3) produces the following specification on the spectral gap of $G$ for sufficient connectivity:

$$\text{spg}(G) \geq \max \left\{ \frac{6}{K}, \frac{2}{K-1} \right\} = \frac{6}{K},$$

where the equality is valid provided $K \geq 2$. To meet this specification, based on Theorem 3.1, it suffices to have $1 - 6/\sqrt{c} \geq 6/K$, which requires $K > 6$ and is equivalent to having $c \geq (1/6 - 1/K)^{-2}$. Using Lemma 3.1, the total number of masks is

$$K + 3\binom{K+1}{2}|A| = K + 6\binom{K+1}{2}|B| \leq K + 9\binom{K+1}{2}c\log M.$$

Setting $c = (1/6 - 1/K)^{-2}$, then $K = 12$ minimizes the coefficient of $\log M$. □

At this point, we note that $2 \cdot 10^5 \cdot \log M$ is a rather large number of Fourier masks. Certainly, the $10^5$ might be an artifact of our analysis—perhaps it could be decreased by leveraging better approximations. However, as the next result shows, the log factor is necessary for the Fourier masks to yield polarization recovery:
Theorem 3.2 Take $G$ as defined in Definition 2.1. Then $\text{spg}(G) > \varepsilon$ only if

$$|A| \geq \frac{\log M}{2 + \log(1/\varepsilon)}.$$ 

Proof. Define $V$ to be the $|A| \times M$ matrix built from taking the rows of $F$ indexed by $A$ and then scaling the columns to have unit norm. Then the inner product between any two columns of $V$ is given by

$$\langle v_m, v_{m'} \rangle = \sum_{a \in A} \left( \frac{1}{\sqrt{|A|}} e^{2 \pi i a/M} \right) \left( \frac{1}{\sqrt{|A|}} e^{-2 \pi i a'/M} \right) = \frac{M}{|A|} \sum_{m'' \in \mathbb{Z}_M} 1_A(m'') e^{-2 \pi i (m'-m)m''/M} = \frac{M}{|A|} (F^*1_A)(m'-m).$$

As such, the worst case coherence between columns of $V$ can be expressed in terms of the Fourier bias of $A$ (and the spectral gap of $G$ by Lemma 2.4):

$$\max_{m, m' \in \mathbb{Z}_M, m \neq m'} |\langle v_m, v_{m'} \rangle| = \frac{M}{|A|} \|A\|_\alpha = 1 - \text{spg}(G) \leq 1 - \varepsilon. \tag{3.4}$$

As one might expect, unit vectors can only be incoherent if there are enough dimensions relative to the number of vectors. To establish this explicitly, view the $v_m$’s as vectors in $\mathbb{C}^{|A|}$ by stacking the real and imaginary parts of each entry. Letting $\delta$ denote the shortest distance between distinct $v_m$’s (this is the same in $\mathbb{C}^{|A|}$ and $\mathbb{R}^{|A|}$), consider the open balls of radius $\delta/2$ centred at each $v_m$. The disjoint union of these $M$ balls is contained in the ball of radius $1 + \delta/2$ centred at the origin, and so a volume comparison gives

$$M \cdot C \left( \frac{\delta}{2} \right)^{2|A|} = \text{Vol} \left( \bigcup_{m \in \mathbb{Z}_M} B \left( v_m, \frac{\delta}{2} \right) \right) \leq \text{Vol} \left( B \left( 0, 1 + \frac{\delta}{2} \right) \right) = C \left( 1 + \frac{\delta}{2} \right)^{2|A|}.$$

Rearranging this inequality then yields

$$|A| \geq \frac{\log M}{2 \log(2/\delta + 1)}. \tag{3.5}$$

To analyse $\delta$, note that, for any $m, m' \in \mathbb{Z}_M$, we have

$$\|v_m - v_{m'}\|^2 = \|v_m\|^2 - 2 \text{Re}\langle v_m, v_{m'} \rangle + \|v_{m'}\|^2 \geq 2 - 2|\langle v_m, v_{m'} \rangle| \geq 2\varepsilon,$$

where the last step is by (3.4). Thus, $\delta^2 \geq 2\varepsilon$, with which we continue (3.5):

$$|A| \geq \frac{\log M}{2 \log(\sqrt{2/\varepsilon} + 1)} = \frac{\log M}{2 + \log(1/\varepsilon)}.$$

Here, the last inequality can be verified using the fact that $\varepsilon \leq 1$. □
4. Numerical simulations

In this section, we compare the polarization method of this paper to a state-of-the-art phase retrieval algorithm known as PhaseLift \cite{10, 13}. The main idea behind PhaseLift is that the intensity measurement $x \mapsto |\langle x, \varphi \rangle|^2$ can be viewed as a linear measurement if the signal $x$ is ‘lifted’ to the outer product $xx^*$ in the real vector space of self-adjoint $M \times M$ matrices. Indeed, the intensity measurement is a Hilbert–Schmidt inner product in this vector space:

$$|\langle x, \varphi \rangle|^2 = \varphi^* xx^* \varphi = \text{Tr}[\varphi^* xx^* \varphi] = \text{Tr}[\varphi \varphi^* xx^*] = \langle xx^*, \varphi \varphi^* \rangle_{\text{HS}}.$$ 

However, this larger vector space has $M^2$ dimensions, and so $M^2$ inner products are necessary to identify any member of the space—that is, unless more information is available. In this case, we know that the desired self-adjoint matrix $xx^*$ is positive semidefinite with rank 1, and so one could seek to minimize rank over the positive semidefinite matrices with the given intensity measurements. Rank minimization tends to be a difficult program to solve, so one is inclined to relax it:

\begin{equation}
\text{(PhaseLift)} \quad \text{minimize} \quad \text{Tr}(X) \quad \text{s.t.} \quad A(X) = b, \ X \succeq 0.
\end{equation}

To date, PhaseLift is known to stably reconstruct all signals in $\mathbb{C}^M$ with only $O(M)$ complex Gaussian measurements \cite{11}. However, very little is known about stability when the measurement vectors do not come from a unitarily invariant distribution (e.g. in the Fourier masks setting). To be fair, the present paper also does not prove stable reconstruction from Fourier masks; indeed, the previous section ‘merely’ established injectivity with efficient recoverability. Therefore, for the sake of completeness, this section will provide numerical simulations that compare the stability of both phase retrieval algorithms.

First, we describe how we implement phase retrieval with polarization. Starting with an ensemble of vertex measurement vectors $\Phi_V$ and edge measurement vectors $\Phi_E$, then, for each edge $(i, j) \in E$, we apply (2.1) to obtain

$$w_{ij} = \frac{1}{3} \sum_{r=0}^{2} \omega^r (|\langle x, \varphi_i + \omega^r \varphi_j \rangle|^2 + \varepsilon_{ijr}),$$

where $\varepsilon_{ijr}$ denotes the noise added to the $(i, j, r)$th edge measurement. Later, we will normalize $w_{ij}$ in order to estimate the relative phase between the measurements at vertices $i$ and $j$, but if $w_{ij}$ is small, this normalization will be particularly susceptible to noise. As such, we first remove vertices which are adjacent to small edge weights so as to promote reliability in the edges.

Algorithm 4.1 Pruning for reliability

Input: Weighted graph $G = (V, E, W)$, pruning parameter $\alpha$

Output: Subgraph $G'$ with a larger smallest edge weight

Initialize $G' \leftarrow G$

for $i = 1$ to $(1 - \alpha)|V|$

\hspace{1em} Find edge $(i, j) \in E$ of minimal weight

\hspace{1em} $G' \leftarrow G' \setminus \{i, j\}$

end

Now that we have isolated a subgraph with reliable edges, we want to find a further subgraph which is sufficiently connected so that its vertices can democratically agree on the relative phases. To find this subgraph, we iteratively remove vertices implicated by spectral clustering \cite{3, 4, 14} until the spectral gap
is sufficiently large. (In practice, we do not observe a noticeable decrease in accuracy when we leave out this step, but we include it anyways to align with the theory.)

**ALGORITHM 4.2** Pruning for connectivity

**Input:** Graph $G' = (V', E')$, pruning parameter $\tau$

**Output:** Subgraph $G''$ with $\text{spg}(G'') \geq \tau$

Initialize $G'' \leftarrow G'$

**while** $\text{spg}(G'') < \tau$ **do**

Take $D$ to be the diagonal matrix of vertex degrees

Compute the Laplacian $L \leftarrow I - D^{-1/2}AD^{-1/2}$

Compute the eigenvector $u$ corresponding to the second eigenvalue of $L$

**for** $i = 1$ **to** $|V''|$ **do**

Let $S_i$ denote the vertices corresponding to the $i$ smallest entries of $D^{-1/2}u$

Let $E(S_i, S_i')$ denote the number of edges between $S_i$ and $S_i'$

$h_i \leftarrow E(S_i, S_i')/\min\{\sum_{v \in S_i} \text{deg}(v), \sum_{v \in S_i'} \text{deg}(v)\}$

**end**

Take $S$ to be the $S_i$ of minimal $h_i$

$G'' \leftarrow G'' \setminus S$

**end**

At this point, our graph has reliable edges and is well connected. We now run angular synchronization [8,26] to reach a consensus on the phases of the vertex measurements, up to a global phase factor.

**ALGORITHM 4.3** Angular synchronization

**Input:** Graph $G'' = (V'', E'', W'')$

**Output:** Vector of phases corresponding to vertex measurements

Let $A_1$ denote the weighted adjacency matrix with the weights normalized

Let $D$ denote the diagonal matrix of vertex degrees

Compute the connection Laplacian $L_1 \leftarrow I - D^{-1/2}A_1D^{-1/2}$

Compute the eigenvector $u$ corresponding to the smallest eigenvalue of $L_1$

Output the phases of the coordinates of $u$

Having estimated the phases of the vertex measurements corresponding to $V''$, we now multiply the square roots of the vertex measurements $\{\sqrt{|\langle x, \varphi_i \rangle|^2 + \varepsilon_i}\}_{i \in V''}$ by these phases to estimate the inner products $\{(x, \varphi_i)\}_{i \in V''}$, and then produce a least-squares estimate for the desired signal $x$. As established in [1], this implementation of the polarization method is stable to noise when the vertex measurement vectors are complex Gaussian. In this section, we run numerical simulations to illustrate stability in the Fourier masks setting.

In the experiments that follow, we run the above implementation of polarization using pruning parameters $\alpha = 0.99$, $\tau = 0.1$, and the following construction of masks: First, we draw $K = 3$ masks whose diagonal entries are independent with distribution $N(0, 1)$. Next, we draw $M - 1$ independent realizations of a random variable $X$ with uniform distribution over the interval $[0, 1]$, and we put $i \in B$ whenever $X_i \leq \log M/M$. This way, $B \subseteq Z_M$ is a subset that does not contain 0, but the non-zero members of $Z_M$ each reside in $B$ with probability $\log M/M$, thereby following the intent of Theorem 2.2. Since $0 \not\in B$, we take $A = B \cup (-B)$, and then we construct the auxiliary masks $\{D_k + \omega^a E^a D_k\}_{r=0}^2$ for every $(k, k') \in \{0, 1, 2\}$, $k \geq k'$ and $a \in A$. Overall, in these experiments, polarization will use a total of $18|A| + 3$ masks, where $|A|$ tends to be approximately $2 \log M$. 
In contrast, PhaseLift offers a lot more flexibility with the number and types of masks used, and this flexibility allows us to run a collection of choice comparisons. We implement PhaseLift using TFOCS Auslender & Teboulle’s single-projection method [5] with the following parameters:

- `opts.alg = 'AT';`
- `opts.maxIts = 1000;`
- `opts.tol = 1e-6;`
- `opts.printEvery = 50;`
- `opts.maxmin = 1.0`

The lambda parameter that balances the trace with the data misfit was set to 0.1. We will run experiments at three noise levels, and for each level, we will compare polarization to three different implementations of PhaseLift. In the first implementation, we will only give PhaseLift the measurements corresponding to the original 3 masks we give to polarization. Next, we will give PhaseLift all 18|A| + 3 of the masks we described in the previous paragraph. For the last comparison, we will give PhaseLift the original 3 masks along with 18|A| additional masks whose diagonal entries are also independent with distribution \( \mathcal{N}(0, 1) \). Based on discussions in [10], this last setup appears to be the intended design of Fourier masks for PhaseLift, and so, in a sense, this last comparison will allow both algorithms to compete with their own ‘home-field advantage’.

In the comparison between polarization and PhaseLift, we use two performance metrics: runtime and relative error of reconstruction, defined in this setting by

\[
\min_{c \in \mathbb{C}, \|c\| = 1} \frac{\|c\hat{x} - x\|_2}{\|x\|_2}.
\]

Here, \( c \) is playing the role of the global phase factor we lost in the intensity measurement process. For each noise level tested \( \sigma^2 \in \{0, 0.1, 1\} \), and for each type of mask ensemble given to PhaseLift (original 3, same 18|A| + 3, random 18|A| + 3), we considered multiple signal dimensions \( M \in \{25, 26, 27, 28, 29\} \). In each of these scenarios, we ran 30 realizations of the following experiment:

- Draw each entry of the signal \( x \) independently from \( \mathcal{N}(0, 1) \).
- Add independent \( \mathcal{N}(0, \sigma^2) \) noise to each intensity measurement \( |\langle x, \varphi \rangle|^2 \).
- Run both polarization and PhaseLift to estimate \( x \) and record the runtime and relative error.

We ran these experiments on an Intel® Core™2 Quad CPU Q9550 at 2.83 GHz with 3.6 GB of memory.

Figure 1 presents the results of the noiseless scenarios (\( \sigma^2 = 0 \)). Here, perhaps the most striking distinction between polarization and PhaseLift is the relative error; indeed, polarization produces an estimate with almost no error (\( \sim 10^{-14} \)), whereas PhaseLift produces noticeable error (\( \sim 10^{-2} \)). In addition, polarization beats PhaseLift when using the same number of masks. For higher dimension signals, this advantage translates to minutes versus hours of running time. Letting \( N \) denote the number of intensity measurements. Then PhaseLift takes \( O(N^{3.5}) \) operations, whereas the computational bottleneck in polarization is the least-squares step takes \( O(N^3) \) operations, but this does not account for the disparity in runtime. Rather, the disparity is indicative of a large constant in front of the \( N^{3.5} \). In comparing the relative error in reconstruction, we note that PhaseLift terminates after the 1000th iteration or when the step size is sufficiently small, so this accounts for the significant difference in performance. Overall,
Fig. 1. Phase retrieval without noise. For the first row of graphs, PhaseLift is given the original $K = 3$ random masks, for the second, it is given all $18|A| + 3$ of the masks given to polarization, and in the last row, PhaseLift is given $18|A| + 3$ random masks. The plotted data illustrate the sample mean plus/minus one sample standard deviation. In some cases, the lower error bar is negative, and so it is not plotted in the log scale.
in the noiseless case, polarization handily defeats PhaseLift, even when PhaseLift is given the random masks it prefers.

Next, Figure 2 illustrates how polarization and PhaseLift perform in the presence of some noise ($\sigma^2 = 0.1$). In this case, PhaseLift continues to be rather slow, and it is clear that polarization enjoys greater stability from the $18|A|$ additional masks when compared with PhaseLift’s performance with just the original 3 random masks. We note that PhaseLift is particularly slow to converge when using few masks, and so the relative error of PhaseLift is especially high in the case of 3 masks (in part, as an artifact of our termination criteria). In [10], it is claimed that 3 random masks are sufficient for PhaseLift to reconstruct typical signals (and this is corroborated with our experiments in the noiseless case), but Figure 2 illustrates that PhaseLift requires more masks in order to reconstruct stably. And indeed, when PhaseLift is given more masks, it performs much better. Specifically, when they use the same masks, polarization and PhaseLift produce estimates with statistically comparable relative error, and PhaseLift performs slightly better when given the random masks it prefers. Intuitively, it makes sense that PhaseLift should be at least slightly more stable than polarization since it leverages all of the measurements in a democratic fashion, whereas polarization disenfranchises the vertex and edge measurements that are too small (i.e. ‘wasting’ measurements).

Figure 3 presents the results of our high-noise scenario ($\sigma^2 = 1$), which are very similar to the results corresponding to $\sigma^2 = 0.1$.

5. Discussion

In this paper, we showed how to leverage the polarization phase retrieval technique to construct Fourier masks that uniquely determine signals up to global phase, and then we used numerical simulations to illustrate the stability of polarization with such masks. At this point, we offer two conclusions:

(i) The polarization technique for phase retrieval is flexible enough to provably accommodate certain measurement design criteria (such as Fourier masks).

(ii) If you have the ability to use polarization instead of PhaseLift, you will gain considerable speedups in runtime at the price of a slight increase in relative error.

We would like to offer a few additional comments in regard to the second conclusion above. First of all, for applications like X-ray crystallography (for which the Fourier masks methodology was originally introduced), the physical experiments involved in the data collection step are very expensive, and so PhaseLift’s runtime is perhaps insignificant by comparison. However, for other applications of phase retrieval (such as correcting phase errors for synthetic aperture radar [23]), it is impractical to use PhaseLift due to its computational complexity. If such an application can instead use polarization-type measurements, then one should expect considerable speedups accordingly. Another comment regarding (ii) above: We have been evaluating the efficiency of PhaseLift in terms of the numerical simulations in the previous section, which used a particular SDP solver with our own choices of precision parameters. Certainly, there may be faster alternatives to solve the PhaseLift SDP (e.g. view it as a feasibility problem [15]), but benchmarking different SDP methods is beyond the scope of this paper, though interesting in its own right.

A number of interesting open questions remain. In general, it is known that $2M$ intensity measurements suffice for almost injectivity over $\mathbb{C}^M$ [6], whereas $4M - 4$ suffice for injectivity [9,17]. For almost injectivity, $2M$ is also known to be necessary [18], and $4M - 4$ is conjectured to be necessary for injectivity [7]. Overall, there is a factor-of-2 discrepancy between the number of measurements
Fig. 2. Phase retrieval with $\mathcal{N}(0, 0.1)$ noise added to each intensity measurement. For the first row of graphs, PhaseLift is given the original $K = 3$ random masks, for the second, it is given all $18|A| + 3$ of the masks given to polarization, and in the last row, PhaseLift is given $18|A| + 3$ random masks. The plotted data illustrate the sample mean plus/minus one sample standard deviation. In some cases, the lower error bar is negative, and so it is not plotted in the log scale.
Fig. 3. Phase retrieval with $\mathcal{N}(0, 1)$ noise added to each intensity measurement. For the first row of graphs, PhaseLift is given the original $K = 3$ random masks, for the second, it is given all $18|A| + 3$ of the masks given to polarization, and in the last row, PhaseLift is given $18|A| + 3$ random masks. The plotted data illustrate the sample mean plus/minus one sample standard deviation. In some cases, the lower error bar is negative, and so it is not plotted in the log scale.
required for almost injectivity and for injectivity, but this discrepancy seems to be much larger when we force the measurements to have Fourier masks structure: Based on [10, Theorem 3.1], 3 Fourier masks suffice for almost injectivity, and Theorem 2.2 gives that $O(\log M)$ suffice for injectivity. Furthermore, Theorem 3.2 gives that $\Omega(\log M)$ are necessary for injectivity with polarization recovery, while [12, Theorem 1.1] gives a PhaseLift guarantee with $O(\log^4 M)$ masks. But are $\Omega(\log M)$ Fourier masks required for injectivity?

In terms of algorithms, it remains to be proved whether stability holds in the Fourier masks setting (for any phase retrieval algorithm). This paper did not exploit the inherent Fourier structure to further speed up reconstruction, though this could very well be possible. Also, the number of masks might be decreased if additional information about the signal were leveraged, as in [16,22]. Finally, now that we have uniform Fourier masks-based performance guarantees for polarization, it is natural to seek similar guarantees for PhaseLift and its modifications [15,28].

Acknowledgement

The authors thank the Erwin Schrödinger International Institute for Mathematical Physics for hosting a workshop on phase retrieval that helped solidify the main ideas in this paper, and the anonymous referees who inspired Theorem 2.1, shortened our proof of Lemma 3.2 and provided several insights that helped shape some of the perspective in the discussion section.

Funding

A.S.B. was supported by AFOSR Grant No. FA9550-12-1-0317 and CMUC/FCT project PTDC/MAT/114394/2009 COMPETE/FEDER; Y.C. was supported by DARPA Grant No. N66001-13-1-4052 and AFOSR Grant No. FA9550-13-1-0076; and D.G.M. was supported by NSF Grant No. DMS-1321779. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense or the U.S. Government.

References


