RECONSTRUCTION OF INTEGERS FROM PAIRWISE DISTANCES

Kishore Jaganathan *Babak Hassibi*

Department of Electrical Engineering
California Institute of Technology, Pasadena

ABSTRACT

Given a set of integers, one can easily construct the set of their pairwise distances. We consider the inverse problem: given a set of pairwise distances, find the integer set which realizes the pairwise distance set. This problem arises in a lot of fields in engineering and applied physics, and has confounded researchers for over 60 years. It is one of the few fundamental problems that are neither known to be NP-hard nor solvable by polynomial-time algorithms. Whether unique recovery is possible also remains an open question.

In many practical applications where this problem occurs, the integer set is naturally sparse (i.e., the integers are sufficiently spaced), a property which has not been explored. In this work, we exploit the sparse nature of the integer set and develop a polynomial-time algorithm which provably recovers the set of integers (up to linear shift and reversal) from the set of their pairwise distances with arbitrarily high probability if the sparsity is $O(n^{1/2-\epsilon})$. Numerical simulations verify the effectiveness of the proposed algorithm.

Index Terms— phase retrieval, turnpike problem, sparse signals

1. INTRODUCTION

We consider the problem of reconstructing a set of integers from the set of their pairwise distances. For example, consider the set $V = \{2, 5, 13, 31, 44\}$. Its pairwise distance set is given by $W = \{0, 3, 8, 11, 13, 18, 26, 29, 31, 39, 42\}$. We look at the problem of recovering the integer set V from the pairwise distance set W[2].

This recovery problem dates back to the origins of the classical phase retrieval problem in the 1930s [1, 2] and has received a lot of attention from researchers. More recently, it has arisen in computational biology, specifically in restriction site mapping of DNA [3]. This problem has also been posed as a computational geometry problem [4].

1.1. Phase Retrieval

Many measurement systems in practice can output only the squared-magnitude of the Fourier transform. Phase information is completely lost, because of which signal recovery is difficult. This is a fundamental problem in many fields, including optics [5], X-ray crystallography [6], astronomical imaging [7], speech processing [8], particle scattering, electron microscopy etc.

Recovering a signal from its Fourier transform magnitude is known as phase retrieval. Since squared-magnitude of the Fourier transform and autocorrelation are Fourier pairs, the phase retrieval problem can be equivalently posed as recovering a signal from its autocorrelation.

Let $\mathbf{x} = \{x_0, x_1, ...x_{n-1}\}$ be a discrete-time signal of length n and sparsity k, where sparsity is defined as the number of non-zero elements. Its autocorrelation, denoted by $\mathbf{a} = \{a_0, a_1, ...a_{n-1}\}$, is defined as

$$a_i \overset{def}{=} \sum_j x_j x_{j+i} = (\mathbf{x} \star \tilde{\mathbf{x}})_i \tag{1}$$

where $\tilde{\mathbf{x}}$ is the time-reversed version of $\mathbf{x}$. Also, let V and W denote the support set of the signal $\mathbf{x}$ and its autocorrelation $\mathbf{a}$ respectively, defined as

$$V = \{i | x_i \neq 0\} \qquad \& \qquad W = \{i | a_i \neq 0\} \tag{2}$$

The phase retrieval problem can be written as

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && \mathbf{x} \star \tilde{\mathbf{x}} = \mathbf{a} \end{aligned} \tag{3}$$

Connection to integer recovery problem: It is often useful to be able to reconstruct the support set of the signal V from the support set of the autocorrelation W. In many applications (e.g, astronomy), the signal's support set is the desired information. In other applications, support knowledge makes signal reconstruction process using available techniques significantly easier [9, 10, 11].

We will assume that if $a_i = 0$, then no two elements in $\mathbf{x}$ are separated by a distance i, i.e.,

$$a_i = 0 \Rightarrow x_j x_{i+j} = 0 \; \forall \; j \tag{4}$$

This work was supported in part by the National Science Foundation under grants CCF-0729203, CNS-0932428 and CCF-1018927, by the Office of Naval Research under the MURI grant N00014-08-1-0747, and by Caltech's Lee Center for Advanced Networking.

[2]If V has a pairwise distance set W, then sets $c \pm V$ also have the same pairwise distance set W for any integer c. These solutions are considered equivalent, and in all the applications it is considered good enough if any equivalent solution, i.e., up to linear translation and flipping, is recovered.

This is a very weak assumption and holds with probability one if the non-zero entries of the signal are chosen from a non-degenerate distribution. With this assumption, the support recovery problem can be posed as

$$\begin{aligned} &\text{find} && V \\ &\text{subject to} && \{|i-j| \mid (i,j) \in V\} = W \end{aligned} \tag{5}$$

Note that V is a set of integers, and W is exactly its pairwise distance set.

1.2. Partial Digest Problem

Over the last few years, there has been a lot of interest in DNA restriction site analysis. A DNA strand is a string on the letters $\{A, T, G, C\}$. Unfortunately, the DNA string cannot be explicitly observed and in order to map it, biochemical techniques which provide indirect information have been developed.

When a particular restriction enzyme is added to a DNA solution, the DNA is cut at particular restriction sites. For example, the enzyme *EcoRI* cuts at locations of the pattern GAATTC. The goal of restriction site analysis is to determine the locations of every site for a given enzyme. In order to do this, a batch of DNA is exposed to a restriction enzyme in limited quantity so that fragments of all possible lengths exist (see Figure 1). Using gel electrophoresis, the fragment lengths can be measured.

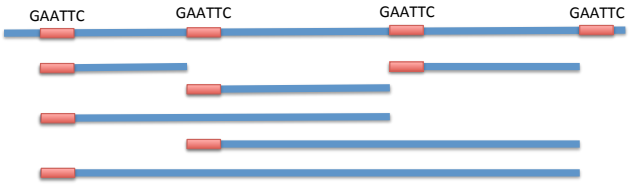

Fig. 1. Partial Digest Problem

Recovering the locations of the restriction sites from the measured fragment lengths is known as the partial digest problem. The locations of the restriction sites correspond to the set of integers V, and the measured fragment lengths correspond to the set of pairwise distances W.

2. CONTRIBUTIONS

Researchers have proposed a wide range of heuristics [12, 13] to solve the phase retrieval problem, a brief summary of which can be found in [14]. [15] provides a theoretical framework to understand the heuristics, which are in essence an alternating projection between a convex set and a non-convex set. The problem with such an approach is that convergence is often to a local minimum (figure 2), hence chances of successful recovery are less. Also, no theoretical guarantees can be provided.

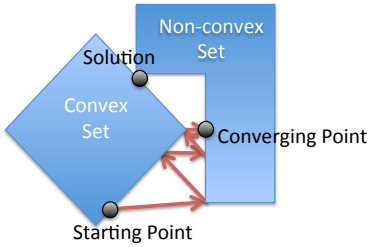

Fig. 2. Alternating projection between a convex and a non-convex set

A variant of the partial digest problem, known as the turnpike problem, which is the problem of recovering an integer set from their pairwise distance multiset (multiplicity information of each pairwise distance also available) is also well studied. The most widely used algorithm to do this recovery is a worst-case exponential algorithm based on backtracking [16]. [17] provides a comprehensive summary of the existing algorithms. The question of unique provable recovery using polynomial-time algorithms remains unanswered, and complicated mechanisms have been used to solve the problem in practice [20, 21].

In many applications of these problems, the underlying signals are naturally sparse. For example, astronomical imaging deals with the locations of the stars in the sky, X-ray crystallography deals with the density of atoms and so on. In DNA restriction site analysis, it is very reasonable to assume that the restriction sites are sparsely distributed.

In our work, we attempt to exploit the sparse nature of the underlying signals. Recently, attempts have been made to exploit sparsity. An alternating projection based heuristic was proposed in [18], a semidefinite relaxation based heuristic was explored in [19]. We develop a polynomial-time algorithm which can provably recover the underlying signals with high probability if the signal is $O(n^{1/2-\epsilon})$ sparse.

3. MAIN RESULT

Suppose $V = \{v_0, v_1, ..., v_{k-1}\}$ is a set of k integers and $W = \{w_0, w_1, ..., w_{K-1}\}$ is its pairwise distance set[3].

Theorem 3.1 (Main Result)**.** *V can be recovered uniquely (upto linear shift and reversal) from W in polynomial-time with probability greater than $1-\delta$ for any $\delta > 0$ if*

(i) $\exists\, n \geq w_{K-1}$ *such that V is chosen uniformly at random from* $\{0, 1, ..., n-1\}$

(ii) $k = O(n^{1/2-\epsilon})$

(iii) $n > n(\epsilon, \delta)$

[3]The elements of V and W are assumed to be in ascending order without loss of generality for convenience of notation, i.e., $v_0 < v_1 < .... < v_{k-1}$ and $w_0 < w_1 < .... < w_{K-1}$

In order to overcome the trivial ambiguity of linear shift and reversal, we attempt to recover the equivalent solution set $U = \{u_0, u_1, ..., u_{k-1}\}$ defined as follows

$$U = \begin{cases} V - v_0 & \text{if} \quad v_1 - v_0 \leq v_{k-1} - v_{k-2} \\ v_{k-1} - V & \text{otherwise} \end{cases} \tag{6}$$

i.e., the equivalent solution set U we attempt to recover has the following properties:

(i) $u_0 = 0$

(ii) $u_1 - u_0 < u_{k-1} - u_{k-2}$

4. ALGORITHM

Let $u_{ij} = |u_i - u_j|$ for $0 \leq i, j \leq k-1$. With this definition, $W = \{u_{ij} : 0 \leq i, j \leq k-1\}$ and $U = \{u_{0j} : 0 \leq j \leq k-1\}$. Note that $U \subseteq W$.

4.1. Intersection Step

The key idea of this step can be summarized as follows: suppose we know the value of u_{0p} for some $1 \leq p \leq k-1$, if U_p and W_p are defined as

$$U_p = \{u_{0j} : p \leq j \leq k-1\} \quad \& \quad W_p = W + u_{0p} \tag{7}$$

then $U_p \subseteq W \cap W_p$. The idea can be extended to multiple intersections. Suppose we know $\{u_{0p} : 1 \leq p \leq t\}$, we can construct $\{W_p : 1 \leq p \leq t\}$ and have

$$U_t \subseteq \left(\bigcap_{p=0}^{t} W_p \right) \tag{8}$$

4.2. Graph Step

For an integer set U whose pairwise distance set is W, consider the set $Z = \{z_0, z_1, ...z_{|Z|-1}\}$ such that $U \subseteq Z \subseteq W$. Construct a graph $G(Z)$ with $|Z|$ vertices such that there exists an edge between z_i and z_j iff the following two conditions are satisfied

(i) $\forall z_g, z_h \in Z, z_g - z_h \neq z_i - z_j \quad$ unless $\quad (i,j) = (g,h)$

(ii) $z_i - z_j \in W$

i.e., there exists an edge between two vertices if their corresponding pairwise distance is unique and belongs to W. For example, consider the integer set $U = \{0, 10, 15, 50\}$ whose pairwise distance set is $W = \{0, 5, 10, 15, 35, 40, 50\}$. Consider the set $Z = \{0, 10, 15, 35, 40, 50\}$. The graph $G(Z)$ looks as shown in Figure 3. Note that there exists an edge between 10 and 40 as it is the only pair of integers with difference 10, there doesn't exist an edge between 0 and 40 as there is another pair $\{10, 50\}$ with difference 40 and so on.

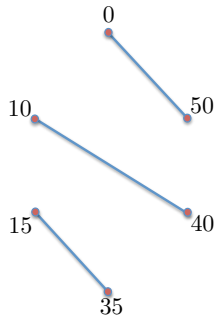

Fig. 3. The graph $G(Z)$ for $Z = \{0, 10, 15, 35, 40, 50\}$, given $W = \{0, 5, 10, 15, 35, 40, 50\}$

The main idea of this step is as follows: suppose we draw a graph $G(Z)$ where $U \subseteq Z \subseteq W$. If there exists an edge between a pair of integers $\{z_i, z_j\} \in Z$ such that $z_i - z_j \in W$, then $\{z_i, z_j\} \in U$. This holds because if $\{z_i, z_j\} \notin U$, then since $z_i - z_j \in W$ there has to be a pair of integers in U which have a pairwise distance $z_i - z_j$, which would contradict the fact that an edge exists between z_i and z_j.

Algorithm 1 Integer Recovery Algorithm

Input: Pairwise distance set W
Output: Integer set U which realizes W

1. Infer u_{01} from W
2. Construct the set $W_1 = W + u_{01}$

   If $k = O(n^{1/4-\epsilon})$
3. Calculate $U_1 = W \cap W_1$
4. Recover $U = \{0\} \cup U_1$

   Else if $k = O(n^{1/2-\epsilon})$
5. Construct the graph $G(\{0\} \cup (W \cap W_1))$ and infer $\{u_{0i_p} : 1 \leq p \leq t = log(k)\}$
6. Construct the set $W_{i_p} = W + u_{0i_p}$ for $1 \leq p \leq t$
7. Calculate $U_{i_t} = W \cap \left(\bigcap_{p=1}^{t} W_{i_p} \right)$
8. Define $\tilde{U} = \{\tilde{u}_0, ...\tilde{u}_{k-1}\}$ as $\tilde{U} = u_{k-1} - U$ and infer $\{\tilde{u}_{0p} : 1 \leq p \leq t\}$ from U_{i_t}
9. Construct the set $\tilde{W}_p = W + \tilde{u}_{0p}$ for $1 \leq p \leq t = log(k)$
10. Calculate $\tilde{U}_t = \left(\bigcap_{p=0}^{t} \tilde{W}_p \right)$
11. Recover $\tilde{U} = \{\tilde{u}_{0p} : 0 \leq p \leq t-1\} \cup \tilde{U}_t$
12. Recover $U = \tilde{u}_{k-1} - \tilde{U}$

5. PROOF OF MAIN THEOREM

In this section, we provide the lemmas necessary to prove the main theorem. Detailed proofs of the lemmas can be found in [22]. Note that the set V of size k is chosen from the set $\{0, 1, ...n-1\}$ uniformly at random.

Lemma 5.1. *u_{01} can be inferred from W.*

Lemma 5.2. *Probability that an integer l belongs to W is*

(i) *1 if $l \in U$.*

(ii) *less than or equal to $\frac{k^2}{n} + o(\frac{k^2}{n})$ if $l \notin U$.*

Lemma 5.3. *The probability that an integer l not in U belongs to $W \cap W_1$ is less than or equal to $\frac{k^4}{n^2} + o(\frac{k^4}{n^2})$*

Lemma 5.4 (Intersection Step)**.** *The probability that an integer l in W and not in U belongs to $W \cap W_1$ is less than $\frac{k^2}{n} + o(\frac{k^2}{n})$.*

Corollary 5.1. *$U_1 = W \cap W_1$ with probability greater than $1 - \delta$ for any $\delta > 0$ if $n > n(\epsilon, \delta)$ and $k = O(n^{1/4-\epsilon})$.*

Lemma 5.5 (Graph Step)**.** *In the graph $G(\{0\} \cup (W \cap W_1))$, integers $\{u_{0p} : 1 \leq p \leq t = log(k)\}$ have an edge with $u_{0,k-1}$ with probability greater than $1 - \delta$ for any $\delta > 0$ if*

(i) *$k = O(n^{1/2-\epsilon})$*

(ii) *$n > n(\epsilon, \delta)$*

Lemma 5.6. *The probability that an integer $l \in W, l \notin U$ belongs to $\left(\bigcap_{p=1}^{t} W_p\right)$ for $t = log(k)$ is less than or equal to $(\frac{k^{2(1+\epsilon)}}{n})^{\sqrt{t}/2} + o(\frac{k^{2(1+\epsilon)}}{n})^{\sqrt{t}/2}$ for $n > n(\epsilon, \delta)$.*

Lemma 5.7. *$U_t = \left(\bigcap_{p=0}^{t} W_p\right)$ with probability greater than $1 - \delta$ for any $\delta > 0$ if*

(i) *$k = O(n^{1/2-\epsilon})$*

(ii) *$t \geq log(k)$, $n > n(\epsilon, \delta)$*

6. NUMERICAL SIMULATIONS

In order to demonstrate the performance of the proposed algorithm, numerical simulations were performed for different values of signal length n and sparsities k. Simulations were performed by choosing k-element subsets V uniformly at random from $\{0, 1, ..., n-1\}$ for different values of n and k. Figure 4 plots the probability of successful recovery for $n = 512$ and $n = 1024$.

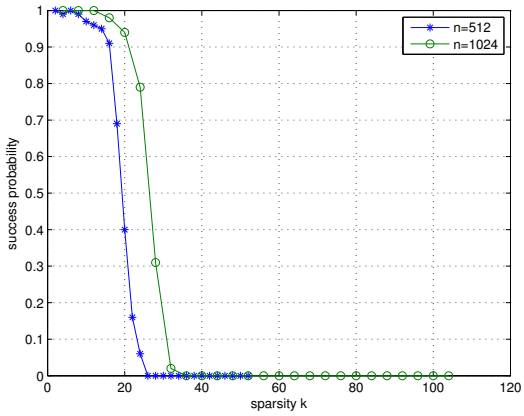

Fig. 4. Probability of successful recovery

In order to understand the various steps in the algorithm, we provide the working details for a particular example:

$$W = \{0, 2, 3, 5, 8, 12, 14, 17, 30, 33, 37, 38, 49, 51,$$
$$52, 54, 57, 60, 68, 71, 76, 89, 90, 94, 97, 101,$$
$$103, 106, 108, 109, 111, 114, 127, 128, 139,$$
$$141, 144, 165, 177, 179, 182\}$$

We can infer $u_{01} = 182 - 179 = 3$ from W. Construct $W_1 = W + u_{01}$ and calculate $W \cap W_1$.

$$W \cap W_1 = \{3, 5, 8, 17, 33, 52, 54, 57, 60, 71,$$
$$97, 106, 109, 111, 114, 144, 182\}$$

Construct $G(\{0\} \cup (W \cap W_1))$ to see that

$$\{182\} \leftrightarrow \{5, 17\}$$

from which we can infer $u_{02} = 5$ and $u_{03} = 17$. Construct $W_2 = W + u_{02}$ and $W_3 = W + u_{03}$ and calculate $\left(\bigcap_{p=0}^{3} W_p\right)$

$$\left(\bigcap_{p=0}^{3} W_p\right) = \{54, 106, 111, 114, 144, 182\}$$

Calculate $U = \{u_{0p} : 0 \leq p \leq 3\} \bigcup \left(\bigcap_{p=0}^{3} W_p\right)$

$$U = \{0, 3, 5, 17, 54, 106, 111, 114, 144, 182\}$$

which is the required integer set.

7. REFERENCES

[1] A. L. Patterson: A direct method for the determination of the components of interatomic distances in crystals, Zeitschr. Krist. 90 (1935) 517-542

[2] A.L. Patterson: Ambiguities in the X-ray analysis of crystal structures, Phys. Review 65 (1944) 195-201.

[3] M. Stefik, "Inferring DNA structures from segmentation data", Artificial Intelligence 11 (1978).

[4] M.I. Shamos, "Problems in computational geometry", CMU, Pittsburgh, PA 1977.

[5] A. Walther, "The question of phase retrieval in optics," Opt. Acta 10, 4149 (1963).

[6] R. P. Millane, "Phase retrieval in crystallography and optics," J. Opt. Soc. Am. A (1990)

[7] J.C. Dainty and J.R. Fienup,"Phase Retrieval and Image Reconstruction for Astronomy," Chapter 7 in H. Stark, ed., Image Recovery: Theory and Application pp. 231-275.

[8] L. Rabiner and B.H. Juang, "Fundamentals of Speech Recognition," Signal Processing Series, Prentice Hall, 1993.

[9] J.R. Fienup, T.R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation", JOSA, Vol. 72, Issue 5, pp. 610-624 (1982).

[10] K. Jaganathan, S. Oymak and B. Hassibi, "Recovery of Sparse 1-D Signals from the Magnitudes of their Fourier Transform". ISIT 2012.

[11] K. Jaganathan, S.Oymak and B. Hassibi, "On Robust Phase Retrieval for Sparse Signals", Allerton Conference on Communication, Control and Computing, 2012.

[12] R. W. Gerchberg and W. O. Saxton. "A practical algorithm for the determination of the phase from image and diffraction plane pictures". Optik 35, 237 (1972).

[13] Fienup J.R, "Reconstruction of an object from the modulus of its Fourier transform", Optics letters (1978)

[14] J. R. Fienup, "Phase retrieval algorithms: a comparison". Appl. Opt. 21 (1982).

[15] H.H. Bauschke, P.L. Combettes and D.R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization" J. Opt. Soc. Am.A (2002).

[16] S.S. Skiena, W.D. Smith and P. Lemke, "Reconstructing Sets from Interpoint Distances (Extended Abstract)", SCG' 90 Proceedings of the sixth annual symposium on computational geometry, Pages 332-339.

[17] T. Dakic, "On the Turnpike Problem", PhD Thesis, Simon Fraser University, 2000.

[18] Y.M. Lu and M. Vetterli. "Sparse spectral factorization: Unicity and reconstruction algorithms". Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on , vol., no., pp. 5976–5979, 22-27 May 2011

[19] Y. Shechtman, Y.C. Eldar, A. Szameit and M. Segev, "Sparsity Based Sub-Wavelength Imaging with Partially Incoherent Light Via Quadratic Compressed Sensing", Optics Express, vol. 19, Issue 16, pp. 14807-14822, Aug. 2011.

[20] R.M. Karp and L.A. Newberg, " An Algorithm for Analyzing Probed Partial Digestion Experiments", Comput Appl Biosci (1995) 11(3): 229-235 doi:10.1093/bioinformatics/11.3.229.

[21] G. Pandurangan and H.Ramesh, "The Restriction Mapping Problem Revisited", Journal of Computer and System Sciences, 2002.

[22] K.Jaganathan and B. Hassibi, "Reconstruction of Integers from Pairwise Distances", arXiv:1212.2386 [cs.DM].