# Lecture 8: Multireference Alignment, Invariant Features, and the Bispectrum

Georgy Noarov

November 5, 2018

## 1 Structure of the Presentation

Today's topic is Multireference Alignment (MRA). It has many flavors, despite being a very simple model. Due to space constraints, thus, we will be unable to cover all the mathematical methods used to obtain the results which we will discuss. We thus decided to focus on broad exposition, while periodically zooming in on some aspects of the theory of MRA so as to encounter some (not all!) of the interesting ideas used in that area. In particular, proofs will be given not-so-often, and if a proof is left out, the reader is welcome to consult the papers referenced in the last Section of these notes for details. There, one can find which sections correspond to which reference sources, for easy navigation.

The outline of the lecture is as follows. We cover in Section 2 the basics of the MRA model, including a simple way to reconstruct the signal called Template Alignment, and we provide motivation for a more advanced approach to signal estimation, known as the Invariant Features method. In Section 3, we introduce the fundamental invariant features of the signal in the MRA framework: its mean, power spectrum and bispectrum. We consider these features' properties, particularly those of the bispectrum. We then determine the outline of the algorithm that estimates the signal based on these invariant features. In Section 4, we define and provide background for two of the algorithms that estimate the Fourier phases of the signal, and thus form the core of the Invariant Features algorithm: Non-Convex Optimization on the Phase Manifold, and Frequency Marching. Finally, in Section 5, devoted to the sample complexity of signal estimation in the MRA model, we have two subsections. The first one deals with the case when the translations of the signal in MRA are assumed uniformly distributed. The other one discusses tight bounds on the sample complexity of MRA for general translation distributions based on their periodicity or aperiodicity.

## 2 The Model

Today we will consider a toy model called Multireference Alignment (or MRA, for short), where we need to recover a signal from observing, multiple times, its shifted versions with added iid Gaussian noise. By "shifted versions of the signal" we mean taking the signal $x$ as a vector and circularly rotating its entries by some fixed amount $r$. That is, the rotated version $R_r x$ of the signal, where $R_r$ is the linear operator of translation by $r$ entries, is such that $(R_r x)[k] = x[k-r]$ for every $k \in \{0, \dots, length(x) - 1\}$.

Formally, the model consists of $M$ observations $\xi_j$, $j \in \{1, \dots, M\}$, of a signal $x$, where $x \in \mathbb{R}^N$ or $x \in \mathbb{C}^N$, such that each observation is given by

$$\xi_j = R_{r_j} x + \varepsilon_j, \text{ with } \varepsilon_j \sim N(0, \sigma^2), \text{ and all } \varepsilon_j \text{ iid.}$$

The amounts by which the signal gets translated, $r_j$, are treated as unknown numbers, or, to equivalently rephrase it, as *latent variables* of this model. We have no direct access to these values. The task is to estimate $x$ from the observations $\xi_j$, $j \in \{1, \ldots, M\}$. At this point notice that due to its definition, the MRA model will only allow us to estimate the signal $x$ *up to a constant shift*.

In the model we consider, the signal is subject to two types of corruption: due to a latent translation, and due to additive Gaussian noise.

- How is this model related to Cryo-EM? We can view it as a toy model of what happens in Cryo-EM, with an easier-to-analyze linear operator acting on the signal which is to be reconstructed.
- In Cryo-EM the signal is the 3D structure of the molecule, and our observations consist of 2D images. If $x$ is the signal which encodes the 3D structure of the molecule and $\xi$ is a 2D image of that structure, then neglecting in-plane translations and other imaging effects, we have that $\xi = Ax + \varepsilon$, where $A$ is the linear transformation that is a composition of a tomographic projection in a certain direction and the group action that rotates the 3D structure of the molecule, and $\varepsilon$ is the additive noise.

## 2.1 A Trivial Approach to Alignment

Our direct goal is to estimate the signal itself, $x$. However, each observation shifts $x$ by an unknown amount $r_j$, so we cannot estimate $x$ directly. To resolve this issue, 2 reasonable approaches are available:

- Either we can first somehow estimate all the latent shifts $r_j$, then shift the respective observations *back* by the estimated amounts $\hat{r}_j$, and average these shifted-back observations, yielding the estimate $\hat{x} = \frac{1}{M}\sum_{j=1}^{M} R_{-\hat{r}_j}\xi_j$ for the signal $x$. The idea of this is that we strive to first align the observations, so that when we are averaging out the noise by considering the sum of all the aligned observations, the shifted versions of the signal contained in the observations are also aligned.
- Or, another approach would be to estimate the signal *without* trying to estimate the latent shifts beforehand. To accomplish this, we would need to extract some characteristics of the signal $x$ from the observations while keeping in mind that, since we do not want to first remove the latent shifts, each observation $\xi_j$ will only allow us to measure the characteristics of a *shifted*, by a unknown amount, version of $x$. Naturally, then, we are led to consider properties of the signal that are *invariant* under all possible shifts of $x$. For such invariant characteristics, then, it doesn't matter what amount $x$ was shifted by in each individual $\xi_j$, and so we can estimate such characteristics *in the same way* from each of the observations and forget about the cyclic shifts associated to all the $\xi_j$'s.

  These considerations will form a basis for the *Invariant Features Approach* to estimating the signal in MRA. Stay tuned until the next section.

From the first bullet point, we have an outline of an approach which estimates the signal after first estimating all latent shifts. A very simple instance of this approach is called *Template Alignment*. Choose one of the observations, say $\xi_1$, as a template. Then, for each other observation $\xi_j$, estimate the latent shift $r_j$ associated with $\xi_j$ by

$$\hat{r}_j = \arg\max_k \mathcal{RE}\left(\sum_{n=0}^{N-1} \xi_1[n]\overline{\xi_j[n+k]}\right).$$

Here, $\mathcal{RE}$ denotes the real part of a complex number.

The idea behind this approach is as follows. Suppose that the noise level is low enough that for any $j > 1$ the major contribution to the correlation between $\xi_1$ and $\xi_j$ comes from the correlation between the shifted versions $R_{r_1}$ and $R_{r_j}$ of the signal $x$ contained in these observations. Given this assumption, we would expect for any $j > 1$ that, since $\xi_1$ contains the signal $x$ shifted by amount $r_1$, and $\xi_j$ - by amount $r_j$, then over all shifted versions $R_k\xi_j$ of this observation $\xi_j$, the highest correlation between $\xi_1$ and $R_k\xi_j$ will be achieved for $k = k_j := r_1 - r_j$. Indeed, both $\xi_1$ and $R_{k_j}\xi_j$ contain the signal $x$ shifted by the same amount $r_1$, which gives the highest possible correlation.

The formula above measures $k_j$, for every $j$, by rotationally aligning $\xi_1$ and $\xi_j$ in all possible $N$ ways and choosing the way that gives the highest correlation: that shift $k \in \{0, \dots, N-1\}$ which achieves the most similarity to the template, as measured by the correlation, is then declared to be the relative latent shift $k_j$ of observation $\xi_j$.

This approach necessitates time complexity $O(MN \log N)$, since the correlations allowing to determine $\hat{r_j}$, for each $j$, can be separately computed in $O(N \log N)$ time, and this is performed for each of the $M$ observations. Thus, the algorithm is *online*.

Template Alignment has one big disadvantage when there is a high level of noise ($\sigma^2$ large), however: Namely, since each observation is aligned with one and the same template observation, the entire estimation procedure becomes very error prone given high noise. One solution to this problem, that would make Alignment more resistant to noise, involves aligning all pairs of observations to each other, but the runtime of that algorithm would be $O(M^2 N \log N)$, by the same token as just discussed, and thus such an approach would be too slow in most realistic scenarios when $M$ is quite large.

## 2.2 Preview: Why Will the Invariant Features Approach Be Better?

As we have seen, the approach outlined above is overly prone to noise; however, it also has two other properties that we can hope to improve on using our Invariant Features approach:

- The time complexity achieved by the Template Alignment approach is $O(MN \log N)$. For $N$ fixed, this depends linearly on the number of observations $M$. However, the coefficient of proportionality of $M$ grows asymptotically as $N \log N$ with increasing $N$. This is potentially a problem for large $N$, since to achieve decent accuracy of estimation, $M$ has to be quite large. The Invariant Features approach will address this problem by using clever preprocessing. Given the set of $M$ observations, we will, initially, constantly many times perform averaging of some particular functions over the $M$ data points, which is a very cheap operation from the computational point of view. Then, we will forget about the $M$ data points and work directly with the computed averages. To these averages, we will apply an algorithm that estimates the signal based on them (see e.g. those algorithms discussed in Section 4). The complexity of this algorithm will now only depend on $N$ but not on the number of data points. Thus, in terms of complexity, the Invariant Features approach will be, unlike Template Alignment, well-tuned to the fact that we require a large number $M$ of data points.

- Also in terms of space, Template Alignment does not seem optimal: it requires to store $MN$ samples corresponding to all the $M$ observations. And indeed it turns out that it is possible to improve on this by using the Invariant Features approach. There, one does not need to store all observations. For each individual observation, we will just need to compute its *invariants* that will later be averaged across all observations. Having computed these invariants, we can then forget about the current observation and move on to the next one. Thus, the Invariant Features approach leads to an algorithm that:

- requires memory constant in $M$, as opposed to Template Alignment,
- is *online* and can be parallelized.

# 3 Invariant Features

## 3.1 The First Few Moments of Signal

Recall from the previous section that we realized it may be possible to avoid direct estimation of the latent shifts in MRA, by instead looking at some invariant characteristics of the signal $x$ under cyclic shifts. The question that we need to resolve now has two levels: 1) Are there such invariant properties? 2) Is there a (small if possible) set of invariant features of $x$ under shifts, such that $x$ could be uniquely reconstructed from a set of estimates of these features?

Question 1) is easy to answer: for example, the mean of the signal $x$ is clearly invariant under the shifts. Question 2), however, will obviously require us to look beyond the mean, since it doesn't suffice to know the average value of a signal in order to reconstruct it. Let us therefore write out the mean along with a few other invariants of higher order:

$$\mu_x = \frac{1}{N}\sum_{n=0}^{N-1} x[n],$$

$$\text{autoCorr}_x[k_1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\overline{x[n-k_1]},$$

$$\text{tripleCorr}_x[k_1,k_2] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\overline{x[n-k_1]}x[n+k_2].$$

That these are all indeed invariants under the action of the group of shifts, is clear. We could try to estimate these from the observations, but that would not be very convenient. Instead, we will consider the Discrete Fourier Transforms of the autocorrelation and the triple correlation. Because the auto- and triple correlation are invariant under shifts, their DFT's too will be invariant under shifts. We shall denote, in what follows, the Fourier coefficients of $x$ by $y[k]$, $k \in \{0,\dots,N-1\}$.

The *power spectrum* $ps_x$ of $x$ is precisely defined to be the Fourier Transform of autoCorr$_x$, and is thus given by $ps_x[k] := |y[k]|^2$ for all Fourier frequencies $k \in \{0,\dots,N-1\}$.

**Definition 1** *The* bispectrum $B_x$ *of signal* $x$ *is defined to be the Fourier Transform of the triple correlation of* $x$*, namely* $B_x[k_1,k_2] := y[k_1]\overline{y[k_2]}y[k_2-k_1]$.

Using these two important Fourier Transforms, we can actually recover the signal. To begin with, note that for each Fourier frequency $k$, the *magnitude* of $y[k]$ can be recovered from the power spectrum as $\sqrt{ps_x[k]}$, $ps_x[k]$ being real nonnegative. Thus, the only unknown thing about $y$, the Fourier transform of $x$, are the *phases* of the Fourier coefficients. In the next Section, we will give explicit algorithms that recover $phase(y[k])$ for every $k \in \{0,\dots,N-1\}$ *based on the estimated bispectrum of the signal* $x$. Thus, those algorithms, coupled with the magnitudes estimated from the power spectrum, will give us complete knowledge of the DFT of $x$ and hence allow us to recover $x$. Presently, we shall state the result that will allow us to estimate the mean, power spectrum and the bispectrum of $x$ from the data $\xi_j$ - so that we can use these estimates to determine approximately the Discrete Fourier Transform of $x$, and thus $x$, by applying the just mentioned algorithms.

**Lemma 2** *The mean, power spectrum, and bispectrum of signal $x$ can be estimated from the $M$ data points $\xi_1, \ldots, \xi_M$ using estimators that are unbiased asymptotically as $M \to \infty$.*

**Proof.** We shall give these estimators here:

- $\hat{\mu_x} = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N}\sum_{n=0}^{N-1}\xi_j[n]\right)$
- $\hat{ps}_x[k] = \frac{1}{M}\sum_{j=1}^{M}(ps_{\xi_j}[k] - N\sigma^2)$
- $\hat{B}_{x-\mu_x} = \frac{1}{M}\sum_{j=1}^{M} B_{\xi_j - \hat{\mu_x}}$

For the estimator of the mean, is is distributed normally as $N(\mu_x, \frac{\sigma^2}{MN})$ hence in particular it is unbiased. The estimator of the power spectrum that we give here can be obtained from the fact that $\mathbb{E}[ps_{\xi_j}[k]] = ps_x[k] + N\sigma^2$, due to independence of the noise from the shift and from the signal. It follows that $\hat{ps}_x[k]$ is unbiased. The variance of the estimator $\hat{ps}_x[k]$ can be shown to be bounded from above by $\frac{\sigma^4}{M}$. Finally, the bispectrum estimator is shown to be asymptotically unbiased in Problem 1 on the Homework. ■

Now, let us state more precisely the result that the bispectrum is, in most cases, enough to uniquely reconstruct the underlying signal $x$. This is the theoretical foundation behind the Invariant Features Approach to MRA. We will state it along with the results that: there exists an estimator for the bispectrum such that the signal recovered using that estimator will have its bispectrum close to the estimator; and that a variant of our estimator for the bispectrum allows to approximate with high probability the actual bispectrum for any given precision $\delta > 0$, provided there are sufficiently many data points $M$.

**Theorem 3** *The following useful properties of the bispectrum hold.*

- (Signal can be recovered from its bispectrum) *We will state this for signals $x \in \mathbb{C}^N$, but there is a similar result for $x \in \mathbb{R}^N$. If $N > 4$, suppose the DFT $y$ of the signal $x$ is such that $y[k] \neq 0$ for any $k \in \{1, \ldots, K\}$, and $y[k] = 0$ for any $k \notin \{0, \ldots, K\}$. Suppose also that $K \geq (N+1)/2$. Then, up to an integer time shift, $x$ is uniquely determined by its bispectrum.*
- (Inverting the bispectrum) *For any $x \in \mathbb{R}^N$ or $\mathbb{C}^N$, such that the DFT of $x$ is nonvanishing, there are precision $\delta(x) > 0$ and sensitivity $L(x) < \infty$ such that if an estimator $\hat{B}_x$ satisfies $\left\|\hat{B}_x - B_x\right\|_F \leq \delta(x)$, then the signal $\hat{x}$ estimated from $\hat{B}_x$ satisfies $\min_{r\in\{0,\ldots,N-1\}} \|x = R_r\hat{x}\|_2 \leq L(x)\left\|\hat{B}_x - B_x\right\|_F$.*
- (Estimating the bispectrum) *Suppose $x$ is real or complex and its DFT is nonvanishing. Consider the estimator $\hat{B}_x = \frac{1}{M}\sum_{j=1}^{M} B_{\xi_j} - \sigma^2 N^2 \hat{\mu}_x A$, where $A$ is a particular constant matrix (we define it in Problem 1 of the Homework, and it comes in two flavors: $A_{\mathbb{C}}$ for complex signals and $A_{\mathbb{R}}$ for real signals). Fix any precision $\delta > 0$ and any probability $p < 1$. Then for any noise level $\sigma$, there is a constant $C(x, p, \delta)$ such that if the number of observations $M$ is at least $C \cdot (\sigma^2 + \sigma^6)$, then $\left\|\hat{B}_x - B_x\right\|_F \leq \delta$ with probability at least $p$.*

## 3.2 Signal Estimation: Invariant Features Approach to MRA

We now state precisely the elements of the Invariant Features Approach to Multireference Alignment. Our Algorithm is as follows:

**Algorithm 1** The Invariant Features Approach for MRA

```
INPUT: Observations $\xi_1, \dots, \xi_M$ according to MRA model with noise level $\sigma$
OUTPUT: $\hat{x}$: estimate of signal $x$
  procedure
    Compute $\hat{\mu}_x, \widehat{ps}_x$, and $\hat{B}_{x-\mu_x}$ according to the formulas in the proof of Lemma 2
    Estimate $y[0]$ from $\hat{\mu}_x$.
    For $k \in \{1, \dots, N-1\}$, estimate the magnitude $|y[k]|$ from the power spectrum, as $\sqrt{\widehat{ps}_x[k]}$
    For $k \in \{1, \dots, N-1\}$, estimate the phase $phase(y[k])$ from the bispectrum $\hat{B}_{x-\mu_x}$, using an algorithm such as those presented in Section 4.
    $\hat{x} \leftarrow Inverse\_DFT(\hat{y})$
    return $\hat{x}$
```

# 4 Algorithms for MRA

We cover in this section two algorithms for estimating Fourier phases of the signal based on the bispectrum: Non-Convex Optimization on the Phase Manifold, and Frequency Marching. The former is an algorithm with superior performance compared to other algorithms estimating Fourier phases from the bispectrum, and thus is important to understand. The latter shows a simpler and more straightforward approach to estimation of phases; it is different from the first algorithm in that it does not require initialization, and besides, it provides some insight into how phase information is stored in the bispectrum.

## 4.1 Non-Convex Optimization on Phase Manifold

The most important and best-performing in practice algorithm for estimating the Fourier phases from the bispectrum is given below. We introduce at this point the following notation which will be used henceforth: for any vector $z \in \mathbb{C}^N$, its *circulant matrix* $C(z)$ is defined to be an $N \times N$ matrix whose first column is $z$ and the $i$th column, for $i \in \{2, \dots, N\}$, is $z$ circularly shifted by $i$ units, that is, $R_i z$.

**Algorithm 2** Non-Convex Optimization on Phase Manifold

```
INPUT: Normalized bispectrum $B_{norm}[k_1, k_2]$, and an $N \times N$ weight matrix $W$
OUTPUT: $\hat{y}$: estimate of normalized Fourier transform $y_{norm}$ of signal
  procedure
    Compute using Riemann Trust Region Method:
    $\hat{y} = \arg\max_{z \in \mathbb{C}^N} \mathcal{RE}\{z^* Q(z) z\}$ subject to $|z[k]| = 1, \forall k$
                                   and additionally subject to $z[k] \in \mathbb{R}, \forall k$, if $x \in \mathbb{R}^N$ ,
    where $Q(z) := W \circ W \circ B_{norm} \circ \overline{C(z)}$
    return $\hat{y}$
```

- The first step is to understand where this (fairly generic) optimization problem comes from, that is, how are $z$ and $Q(z)$ related to the problem of MRA phase estimation. Note that the definition of the bispectrum, $B_x[k_1, k_2] := y[k_1]\overline{y[k_2]}y[k_2 - k_1]$, can be rewritten in matrix form as follows: $B_x = yy^* \circ C(y)$. We can normalize $y$ to obtain $y_{norm}$. Then, the normalized bispectrum will be $B_{norm} = y_{norm} y^*_{norm} \circ C(y_{norm})$. We let denote, for this section, $\hat{B}$ to be the estimate of the *phases* of the bispectrum, thus, of $B_{norm}$, and analogously $\hat{y}$ will denote the estimate of the phases of $y$ that we are seeking. We thus want to find such vector $\hat{y}$ that

$\hat{B}$ is as close as possible to $\hat{y}\hat{y}^* \circ C(\hat{y})$. To be more precise about the metric, we can define a least squares problem:

$$\text{Minimize } \left\| W \circ \left( \hat{B} - zz^* \circ C(z) \right) \right\|_F^2 \text{ for } z \in \mathbb{C}^N \text{ over the torus } \underbrace{S^1 \times S^1 \times \ldots \times S^1}_{N \text{ times}},$$

where $S^1 = \{\alpha \in \mathbb{C} : |\alpha| = 1\}$, and $W$ is an $N \times N$ weight matrix.

Now we can expand the squared Frobenius norm of the objective function, and get

$$\left\| W \circ \left( \hat{B} - zz^* \circ C(z) \right) \right\|_F^2 = \left\| W \circ \hat{B} \right\|_F^2 + \| W \circ zz^* \circ C(z) \|_F^2 - 2\langle W \circ \hat{B}, W \circ zz^* \circ C(z) \rangle.$$

Here, $\langle X, Y \rangle = \mathcal{RE}(trace(X^* Y))$ is the Frobenius inner product. We note that the first two terms in the sum on display are constant (the second one - since $z$ is on the torus), and the third term can be rewritten, by manipulating the trace, into $2\langle z, W \circ W \circ \hat{B} \circ \overline{C(z)} z \rangle$. Thus, we can denote $Q(z) := W \circ W \circ B_{norm} \circ \overline{C(z)}$, and so the original problem becomes a maximization problem on the same torus, with the objective function being $\mathcal{RE}\{z^* Q(z) z\}$. Hence, we have shown how the natural least-squares approach can be equivalently rewritten as the constrained optimization problem above.

- The second step is to see how such optimization problems are solved. The pseudocode says "use the Riemann Trust Region Method", and we will now (very briefly) point the reader in the direction of the related family of Riemannian Gradient methods, on which we will be unable to focus here due to space constraints.

  Note that the optimization is over a torus, which is a smooth manifold in $\mathbb{C}^N$. However, the objective function we are considering is non-convex, which makes it hard to compute global optima. Approaches such as the Riemannian Gradient Descent are known to converge to points that are first-order or second-order critical points. The gradient of our objective is

  $$\nabla \mathcal{RE}\{z^* Q(z) z\} = C(z)z + C(z)^* z + C^{adj}(zz^*)$$

  with $C^{adj}$ being the adjoint to $C(z)$ under the Frobenius inner product we are considering. Using the symmetries of the bispectrum matrix $B$, we can simplify the gradient formula, to yield

  $$\nabla \mathcal{RE}\{z^* Q(z) z\} = 2C(z)z + C(z)^* z.$$

  However, since the optimization is over a submanifold, the first-order conditions for extrema will not simply be $\nabla \mathcal{RE}\{z^* Q(z) z\} = 0$. Rather, we will need to consider the orthogonal projection of the gradient onto the tangent space to our torus at the point of interest. This projection will be called the *Riemannian gradient*.

  Please further consult reference paper [1] for the computation of the tangent space just mentioned and the projection of the gradient onto it, of the Hessian (needed for the second-order conditions), and finally for the discussion of the *retraction* method to perform Riemannian Gradient Ascent in this problem.

## 4.2 Frequency Marching

We now consider an algorithm that requires only the knowledge of the Fourier coefficient $y[1]$, and no further initialization besides $y[0], y[1]$, and the estimated bispectrum $\hat{B}_x[k_1, k_2]$ for all frequencies $k_1, k_2 \in \{0, \ldots, N\}$. This is in contrast to the above manifold optimization algorithm, where we need to initialize $W$. We will first state what Frequency Marching is, and then show that it works. Note that the algorithm deals with the normalized versions $B_{norm}$ and $y_{norm}$ of the bispectrum and Fourier transform, which is convenient for phase estimation. Namely, since each complex number on the unit circle can be given by $e^{ix}$ for some real $x \in [0, 2\pi)$, we can instead of the bispectrum $\hat{B}_x[k_1, k_2]$ consider real numbers (they should be considered $\mod 2\pi$) $\hat{\psi}[k_1, k_2]$, such that for all Fourier frequencies $k_1, k_2$, $\hat{B}_x[k_1, k_2] = e^{i\hat{\psi}[k_1,k_2]}$. Analogously, we can consider, instead of $y$, such $\hat{\psi}[k]$ (real numbers $\mod 2\pi$) for all Fourier frequencies $k$, that $y[k] = e^{i\hat{\psi}[k]}$. This notation turns the definition $B_x[k_1, k_2] := y[k_1]\overline{y[k_2]}y[k_2 - k_1]$ of the bispectrum into

$$\psi[k_1, k_2] = \psi[k_1] - \psi[k_2] + \psi[k_2 - k_1] \mod 2\pi.$$

As we shall see below, Frequency Marching amounts to estimating $\psi[2], \psi[3], \dots$ (in increasing order of frequencies) by using this identity to estimate the next $\psi[k]$ based on $\hat{\psi}[0], \dots, \hat{\psi}[k-1]$ and the estimated bispectrum exponents $\hat{\psi}[\cdot, \cdot]$, all of which are known. Note that as is evident from the below pseudocode, Frequency Marching has runtime $O(N^2)$.

**Algorithm 3** Frequency Marching

```
INPUT: Bispectrum $B[k_1, k_2]$, and FT of signal at frequencies $0$ and $1$: $y[0]$ and $y[1] \neq 0$
OUTPUT: $\hat{y}$: estimate of normalized Fourier transform $y_{norm}$ of signal
  procedure
      Normalize $B$, $y[0]$, $y[1]$: $B_{norm}[k_1, k_2] := e^{i\psi[k_1,k_2]}\ \forall k_1, k_2$, and $y_{norm}[k] := e^{i\psi[k]}, k = 0, 1$
      $\hat{y}[0] \leftarrow y_{norm}[0]$
      $e^{i\hat{\psi}[1]} \leftarrow y_{norm}[1]$
      for $k = 2, \dots, N$ do
          $avePhase \leftarrow phase\left(\sum_{q=1}^{\lfloor k/2 \rfloor} e^{i(\hat{\psi}[q]+\hat{\psi}[k-q]-\hat{\psi}[q,k])}\right)$
          if $avePhase \neq 0$ then
              $e^{i\hat{\psi}[k]} \leftarrow avePhase$
          else
              $e^{i\hat{\psi}[k]} \leftarrow 1$
      $\hat{y} \leftarrow e^{i\hat{\psi}}$
      return $\hat{y}$
```

To shed light onto the average phase computation, note the following. When we are estimating $\hat{\psi}[2]$, there is only one way to do so based on the formula $\psi[k_1, k_2] = \psi[k_1] - \psi[k_2] + \psi[k_2 - k_1] \mod 2\pi$. Namely, we have $\psi[1, 2] = \psi[1] - \psi[2] + \psi[1] \mod 2\pi$. From there, $\hat{\psi}[2] = 2\hat{\psi}[1] - \hat{\psi}[1, 2]$. Having estimated $\hat{\psi}[2]$, we can estimate $\hat{\psi}[3]$ also in just one way, namely from $\psi[1, 3] = \psi[1] - \psi[3] + \psi[2] \mod 2\pi$. However, already for $\hat{\psi}[4]$, we have equations

$$\psi[1, 4] = \psi[1] - \psi[4] + \psi[3] \mod 2\pi$$

and

$$\psi[2, 4] = \psi[2] - \psi[4] + \psi[2] \mod 2\pi$$

which give, in general, different estimates of $\hat{\psi}[4]$ given that the signal is noisy (of course, for noiseless signals, the values for $\psi[4]$ obtained from these two equations would be equal.)

In general, using this sequential approach to estimating Fourier phases, we have, as we invite the reader to check, $\lfloor k/2 \rfloor$ different ways to estimate $\psi[k]$. Namely, these ways would come from equations corresponding to the formulas for $\psi[1, k], \dots, \psi[\lfloor k/2 \rfloor, k]$. This is shown using the symmetry of the bispectrum matrix given by $B[k_2 - k_1, k_2] = B[k_1, k_2]$.

Now, given the $\lfloor k/2 \rfloor$ different estimates of $\psi[k]$, how do we reconcile them? Denote them by $e_1, \dots, e_s$, where $s = \lfloor k/2 \rfloor$. One natural way would be to look for a phase, that is, a number $z$ on the unit circle, such that the sum of the Euclidean distances from $z$ to $e_1, \dots, e_s$ would be minimized:

$$z = \arg\min_{z \in S^1} \sum_{k=1}^{s} \|z - e_k\|_2 \,.$$

Rewriting the norm, we have that $\mathcal{RE}\left(\bar{z} \sum_{k=1}^{s} e_k\right)$ needs to be maximized for $z \in S^1$. Note that $\bar{z} \sum_{k=1}^{s} e_k$ is nonnegative real, for the case when $\sum_{k=1}^{s} e_k \neq 0$, if and only if $z = phase\left(\sum_{k=1}^{s} e_k\right)$. This $z$ therefore delivers the maximum to our new objective, when $\sum_{k=1}^{s} e_k \neq 0$. Otherwise, of course, we can still take $z = phase\left(\sum_{k=1}^{s} e_k\right)$, since all points on the unit circle will maximize the expression.

This argument justifies why in the algorithm we average the phase estimates in that particular way.

# 5 Signal-to-Noise Ratio and Sample Complexity of MRA

An important question is how many measurements $M$ we need in order to estimate $x$ with arbitrary accuracy. This question is answered by the notion of *sample complexity*. For a trivial example of sample complexity, let us consider the following (super simple) model: our signal is a constant $\mu$, and we have $M$ observations $\xi_j$ of that signal corrupted by independent Gaussian noise with $\sigma^2$ variance: $\xi_j = \mu + \varepsilon_j$, with $\varepsilon_j \sim N(0, \sigma^2)$. Now, consider an obvious estimator $\hat{\mu} = \frac{1}{M} \sum_{j=1}^{M} \xi_j$. Now, how many measurements are required in order for $\hat{\mu}$ to be arbitrarily close to $\mu$ in some natural sense (e.g., in the $L_2$ metric)? Notice that $Var(\hat{\mu}) = \frac{\sigma^2}{M}$. Therefore, to drive the variance of the estimator to 0, thus making it arbitrarily close to $\mu$, we need (and it is enough) to have $M$ at least on the order of $\sigma^2$. Thus, the sample complexity of this model is $\sigma^2$.

Now, for any model where each measurement is the sum of a signal and the additive noise, the *Signal-to-Noise Ratio (SNR)* is defined as $\left(\frac{A_{signal}}{A_{noise}}\right)^2$, where $A_{signal}$ and $A_{noise}$ are the root-mean square amplitudes of the signal and of the noise. For the particular simple model just considered above, its SNR is, from this definition, proportional to $\frac{1}{\sigma^2}$. Then, remark that the sample complexity of the model becomes (proportional to) $1/SNR$, as it was shown above to be on the order of $\sigma^2$. In fact, for many models it is the case that the sample complexity is inversely proportional to the $SNR$ of that model.

Going back to MRA, we define $SNR := \frac{\|x\|_2}{\sigma^2}$ for this model. However, we can WLOG assume that the signal is normalized, and so from here on we define, for the MRA model, $SNR := \sigma^{-2}$. Further, as a measure of closeness between the signal $x$ and its estimator $\hat{x}$ we accept the mean square error (MSE), defined for any given asymptotically unbiased estimator $\hat{x}$ as

$$\frac{1}{\|x\|_2^2} \mathbb{E}\left[\min_{r \in \{0, \dots, N-1\}} \|R_r \hat{x} - x\|_2^2\right].$$

Below, we will discuss tight bounds on the sample complexity of MRA under certain further assumptions on the model. Despite the fact that, as stated above, many natural models have sample complexity that is inversely proportional to their Signal-to-Noise Ratio, for MRA this will surprisingly not be the case: the sample complexity of MRA will be $1/SNR^3$, for instance, under the assumption that the latent shifts are distributed uniformly at random. See Section 5.1. There is a quite large class of latent shift distributions, however, for which the sample complexity will be lower, namely $1/SNR^2$. This is discussed in Section 5.2.

## 5.1 Lower and Upper Bounds for Uniform Distribution of Shifts

We begin with the observation that the original MRA problem of finding $x$ given data points $\xi_j = R_{r_j}x + \varepsilon_j$ can be reduced to the following: given uniformly random translations $R_{r_j}$ (meaning that each $r_j$ is sampled from the uniform distribution on $\{0, \dots, N-1\}$), and given $M$ data points $\xi_j = R_{r_j}x + \varepsilon_j$, estimate $x$. Indeed, notice that to each observation $\xi_j$ in the classical MRA setting, we can apply a uniformly random shift $R_{u_j}$, and then $R_{u_j}\xi_j$ is the noisy version of $x$ shifted by $u_j + r_j$ units, where $u_j + r_j$ has the same uniform distribution as $u_j$. Thus, all observations can be turned into observations of $x$ with a uniformly random shift.

First, we shall require the folowing definitions:

**Definition 4** *A* generic signal *is a signal whose Fourier coefficients $y[k] \neq 0$ for any frequency $k \in \{0, \dots, N-1\}$.*

**Definition 5** *The* shift-invariant distance *between vectors $x$ and $z$ is defined as*

$$\rho(x, z) = \min_{r \in \{0, \dots, N-1\}} \|x - R_r z\|_2 .$$

We shall restrict our attention to generic signals for this subsection. There is not much loss in assuming that the signal is generic, since obviously the set of all non-generic signals has Lebesque measure zero. In this setting, surprising results have been obtained: for generic signals, it turns out that determining the original signal $x$ from its shifts has sample complexity $\frac{1}{SNR}^3$ in the following strong sense.

**Theorem 6** *Consider the model $\xi_j = R_{r_j}x + \varepsilon_j$, where the shifts $R_{r_j}$ are by a uniformly random amount. Suppose the signal $x$ is generic. Then the following hold:*

- (Lower Bound) *Suppose $N > 2$, $\sigma^2 \geq 1$, and suppose $\delta > 0$ is sufficiently small. Then for some universal constant $C$, with constant probability for any estimator $\hat{x}$ that is based on $M$ data points, it holds that there exists a generic signal $x \in \mathbb{R}^N$ with $\|x\|_2 = 1$ such that $\rho(x, \hat{x}) \geq \varepsilon$ whenever $M \leq C\varepsilon^{-2}\sigma^6$.*
- (Upper Bound) *There exists an algorithm, called* Jennrich's algorithm*, that solves the MRA problem for generic signals and whose sample complexity is $\frac{1}{SNR}^3$.*

We remark that the lower bound statement implies that if we want to have an estimator $\hat{x}$ that satisfies $\rho(\hat{x}, x) < \varepsilon$ for any given $\varepsilon > 0$ with probability close to 1, then our number of observations $M$ must be $M \geq C\varepsilon^{-2}\sigma^6$, or in other words, the sample complexity of the algorithm must be at least $\frac{1}{SNR}^3$.

Now we describe Jennrich's algorithm, without proof. First, we define the *moment tensors* of the uniform distribution over the possible shifts $\{R_0x, R_1x, \dots, R_{N-1}x\}$ of signal $x$ as

$$T^{(l)}(x) := \frac{1}{N}\sum_{r=0}^{N-1}(R_r x)^{\otimes l}, \text{ where } l = 1, 2, \dots$$

Analogously to the discussion above, where the mean, power spectrum, and bispectrum of a signal were enough to completely specify it, it turns out that the first three moments of the uniform distribution, as just defined, are enough to recover $x$.

Our basic plan will therefore be the following.

- We will estimate the third moment tensor using the estimate

$$\hat{T}^{(3)} = \frac{1}{NM} \sum_{i=1}^{M} \sum_{j=0}^{N-1} \left[ (R_{r_j}\xi_i)^{\otimes 3} - 3sym(R_{r_j}\xi_i \otimes I_N) \right].$$

  Here, the symmetric tensor $sym(A)$ for any tensor $A$ is defined by

$$sym(A)_{a_1,\dots,a_k} = \frac{1}{k!} \sum_{\pi \in S_k} A_{\pi(a_1),\dots,\pi(a_k)}.$$

  It can be shown that $\hat{T}^{(3)}$ is an unbiased estimator of the third moment $T^{(3)}$ and each of its entries has variance of order $\frac{\sigma^6}{M}$, provided $\sigma$ is bounded from below by some positive constant.

- We will now, based on the obtained estimate $\hat{T}^{(3)}$, estimate the original signal $x$ by using the fact that $\hat{T}^{(3)}$ is close to the true value $\frac{1}{N}\sum_{r=0}^{N-1}(R_r x)^{\otimes l}$ of the third moment of $x$ and applying a tensor decomposition algorithm.

As a tensor decomposition algorithm, we shall employ a version of the so-called Jennrich's Algorithm. It takes as input a tensor $T$ which is known to be a noisy version of the sum of tensors $\sum_{j=0}^{N-1} u_j \otimes u_j \otimes u_j = \sum_{j=0}^{N-1} u_j^{\otimes 3}$ for some unknown set of $N \times 1$ vectors $u_0, \dots, u_{N-1}$, and it outputs the matrix $U = [\hat{u}_0, \dots \hat{u}_{N-1}]$ of estimates of the vectors $u_j$. The precise way Jennrich's Algorithm performs this decomposition is as follows:

- Choose two uniformly random unit vectors $a, b \in \mathbb{R}^N$. Form the matrices $A := \sum_{j=0}^{N-1} \langle a, u_j \rangle u_j^{\otimes 2}$ and $B := \sum_{j=0}^{N-1} \langle b, u_j \rangle u_j^{\otimes 2}$, by noticing that for all $i, j$,

$$A_{ij} = \sum_k T_{ijk} a_k$$

  and

$$B_{ij} = \sum_k T_{ijk} b_k.$$

- Perform SVD (singular value decomposition) on matrix $A$, and let $W = [v_0, \dots, v_{N-1}]$, where the $v_j$'s are the $N$ left singular vectors of $A$ corresponding to the first $N$ singular values of $A$.

- Compute $Q = W^T A W [W^T B W]^{-1}$. Find the matrix $P$ such that the eigendecomposition of $Q$ is $PDP^{-1}$ for $D$ diagonal matrix.

- Output $U = WP$.

In order to disentangle $\hat{T}^{(3)} \approx \frac{1}{N}\sum_{r=0}^{N-1}(R_r x)^{\otimes 3}$ and elicit the estimate $\hat{x}$ on $x$, which is our goal, we now simply apply Jennrich's algorithm with vectors $u_r$, such that $u_r = R_r x$ for all $r \in \{0, \dots, N-1\}$. Then, Jennrich's Algorithm will output the estimate $U = [\hat{x}, R_1\hat{x}, \dots, R_{N-1}\hat{x}]$. Recall that we can only estimate the signal $x$ up to a constant shift, therefore it suffices to just take any column of this matrix $U$ to be the recovered signal $x$.

## 5.2 What if Distribution of Translations is Aperiodic?

In the preceding subsection, we replaced the latent shifts of $x$ from the MRA model by random shifts coming from the uniform distribution. We then showed that the optimal sample complexity of an algorithm that recovers $x$ is surprisingly high, namely $1/SNR^3$.

Now let us replace this assumption of uniform distribution and consider the general case where the shifts come from some distribution $\rho$, where $\rho[k]$ is the probability that the shift was by $k$ units, $k \in \{0, \dots, N-1\}$.

Why did we incur such high sample complexity in the case when $\rho \sim Unif(\{0, \dots, N-1\})$? Intuitively, it is always harder to reconstruct a function that is highly aperiodic versus a function possessing a high degree of periodicity. Indeed, for aperiodic functions even local estimation often suffices, while in the periodic case estimating the function on any part of its domain leaves open e.g. the question: what amount is the version of the function that we are estimating shifted by?

Let us introduce the following definition:

**Definition 7** *A distribution $\rho$ is called* periodic *if there is an $n \in \{1, \dots, N-1\}$ such that $\rho[k] = \rho[k+n]$ for all $k \in \{0, \dots, N-1\}$. Otherwise, it is called* aperiodic.

The situation with respect to the periodicity of the distribution $\rho$ of the shifts turns out to correspond quite well to the intuitive explanation provided above. Indeed, it turns out that for generic signals (see definition in the previous section), *the optimal sample complexity of an algorithm to estimate the signal $x$ drops to $1/SNR^2$ in the case when the distribution $\rho$ is aperiodic.*

We shall summarize the lower bound on the sample complexity for aperiodic and for periodic $\rho$, and then we shall constructively, by giving an explicit algorithm, prove the matching upper bound of $1/SNR^2$ on the sample complexity for the special case when $\rho$ has a unique entry.

**Theorem 8** *Consider an (asymptotically) unbiased estimator $\hat{x}$ of $x$. One has the following lower bounds on the MSE of the estimator:*

- (Aperiodic $\rho$) $MSE \geq \frac{1}{8M} SNR^{-2} - O(\frac{1}{M} SNR^{-1.5})$.
- (Periodic $\rho$) *Supposing $\rho$ is periodic with period $n < N/2$, then*

$$MSE \geq \frac{\frac{N}{2n} - 1}{54M} SNR^{-3} - O(\frac{1}{M} SNR^{-2.5}).$$

Notice how the second part of the claim generalizes the lower bound given in the previous subsection for the special case of uniform $\rho$. It shows that for *any* periodic distribution we incur at least the sample complexity of the order $1/SNR^3$.

Now let us quickly understand what will allow us to break the $1/SNR^3$ barrier in the case of aperiodic $\rho$ and reduce it down to $1/SNR^2$. Recall that the reason for having sample complexity $1/SNR^3$ in the case of uniform $\rho$ was in the fact that it was only possible to estimate the underlying signal $x$ if one uses at least the first *three* moments of the uniform distribution. Since the estimate of the third moment was a 3rd degree polynomial in the signal, then understandably the required sample complexity needed to be at least $\sigma^6$, that is, $1/SNR^3$. By contrast, we shall specify how to use only the first *two* moments of the distribution of the translations, when $\rho$ is aperiodic, to exactly recover $x$. Then, since the variance of the estimator of the second moment is bounded by a quantity proportional to $\sigma^4/M$, it becomes possible to estimate the signal $x$ in $1/SNR^2$. Let us see how this is done.

We shall concentrate on the special case of aperiodic distribution $\rho$ having at least one unique entry (that is, for some $k_0$, $\rho[k_0] \neq \rho[k]$ for any $k \neq k_0$). Clearly, if this is the case then $\rho$ is indeed aperiodic. We note here, however, that the tight bound of $1/SNR^2$ holds for all aperiodic distributions, and we are just focusing on the proposed case for simplicity.

As a preliminary, let us recall the general definition of the 1st and 2nd moment of the distribution of shifts of $x$, the special case of which we have seen for the uniform distribution. Namely,

$$T^{(1)} := \mathbb{E}[(R_r x)],$$

and

$$T^{(2)} := \mathbb{E}[(R_r x)^{\otimes 2}] = \mathbb{E}[(R_r x)(R_r x)^T].$$

The expectation is taken over the distribution $r \sim \rho$. Now we have an obvious (asymptotically as $M \to \infty$) unbiased estimator for the first moment, being

$$\hat{T}^{(1)} := \frac{1}{M}\sum_{j=1}^{M} \xi_j.$$

For the second moment, its asymptotically unbiased estimator will be

$$\hat{T}^{(2)} := \frac{1}{M}\sum_{j=1}^{M} \xi_j \xi_j^T - \sigma^2 I_N.$$

Precisely, the statement of the result that the signal $x$ in the case of aperiodic $\rho$ can be determined by its first two moments only is:

**Lemma 9** *Suppose $x$ is generic, and $\rho$ aperiodic. Then for any other signal $x_1$ and distribution $\rho_1$ with the same first two moments as $x$ and $\rho$, one has that $x_1$ and $x$ and $\rho_1$ and $\rho$ are equal, respectively, up to a shift: there is $r \in \{0, \ldots, N-1\}$ such that $x_1 = R_r x$ and $\rho_1 = R_r \rho$.*

**Proof.** Prove it as Exercise 3 on the Homework. ■

Now, we are ready to state the positive algorithmic result for distributions $\rho$ with a unique entry.

**Theorem 10** *Suppose $x$ is generic, and $\rho$ has at least one unique entry, in the sense described above. Then the following Algorithm, which operates based on estimating the first two moments $T^{(1)} := \mathbb{E}[(R_r x)]$, and $T^{(2)} := \mathbb{E}[(R_r x)^{\otimes 2}]$, recovers both $x$ and $\rho$ exactly up to a translation.*

**Algorithm 4** Recovering $x$ and $\rho$ from the first two moments.

```
INPUT: Estimated moments $T^{(1)}$ and $T^{(2)}$
OUTPUT: Signal $x$ and shift distribution $\rho$
  procedure
    $ps[x] \leftarrow Mdiag(FT^{(2)}F^{-1})$
    $u \leftarrow (PS[x])^{-1/2}$
    $Q \leftarrow F^{-1}D_u F$
    $T^{(2)}_{norm} \leftarrow QT^{(2)}Q^*$
    $v \leftarrow UniqueEigenvector(T^{(2)}_{norm})$
    $v_{norm} \leftarrow F^{-1}(ps^{1/2} \circ Fv)$
    $x \leftarrow \frac{Sum(T^{(1)})}{Sum(v_{norm})} v_{norm}$
    $\rho \leftarrow C^{-1}[x]T^{(1)}$
  return $x, \rho$
```

Here, the notation is as follows: $Sum(t)$ denotes the sum of entries of vector $t$; $C[t]$ is a circulant matrix (see previous sections) with vector $t$ as its first column; $D_t$ is the diagonal matrix of vector $t$; $\circ$ denotes Hadamard (entrywise) product; $UniqueEigenvector(A)$ is an eigenvector of matrix $A$ corresponding to an eigenvalue with the dimension of eigenspace 1; $F$ is the Discrete Fourier Transform matrix; $ps[x]$ is the power spectrum of $x$.

**Proof.** Note that $T^{(2)} = C[x]D_\rho C[x]^T$ (see Homework Exercises). Note that the diagonalization of $C[x]$ by the Fourier matrix $F$ is as follows: $C[x] = F^{-1}D_{Fx}F$; indeed, this corresponds exactly to change of basis from the regular to the Fourier basis, in which the frequencies of $x$ are given by $Fx$.

From here, we have that

$$FT^{(2)}F^{-1} = \frac{1}{M}D_{Fx}C[F\rho]D_{\bar{F}x}.$$

Therefore, for every $k$, $(FT^{(2)}F^{-1})_{kk} = \frac{1}{M}(F\rho)_0|(Fx)_k|^2$. Notice that since $\rho$ is a probability distribution, $(F\rho)_0 = \sum_j \rho[j] = 1$. Further, recall the formula for the power spectrum $ps[x]$ is $(ps[x])_k := |(Fx)_k|^2$. Therefore, from the previous display we actually deduce that $(FT^{(2)}F^{-1})_{kk} = \frac{1}{M}(ps[x])_k$, and so the power spectrum of $x$ can be recovered as

$$ps[x] = Mdiag(FT^{(2)}F^{-1}).$$

As the next step, consider the matrix $T^{(2)}_{norm}$ obtained by conjugating $T^{(2)}$ by $F^{-1}D1/|ps[x]|^{1/2}F$. Note that then, $T^{(2)}_{norm} = C[x_{norm}]D\rho C[x_{norm}]^T$, where $x_{norm}$ is the vector whose FT is the normalized FT of $x$, namely $(Fx_{norm})_k = \frac{(Fx)_k}{|(Fx)_k|}$. Because of this property of $x_{norm}$, $C[x_{norm}]$ is orthonormal and real. Therefore, $T^{(2)}_{norm} = C[x_{norm}]D\rho C[x_{norm}]^T$ is in fact an eigendecomposition of $T^{(2)}_{norm}$, and so its eigenvalues are the components of $\rho$ and the eigenvectors are the columns of $C[x_{norm}]$, that is, the translations of $x_{norm}$.

Now, since $\rho$ has a unique entry, then the eigenvector $v$ corresponding to that entry is parallel to a translation of $x_{norm}$, namely $v = \alpha R_r x_{norm}$ for some scalar $\alpha$ and shift $r$. Then, un-normalizing the still normalized FT coefficients, we have by coordinatewise multiplication that

$$v_{norm} = \alpha F^{-1}(FR_r x_{norm} \circ |ps[x]|^{1/2}) = \alpha R_r x.$$

At this point, we can compute $\alpha = Sum(v_{norm})/Sum(x) = Sum(v_{norm})/Sum(T^{(1)})$. This comes from $Sum(x) = Sum(T^{(1)})$, which is because the sum of the elements of the first moment (average) of $x$ must be equal to the sum of the entries of $x$.

Therefore, $R_r x = v_{norm}/\alpha$. As for $\rho$, note that $T^{(1)} = x * \rho$, and therefore $T^{(1)} = C[x]\rho$, hence $\rho = C[x]^{-1}T^{(1)}$, QED. ■

# 6 Bibliography and Where to Refer for Details

## 6.1 List of References and their Corresponding Sections in the Notes

1 T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer: *Bispectrum Inversion with Application to Multireference Alignment.* IEEE Transactions on Signal Processing Vol.66 No.4, 2018

- **Corresponding Sections:** Sections 2-4

2 E. Abbe, T. Bendory, W. Leeb, J. Pereira, N. Sharon, and A. Singer: *Multireference Alignment is Easier with an Aperiodic Translation Distribution*. arXiv:1710.02793v2, 2018

- **Corresponding Sections:** Subsection 5.2

3 A. Perry, J. Weed, A. Bandeira, P. Rigolet, and A. Singer: *The Sample Complexity of Multireference Alignment*. Manuscript, 2018

- **Corresponding Sections:** Subsection 5.1

4 Prof. Singer's Notes for MAT 586, Lecture 12