Sample Complexity of the Boolean Multireference Alignment Problem

Emmanuel Abbe
Princeton University
eabbe@princeton.edu

João Pereira
Princeton University
jpereira@princeton.edu

Amit Singer
Princeton University
amits@math.princeton.edu

***Abstract*—The Boolean multireference alignment problem consists in recovering a Boolean signal from multiple shifted and noisy observations. In this paper we obtain an expression for the error exponent of the maximum A posteriori decoder. This expression is used to characterize the number of measurements needed for signal recovery in the low SNR regime, in terms of higher order autocorrelations of the signal. The characterization is explicit for various signal dimensions, such as prime and even dimensions.**

I. Introduction

The Boolean multireference alignment (BMA) problem consists of estimating an unknown signal $x \in \mathbb{Z}_2^L$, from noisy cyclically shifted copies $Y_1, \dots, Y_N \in \mathbb{Z}_2^L$, i.e.,

$$Y_i = R^{S_i} x \oplus Z_i,\ i \in \{1, \dots, N\}, \tag{1}$$

where the error $Z_i \sim \mathrm{Ber}(p)^L$, the product measure of L Bernoulli variables with parameter p, $\oplus$ denotes addition mod 2, R is the index cyclic shift operator that shifts a vector one element to the right $(x_1, \dots, x_N) \mapsto (x_N, x_1, \dots, x_{N-1})$, R^{S_i} corresponds to applying S_i times the operator R and the shifts $S_i \sim \mathcal{U}(\mathbb{Z}_L)$, the uniform distribution in $\mathbb{Z}_L$.

The motivation to study this problem comes from the classical multireference alignment problem, where the signal and observations are real valued vectors, and the error is Gaussian white noise. Several algorithms were recently proposed to solve the problem, including angular synchronization [1], semidefinite program relaxations of the maximum likelihood decoder [2] and reconstruction using the bispectrum [3]. This problem is also an instance of a larger class of problems, called Non-Unique Games, which also includes the orientation estimation problem in cryo-electron microscopy [4].

Despite these advancements in algorithmic development, not much progress has been made in understanding the fundamental limits of signal recovery. The recent paper [5] investigated fundamental limits of shift recovery in multireference alignment, but not those of signal recovery. We note that estimating the shifts is impossible at low signal-to-noise ratio (SNR) even if an oracle presents us with the true signal. Also, the goal of many applications is signal recovery rather than shift estimation. Our paper aims to fill the gap on signal recovery, by studying the Boolean case. We show here that signal recovery is possible at arbitrarily low SNR, if sufficiently many measurements are available, and quantify this tradeoff.

In BMA the search space is finite, and the maximum A posteriori decoder (MAP) minimizes the probability of error. Our main contribution is an expression for the error exponent of MAP, in the low SNR regime, given in Theorems III.2 and III.3. Our results imply how many measurements are needed, as a function of the SNR, in order to accurately estimate the signal.

The expression depends on the autocorrelations of the signal, defined in (6). Our results connect the order of autocorrelations needed to reconstruct the signal to the number of measurements needed to estimate the signal. This has some connections with previous theoretical work on uniqueness of the bispectrum [6].

We also consider some generalizations of the original problem in order to model some aspects of multireference alignment that arise in applications, such as the introduction of deletions.

II. BMA Problem

In the BMA problem, the errors are i.i.d. Bernoulli of parameter p. If $p = \frac{1}{2}$, then the observations $Y_i \sim \mathrm{Ber}(\frac{1}{2})^L$, regardless of the original signal, and signal recovery is impossible. This corresponds to the case when SNR $= 0$. On the other hand, $p = 0$ or 1 corresponds to the noiseless case. Thus we define

$$\mathrm{SNR} := \left(p - \frac{1}{2}\right)^2. \tag{2}$$

In contrast to proposing an algorithm to solve the BMA problem, our paper focuses on its sample complexity, in the regime when $p \to \frac{1}{2}$ and SNR $\to 0$.

Note that the observations Y_i, $i \in [N]$, given the signal x, are i.i.d., since both the shifts S_i and the errors Z_i are i.i.d. For that reason we will drop the index i when it is more convenient. We rewrite (1), denoting by $x(j)$ the j-th entry of x.

$$Y(j) = x(S + j) \oplus Z(j),\ j \in \mathbb{Z}_L, \tag{3}$$

where '$+$' is addition mod L.

Our paper also considers the sample complexity of the following variations of the basic BMA problem:

- *BMA Problem with consecutive deletions:* In this case the measurements $Y_1, \dots, Y_N$ are in $\mathbb{Z}_2^K$, with $K \leq L$, and

$$Y(j) = x(S + j) \oplus Z(j),\ j \in \mathbb{Z}_K. \tag{4}$$

When $K = L$ we obtain the original BMA problem.

- *BMA Problem with known deletions:* Let $V \subset \mathbb{Z}_L$ be an ordered set of non-deletions, i.e. the set of deletions is $\mathbb{Z}_L \backslash V$. Now the measurements $Y_1, \dots, Y_N$ are in $\mathbb{Z}_2^K$, with $K = |V|$, and:
$$Y(j) = x(S + V_j) \oplus Z(j), \forall j \in \mathbb{Z}_K, \tag{5}$$
where V_j denotes the j-th element of V. When $V = [K]$ we recover the BMA problem with consecutive deletions.
- *BMA Problem (and variations) with non uniform rotations:* Similar to the previous problems, but now the shifts follow some distribution ξ in $\mathbb{Z}_L$.

These variations are motivated by problems similar to multireference alignment. The case of possible deletions is intended to model instances where the observations are only partial, whereas the extension to non-uniform shifts attempts to represent a non-symmetric version of the problem.

III. Results

We start by introducing the following notion of autocorrelation of a signal that is central to our main results.

Definition III.1. *The $(\xi, \mathbf{k})$-autocorrelation of x, with respect to a distribution ξ in $\mathbb{Z}_L$ and $\mathbf{k} = (k_1, k_2, \dots, k_d) \in \mathbb{Z}_L^d$ is defined as*
$$A_{\xi,\mathbf{k}}(x) := \sum_{s=1}^{L} \xi(s) x(k_1 + s) \cdots x(k_d + s). \tag{6}$$
We refer to $d = |k|$ as the order of the auto-correlation. When $\xi \sim \mathcal{U}(\mathbb{Z}_L)$, we simply write $\mathbf{k}$-autocorrelation and $A_\mathbf{k}$. Notice $A_\mathbf{k}$ is shift invariant, that is $A_\mathbf{k}(x) = A_\mathbf{k}(R^s x)$, and in this case we may assume $k_1 = 0$.

We define the minimum autocorrelation order necessary to distinguish x_1 and x_2 under ξ and V as
$$t_{\xi,V}(x_1, x_2) := \inf\{d : A_{\xi,\mathbf{k}}(x_1) \neq A_{\xi,\mathbf{k}}(x_2), \mathbf{k} \in V^d\}, \tag{7}$$
where V^d denotes the vectors in Z_2^d with entries in V. The minimum autocorrelation order necessary to describe all signals in $\mathcal{X}$ is defined as
$$t_{\xi,V}(\mathcal{X}) := \max_{\substack{x_1,x_2 \in \mathcal{X} \\ x_1 \neq x_2}} t_{\xi,V}(x_1, x_2). \tag{8}$$

Given a prior distribution on the signals P_X, with support $\mathcal{X}$, denote by X the random variable with distribution P_X. Given an algorithm for BMA the probability of error is defined as
$$P(\hat{X} \neq X) = \sum_{x_i \in \mathcal{X}} P(\hat{X} \neq x_i) P_X(x_i), \tag{9}$$
where $\hat{X}$ is the answer given by the algorithm. In the BMA problem the search space is finite, thus MAP minimizes the probability of error (9). We obtain results that do not depend on the prior distribution, they depend only on its support.

Theorem III.2. *Consider the BMA problem with known deletions $Z_L \backslash V$ and shift distribution ξ. Let $\mathcal{X} \subset Z_2^L$ be the support of the prior distribution of the signals and μ_x the conditional distribution in $\mathbb{Z}_2^K$ of the observations Y given the signal x, where $K = |V|$. The probability of error of the MAP estimator, denoted by P_e, has the following asymptotic behavior*
$$\lim_{N\to\infty} \frac{1}{N} \log P_e = \min_{\substack{x_1,x_2 \in \mathcal{X} \\ x_1 \neq x_2}} C(\mu_{x_1}, \mu_{x_2}), \tag{10}$$
with
$$C(\mu_{x_1}, \mu_{x_2}) = \frac{2^{4t-3}}{t!} \mathrm{SNR}^t \sum_{\mathbf{k} \in V^t} \Big(A_{\xi,\mathbf{k}}(x_1) - A_{\xi,\mathbf{k}}(x_2)\Big)^2 + O(\mathrm{SNR}^{t+1}), \tag{11}$$
and $t = t_{\xi,V}(x_1, x_2)$.

The theorem implies that the exponent on SNR is $t_{\xi,V}(\mathcal{X})$. In the original problem, with uniform shifts and no deletions, the recovery of the original signal is possible only up to a shift, i.e. we can only recover $R^k x$, where x is the original signal, and k is some shift in $\mathbb{Z}_L$. For that reason, we consider $\mathcal{X}$ to have exactly one element of each class of all the shifts of a signal, i.e., there are no two elements in $\mathcal{X}$ where one is a shift of the other (for example, if L is prime, then there are $2^L - 2$ such elements).

Corollary III.3. *Consider the original problem, with $V = [L]$, $\xi \sim \mathcal{U}(\mathbb{Z}_L)$ and $\mathcal{X}$ as defined above. By inspection one can obtain the error exponent for $L \leq 5$. For $L \geq 6$, we either have*
$$\lim_{N\to\infty} \frac{1}{N} \log P_e = \begin{cases} \frac{2^{10}}{L} \mathrm{SNR}^3 + O(\mathrm{SNR}^4) \\ O(\mathrm{SNR}^4) \end{cases} \tag{12}$$
Also, the first case occurs when L is prime, and the second when $L \geq 12$ and is even. The other values of L remain open.

IV. Proof Techniques

Proof of Theorem III.2: The proof consists of two main parts. The next theorem gives a formula to the error exponent and claim IV.2 makes the connection with autocorrelations.

Theorem IV.1. *Consider the BMA problem with known deletions $Z_L \backslash V$ and shift distribution ξ. Let $\mathcal{X} \subset Z_2^L$ be the space of possible signals and $\mu_x := P_{Y|X}(\cdot|x)$ the conditional distribution in $\mathbb{Z}_2^K$ of the observations given the signal x. The probability of error of the MAP estimator (P_e) has the following asymptotic behavior*
$$\lim_{N\to\infty} \frac{1}{N} \log P_e = \min_{x_1 \neq x_2 \in \mathcal{X}} C(\mu_{x_1}, \mu_{x_2}), \tag{13}$$
with
$$C(\mu_{x_1}, \mu_{x_2}) = \frac{\left(\frac{1}{2} - p\right)^{2s}}{8(s!)^2} \sum_{y \in \mathbb{Z}_2^K} \frac{\left(\mu_{x_1}^{(s)}\left(y; \frac{1}{2}\right) - \mu_{x_2}^{(s)}\left(y; \frac{1}{2}\right)\right)^2}{\mu_{x_1}\left(y; \frac{1}{2}\right)} + O\left(\frac{1}{2} - p\right)^{2s+2}, \tag{14}$$

where $\mu_x^{(m)}(y;p)$ denotes the m-th derivative of $\mu_x(y;p)$ in p, i.e. the derivative of the conditional distribution in y given x in order of the Bernoulli parameter p, and

$$s(x_1,x_2) := \inf\left\{ m : \mu_{x_1}^{(m)}\left(y;\frac{1}{2}\right) \neq \mu_{x_2}^{(m)}\left(y;\frac{1}{2}\right), y \in \mathbb{Z}_2^K \right\}.$$

This theorem follows from Theorems 1 and 2 in [7]. Theorem 1 is a corollary of Sanov Theorem [8], which leads to (13). However the expression obtained by Theorem 1 is rather complex and not very interpretable. In Theorem 2 [7] we Taylor expand (13) and obtain a useful characterization in instances where the SNR is small. We use this expression to obtain (14).

Claim IV.2. *If $\mu_{x_1}^{(m)}\left(y;\frac{1}{2}\right) = \mu_{x_2}^{(m)}\left(y;\frac{1}{2}\right)$ for all $m < n$ and $y \in \mathbb{Z}_2^K$, then the following expressions are equal:*

$$\sum_{y\in\mathbb{Z}_2^K} \frac{\left(\mu_{x_1}^{(n)}\left(y;\frac{1}{2}\right) - \mu_{x_2}^{(n)}\left(y;\frac{1}{2}\right)\right)^2}{\mu_{x_1}\left(y;\frac{1}{2}\right)} \tag{15}$$

and

$$2^{4n} n! \sum_{\mathbf{k}\in V^L} \left(A_{\xi,\mathbf{k}}(x_1) - A_{\xi,\mathbf{k}}(x_2)\right)^2. \tag{16}$$

In fact, since the expressions (15) and (16) are both sum of squares, the claim implies that $t_{\xi,V}(x_1,x_2) = s(x_1,x_2)$, what concludes the proof of theorem III.2. ∎

Proof of Claim IV.2: Denote by $x(V)$ the vector in $\mathbb{Z}_2^K$ ($K = |V|$) that consists of the values of x with indices in V, i.e. the j-th element of $x(V)$ is $x(V_j)$. Also, given $s \in \mathbb{Z}_L$ denote by $s + V$ the ordered set corresponding to the sum of each element in V with s mod L. Equation (5) can then be rewritten, as

$$Y = x(S+V) \oplus Z$$

Then since $Z \sim \mathrm{Ber}(p)^L$, we have

$$\mu_x(y;p|S=s) = (1-p)^{K-w(y\oplus x(s+V))} p^{w(y\oplus x(s+V))},$$

where w denotes the Hamming weight, and since $S \sim \xi$

$$\mu_x(y;p) = \sum_{s=1}^{L} \xi(s)(1-p)^{K-w(y\oplus x(s+V))} p^{w(y\oplus x(s+V))}. \tag{17}$$

In the statement of the theorem we have $x \in \mathbb{Z}_2^L$, however it is convenient for the proof to consider the entries of x to be $-1, 1$, changed by the rule: $a \mapsto 1 - 2a$. We will call

$$u := 1 - 2x \in \Sigma_2^L \tag{18}$$

the corresponding element of x with ± 1 values, where $\Sigma_2 := \{-1, 1\}$, and $v := 1 - 2y$. In analogy to the Hamming weight, we define

$$W(u) := \sum_{s=1}^{L} u(s) = L - 2w(x). \tag{19}$$

With this we rewrite (17)

$$\mu_u(v;p) = \sum_{s=1}^{L} \xi(s)(1-p)^{\frac{K}{2}+\frac{W(v\oplus u(s+V))}{2}} p^{\frac{K}{2}-\frac{W(v\oplus u(s+V))}{2}}, \tag{20}$$

where $\mu_u(v;p) := \mu_x(y;p)$. For simplicity of notation denote

$$W_{v,u,s} := W(v \oplus u(s+V)).$$

The claim is now proved by induction on n. By properties of Jacobi polynomials [9] we have

$$\left(p^{\frac{K}{2}-\frac{b}{2}}(1-p)^{\frac{K}{2}+\frac{b}{2}}\right)^{(m)}_{|p=\frac{1}{2}} = (-2)^{m-K} P_m(b),$$

where P_m is a polynomial with the following property

$$P_m(b) = b^m + Q_m(b), \tag{21}$$

where Q_m has degree at most $m-1$, and $Q_0 \equiv Q_1 \equiv 0$. Thus

$$\mu_u^{(m)}\left(v;\frac{1}{2}\right) = (-2)^{m-K} \sum_{s=1}^{L} \xi(s) P_m(W_{v,u,s}). \tag{22}$$

Then when $m = 1$

$$\sum_{v\in\Sigma_2^K} \frac{\left(\mu_{u_1}^{(1)}\left(v;\frac{1}{2}\right) - \mu_{u_2}^{(1)}\left(v;\frac{1}{2}\right)\right)^2}{\mu_{u_1}\left(v;\frac{1}{2}\right)}$$
$$= 2^{2-K} \sum_{v\in\Sigma_2^K} \left[\sum_{s=1}^{L} \xi(s)\left(W_{v,u_1,s} - W_{v,u_2,s}\right)\right]^2.$$

Now, by the induction hypothesis if $\mu_{u_1}^{(k)}\left(v;\frac{1}{2}\right) = \mu_{u_2}^{(k)}\left(v;\frac{1}{2}\right)$ for all $k \le n-1$, $v \in \Sigma_2^K$

$$\sum_{s=1}^{L} \xi(s) Q_n(W_{v,u_1,s}) = \sum_{s=1}^{L} \xi(s) Q_n(W_{v,u_2,s}),$$

for all $v \in \Sigma_2^K$ since Q_n has degree at most $n-1$. Thus by (21) and (22)

$$\sum_{v\in\Sigma_2^K} \frac{\left(\mu_{u_1}^{(n)}\left(v;\frac{1}{2}\right) - \mu_{u_2}^{(n)}\left(v;\frac{1}{2}\right)\right)^2}{\mu_{u_1}\left(v;\frac{1}{2}\right)} =$$
$$2^{2n-K} \sum_{v\in\Sigma_2^K} \left[\sum_{s=1}^{L} \xi(s)\left(W_{v,u_1,s}^n - W_{v,u_2,s}^n\right)\right]^2 \tag{23}$$

Now splitting the square of the sum on the RHS into a product of two sums and expanding, we obtain terms of the form

$$\sum_{s_1=1}^{L}\sum_{s_2=1}^{L} \xi(s_1)\xi(s_2)(-1)^{\alpha+\beta} \sum_{v\in\Sigma_2^K} W_{v,u_\alpha,s_1}^n W_{v,u_\beta,s_2}^n, \tag{24}$$

where α and β are 1 or 2. By Lemma IV.3 we get

$$\sum_{v\in\Sigma_2^K} W_{v,u_\alpha,s_1}^n W_{v,u_\beta,s_2}^n =$$
$$2^K \sum_{\substack{A\in M_{[2n]} \\ A \text{ is even}}} C_A \prod_{i=1}^{|A|}\left(\sum_{k=1}^{K}\prod_{j=1}^{|a_i|} u_{a_{ij}}(k)\right), \tag{25}$$

Where $u_{a_{ij}}$ is $u_\alpha(s_1+V)$ if $a_{ij} \le n$, and is $u_\beta(s_2+V)$ otherwise. So, since $|a_i|$ is even, as A is an even partition, and the entries of $u_{a_{ij}}$ are ± 1,

$$\sum_{k=1}^{K} \prod_{j=1}^{|a_i|} u_{a_{ij}}(k) = \sum_{k \in V} u_\alpha(s_1+k) u_\beta(s_2+k)$$

if $|a_i \cap [n]|$ is odd, and it is K otherwise. Then

$$\sum_{v \in \Sigma_2^K} W^n_{v,u_\alpha,s_1} W^n_{v,u_\beta,s_2} = R_n\left(\sum_{k \in V} u_\alpha(s_1+k) u_\beta(s_2+k)\right),$$

where R_n is a polynomial with degree n (with coefficients possibly depending on K and n), and $R_1(b) = 2^k b$. It cannot have degree $n+1$ since $|A| \le n$, since it is an even partition of $[2n]$. For it to be a power of order n, we need $|A| = n$, so $|a_i| = 2$ for $i = 1, \dots, n$, thus $C_A = 1$, by the Lemma. Also $|a_i \cap [n]|$ must be odd for all i, thus $|a_i \cap [n]| = 1$. There are exactly $n!$ partitions with this property, so the leading coefficient of R_n is $2^K n!$. We also have

$$\begin{aligned}
&\sum_{s_1=1}^{L} \sum_{s_2=1}^{L} \xi(s_1)\xi(s_2) \left(\sum_{k \in V} u_\alpha(s_1+k) u_\beta(s_2+k)\right)^n \\
&= \sum_{s_1=1}^{L} \sum_{s_2=1}^{L} \xi(s_1)\xi(s_2) \sum_{\mathbf{k} \in V^n} \prod_{i=1}^{n} u_\alpha(s_1+k_i) u_\beta(s_2+k_i) \\
&= \sum_{\mathbf{k} \in V^n} A_{\xi,\mathbf{k}}(u_\alpha) A_{\xi,\mathbf{k}}(u_\beta), \qquad (26)
\end{aligned}$$

Mimicing the argument used in (23), the equation will be true for $n = 1$, since $R_1(b) = 2^k b$, and by the induction hypothesis only the leading coefficient of R_n is of interest, since the other terms will cancel with each other.

$$\sum_{v \in \Sigma_2^K} \left[\sum_{s=1}^{L} \xi(s) \left(W^n_{v,u_1,s} - W^n_{v,u_2,s}\right)\right]^2 = 2^k n! \sum_{\mathbf{k} \in V^n} \left(A_{\xi,\mathbf{k}}(u_1) - A_{\xi,\mathbf{k}}(u_2)\right)^2 \qquad (27)$$

Now through some algebraic manipulation, and using again the argument of the leading coefficient, if $|\mathbf{k}| = n$, then

$$\sum_{\mathbf{k} \in V^n} \left(A_{\xi,\mathbf{k}}(u_1) - A_{\xi,\mathbf{k}}(u_2)\right)^2 = 2^{2n} \sum_{\mathbf{k} \in V^n} \left(A_{\xi,\mathbf{k}}(x_1) - A_{\xi,\mathbf{k}}(x_2)\right)^2 \qquad (28)$$

This together with (23) and (27) concludes the proof. ■

Lemma IV.3. *For any partition $A = \{a_1, \dots, a_{|A|}\}$ of the set $\{1, 2, \dots, m\}$, denote by a_{ij} the j-th entry of a_i and $M_{[m]}$ the set of all such partitions. If $u_1, \dots, u_m \in \Sigma_2^K$*

$$\sum_{v \in \Sigma_2^K} W(u_1 \oplus v) \cdots W(u_m \oplus v) = 2^K \sum_{\substack{A \in M_{[m]} \\ A \text{ is even}}} C_A \prod_{i=1}^{|A|} \left(\sum_{k=1}^{K} \prod_{j=1}^{|a_i|} u_{a_{ij}}(k)\right), \qquad (29)$$

where A is even if all $|a_i|$ are even for $i \in \{1, \dots, |A|\}$. Moreover, C_A is a constant that depends only on the partition A and is always 1 if $|a_i| = 2$ for all $i \in \{1, \dots, |A|\}$.

Proof: Recall (19). We have $W(u \oplus v) = \sum_{k=1}^{K} u(k)v(k)$ and

$$\begin{aligned}
&\sum_{v \in \Sigma_2^K} W(u_1 \oplus v) \cdots W(u_m \oplus v) \\
&= \sum_{k_1=1}^{K} \cdots \sum_{k_m=1}^{K} u_1(k_1) \cdots u_m(k_m) \sum_{v \in \Sigma_2^K} v(k_1) \cdots v(k_m)
\end{aligned} \qquad (30)$$

Suppose $k_1 \ne k_i$ for $i = 2, \dots, m$. Then

$$\begin{aligned}
&\sum_{v \in \Sigma_2^K} v(k_1)v(k_2) \cdots v(k_m) \\
&= \sum_{\substack{v \in \Sigma_2^K \\ v(k_1)=1}} v(k_1)v(k_2) \cdots v(k_m) + \sum_{\substack{v \in \Sigma_2^K \\ v(k_1)=-1}} v(k_1)v(k_2) \cdots v(k_m)
\end{aligned}$$

which is 0. This also occurs if $k_1 = k_2 = \cdots = k_j \ne k_i$, $i > j$ and j is odd. For this not to occur, the classes of k_i's that are equal to each other are required to have all an even number of elements, and in that case, the sum is 2^K. By grouping the k_i's, (30) becomes

$$2^K \sum_{\substack{A \in M_{[m]} \\ A \text{ is even}}} \sum_{\substack{k_1, \dots, k_{|A|}=1 \\ \text{all distinct}}}^{K} \prod_{i=1}^{|A|} \prod_{j=1}^{|a_i|} u_{a_{ij}}(k_i), \qquad (31)$$

Using a combinatorial argument we can rewrite the (31) without the 'all-distinct' condition, at the cost of a constant C_A, which is 1 when $|a_i| = 2$ for $i \in \{1, \dots, |A|\}$.

$$\begin{aligned}
&2^K \sum_{\substack{A \in M_{[m]} \\ A \text{ is even}}} C_A \sum_{k_1, \dots, k_{|A|}=1}^{K} \prod_{i=1}^{|A|} \prod_{j=1}^{|a_i|} u_{a_{ij}}(k_i) = \\
&= 2^K \sum_{\substack{A \in M_{[m]} \\ A \text{ is even}}} C_A \prod_{i=1}^{|A|} \left(\sum_{k=1}^{K} \prod_{j=1}^{|a_i|} u_{a_{ij}}(k)\right)
\end{aligned}$$

■

Proof of Corollary III.3: We first prove equation (12). Recall (6), and denote by

$$B_m(x_1, x_2) := \sum_{\mathbf{k} \in \mathbb{Z}_L^m} \left(A_{\mathbf{k}}(x_1) - A_{\mathbf{k}}(x_2)\right)^2$$

and

$$B_m(L) := \min_{x_1 \neq x_2 \in \mathcal{X}} B_m(x_1, x_2)$$

Note that $B_m(x_1, x_2) = 0$ if $m < t_{\xi,V}(x_1, x_2)$ by (7). For convenience let $B(x_1, x_2) := B_{t_{\xi,V}(x_1,x_2)}(x_1, x_2)$ and $B(L) := B_{t_{\xi,V}(\mathcal{X})}(L)$. Using this notation we rewrite (10) and (11)

$$\lim_{N\to\infty} \frac{1}{N} \log P_e = B(L) \frac{2^{4t_L - 3}}{t_L!} \text{SNR}^{t_L} + O\left(\text{SNR}^{t_L+1}\right)$$

Now equation (12) is equivalent to having $t_{\xi,V}(\mathcal{X}) \geq 3$ and $B_3(L)$ either $\frac{12}{L}$ or 0. Turns out, for $L \geq 6$, if we take

$$x_1^* = (1, 1, 0, 1, \underbrace{0, \ldots, 0}_{L-4 \text{ zeros}}) \text{ and } x_2^* = (1, 0, 1, 1, \underbrace{0, \ldots, 0}_{L-4 \text{ zeros}}),$$

then $t_{\xi,V}(\mathcal{X}) \geq t_{\xi,V}(x_1^*, x_2^*) = 3$ and $B_3(L) \leq B(x_1^*, x_2^*) = \frac{12}{L}$. Also we cannot have $\frac{12}{L} > B_3(L) > 0$. This implies there exists x_1 and x_2 in $\mathcal{X}$ such that $\frac{12}{L} > B(x_1, x_2) > 0$. Since it is positive, there is $\mathbf{k}^* \in \mathbb{Z}_L^3$ such that $A_{\mathbf{k}^*}(x_1) \neq A_{\mathbf{k}^*}(x_2)$. But by definition (6), since $\xi(s) = \frac{1}{L}$, $LA_{\mathbf{k}^*}(x)$ is an integer for $x \in \mathbb{Z}_2^L$, and $L^2(A_{\mathbf{k}^*}(x_1) - A_{\mathbf{k}^*}(x_2))^2 \in \mathbb{Z}$.

Now by the definition we also have $A_{\sigma(\mathbf{k}^*)}(x) = A_{\mathbf{k}^*}(x)$, where σ permutes the entries of $\mathbf{k}^*$. Also, for $s \in \mathbb{Z}_L$, let $s + \mathbf{k}^* := (s + k_1^*, s + k_2^*, s + k_3^*)$, then $A_{s+\mathbf{k}^*}(x) = A_{\mathbf{k}^*}(x)$. There is 6 permutations and L possible values for $s \in \mathbb{Z}_L$, so $B(x_1, x_2)$ is an integer multiple of $\frac{6}{L}$. (we can also have not trivial s and σ such that $s + \mathbf{k}^* = \sigma(\mathbf{k}^*)$ but that case also has the property mentioned). However we cannot have $B(x_1, x_2) = \frac{6}{L}$. That means there exists only one $\mathbf{k}^* \in \mathbb{Z}_L^3$ (with permutations and shifts) such that $A_{\mathbf{k}^*}(x_1) \neq A_{\mathbf{k}^*}(x_2)$. Then

$$\sum_{\mathbf{k}\in\mathbb{Z}_L^3} A_{\mathbf{k}}(x_1) - A_{\mathbf{k}}(x_2) = 6L(A_{\mathbf{k}^*}(x_1) - A_{\mathbf{k}^*}(x_2)) \neq 0 \quad (32)$$

On the other hand

$$\sum_{\mathbf{k}\in\mathbb{Z}_L^3} A_{\mathbf{k}}(x_1) = \frac{1}{L} \sum_{s=1}^{L} \sum_{\mathbf{k}\in\mathbb{Z}_L^3} x(k_1 + s)x(k_2 + s)x(k_3 + s)$$
$$= L^3 A_0(x_1)^3,$$

where A_0 denotes $\mathbf{k}$-autocorrelation with $\mathbf{k} = 0$. Since $t_L > 1$, $A_0(x_1) = A_0(x_2)$, so equation (32) must be 0, and equation (12) follows by contradiction. Now if $L \geq 12$ is even, choose

$$x_1^* = (1, 1, 0, \underbrace{1, \ldots, 1}_{\frac{L}{2}-3 \text{ ones}}, 0, 0, 1, \underbrace{0, \ldots, 0}_{\frac{L}{2}-3 \text{ zeros}})$$

and x_2^* the vector obtained by reversing the entries of x_1^*. Since one is the reverse of the other, they have same 1 and 2 order autocorrelations. Recall (18) and (6) and notice that in this case both $A_{\mathbf{k}}(u_1)$ and $A_{\mathbf{k}}(u_2)$ are 0 when $|\mathbf{k}|$ is odd, since half of the signal is the symmetric of the other half, i.e. $u_1(\{1, \ldots, \frac{L}{2}\}) = -u_1(\{\frac{L}{2} + 1, \ldots, L\})$. Now because of (28) we have $A_{\mathbf{k}}(x_1) = A_{\mathbf{k}}(x_2)$ when $|\mathbf{k}| = 3$, so $t_L \geq 4$, and $B_3(L) = 0$.

Finally, let $L \geq 6$ be prime. We prove by contradiction that $t_L = 3$ and $B_3(L) = \frac{12}{L}$. If this is not true, then it exists x_1^* and x_2^* such that $t_{x_1^*,x_2^*} > 3$, so

$$A_{\mathbf{k}}(x_1^*) = A_{\mathbf{k}}(x_2^*), \quad \mathbf{k} \in \mathbb{Z}_L^n, n \leq 3 \qquad (33)$$

By Theorem 2 of paper [6], if the Fourier coefficients of x_1^* and x_2^* are non-zero, then equation (33) implies one is a shift of the other. Denote by $\{r_j^1\}_{j\in\mathbb{Z}_L}$ and $\{r_j^2\}_{j\in\mathbb{Z}_L}$ the Fourier coefficients of x_1^* and x_2^*, respectively, which are given by

$$r_j^\alpha = \frac{1}{\sqrt{L}} \sum_{s=1}^{L} x_\alpha(s) \omega_L^{-js}, \quad \alpha \in \{1, 2\}, j \in \mathbb{Z}_L, \qquad (34)$$
$$= \frac{1}{\sqrt{L}} \sum_{s:x_\alpha(s)=1} \omega_L^{-js}, \qquad (35)$$

where ω_L is the L'th root of unity. $r_0^\alpha = 0$ implies x_α^* only has zeros, and r_j^α is 0 only if w_L^{-j} is a root of the polynomial

$$\sum_{s:x_\alpha(s)=1} b^s \qquad (36)$$

However, since L is prime, the minimal polynomial of w_L^{-j} in $\mathbb{Q}[x]$, for $L > j > 0$, is $1 + x + \cdots + x^{L-1}$ [10], so this polynomial must divide (36). Thus x_1^* and x_2^* must be the all zeros and all ones signals, but this signals also do not satisfy (33). ∎

Acknowledgment

A. S. and J. P. were partially supported by Award Number R01GM090200 from the NIGMS, FA9550-12-1-0317 from AFOSR, Simons Foundation Investigator Award and Simons Collaborations on Algorithms and Geometry, and the Moore Foundation Data-Driven Discovery Investigator Award.

E. A. is partially supported by NSF CAREER Award CCF-1552131 and ARO grant W911NF-16-1-0051.

References

[1] A. Singer, “Angular synchronization by eigenvectors and semidefinite programming,” *Applied and computational harmonic analysis*, vol. 30, no. 1, pp. 20–36, 2011.

[2] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu, “Multireference alignment using semidefinite programming,” in *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM, 2014, pp. 459–470.

[3] B. M. Sadler and G. B. Giannakis, “Shift- and rotation-invariant object reconstruction using the bispectrum,” *JOSA A*, vol. 9, no. 1, pp. 57–69, 1992.

[4] A. S. Bandeira, Y. Chen, and A. Singer, “Non-unique games over compact groups and orientation estimation in cryo-EM,” *arXiv preprint arXiv:1505.03840*, 2015.

[5] C. Aguerrebere, M. Delbracio, A. Bartesaghi, and G. Sapiro, “Fundamental limits in multi-image alignment,” *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5707–5722, 2016.

[6] R. Kakarala, “The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: a group-theoretic approach,” *Journal of Mathematical Imaging and Vision*, vol. 44, no. 3, pp. 341–353, 2012.

[7] E. Abbe, J. M. Pereira, and A. Singer, “Very noisy Sanov’s theorem and applications to alignment problems,” *arXiv preprint*, 2017.

[8] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2006.

[9] G. Szegő, *Orthogonal polynomials*. American Mathematical Society, 1975, vol. 23.

[10] N. Jacobson, *Lectures in Abstract Algebra: III. Theory of Fields and Galois Theory*. Springer-Verlag, 1964, vol. 32.