

# Binary Discrete Fourier Transform and its Inversion

Howard W. Levinson and Vadim A. Markel

**Abstract**—A binary vector of length  $N$  has elements that are either 0 or 1. We investigate the question of whether and how a binary vector of known length can be reconstructed from a limited set of its discrete Fourier transform (DFT) coefficients. *A priori* information that the vector is binary provides a powerful constraint. We prove that a binary vector is uniquely defined by its two complex DFT coefficients (zeroth, which gives the popcorn, and first) if  $N$  is prime. If  $N$  has two prime factors, additional DFT coefficients must be included in the data set to guarantee uniqueness, and we find the number of required coefficients theoretically. One may need to know even more DFT coefficients to guarantee stability of inversion. However, our results indicate that stable inversion can be obtained when the number of known coefficients is about  $1/3$  of the total. This entails the effect of super-resolution (the resolution limit is improved by the factor of  $\sim 3$ ). Two algorithms for solving the inverse problem numerically are proposed and tested. The first, combinatorial algorithm is suitable for inverting vectors with  $N \lesssim 60$ ; the dependence of its computational complexity on the parameters of the problem is fully characterized in the paper. The second algorithm is based on non-convex optimization and its computational complexity is much harder to estimate; however, the method is demonstrated for a vector with  $N = 199$ . We hope that with further improvements the optimization algorithm can become applicable to two-dimensional problems.

**Index Terms**—Discrete Fourier transform, super-resolution, binary vector

## I. INTRODUCTION

Super-resolution in imaging and signal processing is a subject of significant current importance. Fundamentally, the resolution limit is related to experimental inaccessibility of the spatial Fourier harmonics of an object beyond some band limit [1]. The resulting image is therefore a low-path filtered version of the object. The classical Abbe limit on the resolution of optical systems is of this nature [2, Sec. 13.1.2]. There exist many more examples of imaging or signal reconstruction problems in which the Fourier components of the signal beyond some band limit are lost, corrupted or inaccessible.

One of the most powerful approaches for achieving super-resolution, as well as for image denoising, is based on utilization of prior information [3]. The information can be introduced in the form of a prior probability density in the Bayesian framework [4], or implicitly as a regularizing term in a cost function [5]. In addition to the classical Tikhonov regularization (ridge regression), the developed techniques include Laplacian [6], [7], total variation [8], [9], sparsity-based [10]–[12], and many variants of the above.

Another promising regularization technique is based on the so-called compositional constraints or the  $p$ -species model.

H.W. Levinson is with Department of Mathematics and Computer Science, Santa Clara University, Santa Clara, CA, USA (e-mail: hlevinson@scu.edu)

V.A. Markel is with Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA (e-mail: vmarkel@upenn.edu)

It has been used in a variety of settings including microscopy [13], [14], MRI [15], diffuse optical tomography [16], [17] and electromagnetic tomography [18]. The approach relies on the *a priori* knowledge that the sample consists of two or more known “species”, that is, materials whose properties are known. The spatial distribution of the components is however unknown and must be found by solving an inverse problem. In the context of signal processing, the  $p$ -species model implies that a discrete signal can take only  $p$  known values.

In this paper, we examine what is perhaps the most simple and at the same time fundamental mathematical question of the  $p$ -species model. Assume that there are only two species, i.e., the signal can take only two distinct values. How many Fourier coefficients one needs to know to reconstruct the signal precisely? In other words, assume that the discrete Fourier transform (DFT) coefficients are labeled by  $m$  where  $m \in [-M, M]$  and  $N = 2M + 1$  is the total number of the signal samples. In addition, we know that the signal can take only two distinct allowed values, say,  $a$  and  $b$ . Can we reconstruct the signal precisely if we know only the DFT coefficients within the band limit  $m \in [-L, L]$  where  $L < M$ ? What is the minimum value of  $L$  for which reconstruction is possible?

Below, we will prove that the inverse problem of reconstructing a binary signal from the band-limited DFT data has a unique solution if  $L = 1$  and  $N$  is prime. Moreover, we derive the minimum value of  $L$  that guarantees uniqueness if  $N$  has two prime factors. In the case of a large prime  $N$ ,  $L = 1$  is sufficient to guarantee uniqueness of the inverse solution. However, finding this solution numerically can be difficult or impractical due to high computational complexity and low noise tolerance. The difficulty can be mitigated by including additional DFT coefficients into the data set. As  $L$  is increased, stability of the inverse problem is improved and computational complexity is reduced. In all cases we have considered (with both prime and non-prime  $N$ ), reconstruction becomes practically possible when  $L \gtrsim M/3$ . This corresponds to achieving a three-fold improvement of the resolution limit.

We are interested in reconstructing the signal precisely and not approximately, and the form of the signal is assumed to be completely general. In particular, the signal may be more complicated than one or two compact pulses. Under the circumstances, application of the level-set methods [19] is impractical. Also, the two-species model is related but not identical to the assumption of sparsity. A sparse vector has many of its elements equal to zero but the rest can take any (unknown) values. The binary vectors considered here can have about one half of their elements equal to zero (in the most difficult case) but the rest have the same (known) value.

The rest of this paper is organized as follows. In Section II

we introduce the binary DFT, discuss its basic properties and show that it is sufficient to consider the vectors whose elements take only two known values, 0 and 1. In Section III, we adduce several numerical examples that illustrate the theoretical possibility of reconstructing a binary vector from band-limited DFT data. In Section IV, we prove two key uniqueness theorems. In Section V we discuss stability and in Section VI two algorithms for numerical inversion are introduced and tested. Finally, Section VII contains a discussion and conclusion.

## II. BINARY DFT

Let  $\mathbf{v} = (v_1, v_2 \dots, v_N)$  be a vector of length  $N > 1$ . For simplicity, we assume that  $N$  is odd. The DFT of  $\mathbf{v}$  is given by

$$\tilde{v}_m = \sum_{n=1}^N v_n e^{i\xi mn}, \quad (1a)$$

where

$$\xi = \frac{2\pi}{N}, \quad -M \leq m \leq M, \quad M = \frac{N-1}{2}. \quad (1b)$$

Note that  $\tilde{v}_m$  is defined for any integer  $m$  and is periodic so that  $\tilde{v}_{m+N} = \tilde{v}_m$ . It is therefore sufficient to restrict  $m$  to the interval  $[-M, M]$ . Assuming that we know all Fourier coefficients with indexes in this interval, we can reconstruct  $\mathbf{v}$  by the inverse DFT according to

$$v_n = \frac{1}{N} \sum_{m=-M}^M \tilde{v}_m e^{-i\xi nm}, \quad 1 \leq n \leq N. \quad (1c)$$

We will refer to  $\mathbf{v}$  and  $\tilde{\mathbf{v}} = (\tilde{v}_{-M}, \tilde{v}_{-M+1}, \dots, \tilde{v}_M)$  as to the real-space and Fourier-space vectors.

The questions we wish to address are the following. Assume that we know only some of the Fourier coefficients  $\tilde{v}_m$ , namely, those with the indexes bounded as  $|m| \leq L \leq M$ , and also that  $v_n$  can take only two distinct, *a priori* known values, say,  $a$  and  $b$ . What is the smallest value of  $L$  for which we can reconstruct the whole vector  $\mathbf{v}$  uniquely from the Fourier data? For which values of  $L$  the reconstruction is numerically stable? Finally, we wish to develop a computational algorithm for the reconstruction.

We can simplify the problem by writing

$$v_n = a + (b-a)x_n, \quad (2)$$

where  $x_n$  can take only two values, 0 and 1. We will say that the vector  $\mathbf{x} = (x_1, x_2 \dots, x_N)$  is *binary*. Substituting (2) into (1a), we obtain:

$$\tilde{v}_m = aN\delta_{m0} + (b-a)\tilde{x}_m, \quad (3)$$

where  $\delta_{ml}$  is the Kronecker delta-symbol and  $\tilde{x}_m$  are defined in terms of  $x_n$  analogously to the DFT convention (1a).

Equation (3) can be inverted to yield the relation

$$\tilde{x}_m = \frac{\tilde{v}_m - aN\delta_{m0}}{b-a}. \quad (4)$$

Therefore, if  $\tilde{v}_m$  is known, then  $\tilde{x}_m$  is also known. We thus see that the inverse problem of finding  $\mathbf{v}$  from  $\tilde{\mathbf{v}}$  is mapped onto the problem of finding a binary vector  $\mathbf{x}$  from its DFT  $\tilde{\mathbf{x}}$ .

Therefore, we are interested in reconstructing the full binary vector  $\mathbf{x}$  from a limited set of coefficients  $\tilde{x}_m$  with  $|m| \leq L$ . If  $L = 0$ , the only information about  $\mathbf{x}$  that is present in the data is the popcorn (the total number of 1s in  $\mathbf{x}$ ). Reconstructing  $\mathbf{x}$  with only this information is obviously impossible. However, the popcorn is an important constraint on the possible solutions. In what follows, we assume that the popcorn

$$r \equiv \tilde{x}_0 = \sum_{n=1}^N x_n \quad (5)$$

is known with a high precision so that (5) can be viewed as a hard constraint on the possible inverse solutions. We denote the set of all vectors  $\mathbf{x}$  of length  $N$  containing  $r$  1s by  $\Omega(N, r)$ . The size of this set is

$$S[\Omega(N, r)] = \frac{N!}{r!(N-r)!}. \quad (6)$$

It is sufficient to consider  $r$  in the range  $1 < r \leq (N-1)/2$  since the sets  $\Omega(N, r)$  and  $\Omega(N, N-r)$  are obtained from each other by the substitution  $0 \leftrightarrow 1$ .

The band-limited to  $L$  Fourier-space distance between any two vectors  $\mathbf{x}, \mathbf{y} \in \Omega(N, r)$  is defined as

$$\chi_p(\mathbf{x}, \mathbf{y}; L) = \left[ \frac{1}{L} \sum_{m=1}^L |\tilde{x}_m - \tilde{y}_m|^p \right]^{\frac{1}{p}}, \quad p \geq 1. \quad (7)$$

Note that the term with  $m = 0$  is excluded from the summation. The real-space distance for  $\mathbf{x}, \mathbf{y} \in \Omega(N, r)$  is defined as

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{n=1}^N |x_n - y_n|. \quad (8)$$

If the distance between two vectors in  $\Omega(N, r)$  is  $d$ , one can be obtained from the other by  $d$  pair-wise switches of 0s and 1s. For given  $N$  and  $r$ , the possible values of  $d$  are in the interval  $1 \leq d \leq r$ . If  $L = M$ , the invertibility of DFT implies that the statements  $\chi_p(\mathbf{x}, \mathbf{y}; M) = 0$ ,  $d(\mathbf{x}, \mathbf{y}) = 0$  and  $\mathbf{x} = \mathbf{y}$  are equivalent. However, if  $L < M$ , we do not generally know whether there exist pairs of distinct vectors  $\mathbf{x} \neq \mathbf{y}$  for which  $\chi_p(\mathbf{x}, \mathbf{y}; L) = 0$ .

If  $r = 0$ , the inverse problem is solved trivially for any  $L$ ; we know that all components of  $\mathbf{x}$  are 0s. Similarly, if  $r = N$ , then  $\mathbf{x}$  consists of all 1s. The cases  $r = 1$  and  $r = N-1$  are also easy to solve. If  $r = 1$ , it is sufficient to know only one additional Fourier coefficient, say,  $\tilde{x}_1$ . The inverse problem is then reduced to finding the position  $\nu$  where the single 1 is located; it can be found from the equation  $\tilde{x}_1 = \exp(i\xi\nu)$ . If  $\tilde{x}_1$  is in range of the forward operator, then the above equation has a unique integer solution in the interval  $[1, N]$ . If  $\tilde{x}_1$  is not in range, then the equation has no integer solutions. One can still find the integer  $\nu$  that minimizes the error  $|\tilde{x}_1 - \exp(i\xi\nu)|$ . The case  $r = N-1$  is quite analogous, except that  $\exp(i\xi\nu)$  is replaced by  $-\exp(i\xi\nu)$ . The cases  $r = 2$  and  $r = N-2$  are also easy to analyze. The problem becomes difficult when  $r \sim N/2$  and  $N \gg 1$  so that  $S[\Omega(N, r)]$  is combinatorially large. The remainder of this paper is largely focused on the latter (more difficult) case.

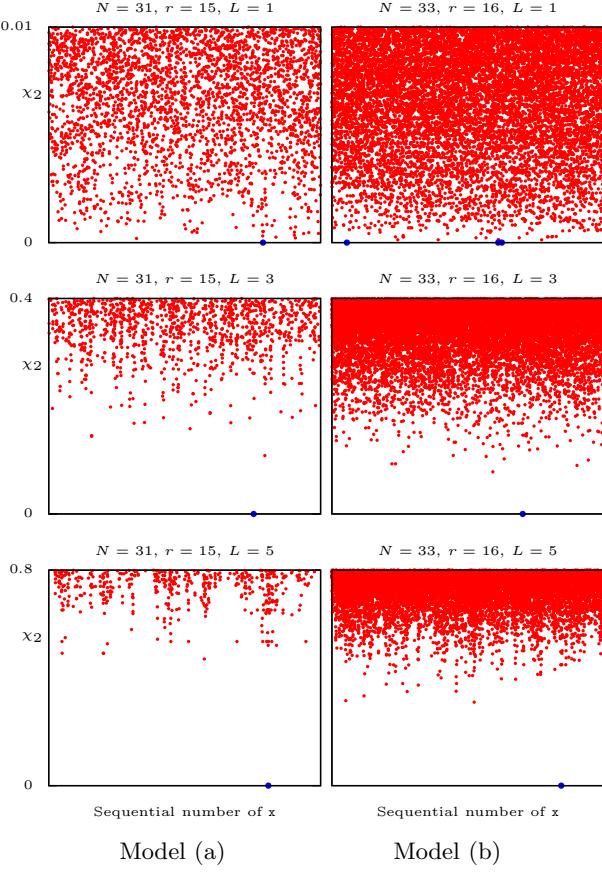


Fig. 1. Fourier-space distances  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; L)$  between various  $\mathbf{x} \in \Omega(N, r)$  and a model vector  $\mathbf{x}_{\text{mod}}$  with the same  $N$  and  $r$ . The left and right columns correspond to the Models (a) and (b), which are defined in the text. Different rows correspond to different values of  $L$ . All data points that fit the vertical scale of each plot are shown. Projections onto the horizontal axis are the sequential numbers of the data points and have no other significance. Due to the finite size of the dots that are used to represent the data points, it may appear that one sequential number (a projection onto the horizontal axis) corresponds to more than one dot; in fact, this is not so. Large blue dots mark exact zeros. Note that, for  $L = 1$ ,  $\chi_p$  is independent of  $p$ .

### III. A NUMERICAL EXAMPLE

As a first step, we can investigate the problem numerically. To this end, we have considered the following two model vectors  $\mathbf{x}_{\text{mod}}$ :

$$(a) N = 31, r = 15, S[\Omega(N, r)] = 2,147,483,648 \\ \mathbf{x}_{\text{mod}} = (1001011000011101101100011010100)$$

and

$$(b) N = 33, r = 16, S[\Omega(N, r)] = 8,589,934,592 \\ \mathbf{x}_{\text{mod}} = (100100110001100111001010100110110)$$

For these values of  $N$ , all elements of  $\Omega(N, r)$  can be constructed explicitly on a computer.

In Fig. 1, we display the quantities  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; L)$  (below some thresholds) for all  $\mathbf{x} \in \Omega(N, r)$ , various  $L$ , and the two model vectors  $\mathbf{x}_{\text{mod}}$  defined above. It can be seen that, for  $N = 31$  and all  $L$  considered, there exists only one  $\mathbf{x} \in \Omega(N, r)$  for which  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; L) = 0$ . This  $\mathbf{x}$  is the true solution, that is, it is identical to  $\mathbf{x}_{\text{mod}}$ . This result implies that the knowledge

of just the first two Fourier coefficients  $\tilde{x}_0$  and  $\tilde{x}_1$  suffices to find the whole  $\mathbf{x}_{\text{mod}}$ . This result is surprisingly strong but, as discussed below, not unexpected. One obvious observation is that 31 is a prime number. We will show that uniqueness of inverse solutions with  $L = 1$  is a general property of all prime  $N$ 's.

In the case  $N = 33$  (not a prime), there are three distinct vectors  $\mathbf{x}$  with  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; 1) = 0$ ; only one of them is equal to  $\mathbf{x}_{\text{mod}}$ . The other two are false solutions. However, uniqueness of the inverse solution is restored by selecting  $L \geq 3$  ( $L = 2$  is still insufficient). In the next section, the theoretical reasons why  $L = 3$  provides the unique solution in this case will be given.

### IV. UNIQUENESS OF INVERSE SOLUTIONS

We now fix a few definitions and prove two key uniqueness theorems.

**Definition 1.** We say that two vectors  $\mathbf{x}, \mathbf{y} \in \Omega(N, r)$  are  $L$ -distinguishable if  $\chi_p(\mathbf{x}, \mathbf{y}; L) > 0$ , where  $1 \leq L \leq M$ , and  $L$ -indistinguishable otherwise. If this property holds for some  $p$ , it holds for all  $p \geq 1$ , including the formal limit  $p = \infty$ . To generalize the definition, we say that vectors of the same length  $N$  but with different popcounts  $r$  are 0-distinguishable<sup>1</sup>. All vectors in  $\Omega(N, r)$  have the same popcount and are therefore 0-indistinguishable. If two vectors are  $L$ -distinguishable, they are also  $L'$ -distinguishable for any  $L' > L$ .

**Definition 2.** We use the acronym IP( $N, r, L$ ) to denote the inverse problem of reconstructing a generic binary vector  $\mathbf{x}$  of known length  $N$  and popcount  $r$  from the set of its DFT coefficients  $\tilde{x}_m$  with  $0 \leq m \leq L$ . Note that  $\tilde{x}_{-m} = \tilde{x}_m^*$ ; therefore, coefficients with negative indexes are not included in the data set. In this paper, we consider only odd  $N$ 's. Correspondingly, the possible values of  $r$  in IP( $N, r, L$ ) are  $0 < r \leq (N+1)/2$ ; the trivial case  $r = 0$  is excluded and the cases with  $r > (N+1)/2$  are equivalent to those with  $r < (N+1)/2$  by symmetry.

**Definition 3.** We say that IP( $N, r, L$ ) is uniquely solvable if all pairs of distinct vectors  $\mathbf{x}, \mathbf{y} \in \Omega(N, r)$  are  $L$ -distinguishable.

**Definition 4.** We say that a binary vector  $\mathbf{x} \in \Omega(N, r)$  contains a regular polygon if there exists integers  $m$  and  $k$  such that  $x_n = 1$  for all  $n \equiv m \pmod{k}$ . We call  $k$  the order of the polygon. Similarly, we say that  $\mathbf{x}$  contains an empty regular polygon if  $x_n = 0$  for all  $n \equiv m \pmod{k}$ . Note that these definitions require  $k$  to divide  $N$ . We say that  $\mathbf{x}$  contains a pair of regular polygons if  $\mathbf{x}$  contains both a regular polygon and an empty regular polygon of the same order. Regular polygons are disjoint in  $\mathbf{x}$  if they do not share any indices  $n$ .

**Theorem 1.** IP( $N, r, 1$ ) is uniquely solvable for all  $r$  in the interval  $0 < r \leq (N+1)/2$  if  $N$  is prime.

*Proof.* Theorem 1 is proved by showing that all vectors in  $\Omega(N, r)$  are pairwise 1-distinguishable for  $N$  prime. Consider two distinct vectors  $\mathbf{x}, \mathbf{y} \in \Omega(N, r)$  and suppose that  $\mathbf{x}$  and  $\mathbf{y}$  are 1-indistinguishable. Let  $\mathbf{z} = \mathbf{x} - \mathbf{y}$  with  $z_n (n = 1, 2, \dots, N)$

<sup>1</sup>Such two vectors do not belong to the same set  $\Omega(N, r)$ .

denoting the real-space components of  $\mathbf{z}$  and  $\tilde{\mathbf{z}}_m$  denoting its DFT coefficients defined according to the convention (1a). Note that  $\mathbf{z}$  is not a binary vector as its entries can take three possible values: 0 and  $\pm 1$ . Therefore,  $\mathbf{z} \notin \Omega(N, r)$ . Since we have assumed that  $\mathbf{x}$  and  $\mathbf{y}$  are 1-indistinguishable, the following equality must hold:

$$0 = \tilde{z}_1 = \sum_{n=1}^N z_n e^{i\xi n}. \quad (9)$$

Since  $N$  is prime, the  $N - 1$  exponential factors  $e^{i\xi n}$  with  $1 \leq n < N - 1$  (excluding the term  $e^{i\xi N} = 1$ ) form the complete set of  $N$ -th primitive roots of unity. Therefore, the  $N$ -th cyclotomic polynomial has the form

$$\Phi_N(x) = \prod_{n=1}^{N-1} (x - e^{i\xi n}) = \sum_{n=0}^{N-1} x^n. \quad (10)$$

By Eisenstein's criterion, this polynomial is irreducible over the rationals [20]. We also introduce the polynomial

$$g(x) = z_N + z_1 x + z_2 x^2 + \dots + z_{N-1} x^{N-1}.$$

Assuming that (9) holds,  $g(x)$  has a root at  $e^{i\xi}$ . We now observe that both  $\Phi_N(x)$  and  $g(x)$  are polynomials with rational coefficients of degree  $N-1$  and with the common root  $e^{i\xi}$ . Since  $\Phi_N(x)$  is irreducible over the rationals, and thus a minimal polynomial, these two polynomials can only differ by a constant. This implies that  $z_n = C$  for some constant  $C$ . Recall  $z_n$  can only take values of 0 and  $\pm 1$ . If  $C = \pm 1$ , then  $\mathbf{x}$  is a vector of all 1s and  $\mathbf{y}$  is a vector of all 0s (or vice versa), which violates the assumption that they are both in  $\Omega(N, r)$ . Hence the only possibility is  $z_n = 0$ , which is equivalent to  $\mathbf{x} = \mathbf{y}$  and contradicts the initial assumption that  $\mathbf{x}$  and  $\mathbf{y}$  are distinct.  $\square$

Theorem 1 implies that, at least theoretically, any binary vector  $\mathbf{x}$  of a prime length  $N$  can be uniquely recovered from its two Fourier coefficients  $\tilde{x}_0 = r$  and  $\tilde{x}_1$ . It does not imply that this problem can always be solved in a numerically stable manner. Stability of inversion is discussed in Section V below.

When  $N$  is not prime, we do not have such a strong statement of uniqueness. However, if  $N$  has two prime factors, we can characterize the false solutions and derive the sufficient conditions of uniqueness assuming that some additional DFT coefficients are known. The following Lemma establishes the necessary and sufficient condition for two binary vectors in  $\Omega(N, r)$  to be 1-distinguishable.

**Lemma 1.** Let  $N = pq$  where  $1 < p \leq q$  are primes (not necessarily distinct). Let  $\mathbf{x} \in \Omega(N, r)$ . Then  $\mathbf{x}$  is 1-indistinguishable from some other (different from  $\mathbf{x}$ ) vector(s) in  $\Omega(N, r)$  iff  $\mathbf{x}$  contains at least one pair of  $p$ - or  $q$ -gons according to Definitions 4. Equivalently,  $\mathbf{x}$  is 1-distinguishable from all other vectors in  $\Omega(N, r)$  iff it does not contain any pairs of  $p$ - or  $q$ -gons.

The proof of Lemma 1 is given in Appendix A. It relies on the algebraic ideas that were used to study the vanishing sums of the roots of unity [21]–[25]. Geometrically, the proof

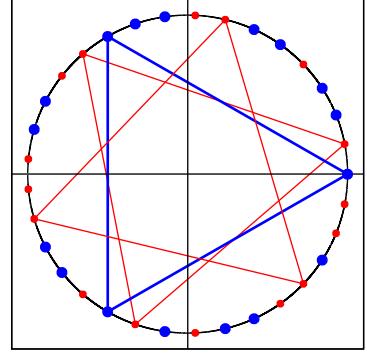


Fig. 2. Geometrical illustration of Lemma 1 for the model vector (b) with  $N = 33$ . Plotted are the points  $e^{i\xi n}$  on the unit circle. The small red and large blue dots are obtained for the values of  $n$  such that  $x_n = 1$  and  $x_n = 0$ , respectively. There are two 3-gons contained in this binary vector, which are shown by thin red lines. There is also one empty 3-gon shown by thick blue lines. This gives two ways in which a pair of polygons can be formed; each distinct pair results in a distinct false solution. The two false solutions shown in Fig. 1 are obtained by switching 1s in the positions corresponding to the vertices of one of the red 3-gons with 0s in the positions corresponding to the vertices of the blue (empty) 3-gon.

states that the only way  $\mathbf{x}$  and  $\mathbf{y}$  in  $\Omega(N, r)$  can be 1-indistinguishable is if they agree element-wise except for the locations that make up an equivalent number of regular  $p$ - or  $q$ -gons. This is illustrated in Fig. 2.

While Lemma 1 establishes the necessary and sufficient condition for two distinct vectors in  $\Omega(N, r)$  to be 1-indistinguishable (in the two prime factors case), the following Lemma provides a potential remedy to the non-uniqueness. Specifically, it tells us how many additional DFT coefficients must be included in the data set to guarantee uniqueness.

**Lemma 2.** Let  $N = pq$  with the same conditions on  $p, q$  as in Lemma 1, and let  $\mathbf{x} \in \Omega(N, r)$ . Let  $L$  be the smallest integer for which  $\mathbf{x}$  contains a pair of  $L$ -gons, which are not subsets of any larger-order pair of regular polygons. Then  $\mathbf{x}$  is  $L$ -distinguishable from all other vectors in  $\Omega(N, r)$ . Moreover,  $L$  is the smallest value of  $k$  for which  $\mathbf{x}$  is  $k$ -distinguishable from all other vectors in  $\Omega(N, r)$ .

The proof of Lemma 2 is given in Appendix B. The geometric concept behind this proof is illustrated in Fig. 3. Lemma 2 implies that, if  $p \neq q$ , then all vectors in  $\Omega(N, r)$  are  $q$ -distinguishable. For  $N = p^2$ , all vectors are  $p$ -distinguishable. Considering again the case  $N = 33, r = 16$ , the best one can assume is that all vectors in  $\Omega(33, 16)$  are 11-distinguishable. This implies that one must use  $L \geq 11$  to guarantee uniqueness of  $\text{IP}(33, r, L)$  for any  $r \geq 11$ . However, the model vector (b) does not contain any 11-gons. Therefore, it can be recovered with  $L = 3$ ; using  $L = 11$  would be an overkill. We note that vectors that contain regular 11-gons make a small subset of  $\Omega(33, 16)$  (are statistically rare). Therefore, using  $L = 3$  for  $N = 33, r = 16$  entails a relatively small risk of running into a false solution. Moreover, Lemma 2 tells us exactly what form these vectors, and the corresponding false solutions, take. The question of *statistically reliable* invertibility – that is, accepting a small risk of finding a false solution – is addressed more

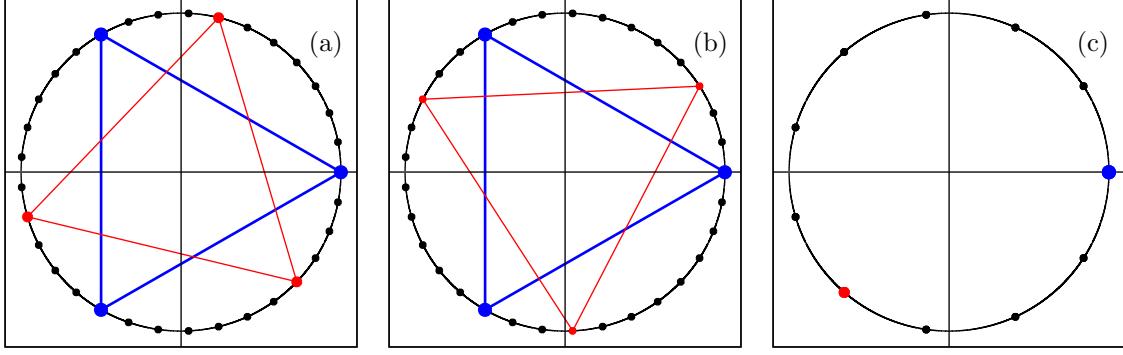


Fig. 3. Geometrical illustration of Lemma 2. Let  $\mathbf{x}$  be given by the model vector (b) ( $N = 33, r = 16$ ) and let  $\mathbf{y}$  be one of the two vectors that are distinct but 1-indistinguishable from  $\mathbf{x}$  (we have chosen for  $\mathbf{y}$  the specific false solution obtained by switching 1s and 0s in the pair of 3-gons shown in Fig. 2 that are geometrically farther apart). Let, as in the proof of Theorem 1,  $\mathbf{z} = \mathbf{x} - \mathbf{y}$ . Plotted are the terms that enter the definition (1a) of  $\tilde{z}_1$  (a),  $\tilde{z}_2$  (b), and  $\tilde{z}_3$  (c). Thus, Panel (a) displays the terms  $z_n e^{i\xi n}$  for  $1 \leq n \leq 33$ . The terms with such  $n$  that  $z_n = 0, z_n = 1$  and  $z_n = -1$  are represented by small black dots, intermediate-size red dots, and large blue dots, respectively. Panel (b) displays the terms  $z_n e^{i\xi 2n}$  according to the same color convention. The additional factor of 2 in the exponent results in a permutation of the dots that are displayed in Panel (a); however, the 3-gons are preserved. Note that the total number of dots in Panels (a) and (b) is the same and equal to 33. It is clear that  $\tilde{z}_1 = \tilde{z}_2 = 0$  since perfect 3-gons sum to zero. In Panel (c), the terms  $z_n e^{i\xi 3n}$  are shown. For  $1 \leq n \leq 33$ , the above terms take only 11 distinct values so that each displayed dot corresponds to a sum of three terms with different  $n$ . However, only dots of the same color (same value of  $z_n$ ) can overlap in this example. It can be seen that each 3-gon of Panel(a) collapses to a single point in Panel (c). Therefore,  $\tilde{z}_3 \neq 0$  implying that  $\mathbf{x}$  and  $\mathbf{y}$  are 3-distinguishable.

systematically in Section V.

In the absence of any *a priori* knowledge about the true solution apart from what is given in the data, we can state the following sufficient condition for uniqueness of  $\text{IP}(N, r, L)$ .

**Theorem 2.** Let  $N = pq$  with the same conditions on  $p, q$  as in Lemma 1,  $r \leq (N+1)/2$ , and

$$L_0 = \max_{\ell, m \in \{0, 1\}} (\{k = p^\ell q^m : k \leq r\}).$$

Then  $\text{IP}(N, r, L)$  is uniquely solvable for any  $L \geq L_0$ .

*Proof.* Theorem 2 is a direct consequence of Lemma 2. For any  $\mathbf{x} \in \Omega(N, r)$ ,  $\mathbf{x}$  needs at least  $L$  1s to form an  $L$ -gon. As there are  $r$  1s in  $\mathbf{x}$ , we must have  $L \leq r$  for any  $L$ -gons contained in  $\mathbf{x}$ . By the definition of  $L_0$ ,  $L_0$  is the largest such order of polygon that can be formed. Hence, by Lemma 2, any  $\mathbf{x} \in \Omega(N, r)$  is  $L_0$ -distinguishable from all other vectors in  $\Omega(N, r)$ . Therefore,  $\text{IP}(N, r, L)$  is uniquely solvable for any  $L \geq L_0$ .  $\square$

To illustrate Theorem 2, consider the case  $N = 143 = 11 \cdot 13$ . If  $r < 11$ , then we have  $L_0 = 1$ . If  $11 \leq r < 13$ , then  $L_0 = 11$ , and if  $13 \leq r \leq (N+1)/2 = 72$ , then  $L_0 = 13$ . Thus,  $L = 13$  guarantees uniqueness of the inverse solution for any binary vector of length  $N = 143$ .

**Remark 1.** Even if  $\text{IP}(N, r, L)$  is not uniquely solvable, some vectors in  $\Omega(N, r)$  can be uniquely recovered from the knowledge of  $\tilde{x}_m$  with  $m = 1, 2, \dots, L$ . For example the following vector

$$(c) N = 35, r = 17, S[\Omega(N, r)] = 4,059,928,950 \\ \mathbf{x} = (01000100101011000110111001001101001)$$

is uniquely recoverable with  $L = 1$  even though  $\text{IP}(35, 17, 1)$  is not uniquely solvable. The reason is that  $\mathbf{x}$  does not contain any 5- or 7-gons.

**Remark 2.** Similar but more complicated results hold for the case when  $N$  has two prime divisors, i.e.,  $N = p^\alpha q^\beta$  and the integers  $\alpha, \beta$  can be larger than 1. Here the difficulty grows from the possibility of an intricate overlapping of different polygon pairs. The case when  $N$  has three or more prime divisors is even more difficult to analyze due to the existence of the so-called asymmetrical minimal vanishing sums of  $N$ -th roots of unity [22], [25].

## V. STABILITY OF INVERSION

Even when an inverse problem  $\text{IP}(N, r, L)$  is uniquely solvable, it is not clear whether finding the solution is a numerically stable procedure. Consider the data points in Fig. 1 for  $N = 31$  and  $L = 1$ . The large (blue) dot corresponds to the true solution  $\mathbf{x}_{\text{mod}}$  and it is the only point with  $\chi_2 = 0$  in the corresponding data set, so that the inverse solutions is unique. However, the Fourier-space distance between the first runner-up to the true solution (let us call it  $\mathbf{y}$ ) and the model is quite small:  $\chi_2(\mathbf{y}, \mathbf{x}_{\text{mod}}, 1) \approx 0.0002$ . On the other hand, the real-space distance between  $\mathbf{y}$  and  $\mathbf{x}_{\text{mod}}$  is not small:  $d(\mathbf{y}, \mathbf{x}_{\text{mod}}) = 10$ . In other words, 10 out of 16 1s in  $\mathbf{y}$  are in the wrong places. This is an obvious sign of instability. A small change in the DFT data can result in a large change of the inverse solution.

In this section, we investigate the stability of  $\text{IP}(N, r, L)$  numerically. To this end, we have selected some values of  $N$  and generated sets of random vectors  $\mathbf{x}_j \in \Omega(N, r)$  for  $r = 2, 3, \dots, M$  and  $j = 1, 2, \dots, J$ , where  $J$  was chosen to be sufficiently large to obtain statistically significant results. The vectors  $\mathbf{x}_j$  were generated as follows. For a particular random realization,  $r$  1s were randomly placed into  $N$  possible positions. Repetitions (identical random realizations) were allowed but occurred very rarely. The number of random realizations in a set,  $J$ , depended on  $r$  and  $N$  and varied from  $10^4$  to  $10^2$  in the most difficult cases such as  $N = 35, r = 17$ .

Then, for each  $\mathbf{x}_j$ , we have computed the Fourier-space distances <sup>2</sup>  $\chi_2(\mathbf{x}_j, \mathbf{y}; L)$  to *all* vectors  $\mathbf{y} \in \Omega(N, r)$  and the minimum distance between  $\mathbf{x}_j$  and any  $\mathbf{y}$  that is not equal to  $\mathbf{x}_j$ . The latter quantity can be formally defined as

$$\kappa(\mathbf{x}_j; L) = \min_{\mathbf{y}, \mathbf{y} \neq \mathbf{x}_j} \chi_2(\mathbf{x}_j, \mathbf{y}; L), \quad \mathbf{x}_j, \mathbf{y} \in \Omega(N, r). \quad (11)$$

Note that  $\kappa(\mathbf{x}_j; L)$  and its averages defined below in (12) depend implicitly on  $N$  and  $r$ . If  $\kappa(\mathbf{x}_j; L) = 0$ , the vector  $\mathbf{x}_j$  is not uniquely recoverable with the particular value of  $L$ . If  $\kappa(\mathbf{x}_j; L) > 0$  but is in some sense small, then  $\mathbf{x}_j$  is uniquely recoverable with the given  $L$  but the inverse solution is numerically unstable. Generally, as  $\kappa(\mathbf{x}_j; L)$  is increased, it becomes easier to recover  $\mathbf{x}_j$ , and the precision requirements on the DFT data become less stringent.

We have also computed the average and the standard deviation of  $\kappa(\mathbf{x}_j; L)$  according to

$$\langle \kappa(L) \rangle = \frac{1}{J} \sum_{j=1}^J \kappa(\mathbf{x}_j; L), \quad (12a)$$

$$\langle \kappa^2(L) \rangle = \frac{1}{J} \sum_{j=1}^J \kappa^2(\mathbf{x}_j; L), \quad (12b)$$

$$\sigma^2(L) = \langle \kappa^2(L) \rangle - \langle \kappa(L) \rangle^2. \quad (12c)$$

These quantities are illustrated in Fig. 4 for  $N = 31, 33, 35$  and  $L$  from 1 to 5. The data displayed in this figure convey how “easy” it is to reconstruct a vector from  $\Omega(N, r)$ . For example, consider the case  $N = 35$ ,  $L = 3$  and  $r = 15$ . We have for these parameters  $\langle \kappa \rangle \approx 0.1$  and  $\sigma \approx 0.03$ . This means that most vectors in  $\Omega(35, 15)$  (those within the the  $\pm 2\sigma$ -interval in the statistical distribution) have the Fourier-space distance to the closest distinct neighbor between 0.04 and 0.16. Therefore, if we find a vector  $\mathbf{y} \in \Omega(35, 15)$  with  $\chi_2(\mathbf{x}, \mathbf{y}; 3) \leq 0.04$ , it is likely to be the true solution. We thus conclude that the Fourier data should be specified with the absolute precision of 0.04 or better.

Not all combinations of  $N$ ,  $L$  and  $r$  allow such simple considerations. The differences  $\langle \kappa \rangle - \sigma$  can be very small or even negative (this is technically possible). In such cases, the lower parts of the error bars in Fig. 4 are outside of the plot frames. Relevant examples include  $N = 35$ ,  $L = 1$  for  $r \geq 5$ . In practical terms, the inverse problems with these combinations of  $N$ ,  $L$  and  $r$  are hard to solve since it is likely that a vector in  $\Omega(N, r)$  has a distinct neighbor that is *almost*  $L$ -indistinguishable from itself. To make the inverse problem better conditioned, one can either increase  $L$  or increase the precision of the DFT data, assuming the inverse solution is unique.

So far, we have not characterized the statistical distribution of  $\kappa(\mathbf{x}_j; L)$ . Therefore, the considerations based on the data of Fig. 4 are qualitative. Computing the full statistical distribution of  $\kappa(\mathbf{x}_j; L)$  is a combinatorially hard problem. Instead, we have introduced the cumulative probabilities

$$P(\epsilon; L) = \frac{1}{J} \sum_{j=1}^J \Theta(\epsilon - \kappa(\mathbf{x}_j; L)), \quad (13)$$

<sup>2</sup>In this section, we rely on the  $L_2$  norm.

where

$$\Theta(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (14)$$

is the step function. Just like  $\kappa(\mathbf{x}_j; L)$ ,  $P(\epsilon; L)$  depends implicitly on  $N$  and  $r$ . It gives the (approximate) fraction of vectors in  $\Omega(N, r)$  that have at least one distinct neighbor that is at least as close in Fourier space (quantified by the  $\chi_2(\mathbf{x}, \mathbf{y}; L)$  metric) as  $\epsilon$ . In particular,  $P(0; L)$  gives the fraction of vectors in  $\Omega(N, r)$  that have at least one distinct but  $L$ -indistinguishable neighbor.

The quantities  $P(\epsilon; 1)$  are shown in Fig. 5 for  $N = 31, 33, 35$ , and  $\epsilon = 0, 10^{-4}, 10^{-3}$ . As expected,  $P(0; 1) = 0$  for  $N = 31$ . This means that, in agreement with Theorem 1, all vectors in  $\Omega(31, r)$  with  $r = 1, 2, \dots, M$  are 1-distinguishable from each other.

Data for larger  $L$  are not shown in Fig. 5 but can be described as follows. In the case  $N = 31$ , we have  $P(10^{-3}; L) = P(10^{-4}; L) = P(0; L) = 0$  for all  $r$  considered and  $L > 1$  (note that  $P(\epsilon; L)$  is a non-decreasing function of  $\epsilon$ ). Therefore, choosing  $L = 2$  already makes the inverse problem stable in this case.

In the case  $N = 33$ ,  $P(10^{-3}; L) = P(10^{-4}; L) = P(0; L) = 0$  (for all  $r$ ) when  $L > 2$ , whereas choosing  $L = 2$  is still not sufficient for making the inverse problem stable. Note that, for  $L < 11$  and  $r > 11$ ,  $P(\epsilon; L)$  are not exactly zero due to the possibility of forming perfect 11-gons (see Lemma 1). However, the values of  $P$  are too small to be determined by the statistical approach used here. These cumulative probabilities can also be computed theoretically by counting all the elements  $\mathbf{x}$  that allow perfect 11-gons.

Finally, in the case  $N = 35$ , we can use Theorem 2 to determine whether there are false solutions. For  $r < 5$ , there are no false solutions already with  $L = 1$ . For  $5 \leq r < 7$ , false solutions are suppressed when  $L \geq 5$ . For  $7 \leq r \leq 17$ , suppressing false solutions requires  $L \geq 7$ . However, even if  $L < 7$ , false solutions are statistically rare. Moreover, when  $L > 1$ , the false solutions appear to be the only source of numerical instability. Therefore, the inverse problem with  $N = 35$  and arbitrary  $r$  can be solved in practice even with  $L < 7$ . For example, choosing  $L = 3$  entails the probability of running into a false solution of the order of 0.01 or smaller.

## VI. INVERSION

Even if the solution to the inverse problem is unique and we know the DFT data with sufficient precision to guarantee stability, finding the solution is a nontrivial task. In particular, iterative methods that use pair-wise swaps of 0s and 1s and seek to optimize  $\chi_p$  are unlikely to work. The reason for this is illustrated in Figs. 6 and 7.

In Fig. 6 we plot the real-space distance  $d(\mathbf{x}, \mathbf{x}_{\text{mod}})$  between various vectors  $\mathbf{x} \in \Omega(N, r)$  and the model  $\mathbf{x}_{\text{mod}}$  vs the corresponding Fourier-space distance  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; L)$ . It can be seen that the two distances do not correlate well. It is possible to have small  $\chi_2$  for a large  $d$  and small  $d$  with a large  $\chi_2$ . Any optimization technique that works directly with the real-space vectors  $\mathbf{x} \in \Omega(N, r)$  and tries to reduce the error  $\chi_2$  iteratively is therefore not likely to solve the problem: the iterations will

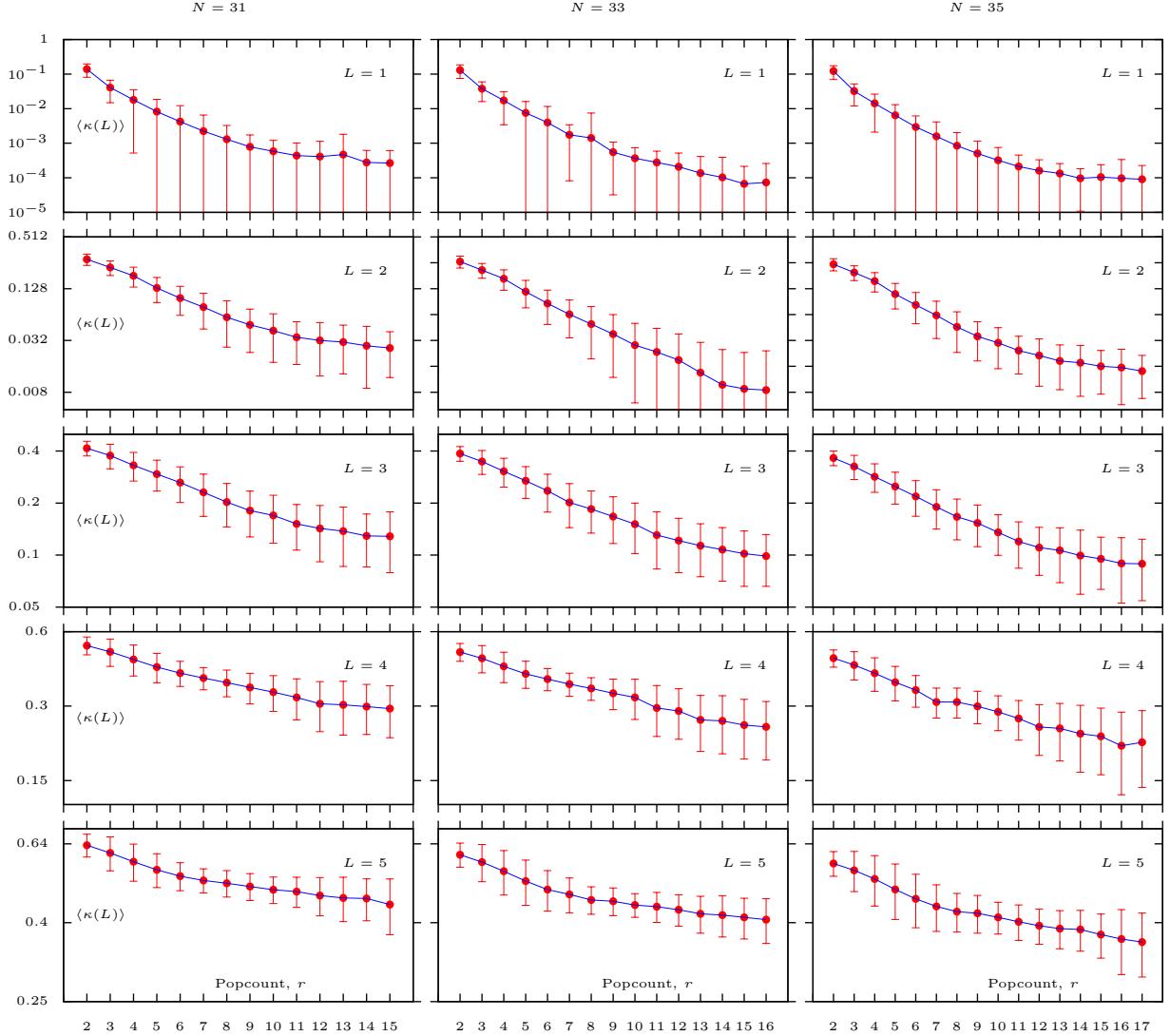


Fig. 4. Averages  $\langle \kappa(L) \rangle$  as functions of  $r$  for  $N = 31, 33, 35$  and  $L$  from 1 to 5. Error bars are shown at the level of one standard deviation,  $\sigma(L)$ , as defined in (12). The vertical axes in all plots are logarithmic. The trivial case  $r = 1$  is not shown.

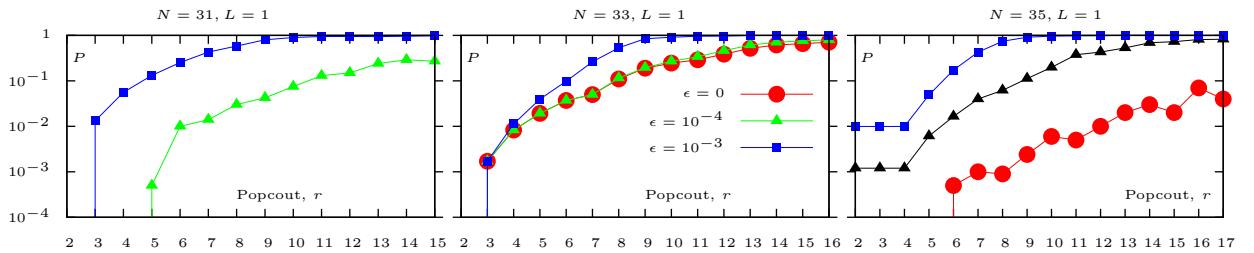


Fig. 5. Cumulative probabilities  $P(\epsilon; 1)$  for  $\epsilon$  as labeled and  $N = 31$  (a),  $N = 33$  (b), and  $N = 35$  (c) as functions of  $r$ . The trivial case  $r = 1$  is not shown.

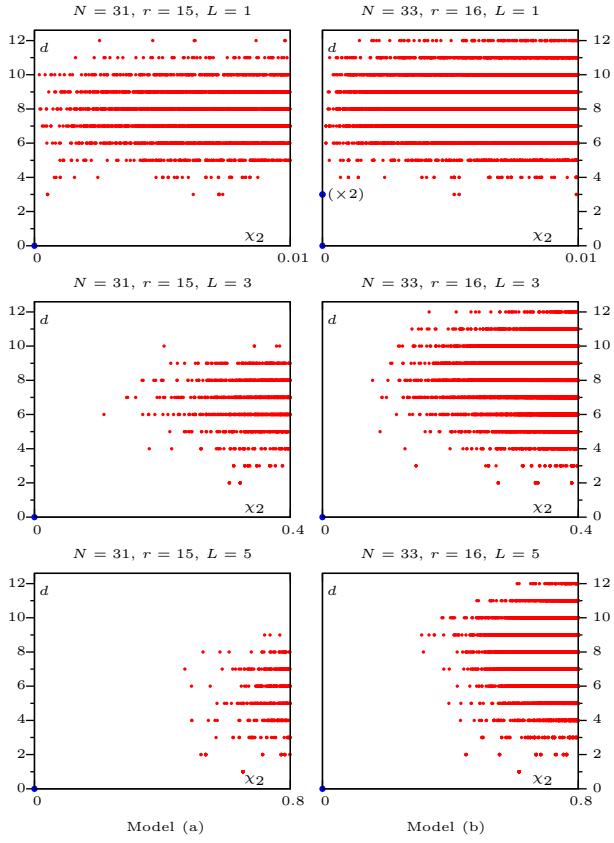


Fig. 6. Real-space distance  $d(\mathbf{x}, \mathbf{x}_{\text{mod}})$  vs Fourier-space distance  $\chi_2(\mathbf{x}, \mathbf{x}_{\text{mod}}; L)$  for the same data points as in Fig. 1. Symbol  $(\times 2)$  indicates that the data point corresponds to two distinct false solutions. Both false solutions have the same real-space distance to the model,  $d = 3$ .

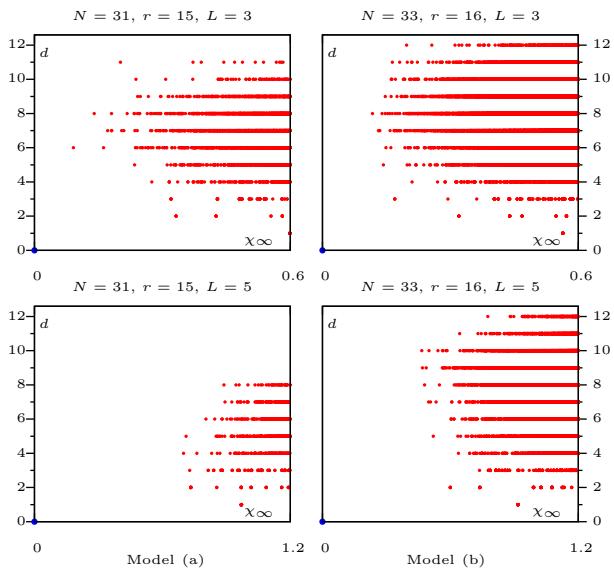


Fig. 7. Same as in Fig. 6 but for  $\chi_\infty$ . The cases  $L = 1$  are not shown since in this case  $\chi_2$  and  $\chi_\infty$  coincide.

inevitably end up in one of the many local minima of  $\chi_2$ , which, in real space, are still very far from the true solution. The difficulty is not removed if we use  $\chi_\infty$  instead of  $\chi_2$ , as is illustrated in Fig. 7. In fact, we have tested numerically several iterative methods that switch one, two or three pairs of 0s and 1s at one iteration step (thus keeping  $r$  unchanged) and try to reduce  $\chi_2$ . None of these approaches were successful.

**Remark 3.** If we could find a vector  $\mathbf{y}$  such that  $d(\mathbf{y}, \mathbf{x}_{\text{mod}}; L) = 1$ , we would be able to find, starting from this result, the true solution (assuming it is unique) in at most  $r(N - r)$  deterministic steps. But as can be seen from the data of Figs. 6 and 7, small  $\chi_p$  does not imply small  $d$ . It is almost as hard to find a vector  $\mathbf{y}$  with the above property as it is to find  $\mathbf{x}_{\text{mod}}$  itself.

Thus, we have encountered a somewhat paradoxical result. The inverse problem can be linear and have a unique and stable solution; yet, any method that updates iteratively  $\mathbf{x} \in \Omega(N, r)$  in an attempt to minimize  $\chi_p$  does not work. The mathematical reason for this difficulty is that the forward DFT maps a discrete set  $\Omega(N, r)$  onto a continuous linear space of Fourier data; not every point in the data space is an image of an element in  $\Omega(N, r)$ . Iterative methods that work well for continuous maps do not work here. However, we describe below two approaches to solving the inverse problem that work reasonably well. One is combinatorial and the other is based on non-convex optimization of a continuous map.

#### A. Combinatorial algorithm

For relatively small values of  $N$  (i.e.,  $\lesssim 100$ ), we can use the following approach. To generate an initial guess, we first compute the low-path filtered (up to  $L$ ) inverse DFT of the forward data  $\mathbf{g} = (g_1, \dots, g_N)$ , defined as

$$g_n = \frac{1}{N} \sum_{m=-L}^L \tilde{x}_m e^{-i\xi nm}, \quad (15)$$

where we have used all the available data points  $\tilde{x}_m$  including  $\tilde{x}_0 = r$ . Note that  $\mathbf{g}$  is consistent with the data but not binary. We then round off the  $r$  largest elements of  $\mathbf{g}$  to 1 and the rest to 0. We refer to this procedure as to “roughening” and write  $\mathbf{b} = \mathcal{R}[\mathbf{g}]$ , where  $\mathcal{R}[\cdot]$  is the roughening operator. The result is a binary vector  $\mathbf{b}$  with the correct length and popcount, so that  $\mathbf{b} \in \Omega(N, r)$ . However,  $\mathbf{b}$  is not consistent with the rest of the data points.

Next, starting from  $\mathbf{x} = \mathbf{b}$ , we invoke a recursive procedure that seeks a vector  $\mathbf{x}_{\text{rec}}$  such that  $\chi_p(\mathbf{x}_{\text{rec}}, \mathbf{x}_{\text{mod}}; L) \leq \epsilon$ , where  $\epsilon$  is a pre-determined small constant. The recursion can be programmed in such a way that the algorithm first tests all vectors  $\mathbf{x}$  such that  $d(\mathbf{x}, \mathbf{b}) = 1$ , then all vectors such that  $d(\mathbf{x}, \mathbf{b}) = 2$ , etc., until the solution is found. To guarantee finding the solution, the recursion must be run to the maximum depth  $d_{\max} = d(\mathbf{x}_{\text{mod}}, \mathbf{b})$ . This underscores the importance of making the initial guess  $\mathbf{b}$  as close to the true solution as possible. In Fig. 8, we plot the average values  $\langle d_{\max} \rangle$  and the corresponding standard deviations for  $10^4$  random model vectors of the length  $N = 61$  and different  $L$  as functions of the popcount  $r$ . It can be seen that the typical values

of  $d_{\max}$  are significantly smaller than  $r$  and, as expected, decrease with  $L$ . The maximum value of  $d_{\max}$  is theoretically equal to  $r$ , but such cases are statistically rare and are further suppressed when  $L > 1$ . We emphasize that the results shown in Fig. 8 pertain to random model vectors without any particular structure. The vectors in which 1s are grouped, i.e., representing a single rectangular pulse or a few such pulses tend to have smaller values of  $d_{\max}$ .

The computational complexity of the algorithm (the number of vectors that must be tested to find the solution assuming no *a priori* knowledge of  $d_{\max}$ ) scales as  $O(C)$ , where

$$C = \sum_{k=1}^{d_{\max}} \frac{r!(N-r)!}{(k!)^2(r-k)!(N-r-k)!} . \quad (16)$$

This function is illustrated in Fig. 9 for several values of  $N$  and  $r$ . It can be seen that the complexity can be much smaller than that of global optimization [that is, testing all vectors in  $\Omega(N, r)$ ] if  $d_{\max}$  is in some sense small. For example, for  $N = 61$ , the complexity is still manageable and well below that of global optimization for  $d_{\max} \lesssim 10$  and all possible values of  $r$ . For a given set of parameters  $N, r, L$ , one can refer to Fig. 8 or similar easily computable data sets to be informed of what values of  $d_{\max}$  can be expected. It should be kept in mind that the complexity given in (16) is reached only in the worst-case scenario; in practice, the solution is almost always found in fewer steps.

The algorithm described in this subsection was able to find all three model vectors (a), (b) and (c) defined above with  $L = 1$  by using the condition  $\chi_\infty \leq \epsilon = 10^{-6}$ . We note that IP(33, 16, 1) is not uniquely solvable but the algorithm has found the true solution anyway. This occurred by chance; the search could have ended up with one of the two false solutions. Also, IP(35, 17, 1) is not uniquely solvable but, as was mentioned in Remark 1, the particular model vector (c) is uniquely recoverable with  $L = 1$ . Of course, as  $L$  is increased, the value of  $d_{\max}$  tends to decrease, and the reconstruction becomes faster. In Fig. 10, we show the typical reconstruction steps for the model vector (c) using  $L = 5$ . The band-limited inverse DFT,  $g$ , does not have a well defined structure but, when roughened, is only 4 pair-wise swipes of 0s and 1s away from the model so that  $d_{\max} = 4$ . At this value of  $d_{\max}$ , combinatorial inversion is very fast. However, for the parameters considered, the inversion also works in a reasonable time (on the order of one minute) for  $L = 1$ , when  $d_{\max} = 8$  (data are not shown).

### B. Non-convex optimization

When  $N$  increases past  $\sim 100$ , the combinatorial approach of the previous subsection may become impractical. In particular, it is hardly applicable to two-dimensional images. We now describe an alternative approach that relies on the continuity of the conventional (unrestricted and unconstrained) DFT to define an iterative scheme that does not involve combinatorial complexity. Let  $v = (v_1, \dots, v_N)$  be a general real-valued

unconstrained vector of length  $N$ . We define the cost function  $F[v]$  that quantifies how close  $v$  is to a binary vector as

$$F[v] = \sum_{n=1}^N v_n^2(v_n - 1)^2 . \quad (17)$$

We then seek to minimize  $F[v]$  while keeping  $v$  consistent with the data. To this end, we start with the band-limited initial guess  $v = g$  [defined in (15)] and update  $v$  iteratively according to

$$v \leftarrow v + q , \quad (18a)$$

where  $q = (q_1, \dots, q_N)$  is of the form

$$q_n = \sum_{m=L+1}^M [c_m \cos(\xi mn) + s_m \sin(\xi mn)] . \quad (18b)$$

Here  $c_m$  and  $s_m$  are some real-valued coefficients, which must be determined at each iteration step independently. It can be seen that the vector  $v$  updated according to (18) remains consistent with the data. To determine  $q$ , we use the steepest descent approach. The derivatives of  $F[v + q]$  with respect to the set of coefficients  $c_m, s_m$  evaluated at  $q = 0$  are given by

$$F_m^{(c)}[v] \equiv \left. \frac{\partial F[v + q]}{\partial c_m} \right|_{q=0} = 2 \sum_{n=1}^N u_n \cos(\xi mn) , \quad (19a)$$

$$F_m^{(s)}[v] \equiv \left. \frac{\partial F[v + q]}{\partial s_m} \right|_{q=0} = 2 \sum_{n=1}^N u_n \sin(\xi mn) , \quad (19b)$$

where

$$u_n = v_n(v_n - 1)(2v_n - 1) . \quad (20)$$

Therefore, we have determined the gradient of  $F[v]$  with respect to the unknown coefficients  $c_m$  and  $s_m$ . Denoting the gradient vector by  $p = (p_1, \dots, p_N)$ , we have for the components:

$$p_n = \sum_{m=L+1}^M \left[ F_m^{(c)}[v] \cos(\xi mn) + F_m^{(s)}[v] \sin(\xi mn) \right] . \quad (21)$$

According to the general strategy of steepest descent, we select  $q = \alpha p$  in the iteration step (18a), where  $\alpha$  is a scalar to be determined. By direct substitution, we find that

$$f(\alpha) \equiv \frac{\partial F[v + \alpha p]}{\partial \alpha} = A_0 + A_1 \alpha + A_2 \alpha^2 + A_3 \alpha^3 , \quad (22a)$$

where

$$A_0 = \sum_{n=1}^N v_n(v_n - 1)(2v_n - 1)p_n = \sum_{n=1}^N u_n p_n , \quad (22b)$$

$$A_1 = \sum_{n=1}^N [2v_n(v_n - 1) + (2v_n - 1)^2] p_n^2 , \quad (22c)$$

$$A_2 = 3 \sum_{n=1}^N (2v_n - 1)p_n^3 , \quad A_3 = 2 \sum_{n=1}^N p_n^4 . \quad (22d)$$

The function  $F[v + \alpha p]$  is a fourth-order polynomial in  $\alpha$  with a positive senior coefficient. Consequently, either it has

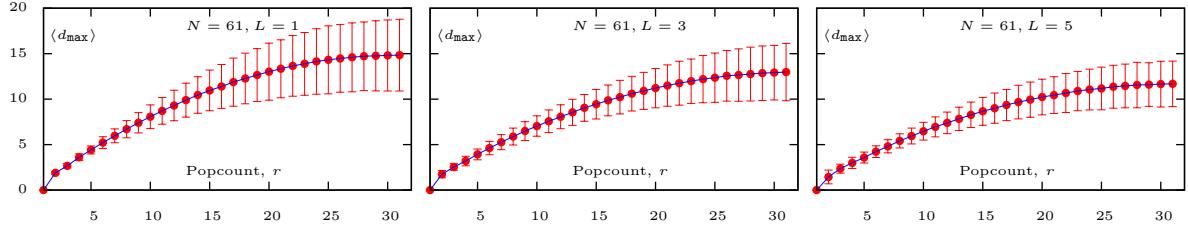


Fig. 8. Average values  $\langle d_{\max} \rangle$  for  $10^4$  random model vectors of length  $N = 61$  and different  $L$  as functions of  $r$ . Error bars are shown at the level of one standard deviation.

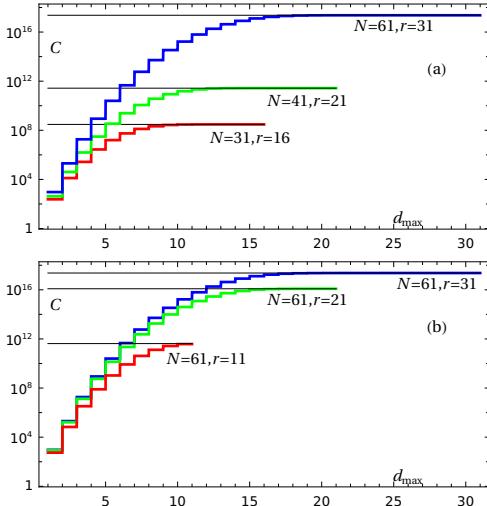


Fig. 9. Computational complexity of the combinatorial optimization algorithm described in Section VI-A as a function of the distance between the initial guess and the solution,  $d_{\max}$ , for several values of  $N$  and  $r$ , as labeled. Thin black lines indicate the computational complexity of global optimization, i.e., testing all vectors in  $\Omega(N, r)$ . Panel (a) shows the dependence for several values of  $N$  and the maximum value of  $r$  for each  $N$ . Panel (b) shows the dependence for  $N = 61$  and several values of  $r$ . The blue lines for  $N = 61, r = 31$  in the two panels are identical.

one real minimum or two minima and one maximum. These special values of  $\alpha$  can be determined by solving the cubic equation  $f(\alpha) = 0$ . If the cubic has a single real root  $\alpha_1$ , we set  $\alpha = \alpha_1$ . If there are three real roots  $\alpha_1 \leq \alpha_2 \leq \alpha_3$ , then  $F[v + \alpha p]$  has a maximum at  $\alpha_2$  and two minima at  $\alpha_1$  and  $\alpha_3$ . To avoid “jumping” over a maximum, we set  $\alpha = \alpha_1$  if  $\alpha_2 > 0$  and  $\alpha = \alpha_3$  otherwise. This completely defines each iteration step and guarantees that  $F[v]$  is decreased in each iteration until a local minimum of  $F[v]$  is reached. The condition for the local minimum is

$$\sum_{n=1}^N u_n \cos(\xi mn) = \sum_{n=1}^N u_n \sin(\xi mn) = 0 \quad \text{for } L < m \leq M,$$

where  $u_n$  are defined in (20). The steepest descent algorithm described here can find a local minimum satisfying the above condition efficiently and with high precision. Note that, in most iteration steps, the cubic has only one real root. The problem is however that  $F[v]$  has many local minima. It is unlikely that any given local minimum is close to the true solution.

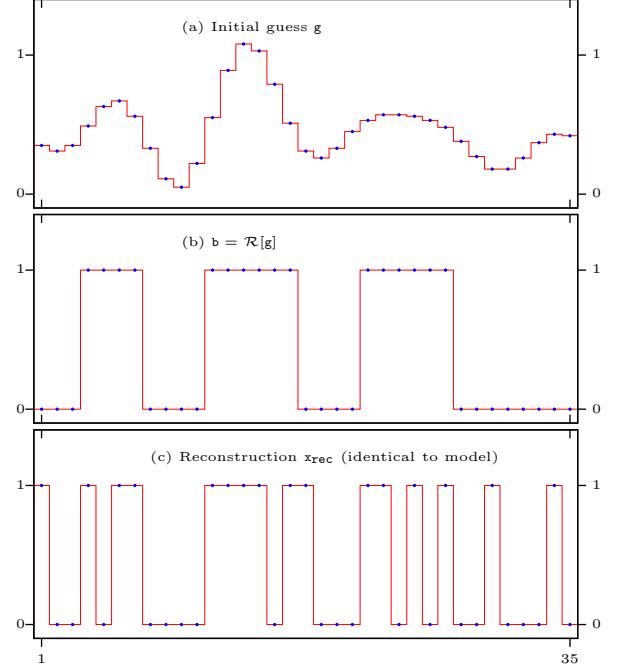


Fig. 10. Intermediate steps in the reconstruction with  $N = 35$ ,  $r = 17$  and  $L = 5$ . Initial guess  $g$  (a), initial guess after roughening,  $b = R[g]$  (b) and the reconstruction  $x_{\text{rec}}$  (identical to model) (c).

To overcome the above difficulty, we can adopt the following approach. Once a local minimum  $\bar{v}_k$  is found ( $k$  labels different local minima), we compute the cost function  $\bar{F}_k = F[\bar{v}_k]$  and check the condition  $\bar{F}_k \leq \epsilon$ , where  $\epsilon$  is some pre-determined small constant. If the condition holds, we apply the roughening operation  $x_{\text{rec}} = \mathcal{R}[\bar{v}_k]$  and assume that the binary vector  $x_{\text{rec}}$  is the solution. If the condition does not hold, we perturb  $\bar{v}_k$  by adding to it a random vector  $q_{\text{rand}}$  of the form (18b). In this way, we select a new initial guess for  $v$ , from which we run the steepest descent algorithm again. Depending on the length and direction of  $q_{\text{rand}}$ , we will end up either in the same or a different local minimum. If, after some number of trials, new local minima are not found, we can increase the average length of  $q_{\text{rand}}$ . Generally, this approach samples many local minima of  $F[v]$  until one is found that satisfies  $\bar{F}_{k_0} \leq \epsilon$ . Once this condition is met, we compute  $x_{\text{rec}} = \mathcal{R}[\bar{v}_{k_0}]$  and assume that  $x_{\text{rec}}$  is the solution. Note that, in the case of ideal data,  $\bar{F}_{k_0}$  can be very small and the vector  $\bar{v}_{k_0}$  very close to being binary. In this case, the roughening

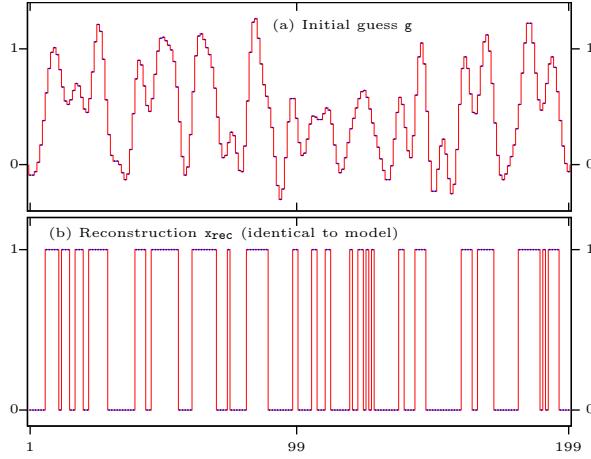


Fig. 11. Reconstruction with  $N = 199$ ,  $r = 90$  and  $L = 29$ . Initial guess (a) and the reconstruction  $x_{\text{rec}}$  (b).

operation is not truly required or rather is trivial. However, with non-ideal data, the roughening operation is required to obtain a binary vector.

To illustrate the method, we have reconstructed a vector of the length  $N = 199$  with  $r = 90$  using  $L = 29$  (so that the ratio  $L/M = 29/99 \approx 1/3$ ) and the stop condition  $\bar{F}_k \leq \epsilon = 10^{-6}$ . The initial guess  $g$  for this simulation is shown in Fig. 11(a) and the reconstruction (identical to the model) in Fig. 11(b). If we apply the roughening operation directly to the initial guess,  $b = \mathcal{R}[g]$ , we would obtain  $d_{\max} = d(b, x_{\text{mod}}) = 8$ . Although the roughened initial guess and the model are not too far apart in real space, the complexity of finding the solution by the combinatorial method of Section VI-A is  $\sim 3 \cdot 10^{22}$  operations. This is beyond the reach for any realistic computer. The optimization method of this subsection however finds the solution in under one minute on an average notebook.

Note that  $N = 199$  is prime and  $\text{IP}(199, r, L)$  is uniquely solvable for any  $L > 0$  and  $0 < r \leq 100$ . However, finding the inverse solution numerically with  $L < 29$  for the model vector displayed in Fig. 11(b) proved to be difficult. Perhaps, the dimensionality of the space of unknown coefficients is too large in this case. It is difficult to state a general condition of convergence to the true solution or the computational complexity of the method described in this subsection. Selection of the random vectors  $q_{\text{rand}}$  can be done in many different ways using different probability distribution functions. This can influence convergence and efficiency. However, we do know for sure that, if the condition  $\bar{F}[v_k] \leq \epsilon$  is not satisfied, then  $\bar{v}_k$  is not the solution.

## VII. DISCUSSION AND CONCLUSIONS

We have shown that binary compositional constraints are powerful priors that allow one to obtain a well-pronounced super-resolution effect. In particular, we have proved that a band limited DFT of a binary vector of length  $N$  is uniquely invertible from the knowledge of just two complex DFT coefficients – the zeroth and the first – if  $N$  is prime. If  $N$  has two prime factors, then Theorem 2 tells us how many additional DFT coefficients must be known to guarantee uniqueness.

The above result can be useful if applied to analysis of two-dimensional images.

Although Theorems 1 and 2 establish the sufficient condition for invertibility assuming only that the unknown vector is consistent with the data, many (but not all) binary vectors are uniquely recoverable even if conditions of the Theorems do not hold.

We have further investigated stability of inversion and conditions under which reconstructing a binary vector from a limited set of DFT coefficients is practically feasible. Preliminary numerical data indicate that stable reconstructions can be obtained when about  $1/3$  of all DFT coefficients are known. This entails an approximately three-fold super-resolution effect.

Although binarity is a powerful constraint, devising practical reconstruction algorithms is a nontrivial task. We provided and demonstrated two such algorithms in the paper. The first (combinatorial) algorithm is guaranteed to find a solution if run consistently but is NP-hard. Still, it involves fewer or much fewer operations than global optimization. This approach is applicable to vectors with  $N \lesssim 100$ . For longer vectors, we have developed an algorithm based on optimization of the non-convex cost function  $F[\cdot]$  (17). The local minima of this function can be easily found using the steepest descent approach. However, not every local minimum of a non-convex function is a global minimum. To overcome this difficulty, we have introduced random perturbations, which allow one to sample many local minima and finally find one with sufficient depth, which then coincides with the global minimum. We note that convexification of the inverse problem, that is, finding a relevant convex cost function, proved to be difficult and is likely impossible.

Although we have obtained numerical reconstructions for vectors up to  $N = 199$  with about  $1/3$  of DFT coefficients considered to be known, it is clear that the developed algorithms are not optimal and leave a lot of space for improvement. We hope that, with further refinements, these results can be used to obtain super-resolution effect in two-dimensional black-and-white images.

## REFERENCES

- [1] A. A. Maznev and O. B. Wright, "Upholding the diffraction limit in the focusing of light and sound," *Wave Motion*, vol. 68, p. 182–189, 2017.
- [2] M. Born and E. Wolf, *Principles of Optics*. Cambridge, 1999.
- [3] Y. Romano, M. Elad, and M. Peyman, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imag. Sci.*, vol. 10, p. 1804–1844, 2017.
- [4] D. Watzinger, *e & i Elektrotechnik und Informationstechnik*. Springer, 2007, vol. 124, no. 7-8, ch. Bayesian inference for inverse problems - statistical inversion, pp. 240–247.
- [5] H. W. Engl and P. Kugler, *Multidisciplinary Methods for Analysis Optimization and Control of Complex Systems Mathematics in Industry*. Springer, 2005, vol. 6, ch. Nonlinear inverse problems: Theoretical aspects and some industrial applications, pp. 3–47.
- [6] R. L. Lagendijk and J. Biemond, *Iterative identification and restoration of images*. Springer, 1990.
- [7] X. Liu, D. Zhai, D. Zhao, G. Zhai, and W. Gao, "Progressive image denoising through hybrid graph laplacian regularization: A unified framework," *IEEE Trans. Imag. Proc.*, vol. 23, pp. 1491–1503, 2014.
- [8] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

- [9] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imag. Sci.*, vol. 2, p. 183–202, 2009.
- [10] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Info. Theor.*, vol. 53, pp. 4655–4666, 2007.
- [11] T. Blumensath, “Accelerated iterative hard thresholding,” *Sign. Proc.*, vol. 92, pp. 752–756, 2012.
- [12] ———, “Compressed sensing with nonlinear observations and related nonlinear optimization,” *IEEE Trans. Info. Theor.*, vol. 59, pp. 3466–3474, 2013.
- [13] B. Deutsch, R. Reddy, D. Mayerich, H. Bhagrava, and P. S. Carney, “Compositional prior information in computed infrared spectroscopic imaging,” *J. Opt. Soc. Am. A*, vol. 32, pp. 1126–1131, 2015.
- [14] L. Pfister, R. Bhagrava, Y. Bresler, and P. S. Carney, “Composition-aware spectroscopic tomography,” *Inverse Problems*, 2020, in press.
- [15] Z. Liang, “Spatiotemporal imaging with partially separable functions,” in *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2007, pp. 988–991.
- [16] A. Corlu, J. Durduran, R. Choe, M. Schweiger, E. Hillman, S. Arridge, and A. Yodh, “Uniqueness and wavelength optimization in continuous-wave multispectral diffuse optical tomography,” *Opt. Lett.*, vol. 28, no. 23, pp. 2339–2402, 2003.
- [17] ———, “Diffuse optical tomography with spectral constraints and wavelength optimization,” *Appl. Opt.*, vol. 44, no. 11, pp. 2082–2090, 2005.
- [18] T. Zhang, C. Godavarthi, P. C. Chaumet, G. Maire, H. Giovannini, A. Talneau, M. Allain, K. Belkebir, and A. Sentenac, “Far-field diffraction microscopy at  $\lambda/10$  resolution,” *Optica*, vol. 3, p. 609, 2016.
- [19] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedland, and S. Roy, “Level-set methods for convex optimization,” *Math. Prog.*, vol. 174, pp. 359–390, 2018.
- [20] K. Ireland, M. Rosen, and M. Rosen, *A classical introduction to modern number theory*, ser. Graduate texts in mathematics. Springer, 1990.
- [21] H. B. Mann, “On linear relations between roots of unity,” *Mathematika*, vol. 12, pp. 107–117, 1965.
- [22] J. Conway and A. Jones, “Trigonometric Diophantine equations (on vanishing sums of roots of unity),” *Acta Arithmetica*, vol. 30, pp. 229–240, 1976.
- [23] H. W. Lenstra, “Vanishing sums of roots of unity,” in *Proc. Bicentennial Congress Wiskundig Genootschap, Part II*. Vrije Univ. Amsterdam, 1978, pp. 249–268.
- [24] T. Y. Lam and K. H. Leung, “Vanishing sums of  $m$ th roots of unity in finite fields,” *Finite Fields and Their Applications*, vol. 2, pp. 422–438, 1996.
- [25] ———, “On vanishing sums of roots of unity,” *J. Algebra*, vol. 224, pp. 91–109, 2000.

## APPENDIX A PROOF OF LEMMA 1

Suppose  $x$  is not pairwise 1-distinguishable from some distinct vector  $y$  in  $\Omega(N, r)$ . By definition, we have  $\tilde{x}_1 = \tilde{y}_1$ . Following the same notation as in the proof of Theorem 1, we define  $z = x - y$  and conclude that (9) must hold, where each  $z_n$  takes a value in  $\{0, \pm 1\}$ . As  $x \neq y$ , not all entries of  $z_n$  can be 0. Furthermore, as  $x$  and  $y$  are binary vectors with the same popcount, by construction  $\tilde{z}_0 = 0$ .

We will show that  $z$  either has an equivalent number of  $p$ -gons and “negative”  $p$ -gons (a polygon with entries of  $-1$ ), or that  $z$  has an equivalent number of  $q$ -gons and “negative”  $q$ -gons. As each positive  $p$ - or  $q$ -gon in  $z$  corresponds to a  $p$ - or  $q$ -gon in  $x$  (and an empty  $p$ - or  $q$ -gon in  $y$ ), and each negative  $p$ - or  $q$ -gon in  $z$  corresponds to an empty  $p$ - or  $q$ -gon in  $x$  (and a  $p$ - or  $q$ -gon in  $y$ ), this will prove the forward direction.

We consider two cases:  $p = q$  and  $p \neq q$ . First, consider the case  $p \neq q$ . Letting  $\zeta_p$  and  $\zeta_q$  be primitive roots of unity of  $p$ -

th and  $q$ -th order, respectively, we rewrite (9) using Theorem 2.3 in [23] as

$$0 = \sum_{\ell=1}^q \sum_{k=1}^p z_{(k\ell)} \zeta_p^k \zeta_q^\ell. \quad (23)$$

Here we have introduced the composite index  $(k\ell)$  from 1 to  $pq$ . Note that as  $\zeta_p^k \zeta_q^\ell = \zeta_N^{p\ell+qk}$ , by fixing either a value of  $\ell$  or  $k$ , we vary over roots of unity to make a regular  $p$ - or  $q$ -gon, respectively. We can now apply Lemma 1 of [22] to conclude that (23) holds iff the inner sum is constant for all  $\ell$ . That is, for all  $\ell$ ,

$$\sum_{k=1}^p z_{(k\ell)} \zeta_p^k = \sum_{k=1}^p z_{(k1)} \zeta_p^k.$$

Similar to the proof of Theorem 1, as  $\zeta_p$  is a root of unity of a prime order, this equation can only hold if, for each  $\ell$ , there is a fixed constant difference between  $z_{(k\ell)}$  and  $z_{(k1)}$  for all  $k$ . Thus we have, for all  $k$  and  $\ell$ ,

$$z_{(k\ell)} - z_{(k1)} = C_\ell, \quad (24)$$

where each  $C_\ell \in \{0, \pm 1, \pm 2\}$ . Now, by summing over all  $k$  and  $\ell$ , we obtain

$$\begin{aligned} \sum_{k=1}^p \sum_{\ell=1}^q (z_{(k\ell)} - z_{(k1)}) &= \sum_{k=1}^p \sum_{\ell=1}^q C_\ell, \\ -q \sum_{k=1}^p z_{(k1)} &= p \sum_{\ell=1}^q C_\ell, \end{aligned}$$

where, in the second line, we have used the fact that  $\tilde{z}_0 = 0$ . As  $p$  and  $q$  are relatively prime,  $\sum z_{(k1)}$  must be a multiple of  $p$ . As each  $z_n$  is in  $\{0, \pm 1\}$ , we must have  $\sum z_{(k1)}$  in  $\{0, \pm p\}$ . We break this into two cases.

If  $\sum z_{(k1)} = \pm p$ , then  $z_{(k1)}$  is constant for all  $k$  with a value of 1 or -1. By our indexing and Definition 4,  $z$  contains either a  $p$ -gon or a negative  $p$ -gon. By (24), for  $2 \leq \ell \leq q$ ,  $z_{(k\ell)}$  must form either a  $p$ -gon, an empty  $p$ -gon, or a negative  $p$ -gon. As  $\tilde{z}_0 = 0$ , there must be an equivalent number of positive and negative  $p$ -gons, as desired.

If  $\sum z_{(k1)} = 0$ , we first note that, if all the  $z_{(k1)} = 0$ , then since not all entries of  $z_n$  can be 0, there must be at least one nonzero value of  $C_\ell$  in (24). This yields either a positive or negative  $p$ -gon. Identical reasoning ( $\tilde{z}_0 = 0$ ) leads to an equivalent number of positive and negative  $p$ -gons in this case.

The remaining option is for  $z_{(k1)}$  to have equal number of +1 and -1’s. In this case, as  $z_{(k\ell)}$  can only have entries in  $\{0, \pm 1\}$ ,  $C_\ell$  must be equal to 0 in Eq. (24) for all  $\ell$ . By our indexing, values of  $k$  for which  $z_{(k1)} = \pm 1$  yield  $q$ -gons or “negative”  $q$ -gons, respectively. As we have equal number of +1 and -1 values, we have an equivalent number of positive and negative  $q$ -gons, finishing the proof in the  $p \neq q$  case.

For the  $p = q$  case, we apply Theorem 2.2 in [23], which states that the sum in (9) vanishes iff, for all  $1 \leq j \leq p$ ,

$$0 = \sum_{n=1}^p z_{(jn)} \zeta_p^n,$$

where  $(jn)$  is a composite index with  $n$  referring to  $p$ -periodic locations in the original vector  $\mathbf{z}$ . As in the proof of Theorem 1, this implies that for each  $j$ , either  $z_{(jn)}$  is constant for all  $n$  with a value of 0 or  $\pm 1$ . Similar to the previous case, as not all entries  $z_n$  can be 0, there must be at least one value of  $j$  for which  $z_{(jn)} = \pm 1$ . Applying the condition  $\tilde{z}_0 = 0$  then results in equivalent number of positive and negative  $p$ -gons.

For the reverse direction, we note that, if  $\mathbf{x}$  contains pairs of  $p$ - or  $q$ -gons and empty  $p$ - or  $q$ -gons, then  $\mathbf{y}$ , which flips the full and empty  $p$ - or  $q$ -gons and agrees at all other entries with  $\mathbf{x}$ , satisfies  $\tilde{x}_0 = \tilde{y}_0$  and  $\tilde{x}_1 = \tilde{y}_1$ .

## APPENDIX B PROOF OF LEMMA 2

The case  $L=1$  is a direct consequence of Lemma 1. For  $L > 1$ , Lemma 1 states that, if  $\mathbf{y}$  is not 1-distinguishable from  $\mathbf{x}$ , then  $\mathbf{z} = \mathbf{x} - \mathbf{y}$  has a pair of positive and negative  $L$ -gons. This may not be the only such pair of  $L$ -gons. However, the proof of Lemma 1 showed that  $\mathbf{z}$  cannot also have any positive and negative  $L'$ -gon pairs for  $L' \neq L$ . We will rewrite the DFT coefficients of  $\mathbf{z}$  by grouping the roots of unity which form each  $L$ -gon together. (with all other entries being 0). Each  $L$ -gon uses the roots (for some  $\ell$ )

$$\{\zeta_N^{\ell+kN/L} : 1 \leq k \leq L\} = \{\zeta_N^\ell \zeta_L^k : 1 \leq k \leq L\}. \quad (25)$$

We thus write

$$\tilde{z}_n = \left( \sum_{\ell \in \mathcal{S}^+} \zeta_N^\ell - \sum_{\ell \in \mathcal{S}^-} \zeta_N^\ell \right) \sum_{k=1}^L \zeta_L^{nk}, \quad (26)$$

where  $\mathcal{S}^+$  and  $\mathcal{S}^-$  are index sets that track the appropriate positive and negative coefficients for the polygons. As the polygons come in pairs,  $|\mathcal{S}^+| = |\mathcal{S}^-|$ . We claim that

$$\sum_{k=1}^L \zeta_L^{nk} = \begin{cases} L, & n = L \\ 0, & 1 \leq n < L \end{cases} \quad (27)$$

The case  $n = L$  is straightforward. As  $\zeta_L$  is an  $L$ -th root of unity,  $\zeta_L^{kL} = 1$  for all  $k$ . For  $n < L$ , we can analyze the sums using a group theory approach.

As  $\zeta_L$  is a generator of the cyclic group of order  $L$ , we have  $\zeta_L^{nk} = \zeta_L^{nk'}$  iff  $nk \equiv nk' \pmod{L}$  [20]. This is equivalent to

$$k \equiv k' \pmod{\frac{L}{\gcd(L, n)}}. \quad (28)$$

As  $L$  must be prime (it is either  $p$  or  $q$ ) and  $n < L$ ,  $\gcd(L, n) = 1$ . As  $1 \leq k, k' \leq L$ , (28) implies that  $\zeta_L^{nk} \neq \zeta_L^{nk'}$  for all  $k$  and  $k'$ . Thus, (27) is just a permuted sum of all  $L$ -th roots of unity, which is still 0.

With (27) now verified, we can apply this to (26) as

$$\tilde{z}_n = \begin{cases} L \left( \sum_{\ell \in \mathcal{S}^+} \zeta_N^\ell - \sum_{\ell \in \mathcal{S}^-} \zeta_N^\ell \right), & n = L \\ 0, & 1 \leq n < L \end{cases}. \quad (29)$$

All that is left to do is to verify that the sum for  $\tilde{z}_L$  in (29) is nonzero. This is a sum of roots of unity of  $N$ th order with equal number of coefficients of  $\pm 1$ . As in the proof of Lemma 1, this term vanishes iff  $\zeta_N^\ell \in \mathcal{S}^+$  and the  $\zeta_N^\ell \in \mathcal{S}^-$  are all used to form an equivalent number of  $p$ - and  $q$ -gons.

If  $\zeta_N^\ell \in \mathcal{S}^+$ , then  $\zeta_N^{\ell+kN/L} \notin \mathcal{S}^+$  for all  $k$  as these rotations would give the same  $L$ -gon. Thus by the form given by (25),  $\mathcal{S}^+$  cannot have any  $L$ -gons. Likewise,  $\mathcal{S}^+$  cannot have any  $L'$ -gons, as this would require  $|\mathcal{S}^+| = L'$ . This implies there are  $L'$  many  $L$ -gons in  $\mathbf{z}$ . As  $L'$  can only be  $p$  or  $q$ ,  $LL' = N$  implying that  $\tilde{z}_0 = N$  which is not true. Identical reasoning applies to  $\mathcal{S}^-$ , and hence  $\tilde{z}_L \neq 0$ . Therefore,  $\mathbf{x}$  is pairwise  $L$ -distinguishable from all other vectors in  $\Omega(N, r)$ .