Multi-reference alignment in high dimensions: sample complexity and phase transition

Elad Romanov *, Tamir Bendory†, and Or Ordentlich‡

Abstract. Multi-reference alignment entails estimating a signal in $\mathbb{R}^L$ from its circularly-shifted and noisy copies. This problem has been studied thoroughly in recent years, focusing on the finite-dimensional setting (fixed L). Motivated by single-particle cryo-electron microscopy, we analyze the sample complexity of the problem in the high-dimensional regime $L \to \infty$. Our analysis uncovers a phase transition phenomenon governed by the parameter $\alpha = L/(\sigma^2 \log L)$, where σ^2 is the variance of the noise. When $\alpha > 2$, the impact of the unknown circular shifts on the sample complexity is minor. Namely, the number of measurements required to achieve a desired accuracy ε approaches σ^2/ε for small ε; this is the sample complexity of estimating a signal in additive white Gaussian noise, which does not involve shifts. In sharp contrast, when $\alpha \le 2$, the problem is significantly harder and the sample complexity grows substantially quicker with σ^2.

Key words. multi-reference alignment, information-theoretic lower bounds, estimation in high dimensions, mathematics of cryo-EM imaging

AMS subject classifications. 62B10, 94A15, 94A12, 62F99

1. Introduction.

We study the sample complexity of the multi-reference alignment (MRA) model: the problem of estimating a signal from its circularly-shifted and noisy copies. Specifically, let $X \sim \mathcal{N}(0, I)$ be an L-dimensional vector with i.i.d. standard normal entries. We collect n independent measurements of random cyclic shifts of X, corrupted by additive white Gaussian noise:

$$Y_i = R_{\ell_i} X + \sigma Z_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where R_ℓ denotes a cyclic shift, namely, $(R_\ell X)_j = X_{(j+\ell) \bmod L}$ for all $j = 0, \dots, L-1$, $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, and $\ell_i \overset{i.i.d.}{\sim} \text{Uniform}(\{0, \dots, L-1\})$ are statistically independent of X. Given the measurements $Y^n = (Y_1, \dots, Y_n)$, one is interested in constructing an estimator $\hat{X} = \hat{X}(Y^n)$ of the signal. Importantly, the unknown shifts $\ell_1, \dots, \ell_n$—while their estimation might be a means to an end—are nuisance variables. Figure 1 shows an example of a measurement drawn from (1.1).

This paper focuses on the high-dimensional regime, where the dimension of the signal grows indefinitely $L \to \infty$. In this setting, we wish to characterize the relations between the number of measurements n, the length of each observation L, and the noise level σ^2 that allow estimating X to a prescribed accuracy. This is in contrast to previous works, surveyed in Section 3, which analyzed the interplay between n and σ, while considering a fixed L.

*School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel (elad.romanov@mail.huji.ac.il).

†School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel (bendory@tauex.tau.ac.il).,

‡School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel (or.ordentlich@mail.huji.ac.il).

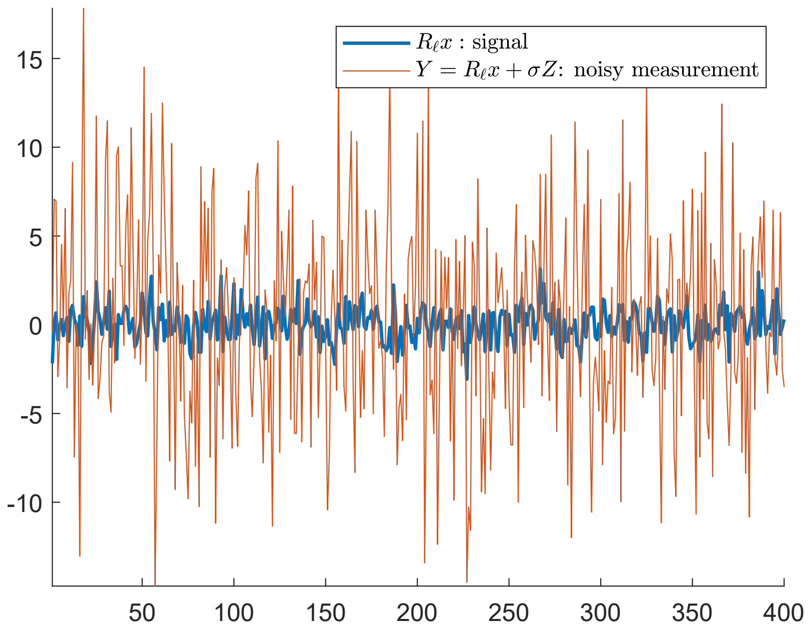

Figure 1. *An example of a measurement drawn from (1.1) for $\alpha = 2$ and $L = 400$. The corresponding noise level is $\sigma^2 = 33.38$.*

It is important to note that given the measurements, there is no way to distinguish between X and its cyclic shift since $P_{Y^n|X=x} = P_{Y^n|X=R_1 x} = \cdots = P_{Y^n|X=R_{L-1}x}$. Therefore, we can only estimate the orbit of X under the group of circular shifts $\mathbb{Z}_L$. Accordingly, we use the following distortion measure

$$\rho(X, \hat{X}) = \frac{1}{L} \min_{\ell=0,\ldots L-1} \|X - R_\ell \hat{X}\|^2. \tag{1.2}$$

In the sequel, we loosely say that we aim to estimate X rather than its orbit, and refer to $\mathbb{E}\rho(X, \hat{X})$ as the MSE.

Sample complexity. Our goal in this paper is to characterize the smallest possible number of measurements required to achieve a desired MSE in terms of the dimension L and the noise level σ^2. To that end, we define the smallest MSE attainable by any estimator as

$$\mathrm{MSE}^*_{\mathrm{MRA}}(L, \sigma^2, n) := \inf_{\hat{X}} \mathbb{E}\rho(X, \hat{X}(Y^n)), \tag{1.3}$$

and the sample complexity of the MRA problem

$$n^*_{\mathrm{MRA}}(L, \sigma^2, \varepsilon) := \min \left\{ n \ : \mathrm{MSE}^*_{\mathrm{MRA}}(L, \sigma^2, n) \le \varepsilon \right\}. \tag{1.4}$$

We define the signal-to-noise ratio (SNR) by

$$\mathrm{SNR} := \frac{\mathbb{E}\|X\|^2}{\sigma^2} = \frac{L}{\sigma^2}. \tag{1.5}$$

This definition is consistent with previous works which considered a fixed L and $\sigma \to \infty$, implying SNR$\to 0$; see Section 3.

The asymptotics in our model turn out to be particularly interesting when the dimension, the noise level, and the SNR are simultaneously large. In particular, it will be convenient to parametrize the noise variance by

$$\sigma^2(\alpha) = \frac{L}{\alpha \log L} \quad \Longleftrightarrow \quad \alpha = \frac{L}{\sigma^2 \log L} = \frac{\text{SNR}}{\log L}. \tag{1.6}$$

Accordingly, we define $\text{MSE}^*_{\text{MRA}}(L, \alpha, n) := \text{MSE}^*_{\text{MRA}}(L, \sigma^2(\alpha), n)$ and $n^*_{\text{MRA}}(L, \alpha, \varepsilon) := n^*_{\text{MRA}}(L, \sigma^2(\alpha), \varepsilon)$.

Motivation. The MRA model is mainly motivated by single-particle cryo-electron microscopy (cryo-EM)—a leading technology to constitute the 3-D structure of biological molecules. In its most simplified version, the cryo-EM problem involves reconstructing a 3-D structure from its multiple noisy tomographic projections, taken after the structure has been rotated by an unknown 3-D rotation. In analogy, in the MRA model (1.1) the signal X is measured after an unknown circular shift. In Theorem 2.3, we extend the basic model to include a projection; we refer to this model as the projected MRA model. This projection plays the role, to some extent, of the tomographic projection in cryo-EM. Section 7 discusses further potential extensions.

The correspondence between MRA and cryo-EM, while admittedly not perfect, has motivated an extensive study of the MRA problem in recent years. For example, the resolution limitations of MRA were analyzed in [12] in order to draw an analogy to the achievable resolution of cryo-EM—a crucial aspect from a biological standpoint. More relevant to this work, in [3,6,8,32], the sample complexity of the MRA and cryo-EM models were analyzed for a fixed dimension L. Remarkably, it was shown that in the low noise regime (small σ), the number of measurements should scale like σ^2, while in the high noise regime (large σ) n must increase with σ^6; see further discussion in Section 3.

Our high-dimensional analysis is motivated by the size of modern cryo-EM datasets. In a typical cryo-EM experiment, the number of measurements and the dimension of the 3-D structure are of the same order of a few millions. For example, a 3-D structure of size $200 \times 200 \times 200$ voxels resulting in $8,000,000$ parameters to be estimated. Since a typical noise level in a cryo-EM dataset is $\sigma^2 \approx 100$, the anticipated parameter regime is $\alpha \gg 1$. We do emphasize, however, that these numbers should be taken with some degree of skepticism: while cryo-EM is a motivation for studying the MRA problem, these are ultimately quite different problems, and practical cryo-EM setups involve additional complications, that are not captured by MRA [10]. In fact, high-dimensional statistical analysis has been already proven to be effective for cryo-EM data processing. For example, a covariance estimation technique based on high-dimensional analysis (the so-called spiked model) has significantly improved image denoising [14].

Information-theoretic background and asymptotic notation. The analysis of this work is greatly based on information-theoretic notions and techniques. For completeness, we review the relevant definitions in supporting information (SI) appendix, Section SM1.

We also repeatedly use asymptotic notation. For sequences $a = a(L)$ and $b = b(L)$, we write $a(L) = O(b(L))$ if there exists a constant $C > 0$ such that $a(L) \leq Cb(L)$ for all L. Similarly, $a(L) = \Omega(b(L))$ means $a(L) \geq Cb(L)$. Occasionally, we use $a(L) = O_\beta(b(L))$ to signify explicitly that C depends on some parameter β. The notation $a(L) = o(b(L))$ means $a(L)/b(L) \to 0$ as $L \to \infty$. In particular, if $a(L) = o(1)$ then $a(L) \to 0$ asymptotically. Similarly, $a(L) = \omega(b(L))$ means $a(L)/b(L) \to \infty$.

Reproducibility. The code to reproduce the figures is publicly available at https://github.com/TamirBendory/high-dimensional-mra-bounds.[1]

Supporting information (SI). Due to space constraints, we have relegated the proofs of several technical claims to the SI appendix. In addition to those, the SI contains a brief review of all information-theoretic notions necessary to follow this work (Section SM1), as well as some additional discussion which is somewhat tangential to our main results (Section SM2).

2. Main results and discussion.

Phase transition.. This work focuses on the asymptotic setting where L tends to infinity. Our first main finding is that in this asymptotic limit there is a transition in terms of the behavior of the sample complexity. For $\alpha > 2$, the MRA problem is essentially as easy as estimating a signal in additive white Gaussian noise (AWGN), with no random shifts. More precisely, for sufficiently small distortion ε, the sample complexity tends to the sample complexity of estimating a signal in AWGN, $n^*_{\text{AWGN}}(L, \alpha, \varepsilon) = \lceil \left(\frac{1}{\varepsilon} - 1\right) \sigma^2(\alpha) \rceil$, which behaves as $\frac{\sigma^2(\alpha)}{\varepsilon}$ for small ε. In sharp contrast, for $\alpha \leq 2$ the problem becomes substantially harder.

Theorem 2.1. *The sample complexity of the MRA model* (1.1) *obeys:*

1. *For any $\alpha > 2$ we have*

$$\lim_{\varepsilon \to 0} \lim_{L \to \infty} \frac{n^*_{MRA}(L, \alpha, \varepsilon)}{\sigma^2(\alpha)/\varepsilon} = \lim_{\varepsilon \to 0} \lim_{L \to \infty} \frac{n^*_{MRA}(L, \alpha, \varepsilon)}{n^*_{AWGN}(L, \alpha, \varepsilon)} = 1.$$

2. *For any $\alpha \leq 2$ and any $\varepsilon < 1$ we have*

$$n^*_{MRA}(L, \alpha, \varepsilon) = \omega\left(\sigma^2 \log(1/\varepsilon)\right).$$

In particular, for fixed ε,

$$\lim_{L \to \infty} \frac{n^*_{MRA}(L, \alpha, \varepsilon)}{n^*_{AWGN}(L, \alpha, \varepsilon)} = \infty.$$

In part 1 of Theorem 2.1, the lower bound $\frac{n^*_{\text{MRA}}(L,\alpha,\varepsilon)}{n^*_{\text{AWGN}}(L,\alpha,\varepsilon)} \geq 1$ is trivial: estimating in the MRA model is harder than estimating a signal in AWGN (namely, when the shifts are known). A small subtlety is that the distortion measure $\mathbb{E}\rho(X, \hat{X})$ is a bit weaker than the standard definition of MSE, $\mathbb{E}\|X - \hat{X}\|^2$, as it allows for any cyclic shift. However, we show in Section 5 that, as expected, this has a vanishing effect for large L. In order to show that $\lim_{\varepsilon \to 0} \lim_{L \to \infty} \frac{n^*_{\text{MRA}}(L,\alpha,\varepsilon)}{n^*_{\text{AWGN}}(L,\alpha,\varepsilon)} \leq 1$ we introduce an algorithm that for any $\alpha > 2$ requires about $\sigma^2(\alpha)/\varepsilon$ samples to achieve $\mathbb{E}\rho(X, \hat{X}) \leq \varepsilon$, provided that ε is sufficiently small and L is sufficiently large. The sole purpose of the estimation procedure is establishing an upper bound; its computational complexity is exponential in L and thus the procedure is far from being efficient. More specifically, it is based on a two-step procedure. First, we construct a δ-net

[1] Our expectation-maximization implementation is based on the code of [11].

that, by definition, contains a member close to X and look for the most likely candidate within that net given the measurements. Second, we use this candidate in order to determine almost all shifts $\hat{\ell}_i$, and then estimate the signal by alignment and averaging $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} R_{-\hat{\ell}_i} Y_i$. The details are given in Section 6.

In order to establish part 2 of Theorem 2.1, we show that for $\alpha \leq 2$ the mutual information (MI) $I(X;Y)$ between X and a single MRA measurement grows with L significantly slower than $I(X; X + \sigma Z)$, as in estimating a signal in AWGN. The details are given in Section 5.

Although our results are *asymptotic* in L, the transition in the difficulty of the problem around $\alpha = 2$, as predicted by Theorem 2.1, is evident already for relatively small L. Figure 2 presents the root MSE (RMSE) as a function of α for different values of L. We take our estimator $\hat{X}$ to be the output of the expectation-maximization (EM) algorithm [11, 20], which is the standard choice for MRA; see details in Section 3. For large values of L and large α, the error of EM tends to that of estimating a signal in AWGN, implying that it detects the shifts accurately. For smaller values of α, the error grows rapidly, especially when $\alpha < 2$. We note that the observed transition in the vicinity of $\alpha = 2$, at the values of L considered in Figure 2 (few 100s), appears to not be very sharp. Our proofs suggest that perhaps this behavior is to be expected: the concentration rates we are able to derive for some of the quantities relevant to the analysis is quite slow (inverse polynomial in L, with a very small exponent when α is close to 2).

Connection with template matching. At this point, the reader may wonder what is the intuitive interpretation of $\alpha = 2$. To answer this question we now introduce the *template matching problem*, which is studied in detail in Section 4. In this problem, we are given X and one MRA measurement $Y = R_\ell X + Z$, where X, R_ℓ and Z are distributed as above, and our goal is to recover the shift R_ℓ. We will see that in the asymptotic setting, $\alpha = 2$ is the critical threshold for this problem. That is, the error probability in recovering R_ℓ from (X, Y) approaches 0 for all $\alpha > 2$, and approaches 1 for all $\alpha < 2$.

In the MRA problem, recovering the shifts is harder, as we do not have access to X. We nevertheless show that for $\alpha > 2$, given enough measurements, it is possible to recover a fraction approaching 1 of the shifts correctly. On the other hand, recovering a large fraction of the shifts correctly for $\alpha < 2$ is impossible since it is impossible even in the template matching model. Intuitively, if we cannot recover almost all shifts, the attained MSE should be much worse than in estimating a signal in AWGN, which means that the sample complexity should be much higher for $\alpha < 2$. Our bounds in Section 5 formalize this intuition.

To illustrate the phase transition for template matching, we conducted a "genie-aided" experiment, presented in Figure 3. In this experiment, we use the true X (the "genie") in order to estimate the shifts by $\hat{\ell}_i = \arg\max_{\ell \in \{0,\ldots,L-1\}} \langle R_\ell X, Y_i \rangle$. Then, we estimate the signal by aligning the measurements and averaging $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} R_{-\hat{\ell}_i} Y_i$. For large values of α, the recovery error converges to the error of estimating a signal in AWGN. For smaller α values, and in particular $\alpha < 2$, the recovery error rapidly increases.

Tighter lower bound for the low SNR regime. Theorem 2.1 shows that for all $\alpha \leq 2$ and fixed $\varepsilon < 1$ the shifts make a difference: the sample complexity with unknown shifts (i.e., the MRA problem) is $\omega\left(\sigma^2(\alpha)\log(1/\varepsilon)\right)$, and is therefore substantially greater than the sample complexity when the shifts are known. For $\alpha < 1$, we were able to prove a much stronger lower

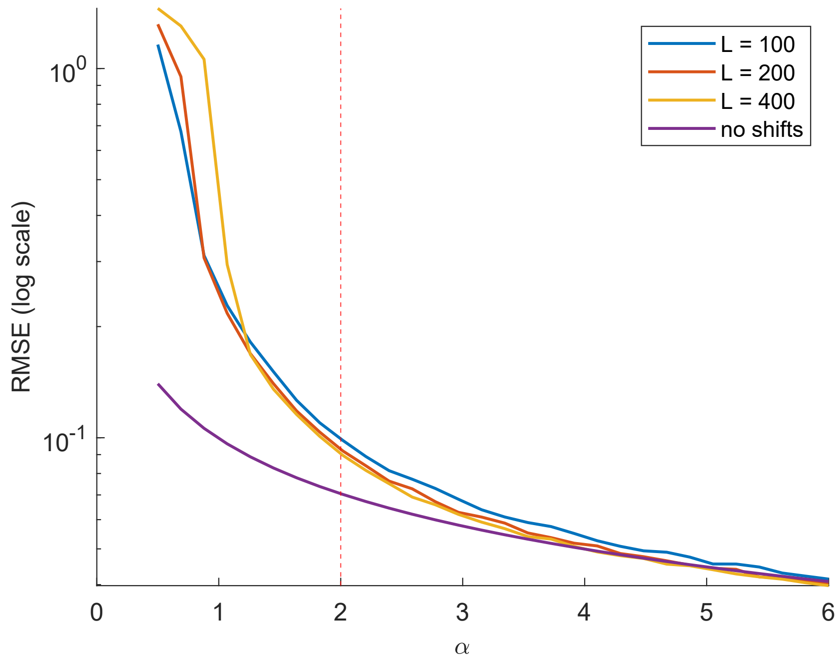

Figure 2. *The RMSE of EM (averaged over 100 trials) as a function of α for different values of L. The number of measurements was set to be $n(L) = 100L/\log(L)$. An example of a single measurement appears in Figure 1. For large values of α, the error reduces to the error of estimating a signal in AWGN, $\sqrt{\frac{\sigma^2}{\sigma^2+n}} = \frac{1}{\sqrt{1+100\alpha}}$, suggesting that EM performs as if the shifts were known. For small values of α, and in particular $\alpha < 2$, the error rapidly increases.*

bound on the sample complexity.

Theorem 2.2. *For any $0 < \alpha < 1$, and $0 < \varepsilon < 1$,*

$$n^*_{MRA}(L, \alpha, \varepsilon) = \Omega\left(L^{2-\alpha}\log(1/\varepsilon)\right) . \tag{2.1}$$

Theorems 2.1 and 2.2 are proved in Section 5.

The sample complexity of the projected MRA model. Recall that MRA serves as a toy model of the cryo-EM reconstruction problem. An additional complication arising in cryo-EM is a fixed tomographic projection, a line integral, also known as the X-ray transform. To account for this effect, we extend our basic model (1.1) to the *projected multi-reference alignment problem* (PMRA) model:[2]

$$Y_i = \pi_S R_{\ell_i} X + \sigma Z_i. \tag{2.2}$$

Here, $\pi_S : \mathbb{R}^L \to \mathbb{R}^{L'}$ is matrix projecting a vector in $\mathbb{R}^L$ to $\mathbb{R}^{L'}$ by keeping only the coordinates that belong to a subset $S \subset [L]$ of size $L' \leq L$ and discarding the rest, and $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ are

[2]We mention that other projected MRA models were studied in [6, 12].

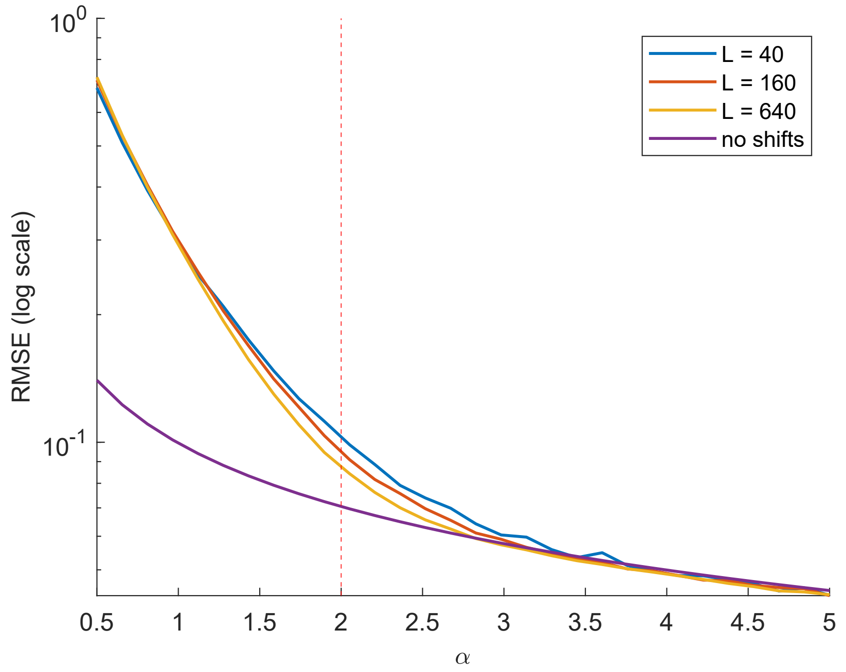

Figure 3. *A "genie-aided" experiment: the true X is used to estimate the shifts $\hat{\ell}_1, \ldots, \hat{\ell}_n$, as in the template matching problem, and then the signal is estimated by aligning all measurements and averaging $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} R_{-\hat{\ell}_i} Y_i$. The figure presents the RMSE (averaged over 50 trials) as a function of α for different values of L. The number of measurements was set to be $n(L) = 100L/\log(L)$. For large values of α, the error reduces to the error of estimating a signal in AWGN (i.e., when the shifts are known) $\sqrt{\frac{\sigma^2}{\sigma^2+n}} = \frac{1}{\sqrt{1+100\alpha}}$. For small values of α, and in particular $\alpha < 2$, the template matching error quickly increases.*

L'-dimensional i.i.d. Gaussian vectors. We assume that S is fixed and known to the estimator. As in MRA without the projection, the goal is to reconstruct X up to a circular shift, that is, produce an estimate $\widehat{X}$ such that $\mathbb{E}\rho(X, \widehat{X})$ is as small as possible.

We study the PMRA problem in an asymptotic setting where $L, L', \sigma^2 \to \infty$ simultaneously. It makes sense to adopt a slightly different scaling for the noise in PMRA, as

$$\sigma^2 = \sigma^2_{\mathrm{PMRA}}(\alpha) = \frac{L'}{\alpha \log(L)}. \tag{2.3}$$

The reason for this particular scaling will be made clear from the analysis: the numerator is the total signal energy available in a single measurement, $\mathbb{E}\|\pi_S R_{\ell_i} X\|^2 = L'$; the $\log(L)$ factor is log the size of the group of shifts. In Section 7 we provide some remarks as to how to extend our results to other groups. Similarly to our notation for the MRA model, we denote the smallest attainable MSE in the PMRA model as $\mathrm{MSE}^*_{\mathrm{PMRA}}(L, \alpha, n)$, and the sample complexity as $n^*_{\mathrm{PMRA}}(L, \alpha, \varepsilon)$.

Theorem 2.3. *Suppose that $\sigma^2_{\mathrm{PMRA}}(\alpha)$ is scaled as in (2.3), and $L, L' \to \infty$, so that $L' \leq L$ and $L' = \omega(\log(L))$ (that is, L grows strictly less than exponentially fast in L'). The sample complexity of the PMRA model (2.2) obeys the following lower bounds:*

1. *For any* $\alpha > 2$ *and* $0 < \varepsilon < 1$ *we have that*

$$n^*_{PMRA}(L,\alpha,\varepsilon) \geq \frac{L}{L'}\left(\frac{1}{\varepsilon}-1\right)\sigma^2_{\mathrm{PMRA}}(\alpha)(1+o(1)). \tag{2.4}$$

2. *For any* $\alpha \leq 2$ *and* $0 < \varepsilon < 1$ *we have that*

$$n^*_{PMRA}(L,\alpha,\varepsilon) = \omega\left(\frac{L}{L'}\sigma^2_{\mathrm{PMRA}}(\alpha)\log(1/\varepsilon)\right). \tag{2.5}$$

The proof of the theorem relies heavily on the proof of Theorem 2.1. Due to space constraints, a proof sketch is relegated to the SI appendix, see Section SM6. We conjecture that at high SNR ($\alpha > 2$), the lower bound given in Theorem 2.3 is in fact tight at very low MSE (formally $\varepsilon \to 0$, as in Theorem 2.1).

Extension to other signal priors and group actions. In section 7 we describe briefly how one could modify our proofs to account for other i.i.d. signal priors (besides Gaussian) and finite group actions.

3. Prior art.

The multi-reference alignment problem was introduced by [7], and fully formulated in [8]. The general MRA model reads

$$Y_i = T_i(g_i \circ X) + \sigma Z_i, \qquad i = 1,\dots,n, \tag{3.1}$$

where g_i is a random element of a compact group G (drawn from a possibly unknown distribution over G) acting on a vector space $X \in \mathbb{X}$, and T_i, $i = 1,\dots,n$, are known linear operators. If $T_i = I$ for all i, g_i are drawn uniformly from the group of cyclic shifts $\mathbb{Z}_L$, and $X \sim \mathcal{N}(0,I)$, then (3.1) reduces to the MRA model (1.1). This model can be thought of as a special case of a Gaussian mixture model, where all centers are connected through a group action (i.e., a cyclic shift). If $T_i = \pi_S$ for all i, we get the projected MRA model (2.2). In cryo-EM—the main motivation of this work—G is the group of 3-D rotations $SO(3)$, $\mathbb{X}$ is the space of 3-D "band-limited" functions (that is, functions that can be expanded by finitely many basis functions), and T_i encodes the (fixed) tomographic projection, as well as other linear effects, such as the microscope's point spread function (which varies across images) and sampling [10, 41].

The sample complexity of the MRA model (1.1), in the minimax sense, was first studied in [9, 32]. The focus of these works, as well as the rest of the works mentioned in this section, is on the regime where the noise level σ and number of measurements n diverge, while the dimension of each measurement L is fixed, implying SNR $\to 0$. These results were extended to the general MRA model (3.1) by [6] and [3] (the latter generalizes the framework proposed in [1]). These papers constitute an intimate connection between the MRA model and the method of moments—a classical estimation technique. Let $\bar{d}$ be the lowest order moment that distinguishes two different signals (signals that are not in the same orbit) given a specific MRA model (namely, fixed $T_i, \mathbb{X}$, and a distribution over G). Then, unless $n \cdot \mathrm{SNR}^{\bar{d}} \to \infty$, the MSE is bounded from below. More informally, the moments determine the optimal (minimax) estimation rate of the problem. For example, for the MRA model (1.1) it is known that the third moment determines a generic signal uniquely (in this work we only consider normal i.i.d. signals that fall into this category), i.e., $\bar{d} = 3$, and thus $n \cdot \mathrm{SNR}^3 \gg 1$ is a necessary condition.

Remarkably, this phenomenon was observed empirically in context of cryo-EM early on by Sigworth [39].

In this work, we propose an alternative explanation for the statistical difficulty of MRA at low SNR, in a setting where the signal X is "generic" (specifically, $X \sim \mathcal{N}(0, I)$) and the dimension is very large. The separation between the two SNR regimes we identify is *not* given in terms of moments; instead, it is characterized in terms of a very natural estimation-theoretic question: is it possible, in an information-theoretic sense, to consistently recover the unknown shifts (nuisance parameters) themselves? As we scale $\mathrm{SNR} = \alpha \log L$, the threshold $\alpha = 2$, separating the high and low SNR regimes, is exactly the threshold for the shift recovery problem. Note that in this high-dimensional setting, we find that the low SNR regime in fact extends beyond the case $\mathrm{SNR} \to 0$ to unbounded values of SNR (provided that it grows slowly enough with L)—this is in contrast to previous works that study MRA in fixed dimension.

From the algorithmic perspective, two main computational frameworks were applied to MRA problems. The first approach is based on expectation-maximization (EM)—a popular heuristic to maximize the posterior distribution [20]. EM is the most popular and successful methodology to elucidate high-resolution 3-D structures using cryo-EM [10, 37], and it was successfully applied to a variety of MRA setups [1, 11, 12, 16, 31]. A recent work [22] studies the likelihood landscape for the general MRA model (3.1), where G is a discrete group and $T_i = I$. The latter paper shows that when the dimension is fixed and the SNR is sufficiently high, the log likelihood has certain favorable features from an optimization perspective; their results give a compelling argument for why EM seems to give good performance for MRA in high SNR. In [17], it is shown that usually maximum likelihood achieves the parametric rate $\rho(X, \widehat{X}_{\mathrm{MLE}}) \sim 1/n$, although in some cases the rate can be $\sim 1/\sqrt{n}$.

The second algorithmic framework is based on the method of moments. This approach has an appealing property: it requires only one pass over the measurements, and thus its computational load is relatively low, unless L is large [1, 11, 16, 31, 32, 35]. In addition, as mentioned, it achieves the optimal estimation rate when L is fixed and $\mathrm{SNR} \to 0$. Consequently, a variety of moment-based algorithms were proposed. For example, the authors of [32] suggest estimating the third-order tensor moment of the signal $T^{(3)} = L^{-1} \sum_{\ell=0}^{L-1} (R_\ell X)^{\otimes 3}$, from which X can be recovered by Jenrich's method [24, 29]. Using the robustness analysis of [23], they were able to show that $n = O\left(\varepsilon^{-1}\sigma^6 \mathrm{poly}(L)\right)$ samples suffice to achieve $\rho(X, \widehat{X}) \le \varepsilon$ with constant probability. This bound depends polynomially on both the dimensional and on the inverse smallest DFT coefficient of X; when $X \sim \mathcal{N}(0, I)$, one can verify that typically all the DFT coefficients of X are greater than $\Omega(L^{-1/2})$. The $\mathrm{poly}(L)$ dependence is not computed explicitly, but to the best of our understanding, the analysis of [23] provides a significantly worse dimensional scaling than the $\Omega(L^2)$ in our lower bound (as $\alpha \to 0$). Another work [11] studies recovery by bispectrum inversion, which is equivalent to the third-order moment if the distribution of shifts is uniform. They argue that when L is fixed, the sample complexity should scale like $O(\sigma^6)$, hiding an implicit dependence on L. The method of moments was also applied to cryo-EM and related technologies, see for example [21, 26, 30, 38], as well as to additional MRA setups [2, 5, 25].

A recent work [27] establishes an enticing connection between likelihood-based techniques and the method of moments for the general MRA model (3.1) for fixed L, $\mathrm{SNR} \to 0$, and $T_i = I$. Specifically, it was shown that likelihood optimization in the low SNR regime reduces to a sequence of moment matching problems. In addition, the method of moments is also closely-related to invariant theory and thus tools from the latter field can be applied to analyze MRA models; see in particular [6].

4. Phase transition of template matching.

Suppose that the shifts R_{ℓ_i} are all known. In this scenario, estimating the signal is easy: one needs to align each observation $R_{\ell_i}^{-1} y_i$ and average out the noise. Therefore, if possible, it makes sense to try and estimate the shifts. In this section, we study the problem of estimating a shift when the signal is assumed to be known (which is not the case in MRA); we refer to this problem as *template matching*. Specifically, suppose that one has access to a signal, a "template" $X \in \mathbb{R}^L$, and observes a single sample $Y = R_\ell X + \sigma Z$, where $X \sim \mathcal{N}(0, I)$, $R_\ell \sim \text{Uniform}(\{0, \dots, L-1\})$ is a random uniform shift, $Z \sim \mathcal{N}(0, I)$, and R_ℓ, Z and X are mutually independent. The goal, then, is to recover R_ℓ from X and Y.[3]

While the template matching problem seems to be significantly easier than the MRA problem, we show a surprising phenomenon: in high dimensions, template matching and MRA share the exact same phase transition point. In particular, it turns out that in high dimensions, under our parameterization $\sigma^2(\alpha)$, which amounts to $L/\sigma^2 = \alpha \log(L)$, the template matching problem displays a *sharp recoverability threshold*. That is: (i) whenever $\alpha > 2$, the random shift can be recovered with error probability $p_e \to 0$ as $L \to \infty$; (ii) whenever $\alpha < 2$, the shift cannot be consistently recovered, and in fact for any estimator, $p_e \to 1$.

Observe that the optimal estimator (in the sense of maximum a posteriori probability) for R_ℓ is given by:

$$\widehat{R}_{\text{MAP}} = \underset{\ell'}{\text{argmin}}\, \|X - R_{\ell'}^{-1} Y\|^2 = \underset{\ell'}{\text{argmax}}\, \frac{\langle X, R_{\ell'}^{-1} Y\rangle}{\|X\|^2} \,. \tag{4.1}$$

Denote its error probability by

$$p_e = \Pr\left(R_\ell \neq \widehat{R}_{\text{MAP}}\right). \tag{4.2}$$

We start by establishing that with overwhelming probability, the template X is "incoherent", in the sense that the correlations $\langle X, R_{\ell'} X\rangle / \|X\|^2$ are very small, unless $\ell' = 0$. The lemma is proved in Appendix SM3.

Lemma 4.1. *For $\kappa > 0$, let $\mathcal{A}(\kappa)$ be the event that*

$$\left|L^{-1}\|X\|^2 - 1\right| < \kappa \quad \textit{and} \quad \max_{\ell' \neq 0} L^{-1} \left|\langle X, R_{\ell'} X\rangle\right| \leq \kappa,$$

and let $\overline{\mathcal{A}(\kappa)}$ be its complement. Then,

$$\Pr(\overline{\mathcal{A}(\kappa)}) \leq 2L \exp\left(-cL \min(\kappa, \kappa^2)\right),$$

for a universal constant $c > 0$. In particular, one can choose a sequence $\kappa = \kappa_L$ such that $\kappa \to 0$ sufficiently slowly, for example, $\kappa = CL^{-1/2}\log(L)$ for $C > 0$ large enough, so that $\Pr(\mathcal{A}_L(\kappa_L)) = 1 - o(1)$.

[3] A more general setting, where X is not necessarily Gaussian, and $R_\ell X$ goes through some general channel, not necessarily Gaussian, was studied by Wang, Hu, and Shayevitz [45], but under different asymptotics.

Let

$$\Theta_{\ell'} = \frac{\langle X, R_{\ell'}^{-1} Y\rangle}{\|X\|^2} = \frac{\langle X, R_{\ell-\ell'} X\rangle}{\|X\|^2} + \frac{\sigma \langle X, R_{\ell'}^{-1} Z\rangle}{\|X\|^2}, \tag{4.3}$$

and

$$W_{\ell'} = \|X\|^{-1} \langle X, R_{\ell'}^{-1} Z\rangle. \tag{4.4}$$

Recalling that $\widehat{R}_{\text{MAP}} = \text{argmax}_{\ell'} \Theta_{\ell'}$, and plugging $\sigma^2 = (\alpha \log(L))^{-1} L$, Lemma 4.1 implies that with high probability,

$$\Theta_{\ell'} = \begin{cases} 1 + (1 + o(1)) \frac{1}{\sqrt{\alpha \log(L)}} \cdot W_\ell & \text{if } \ell' = \ell, \\ o(1) + (1 + o(1)) \frac{1}{\sqrt{\alpha \log(L)}} \cdot W_{\ell'} & \text{if } \ell' \neq \ell. \end{cases} \tag{4.5}$$

Notice that for every ℓ', $W_{\ell'} \sim \mathcal{N}(0,1)$, being the projection of $R_{\ell'}^{-1} Z \sim \mathcal{N}(0, I)$ onto a unit vector $X/\|X\|$. This clearly implies that $\Theta_\ell \xrightarrow{p} 1$ as $L \to \infty$. Thus, to analyze the error of the MAP estimator, it simply remains to understand the behavior of $\max_{\ell'} W_{\ell'}$. To this end, we recall the following three results. We start with a well-known fact about the maximum of i.i.d. standard Gaussians:

Lemma 4.2. *Let $Z_1, \ldots, Z_L$ be i.i.d $\mathcal{N}(0,1)$ random variables. Then, as $L \to \infty$,*

$$\mathbb{E}\left[\max_\ell Z_\ell\right] / \sqrt{2 \log(L)} \to 1.$$

The upper bound $\mathbb{E}\left[\max_\ell Z_l\right] \leq \sqrt{2\log(L)}$ is elementary, and holds even when $Z_1, \ldots, Z_L$ are not independent. The proof follows from $\mathbb{E} \max_\ell Z_\ell \leq \beta^{-1} \log \mathbb{E} \max_\ell e^{\beta Z_\ell} \leq \beta^{-1} \log \mathbb{E} \sum_{\ell=1}^L e^{\beta Z_\ell} = \beta/2 + \beta^{-1} \log(L)$, which holds for all $\beta > 0$; now take $\beta = \sqrt{2 \log(L)}$. The proof of the matching lower bound, on the other hand, is more involved and follows from results in extreme value theory, see, for instance, Example 1.1.7 in [19]. We also use the following "quantitative" version of the Sudakov-Fernique inequality:

Lemma 4.3 (Theorem 2.2.5 in [4]). *Let $(X_1, \ldots, X_L)$ and $(Y_1, \ldots, Y_L)$ be Gaussian vectors so that $\mathbb{E}[X_i] = \mathbb{E}[Y_i]$ for all i. Set*

$$\gamma_{i,j}^X = \mathbb{E}(X_i - X_j)^2, \quad \gamma_{i,j}^Y = \mathbb{E}(Y_i - Y_j)^2,$$

and $\gamma = \max_{i,j} |\gamma_{i,j}^X - \gamma_{i,j}^Y|$. Then

$$\left| \mathbb{E}\left[\max_i X_i\right] - \mathbb{E}\left[\max_i Y_i\right] \right| \leq \sqrt{2\gamma \log(L)}.$$

To get concentration around the mean, we use (a simple case of) the Borell-TIS inequality:

Lemma 4.4. *Let $(X_1, \ldots, X_L)$ be a Gaussian vector, and set $\sigma^2 = \max_i \mathbb{E}[X_i^2]$. Then*

$$\Pr\left(\left| \max_i X_i - \mathbb{E}\left[\max_i X_i\right] \right| \geq t \right) \leq 2e^{-t^2/2\sigma^2}.$$

See, e.g., [4, Theorem 2.1.1] (there only a one sided bound is stated; the other side follows the same way). The following is now an immediate corollary of Lemmas 4.1, 4.2,4.3 and 4.4:

Theorem 4.5 (Sharp threshold for template matching). *If $\alpha > 2$, then $p_e \to 0$ as $L \to \infty$. Conversely, if $\alpha < 2$, then $p_e \to 1$.*

Proof. We start by estimating $\mathbb{E}\max_{\ell'} W_{\ell'}$. Choose $\kappa = o(1)$ such that the event $\mathcal{A}(\kappa)$ of Lemma 4.1 holds with probability $1 - o(1)$. Conditioned on X, $\{W_{\ell'}\}_{\ell'=0,\ldots,L-1}$ is a centered Gaussian vector, with covariance

$$C_{i,j}(X) = \mathbb{E}[W_i W_j \mid X] = \|X\|^{-2}\langle R_i X, R_j X\rangle,$$

whereby under $\mathcal{A}$, $|C_{i,j}(X) - \delta_{i,j}| = o(1)$.

Let $(\tilde{W}_1, \ldots, \tilde{W}_{L-1})$ be i.i.d $\mathcal{N}(0,1)$ random variables. By Lemmas 4.2 and 4.3, conditioned on X and under $\mathcal{A}$,

$$\mathbb{E}[\max_{\ell'} W_{\ell'} \mid X, \mathcal{A}] = \mathbb{E}[\max_{\ell'} \tilde{W}_{\ell'}] + o(\sqrt{\log(L)}) = \sqrt{(2+o(1))\log(L)}.$$

Lemma 4.4 gives us a uniform (in X) concentration inequality, conditioned on X and under $\mathcal{A}$,

$$\Pr\left(\left|\max_{\ell'} W_{\ell'} - \sqrt{2\log(L)}\right| \geq \sqrt{\varepsilon\log(L)} \,\middle|\, X, \mathcal{A}\right) \leq 2L^{-(\varepsilon+o(1))/2},$$

so that

$$\Pr\left(\left|\max_{\ell'} W_{\ell'} - \sqrt{2\log(L)}\right| \geq \sqrt{\varepsilon\log(L)}\right) \leq 2L^{-(\varepsilon+o(1))/2} + \Pr\left(\overline{\mathcal{A}}\right) = o_\varepsilon(1).$$

Thus, we have shown that $\max_{\ell'} W_{\ell'}/\sqrt{2\log(L)} \xrightarrow{p} 1$. Using equation (4.5), we deduce that $\Theta_\ell \xrightarrow{p} 1$ whereas $\max_{\ell'\neq\ell} \Theta_{\ell'} \xrightarrow{p} \sqrt{2/\alpha}$. Since $\widehat{R}_{\mathrm{MAP}} = \operatorname{argmax}_{\ell'} \Theta_{\ell'}$, we conclude that $p_e \to 0$ when $\alpha > 2$ and $p_e \to 1$ when $\alpha < 2$. ∎

A remark on the relation between template matching and synchronization.. In the MRA model, one does not have access to the true template and thus needs to estimate the relative shifts based solely on the data; this problem is referred to as *synchronization.*

For simplicity, let us assume we are given two measurements $Y_1 = X + \sigma Z_1$ and $Y_2 = R_\ell X + \sigma Z_2$, and would like to estimate R_ℓ (recall that X is unknown). The optimal (MAP) estimator is $\widehat{R}_{\mathrm{syn}} = \operatorname{argmax}_{\ell'} \Pr(R_{\ell'}|Y_1, Y_2)$. It is straightforward to show that

$$\begin{aligned}\widehat{R}_{\mathrm{syn}} &= \operatorname*{argmax}_{\ell'}\langle Y_1, R_{\ell'}^{-1} Y_2\rangle = \operatorname*{argmax}_{\ell'}\langle (X+\sigma Z_1), R_{\ell'}^{-1}(R_\ell X + \sigma Z_2)\rangle \\ &= \operatorname*{argmax}_{\ell'}\left\{\langle X, R_{\ell-\ell'}X\rangle + \sigma\langle X, R_{\ell'}^{-1} Z_2\rangle + \sigma\langle X, R_{\ell-\ell'}^{-1} Z_1\rangle + \sigma^2\langle Z_1, R_{\ell'}^{-1} Z_2\rangle\right\}.\end{aligned}$$

In order for this to consistently return the true relative shift R_ℓ, one needs to ensure that the "noise" term,

$$\sigma\langle X, R_{\ell'}^{-1} Z_2\rangle + \sigma\langle X, R_{\ell-\ell'}^{-1} Z_1\rangle + \sigma^2\langle Z_1, R_{\ell'}^{-1} Z_2\rangle$$

is small compared to $\|X\|^2 \sim L$. The "typical" size of the first two terms is $\sigma\langle X, R_{\ell'}^{-1} Z_2\rangle + \sigma\langle X, R_{\ell-\ell'}^{-1} Z_1\rangle \sim \sigma\sqrt{L}$, whereas the third is $\sigma^2\langle Z_1, R_{\ell'}^{-1} Z_2\rangle \sim \sigma^2\sqrt{L}$, and is therefore the dominant one for large σ. Thus, to succeed with non-vanishing probability, we need that $\sigma^2\sqrt{L} \lessapprox L$, that is, $\sigma^2 \lessapprox \sqrt{L}$. In the regime we are interested in, the noise level is $\sigma^2 \sim L/\log(L)$, and this turns out to be far too large.

We mention in passing that if many measurements are available, one can leverage the redundancy in the data to recover the true relative shifts in challenging environments; see for example [15, 33, 36, 40, 42].

5. Sample complexity lower bounds.

5.1. The information-theoretic method for estimation lower bounds. We employ a standard information-theoretic method of obtaining estimation error lower bounds, via rate-distortion theory (see e.g. [34]). We refer the reader to SI Appendix SM1 for a basic review of the information-theoretic definitions and facts we use in this section. Let $\widehat{X}$ be an estimator of X from the measurements $Y^n = (Y_1, \ldots, Y_n)$, which achieves expected error ("distortion")

$$\mathbb{E}\rho(X, \widehat{X}) = L^{-1}\mathbb{E} \min_{\ell=0,\ldots,L-1} \|X - R_\ell^{-1}\widehat{X}\|^2 \le \varepsilon. \tag{5.1}$$

Since the estimator depends only on the measurements, and not on X, the triplet $X - Y^n - \widehat{X}$ constitutes a Markov chain. Hence, by the data processing inequality (Proposition SM1.3 item 3) we have that $I(X; \widehat{X}) \le I(X; Y^n)$. We lower-bound $I(X; \widehat{X})$ by the *rate distortion function* (RDF) $R(\cdot)$ associated with the source $X \sim \mathcal{N}(0, I)$, and distortion measure $\rho(\cdot, \cdot)$:

$$R(\varepsilon) = \min_{P_{W|X}:\mathbb{E}\rho(X,W)\le\varepsilon} I(X; W).$$

The minimization here is done over conditional distributions $P_{W|X}$, or equivalently, over joint distributions $P_{X,W}$ whose X-marginal is P_X—in our case $\mathcal{N}(0, I)$—obeying the average distortion constraint $\mathbb{E}\rho(X, W) \le \varepsilon$. Since the conditional distribution $P_{\widehat{X}|X}$ is, by definition, feasible for this minimization problem, we have $R(\varepsilon) \le I(X; \widehat{X})$. Combining this with the upper bound $I(X; \hat{X}) \le I(X; Y^n)$, we get

$$R(\varepsilon) \le I(X; Y^n), \tag{5.2}$$

and we shall next derive a lower bound for $R(\varepsilon)$ in terms of ε.

5.2. A lower bound on the rate-distortion function. We start by obtaining a lower bound on the RDF. While the RDF problem for a Gaussian source under MSE distortion measure is classical, the MSE up to the best alignment (the distortion measure we consider) is somewhat non-standard. Obtaining a precise expression for the true RDF seems difficult, but a simple lower bound can be obtained as follows.

Proposition 5.1. *For an L dimensional i.i.d. Gaussian vector $X \sim \mathcal{N}(0, I)$, and distortion measure $\rho(\cdot, \cdot)$ as defined in* (1.2)*, the rate distortion function satisfies*

$$R(\varepsilon) \ge \frac{L}{2} \log\left(\frac{1}{\varepsilon}\right) - \log(L).$$

Proof. By definition of the rate distortion function, to establish the claim we need to show that for any conditional distribution ("test-channel") $P_{W|X}$ that satisfies the constraint $\mathbb{E}\rho(X,W) \le \varepsilon$, where $\rho(X,W) = L^{-1}\min_{\ell=0,\dots L-1}\|X - R_\ell^{-1}W\|^2$, it holds that $I(X;W) \ge \frac{L}{2}\log\left(\frac{1}{\varepsilon}\right) - \log(L)$. To that end, let $R = R(X,W) = \operatorname{argmin}_{\ell' \in [0,\dots,L-1]}\|X - R_{\ell'}W\|$ be the difference minimizing shift. By the chain law of MI (Proposition SM1.3 item 2),

$$I(X;W) = I(X;W,R) - I(X;R|W) \ge I(X;W,R) - \log(L), \tag{5.3}$$

where we used $I(X;R|W) \le H(R|W) \le \log(L)$; the former follows from the definition of MI and non-negativity of entropy (Proposition SM1.1 item 1), and the latter follows from Proposition SM1.1 item 2 as the random variable R can take at most L values. Recall that $L^{-1}\mathbb{E}\|X - RW\|^2 \le \varepsilon$ by definition of R. We therefore have that

$$I(X;RW) \ge \min_{P_{W'|X}: L^{-1}\mathbb{E}\|X-W'\|^2 \le \varepsilon} I(X;W') = \frac{L}{2}\log\left(\frac{1}{\varepsilon}\right),$$

where in the second equality we have used the well-known expression for the quadratic Gaussian rate distortion function (Proposition SM1.4). Thus, using the data processing inequality (Proposition SM1.3 item 3), we have

$$I(X;W,R) \ge I(X;RW) \ge \frac{L}{2}\log\left(\frac{1}{\varepsilon}\right).$$

Substituting this into (5.3) establishes the claim. ∎

Combining Proposition 5.1 with equation (5.2), we get

$$I(X;Y^n) \ge R(\varepsilon) \ge \frac{L}{2}\log\left(\frac{1}{\varepsilon}\right) - \log(L).$$

Setting $\varepsilon = \mathbb{E}\rho(X,\widehat{X})$, we have obtained the following bound:

Corollary 5.2. *Suppose that $X \sim \mathcal{N}(0,I)$ is an L dimensional i.i.d. Gaussian vector, $\widehat{X}$ is any estimator of X from $Y_1,\dots,Y_n$, and $\rho(\cdot,\cdot)$ is as defined in (1.2). Then*

$$\mathbb{E}\rho(X,\widehat{X}) \ge \exp\left(-\frac{2I(X,Y^n) + 2\log(L)}{L}\right) = \exp\left(-2L^{-1}\cdot I(X,Y^n) + o(1)\right).$$

Equivalently,

$$\mathrm{MSE}^*_{MRA}(L,\alpha,n) \ge \exp\left(-\frac{2I(X,Y^n) + 2\log(L)}{L}\right) = \exp\left(-2L^{-1}\cdot I(X,Y^n) + o(1)\right).$$

Corollary 5.2 tells us that an upper bound on the MI $I(X;Y^n)$ would give us a lower bound on the expected error of any estimator of X from $Y^n = (Y_1,\dots,Y_n)$. We devote the next section to deriving such upper bounds.

5.3. Upper bounds on the mutual information. We start with the rather trivial observation that the MI between the signal X and the measurements Y^n is smaller than the MI in a problem where there are no random shifts, which is equal to $\frac{L}{2}\log(1+n\sigma^{-2})$. The next lemma formalizes this intuition and quantifies the MI difference between the two problems.

Lemma 5.3. *The mutual information between the signal X and measurements $Y_1,\dots,Y_n$ is*

$$I(X;Y^n) = \frac{L}{2}\log(1+n\sigma^{-2}) - I(R^n;X|Y^n), \tag{5.4}$$

where $R^n = (R_{\ell_1},\dots,R_{\ell_n})$. In particular, $I(X;Y^n) \le \frac{L}{2}\log(1+n\sigma^{-2})$.

Proof. Let $\tilde{Y}_i = R_{\ell_i}^{-1}Y_i = X + \sigma R_{\ell_i}^{-1}Z_i$. We may write

$$\begin{aligned} I(X;Y^n) &= I(X;Y^n,R^n) - I(X;R^n|Y^n) \\ &= I(X;\tilde{Y}^n,R^n) - I(X;R^n|Y^n) \\ &= I(X;\tilde{Y}^n) + I(X;R^n|\tilde{Y}^n) - I(X;R^n|Y^n), \end{aligned}$$

where the first and third equalities follow by the chain rule for MI (Proposition SM1.3 item 2), and the second follows from Proposition SM1.3 item 4, and the fact that the mapping $(Y^n,R^n)\mapsto(\tilde{Y}^n,R^n)$ is invertible. By the fact that the Gaussian distribution is rotation invariant, and in particular $R_{\ell_i}^{-1}Z \sim \mathcal{N}(0,I)$, we have that R^n is statistically independent of $(X,\tilde{Y}^n)$, and consequently

$$I(X;R^n|\tilde{Y}^n) = H(R^n|\tilde{Y}^n) - H(R^n|\tilde{Y}^n,X) = H(R^n) - H(R^n) = 0,$$

where the first equality follows by definition of conditional mutual information and the second by Proposition SM1.3.5. It remains to compute $I(X;\tilde{Y}^n)$. To this end, note that conditioned on $X=x$, the measurements $\tilde{Y}_1,\dots,\tilde{Y}_n$ are simply i.i.d. Gaussian measurements $Y_i \sim \mathcal{N}(x,\sigma^2 I)$. It is well-known that in this case, the sample mean $\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i = X$ is a sufficient statistic of $\tilde{Y}^n$ for X. Conditioned on $X=x$, the sample mean has distribution $\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i \sim \mathcal{N}(x,\sigma^2/n\cdot I)$, therefore,

$$I(X;\tilde{Y}^n) = I\left(X;\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i\right) = I\left(X;X+\mathcal{N}(0,\sigma^2/n\cdot I)\right) = \frac{L}{2}\log(1+n\sigma^{-2}), \tag{5.5}$$

where the last equality follows from Proposition SM1.3 item 6. ∎

Combining Corollary 5.2 and Lemma 5.3, we obtain the following lower bound, that essentially says the MSE in the MRA model is no better than in estimating a signal in AWGN.

Corollary 5.4. *The smallest attainable MSE in the MRA model satisfies*

$$\mathrm{MSE}^*_{MRA}(L,\sigma^2,n) \ge \frac{L^{-\frac{2}{L}}}{1+n\sigma^{-2}} = \frac{1}{1+n\sigma^{-2}}(1+o(1)),$$

and the sample complexity satisfies

$$n^*_{MRA}(L,\sigma^2,\varepsilon) \ge \left\lceil \left(\frac{L^{-\frac{2}{L}}}{\varepsilon} - 1\right)\sigma^2 \right\rceil = n^*_{AWGN}(L,\sigma^2,\varepsilon)(1+o(1)).$$

Lemma 5.3 tells us that the gap between $I(X;Y^n)$ and the MI in estimating a signal in AWGN, without the shifts, $\frac{L}{2}\log(1+n\sigma^{-2})$, is $I(X;R^n|Y^n)$. This quantity is intimately related to a multi-sample version of the template matching problem, as was considered in Section 4. This connection will be exploited later on, when we derive an upper bound on the single sample MI $I(X;Y_i)$.

Information combining. Observe that the measurements $Y_1,\dots,Y_n$ are mutually independent and identically distributed conditioned on X; that is, the samples are obtained by passing the same signal X independently through a memoryless channel. By Proposition SM1.3 item 5, this implies that

$$I(X;Y^n) \le \sum_{i=1}^{n} I(X;Y_i) = nI(X;Y), \tag{5.6}$$

where $Y = R_\ell X + \sigma Z$ is a single measurement in the MRA model. Substituting (5.6) into Corollary 5.2, yields the following.

Proposition 5.5. *The smallest attainable MSE in the MRA model satisfies*

$$\mathrm{MSE}^*_{MRA}(L,\sigma^2,n) \ge L^{-\frac{2}{L}}\exp\left(-n\frac{2}{L}I(X;Y)\right) = \exp\left(-n\frac{2}{L}I(X;Y)\right)(1+o(1)),$$

and the sample complexity satisfies

$$n^*_{MRA}(L,\sigma^2,\varepsilon) \ge \frac{L}{2}\cdot\frac{\log\left(\frac{1}{\varepsilon}\right)-\frac{2\log(L)}{L}}{I(X;Y)} = \log\left(\frac{1}{\varepsilon}\right)\cdot\frac{L}{2I(X;Y)}(1+o(1)),$$

where $Y = R_\ell X + \sigma Z$ is a single measurement in the MRA model.

It is important to emphasize at this point that the bound in (5.6) becomes very loose for n sufficiently large. Indeed, Lemma 5.3 implies that $I(X;Y^n)$ should scale at best logarithmically, rather than linearly, with n. Consequently, the lower bound on $\mathrm{MSE}^*_{\mathrm{MRA}}(L,\sigma^2,n)$ in Proposition 5.5 decreases exponentially fast with n, whereas we know from Corollary 5.4 that it cannot decrease faster than the parametric rate of $1/n$ as in estimating a signal in AWGN. Despite its grossly wrong dependence on n, the upper bound $I(X;Y^n) \le nI(X;Y)$ does suffice to say something non-trivial about the sample complexity of the problem. As seen from Proposition 5.5: in order for the estimation error to be strictly bounded away from one, one needs at least $\Omega(L\cdot I(X;Y)^{-1})$ samples. We will see that this rather "naïve" analysis is already enough to accurately separate between a "high SNR" and a "low SNR" regime, where the behavior of the MRA problem is qualitatively different. Intuitively, as the measurements $Y_1,\dots,Y_n$ are only dependent through the random variable X, if n is so small that it is impossible to learn much about X from Y^n, the dependence between $Y_1,\dots,Y_n$ must be weak. Thus, in that regime, ignoring this dependence and bounding $I(X;Y^n) \le nI(X;Y)$ is a rather accurate estimate.

The problem of obtaining a stronger bound on multi-sample MI $I(X;Y^n)$ in terms of the single-sample MI $I(X;Y)$ is an instance of a so-called *information combining* problem. Several problems of this type have been studied in the information theory literature, mostly dealing with binary channels [28, 43]. In our case, we believe this problem to be quite hard, at least in the low SNR regime, and thus we could not obtain a tighter bound. Deriving such bounds can yield stronger lower bounds on $\mathrm{MSE}^*_{\mathrm{MRA}}(L,\alpha,n)$ in the low-SNR regime ($\alpha < 2$) than the ones we obtain here using the simple bound $I(X;Y^n) \le nI(X;Y)$).

Roadmap. We will devote the rest of this section to deriving upper bounds on $I(X;Y)$. These bounds, together with Proposition 5.5, will immediately imply lower bounds on the MSE and the sample complexity. In particular, we will derive two bounds, using different methods, that will be effective in two SNR regimes.

- We estimate the mutual information using Jensen's inequality to facilitate the computation of several expectations. One could expect this method to give somewhat tight results when $I(X;Y)$ is very small, and indeed, we shall see that when $0 < \alpha < 1$, we obtain a bound $I(X;Y) = O(L^{\alpha-1})$, which tends to 0 as $L \to \infty$. For $\alpha \ge 1$, the obtained bound will turn out to be too loose.
- In Lemma 5.3 we have found that $I(X;X+\sigma Z) - I(X;Y) = I(X, R_\ell|Y)$. We lower bound this gap using a Fano-like inequality, which in the case $\alpha < 2$ amounts to "quantifying" how well R_ℓ can be estimated from X and Y, in a somewhat more precise sense than Theorem 4.5 (which tells us that in this case, the error is $p_e = 1-o(1)$). This will allow us to show that when $\alpha < 2$, $I(X;Y) = o(\log(L))$. We will not, however, be able to recover the estimate in the case of $0 < \alpha < 1$ using this method.

5.3.1. MI bound at very low SNR ($\alpha < 1$).

We first express $I(X;Y)$ in the following way:

Lemma 5.6. *Suppose that $X \sim \mathcal{N}(0,I)$, $Z \sim \mathcal{N}(0,I)$, and $R \sim \mathrm{Uniform}(\{R_0,\dots,R_{L-1}\})$ are mutually independent. Then,*

$$I(X;Y) = \frac{L}{2}\log(1+\sigma^{-2}) - L\sigma^{-2} + \mathbb{E}_{X,Z}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle X+\sigma Z, RX\rangle\right)\right].$$

Proof. Write $I(X;Y) = h(Y) - h(Y|X)$. Note that for any shift R_ℓ, $R_\ell X \sim \mathcal{N}(0,I)$ and therefore $Y \sim \mathcal{N}(0,(1+\sigma^2)I)$; this means that $Y = R_\ell X + \sigma Z$ is independent of R_ℓ. The differential entropy of Y is $h(Y) = h(\mathcal{N}(0,(1+\sigma^2)I) = \frac{L}{2}\log(2\pi e) + \frac{L}{2}\log(1+\sigma^2)$, by Proposition SM1.1 item 3.

Let us now write the conditional differential entropy explicitly. The conditional density of Y given X is $p_{Y|X}(y|x) = \mathbb{E}_R\left[(2\pi\sigma^2)^{-L/2}\exp\left(-\frac{1}{2\sigma^2}\|y-Rx\|^2\right)\right]$ for uniform R. The conditional entropy is then simply

$$\begin{aligned} h(Y|X) &= \mathbb{E}_{X,Y}\left[-\log p_{Y|X}(Y|X)\right] \\ &= \frac{L}{2}\log(2\pi\sigma^2) - \mathbb{E}_{X,Y}\left[\log \mathbb{E}_R \exp\left(-\frac{1}{2\sigma^2}\|Y-RX\|^2\right)\right] \\ &= \frac{L}{2}\log(2\pi\sigma^2) - \mathbb{E}_{X,Y}\left[\log \mathbb{E}_R \exp\left(-\frac{1}{2\sigma^2}\left(\|Y\|^2+\|X\|^2 - 2\langle Y, RX\rangle\right)\right)\right] \\ &= \frac{L}{2}\log(2\pi\sigma^2) + \frac{L+(1+\sigma^2)L}{2\sigma^2} - \mathbb{E}_{X,Y}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle Y, RX\rangle\right)\right]. \end{aligned}$$

It remains to compute the expectation with respect to the joint distribution of X and Y in the last term. Recall that we can write $Y = R'X + \sigma Z$ for $R' \sim \mathrm{Uniform}(\{R_0,\dots,R_{L-1}\})$

and $Z \sim \mathcal{N}(0, I)$, both independent of X. Alternatively, we could also write $Y = R'(X + \sigma Z)$, which defines the exact same joint distribution between X and Y, due to the orthogonal invariance of $Z \sim \mathcal{N}(0, I)$; this second form is slightly more convenient in what follows. Since R is uniformly distributed,

$$\mathbb{E}_{X,Z,R'}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle R'(X+\sigma Z), RX\rangle\right)\right] = \mathbb{E}_{X,Z,R'}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle (X+\sigma Z), (R')^{-1}RX\rangle\right)\right]$$
$$= \mathbb{E}_{X,Z}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle (X+\sigma Z), RX\rangle\right)\right],$$

that is, we can "drop" R'. The claimed formula now readily follows. ■

The following proposition is the main estimate of this section. The proof uses some properties of the spectrum of R_ℓ, stated and proved in Appendix SM4.

Proposition 5.7. *We have the following upper bound on the single sample MI:*

$$I(X;Y) \le \log\left(1 + L^{-1}e^{\sigma^{-2}L}\right) + O(\sigma^{-4}L).$$

In particular, if $\sigma^{-2}L = \alpha \log(L)$ for $0 < \alpha < 1$, then the MI asymptotically vanishes as $L \to \infty$ with $I(X;Y) \le L^{-1+\alpha}(1 + o(1))$.

Proof. By the concavity of the log function, we always have $\mathbb{E}_W \log(W) \le \log(\mathbb{E}W)$. Thus,

$$\mathbb{E}_{X,Z}\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle X+\sigma Z, RX\rangle\right)\right] \le \mathbb{E}_X\left[\log \mathbb{E}_{Z,R} \exp\left(\frac{1}{\sigma^2}\langle X+\sigma Z, RX\rangle\right)\right]$$
$$= \mathbb{E}_X\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle + \frac{1}{2\sigma^2}\|RX\|^2\right)\right]$$
$$= \mathbb{E}_X\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle + \frac{1}{2\sigma^2}\|X\|^2\right)\right]$$
$$= \frac{1}{2}\sigma^{-2}L + \mathbb{E}_X\left[\log \mathbb{E}_R \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle\right)\right]$$
$$\le \frac{1}{2}\sigma^{-2}L + \log \mathbb{E}_{R,X} \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle\right).$$

Plugging into the expression in Lemma 5.6, we get

$$I(X;Y) \le \frac{L}{2}\log(1+\sigma^{-2}) - \frac{1}{2}L\sigma^{-2} + \log \mathbb{E}_{R,X} \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle\right).$$

Note that as $L, \sigma^2 \to \infty$, already $\frac{L}{2}\log(1+\sigma^{-2}) - \frac{1}{2}L\sigma^{-2} = O(\sigma^{-4}L)$. Observe that $\langle X, RX\rangle = \langle X, R^\top X\rangle = \frac{1}{2}\langle X, (R+R^\top)X\rangle$. By Lemma SM4.1, all the matrices $R_\ell + R_\ell^\top$ are diagonalized by some orthonormal basis with eigenvalues $\{2\cos\left(\frac{2\pi}{L}k\ell\right)\}_{k=0}^{L-1}$. By the orthogonal invariance of $X \sim \mathcal{N}(0, I)$, there are i.i.d. $W_{k,\ell} \sim \mathcal{N}(0, 1)$ such that for all ℓ,

$$\sigma^{-2}\langle X, R_\ell X\rangle = \sigma^{-2}\sum_{k=0}^{L-1}\cos\left(\frac{2\pi}{L}k\ell\right)W_{k,\ell}^2.$$

Recall that the moment generating function of a χ^2 random variable is

$$\mathbb{E}_{W\sim\mathcal{N}(0,1)}[e^{tW^2}] = (1-2t)^{-1/2} \quad \text{for } t > 1/2\,,$$

see, e.g, [18, page 621]. Therefore, assuming σ^2 is sufficiently large (e.g., $\sigma^2 > 2$),

$$\begin{aligned} \log \mathbb{E}_{R,X} \exp\left(\frac{1}{\sigma^2}\langle X, RX\rangle\right) &= \log\left[L^{-1}\sum_{\ell=0}^{L-1}\prod_{k=0}^{L-1}\left(1-2\sigma^{-2}\cos\left(\frac{2\pi}{L}k\ell\right)\right)^{-1/2}\right] \\ &= \log\sum_{\ell=0}^{L-1} e^{\psi_\ell} - \log(L), \end{aligned}$$

where

$$\psi_\ell = -\frac{1}{2}\sum_{k=0}^{L-1}\log\left(1-2\sigma^{-2}\cos\left(\frac{2\pi}{L}k\ell\right)\right).$$

Expanding the log function to first order around 1 and noting that $\sum_{k=0}^{L-1}\cos\left(\frac{2\pi}{L}k\ell\right) = L\cdot\mathbb{1}_{\{\ell=0\}}$ (see Lemma SM4.1), for large values of L and σ^2, we get

$$\psi_\ell = \sum_{k=0}^{L-1}\sigma^{-2}\cos\left(\frac{2\pi}{L}k\ell\right) + O(\sigma^{-4}L) = \begin{cases}\sigma^{-2}L + O(\sigma^{-4}L) & \text{if } \ell = 0,\\ O(\sigma^{-4}L) & \text{otherwise.}\end{cases}$$

Thus, we have the estimate

$$\begin{aligned} \log\sum_{\ell=0}^{L-1} e^{\psi_\ell} - \log(L) &= \log\left(\frac{1}{L}e^{\sigma^{-2}L+O(\sigma^{-4}L)} + \frac{L-1}{L}e^{O(\sigma^{-4}L)}\right) \\ &= \log\left(1+L^{-1}e^{\sigma^{-2}L}\right) + O(\sigma^{-4}L), \end{aligned}$$

from which the claimed result immediately follows. ∎

Observe that for $\alpha > 1$, Proposition 5.7 gives an upper bound of the order $I(X;Y) = O(\log(L))$. It will turn out that when $\alpha > 2$, this is indeed the right order of magnitude. However, for $1 < \alpha \leq 2$ the bound is too loose, and in fact $I(X;Y) = o(\log(L))$.

5.3.2. MI bound using template matching. We start from Lemma 5.3 which gives, for $n = 1$ and $Y = RX + \sigma Z$, $I(X;Y) = \frac{L}{2}\log(1+\sigma^{-2}) - I(R;X|Y)$. We make the important observation that R and Y are independent; indeed, regardless of R, it holds that $Y|R \sim \mathcal{N}(0,(1+\sigma^2)I)$. We remark, however, that when $n > 1$, Y^n is not independent of R^n. We can therefore use Proposition SM1.1 item 5, and Proposition SM1.1 item 2 to write

$$I(R;X|Y) = H(R|Y) - H(R|X,Y) = H(R) - H(R|X,Y) = \log(L) - H(R|X,Y),$$

so that

$$I(X;Y) = \frac{L}{2}\log(1+\sigma^{-2}) - \log(L) + H(R|X,Y). \quad (5.7)$$

The following is now an immediate consequence of Fano's inequality (Proposition SM1.2) and Theorem 4.5.

Proposition 5.8. *Suppose that $\sigma^{-2}L = \alpha\log(L)$ with $\alpha > 2$. Then,*

$$I(X;Y) = \frac{L}{2}\log(1+\sigma^{-2}) - (1+o(1))\log(L)$$
$$= \left(\frac{\alpha}{2} - 1 + o(1)\right)\log(L) + O(\sigma^{-4}L).$$

Proof. We estimate $H(R|X,Y)$. Clearly, $H(R|X,Y) \geq 0$ by non-negativity of entropy (Proposition SM1.1 item 1). As for an upper bound, by Fano's inquality (Proposition SM1.2), for any estimator $\widehat{R}$ of R from X, Y, the error probability $p_e = \Pr(R \neq \widehat{R})$ satisfies

$$H(R|X,Y) \leq \log 2 + p_e \log(L).$$

By Theorem 4.5, $\widehat{R}_{\mathrm{MAP}}$ has error $p_e \to 0$, which means that $H(R|X,Y) = o(1)\cdot\log(L) = o(\log(L))$. Plugging this into equation (5.7) and expanding $\frac{L}{2}\log(1+\sigma^{-2}) = \frac{\alpha}{2}\log(L) + O(\sigma^{-4}L)$, we obtain the desired estimate for $I(X;Y)$. ■

Proposition 5.8 above will not be needed for our main results, but its proof serves as good exposition towards bounding the conditional entropy $H(R|X,Y)$ in the harder case $\alpha \leq 2$. When $\alpha < 2$ we have $p_e \to 1$, so that it is no longer true that $H(R|X,Y) = o(\log(L))$. Indeed, since $I(X;Y) = (\alpha/2-1)\log(L) + O(\sigma^{-4}L) + H(R|X,Y)$, we must have that $H(R|X,Y) \geq (1-\alpha/2-o(1))\log(L)$, since the MI is non-negative. While, indeed, in this regime R cannot be recovered from X, Y, we can still obtain a non-trivial upper bound (of the form $c(\alpha)\log(L)$ for some $c(\alpha) < 1$) on the conditional entropy $H(R|X,Y)$; the idea is that given X, Y, we can form a relatively small list that contains R with high probability.

Our goal, then, is to non-trivially upper bound $H(R|X,Y)$ in the regime $\alpha \leq 2$ where $p_e \not\to 0$. Let $\tau > 0$, and denote by S_τ the set of τ-likely shifts:

$$\mathcal{S}_\tau = \left\{ R' : \frac{\langle X, (R')^{-1}Y\rangle}{\|X\|^2} \geq 1-\tau \right\}. \tag{5.8}$$

The analysis of Section 4 tells us that for any $\tau > 0$, the true shift R belongs with high probability to the set $\mathcal{S}_\tau$. Moreover, when $\alpha > 2$ (and $\tau > 0$ is a sufficiently small constant), in fact with high probability $S_\tau = \{R\}$. When $\alpha \leq 2$ this will no longer be the case; nonetheless, we show that $|\mathcal{S}_\tau|$ is with high probability significantly smaller than L. This means that given X and Y, we can produce a list of likely candidates for R which is much smaller than the entire group of shifts. The following lemma is proved in the SI Appendix, Section SM5.

Lemma 5.9. *Let $\kappa, \tau, \zeta > 0$. Set $M = L^{1-\frac{1}{2}\alpha(1-\kappa)\left(1-\tau-\frac{\kappa}{1-\kappa}\right)^2+\zeta}$, and assume that $\alpha \leq 2$. Then*

$$\Pr\left(R \notin \mathcal{S}_\tau \text{ or } |\mathcal{S}_\tau| > M\right) \leq 2Le^{-cL\min(\kappa,\kappa^2)} + L^{-\frac{1}{2}\alpha(1-\kappa)\left(1-\tau-\frac{\kappa}{1-\kappa}\right)^2} + 2L^{-\zeta}, \tag{5.9}$$

where $c > 0$ is the universal constant of Lemma 4.1.

Lemma 5.9 implies that there are slowly decaying sequences $\tau = \tau_L = o(1), \delta = \delta_L = o(1)$ such that the event

$$\mathcal{B} = \left\{ R \in \mathcal{S}_{\tau_L} \text{ and } |\mathcal{S}_{\tau_L}| \leq L^{1-\frac{1}{2}\alpha+\delta_L} \right\}$$

holds with high probability of $\Pr(\mathcal{B}) = 1 - o(1)$. We use this to bound the conditional entropy $H(R|X,Y)$, and obtain a bound on the MI:

Proposition 5.10. *Suppose that $\alpha \le 2$. Then,*

$$I(X;Y) = o(\log(L)).$$

Proof. We upper bound the conditional entropy $H(R|X,Y)$ using a "Fano-like" argument. Let E be the indicator for the event $\mathcal{B}$ above. Since E is completely deterministic given (R, X, Y), we have that $H(E|R,X,Y) = 0$ by Proposition SM1.1 item 1 and by the chain rule of entropy (Proposition SM1.1 item 4) we have

$$\begin{aligned} H(R|X,Y) &= H(R|X,Y) + H(E|R,X,Y) \\ &= H(R,E|X,Y) \\ &= H(E|X,Y) + H(R|X,Y,E) \\ &\le H(E) + H(R|X,Y,E=1)\Pr(E=1) + H(R|X,Y,E=0)\Pr(E=0), \end{aligned}$$

where we have bounded $H(E|X,Y) \le H(E)$ using Proposition SM1.1 item 5, and expanded $H(R|X,Y,E)$ according to the definition of conditional entropy, averaging only with respect to E.

Now, given that $E = 1$, we know that R belongs to $\mathcal{S}_{\tau_L}$, which has size $|\mathcal{S}_{\tau_L}| \le M = L^{1-\frac{1}{2}\alpha+\delta_L}$. Hence, $H(R|X,Y,E=1) \le \log(M) = \left(1 - \frac{1}{2}\alpha + \delta_L\right)\log(L)$ by Proposition SM1.1 item 2, and by the same reason $H(R|X,Y,E=0) \le \log(L)$. By definition, $\Pr(E=1) = \Pr(\mathcal{B}) = 1 - o(1)$, and $H(E) \le \log(2)$ by Proposition SM1.1 item 2. Thus, $H(R|X,Y) \le \left(1 - \frac{1}{2}\alpha + o(1)\right)\log(L)$. Plugging this into Eq. (5.7),

$$\begin{aligned} I(X;Y) &= \frac{L}{2}\log(1+\sigma^{-2}) - \log(L) + H(R|X,Y) \\ &= \left(\frac{\alpha}{2} - 1 + o(1)\right)\log(L) + O(\sigma^{-4}L) + \left(1 - \frac{\alpha}{2} + o(1)\right)\log(L) \\ &= o(\log(L)) + O(\sigma^{-4}L), \end{aligned}$$

as claimed. ■

Remark 5.11. One might wonder if the argument above (if carried out delicately enough) can match the estimate $I(X;Y) = O(L^{-1+\alpha})$ we have already seen for $\alpha < 1$. Unfortunately, the bound $\Pr(|S_\tau| \ge M) \le 2L^{-\delta}$ (using Markov's inequality; see the proof of Lemma 5.9 in SI Appendix, Section SM5) is already too crude for that purpose: since we need to choose $\delta = o(1)$, the $o(1)$ correction above must decay slower than L^{-c} (for any $c > 0$).

5.3.3. Proof of main results. We are ready to prove Theorem 2.2 and the sample complexity lower bounds of Theorem 2.1.

Proof of Theorems 2.1 (lower bounds) and 2.2..

- Theorem 2.1, $\alpha > 2$ (lower bound): Corollary 5.4 immediately implies that $\lim_{\varepsilon\to 0}\lim_{L\to\infty} \frac{n^*_{\mathrm{MRA}}(L,\alpha,\varepsilon)}{\sigma^2/\varepsilon} \ge 1$.
- Theorem 2.1, $\alpha \le 2$: Combining Proposition 5.5 and Proposition 5.10, give $n^*_{\mathrm{MRA}}(L,\alpha,\varepsilon) = \omega\left(\frac{L}{\log(L)}\log(1/\varepsilon)\right) = \omega\left(\sigma^2\log(1/\varepsilon)\right)$.
- Theorem 2.2, $\alpha < 1$: Combining Proposition 5.5 and Proposition 5.7 yield $n^*_{\mathrm{MRA}}(L,\alpha,\varepsilon) = \Omega(L^{2-\alpha}\log(1/\varepsilon))$.

The proof of the upper bound $\lim_{\varepsilon\to 0}\lim_{L\to\infty}\frac{n^*_{\mathrm{MRA}}(L,\alpha,\varepsilon)}{\sigma^2/\varepsilon} \le 1$ for $\alpha > 2$ (item (1) of Theorem 2.1) appears in Section 6.

6. Sample complexity upper bound for $\alpha > 2$ via brute-force template matching.

In this section we propose a recovery algorithm for the high SNR regime $\alpha > 2$, which essentially matches our $\Omega(L/\log L)$ lower bound on the sample complexity. Our goal here is not to propose a new MRA algorithm, but rather to establish a matching upper bound on the *statistical difficulty* of the problem; that is, we are studying the fundamental information-theoretic (rather than computational) limits of MRA. [4] In particular, the proposed algorithm is computationally intractable, and involves a brute-force search on an exponentially sized set of candidates. Moreover, our approach is tailored to the case $\alpha > 2$, which is exactly the SNR regime where template matching is statistically possible.

Outline of our algorithm. Before diving into the technical details of our proposed scheme, we give a brief outline of the approach. The estimation algorithm works in two stages. Suppose we are given n independent samples. We divide them into two subsamples of sizes n_1 and n_2, $n_1 + n_2 = n$. We do this so to ensure that the estimator $\widehat{Q}$ produced in step 1 is statistically independent of the additive noise in the samples used for step 2. This simplifies our analysis considerably. The two stages performed by the algorithm are the following.

1. *Brute-force search for a template:* In the first stage, we use the first n_1 samples to find some direction $\widehat{Q} \in \mathbb{S}^{L-1}$ (here $\mathbb{S}^{L-1}$ is the unit sphere in $\mathbb{R}^L$) such that $\widehat{Q}$ is sufficiently well-aligned with some shift of the true signal, that is, $\max_\ell L^{-1/2}\langle X, R_\ell^{-1}\widehat{Q}\rangle \ge 1-\eta$, where $\eta = \eta(\alpha)$ is small. To do this, we consider a fine-enough cover of the sphere, $\mathcal{N} \subset \mathbb{S}^{L-1}$, and take $\widehat{Q} \in \mathcal{N}$ as the minimizer of a certain score: $\widehat{Q} = \mathrm{argmin}_{Q\in\mathcal{N}} \sum_{i=1}^{n_1} s_i(Q)$, where $s_i(Q)$ is computed from the i-th sample Y_i. Minimizing $\sum_{i=1}^{n_1} s_i(Q)$ over $\mathbb{S}^{L-1}$ boils down to a brute-force search over the cover, whose size is exponential in L. Hence, this algorithm is not efficient. In principle, one could take at this point $\sqrt{L}\widehat{Q} \approx \|X\|\widehat{Q}$ as an estimator for X. Unfortunately, the MSE of this estimator decays at a suboptimal rate with respect to the number of samples n; this is remedied by the second step.
2. *Alignment and averaging:* Using $\widehat{Q}$ from the previous step, we perform template matching on the remaining n_2 samples $Y_1, \ldots, Y_{n_2}$ in order to estimate their shifts relative to $\widehat{Q}$:
$$\widehat{R}_{\ell_i} = \underset{\ell}{\mathrm{argmax}}\langle Y_i, R_\ell \widehat{Q}\rangle.$$
The final estimator for X is then the average of the aligned measurements:
$$\widehat{X} = \frac{1}{n_2}\sum_{i=1}^{n_2} \widehat{R}_{\ell_i}^{-1} Y_i.$$

All the missing technical details are provided in the next two sections. Due to space constraints, the proofs of all lemmas are given in the SI Appendix, Section SM7.

[4] This distinction is not trivial in general. In the context of MRA, for instance, previous papers conjectured that a natural extension of the MRA model, called heterogeneous MRA, suffers from a fundamental computational-statistical gap [6,16]. We *do not* claim, however, that such a computational-statistical gap holds for the MRA model considered in this paper, with α close to 2.

Main result of this section.. The main result of this section is the following:

Proposition 6.1. *Suppose that $\alpha > 2$, fix $\varepsilon > 0$, and let $n, L \to \infty$. Then, there exists some $c(\alpha) > 0$ depending on α such that if*

$$n_1 = c(\alpha)\sigma^2, \quad n_2 = (1 + o(1))\frac{\sigma^2}{\varepsilon},$$

then the estimator $\widehat{X}$ returned by our algorithm satisfies $\rho(X, \widehat{X}) \leq \varepsilon$ with probability $1 - o(1)$.

Note that when $\varepsilon > 0$ is small, the sample complexity is dominated by n_2:

$$n = c(\alpha)\sigma^2 + (1 + o(1))\frac{\sigma^2}{\varepsilon} \approx (1 + o(1))\frac{\sigma^2}{\varepsilon},$$

and thus almost independent of the constant $c(\alpha)$. Proposition 6.1 should be compared with the optimal achievable MSE for estimating a signal in AWGN, without the shifts $L^{-1}\mathbb{E}\|X - \widehat{X}_{\text{MMSE}}\|^2 = \frac{\sigma^2}{\sigma^2+n}$.

Proof of Theorem 2.1 (upper bound). The upper bound for $\alpha > 2$ follows readily from Proposition 6.1. To show this, we construct a new estimator $[\widehat{X}]$ as follows: $[\widehat{X}] = \widehat{X}$ if $\|\widehat{X}\| \leq 10\sqrt{L}$ and $[\widehat{X}] = 0$ otherwise. Note that under the high-probability event $\|X\| \leq 2\sqrt{L}$, necessarily $\rho(X, [\widehat{X}]) \leq \rho(X, \widehat{X})$. Write

$$\mathbb{E}\rho(X, [\widehat{X}]) = \mathbb{E}\left[\rho(X, [\widehat{X}])\mathbb{1}_{\|X\| \leq 2\sqrt{L}}\right] + \mathbb{E}\left[\rho(X, [\widehat{X}])\mathbb{1}_{\|X\| > 2\sqrt{L}}\right].$$

Under $\|X\| \leq 2\sqrt{L}$, the random variable $\rho(X, [\widehat{X}])$ is bounded by a constant, hence by Proposition 6.1,

$$\mathbb{E}\left[\rho(X, [\widehat{X}])\mathbb{1}_{\|X\| \leq 2\sqrt{L}}\right] \leq \varepsilon + o(1),$$

since $\rho(X, \widehat{X}) \leq \varepsilon$ holds w.p. $1 - o(1)$. As for the other term,

$$\mathbb{E}\left[\rho(X, [\widehat{X}])\mathbb{1}_{\|X\| > 2\sqrt{L}}\right] \leq \mathbb{E}\left[L^{-1/2}(\|X\| + 10L^{1/2})\mathbb{1}_{\|X\| > 2\sqrt{L}}\right] \leq 6\mathbb{E}\left[L^{-1/2}\|X\|\mathbb{1}_{L^{-1/2}\|X\| > 2}\right],$$

so that by Cauchy-Schwartz,

$$\mathbb{E}\left[L^{-1/2}\|X\|\mathbb{1}_{L^{-1/2}\|X\| > 2}\right] \leq \left(L^{-1}\mathbb{E}[\|X\|^2]\right)^{1/2}\left(\Pr(\|X\| > 2\sqrt{L})\right)^{1/2} = o(1).$$

Thus, $[\widehat{X}]$ uses $n = [(1 + o(1))/\varepsilon + c(\alpha)]\,\sigma^2$ samples and achieves $\mathbb{E}\rho(X, [\widehat{X}]) \leq \varepsilon + o(1)$, so that

$$\limsup_{L\to\infty} \frac{n^*_{\text{MRA}}(L, \alpha, \varepsilon)}{\sigma^2/\varepsilon} \leq 1 + O_\alpha(\varepsilon).$$

Class of "nice signals.". Before getting to the details of the algorithm, in the analysis that follows, it is convenient to treat the signal X as fixed and belonging some class of "nice" signals. Specifically, we require that: (i) the signal is sufficiently uncorrelated with its shifts, in that $L^{-1}\langle X, R_\ell X\rangle \approx 0$ for all $\ell \neq 0$, and its norm is concentrated around $L^{-1}\|X\|^2 \approx 1$; (ii) The Fourier (DFT) coefficients of X are uniformly bounded.

Let $f_0, \dots, f_{L-1} \in \mathbb{C}^L$ be the DFT basis vectors, that is, $(f_\ell)_j = L^{-1/2} e^{\frac{2\pi i}{L}\ell j}$, and $\mathcal{F} \in U(L)$ be the matrix whose columns are $f_0, \dots, f_{L-1}$, so that $\mathcal{F}^* X \in \mathbb{C}^L$ are the Fourier coefficients of X (here $\mathcal{F}^*$ denotes the Hermitian conjugate of $\mathcal{F}$.) For $\kappa > 0$, we formally consider the set (6.1)

$$\mathbb{X}_\kappa = \left\{ X \in \mathbb{R}^L \quad : \quad \max_\ell \left| L^{-1} \langle X, R_\ell X \rangle - \mathbb{1}_{\{\ell=0\}} \right| \le \kappa, \quad \text{and } \|\mathcal{F}^* X\|_\infty \le \sqrt{10 \log(L)} \right\},$$

where $\mathbb{1}_{\{\ell=0\}} = 1$ when $\ell = 0$ and is zero otherwise. We take $\kappa = o(1)$ sufficiently large so to ensure that when $X \sim \mathcal{N}(0, I)$, the constraint $\max_\ell \left| L^{-1}\langle X, R_\ell X\rangle - \mathbb{1}_{\{\ell=0\}} \right| \le \kappa$ holds with probability $1 - o(1)$ as $L \to \infty$; by Lemma 4.1, we may choose $\kappa = c \log(L)/\sqrt{L}$ for $c > 0$ a large enough constant. Let $\mathbb{X}$ be the set corresponding to such choice. To lighten the notation, we will not keep track of κ explicitly, instead referring to all vanishing terms as $o(1)$. For the other constraint, the exact bound $\|\mathcal{F}^* X\|_\infty \le \sqrt{10 \log(L)}$ is somewhat arbitrary, in that 10 can be replaced with any constant greater than 4. The following is quite immediate at this point:

Lemma 6.2. *Suppose that $X \sim \mathcal{N}(0, I)$. Then,* $\Pr(X \notin \mathbb{X}) = o(1)$.

We note that it is likely that without assuming that the estimation is over a class of "nice" signals (for example, the class $\mathbb{X}_\kappa$), the situation changes. On that note, we mention the work [17], where it is shown that there are signals X for which the MLE only attains the rate $\rho(X, \widehat{X}_{\mathrm{MLE}}) \sim n^{-1/2}$.

6.1. Step 1: Brute force template matching.

Recall that our intermediate goal here is to find a direction $\widehat{Q} \in \mathbb{S}^{L-1}$ such that $\max_\ell L^{-1/2}\langle X, R_\ell^{-1}\widehat{Q}\rangle \ge 1 - \eta$, where $\eta > 0$ is some desired accuracy level. Since, assuming $X \in \mathbb{X}$, for any $Q \in \mathbb{S}^{L-1}$,

$$\left\| \frac{X}{\|X\|} - R_\ell^{-1} Q \right\|^2 = 2 - 2\left\langle \frac{X}{\|X\|}, R_\ell^{-1} Q \right\rangle = 2 - 2L^{-1/2}\langle X, R_\ell^{-1} Q\rangle + o(1),$$

then taking $\mathcal{N}$ to be a $\sqrt{\eta}$-cover of $\mathbb{S}^{L-1}$, it must contain some $Q \in \mathcal{N}$ with $L^{-1/2}\langle Q, R_\ell^{-1} X\rangle \ge 1 - \frac{1}{2}\eta + o(1)$. It is well known that one can find a cover of the sphere which is not too large:

Lemma 6.3. *[Lemma 5.13 in [44]] There exists an $\sqrt{\eta}$-cover $\mathcal{N}$ of $\mathbb{S}^{L-1}$ of size $|\mathcal{N}| \le (3/\sqrt{\eta})^L$. That is, there exists a set $\mathcal{N} \subset \mathbb{S}^{L-1}$ of size $|\mathcal{N}| \le (3/\sqrt{\eta})^L$, such that $\forall X \in \mathbb{S}^{L-1}, \exists Q \in \mathcal{N}$ with $\|X - Q\| \le \sqrt{\eta}$.*

For each $Q \in \mathcal{N}$, we define its per-sample score:

$$s_i(Q) = s_i^\eta(Q) = \mathbb{1}\left[\max_\ell L^{-1/2}\langle Y_i, R_\ell^{-1} Q\rangle \ge 1 - \frac{3}{4}\eta \right],$$

and the total score $s(Q) = \sum_{i=1}^{n_1} s_i(Q)$, n_1 being the number of samples allocated for this step. That is, $s(Q)$ is the number of samples Y_i such that $L^{-1/2}\langle Q, R_\ell^{-1} Y_i\rangle \ge 1 - \frac{3}{4}\eta$ for some ℓ. The returned estimator is then simply

$$\widehat{Q} = \operatorname*{argmax}_{Q \in \mathcal{N}} s(Q).$$

Note that $s_i(\cdot)$ could be thought of as a discontinuous proxy for the log-likelihood (restricted to $X \in \mathbb{S}^{L-1}$): $\log P(Y_i|X) = \log \sum_{\ell=0}^{L-1} \exp\left(\frac{1}{\sigma^2}\langle X, R_\ell^{-1} Y_i\rangle\right) + \text{constant}$. When σ is small, the log-likelihood is essentially dominated by $\max_\ell \sigma^{-2}\langle X, R_\ell^{-1} Y_i\rangle$. Maximizing the likelihood is computationally more straightforward (in the sense that this is a continuous optimization problem, no need to quantize the domain as we do); however, analyzing the MLE directly appears to be difficult [22, 27].

We start by showing that there are only a few shifts ℓ such that $L^{-1/2}\langle X, R_\ell^{-1} Q\rangle$ are all large.

Lemma 6.4. *Suppose that $X \in \mathbb{X}$. For $Q \in \mathbb{S}^{L-1}$, let*

$$N_Q(h) = \left|\left\{\ell \,:\, L^{-1/2}\left|\langle X, R_\ell^{-1} Q\rangle\right| \geq h\right\}\right|.$$

Then, $N_Q(h) \leq h^{-2}\|\mathcal{F}^* X\|_\infty^2 \leq h^{-2} \cdot 10\log(L)$.

We next show that if $\max_\ell L^{-1/2}\langle X, R_\ell^{-1} Q\rangle$ is small, then with high probability the score $s(Q)$ is not large.

Lemma 6.5. *Assume that $X \in \mathbb{X}$, $\alpha > 2$, $\eta < 1 - \sqrt{2/\alpha}$, and L is large enough so that $\log(L) \leq L^{3\eta^2\alpha/128}$. Suppose that $Q \in \mathbb{S}^{L-1}$ is such that $\max_\ell L^{-1/2}\langle X, R_\ell^{-1} Q\rangle \leq 1 - \eta$, then*

$$\Pr\left(s(Q) \geq n_1/2\right) \leq \left[16\left(2 + \frac{640}{\left(1 - \sqrt{\frac{2}{\alpha}}\right)^2}\right) L^{-\eta^2\alpha/128}\right]^{n_1/2}.$$

Next, we prove that if $\max_\ell\langle X, R_\ell^{-1} Q\rangle$ is sufficiently large, then $s(Q)$ is large with high probability.

Lemma 6.6. *Assume that $X \in \mathbb{X}$, $\alpha > 2$, and L is large enough so that $L^{\eta^2\alpha/64} \geq 4$. Suppose that $Q \in \mathbb{S}^{L-1}$ is such that $\max_\ell\langle X, R_\ell^{-1} Q\rangle \geq 1 - 5\eta/8$. Then,*

$$\Pr(s(Q) < n_1/2) \leq e^{-n_1/32}.$$

We are now ready to conclude the analysis of Step 1 of our algorithm.

Proposition 6.7. *Assume that $X \in \mathbb{X}$, $\alpha > 2$, and $\eta < 1 - \sqrt{2/\alpha}$. Then, there is constant $c > 0$, such that whenever*

$$n_1 \geq c\frac{L\log(1/\eta)}{\alpha\eta^2\log(L)} = c\frac{\sigma^2\log(1/\eta)}{\eta^2},$$

the vector $\widehat{Q} = \operatorname{argmax}_{Q\in\mathcal{N}} s(Q)$ satisfies $\max_\ell\langle X, R_\ell^{-1} Q\rangle \geq 1 - \eta$ with probability $1 - o(1)$ as $n_1, L \to \infty$. In fact, the error probability decays exponenentially fast with n_1.

Proof. As argued in the beginning of this section, the $\sqrt{\eta}$-cover $\mathcal{N}$ contains some $Q \in \mathbb{S}^{L-1}$ such that $L^{-1/2}\langle X, R_\ell^{-1} Q\rangle \geq 1 - \eta/2 - o(1) \geq 1 - 5\eta/8$ for some ℓ. By Lemma 6.6, with probability greater than $1 - e^{-n_1/32}$, this vector has score $s(Q) \geq n_1/2$. It therefore suffices to show that with high probability, all the vectors $Q \in \mathcal{N}$ that are bad, meaning that $\max_\ell L^{-1/2}\langle X, R_\ell^{-1}Q\rangle < 1-\eta$, have score $s(Q) < n_1/2$. By Lemmas 6.3 and 6.5,

$$\begin{aligned}\Pr\left(\exists \text{bad } Q \in \mathcal{N} \,:\, s(Q) \ge n_1/2\right) &\le |\mathcal{N}| \cdot \Pr\left(s(Q) \ge n_1/2 \,\middle|\, Q \text{ is bad}\right) \\ &\le (9/\eta)^{L/2} \cdot \left[16\left(2 + \frac{640}{\left(1-\sqrt{\frac{2}{\alpha}}\right)^2}\right) L^{-\eta^2\alpha/128}\right]^{n_1/2} \\ &\le \left(C(\alpha) e^{-c_1\eta^2\alpha\log(L) + c_2\frac{L}{n}\log(1/\eta)}\right)^{n_1},\end{aligned}$$

where $c_1, c_2 > 0$ are absolute constants, and $C(\alpha)$ depends on α. Then, this probability tends to 0 as $n_1, L \to \infty$ (exponentially fast in n_1) whenever $n_1 \ge c\frac{L\log(1/\eta)}{\alpha\eta^2\log(L)}$ for some other $c > 0$. ■

Note that at this point we could take $\widehat{X} = L^{1/2} \cdot \widehat{Q}$ as an estimator for X, so that

$$\rho(X, \widehat{X}) = \min_\ell \|L^{-1/2}X - R_\ell^{-1}Q\|^2 \le 2\eta + o(1),$$

holds with high probability. For *fixed* η, this estimator indeed captures the correct dimensional scaling of the sample complexity, namely, that $n = O(L/(\alpha\log L)$ samples are sufficient to get non-trivial alignment error. However, its dependence on η is seemingly quite bad: for estimating a signal in AWGN, without the shifts, the optimal dependence on η should look like $O(L/(\alpha\log L)\cdot\eta^{-1})$, rather than the much worse $O\left(L/(\alpha\log L)\cdot\eta^{-2}\log(1/\eta)\right)$ we were able to show. In the next section, we see how to achieve this "correct" rate by essentially recovering the shifts on all but a vanishing fraction of the samples, and averaging the properly aligned measurements.

6.2. Step 2: Achieving optimal MSE decay rate by alignment and averaging. Suppose that one has access to a known template $Q \in \mathbb{S}^{L-1}$, such that $\langle X, Q\rangle \ge 1-\eta$. Since $L^{-1}\|X\|^2 = 1 + o(1)$, this is the same as having $\|L^{-1/2}X - Q\|^2 \le 2\eta + o(1)$, and since $\max_{\ell\neq 0} L^{-1}|\langle X, R_\ell X\rangle| = o(1)$, we see that for any $\ell \neq 0$,

$$\|L^{-1/2}R_\ell X - Q\| \ge \|L^{-1/2}[R_\ell X - X]\| - \|L^{-1/2}X - Q\| \ge \sqrt{2} - \sqrt{2\eta} - o(1).$$

In particular, we see that when $\sqrt{2\eta} < \sqrt{2} - \sqrt{2\eta}$, that is, $\eta < 1/4$ (and L is sufficiently large), there is a *unique* ℓ (specifically, $\ell = 0$) such that $\|L^{-1/2}X - R_\ell Q\|^2 \le 2\eta + o(1)$. In that case, the idea of matching a sample $Y_i = R_{\ell_i}X + \sigma Z$ against the template Q becomes well-posed, in the sense that its desired outcome is clear: we would like to recover the shift R_{ℓ_i}.

Lemma 6.8. *Assume that $X \in \mathbb{X}$ and $\alpha > 2$. Let $Y = R_\ell X + \sigma Z$, and suppose that $Q \in \mathbb{S}^{L-1}$ is independent of Y and satisfies $\max_{\ell'} L^{-1/2}\langle X, R_{\ell'}^{-1}Q\rangle \ge 1-\eta$, where*

$$\sqrt{\eta} < \frac{1}{2}(1 - \sqrt{2/\alpha}).$$

Denote the maximizing shift by ℓ^*. Let $\widehat{\ell} = \operatorname{argmax}_{\ell'}\langle Y, R_{\ell'}Q\rangle$. Then

$$\Pr\left(\widehat{\ell} \neq \ell - \ell^*\right) \le 2L^{-\frac{1}{2}\alpha\left(1/2 - 1/\sqrt{2\alpha} - \sqrt{\eta}\right)^2 + o(1)}.$$

Given Lemma 6.8, we propose the following estimation strategy. Suppose we would like to estimate X up to error $\rho(X,\widehat{X}) \le \varepsilon < 1$. Fix some $\eta > 0$ with $\sqrt{\eta} < (1-\sqrt{2/\alpha})/2$ (for concreteness, say $\eta = (1-\sqrt{2/\alpha})^2/16$). We first apply the algorithm of Step 1 (Setion 6.1) to obtain $\widehat{Q} \in \mathbb{S}^{L-1}$ such that $\max_\ell \langle X, R_\ell^{-1}\widehat{Q}\rangle \ge 1-\eta$. Assuming that $n_1 \ge \frac{c\log(1/\eta)}{\eta^2}\sigma^2 = c_\eta \sigma^2$, we are successful with probability $1-o(1)$. Let ℓ^* be such that $\langle X, R_{\ell^*}^{-1}Q\rangle \ge 1-\eta$. Next, for n_2 new independent samples, we compute for each measurement $\widehat{\ell}_i = \operatorname{argmax}_\ell \langle Y_i, R_\ell \widehat{Q}\rangle$ and return the aligned sample average:

$$\widehat{X} = \frac{1}{n_2}\sum_{i=1}^{n_2} R_{\widehat{\ell}_i}^{-1} Y_i. \tag{6.2}$$

Lemma 6.8 tells us that we should expect most of the aligned measurements $R_{\widehat{\ell}_i}^{-1}Y_i$ to be well-aligned with $R_{\ell^*}X$, that is, $R_{\widehat{\ell}_i}^{-1}Y_i = R_{\ell^*}X + \mathcal{N}(0,\sigma^2 I)$. This means that, $\widehat{X} \approx R_{\ell^*}X + \mathcal{N}(0,(\sigma^2/n_2)I)$, hence $\rho(X,\widehat{X}) \le L^{-1}\|R_{\ell^*}X - \widehat{X}\|^2 \approx \sigma^2/n_2$, which is smaller than ε if $n_2 \ge \sigma^2/\varepsilon$. We make this argument precise below:

Proposition 6.9. *Assume that $X \in \mathbb{X}$ and $\alpha > 2$. Fix $\varepsilon > 0$ and some $\eta < \frac{1}{2}(1-\sqrt{2/\alpha})^2$. Let $\widehat{Q} \in \mathbb{S}^{L-1}$ be the output of Step 1 (run with a tuning parameter η and n_1 samples). Let $\widehat{X}$ be as in equation (6.2), computed from n_2 new samples. Suppose that $n_1, n_2, L \to \infty$ with*

$$n_1/\sigma^2 \to \gamma_1, \quad n_2/\sigma^2 \to \frac{\gamma_2}{\varepsilon},$$

where γ_1 and γ_2 are constants satisfying

$$\gamma_1 = \gamma_1(\eta) \ge \frac{c\log(1/\eta)}{\eta^2}, \quad \gamma_2 > 1,$$

(c being the universal constant from Proposition 6.7). Then,

$$\Pr\left(\rho(X,\widehat{X}) \le \varepsilon\right) \to 1.$$

Proposition 6.1 now immediately follows from Lemma 6.2 and Proposition 6.9.

7. Conclusions and extensions.

In this work we have studied the sample complexity of the MRA problem in the limit of large L. In this regime, we have shown that the parameter $\alpha = \frac{\sigma^2\log L}{L}$ plays a crucial role in characterizing the best attainable performance of any estimator.

As mentioned above, the MRA model is primarily motivated by the cryo-EM technology to constitute the 3-D structure of biological molecules. In the cryo-EM literature, it was shown that it is effective to assume that the molecule was drawn from a Gaussian prior with decaying power spectrum [37]. In addition, the 3-D rotations are usually not distributed uniformly over the group $SO(3)$. We now discuss briefly how these different aspects can be potentially incorporated into our framework.

Prior on the signal. Our model assumes a Gaussian i.i.d. prior on the signal X to be reconstructed. While this assumption lends itself to a relatively clean analysis, and allows to compare our bounds on $n^*_{\mathrm{MRA}}(L,\alpha,\varepsilon)$ to the simple benchmark $n^*_{\mathrm{AWGN}}(L,\alpha,\varepsilon)$, many of our results can be generalized to treat other priors on X. In particular, all of our sample complexity lower bounds are based on lower bounding the mutual information between X and $\hat{X}$ under the constraint $\mathbb{E}[\rho(X,\hat{X})] \le \varepsilon$ on the one hand, and upper bounding $I(X;Y^n)$ under the MRA model, on the other hand. In Proposition 5.1 we have relied on the Gaussian rate distortion function to lower bound $I(X;\hat{X})$ for any estimator that achieves MSE at most ε. For X whose distribution is not $\mathcal{N}(0,I)$, we can either compute the corresponding rate distortion function explicitly, or simply apply Shannon's lower bound $R(D) \ge h(X) - \frac{L}{2}\log(2\pi e D)$, see [13]. Our upper bounds on $I(X;Y^n)$ in the regime $\alpha > 1$ are based on Lemma 5.3, followed by lower bounding $I(R^n;X|Y^n)$ using Fano-like arguments. It is easy to see that (5.4) continues to hold, with $\le$ instead of $=$, for any random variable X with $\mathbb{E}\|X\|^2 \le L$. Furthermore, the lower bounds on $I(R^n;X|Y^n)$ we derive in Section 5.3.2 remain valid whenever $\frac{\|X\|}{L}$ is sufficiently concentrated around 1 and $\frac{\langle X, R_\ell X\rangle}{L}$ is sufficiently concentrated around 0 for all $\ell = 1,\dots,L-1$. In particular, this is the case for (sufficiently light-tailed) i.i.d. zero-mean and unit variance distributions. In light of the discussion above, we see that the parameter $\alpha = \frac{\sigma^2 \log L}{L}$ is of great importance whenever the random signal X satisfies the above concentration requirements and has differential entropy proportional to L.

Shift distribution. Assuming uniform prior on the i.i.d. shifts $R_{\ell_1},\dots,R_{\ell_n}$ is a worst-case analysis. Indeed, for any given distribution, shifting all measurements again $R_{u_i}Y_i$, for $u_i \overset{i.i.d.}{\sim}$ Uniform$(\{0,\dots,L-1\})$ before feeding them to the estimator leads to (1.1). However, previous works (for fixed L) showed that harnessing non-uniformity can make a big difference in the sample complexity [1,38]. With some effort, our upper bounds on $I(X;Y^n)$ in the regime $\alpha > 1$ should also extend to treat this case. Here, the main challenge is to generalize Lemma 5.9 to the case of non-uniform distribution, i.e., to find a sharp estimate on the smallest possible size of a list of candidates for the true shift, which contains the true shift with high probability.

Extension to other groups. We believe that many aspects of our information-theoretical analysis can be generalized to other (families of) discrete groups, denoted here by $\mathcal{G}_L$, which satisfy the following properties (roughly speaking): (i) If X is suitably generic and $g \ne h$, then $\langle gX, hX\rangle$ is very small - concretely, if $X \sim \mathcal{N}(0,I)$, then $\mathbb{E}[\langle gX, hX\rangle] = 0$; (ii) The size of the group $|\mathcal{G}_L|$ does not grow too fast (strictly less than exponentially fast in L). These conditions imply that whenever X is isotropic and sufficiently light-tailed (e.g., sub-Gaussian), $\{gX\}_{g\in\mathcal{G}}$ are "almost orthogonal." The proper noise scaling to consider would then be $\sigma^2 = \frac{L}{\alpha \log|\mathcal{G}_L|}$, with $\alpha = 2$ being the critical noise level—this comes from the fact that $\max_{g\in\mathcal{G}_L}\langle gX, Z\rangle \approx \sqrt{2\log|\mathcal{G}_L|}$. For continuous compact groups , we suspect that one might be able to apply some of our arguments by cleverly discretizing the suitable group action. Carrying out a program of this type seems as a promising direction for future research.

Acknowledgment. E.R. and O.O. are supported in part by the ISF under Grant 1791/17. E.R. is supported in part by an Einstein-Kaye fellowship from the Hebrew University of Jerusalem. T.B. is supported in part by NSF-BSF grant no. 2019752, and the Zimin Institute for Engineering Solutions Advancing Better Lives.

REFERENCES

[1] E. Abbe, T. Bendory, W. Leeb, J. M. Pereira, N. Sharon, and A. Singer, *Multireference alignment is easier with an aperiodic translation distribution*, IEEE Transactions on Information Theory, 65 (2018), pp. 3565–3584.

[2] E. Abbe, J. M. Pereira, and A. Singer, *Sample complexity of the boolean multireference alignment problem*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 1316–1320.

[3] E. Abbe, J. M. Pereira, and A. Singer, *Estimation in the group action channel*, in 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018, pp. 561–565.

[4] R. J. Adler and J. E. Taylor, *Random fields and geometry*, Springer Science & Business Media, 2009.

[5] Y. Aizenbud, B. Landa, and Y. Shkolnisky, *Rank-one multi-reference factor analysis*, arXiv preprint arXiv:1905.12442, (2019).

[6] A. S. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. S. Wein, *Estimation under group actions: recovering orbits from invariants*, arXiv preprint arXiv:1712.10163, (2017).

[7] 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.

[8] A. S. Bandeira, Y. Chen, R. R. Lederman, and A. Singer, *Non-unique games over compact groups and orientation estimation in cryo-EM*, Inverse Problems, 36 (2020), p. 064002.

[9] A. S. Bandeira, P. Rigollet, and J. Weed, *Optimal rates of estimation for multi-reference alignment*, arXiv preprint arXiv:1702.08546, (2017).

[10] T. Bendory, A. Bartesaghi, and A. Singer, *Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities*, IEEE Signal Processing Magazine, 37 (2020), pp. 58–76.

[11] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, *Bispectrum inversion with application to multireference alignment*, IEEE Transactions on signal processing, 66 (2017), pp. 1037–1050.

[12] T. Bendory, A. Jaffe, W. Leeb, N. Sharon, and A. Singer, *Super-resolution multi-reference alignment*, arXiv preprint arXiv:2006.15354, (2020).

[13] T. Berger, *Rate distortion theory: A mathematical basis for data compression*, Prentice-Hall, 1971.

[14] T. Bhamre, T. Zhang, and A. Singer, *Denoising and covariance estimation of single particle cryo-EM images*, Journal of structural biology, 195 (2016), pp. 72–81.

[15] N. Boumal, *Nonconvex phase synchronization*, SIAM Journal on Optimization, 26 (2016), pp. 2355–2377.

[16] N. Boumal, T. Bendory, R. R. Lederman, and A. Singer, *Heterogeneous multireference alignment: A single pass approach*, in 2018 52nd Annual Conference on Information Sciences and Systems (CISS), IEEE, 2018, pp. 1–6.

[17] V.-E. Brunel, *Learning rates for gaussian mixtures under group action*, in Conference on Learning Theory, 2019, pp. 471–491.

[18] G. Casella and R. Berger, *Statistical Inference*, Duxbury advanced series, Duxbury Thomson Learning, 2 ed., 2002.

[19] L. De Haan and A. Ferreira, *Extreme value theory: an introduction*, Springer Science & Business Media, 2007.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), 39 (1977), pp. 1–22.

[21] J. J. Donatelli, P. H. Zwart, and J. A. Sethian, *Iterative phasing for fluctuation X-ray scattering*, Proceedings of the National Academy of Sciences, 112 (2015), pp. 10286–10291.

[22] Z. Fan, Y. Sun, T. Wang, and Y. Wu, *Likelihood landscape and maximum likelihood estimation for the discrete orbit recovery model*, arXiv preprint arXiv:2004.00041, (2020).

[23] N. Goyal, S. Vempala, and Y. Xiao, *Fourier PCA and robust tensor decomposition*, in Proceedings of the forty-sixth annual ACM symposium on Theory of computing, 2014, pp. 584–593.

[24] R. A. Harshman, *Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis*, (1970).

[25] M. Hirn and A. Little, *Wavelet invariants for statistically robust multi-reference alignment*, arXiv

preprint arXiv:1909.11062, (2019).
[26] Z. Kam, *The reconstruction of structure from electron micrographs of randomly oriented particles*, Journal of Theoretical Biology, 82 (1980), pp. 15–39.
[27] A. Katsevich and A. Bandeira, *Likelihood maximization and moment matching in low SNR Gaussian mixture models*, arXiv preprint arXiv:2006.15202, (2020).
[28] I. Land, S. Huettinger, P. A. Hoeher, and J. B. Huber, *Bounds on information combining*, IEEE Transactions on Information Theory, 51 (2005), pp. 612–619.
[29] S. E. Leurgans, R. T. Ross, and R. B. Abel, *A decomposition for three-way arrays*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 1064–1083.
[30] E. Levin, T. Bendory, N. Boumal, J. Kileel, and A. Singer, *3D ab initio modeling in cryo-EM by autocorrelation analysis*, in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1569–1573.
[31] C. Ma, T. Bendory, N. Boumal, F. Sigworth, and A. Singer, *Heterogeneous multireference alignment for images with application to 2D classification in single particle reconstruction*, IEEE Transactions on Image Processing, 29 (2019), pp. 1699–1710.
[32] A. Perry, J. Weed, A. S. Bandeira, P. Rigollet, and A. Singer, *The sample complexity of multireference alignment*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 497–517.
[33] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, *Message-passing algorithms for synchronization problems over compact groups*, Communications on Pure and Applied Mathematics, 71 (2018), pp. 2275–2322.
[34] Y. Polyanskiy and Y. Wu, *Lecture notes on information theory*, (2019). http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf.
[35] T. Pumir, A. Singer, and N. Boumal, *The generalized orthogonal Procrustes problem in the high noise regime*, arXiv preprint arXiv:1907.01145, (2019).
[36] E. Romanov and M. Gavish, *The noise-sensitivity phase transition in spectral group synchronization over compact groups*, Applied and Computational Harmonic Analysis, (2019).
[37] S. H. Scheres, *RELION: implementation of a Bayesian approach to cryo-EM structure determination*, Journal of structural biology, 180 (2012), pp. 519–530.
[38] N. Sharon, J. Kileel, Y. Khoo, B. Landa, and A. Singer, *Method of moments for 3D single particle ab initio modeling with non-uniform distribution of viewing angles*, Inverse Problems, 36 (2020), p. 044003.
[39] F. J. Sigworth, *A maximum-likelihood approach to single-particle image refinement*, Journal of structural biology, 122 (1998), pp. 328–339.
[40] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Applied and computational harmonic analysis, 30 (2011), pp. 20–36.
[41] A. Singer, *Mathematics for cryo-electron microscopy*, Proceedings of the International Congress of Mathematicians, (2018).
[42] A. Singer and Y. Shkolnisky, *Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming*, SIAM journal on imaging sciences, 4 (2011), pp. 543–572.
[43] I. Sutskover, S. Shamai, and J. Ziv, *Extremes of information combining*, IEEE Transactions on Information Theory, 51 (2005), pp. 1313–1325.
[44] R. van Handel, *Probability in high dimension*, tech. report, PRINCETON UNIV NJ, 2014.
[45] L. Wang, S. Hu, and O. Shayevitz, *Quickest sequence phase detection*, IEEE Transactions on Information Theory, 63 (2017), pp. 5834–5849.