Foundations of Computational Mathematics (2023) 23:1851–1898
https://doi.org/10.1007/s10208-022-09584-6

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Sparse Multi-Reference Alignment: Phase Retrieval, Uniform Uncertainty Principles and the Beltway Problem

Subhroshekhar Ghosh[1] · **Philippe Rigollet**[2]

Received: 24 June 2021 / Revised: 10 March 2022 / Accepted: 23 June 2022 /
Published online: 1 August 2022

Abstract

Motivated by cutting-edge applications like cryo-electron microscopy (cryo-EM), the Multi-Reference Alignment (MRA) model entails the learning of an unknown signal from repeated measurements of its images under the latent action of a group of isometries and additive noise of magnitude σ. Despite significant interest, a clear picture for understanding rates of estimation in this model has emerged only recently, particularly in the high-noise regime $\sigma \gg 1$ that is highly relevant in applications. Recent investigations have revealed a remarkable asymptotic sample complexity of order σ^6 for certain signals whose Fourier transforms have full support, in stark contrast to the traditional σ^2 that arise in regular models. Often prohibitively large in practice, these results have prompted the investigation of variations around the MRA model where better sample complexity may be achieved. In this paper, we show that *sparse* signals exhibit an intermediate σ^4 sample complexity even in the classical MRA model. Further, we characterize the dependence of the estimation rate on the support size s as $O_p(1)$ and $O_p(s^{3.5})$ in the dilute and moderate regimes of sparsity respectively. Our techniques have implications for the problem of *crystallographic phase retrieval*, indicating a certain local uniqueness for the recovery of sparse signals from their power spectrum. Our results explore and exploit connections of the MRA estimation problem with two classical topics in applied mathematics: the *beltway problem* from combinatorial optimization, and *uniform uncertainty principles* from harmonic analysis. Our

Communicated by Rachel Ward.

✉ Subhroshekhar Ghosh
subhrowork@gmail.com

Philippe Rigollet
rigollet@math.mit.edu

[1] Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076, Singapore

[2] Department of Mathematics, Massachusetts Institute of Technology, 182 Memorial Drive, Cambridge, MA 02142, USA

techniques include a certain enhanced form of the probabilistic method, which might be of general interest in its own right.

Keywords Multi reference alignment · Cryo electron microscopy · Phase retrieval · Uncertainty principles · Fourier analysis · Beltway problem · Combinatorial optimization · Sparse signal processing · Sample complexity · Probabilistic method

Mathematics Subject Classification 62F12 · 62R99 · 68Q32 · 68Q87 · 05B10 · 42A16 · 42A61 · 94A15 · 94A12 · 92E10

1 Introduction

1.1 The MRA Problem

The Multi-Reference Alignment (MRA) problem is a simple model that captures fundamental characteristics of various statistical models with latent group actions. It arises in various questions across science and engineering such as structural biology [22, 38, 39, 49, 50, 56], image recognition [17, 23, 30, 45], robotics [47] and signal processing [31, 59]. This problem also serves as a simplification for more complex ones that feature repeated observations of a signal subject to latent group actions and additive measurement noise. Such problems include, for example, the three-dimensional reconstruction of molecules using cryo-electron microscopy (cryo-EM) [6, 40, 53]. Such models have gained salience in recent years with the remarkable growth in the scope and capabilities of data-intensive procedures in science and technology.

The MRA problem [3, 40, 44] consists in a signal $\theta : \mathbb{Z}_L \mapsto \mathbb{R}$ (equivalently, a vector $\theta \in \mathbb{R}^L$), and n independent noisy observations $y_1, \ldots, y_n$ that satisfy

$$y_i = R_i\theta + \sigma\xi_i, \tag{1.1}$$

where the R_i-s are isometries of $\mathbb{R}^L$, and the random variables ξ_i are i.i.d. L-dimensional standard Gaussians $N(0, I_L)$, and $\sigma > 0$ is the scale of the noise. The R_i-s are taken random, sampled from the group of cyclic shifts $\mathcal{G}$ on $\mathbb{R}^L$, and are independent as random variables from the noise $\{\xi_i\}_i$.

The group of cyclic co-ordinate shifts is given by $(R_\ell\theta)_k = \theta_{k+\ell \,(\mathrm{mod} L)}$, where $(v)_k$ denotes the kth co-ordinate for a vector $v \in \mathbb{R}^L$. The canonical distribution for the isometries R_i is uniform on the group $\mathcal{G}$, although other distributions have been considered [1].

Of course, due to the latent group actions, it is not possible to recover θ exactly. Instead, our goal is to obtain an estimator $\tilde{\theta}$ whose distance to the orbit of θ under the action of the group $\mathcal{G}$, as defined by

$$\varrho(\tilde{\theta}, \theta) := \min_{G\in\mathcal{G}} \frac{1}{\sqrt{L}} \|\tilde{\theta} - G\theta\|_2 \tag{1.2}$$

is typically small with growing number of samples n.

On a related note, we also define the distance ρ below, which will enable us to invoke results from the literature on the MRA model.

$$\rho(\theta, \varphi) := \min_{G \in \mathcal{G}} \|\theta - G\varphi\|_2 \tag{1.3}$$

Observe that $\varrho(\theta, \varphi) = \frac{1}{\sqrt{L}}\rho(\theta, \varphi)$; so results in the two metrics are simple scalings of each other by a factor of $\sqrt{L}$.

In this work, we focus on the statistical performance achievable in the MRA model. Of key interest is the dependence on the typical behaviour of ϱ on the quantities n and σ for the asymptotics $n, \sigma \to \infty$ which are well justified by applications such as molecular spectroscopy [40, 52]. In this regime, statistical rates of convergence are of the form

$$\mathbb{E}\varrho(\tilde{\theta}, \theta) \leq C(L, s)\frac{\sigma^\alpha}{\sqrt{n}}$$

where α is an exponent that critically controls the performance of $\tilde{\theta}$ in the regime of interest [4, 40]. We also keep track of other important quantities such as the dimension L or the sparsity s of the signal θ but only insofar as they appear in leading terms. Dual to the above *rate*, one may consider the *sample complexity* of $\tilde{\theta}$, which is the number n of samples required to achieve accuracy ε. Equating the right-hand side of the above display with ε and solving for n yields a sample complexity of $\sigma^{2\alpha}/\varepsilon^2$. In this work, we always achieve parametric rates where the dependence on n is $n^{-1/2}$ and hence, the sample complexity thus scales as ε^{-2} in ε. As a result, we refer to $\sigma^{2\alpha}$ as the sample complexity of $\tilde{\theta}$.

While the MRA problem has been mostly attacked using the synchronization approach [3], it is only recently that it was recognized as a Gaussian mixture model [4] which has enabled the use of various methods such as the method of moments [16, 40] and expectation-maximization [11] to recover the signal of interest f. For a detailed discussion on the likelihood landscape of such models, we refer the reader to [18, 25, 26, 34].

Using the Gaussian structure of the noise, it is straightforward to write down an expression for the likelihood function for given observations $\{y_1, \ldots, y_n\}$, where the likelihood function is parametrized by the signal parameter θ. If the density corresponding to θ for an observation y is given by p_θ, then we can write

$$p_\theta(y) = \frac{1}{|\mathcal{G}|}\sum_{R \in \mathcal{G}} \frac{1}{(\sqrt{2\pi}\sigma)^L} \exp\left(-\frac{\|y - R\theta\|_2^2}{2\sigma^2}\right) \tag{1.4}$$

and the log likelihood corresponding to the data $\{y_1, \ldots, y_n\}$ as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i). \tag{1.5}$$

The perspective of Gaussian mixture models has enabled the discovery of a singular statistical phenomenon due to the presence of the latent isometries [4, 40]. To recall this result, we introduce some notation.

We consider vectors in $\mathbb{R}^L$ as functions mapping $\mathbb{Z}_L$ to $\mathbb{R}$, and consider $\mathbb{Z}_L$ in the standard parametrization 1.7. Let $\hat{\theta} \in \mathbb{R}^L$ denote the (discrete) Fourier transform of θ, also considered as a function $\hat{\theta} : \mathbb{Z}_L \mapsto \mathbb{R}$, where $\mathbb{Z}_L$ is viewed in the standard parametrization. Since the signal θ is real, $\hat{\theta}$ is symmetric about the origin. We define the positive support of $\hat{\theta}$ by

$$\mathrm{psupp}(\hat{\theta}) = \{j \mid j \in \{1, \dots, \lfloor (L-1)/2 \rfloor\}, \hat{\theta}_j \neq 0\}\,.$$

To prohibit $\hat{\theta}$ to scale with the sample size n, it is reasonable to assume that there exists two positive constants c and c_0, such that $c^{-1} \leq \|\theta\| \leq c$ and $|\hat{\theta}_j| \geq c_0$ for all $j \in \mathrm{psupp}(\hat{\theta})$. The group action under consideration is the group of shifts $\mathbb{Z}_L$.

We now discuss the minimax lower bound proved in [4], which is also shown to give the optimal rate. To be precise, the results in [4] are stated in the setting of the closely related phase-shift model (essentially, a continuous version of the MRA), but the broad implications of the result are understood to also capture the behaviour of the MRA model. [4] gives us the following minimax lower bound on the estimation error.

Theorem 1 [4, Theorem 1] *Let $2 \leq s \leq L/2$. Let $\mathcal{T}_s$ be the set of vectors $\theta \in \mathcal{T}$ satisfying $\mathrm{psupp}(\hat{\theta}) \subset [s]$. For any $\sigma \geq \max_{\theta \in \mathcal{T}_s} \|\theta\|$, the phase shift model satisfies*

$$\inf_{T_n} \sup_{\theta \in \mathcal{T}_s} \mathbb{E}[\varrho(T_n, \theta)] \gtrsim \frac{\sigma^{2s-1}}{\sqrt{n}} \wedge 1\,, \tag{1.6}$$

where the infimum is taken over all estimators T_n of θ.

A careful inspection of the proof of this lower bound indicates that it is in fact driven by specific cancellations of the Fourier coefficients of θ. Indeed, if $\mathrm{psupp}(\hat{\theta}) \subset [(L-1)/2]$, there exists specific sparsity patterns for the Fourier transform of θ that make it hard to estimate: in this case, Theorem 1 indicates a worst-case lower bound with a terrible sample complexity: σ^{2L-2}. This result is mitigated in [40] where it is proved that if $\mathrm{psupp}(\hat{\theta}) = [L/2]$, that is if $\hat{\theta}$ has *full support*—recall that we assumed $|\hat{\theta}_j| \geq c_0$ for all $j \in \mathrm{supp}(\hat{\theta})$—then a sample complexity of σ^6 may be achieved. The unusual exponent $6 = 2 \cdot 3$ comes from the fact that in this case, the orbit of θ may be identified from the first three moments of Y.

While σ^6 is a significant improvement over σ^{2L-2}, this scaling is still inauspicious in applications where σ is large [notice that the dependence of sample complexity on σ scales like the square of that of the estimation rate as in (1.6)]. This situation has prompted the investigation of settings where the orbit of θ could be recovered robustly only from its first two moments, thus leading to the a sample complexity σ^4. This is the case for example if $\hat{\theta}$ has full support and the distribution of the isometries on the group $\mathcal{G}$ is *not uniform* but follows some specific distribution instead [1].

In this paper, we unveil generic conditions on the signal θ under which a sample complexity of σ^4 can be achieved in the original MRA model, where the distribution

of isometries from the group $\mathcal{G}$ *is uniform*. Interestingly, these results are built on connections with other well-studied questions in applied mathematics, in particular the *beltway problem* from combinatorial optimization and *uniform uncertainty principles* from harmonic analysis.

In methodological terms, obtaining a $\sigma^2/\sqrt{n}$ rate will be found to be related to our ability to recover a signal from the second moment tensor, and in turn, from the modulus of the Fourier coefficients of its observations in the MRA model. This will eventually be made possible by the sparsity of the signal. In a related vein, we note that the well-known problem of phase retrieval, albeit in a different context, examines signal recovery from the modulus of its random linear measurements. However, it may be noted that our observational setting in the MRA model with the latent group action has a very different and more complicated structural setup than the existing literature on phase retrieval, which largely focuses on a specific setting akin to compressed sensing with limited information. Nonetheless, there are natural connections to phase retrieval, especially to the so-called *crystallographic phase retrieval* problem [12]; this is discussed in detail in Sect. 1.4.

1.2 Estimation Rates for Generic Sparse Signals

In the present work, we focus attention on the class of sparse signals in the context of the MRA problem, and explore rates if estimation for such signals. Our investigations naturally demarcate the set of sparse signals into two regimes, marked by results of differing nature.

On the one hand, we have the *dilute* regime of sparsity (roughly, of the order $L^{1/3}$), where a randomly chosen subset of $\mathbb{Z}_L$ of that size does not have any multiplicities in its mutual differences. This condition is referred to as the collision-free property of the subset. In the dilute regime of sparsity, we establish $O(\sigma^4)$ sample complexity. This is complemented by the *moderate* regime of sparsity, which extends all the way up to order $L/\log^5 L$ where we show that the improved sample complexity may also be achieved.

We also unveil the dependence structure of the estimation rate asymptotics on the sparsity s of the signal. In the dilute regime, there is an $O_p(1)$ dependency, whereas in the moderate regime of sparsity, the dependency is $O_p(s^{3.5})$. Observe that we are considering asymptotic rates of estimation which are by nature local to the true signal; this is different from non-asymptotic rates which usually involve additional logarithmic dependence on L.

The relative difficulty in obtaining $O(\sigma^4)$ sample complexity with increasing size of the signal support is reflected in the dependence structure of the asymptotics on the sparsity, as well as in the additional assumptions required in the moderate regimes. Such behaviour is perhaps well anticipated in view of the fact that, in the regime of full signal support, sample complexity of order better than σ^6 generically not possible, a result which we also establish in this work.

In the dilute regime, our methodological ingredients include exploiting collision-freeness, whereas in the moderate regime they include repeated, nested applications

of the probabilistic method in order to find frequency sets conducive to our analysis in the Fourier space, aided by the tool of uniform uncertainty principles.

In order to discuss our results in detail, we first introduce a few notations and concepts.

1.2.1 Some Notations and Concepts

In this work, we will set $\mathcal{G}$ to be the group of rotations by the elements of $\mathbb{Z}_L$, that is, for each $g \in \mathcal{G}$ and $v : \mathbb{Z}_L \mapsto \mathbb{C}$, we define the action $[g \cdot v]$ as $[g \cdot v](i) = v(i + g) \quad \forall i \in \mathbb{Z}_L$.

We note in passing that the results of this paper would also hold under the action of a richer group of isometries $\mathcal{G}$ where the rotations of $\mathbb{Z}_L$ are augmented with a reflection or "flip", that is, the operation α acting on $\mathbb{Z}_L$ that sends $x \mapsto \check{x} = -x$; in other words, the group $\mathcal{G} = \mathbb{Z}_L \rtimes \mathbb{Z}_2$. In fact, there has been recent interest focussed on *dihedral* multi-reference alignment [13]. However, for purposes of presentation, we will adhere to the isometry group $\mathcal{G}$ given by the rotations of $\mathbb{Z}_L$.

We view the signal θ as a function on the discretized circle $\mathbb{Z}_L$, where the elements of the latter are enumerated as

$$\mathbb{Z}_L = \{\lfloor -(L-1)/2 \rfloor, \lfloor -(L-1)/2 \rfloor + 1, \ldots, \lfloor (L-1)/2 \rfloor - 1, \lfloor (L-1)/2 \rfloor\} \tag{1.7}$$

We call this parametrization the *standard parametrization* of $\mathbb{Z}_L$. The *positive* part of $\mathbb{Z}_L$ may then be defined as

$$\mathbb{Z}_L^+ = \{0, 1, \ldots, \lfloor (L-1)/2 \rfloor - 1, \lfloor (L-1)/2 \rfloor\}. \tag{1.8}$$

We include here a discussion on restricted MLEs, which will constitute the main estimators in describing our statistical results. For a deterministic set of signals $\mathcal{T}$ (where the true signal is known to belong), it is natural to maximize the log likelihood (1.5) over $\theta \in \mathcal{T}$. We refer to such MLEs as restricted MLE. In the setting where the signal is sampled from generative models, we consider MLEs restricted to signal classes $\mathcal{T}$ that are events of high probability under the respective generative model (as relevant model parameters tend to ∞). For more on the relationship between deterministic classes of *good* signals and generative models, we direct the reader to Remark 11.

Further notations and concepts used in this paper that are of a more generic nature can be found in the Appendix A.

1.2.2 The *Dilute* Regime: Sparse Collision-Free signals

We first define the notion of collision-free property of the support of a signal, and subsequently use it for introducing the appropriate signal class for the dilute regime, which, roughly speaking, consists of signals that can at best be of size $O(L^{1/2})$ and typically of size $O(L^{1/3})$.

Definition 2 For a vector $v \in \mathbb{R}^L$, we will denote by $\mathcal{D}(v)$ the (multi-)set of differences $\{v(i) - v(j) : 1 \le i, j \le d\}$. In general, this is a set of differences with multiplicities. In case the multiplicity is exactly 1 for each difference appearing in $\mathcal{D}(v)$, we call the vector v *collision-free*, that is there are no repeated differences in its support.

Notice that being collision-free is really a property of the support $\mathrm{supp}(v)$ of the vector v.

We are now ready to define the signal class that we will investigate in the dilute regime.

Definition 3 We consider the set $\mathcal{T} \subset \mathbb{R}^L$ to consist of the signals $\theta : \mathbb{Z}_L \mapsto \mathbb{R}$ that satisfy the conditions outlined below.

(i) θ is collision free
(ii) There exist positive numbers $m, M, \varepsilon > 0$ (uniform for the set $\mathcal{T}$) such that $m \le |\theta(i)| \le M$ on $\mathrm{supp}(\theta)$, and $s := |\mathrm{supp}(\theta)| \ge (2+\varepsilon)M^2/m^2$.

We can then state the following theorem.

Theorem 4 *Let $\mathcal{T}$ be the set of signals as in Definition 3. Then for σ bigger than a threshold $\sigma_0(L)$, for any signal $\theta_0 \in \mathcal{T}$, the restricted MLE $\tilde{\theta}_n$ for the MRA problem satisfies $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = O_p(\sigma^2)$ as $n \to \infty$.*

A crucial ingredient in the proof of Theorem 4 is the following curvature lower bound on the second moment tensor, which we state below as a result of independent interest.

Lemma 5 *Let $\mathcal{T}$ be the set of vectors $\theta \in \mathbb{R}^L$ as in Definition 3. Then, for any $\theta, \theta_0 \in \mathcal{T}$, we have*

$$\|\mathbb{E}_G[(G\theta)^{\otimes 2}] - \mathbb{E}_G[(G\theta_0)^{\otimes 2}]\|_F \ge \sqrt{\frac{2\varepsilon}{2+\varepsilon}} \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{s} \cdot \rho(\theta, \theta_0)$$

for some universal constant c.

The collision-free property of the support of the signal, as enunciated in this section, is typically associated with the signal being considerably sparse; hence the name *dilute* regime. In fact, it may be shown that for the signal support to be collision-free, the size s of the support cannot exceed $O(L^{1/2})$. On the other hand, it can also be shown that typical subsets (e.g., chosen uniformly at random from the co-ordinates) the collision free property holds with high probability if $s = o(L^{1/3})$. We refer the reader to Appendix D for further on these size bounds.

1.2.3 The *Moderate* Regime: Generic Sparse Symmetric Signals

In this section, we demonstrate that we can extend far beyond the dilute regime and obtain a sample complexity of σ^4 for generic symmetric signals in the so-called *moderate* regime of sparsity, extending all the way up to support size $s = O(L/\log^5 L)$. In doing so, we invoke uncertainty principles from Fourier analysis as an effective technique for the studying MRA problem.

To this end, we define the notion of the Bernoulli–Gaussian distribution, and the symmetric version thereof. The Bernoulli–Gaussian distribution is a popular model for modelling generic or typical sparse signals in statistics and signal processing [35, 54]. In this work, we use the symmetric Bernoulli–Gaussian distribution in order to model sparse symmetric signals for investigating estimation rates under the MRA model. Such symmetry hypothesis is well-motivated by the fact that many natural objects of interest, such as molecules, exhibit symmetries that are of significance in spectroscopy [20, 43, 58]; this includes reflection symmetries that are related to the important notion of chirality [5, 19].

A signal following the Bernoulli–Gaussian distribution with variance ζ^2 and sparsity s consists in generating the signal support via independent random sampling of points in $\mathbb{Z}_L$ with probability s/L each, and then independently generating the signal values on the support via a $N(0, \zeta^2)$ distribution for each point. The symmetric Bernoulli–Gaussian distribution differs from the general case defined above only in the fact that its support is constrained to be symmetric. To obtain this, we consider $\mathbb{Z}_L$ in its standard parametrization (1.7), and pick the *positive* part A_+ of the support by independent random sampling from $\mathbb{Z}_L^+$ with probability s/L, and then obtain the full symmetric support A via reflection about the origin, i.e. $A = A_+ \cup (-A_+)$.

While the Bernoulli–Gaussian distribution is standard for modelling sparse signals, our results apply to far more general signal classes. This includes the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ distribution, which entails that the support is $[-s, s]$ and the signal values are independent $N(0, \zeta^2)$ random variable. In fact, other than independent Gaussian values, our results only require that the signal support be sparse and sufficiently *generic*, in a precise arithmetic sense that we call *cosine genericity*.

We call such signals *generic sparse symmetric signals*. Our main estimation rate results will be stated below in terms of this signal class; the precise and detailed definitions for it are provided in Appendix C.

Theorem 6 *Let* $\log^9 L \leq s \leq L/\log^5 L$. *Consider a generic s-sparse symmetric signal* $\theta_0 : \mathbb{Z}_L \mapsto \mathbb{R}$, *with dispersion* ζ^2, *sparsity constants* (α, β) *and index* $\tau > 0$. *Then for* σ *bigger than a threshold* $\sigma_0(L)$, *with high probability in* θ_0, *the restricted MLE* $\tilde{\theta}_n$ *for the MRA problem satisfies* $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = O_p(\sigma^2)$ *as* $n \to \infty$.

Theorem 6 enables us to deduce an improved sample complexity of order σ^4 for signals sampled from the symmetric Bernoulli–Gaussian distribution.

Corollary 7 *Let* $\log^9 L \leq s \leq L/\log^5 L$. *Consider a signal* θ_0 *sampled from either of:*

(i) *The symmetrized Bernoulli–Gaussian distribution on* $\mathbb{Z}_L$ *with mean zero, variance* ζ^2 *and sparsity* s, *or*
(ii) *The* $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ *distribution.*

Then, for σ *bigger than a threshold* $\sigma_0(L)$, *with high probability in* θ_0, *the restricted MLE* $\tilde{\theta}_n$ *for the MRA problem satisfies* $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = O_p(\sigma^2)$ *as* $n \to \infty$.

1.2.4 Dependence on Sparsity and Ambient Dimension

In important signal classes, we can obtain the dependence of asymptotic estimation rates on the parameters (s, L). We record them in the following result.

Theorem 8 *Let θ_0 be the signal in the MRA model, and σ bigger than a threshold $\sigma_0(L)$.*

(i) *If θ_0 is as in Definition 3 with $|\operatorname{supp}(\theta_0)| = s$, then we have the $\lim_{n\to\infty} \sqrt{n}\varrho(\tilde{\theta}_n, \theta_0)/\sigma^2 = O_p(1)$ as a function of s, L.*

(ii) *If θ_0 is sampled from the symmetric Bernoulli–Gaussian distribution with mean 0, variance ζ^2 and sparsity parameter s with $\log^9 L \leq s \leq L/\log^5 L$, then with high probability in θ_0, we have $\lim_{n\to\infty} \sqrt{n}\varrho(\tilde{\theta}_n, \theta_0)/\sigma^2 = O_p(s^{3.5})$ as a function of s, L.*

We emphasize here that our rate bounds are asymptotic in n, and therefore necessarily *local* in character, focusing on a small enough neighbourhood of the signal (that will generally depend on L). In contrast, non-asymptotic bounds that are generally *global* over the set of allowable signals, and therefore will usually exhibit an L dependence, such as the standard $\sqrt{\log L}$ dependence in much of the signal processing literature. On a related note, in this work we concern ourselves with the leading, dominant term in an asymptotic expansion of $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0)/\sigma^2$ in the regime of large L; higher order terms in this expansion will generically depend on L.

We observe that the asymptotic upper bound in the dilute regime is $O_p(1)$, but in the regime of moderate sparsity it is growing as $s^{3.5}$. This is perhaps to be expected, in tune with the fact that a σ^2 dependence of the estimation rate eventually breaks down for signals with full support.

It is of interest to make explicit the role of the ambient dimension L as a quantifier in the main results of this paper. The main results, such as Theorems 6, 8 and 4 and Corollary 7, entail statements regarding generic signals. This genericity refers to the fact that the statements of these results hold for a set of signals that, under suitable distributions on the signal space (whose specifics are clarified for each result), has probability at least $1 - p(L)$, where $p(L) \to 0$ as L tends to infinity. While a precise bookkeeping of our arguments would indeed yield explicit formulae for such sequences $p(L)$ for the relevant distributions considered in this paper, we prefer not to pursue that route so as to maintain brevity, given that the statements in their present forms already capture the main qualitative phenomena and key dependencies.

Thus, our results are in particular not asymptotic in L: indeed, the results hold for each L (bigger than some threshold L_0) for a class of signals $\mathcal{S}(L)$ that depends on L. As L grows, the probability measure of $\mathcal{S}$ under a natural distribution converges to 1. However, the results do have a clear interpretation even without letting $L \to \infty$, which is the point of view we take in this paper. In the recent work [46], the authors undertake an examination of the MRA problem in the *high dimensional* regime, where $L, \sigma \to \infty$ jointly with the special parametric dependence $\sigma = L/\alpha \log L$ for a parameter $\alpha > 0$. In such a situation, the order of the limits $n, L \to \infty$ becomes important. In the present work, we interpret the results as above—with fixed L and $n \to \infty$. Thus, the results in this paper are of a different flavour and indeed not comparable to the high dimensional scenario as in [46]. A synthesis of these two points of views, however, remains an interesting question for future work.

1.2.5 A Technical Lemma

A useful ingredient in the proof of Theorem 6 is a deterministic technical lemma which we state below.

Lemma 9 *Consider the set of signals $\mathcal{T}$ on $\mathbb{Z}_L$ in the standard enumeration (1.7), defined as follows. For each $\theta \in \mathcal{T}$, we have*

(i) *θ is symmetric, i.e. $\theta(i) = \theta(-i) \forall i \in \mathbb{Z}_L$.*
(ii) *There exist $m_{\mathcal{T}}, M_{\mathcal{T}} > 0$, uniform in θ, such that $m_{\mathcal{T}} \leq |\theta(i)| \leq M_{\mathcal{T}}$ on $\mathrm{supp}(\theta)$.*
(iii) *There exists $\Lambda \subset \mathbb{Z}_L$, possibly depending on θ, such that:*
  (a)

$$c_1 \cdot \frac{1}{L}\|h\|_2^2 \leq \frac{1}{|\Lambda|}\sum_{\xi \in \Lambda} |\hat{h}(\xi)|^2 \leq c_2 \cdot \frac{1}{L}\|h\|_2^2$$

  for all h with $\mathrm{supp}(h) \subseteq \mathrm{supp}(\theta)$, with positive constants c_1, c_2 uniform in h, θ.
  (b) *$\min_{\xi \in \Lambda} |\hat{\theta}(\xi)| \geq \mathfrak{m}(\mathcal{T})$, for some $\mathfrak{m}(\mathcal{T}) > 0$ that is uniform in θ.*

Then, for σ bigger than a threshold $\sigma_0(L)$ and any signal $\theta_0 \in \mathcal{T}$, the restricted MLE $\tilde{\theta}_n$ satisfies $\sqrt{n}\varrho(\hat{\theta}_n, \theta_0) = O_p(\sigma^2)$ as $n \to \infty$.

Remark 10 A significant setting in the context of the conditions (i)–(iii) above is when the signal class $\mathcal{T}$ is sparse with typical support size s, and $m_{\mathcal{T}}$, $M_{\mathcal{T}}$, $\mathfrak{m}(\mathcal{T})$ as well as the constants c_1, c_2 all depend only on the sparsity s and not on L. This is, in fact, true for both the sparse symmetric Bernoulli–Gaussian distribution and the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ distributions that we consider in this paper, and lead to estimation rates that depend asymptotically only on the sparsity and not on the ambient dimension.

Remark 11 We observe that $\mathcal{T}$ in Lemma 9 is a deterministic subset of the space of signals, whereas Theorem 6 and Corollary 7 consider a typical signal from certain generative models. We can, however, easily reconcile the two by considering a large, compact set $\mathcal{T}$ of signals that carries a high probability measure under the respective generative distribution and satisfies the conditions of Lemma 9. This allows us to deduce improved rates of estimation for *typical* signals under generative distributions by making use of the deterministic Lemma 9.

1.2.6 Estimation Rates Beyond Sparsity

Theorems 4 and 6 establishes a sample complexity of order σ^4 for signals with sparse support. Complementary to Theorem 4, in this section we examine classes of signals with very different structural properties compared to sparse support, and show that in such settings, $\sigma^2/\sqrt{n}$ rates of estimation is generically impossible.

Theorem 12 *Let $\mathcal{T}$ be the set of vectors $\theta \in \mathbb{R}^L$ is such that its support $\mathrm{supp}(\theta)$ is all of $\mathbb{Z}_L$ and $m \leq |\theta(i)| \leq M$ on $\mathrm{supp}(\theta)$. Then, for any signal $\theta_0 \in \mathcal{T}$ the restricted MLE $\hat{\theta}_n$ satisfies $\sqrt{n}\varrho(\hat{\theta}_n, \theta_0) = \Omega_p(\sigma^3)$ as $n \to \infty$.*

This will follow from the following Lemma, to state which we need to introduce the following notation. For a signal $\varphi = (\varphi^{(1)}, \dots, \varphi^{(L)}) \in \mathbb{R}^L$, we define the average

$$\overline{\varphi} := \frac{1}{L} \sum_{i=1}^{L} \varphi^{(i)}$$

Lemma 13 *Let $\mathcal{T}$ be the set of vectors $\theta \in \mathbb{R}^L$ is such that its support* $\mathrm{supp}(\theta)$ *is all of $\mathbb{Z}_L$ and $m \leq |\theta(i)| \leq M$ on* $\mathrm{supp}(\theta)$*. Then, there exist a sequence $\{\theta_k\}_{k>0} \subset \mathcal{T}$ such that $\rho(\theta_k, \theta_0) \to 0$ as $k \to \infty$, $\overline{\theta}_k = \overline{\theta}_0$, and we have*

$$\|\mathbb{E}_G[(G\theta_k)^{\otimes 2}] - \mathbb{E}_G[(G\theta_0)^{\otimes 2}]\|_F \leq C\varrho(\theta_k, \theta_0)^2. \tag{1.9}$$

We may compare the statement of Lemma 13 with that of Lemma 5, and notice that the lower bound in Lemma 5 is linear in $\varrho(\theta, \theta_0)$; whereas the upper bound in Lemma 13 is quadratic in $\varrho(\theta, \theta_0)$.

In summary, attaining improved rates of estimation via the second moment is not possible when the signal has full support.

1.3 Main Ideas and Ingredients

Our investigation of the asymptotic estimation rates for the MRA model connects to several classical topics in applied mathematics, including the beltway problem related to combinatorial optimization and uniform uncertainty principles from harmonic analysis.

1.3.1 The Beltway Problem

The beltway problem consists in recovering a set $\mathcal{S}$ of numbers from their pairwise differences $\mathcal{D}$, up to the trivial symmetry of translating all the numbers in the set by the same amount. It is closely related the so-called *turnpike problem* or the *partial digest problem*, and is of interest in computational biology, where it arises naturally in DNA restriction site analysis among other problems. A set of integers is said to be *collision-free* if all the pairwise distances obtained from that set are distinct. In 1939, Piccard [41] conjectured that, if two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ of integers have the same set of pairwise differences $\mathcal{D}$, and the pairwise differences are known to be unique (i.e., $\mathcal{S}_1$ and $\mathcal{S}_2$ are collision-free), then the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ must be translates of each other.

Following major advances by Bloom [14], a description of the complete landscape of Piccard's conjecture was obtained by Bekir and Golomb [8, 9], who demonstrated in particular that the conjecture is true for all sets of cardinality greater than 6.

The upshot of these developments is that if $\mathcal{S}$ is a set of integers with $|\mathcal{S}| \geq 7$ and such that the pairwise distances of the numbers in $\mathcal{S}$ are distinct (in other words, $\mathcal{S}$ is collision-free), then the set $\mathcal{S}$ is uniquely determined (up to translations) by its pairwise distances. This will be exploited in our investigations of the MRA estimation rates. In particular, the beltway problem motivates our definition of the "collision-free"

condition on the support of the signal, which will be used for obtaining improved estimation rates in the dilute regime of sparsity.

1.3.2 Uniform Uncertainty Principles

Uncertainty principles have a long history in harmonic analysis and in applied mathematics, starting from Heisenberg's celebrated Uncertainty Principle in quantum mechanics [27, 29]. Roughly speaking, these entail that a function cannot both be simultaneously localized (i.e., have a 'small support' in an appropriate sense) both in the physical space and the Fourier space. This would imply that for a function with a small support in the physical space (e.g., a sparse signal), the Fourier transform would be overwhelmingly non-vanishing, and therefore we would need almost all of the Fourier coefficients in order to fully capture the 'energy' (i.e. the L^2 norm) of the function, by the Parseval–Plancherel Theorem.

However, if our goal is to approximate the function (up to a limited multiplicative error in the total energy), it may actually suffice to work with a relatively small subset of Fourier coefficients. In fact, such an appropriate set of Fourier coefficients may be obtained via random sampling, and furthermore, such a 'good' set of frequencies may be shown to provide a good approximation simultaneously for all sparse signals. Results in this vein are referred to as *uniform uncertainty principles*; for an expository account we refer the reader to [55] (Chap. 3.2). These are also closely related to the so-called *Restricted Isometry Property* (RIP) for random sub-sampling of Fourier matrices (c.f., [48]).

1.4 Connection to Phase Retrieval

1.4.1 Generalities

Phase retrieval is a central and long-standing question in applied mathematics that find applications in a variety of domains such as astronomy, electron microscopy and optical imaging, and has emerged in recent years as a widely studied question in the field of signal processing [21, 28, 51].

A key connection between the MRA model and the phase retrieval problem arises via second moment tensors. Theorem 16 (Theorem 9, [4]), as well as the related work [40], establishes a clear connection between the $O(\sigma^4)$ sample complexity in the MRA problem and being able to recover the true signal from (estimates of) its second moment tensor, via an expansion of the Kullback–Leibler divergence in terms of moment tensors. The second moment tensor of a signal θ is a circulant matrix related to the convolution $\theta \star \check{\theta}$, whose Fourier transform $\widehat{\theta \star \check{\theta}} = |\hat{\theta}|^2$ as functions.

The problem of (Fourier) phase retrieval entails signal recovery from the modulus of its random linear measurements. This problem has been well-investigated in recent years, as indicated by a substantial literature [7, 10, 32, 33, 37, 42].

The recent work [12] examines the question of recovering a sparse signal from its power spectrum in the context of crystallographic phase retrieval in a non-randomized setting. The connection of such questions to the turnpike problem was considered in

the earlier work [42]. The paper [46] considers the sample complexity of MRA in high dimensions (under a Gaussian prior), exploring in particular the interplay between the dimension and the noise level.

However, as noted earlier, our observational setting in the MRA model with the latent group action has a very different and more complicated structural setup than the existing literature on phase retrieval, which largely focuses on a specific setting akin to compressed sensing with limited information.

The present work focusses on statistical rates of estimation, leaving aside the question of algorithmic implementation for future work. The elaborate literature on phase retrieval, on the other hand, makes a detailed exploration of algorithmic issues, which might be of natural interest in this regard.

1.4.2 Crystallographic Phase Retrieval

The results and methods in the present work have applications to phase retrieval, in particular the problem of crystallographic phase retrieval. The latter problem is believed to be perhaps the most important phase retrieval setup, where the interest is in recovering a sparse signal from its *power spectrum*, i.e., the magnitude of its Fourier transform at various frequencies [12, 24]. This is equivalent to recovering a signal from its *periodic auto-correlation*. The problem is motivated by X-ray crystallography, a technique for determining molecular structure [36].

Formally, let θ_0 be an s-sparse signal in ambient dimension L. In crystallographic phase retrieval, we are interested in recovering θ_0 from the magnitude of its discrete Fourier transform, namely $\{|\hat{\theta}_0(j)|^2\}$ at measurement frequency j. Equivalently, the interest is in recovering θ_0 from its periodic auto correlations, which are given by

$$\mathcal{A}_{\theta_0}(l) = \sum_{i \in \mathbb{Z}_L} \theta_0(i)\theta_0(i + l \bmod L) \quad \text{for } i \in \mathbb{Z}_L.$$

Clearly, the power spectrum of a signal remains invariant under rotations and reflections, and objective is to recover the signal up to these *intrinsic* symmetries.

The theoretical understanding of crystallographic phase retrieval is rather limited (c.f. [12, 42]). In particular, even the fundamental question of uniqueness is poorly understood. In [12], conjectures were laid out regarding the uniqueness of sparse signal recovery from its power spectrum, that predicted in particular that under general conditions unique recovery should be possible when $s/L \leq 1/2$.

Our investigations in this paper have implications for the crystallographic phase retrieval problem, indicating local uniqueness of signal recovery from power spectrum for sparse collision-free signals (i.e., in the dilute regime of sparsity) and for generic sparse symmetric signals with $s = O(L/\log^5 L)$ (i.e., in the moderate regime of sparsity).

This follows from the fact that our $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = O_P(\sigma^2)$ estimation rates for such signals are a consequence of lower bounds on the second moment difference tensors, such as in Lemma 5 and (5.6). The second moment tensor $\mathbb{E}_{\mathcal{G}}[(g \cdot \theta)^{\otimes 2}]$ is a matrix whose entries are precisely the periodic auto correlations $\mathcal{A}_\theta$, so lower bounds such as Lemma 5 and (5.6) indeed demonstrate unique recovery of θ from $\mathcal{A}_\theta$. The

uniqueness is *local*, because our estimation rates are local in character, which entails that lower bounds such as (5.6) hold in a neighbourhood of the true signal θ_0. Further, the uniqueness is among a class of signals that share similar sparsity features as the true signal, e.g. the signal class $\mathcal{T}$ in Lemma 5. Application of the techniques of the present work to obtain more extensive uniqueness guarantees for the crystallographic phase retrieval problem is an interesting problem for future research.

2 Rates of Estimation and Curvature of the KL Divergence

2.1 Estimation Rates and Curvature

In this work, we establish, under very general conditions, quadratic rates of estimation (i.e., scaling as σ^2) in the MRA problem, in the context of Theorem 1 (and (1.6) in particular).

Our point of departure is the population risk of the MRA model, given by

$$R(\theta) = -\mathbb{E}_{p_{\theta_0}}[\log p_\theta(Y)] + C, \tag{2.1}$$

where C is a universal constant. Clearly, we have

$$\begin{aligned} R(\theta) &= -\int \log p_\theta(y) p_{\theta_0}(y) \mathrm{d}y + C \\ &= \int \log\left(\frac{p_{\theta_0}(y)}{p_\theta(y)} \cdot \frac{1}{p_{\theta_0}(y)}\right) p_{\theta_0}(y)\mathrm{d}y + C \\ &= D_{KL}(p_{\theta_0}||p_\theta) - \left(\int p_{\theta_0}(y) \log p_{\theta_0}(y)\mathrm{d}y\right) + C \end{aligned}$$

where $D_{KL}(p_{\theta_0}||p_\theta)$ is the Kullback–Leibler divergence between p_{θ_0} and p_θ. Since θ_0 is fixed, as a function of θ, the population risk $R(\theta)$ equals

$$R(\theta) = D_{KL}(p_{\theta_0}||p_\theta) + C(\theta_0), \tag{2.2}$$

where $C(\theta_0)$ is a function of θ_0.

The Fisher information matrix of the MRA model is given by

$$I(\theta_0) = -\mathbb{E}[\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta_0}] = \nabla_\theta^2 R(\theta_0), \tag{2.3}$$

where ∇_θ^2 denotes the Hessian with respect to the variable θ. It has been demonstrated [2] that the MLE $\tilde{\theta}_n$ is an asymptotically consistent estimate for the true signal θ_0 in the MRA model. This immediately enables us to invoke standard asymptotic normality theory for maximum likelihood estimators and conclude that:

$$\sqrt{n}(\tilde{\theta} - \theta_0) \quad \text{is asymptotically normal with mean 0 and variance} \quad I(\theta_0)^{-1}. \tag{2.4}$$

From the considerations above, we may conclude that the asymptotic covariance is given by $\left(\nabla_\theta^2 D_{KL}(p_{\theta_0}||p_\theta)\right)^{-1}$. For a detailed discussion on such asymptotic normality, we refer the reader to [57], in particular Sects. 5.3 and 5.5 therein.

We observe that the probability distribution p_θ as well as $D_{KL}(p_\theta \| p_\varphi)$ are invariant under the action of $\mathcal{G}$, i.e., invariant under the transformations $\theta \mapsto G \cdot \theta$ for $G \in \mathcal{G}$. As a result, for $\varrho(\theta, \theta_0)$ small enough (equivalently, $\|\theta - \theta_0\|_2$ small enough), we may assume without loss of generality that $\varrho(\theta, \theta_0) = \frac{1}{\sqrt{L}}\|\theta - \theta_0\|_2$ (c.f., [4]; esp. the proof of Theorem 4 therein). Since $\|\tilde{\theta}_n - \theta_0\|_2 \to 0$ as $n \to \infty$, this will be true for $\varrho(\tilde{\theta}_n, \theta_0)$ with high probability.

The upshot of the asymptotic normality discussed above is that, as $n \to \infty$, the quantity $\rho(\tilde{\theta}_n, \theta_0)$ (which equals $\frac{1}{\sqrt{L}}\|\tilde{\theta}_n - \theta_0\|_2$ with high probability), is of the order

$$n^{-1/2}\sqrt{\mathrm{Tr}\left[\frac{1}{L}\cdot I(\theta)^{-1}\right]} = n^{-1/2}\sqrt{\mathrm{Tr}\left[\frac{1}{L}\cdot \nabla_\theta^2 D_{KL}(p_{\theta_0}||p_\theta)^{-1}\right]},$$

where $\mathrm{Tr}[\cdot]$ denotes the trace. This is related to the fact that if $\mathbb{X} \sim N(0, \Sigma)$, then $\mathbb{E}[\|\mathbb{X}\|_2^2] = \mathbb{E}[\mathrm{Tr}(\mathbb{X}^*\mathbb{X})] = \mathrm{Tr}(\Sigma)$.

Thus, the estimation rate for the MRA problem asymptotically depends on σ via the dependence of $\sqrt{\mathrm{Tr}[\frac{1}{L}\cdot\nabla_\theta^2 D_{KL}(p_{\theta_0}||p_\theta)^{-1}]}$ on σ. In view of this, curvature bounds on $D_{KL}(p_{\theta_0}||p_\theta)$ assume significance. We record this in the following proposition.

To this end, we recall the metric $\rho(\cdot,\cdot)$ (1.3), which is essentially a scaling of ϱ: indeed, $\rho(\cdot,\cdot) = \sqrt{L}\varrho(\cdot,\cdot)$.

Proposition 14 *We have the following relations between curvature bounds on D_{KL} and the asymptotic behaviour of $\varrho(\tilde{\theta}_n, \theta_0)$:*

(i) *If $D_{KL}(p_{\theta_0}||p_\theta) \geq K_1(\sigma)\rho(\theta,\theta_0)^2$ for θ in a neighbourhood U of θ_0, then $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = O_p\left(\frac{K_1(\sigma)^{-1/2}}{\sqrt{L}}\right)$.*

(ii) *If $D_{KL}(p_{\theta_0}||p_\theta) \leq K_2(\sigma)\rho(\theta,\theta_0)^2$ for θ in a neighbourhood U of θ_0, then $\sqrt{n}\varrho(\tilde{\theta}_n, \theta_0) = \Omega_p\left(\frac{K_2(\sigma)^{-1/2}}{\sqrt{L}}\right)$.*

We defer the proof of Proposition 14 to Sect. 6.

Curvature bounds as in Proposition 14 would, in particular, be implied by upper and lower bounds on $D_{KL}(p_{\theta_0}||p_\theta)$ in the form of $K(\sigma)\|\theta - \theta_0\|^2$ (valid on some neighbourhood U of θ_0)—in which case, the asymptotic estimation rate in the MRA problem would scale as $\frac{K(\sigma)^{-1/2}}{\sqrt{L}}\cdot\frac{1}{\sqrt{n}}$.

We introduce the difference of the mth moment tensors corresponding to two signals θ, φ:

$$\Delta_m(\theta,\varphi) := \mathbb{E}[(G\theta)^{\otimes m}] - \mathbb{E}[(G\varphi)^{\otimes m}].$$

Furthermore, by the (Frobenius) norm $\|\cdot\|$ for a tensor, we will denote its Hilbert Schmidt norm. In what follows, we will invoke two results from [4], in which we will make use of the distance ρ (c.f. 1.3).

This allows us to state the following results from [4].

Lemma 15 [4][Lemma 8] *If* $\tilde{\theta} = \theta - \mathbb{E}_{\mathcal{G}}[G\theta]$ *and* $\tilde{\varphi} = \varphi - \mathbb{E}_{\mathcal{G}}[G\varphi]$, *then*

$$D_{KL}(p_\theta || p_\varphi) = D_{KL}(p_{\tilde{\theta}} || p_{\tilde{\varphi}}) + \frac{1}{2\sigma^2}\|\Delta_1(\theta,\varphi)\|^2.$$

Theorem 16 [4][Theorem 9] *Let* $\theta, \varphi \in \mathbb{R}^L$ *satisfy* $3\rho(\theta,\varphi) \le \|\theta\| \le \sigma$ *and* $\mathbb{E}[G\theta] = \mathbb{E}[G\varphi] = 0$. *Let* $\Delta_m = \Delta_m(\theta,\varphi) = \mathbb{E}[(G\theta)^{\otimes m}] - \mathbb{E}[(G\varphi)^{\otimes m}]$. *For any* $k \ge 1$, *there exist universal constants* $\underline{C}$ *and* $\overline{C}$ *such that*

$$\underline{C}\sum_{m=1}^{\infty}\frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m}m!} \le D_{KL}(p_\theta||p_\varphi) \le 2\sum_{m=1}^{k-1}\frac{\|\Delta_m\|^2}{\sigma^{2m}m!} + \overline{C}\frac{\|\theta\|^{2k-2}\rho(\theta,\varphi)^2}{\sigma^{2k}}.$$

We now use Lemma 15 and Theorem 16 in order to obtain bounds on $D_{KL}(p_\theta \| p_\varphi)$ that are tailored to our specific requirements in the present paper, focussing mostly on the second moment difference tensor in the context of Theorem 16.

To this end, we consider the following results. For notational simplicity, we will use the notation $D_{KL}(\theta_1 \| \theta_2)$ to denote $D_{KL}(p_{\theta_1} \| p_{\theta_2})$. Further, for any $\theta = (\theta^{(1)}, \ldots, \theta^{(L)}) \in \mathbb{R}^L$, we denote $\overline{\theta} = \frac{1}{L}\sum_{i=1}^{L}\theta^{(i)}$.

Proposition 17 *Let* $\theta, \varphi \in \mathcal{T} \subset \mathbb{R}^L$ *belong to a bounded set* $\mathcal{T}$ *of signals. Then for* σ *bigger than a threshold* $\sigma_0(L)$, *and* $\varrho(\theta,\varphi)$ *small enough, we have*

$$D_{KL}(\theta||\varphi) \ge C\sigma^{-4}\cdot\|\Delta_2(\theta,\varphi)\|_F^2.$$

Proposition 18 *Let* $\theta, \varphi \in \mathcal{T} \subset \mathbb{R}^L$ *belong to a bounded set* $\mathcal{T}$ *of signals, such that* $\overline{\theta} = \overline{\varphi}$ *and* $\|\Delta_2(\theta,\varphi)\|_F \le c\rho(\theta,\varphi)^2$. *Then, for* $\varrho(\theta,\varphi)$ *small enough we have*

$$D_{KL}(\theta||\varphi) \le C\frac{\|\theta\|_2^4}{\sigma^6}\cdot\rho(\theta,\varphi)^2$$

for some positive number C.

We defer the proof of Propositions 17 and 18 to Sect. 6.

3 The Dilute Regime of Sparsity and the Beltway Problem

3.1 Proof of Theorem 4

In this section, we will establish Theorem 4 and Lemma 5.

Proof of Theorem 4 We combine Lemma 5 with Proposition 17 to deduce that, for $\sigma \ge \sigma_0(L)$ and any $\theta, \theta_0 \in \mathcal{T}$ such that $\varrho(\theta,\theta_0)$ is small enough, the Kullback–Leibler divergence

$$D_{KL}(\theta||\theta_0) \ge c\sigma^{-4}\frac{s}{L}\cdot\rho(\theta,\varphi)^2$$

for some positive number c. This enables us to invoke Proposition 14 part (i) and deduce the desired asymptotic rate of estimation. This completes the proof. □

As is evident from the proof of Theorem 4, the key phenomenon to understand in the setting of the present section is the curvature lower bound as encapsulated in Lemma 5. We now proceed to the proof of this important result.

Proof of Lemma 5 As in the statement of the lemma, we focus on the situation where $\varrho(\theta, \theta_0)$ is small, and we recall that, for $\varrho(\theta, \theta_0)$ small enough, we may take $\varrho(\theta, \theta_0) = \|\theta - \theta_0\|_2$ because of the $\mathcal{G}$ invariance of D_{KL} and the moment tensors Δ_m. This is the setting in which we will work.

Notice that in this setting, θ and θ_0 have the same support. This follows from the assumption on our signal class $\mathcal{T}$ that for any $u \in \mathcal{T}$, we have $|u(i)| \geq m \quad \forall i \in \text{supp}(u)$. As a result, if $i \in \text{supp}(\theta) \Delta \text{supp}(\theta_0)$, then $|\theta(i) - \theta_0(i)| \geq m$, which implies that $\varrho(\theta, \theta_0) = \|\theta - \theta_0\|_2 \geq m$, which contradicts the smallness of $\varrho(\theta, \theta_0)$.

We set $h = \theta - \theta_0$, to be thought of has having small L^2-norm. Notice that the above discussion implies $\text{supp}(h) \subseteq \text{supp}(\theta_0)$. We consider

$$\|\mathbb{E}_{\mathcal{G}}[G\theta^{\otimes 2}G^*] - \mathbb{E}_{\mathcal{G}}[G\theta_0^{\otimes 2}G^*]\|_F = \|\mathbb{E}_{\mathcal{G}}[G(\theta_0 + h)^{\otimes 2}G^*] - \mathbb{E}_{\mathcal{G}}[G\theta_0^{\otimes 2}G^*]\|_F.$$

To the leading order in h, this is $\|\mathbb{E}_{\mathcal{G}}[\theta_0 h^* + h\theta_0^*]\|_F$, where * denotes transpose. Since $\|h\|_2$ is small, it suffices to consider this leading order term, and demonstrate that this is $\geq c\sqrt{s}\|h\|_2$. Henceforth, we focus on this objective.

We then have

$$\left(\mathbb{E}_{\mathcal{G}}[G(\theta_0 h^* + h\theta_0^*)G^*]\right)_{i,j} = \frac{1}{L}\sum_{g \in \mathbb{Z}_L}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)].$$

In our subsequent considerations, we will use the symbol J to denote the matrix $\mathbb{E}_{\mathcal{G}}[G(\theta_0 h^* + h\theta_0^*)G^*]$. Observe that J is a Toeplitz matrix. We can therefore denote the entries of J as $J_{i,j} = J_{j-i}$. Furthermore, for each i, j we have $J_{i,j} = J_{i+k,j+k}$, for any $k \in \mathbb{Z}_L$ and the sums $i + k$, $j + k$ in the indices being interpreted to be sums in $\mathbb{Z}_L$. In view of this, we can write $\|J\|_F^2 = L\sum_{k=0}^{L-1}|J_k|^2$. From here on, we will focus on estimating from below the sum $\sum_{k=0}^{L-1}|J_k|^2$.

Note that $\theta_0(i+g)h(j+g)$ or $h(i+g)\theta_0(j+g)$ is non-zero only if both $i+g$ and $j+g$ belong to the support of θ_0 (which contains the support of h). But since all non-zero differences occur exactly once (collision-free property), there exists a unique $g = g(i, j)$ such that $[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$ can possibly be non-zero. In particular, this means that, for J_k to be non-zero, k has to belong to $\mathcal{D}(\theta_0)$.

Suppose $0 \neq k \in \mathcal{D}(\theta_0)$ and let $i, j \in \text{supp}(\theta_0)$ be such that $j - i = k$. By the collision-free property, there is exactly one such pair (i, j). Therefore

$$|J_k|^2 = \frac{1}{L^2}[\theta_0(i)h(j) + h(i)\theta_0(j)]^2.$$

Conversely, if $i \neq j$ are such that $j - i \in \mathcal{D}(\theta_0)$, then there is a (unique) contribution to the sum $\sum_k |J_k|^2$ by an amount $\frac{1}{L^2}[\theta_0(i)h(j) + h(i)\theta_0(j)]^2$.

Finally, note that in case either i or j does not belong to $\mathrm{supp}(\theta_0)$, we have

$$\frac{1}{L^2}[\theta_0(i)h(j) + h(i)\theta_0(j)]^2 = 0.$$

Putting together all of the above, and denoting $s := |\mathrm{supp}(\theta_0)|$ we have

$$\begin{aligned}
\|J\|_F^2 &= L\sum_{k=0}^{L-1}|J_k|^2 \\
&\geq L\sum_{k=1}^{L-1}|J_k|^2 \quad \geq \frac{1}{L}\cdot\sum_{i\neq j}[\theta_0(i)h(j)+h(i)\theta_0(j)]^2 \\
&= \frac{1}{L}\cdot\sum_{i\neq j}[\theta_0(i)^2h(j)^2 + h(i)^2\theta_0(j)^2 + 2\theta_0(i)h(i)\theta_0(j)h(j)]. \\
&= \frac{1}{L}\cdot\left[2\left(\sum_i\theta_0(i)^2\right)\left(\sum_j h(j)^2\right) - 2\sum_i\theta_0(i)^2h(i)^2 + 2\left(\sum_i\theta_0(i)h(i)\right)^2\right. \\
&\quad \left. -2\sum_i\theta_0(i)^2h(i)^2\right] \\
&\geq \frac{2}{L}\cdot\left[\|\theta_0\|^2\|h\|^2 - 2\sum_i\theta_0(i)^2h(i)^2\right] \geq \frac{2}{L}\cdot\left[\|\theta_0\|^2\|h\|^2 - 2M^2\sum_i h(i)^2\right] \\
&\geq \frac{2}{L}\cdot\left(\|\theta_0\|^2 - 2M^2\right)\|h\|^2 \geq \frac{2}{L}\cdot\left(sm^2 - 2M^2\right)\|h\|^2 \\
&= \frac{2s}{L}(m^2 - \frac{2M^2}{s})\|h\|^2, \qquad (3.1)
\end{aligned}$$

where, in the last few steps, we have used the fact that $m \leq |\theta_0(i)| \leq M\ \forall i \in \mathrm{supp}(\theta_0)$.

For $s \geq (2+\varepsilon)M^2/m^2$ with $\varepsilon > 0$, the lower bound in (3.1) can be further bounded below by $\frac{2\varepsilon}{2+\varepsilon}\cdot\frac{s}{L}\cdot\|h\|^2$, as desired. □

4 Curvature of KL Divergence and the Fourier Transform

In this section, we will show that, without additional structural assumptions on the signal (such as sparsity), the second moment is *generically* insufficient to achieve $O(\sigma^4)$ sample complexity, as indicated in Theorem 12. In doing so, we will study the MRA problem in general, and the second moment tensor in particular, from the point of view of the Fourier transform of the signal θ.

Proof of Theorem 12 We begin with the fact that and that $D_{KL}(p_{\theta_0}\|p_\theta)$ has a local minimum at $\theta = \theta_0$, and on a related note, we have

$$D_{KL}(p_{\theta_0}\|p_\theta)\big|_{\theta=\theta_0} = 0, \quad \nabla_\theta D_{KL}(p_{\theta_0}\|p_\theta)\big|_{\theta=\theta_0} = 0, \quad \nabla_\theta^2 D_{KL}(p_{\theta_0}\|p_\theta)\big|_{\theta=\theta_0} = I(\theta_0).$$

We can consider a second order Taylor series expansion of $D_{KL}(p_{\theta_0}\|p_\theta)$ in the variable θ in a small enough neighbourhood of θ_0, and obtain

$$D_{KL}(p_{\theta_0}\|p_\theta) \geq c\sigma_{\min}(I(\theta_0))\varrho(\theta,\theta_0)^2, \tag{4.1}$$

where $\sigma_{\min}(I(\theta_0))$ is the smallest singular value of $I(\theta_0)$.

We combine the above observation with Lemma 13 and Proposition 18. For θ_k as in Lemma 13, we may deduce via Proposition 18 that

$$D_{KL}(p_{\theta_0}\|p_\theta) \leq c\|\theta_0\|_2^4\sigma^{-6}\varrho(\theta,\theta_0)^2 \leq c(L)\sigma^{-6}\varrho(\theta,\theta_0)^2, \tag{4.2}$$

where we have used the fact that the boundedness of the signal class $\mathcal{T}$ implies that $\|\theta\|_2 \leq C(L)$ uniformly for $\theta \in \mathcal{T}$ for some positive number $C(L)$.

Combining (4.1) and (4.2), we obtain

$$\sigma_{\min}(I(\theta_0)) \leq c(L)\sigma^{-6}. \tag{4.3}$$

We recall from Sect. 2.1, in particular (2.4), that $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to N(0, I(\theta_0)^{-1})$ as $n \to \infty$. Let $\mathcal{Z}$ denote a standard Gaussian vector of the same dimension as θ_0. Then we may write $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to I(\theta_0)^{-1/2}\mathcal{Z}$. Write $I(\theta_0) = U^*\Sigma U$ as the spectral decomposition of $I(\theta_0)$ with the eigenvalues $\{\sigma_i(I(\theta_0))\}_i$. Notice that, by rotational invariance, $\mathcal{Z}' := U\mathcal{Z}$ is also a standard Gaussian of the same dimension (with coordinates $\{\mathcal{Z}'_i\}_i$ being distributed as a standard $N(0,1)$ variable z). We may then deduce that

$$\begin{aligned}
&\sqrt{n}\|\theta - \theta_0\|_2 \to \|I(\theta_0)^{-1/2}\mathcal{Z}\|_2 \\
&= \|U^*\Sigma^{-1/2}U\mathcal{Z}\|_2 = \|\Sigma^{-1/2}\mathcal{Z}'\|_2 \\
&= \sqrt{\sum_i \sigma(I(\theta_0))^{-1}|\mathcal{Z}'_i|^2} \geq \sqrt{[\sigma_{\min}(I(\theta_0))]^{-1}}|z| \\
&\geq c_1(L)\sigma^3|z| = \Omega_p(\sigma^3),
\end{aligned}$$

where in the last inequality we have used (4.3).

Finally, in a small enough neighbourhood of θ_0 we may identify $\varrho(\theta,\theta_0)$ as $\frac{1}{\sqrt{L}}\|\theta - \theta_0\|_2$, and conclude that $\sqrt{n}\varrho(\tilde{\theta}_n,\theta_0) = \Omega_p(\sigma^3)$ as $n \to \infty$, as desired. □

We henceforth focus our attention on proving Lemma 13.

In order to carry out our investigations, we will utilize the Fourier transform as a key tool, and make repeated use of the renowned Parseval–Plancherel Theorem regarding the isometry properties of the Fourier transform. For convenience, we recall below these notions relevant for our analysis on $\mathbb{Z}_L$.

Definition 19 For $\theta = (\theta_1, \ldots, \theta_L) \in \mathbb{R}^L$, i.e. $\theta : \mathbb{Z}_L \mapsto \mathbb{R}$, we define the Fourier transform $\hat{\theta} : \mathbb{Z}_L \mapsto \mathbb{C}$ as

$$\hat{\theta}_j = \sum_{k=1}^{L} \theta_k \exp\left(-\frac{2\pi i j k}{L}\right).$$

The inverse Fourier transform of θ, denoted $\check{\theta} : \mathbb{Z}_L \mapsto \mathbb{C}$ is defined as

$$\check{\theta}_j = \frac{1}{L} \sum_{k=1}^{L} \theta_k \exp\left(\frac{2\pi i j k}{L}\right).$$

Theorem 20 (Parseval–Plancherel Theorem) *For $\theta = (\theta_1, \ldots, \theta_L) \in \mathbb{R}^L$, i.e. $\theta : \mathbb{Z}_L \mapsto \mathbb{R}$, we have*

$$\|\theta\|_2^2 = \sum_{k=1}^{L} |\theta_k|^2 = \frac{1}{L} \sum_{k=1}^{L} |\hat{\theta}_k|^2 = \frac{1}{L} \|\hat{\theta}\|_2^2.$$

Equivalently, for $\theta, \psi \in \mathbb{R}^L$, we may write

$$\sum_{k=1}^{L} \theta_k \overline{\psi}_k = \frac{1}{L} \sum_{k=1}^{L} \hat{\theta}_k \overline{\hat{\psi}_k} = \frac{1}{L} \cdot \langle \hat{\theta}, \hat{\psi} \rangle.$$

We begin with a result that obtains a succinct expression for the second moment tensor that is valid for any signal in $\mathbb{R}^L$, and therefore of general interest. For stating our result, we introduce the following notation: for any vector $v \in \mathbb{R}^L$, we denote by $\mathcal{M}(v)$ the matrix

$$[\mathcal{M}(v)]_{ij} := v(i - j).$$

This identifies the matrix $\mathcal{M}(v)$ as the Toeplitz matrix with symbol $\hat{v}$, the Fourier transform of v. Furthermore, as is common in our context, we will view any vectors $u, v \in \mathbb{R}^L$ to be functions mapping $\mathbb{Z}_L \mapsto \mathbb{R}$, and by the convolution $u * v$ we will denote the convolution of these two functions (under the action of the rotation group $\mathbb{Z}_L$). Namely,

$$[u * v](k) = \sum_{g \in \mathbb{Z}_L} u(g) v(k - g).$$

We recall that for any $v : \mathbb{Z}_L \mapsto \mathbb{R}$, the vector $\check{v}$ is given by $[\check{v}](i) = v(-i)$.

Lemma 21 *For any $v_1, v_2 \in \mathbb{R}^L$, we have*

$$\mathbb{E}_{\mathcal{G}}[G(v_1 \otimes v_2)G^*] = \frac{1}{L} \cdot \mathcal{M}(v_1 * \check{v}_2).$$

Proof Observe that

$$\mathbb{E}_{\mathcal{G}}[G(v_1 \otimes v_2)G^*]_{ij} = \frac{1}{L} \cdot \left(\sum_{g \in \mathbb{Z}_L} v_1(i+g)v_2(j+g) \right),$$

where as is usual in this context, the indices $i+g$, $j+g$ for the co-ordinates of the vectors are interpreted in the cyclic group $\mathbb{Z}_L$. Setting $i' = i+g$, we can re-write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{G}}[G(v_1 \otimes v_2)G^*]_{ij} \\
&= \frac{1}{L} \cdot \left(\sum_{i' \in \mathbb{Z}_L} v_1(i')v_2(j-i+i') \right) \\
&= \frac{1}{L} \cdot \left(\sum_{i' \in \mathbb{Z}_L} v_1(i')\check{v}_2(i-j-i') \right) \\
&= \frac{1}{L} \cdot \left[\mathcal{M}(v_1 * \check{v}_2) \right]_{ij},
\end{aligned}$$

as desired. □

Another result which would be useful for us subsequently is the following lemma.

Lemma 22 *For any $v \in \mathbb{R}^L$, we have*

$$\|\mathcal{M}(v)\|_F = \sqrt{L}\|v\|_2 = \|\hat{v}\|_2^2. \tag{4.4}$$

More generally,

$$\mathrm{Tr}[\mathcal{M}(v)\mathcal{M}(w)^*] = L\langle v, w\rangle = \langle \hat{v}, \hat{w}\rangle. \tag{4.5}$$

Proof We have,

$$\begin{aligned}
&\|\mathcal{M}(v)\|_F^2 \\
&= \sum_{i=1}^{L}\sum_{j=1}^{L} |v(i-j)|^2 \\
&= \sum_{k \in \mathbb{Z}_L} L|v(k)|^2 \\
&= L\|v\|_2^2 \\
&= \|\hat{v}\|_2^2 \quad \text{[by the Parseval–Plancherel Theorem]}
\end{aligned}$$

which completes the proof of (4.4).

The equality (4.5) follows from (4.4) via polarization, wherein we make use of the fact that $\|\mathcal{M}(v)\|_F^2 = \mathrm{Tr}[\mathcal{M}(v)\mathcal{M}(v)^*]$ and that the mapping $v \mapsto \mathcal{M}(v)$ is linear. □

We are now ready to state the following lemma.

Lemma 23 *For any $\theta, \varphi \in \mathbb{R}^L$ such that $h = \varphi - \theta$, we have*

$$\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] = \frac{1}{L} \cdot \left(\mathcal{M}(\theta * \check{h}) + \mathcal{M}(\check{\theta} * h) + \mathcal{M}(h * \check{h}) \right).$$

Proof We have,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[(G(\theta + h))^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] + \mathbb{E}_{\mathcal{G}}[(Gh)^{\otimes 2}] + \mathbb{E}_{\mathcal{G}}[G(\theta \otimes h)G^*] + \mathbb{E}_{\mathcal{G}}[G(h \otimes \theta)G^*] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[G(\theta \otimes h)G^*] + \mathbb{E}_{\mathcal{G}}[G(h \otimes \theta)G^*] + \mathbb{E}_{\mathcal{G}}[(Gh)^{\otimes 2}] \\
&= \frac{1}{L} \cdot \left[\mathcal{M}(\theta * \check{h}) + \mathcal{M}(h * \check{\theta}) + \mathcal{M}(h * \check{h}) \right] \\
&\quad \text{(using Lemma 21)}
\end{aligned}$$

as desired. □

Lemma 23 shows, in particular, that in the regime of small h, the linearization (in h) of the second moment difference tensor is given by $J(\theta, h) := \frac{1}{L} \cdot \left(\mathcal{M}(\theta * \check{h}) + \mathcal{M}(\check{\theta} * h) \right)$, which we will focus on in the proof of Lemma 13 that follows.

Proof of Lemma 13 In order to construct a sequence $\{\theta_k\}_k$ as in the statement of the present Lemma, it would suffice to demonstrate the existence of θ arbitrarily close to θ_0 such that $\overline{\theta} = \overline{\theta}_0$ and (1.9) is satisfied . We recall that, for $\varrho(\theta, \theta_0)$ small enough, we may take $\varrho(\theta, \theta_0) = \|\theta - \theta_0\|_2$.

In what follows, we will set $h := \theta - \theta_0$.

We begin with

$$\begin{aligned}
&\|\mathcal{M}(\theta_0 * \check{h}) + \mathcal{M}(\check{\theta}_0 * h)\|_F^2 \\
&\quad = \|\mathcal{M}(\theta_0 * \check{h})\|_F^2 + \|\mathcal{M}(\check{\theta}_0 * h)\|_F^2 + \mathrm{Tr}[\mathcal{M}(\theta_0 * \check{h})\mathcal{M}(\check{\theta}_0 * h)^*] \\
&\qquad + \mathrm{Tr}[\mathcal{M}(\theta_0 * \check{h})^* \mathcal{M}(\check{\theta}_0 * h)] \\
&\quad = \|\mathcal{M}(\theta_0 * \check{h})\|_F^2 + \|\mathcal{M}(\check{\theta}_0 * h)\|_F^2 + 2\Re\left(\mathrm{Tr}[\mathcal{M}(\theta_0 * \check{h})\mathcal{M}(\check{\theta}_0 * h)^*] \right) \qquad (4.6)
\end{aligned}$$

Using Lemma 22, we deduce that $\|\mathcal{M}(\theta_0 * \check{h})\|_F = \sqrt{L}\|\theta_0 * \check{h}\|_2$. Using Parseval–Plancherel's Theorem, we deduce that

$$\|\theta_0 * \check{h}\|_2 = \frac{1}{\sqrt{L}} \|\widehat{\theta_0 * \check{h}}\|_2 = \frac{1}{\sqrt{L}} \|\hat{\theta} \cdot \hat{\check{h}}\|_2,$$

where for two vectors $u, v \in \mathbb{R}^L$, the quantity $u \cdot v$ denotes their co-ordinate wise product. We further introduce the notations that, for any vector $v \in \mathbb{R}^L$, we denote

by $|v|$ the vector given by $|v|(i) = |v(i)|\ \forall i \in \mathbb{Z}_L$, and by v^2 we denote the vector given by $v^2(i) = v(i)^2\ \forall i \in \mathbb{Z}_L$. We may deduce from definition that, for any vector $v \in \mathbb{R}^L$, we have $\hat{\check{v}} = \overline{\hat{v}}$, which further leads to

$$\|\mathcal{M}(\theta_0 * \check{h})\|_F^2 = L \cdot \frac{1}{L}\|\hat{\theta} \cdot \overline{\hat{h}}\|_2^2 = \langle |\hat{\theta}|^2, |\hat{h}|^2 \rangle. \tag{4.7}$$

Similarly, we can deduce that

$$\|\mathcal{M}(\check{\theta}_0 * h)\|_F^2 = \|\hat{\check{\theta}} \cdot \hat{h}\|_2^2 = \langle |\hat{\theta}|^2, |\hat{h}|^2 \rangle. \tag{4.8}$$

and

$$\begin{aligned}
&\Re\left(\mathrm{Tr}[\mathcal{M}(\theta_0 * \check{h})\mathcal{M}(\check{\theta}_0 * h)^*]\right) \\
&= L \cdot \Re\langle \theta_0 * \check{h}, \check{\theta}_0 * h \rangle \quad \text{[Using Lemma 22]} \\
&= L \cdot \frac{1}{L}\Re\langle \widehat{\theta_0 * \check{h}}, \widehat{\check{\theta}_0 * h} \rangle \quad \text{[Using the Parseval–Plancherel Theorem]} \\
&= \Re\langle \hat{\theta}_0 \cdot \hat{\check{h}}, \hat{\check{\theta}}_0 \cdot \hat{h} \rangle \\
&= \Re\langle \hat{\theta}_0 \cdot \overline{\hat{h}}, \overline{\hat{\theta}}_0 \cdot \hat{h} \rangle.
\end{aligned} \tag{4.9}$$

Combining (4.6)–(4.9), we may deduce that

$$\begin{aligned}
&\|\mathcal{M}(\theta_0 * \check{h}) + \mathcal{M}(\check{\theta}_0 * h)\|_F^2 \\
&= 2\langle |\hat{\theta}|^2, |\hat{h}|^2 \rangle + 2\Re\langle \hat{\theta}_0 \cdot \overline{\hat{h}}, \overline{\hat{\theta}}_0 \cdot \hat{h} \rangle \\
&= \left[\sum_{\xi \in \mathbb{Z}_L} 2\left(|\hat{\theta}(\xi)|^2 |\hat{h}(\xi)|^2 + \Re\left(\hat{\theta}(\xi)^2 [\overline{\hat{h}(\xi)}]^2 \right) \right) \right]
\end{aligned} \tag{4.10}$$

For $\theta_0 \in T$ as in the statement of the Theorem, we propose to choose h such that

$$\left(|\hat{\theta}(\xi)|^2 |\hat{h}(\xi)|^2 + \hat{\theta}(\xi)^2 [\overline{\hat{h}(\xi)}]^2 \right) = 0\ \forall \xi \in \mathbb{Z}_L. \tag{4.11}$$

This would be possible because of the following reasons; using the fact that the Fourier transform is a bijection, we will determine the choice of h in the Fourier domain..

First, we set $\hat{h}(0) = 0$, which will come in handy later. To set the coordinates $\hat{h}(\xi)$ for $\xi \neq 0$, we proceed as follows. Recall that the only restriction on vectors in the signal class $\mathcal{T}$ is that they have full support and their co-ordinates assume real values between m and M. This translates into the fact that the only restriction on the difference h of two such signals in the interior of $\mathcal{T}$ is that h has real co-ordinates (as long as $\|h\|_2$ is small enough). This implies that the only restriction on the Fourier transform $\hat{h}$ is that $\hat{h}$ is symmetric (the essential reason for which is that the Fourier transform is

surjective from $\mathbb{R}^L$ to symmetric vectors in $\mathbb{C}^L$). So, for any given $\theta_0 \in T$, we choose $\hat{h}(\xi)$ (for $\xi \neq 0$) such that $|\hat{h}(\xi)|$ is small enough and

$$2\mathrm{Arg}(\hat{h}(\xi)) = \pi - 2\mathrm{Arg}(\hat{\theta}_0(\xi)), \tag{4.12}$$

which ensures that $\left(|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2 + \hat{\theta}(\xi)^2[\overline{\hat{h}(\xi)}]^2\right) = 0$. The symmetry condition on $\hat{h}$ can be satisfied because, $\theta_0 \in T$ implies that $\hat{\theta}_0$ is symmetric, which allows us to choose $\hat{h}$ as in (4.12) so that the vector $\hat{h}$ is indeed symmetric.

The upshot of the (4.11) is that $\|\mathcal{M}(\theta_0 * \check{h}) + \mathcal{M}(\check{\theta}_0 * h)\|_F = 0$, which implies that $\mathcal{M}(\theta_0 * \check{h}) + \mathcal{M}(\check{\theta}_0 * h) = 0$. Thus, Lemma 23 implies that

$$\|\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]\|_F = \|\mathcal{M}(h * \check{h})\|_F.$$

But then we have

$$\begin{aligned}
\|\mathcal{M}(h * \check{h})\|_F^2 &= L\|h * \check{h}\|_2^2 \quad \text{[Using Lemma 22]} \\
&= L \cdot \frac{1}{L}\|\widehat{h * \check{h}}\|_2^2 = \|\hat{h} \cdot \overline{\hat{h}}\|_2^2 \quad \text{[By Parseval–Plancherel Theorem]} \\
&= \left(\sum_{\xi \in \mathbb{Z}_L} |\hat{h}(\xi)|^4\right) \qquad (4.13) \\
&\leq \left(\sum_{\xi \in \mathbb{Z}_L} |\hat{h}(\xi)|^2\right)^2 \quad \text{[By Cauchy–Schwarz]} \\
&= \|\hat{h}\|_2^4 = L^2\|h\|_2^4. \quad \text{[By Parseval–Plancherel]} \qquad (4.14)
\end{aligned}$$

The upshot of this is that $\|\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]\|_F \leq C_L\|h\|_2^2$, thereby verifying the condition (1.9).

It remains to verify the condition $\overline{\theta} = \overline{\theta}_0$; equivalently, $\overline{h} = 0$. Observe that $\hat{h}(0) = \sum_{x \in \mathbb{Z}_L} h(x) = L\overline{h}$. However, we have already set $\hat{h}(0) = 0$, which therefore implies that $\overline{h} = 0$. This entails that $\overline{\theta} = \overline{\theta}_0$, thereby completing the proof.

□

5 The Regime of Moderate Sparsity and Uncertainty Principles

In this section, we establish Theorems 6 and Lemma 9, in the process invoking Uniform Uncertainty Principles from the discrete Fourier analysis of sparse vectors. We will proceed as follows. First, we will establish the technical Lemma 9. Next, we will verify that the generic signals considered in Theorem 6 satisfy the conditions of Lemma 9, thereby completing the proof of Theorem 6. Finally, we will demonstrate the genericity of support for the symmetric Bernoulli–Gaussian and $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ distributions.

5.1 Proof of Theorem 6 and Lemma 9

Proof of Lemma 9 Our strategy would involve demonstrating a lower bound on the norm of the second moment difference tensor $\Delta_2(\varphi,\theta) = \mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]$ that is linear in the distance $\rho(\varphi,\theta)$. That is, $\|\Delta_2(\varphi,\theta)\|_F \geq C(L)\cdot\rho(\varphi,\theta)$. We will do so for $\varphi \in \mathcal{T}$ lying in a neighbourhood of θ. Once such a lower bound is obtained, the theorem will follow from Proposition 17 and Proposition 14 part (i).

We will work in the local neighbourhood of θ, so that $\varrho(\varphi,\theta) \leq r(L)$; the precise size of $r(L)$ will be specified later. If $r(L)$ is small enough, then without loss of generality, we may write $\rho(\varphi,\theta) = \|\varphi-\theta\|_2 = \|h\|_2$, which is the formulation that we will work with.

Notice that, because of the lower bound in part (ii) of the statement of this Lemma, we have both $\min_{\{}\, i \in \mathrm{supp}(\varphi)\}|\varphi(i)|, \min_{\{}\, i \in \mathrm{supp}(\theta)\}|\theta(i)| \geq m_{\mathcal{T}}$. As such, if $r(L)$ is small enough such that $r(L) < m_{\mathcal{T}}$, we have $\|\varphi-\theta\|_2 = \|h\|_2 < m_{\mathcal{T}}$. This implies that, for $r(L)$ small enough, the supports of φ and θ must coincide. In particular, the difference $h = \varphi - \theta$ must satisfy $\mathrm{supp}(h) \subseteq \mathrm{supp}(\theta)$. This enables us to invoke condition (iii-a) in the statement of the Lemma for such h, which we shall use below.

Since θ, φ are symmetric, $h = \varphi - \theta$ are also symmetric. This implies that $\theta = \check{\theta}$ and $h = \check{h}$, and both Fourier transforms $\hat{\theta}$ and $\hat{h}$ are real-valued. From Lemma 23 we may deduce that

$$\|\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]\|_F \geq \frac{2}{L}\cdot\|\mathcal{M}(\theta * h)\|_F - \frac{1}{L}\|\mathcal{M}(h * h)\|_F. \quad (5.1)$$

Using (4.10) and (4.13), we may further simplify this to

$$\begin{aligned}\|\Delta(\varphi,\theta)\|_F &= \|\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]\|_F \\ &\geq \frac{1}{L}\left(4\cdot\sum_{\xi\in\mathbb{Z}_L}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2\right)^{1/2} - \frac{1}{L}\left(\sum_{\xi\in\mathbb{Z}_L}|\hat{h}(\xi)|^4\right)^{1/2}. \quad (5.2)\end{aligned}$$

The second term on the right-hand side will be controlled using the fact that $\|h\|_2 \leq r(L)$, a consequence of the fact that $\varrho(\varphi,\theta) \leq r(L)$. To demonstrate this, we consider $\|h\|_\infty = \sup_{\xi\in\mathbb{Z}_L}|\hat{h}(\xi)|$. For any $\xi \in \mathbb{Z}_L$, we observe via the Cauchy Schwarz inequality that

$$|\hat{h}(\xi)| = \left|\sum_{k\in\mathbb{Z}_L} h(k)\exp\left(\frac{2\pi i k}{L}\right)\right| \leq |\mathrm{supp}(h)|\cdot\|h\|_2 \leq L\cdot\|h\|_2. \quad (5.3)$$

Using the Parseval–Plancherel Theorem, we may proceed as

$$\begin{aligned}
&\left(\sum_{\xi\in\mathbb{Z}_L}|\hat{h}(\xi)|^4\right)^{1/2}\\
&\leq \|h\|_\infty\left(\sum_{\xi\in\mathbb{Z}_L}|\hat{h}(\xi)|^2\right)^{1/2}\\
&\leq L\,\|h\|_2\|\hat{h}\|_2 \quad \text{[using (5.3)]}\\
&\leq L\sqrt{L}\,\|h\|_2^2 \quad \text{[via Parseval–Plancherel Theorem]}\\
&\leq L\sqrt{L}\,r(L)\cdot\|h\|_2. \quad \text{[Since } \|h\|_2\leq r(L)\text{]} \qquad (5.4)
\end{aligned}$$

Thus, if we are able to show in the context of (5.2) that $\left(4\sum_{\xi\in\mathbb{Z}_L}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2\right)^{1/2}$ is bounded below by $c(L)\|h\|_2$, then as soon as $r(L)$ is chosen to be small enough such that $L\sqrt{L}\,r(L)\leq\frac{1}{2}c(L)$, we will be done [via (5.2)] with an overall lower bound on $\|\Delta(\varphi,\theta)\|_F$ by $\frac{1}{2L}\cdot c(L)\cdot\|h\|_2$.

In view of this, we will henceforth focus attention to lower-bounding $\left(4\cdot\sum_{\xi\in\mathbb{Z}_L}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2\right)^{1/2}$.

To this end, we recall the set Λ and the quantity $\mathfrak{m}(\mathcal{T})$ from the defining criteria of $\mathcal{T}$, and proceed as

$$\begin{aligned}
\frac{1}{L}\cdot\sum_{\xi\in\mathbb{Z}_L}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2 &= \frac{1}{L}\cdot\sum_{\xi\in\mathbb{Z}_L}|\widehat{\theta*h}(\xi)|^2\\
&\geq c_2^{-1}\frac{1}{|\Lambda|}\cdot\sum_{\xi\in\Lambda}|\widehat{\theta*h}(\xi)|^2 \quad \text{(since } \theta*h \text{ is 4s-sparse)}\\
&= c_2^{-1}\cdot\frac{1}{|\Lambda|}\cdot\sum_{\xi\in\Lambda}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2 \quad \geq c_2^{-1}\cdot\frac{1}{|\Lambda|}\cdot\sum_{\xi\in\Lambda}\mathfrak{m}(\mathcal{T})^2|\hat{h}(\xi)|^2\\
&= c_2^{-1}\mathfrak{m}(\mathcal{T})^2\cdot\frac{1}{|\Lambda|}\cdot\sum_{\xi\in\Lambda}|\hat{h}(\xi)|^2 \quad \geq c_1c_2^{-1}\mathfrak{m}(\mathcal{T})^2\cdot\frac{1}{L}\cdot\sum_{\xi\in\mathbb{Z}_L}|\hat{h}(\xi)|^2\\
&= c_3^2\cdot\mathfrak{m}(\mathcal{T})^2\|h\|_2^2, \qquad (5.5)
\end{aligned}$$

where $c_3=\sqrt{c_1c_2^{-1}}$.

We consider (5.5) in the context of (5.4) and the discussion immediately following it. With $c(L)=2c_3\cdot\sqrt{L}\cdot\mathfrak{m}(\mathcal{T})$, we may conclude that $\left(4\sum_{\xi\in\mathbb{Z}_L}|\hat{\theta}(\xi)|^2|\hat{h}(\xi)|^2\right)^{1/2}\geq c(L)\|h\|_2$, and therefore $\|\Delta(\varphi,\theta)\|_F\geq\frac{1}{2L}\cdot c(L)\cdot\|h\|_2\geq c_4\cdot\frac{\mathfrak{m}(\mathcal{T})}{\sqrt{L}}\cdot\|h\|_2$ for a suitable positive constant c_4. Recalling that $\rho(\varphi,\theta)=\|h\|_2$, we obtain the desired lower bound on the second moment tensor difference in a neighbourhood of θ:

$$\|\Delta(\varphi,\theta)\|_F=\|\mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}]-\mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}]\|_F\geq c_4\cdot\frac{\mathfrak{m}(\mathcal{T})}{\sqrt{L}}\cdot\rho(\varphi,\theta). \qquad (5.6)$$

Combined with the discussion at the beginning of this proof, this completes the argument. □

We now discuss the proof of Theorem 6.

Proof of Theorem 6 Our approach to this proof will involve demonstrating that the set of signals $\mathcal{T} = \mathcal{T}_s$, as in the statement of Lemma 9, has high probability under the generative model for the signal in the present theorem. We will do this by showing below that each of the criteria (i)–(iii) in Lemma 9 has high probability under the conditions of our current theorem.

(i): This condition is trivially satisfied by the present generative model, by definition.

(ii): We now consider the upper and lower bounds on the signal θ on $\text{supp}(\theta) = \Xi$. In doing so, we will use the fact that the support Ξ is *typically s-sparse* with sparsity constants (α, β), which implies that $\alpha s \leq |\Xi| \leq \beta s$ with high probability.

But $\max_{k\in\text{supp}(\theta)} |\theta(k)| \leq \max\{\xi_k : k \in \text{supp}(\theta)\}$, where the ξ_k-s are i.i.d. centred Gaussians with variance ζ^2, is given by $O_p(\zeta\sqrt{2\log|\Xi|}) = O_p(\zeta\sqrt{\log s})$. This enables us to set $M_{\mathcal{T}_s} = c\zeta \log s$ in order to ensure that the maximum condition is satisfied with high probability.

Similarly, $\min_{k\in\text{supp}(\theta)} |\theta(k)| \leq \min\{\xi_k : k \in \text{supp}(\theta)\}$ will decay (in $|\Xi|$) as x such that $x|\Xi|/\zeta = O_p(1)$; this follows from the functional form of the Gaussian density $N(0, \zeta^2)$. Thus, it suffices to take $m_{\mathcal{T}_s} = c \min_\Xi (\zeta/|\Xi| \log|\Xi|) = c\zeta/s \log s$ in order to ensure that the minimum condition is satisfied with high probability.

(iii): We now come to the consideration of the desirable set of frequencies Λ. We will demonstrate the existence of such a frequency set by an *enhanced* version of the probabilistic method. While we will use randomness for finding the desirable set Λ, it may be observed that the typical random frequency set of the right size would not satisfy the stipulated conditions on Λ. Instead, we will require additional considerations in order to show the *existence* of *one such (atypical) set*. To this end, we would require certain auxiliary technical results, which are encapsulated in Lemmas 26, 27 and 28.

To begin with, we

$$\begin{aligned}&\text{Draw a random subset } \Lambda \subset \mathbb{Z}_L \text{ having expected size } a, \text{ that is}\\ &\text{each element of } \mathbb{Z}_L \text{ can be in } \Lambda \text{ independently with probability } a/L,\end{aligned} \tag{5.7}$$

where a is a positive number slightly larger than the sparsity parameter s, to be specified in detail later. By Lemma 26, such a set Λ will satisfy the condition (iii)(a) in the present Lemma for all s-sparse vectors h with probability

$$\mathbb{P}[\Lambda \text{ satisfies (iii)(a) for all } s-\text{sparse vectors } h] \geq 1 - 5\exp(-ca(s\log^4 L)^{-1}), \tag{5.8}$$

the probability in question being in the randomness of Λ. We say that a subset $A \subset \mathbb{Z}_L$ satisfies the Uniform Uncertainty Principle for s-sparse vectors (abbrv. s-UUP), if it satisfies (5.12). Observe that a, being equal to $|\Lambda|$ and $\Lambda \subseteq \mathbb{Z}_L$, needs to satisfy $a \leq L$. In view of this, to ensure that Λ satisfies (iii)(a) with high probability (as s, L grow large), we need to have $a(s\log^4 L)^{-1} \to \infty$ in (5.8), which is equivalent

to $s \ll L/(\log^4 L)$, which in turn is ensured by the condition $L \leq L/(\log^5 L)$ for L large enough.

We now work towards showing that with positive, albeit vanishingly small probability, the condition (iii)(b) is also satisfied. To that end, we notice that while (iii)(a) is valid for all s-sparse vectors h, we need to verify (iii)(b) only for the signal class of our interest - fact that we will crucially exploit in our considerations.

We begin with the observation that, since Ξ is (s^τ)-cosine generic, with high probability in the set Ξ we have the inequality $\min_{\xi\in\mathbb{Z}_L} |\mathcal{V}(\Xi,\xi)| \geq s^\tau$. For such a set Ξ, and any $\eta \in (0,1)$, we invoke the quantity $\mathfrak{a}(\Xi,\eta,\zeta) = C(1-\eta)^{-1}\zeta^{-\eta}\left(\min_{\xi\in\mathbb{Z}_L}\mathcal{V}(\Xi,\xi)\right)^{-\frac{1}{2}\eta}$ [c.f. (5.14)], which immediately leads to the bound $\mathfrak{a}(\Xi,\eta,\zeta) \leq C(1-\eta)^{-1}\zeta^{-\eta}s^{-\frac{1}{2}\tau\eta}$.

For $\kappa > 0$ to be specified later, notice that this implies, with high probability in the signal θ, that

$$\mathfrak{a}(\Xi,\eta,\zeta)|\Xi|^{-\frac{1}{2}\kappa\eta} \leq C_\eta \zeta^{-\eta} s^{-\frac{1}{2}(\kappa+\tau)\eta}$$

with $C_\eta = C(1-\eta)^{-1}$, which is small in the regime of large s (and fixed ζ) as soon as $\kappa+\tau > 0$. In light of Lemma 27 and the defining equation (5.13), we may therefore conclude that, with high probability in the signal θ, we have the inequality $|\hat{\theta}(\xi)| \geq s^{-\kappa}$ for a set $\mathfrak{S}(\kappa)$ of frequencies ξ satisfying $|\mathfrak{S}(\kappa)| \geq L(1 - C_\eta\zeta^{-\eta}s^{-\frac{1}{2}(\kappa+\tau)\eta})$.

For a given τ, we now select $\kappa = \max\{4-\tau, 0\}$, which implies in particular that $\frac{1}{2}(\kappa+\tau) \geq 2$ and automatically $\kappa+\tau > 0$. We then choose $\eta = 3/4$, leading to the bound $|\mathfrak{S}(\kappa)| \geq L(1 - c\zeta^{-1/2}s^{-3/2})$. On this set $\mathfrak{S}(\kappa)$, with high probability in the signal θ, we have $|\hat{\theta}(\xi)| \geq c\min\{s^{\tau-4}, 1\}$.

We now examine carefully a randomly sampled subset $\Lambda \subset \mathbb{Z}_L$ with average size a, as in (5.7). We want to understand $\mathbb{P}[\Lambda \subset \mathfrak{S}(\kappa)]$, equivalently, $\mathbb{P}[\mathfrak{S}(\kappa)^\complement \subset \Lambda^\complement]$. Observe from the discussion above that $|\mathfrak{S}(\kappa)^\complement| \leq cL\zeta^{-1/2}s^{-3/2}$, and note that the probability of a particular frequency ξ to belong to $\Lambda^\complement$ is $(1-a/L)$. Since each frequency in $\mathbb{Z}_L$ is chosen to belong to Λ independently of each other, we may deduce that, as long as a/L remains bounded away from 1, we have

$$\begin{aligned}
\mathbb{P}[\Lambda \subset \mathfrak{S}(\kappa)] &= \mathbb{P}[\mathfrak{S}(\kappa)^\complement \subset \Lambda^\complement] \\
&= (1-a/L)^{|\mathfrak{S}(\kappa)^\complement|} \\
&\geq \exp(-c'\frac{a}{L}\cdot cL\zeta^{-1/2}s^{-3/2}) \\
&= \exp(-c''\zeta^{-1/2}as^{-3/2}). \qquad (5.9)
\end{aligned}$$

We may then proceed as

$$\begin{aligned}
&\mathbb{P}[\{\Lambda \text{ is s-UUP }\} \cap \{\Lambda \subset \mathfrak{S}(\kappa)\}] \\
&\quad = \mathbb{P}[\{\Lambda \subset \mathfrak{S}(\kappa)\} \setminus \{\Lambda \text{ is s-UUP }\}^\complement] \\
&\quad \geq \mathbb{P}[\Lambda \subset \mathfrak{S}(\kappa)] - \mathbb{P}[\{\Lambda \text{ is s-UUP }\}^\complement]
\end{aligned}$$

$$\geq \exp(-c''\zeta^{-1/2}as^{-3/2}) - 5\exp(-c_3as^{-1}\log^{-4} L))[\text{c.f. (5.9) and Lemma 26}] \tag{5.10}$$

The last expression in (5.10) is positive as soon as $\zeta^{-1/2}s^{-3/2} \ll s^{-1}\log^{-4} L$, equivalently $s \gg \log^8 L/\zeta$ in the regime of large L. The latter condition, in turn, is guaranteed by $s \geq \log^9 L$, as soon as L is large enough.

In this regime, i.e. when $s \gg \log^8 L/\zeta$, we have

$$\mathbb{P}[\{\Lambda \text{ is s-UUP }\} \cap \{\Lambda \subset \mathfrak{S}(\kappa)\}] > 0,$$

implying that there exists a realization of the subset Λ such that:

(i) Λ satisfies (5.12) for all s-sparse signals.
(ii) $\min_{\xi\in\Lambda} |\hat{\theta}(\xi)| \geq c\min\{s^{\tau-4}, 1\}$.

These two facts together establish the existence of a frequency set Λ as required in condition (iii) of Lemma 9, with

$$\mathfrak{m}(\mathcal{T}_s) = c\min\{s^{\tau-4}, 1\} \tag{5.11}$$

as above. This completes the argument for (iii), and therefore completes the proof of the present theorem. □

Remark 24 We observe that in (5.11) the lower bound $\mathfrak{m}(\mathcal{T}_s) = c\min\{s^{\tau-4}, 1\}$, in fact, depends only on s (and not on L).

Remark 25 In view of (5.8) and the discussion immediately thereafter, we note that it suffices to have $s \leq L/(\log^5 L)$ for L large enough. On the other hand, in view of (5.10) and the ensuing discussion, it suffices to have $s \geq \log^9 L$. Combining these two observations, we work in the regime of L such that

$$\log^9 L \leq s \leq L/(\log^5 L).$$

It remains to establish Lemmas 27 – 29, which we take up in the next section.

5.2 Proofs of Lemmas 27, 28 and 29

To begin with, we invoke the following *Uniform Uncertainty Principle* from [48].

Lemma 26 ([48]) *Let $\mathfrak{F}$ be a random set of frequencies in $\mathbb{Z}_L$ having expected size a, that is each element of $\mathbb{Z}_L$ can be in $\mathfrak{F}$ independently with probability a/L. Then there are fixed numbers c_1, c_2, c_3 such that, simultaneously for all $f : \mathbb{Z}_L \mapsto \mathbb{R}$ that is $s-sparse$, the event*

$$c_1 \cdot \frac{1}{L}\sum_{\xi\in\mathbb{Z}_L} |\hat{f}(\xi)|^2 \leq \frac{1}{|\mathfrak{F}|}\sum_{\xi\in\mathfrak{F}} |\hat{f}(\xi)|^2 \leq c_2 \cdot \frac{1}{L}\sum_{\xi\in\mathbb{Z}_L} |\hat{f}(\xi)|^2 \tag{5.12}$$

holds with probability $\geq 1 - 5\exp(-c_3a(s\log^4 L)^{-1})$.

Next, for $\kappa > 0$ and a function $f : \mathbb{Z}_L \to \mathbb{R}$ with $\mathrm{supp}(f) = \Xi$, we define the set of frequencies $\mathfrak{S}_f(\kappa)$ as

$$\mathfrak{S}_f(\kappa) = \{\xi \in \mathbb{Z}_L : |\hat{f}(\xi)| \geq |\Xi|^{-\kappa}\}. \tag{5.13}$$

Then we are ready to state the following lemma.

Lemma 27 *Let $\Xi \subset \mathbb{Z}_L$, let*

$$\mathfrak{a}(\Xi, \eta, \zeta) = C(1-\eta)^{-1}\zeta^{-\eta}\left(\min_{\xi \in \mathbb{Z}_L} \mathcal{V}(\Xi, \xi)\right)^{-\frac{1}{2}\eta} \tag{5.14}$$

for $C > 0$ as in Lemma 28 and any $0 < \eta < 1$, and let $f \sim N_\Xi(\mathbf{0}, \zeta^2\mathbf{I})$. Then, for $\kappa > 0$, we have

$$|\mathfrak{S}_f(\kappa)| \geq L\left(1 - \mathfrak{a}(\Xi, \eta, \zeta)|\Xi|^{-\frac{1}{2}\kappa\eta}\right)$$

with probability $\geq 1 - \mathfrak{a}(\Xi, \eta, \zeta)|\Xi|^{-\frac{1}{2}\kappa\eta}$.

Proof We will approach this result by upper bounding the size of the set $\mathfrak{S}(\kappa)^\complement$. Observe that, $\xi \in \mathfrak{S}(\kappa)^\complement$ implies that $|\hat{f}(\xi)| \leq |\Xi|^{-\kappa}$; equivalently, $|\hat{f}(\xi)|^{-\eta} \geq |\Xi|^{\kappa\eta}$.

This implies that,

$$\begin{aligned} |\Xi|^{\kappa\eta}|\mathfrak{S}(\kappa)^\complement| &= \sum_{\xi \in \mathfrak{S}(\kappa)^\complement} |\Xi|^{\kappa\eta} \\ &\leq \sum_{\xi \in \mathfrak{S}(\kappa)^\complement} |\hat{f}(\xi)|^{-\eta} \\ &\leq \sum_{\xi \in \mathbb{Z}_L} |\hat{f}(\xi)|^{-\eta}. \end{aligned}$$

Therefore, we may proceed as

$$|\Xi|^{\kappa\eta}\mathbb{E}[|\mathfrak{S}(\kappa)^\complement|] \leq \mathbb{E}[\sum_{\xi \in \mathbb{Z}_L} |\hat{f}(\xi)|^{-\eta}] \leq \sum_{\xi \in \mathbb{Z}_L} \mathbb{E}[|\hat{f}(\xi)|^{-\eta}] \leq L\mathfrak{a}(\Xi, \eta, \zeta), \tag{5.15}$$

where, in the last step, we make use of the definition (5.14) and of Lemma 28; in particular choosing C to be as in that Lemma.

We may restate this as

$$\mathbb{E}[|\mathfrak{S}(\kappa)^\complement|] \leq L\mathfrak{a}(\Xi, \eta, \zeta)|\Xi|^{-\kappa\eta}. \tag{5.16}$$

We may then proceed via Markov's inequality as

$$\begin{aligned}\mathbb{P}[|\mathfrak{S}(\kappa)| &\le L(1-|\Xi|^{-\frac{1}{2}\kappa\eta})] \\ &= \mathbb{P}[|\mathfrak{S}(\kappa)^{\complement}| \ge L|\Xi|^{-\frac{1}{2}\kappa\eta}] \\ &\le \mathbb{E}[|\mathfrak{S}(\kappa)^{\complement}|]/L|\Xi|^{-\frac{1}{2}\kappa\eta} \\ &\le L^{-1}|\Xi|^{\frac{1}{2}\kappa\eta}\cdot L\mathfrak{a}(\Xi,\eta,\zeta)|\Xi|^{-\kappa\eta} \quad \text{[using (5.16)]} \\ &= \mathfrak{a}(\Xi,\eta,\zeta)|\Xi|^{-\frac{1}{2}\kappa\eta},\end{aligned}$$

as desired. □

For $\Xi \subset \mathbb{Z}_L$ and $a \in \mathbb{Z}_L$, we recall (C.1):

$$\mathcal{V}(\Xi, a) = \mathbb{1}_{\{0\in\Xi\}} + 2\sum_{k\in\Xi\setminus\{0\}} \cos^2(2\pi ak/L),$$

where $\mathbb{1}_A$ denotes the indicator function of the event A.

Lemma 28 *Consider two subsets $\Xi, A \subset \mathbb{Z}_L$. Let $f \sim N_\Xi(\mathbf{0}, \zeta^2\mathbf{I})$. Then there is a positive number C such that for any $0 < \eta < 1$, we have*

$$\mathbb{E}[|\hat{f}(\xi)|^{-\eta}] \le C(1-\eta)^{-1}\zeta^{-\eta}\left(\min_{\xi\in A}\mathcal{V}(\Xi,\xi)\right)^{-\frac{1}{2}\eta}$$

for all $\xi \in A$.

Proof We invoke Lemma 29 in order to conclude that, for any $\xi \in \mathbb{Z}_L$, we have $\hat{f}(\xi) \sim N(0, \zeta^2\mathcal{V}(\Xi,\xi))$. In other words, $\hat{f}(\xi)$ is a 1D Gaussian random variable with variance $\mathcal{V}(\Xi,\xi)$. Now, it follows from the 1D standard Gaussian density formula that for any $\eta < 1$, a 1D standard Gaussian Z has a finite negative moment $\mathbb{E}[|X|^{-\eta}]$ by $C(1-\eta)^{-1}$ for some positive number C. It follows that, if $Y \sim N(0, a^2)$, then $\mathbb{E}[|Y|^{-\eta}] = a^{-\eta}\mathbb{E}[|X|^{-\eta}] \le C(1-\eta)^{-1}a^{-\eta}$. This completes the proof. □

We can then state

Lemma 29 *Let $f \sim N_\Xi(\mathbf{0}, \zeta^2\mathbf{I})$. Then, for any $\xi \in \mathbb{Z}_L$, we have $\hat{f}(\xi) \sim N(0, \zeta^2\mathcal{V}(\Xi,\xi))$.*

Proof For any $\xi \in \mathbb{Z}_L$, we can write

$$\hat{f}(\xi) = \sum_{\kappa\in\Xi} f(k)\exp(2\pi i\xi k/L);$$

using the fact that Ξ is symmetric about the origin and denoting $\Xi_+ = \Xi \cap \{1,\dots,L/2\}$, this may be rewritten as

$$\hat{f}(\xi) = f(0)\mathbb{1}_{\{0\in\Xi\}} + 2\sum_{\kappa\in\Xi_+} f(k)\cos(2\pi\xi k/L). \tag{5.17}$$

Since $\{f(k)\}_{k\in\Xi}$ is a collection of i.i.d. $N(0,\zeta^2)$ random variables (and 0 for $k \notin \Xi$), we deduce that $\hat{f}(\xi)$ is Gaussian with mean 0 and variance $\zeta^2 \mathbb{1}_{\{0\in\Xi\}} + 4\sum_{k\in\Xi_+} \zeta^2 \cos^2(2\pi\xi k/L) = \zeta^2 \mathcal{V}(\Xi,\xi)$, as desired. □

5.3 Genericity of Support for Symmetric Bernoulli–Gaussian and $N^{\text{symm}}_{[-s,s]}(0,\zeta^2 I)$ Distributions

In this section, we demonstrate that two major classes of distributions in the regime of moderate sparsity—namely, the sparse symmetric Bernoulli–Gaussian distribution and the $N^{\text{symm}}_{[-s,s]}(0,\zeta^2 I)$ distribution—exhibit cosine-genericity of support.

To begin with, we recall the definitions of support sets of signals being typically s-sparse (Definition 42) and Γ-cosine generic (Definition 43). We would also need to make use of Bernstein's inequality, which we state below.

Lemma 30 (Bernstein's Inequality) [15] *Let $X_1,\ldots,X_n$ be mean zero random variables, and let $|X_i| \le M$ for all $1 \le i \le n$. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{1}{3}Mt}\right).$$

We first take up the support properties of the symmetric Bernoulli–Gaussian distribution.

Lemma 31 *For L large enough and $\log^9 L \le s \le L/\log^5 L$, the sparse symmetric Bernoulli distribution with mean 0, variance ζ^2 and sparsity parameter s is typically s-sparse with sparsity constants $(1/2, 2)$, and is $s/32$-cosine generic.*

Proof We first show that the sparse symmetric Bernoulli distribution is typically s-sparse, which amounts to showing that $|\Xi|$ is of order s with high probability. But we observe that $|\Xi| = Y_0 + 2\sum_{i=1}^{\lfloor (L-1)/2\rfloor} Y_i$, where each Y_i is a Bernoulli(s/L) random variable. We immediately conclude that $\mathbb{E}[|\Xi|] = s+1$. Applying Bernstein's inequality (c.f. Lemma 30) to the centred random variables $X_i = 2(Y_i - \mathbb{E}[Y_i])$ (for $i \ge 1$) and $X_0 = Y_0 - \mathbb{E}[Y_0]$ with $t = \frac{1}{2}s$ and $M = 2$, we obtain

$$\mathbb{P}\left(\big||\Xi| - \mathbb{E}[|\Xi|]\big| \ge t\right) \le \exp\left(-\frac{\frac{1}{8}s^2}{4(\frac{L}{2}+1)\frac{s}{L}(1-\frac{s}{L}) + \frac{1}{3}s}\right) = \exp(-cs(1+o_L(1)))$$

for some positive number c. Since $\mathbb{E}[\Xi] = s+1$, we deduce that $\frac{1}{2}s \le |\Xi| \le 2s$ with probability $1 - o_L(1)$, implying that the sparse symmetric Bernoulli distribution is typically s-sparse, with sparsity constants $(1/2, 2)$.

To demonstrate that the symmetric Bernoulli distribution is cosine generic with the parameters as claimed in the statement of this lemma, we first compute, for a fixed

$\xi \in \mathbb{Z}_L$, the expectation $\mathbb{E}[\mathcal{V}(\Xi, \xi)]$. To this end, we may write

$$\mathcal{V}(\Xi, \xi) = Y_0 + 4 \sum_{k=1}^{\lfloor (L-1)/2 \rfloor} \cos^2(2\pi \xi k/L) Y_i, \tag{5.18}$$

where the random variables Y_i are defined as above. Then

$$\mathbb{E}[\mathcal{V}(\Xi, \xi)] = \frac{s}{L} + \frac{4s}{L} \sum_{k=1}^{\lfloor (L-1)/2 \rfloor} \cos^2(2\pi \xi k/L). \tag{5.19}$$

Setting $\mu = \exp(2\pi i \xi/L)$, this reduces to

$$\begin{aligned} \mathbb{E}[\mathcal{V}(\Xi, \xi)] &= \frac{s}{L} + \frac{s}{L} \sum_{k=1}^{\lfloor (L-1)/2 \rfloor} |\mu^k + \mu^{-k}|^2 = \frac{s}{L} + \frac{s}{L} \sum_{k=1}^{\lfloor (L-1)/2 \rfloor} (2 + 2\Re(\mu^{2k})) \\ &= s(1 + o_L(1)) + s \cdot \frac{1}{L} \Re \left(\sum_{k=1}^{\lfloor (L-1)/2 \rfloor} \mu^{2k} \right) \\ &= s \cdot \left[1 + \frac{2}{L} \Re \left(\mu^2 \cdot \frac{1 - \mu^{L-\alpha}}{1 - \mu^2} \right) + o_L(1) \right], \end{aligned} \tag{5.20}$$

where $\alpha = 1$ or 2, depending on whether L is odd or even.

By considering the magnitude of the quantity $\left(\mu^2 \cdot \frac{1-\mu^{L-\alpha}}{1-\mu^2} \right)$, we deduce from (5.20) that for large s, L the expectation

$$\begin{aligned} &\mathbb{E}[\mathcal{V}(\Xi, \xi)] \geq s/2 \\ &\text{unless } |1 - \mu^2| < 8/L \iff |2\xi/L - \delta| < 8/L \text{ for } \delta = 0, \pm 1 \\ &\iff |\xi - \delta \cdot \frac{L}{2}| < 4 \text{ for } \delta = 0, \pm 1. \end{aligned} \tag{5.21}$$

It remains to deal with the frequencies ξ that satisfy (5.21). We will demonstrate the details for the case $\delta = 0$; the computations for $\delta = \pm 1$ are similar, and indeed can be reduced to the consideration of $\delta = 0$ by making a change of variables $\hat{\xi} = \xi - \delta \cdot \frac{L}{2}$ and observing that $\cos^2(2\pi \hat{\xi} k/L) = \cos^2(2\pi \xi k/L)$.

Therefore, we reduce ourselves to considering the frequencies ξ [in the context of (5.21)] such that $|\xi| < 4$. We then invoke (5.19) and lower bound

$$\mathbb{E}[\mathcal{V}(\Xi, \xi)] \geq \frac{4s}{L} \sum_{k=1}^{L/32} \cos^2(2\pi \xi k/L) \geq \frac{4s}{L} \sum_{k=1}^{L/32} \cos^2(2\pi/8) = \frac{4s}{L} \cdot \frac{L}{32} \cdot \frac{1}{2} = s/16. \tag{5.22}$$

We combine our analyses of the two classes of frequencies, summarize it as:

$$\mathbb{E}[\mathcal{V}(\Xi,\xi)] \geq s/16 \quad \forall \xi \in \mathbb{Z}_L. \tag{5.23}$$

We centre the Y_i-s in (5.18) by their expectations, and define the centered random variables $X_0 = Y_0 - \mathbb{E}[Y_0]$ and for $1 \leq i \leq \lfloor (L-1)/2 \rfloor$, $X_i = 4\cos^2(2\pi\xi k/L)(Y_i - \mathbb{E}[Y_i])$. Then we may write

$$\mathcal{V}(\Xi,\xi) - \mathbb{E}[\mathcal{V}(\Xi,\xi)] = \sum_{i=0}^{\lfloor (L-1)/2 \rfloor} X_i.$$

Notice that, for any $i \geq 0$, $\mathrm{Var}[X_i] \leq 4s/L(1-s/L)$, so that

$$\sum_{i=0}^{\lfloor (L-1)/2 \rfloor} \mathbb{E}[X_i^2] \leq 2s(1-s/L)(1+o_L(1)).$$

Applying Bernstein's inequality (c.f. Lemma 30) with $t = s/32$ and $M = 4$, we proceed as

$$\begin{aligned}
&\mathbb{P}\left(\left|\mathcal{V}(\Xi,\xi) - \mathbb{E}[|\mathcal{V}(\Xi,\xi)]\right| \geq s/32\right) \\
&\quad \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=0}^{\lfloor (L-1)/2 \rfloor} \mathbb{E}[X_i^2] + \frac{1}{3}Mt}\right) \\
&\quad \leq \exp\left(-\frac{\frac{1}{2048}s^2}{2s(1-s/L)(1+o_L(1)) + \frac{1}{24}s}\right). \\
&\quad \leq \exp(-s/10^4(1+o_L(1))).
\end{aligned}$$

By a union bound, we may further deduce that

$$\mathbb{P}\left(\exists \xi \in \mathbb{Z}_L \text{ such that } \left|\mathcal{V}(\Xi,\xi) - \mathbb{E}[\mathcal{V}(\Xi,\xi)]\right| \geq s/32\right) \leq L\exp(-s/10^4(1+o_L(1))). \tag{5.24}$$

The right-hand side is $o_L(1)$ as soon as $s \gg 10^4 \log L$.

On the complement of the event in (5.24), that is, when $\{\left|\mathcal{V}(\Xi,\xi) - \mathbb{E}[\mathcal{V}(\Xi,\xi)]\right| < s/32 \quad \forall \xi \in \mathbb{Z}_L\}$, we may deduce from (5.21) that for all $\xi \in \mathbb{Z}_L$

$$\begin{aligned}
&\mathcal{V}(\Xi,\xi) \\
&\quad \geq \mathbb{E}[\mathcal{V}(\Xi,\xi)] - \left|\mathcal{V}(\Xi,\xi) - \mathbb{E}[\mathcal{V}(\Xi,\xi)]\right| \\
&\quad \geq s/16 - s/32 \\
&\quad = s/32.
\end{aligned}$$

This shows that $\min_{\xi\in\mathbb{Z}_L} \mathcal{V}(\Xi,\xi) \geq s/32$ with probability $1-o_L(1)$ (in the random subset Ξ), thereby establishing the claim that for $s \geq c\log L$ and L large enough, the random subset Ξ is $s/32$-cosine generic. □

Finally, we end this section with a study of the support properties of the $N^{\mathrm{symm}}_{[-s,s]}(0,\zeta^2 I)$ distribution.

Lemma 32 *For s, L large enough, the deterministic subset $\{[-s,s]\cap\mathbb{Z}_L\}$ is typically s-sparse and is $s/16$-cosine generic.*

Proof The (deterministic) subset Ξ has size exactly $2s+1$, therefore Ξ is trivially *typically s-sparse*. It remains to show the cosine genericity of Ξ.

For any $\xi\in\mathbb{Z}_L$, we have

$$\begin{aligned}\mathcal{V}(\Xi,\xi)&=1+2\sum_{k\in[-s,s]\setminus\{0\}}\cos^2(2\pi\xi k/L)\\&=-1+2\sum_{k=-s}^{s}\cos^2(2\pi\xi k/L)\\&=-3+\sum_{k=0}^{s}|\exp(2\pi i\xi k/L)+\exp(-2\pi i\xi k/L)|^2.\end{aligned}\tag{5.25}$$

Setting $\omega=\exp(2\pi i\xi/L)$, we may proceed as

$$\begin{aligned}\mathcal{V}(\Xi,\xi)&=-3+\sum_{k=0}^{s}|\omega^k+\omega^{-k}|^2=-3+\sum_{k=0}^{s}(2+2\Re(\omega^{2k}))\\&=2s-1+2\Re\left(\sum_{k=0}^{s}\omega^{2k}\right)=2s-1+2\Re\left(\frac{1-\omega^{2s+2}}{1-\omega^2}\right).\end{aligned}$$

The last equation implies that

$$\mathcal{V}(\Xi,\xi)\geq s(1+o(1))\tag{5.26}$$

unless $\left|\Re\left(\frac{1-\omega^{s+2}}{1-\omega^2}\right)\right|>s/4$, which would in particular imply that

$$\left|\frac{1-\omega^{s+2}}{1-\omega^2}\right|>s/4.\tag{5.27}$$

Recalling the definition of ω, and observing that $|1-\omega^{s+2}|\leq 2$ we may deduce that (5.27) is true only if $|1-\omega^2|<8/s$. Recalling that $\omega=\exp(2\pi i\xi/L)$, we deduce that for large enough s, the inequality holds $|1-\omega^2|<8/s$ only if $|2\xi/L-\delta\cdot\frac{L}{2}|\leq 8/s(1+o_s(1))$, where $\delta=0,\pm1$. As in the proof of Lemma 31, we focus on the case

$\delta = 0$, noting in passing that the cases $\delta = \pm 1$ are similar and are easily dealt with using a simple change of variables from ξ.

When $\delta = 0$, we are considering frequencies ξ such that $\xi/L < 4/s$. This in particular implies that for all $|k| \leq s/32$, we have $|2\pi\xi k/L| < \pi/4$, implying $\cos^2(2\pi\xi k/L) \geq 1/2$.

We now proceed to lower bound $\mathcal{V}(\Xi, \xi)$ for $\xi \in \mathbb{Z}_L$. If $\xi \in \mathbb{Z}_L$ is such that $\omega = \exp(2\pi i \xi/L)$ *does not* satisfy (5.27), then by (5.26) we conclude that $\mathcal{V}(\Xi, \xi) \geq s(1 + o(1))$.

If $\xi \in \mathbb{Z}_L$ is such that $\omega = \exp(2\pi i \xi/L)$ satisfies (5.27), then we proceed as follows. Using (5.25), we may lower bound $\mathcal{V}(\Xi, \xi)$ as

$$\begin{aligned}
&\mathcal{V}(\Xi, \xi) \\
&\quad = -1 + 2\sum_{k=-s}^{s} \cos^2(2\pi\xi k/L) \\
&\quad \geq 1 + 2 \sum_{1 \leq |k| \leq s/32} \cos^2(2\pi\xi k/L) \\
&\quad \geq s/16. \qquad (5.28)
\end{aligned}$$

Combining (5.26) and (5.28), we deduce that $\mathcal{V}(\Xi, \xi) \geq s/16 \quad \forall \xi \in \mathbb{Z}_L$, thereby showing that Ξ is $s/16$-cosine generic and completing the proof of the lemma. □

6 Results on the Curvature of D_{KL}

In this section, we provide the proofs of several propositions pertaining to the curvature of the KL divergence for the MRA model.

6.1 Moment Difference Tensors and D_{KL}

Proof of Proposition 14 We discuss (i); the case of (ii) would be similar. We recall that the probability distribution p_θ as well as $D_{KL}(p_\theta \| p_\varphi)$ are invariant under the action of $\mathcal{G}$, i.e., invariant under the transformations $\theta \mapsto G \cdot \theta$ for $G \in \mathcal{G}$. As a result, for $\varrho(\theta, \theta_0)$ small enough (equivalently, $\|\theta - \theta_0\|_2$ small enough), we may assume without loss of generality that $\varrho(\theta, \theta_0) = \frac{1}{\sqrt{L}}\|\theta - \theta_0\|_2$ (c.f., [4]; esp. the proof of Theorem 4 therein).

The dimension of the Hessian of $D_{KL}(p_{\theta_0} || p_\theta)$ depends on the local dimension of the parameter space at the point θ_0, which is the same as $k = |\operatorname{supp}(\theta_0)|$. The lower bound on D_{KL} in (i) implies that

$$K_1(\sigma)^{1/2}\mathrm{Id}_k \preceq I(\theta_0) \iff I(\theta_0)^{-1} \preceq K_1(\sigma)^{-1/2}\mathrm{Id}_k,$$

where Id_k is the $k \times k$ identity matrix, and $\preceq$ denotes domination in the sense of non-negative definite matrices. Since $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to N(0, I(\theta_0)^{-1})$, setting $\mathcal{Z}_k \sim$

$N(0, \mathrm{Id}_k)$, we may deduce that as $n \to \infty$ we have the distributional convergence

$$\sqrt{n}\|\tilde{\theta}_n - \theta_0\|_2 = \sqrt{n\|\tilde{\theta}_n - \theta_0\|_2^2} \to \|I(\theta_0)^{-1/2}\mathcal{Z}_k\|_2. \tag{6.1}$$

On the other hand, we have

$$\|I(\theta_0)^{-1/2}\mathcal{Z}_k\|_2 = \sqrt{\langle \mathcal{Z}_k, I(\theta_0)^{-1}\mathcal{Z}_k\rangle} \leq K_1(\sigma)^{-1/2}\|\mathcal{Z}_k\|_2. \tag{6.2}$$

Thus, $\sqrt{n}\varrho_\theta, \theta_0 = \sqrt{n}\frac{1}{\sqrt{L}} \cdot \|\tilde{\theta}_n - \theta_0\|_2 = O_p(K_1(\sigma)^{-1/2}/\sqrt{L})$, as desired.

We note in passing that $\|\mathcal{Z}_k\|_2$ is a $\sqrt{\chi^2(k)}$ distribution. □

Recall that for any $\theta = (\theta_1, \ldots, \theta_L) \in \mathbb{R}^L$, we denote $\overline{\theta} = \frac{1}{L}\sum_{i=1}^L \theta_i$. This leads us to the fact that $\theta^* := \mathbb{E}_{\mathcal{G}}[G\theta] = \overline{\theta} \cdot \mathbb{1}$, where $\mathbb{1} = (1, 1, \ldots, 1) \in \mathbb{R}^L$ is the all ones vector in L dimensions. Finally, we denote by $\tilde{\theta}$ the centred version of θ, that is, $\tilde{\theta} = \theta - \theta^* = \theta - \mathbb{E}_{\mathcal{G}}[G\theta]$. We observe that

$$\mathbb{E}_{\mathcal{G}}[G\tilde{\theta}] = \mathbb{E}_{\mathcal{G}}[\theta - \theta^*] = \mathbb{E}_{\mathcal{G}}[\theta] - \mathbb{E}_{\mathcal{G}}[\theta^*] = \theta^* - \theta^* = 0. \tag{6.3}$$

Notice further that, with the above notations, we may write

$$\Delta_1(\theta, \varphi) = (\overline{\theta} - \overline{\varphi})\mathbb{1}. \tag{6.4}$$

Towards the proofs of Propositions 17 and 18, we will now present a comparison between the second moment difference tensors for the centred and uncentred versions of two vectors θ and φ. To this end, we state the following Proposition.

Proposition 33 *We have,*

$$\Delta_2(\theta, \varphi) = \Delta_2(\tilde{\theta}, \tilde{\varphi}) + (\overline{\theta}^2 - \overline{\varphi}^2) \cdot \mathbb{1} \otimes \mathbb{1}.$$

Proof We have,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[(G(\tilde{\theta} + \theta^*))^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta} + G\theta^*))^{\otimes 2}] \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta} + \theta^*))^{\otimes 2}] \quad [\text{since } \theta^* \text{ is } \mathcal{G}\text{-invariant}] \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta})^{\otimes 2}] + \mathbb{E}_{\mathcal{G}}[G\tilde{\theta} \otimes \theta^*] + \mathbb{E}_{\mathcal{G}}[\theta^* \otimes G\tilde{\theta}] + \theta^* \otimes \theta^* \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta})^{\otimes 2}] + \mathbb{E}_{\mathcal{G}}[G\tilde{\theta}] \otimes \theta^* + \theta^* \otimes \mathbb{E}_{\mathcal{G}}[G\tilde{\theta}] + \theta^* \otimes \theta^* \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta})^{\otimes 2}] + \theta^* \otimes \theta^* \quad [\text{using (6.3)}] \\
&= \mathbb{E}_{\mathcal{G}}[(G\tilde{\theta})^{\otimes 2}] + \overline{\theta}^2 \mathbb{1} \otimes \mathbb{1}.
\end{aligned} \tag{6.5}$$

In view of (6.5), we may write

$$\Delta_2(\theta,\varphi) = \mathbb{E}_{\mathcal{G}}[(G\theta)^{\otimes 2}] - \mathbb{E}_{\mathcal{G}}[(G\varphi)^{\otimes 2}] = \Delta_2(\tilde{\theta},\tilde{\varphi}) + (\overline{\theta}^2 - \overline{\varphi}^2)\cdot \mathbb{1}\otimes\mathbb{1}, \quad (6.6)$$

as desired. □

We now proceed to establish Proposition 17.

Proof of Proposition 17 Observe that, $|\overline{\theta}^2 - \overline{\varphi}^2| = |\overline{\theta}+\overline{\varphi}|\cdot|\overline{\theta}-\overline{\varphi}| \le (\|\theta\|_\infty + \|\varphi\|_\infty)\cdot |\overline{\theta}-\overline{\varphi}|$. This implies, in particular, that

$$\|\Delta_2(\tilde{\theta},\tilde{\varphi})\|_F \ge \|\Delta_2(\theta,\varphi)\|_F - (\|\theta\|_\infty + \|\varphi\|_\infty)\cdot|\overline{\theta}-\overline{\varphi}|\cdot\|\mathbb{1}\otimes\mathbb{1}\|_F. \quad (6.7)$$

Now, Theorem 16 implies that

$$\begin{aligned} &D_{KL}(\tilde{\theta}||\tilde{\varphi}) \\ &\quad\ge C\cdot\|\Delta_2(\tilde{\theta},\tilde{\varphi})\|_F^2/\sigma^4 \\ &\quad\ge C\cdot\left(\|\Delta_2(\theta,\varphi)\|_F - (\|\theta\|_\infty+\|\varphi\|_\infty)\cdot|\overline{\theta}-\overline{\varphi}|\cdot\|\mathbb{1}\otimes\mathbb{1}\|_F\right)^2/\sigma^4 \\ &\quad\ge C\cdot\left(\frac{3}{4}\|\Delta_2(\theta,\varphi)\|_F^2 - 3(\|\theta\|_\infty+\|\varphi\|_\infty)^2\cdot|\overline{\theta}-\overline{\varphi}|^2\cdot\|\mathbb{1}\otimes\mathbb{1}\|_F^2\right)/\sigma^4 \end{aligned}$$

for a positive number C, where in the last step we use Proposition 35.

Combining the above with Lemma 15 we obtain

$$\begin{aligned} &D_{KL}(\theta||\varphi) \\ &\ge \frac{1}{2}|\overline{\theta}-\overline{\varphi}|^2\|\mathbb{1}\|_2^2\cdot\sigma^{-2} + C\left(\frac{3}{4}\|\Delta_2(\theta,\varphi)\|_F^2 - 3(\|\theta\|_\infty+\|\varphi\|_\infty)^2|\overline{\theta}-\overline{\varphi}|^2\|\mathbb{1}\otimes\mathbb{1}\|_F^2\right) \\ &\ge \sigma^{-4}\frac{3C}{4}\|\Delta_2(\theta,\varphi)\|_F^2 + |\overline{\theta}-\overline{\varphi}|^2\left(\frac{1}{2}\sigma^{-2}\|\mathbb{1}\|_2^2 - 3C\sigma^{-4}(\|\theta\|_\infty+\|\varphi\|_\infty)^2\|\mathbb{1}\otimes\mathbb{1}\|_F^2\right). \end{aligned} \quad (6.8)$$

We now make use of the fact that the signal class $\mathcal{T}$ is bounded (in the deterministic setting), and in the case of generative models, the random signal is bounded with high probability.

We then consider the term

$$\left(\frac{1}{2}\sigma^{-2}\|\mathbb{1}\|_2^2 - 3C\sigma^{-4}(\|\theta\|_\infty+\|\varphi\|_\infty)^2\|\mathbb{1}\otimes\mathbb{1}\|_F^2\right)$$

on the right-hand side of (6.8), and observe that when σ is large enough—that is, $\sigma \ge \sigma_0(L)$ for some threshold $\sigma_0(L)$, we have

$$\left(\sigma^{-2}\|\mathbb{1}\|_2^2 - 3\sigma^{-4}(\|\theta\|_\infty+\|\varphi\|_\infty)^2\|\mathbb{1}\otimes\mathbb{1}\|_F^2\right) \ge \frac{1}{4}\sigma^{-2}\|\mathbb{1}\|_2^2. \quad (6.9)$$

Combining (6.8) and (6.9), we obtain

$$\begin{aligned} &D_{KL}(\theta||\varphi) \\ &\quad \geq \sigma^{-4} \cdot \frac{3C}{4} \|\Delta_2(\theta,\varphi)\|_F^2 + \sigma^{-2} \cdot \frac{1}{4} |\overline{\theta} - \overline{\varphi}|^2 \|\mathbb{1}\|_2^2 \\ &\quad \geq \sigma^{-4} \cdot \frac{3C}{4} \|\Delta_2(\theta,\varphi)\|_F^2. \end{aligned} \tag{6.10}$$

□

Remark 34 We observe that equality can hold in (6.10), whenever $\overline{\theta} = \overline{\varphi}$. This is indeed possible for specific directions of approach of φ to the signal θ when θ lies in the interior of the signal class. The standard signal classes considered in MRA, in this paper as well as otherwise, and also the generative models considered in this paper, have their interiors account for their full Lebesgue measure, so nearly all signals θ do in fact have such a bad direction of approach where equality in (6.10) holds.

We continue on to the proof of Proposition 18.

Proof of Proposition 18 When, for some θ, φ, we have $\|\Delta_2(\theta,\varphi)\|_F \leq c\rho(\theta,\varphi)^2$, then we may proceed to analyse the order of $D_{KL}(\theta\|\varphi)$ as follows. Combining Lemma 15 and Theorem 16 applied with $k = 3$, and noting that $\Delta_1(\tilde{\theta}\|\tilde{\varphi}) = 0$, we may proceed as

$$\begin{aligned} &D_{KL}(\theta||\varphi) \\ &\quad = D_{KL}(p_{\tilde{\theta}}||p_{\tilde{\varphi}}) + \frac{1}{2\sigma^2}\|\Delta_1(\theta,\varphi)\|^2 \\ &\quad \leq 2\sum_{m=1}^{2} \frac{\|\Delta_m(\tilde{\theta},\tilde{\varphi})\|^2}{\sigma^{2m} m!} + C\frac{\|\tilde{\theta}\|_2^4 \rho(\tilde{\theta},\tilde{\varphi})^2}{\sigma^6} + \frac{1}{2\sigma^2}\|\Delta_1(\theta,\varphi)\|^2 \\ &\quad = \frac{1}{2\sigma^2} \cdot |\overline{\theta} - \overline{\varphi}|^2 \|\mathbb{1}\|_2^2 + \frac{1}{\sigma^4} \cdot \|\Delta_2(\tilde{\theta},\tilde{\varphi})\|_F^2 + C\frac{\|\tilde{\theta}\|_2^4 \rho(\tilde{\theta},\tilde{\varphi})^2}{\sigma^6} \quad \text{[using (6.4)]} \\ &\quad = \frac{1}{2\sigma^2} \cdot |\overline{\theta} - \overline{\varphi}|^2 \|\mathbb{1}\|_2^2 \left(1 + \frac{C_1\|\theta\|_2^4}{\sigma^4}\right) + \frac{1}{\sigma^4} \cdot \|\Delta_2(\tilde{\theta},\tilde{\varphi})\|_F^2 + \frac{2C\|\theta\|_2^4}{\sigma^6}\rho(\theta,\varphi)^2, \end{aligned} \tag{6.11}$$

where, in the last step, we have used Proposition 36.

Therefore, if θ, φ are such that $\overline{\theta} = \overline{\varphi}$ and $\|\Delta_2(\theta,\varphi)\|_F \leq c\rho(\theta,\varphi)^2$, we may conclude from (6.11) that

$$D_{KL}(\theta||\varphi) \leq \frac{1}{\sigma^4} \cdot C \cdot \rho(\theta,\varphi)^4 + 2C\frac{\|\theta\|_2^4}{\sigma^6} \cdot \rho(\theta,\varphi)^2.$$

For small enough $\rho(\theta, \varphi)$, the quadratic term involving $\rho(\theta, \varphi)^2$ dominates in the above, and we have

$$D_{KL}(\theta||\varphi) \leq 4C \frac{\|\theta\|_2^4}{\sigma^6} \cdot \rho(\theta, \varphi)^2 \tag{6.12}$$

for some positive number C and small enough $\rho(\theta, \varphi)$. □

We complete this section with the auxiliary Propositions 35 and 36.

Proposition 35 *Let $a, b > 0$. Then we have*

$$(a - b)^2 \geq \frac{3}{4}a^2 - 3b^2.$$

Proof We observe that

$$2ab = 2 \cdot \frac{1}{2}a \cdot 2b \leq \frac{1}{4}a^2 + 4b^2. \tag{6.13}$$

We may then expand $(a - b)^2 = a^2 + b^2 - 2ab$ and use (6.13) to lower bound the $-2ab$ term. This completes the proof. □

Proposition 36 *We have,*

$$\rho(\tilde{\theta}, \tilde{\varphi})^2 \leq 2\rho(\theta, \varphi)^2 + 2|\overline{\theta} - \overline{\varphi}|^2 \|\mathbb{1}\|_2^2.$$

Proof Recall that $\theta^* = \mathbb{E}_{\mathcal{G}}[G\theta]$ is $\mathcal{G}$-invariant, i.e., $G\theta^* = \theta^* \forall G \in \mathcal{G}$; the same holds true for φ^*. For any $G \in \mathcal{G}$, we may write

$$\begin{aligned}
&\|\tilde{\theta} - G\tilde{\varphi}\|_2^2 \\
&\quad = \|(\theta - \theta^*) - G(\varphi - \varphi^*)\|_2^2 \\
&\quad = \|(\theta - G\varphi) - (\theta^* - \varphi^*)\|_2^2 \quad \text{[Since } \theta^*, \varphi^* \text{ are } \mathcal{G}\text{-invariant]} \\
&\quad \leq 2\|\theta - G\varphi\|_2^2 + 2\|\theta^* - \varphi^*\|_2^2.
\end{aligned}$$

We summarize the above computations as

$$\|\tilde{\theta} - G\tilde{\varphi}\|_2^2 \leq 2\|\theta - G\varphi\|_2^2 + 2\|\theta^* - \varphi^*\|_2^2 \quad \forall G \in \mathcal{G}. \tag{6.14}$$

This implies that, for any $G \in \mathcal{G}$ we have

$$\rho(\tilde{\theta}, \tilde{\varphi})^2 = \min_{\mathfrak{g} \in \mathcal{G}} \|\tilde{\theta} - \mathfrak{g}\tilde{\varphi}\|_2^2 \leq \|\tilde{\theta} - G\tilde{\varphi}\|_2^2 \leq 2\|\theta - G\varphi\|_2^2 + 2\|\theta^* - \varphi^*\|_2^2. \tag{6.15}$$

Taking minimum over $G \in \mathcal{G}$ on the right-hand side of (6.15) and noting that $\theta^* = \overline{\theta}\mathbb{1}$ (and similarly for φ^*), we obtain

$$\rho(\tilde{\theta}, \tilde{\varphi})^2 \leq 2\rho(\theta, \varphi)^2 + 2|\overline{\theta} - \overline{\varphi}|^2 \|\mathbb{1}\|_2^2,$$

as desired. □

6.2 *L*, *s* Dependence of Estimation Rates

It may be noted that, in the context of Proposition 14 part (i), if we have additional information on the dependence of $K_1(\sigma)$ on L and/or s, then we can have informative asymptotic upper bounds on $\varrho(\tilde{\theta}_n, \theta)$ vis-a-vis its dependence on L and/or s.

Proof of Theorem 8 We begin by recalling that by the $\mathcal{G}$-invariance of $D_{KL}(p_\theta \| p_\varphi)$, for $\varrho(\theta, \theta_0)$ (equivalently, $\|\theta - \theta_0\|_2$) small enough, we may assume without loss of generality that $\varrho(\theta, \theta_0) = \frac{1}{\sqrt{L}}\|\theta - \theta_0\|_2$. Since $\|\tilde{\theta}_n - \theta_0\|_2 \to 0$ as $n \to \infty$, This will be true for $\varrho(\tilde{\theta}_n, \theta_0)$ with high probability.

The Dilute Regime. In view of Lemma 5 and Proposition 17, we obtain a local curvature estimate on $D_{KL}(p_\theta \| p_{\theta_0})$ as

$$D_{KL}(p_\theta \| p_{\theta_0}) \geq C \cdot \frac{s}{L\sigma^4} \cdot \|\theta - \theta_0\|_2^2.$$

Thus, we are in the setting of Proposition 14 part (i) with $K_1(\sigma) = C \cdot \frac{s}{L\sigma^4}$. In view of (6.1) and (6.2), we conclude that the limiting distribution of $\sqrt{n}\varrho(\tilde{\theta}, \theta_0)/\sigma^2$ is stochastically dominated by a $C_1\sqrt{\chi^2(s)/s}$ random variable, for a constant $C_1 > 0$. We conclude by noting that the latter random variable is $O_p(1)$.

The Regime of Moderate Sparsity. In the case of generic sparse symmetric signals, it may be seen from (5.6) and Remark 24 that if the support typically s-sparse and is s^τ cosine-generic, then for $\varrho(\theta, \theta_0)$ small we have

$$\|\Delta_2(\theta, \theta_0)\|_F \geq \frac{s^{\tau-4}}{\sqrt{L}}\|\theta - \theta_0\|_2.$$

Hence, by Theorem 16, for such signals we have

$$D_{KL}(p_\theta \| p_{\theta_0}) \geq C \cdot \frac{s^{2(\tau-4)}}{L\sigma^4} \cdot \|\theta - \theta_0\|_2^2.$$

Once again, this places us in the context of Proposition 14 part (i), with

$$K_1(\sigma) = C \cdot \frac{s^{2(\tau-4)}}{L\sigma^4}.$$

Furthermore, in view of (6.1) and (6.2), we may conclude that the limiting distribution of $\sqrt{n}\varrho(\tilde{\theta}, \theta_0)/\sigma^2$ is stochastically dominated by a $C_2 s^{4-\tau}\sqrt{\chi^2(s)}$ random variable, for a constant $C_2 > 0$. The latter random variable is $O_p(s^{4.5-\tau})$.

We finally observe that two significant examples of generic sparse symmetric signals—namely, the Bernoulli–Gaussian distribution and the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ have supports that are typically s-sparse and constant times s-cosine generic. We provide the details in the case of the Bernoulli–Gaussian; the case of the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ is similar. We invoke Lemma 31 to conclude that the symmetric Bernoulli–Gaussian distribution with sparsity s and variance ζ^2 is typically s-sparse (with sparsity constants $(1/2, 2)$) and $s/32$ cosine generic. The analogous Lemma to be applied for the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ distribution is Lemma 32.

Thus, for these two signal distributions, $\tau = 1$ in these settings in the context of the discussion immediately above.

In view of this fact, and the discussion above, the Bernoulli–Gaussian signal ensemble and the $N^{\mathrm{symm}}_{[-s,s]}(0, \zeta^2 I)$ entail estimation rates that, upon scaling by $\sigma^2/\sqrt{n}$, are $O_p(s^{3.5})$. □

Acknowledgements SG was supported in part by the MOE grants R-146-000-250-133, R-146-000-312-114 and MOE-T2EP20121-0013. PR was supported by the NSF awards DMS-1712596, IIS-1838071, DMS-2022448, and DMS-210637. The authors would like to thank Victor-Emmanuel Brunel for stimulating discussions that shaped the direction of this project, Tamir Bendory for bringing to their attention recent literature on phase retrieval, and Michel Goemans for pointing them to the partial digest problem. The authors are grateful to the anonymous referees for their meticulous reading of the manuscript and their prescient suggestions towards its improvement, and especially for pointing out important connections of the present work to the problem of crystallographic phase retrieval.

Appendix

A Appendix: Additional Notations

Definition 37 Let $\{X_n\}_{n\geq 1}$ be a sequence of non-negative random variables, and $\{a_n\}_{n\geq 1}$ is a sequence of positive numbers (deterministic or random). Then:

- By the statement $X_n = O_p(a_n)$ we mean that, for every $\varepsilon > 0$, there exists $0 < C(\varepsilon) < \infty$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left[X_n/a_n \leq C(\varepsilon)\right] \geq 1 - \varepsilon.$$

- By the statement $X_n = \Omega_p(a_n)$ we mean that, for every $\varepsilon > 0$, there exists $0 < c(\varepsilon) < \infty$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left[X_n/a_n \geq c(\varepsilon)\right] \geq 1 - \varepsilon.$$

- By the statement $X_n = \Theta_p(a_n)$ we mean that for every $\varepsilon > 0$, there exist $0 < c(\varepsilon) < C(\varepsilon) < \infty$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left[c(\varepsilon) \leq X_n/a_n \leq C(\varepsilon)\right] \geq 1 - \varepsilon.$$

Further, $\|\cdot\|_F$ will denote the Frobenius norm of a matrix, and the expectation $\mathbb{E}_G$ will be taken with respect to G chosen uniformly from the group of isometries $\mathcal{G}$.

For any positive integer m, by the symbol $[m]$ we denote the set $\{1, \ldots, m\}$.

For two sequences of positive numbers $(a_k)_{k>0}$ and $(b_k)_{k>0}$, we write $a_k \ll b_k$ when we have $b_k/a_k \to \infty$ as $k \to \infty$.

A sequence of events $\{E_m\}_{m\geq 1}$, defined with respect to probability measures $\mathbb{P}_m$, is said to occur with high probability if $\mathbb{P}_m[E_m] \to 1$ as $m \to \infty$.

For any $\theta = (\theta_1, \ldots, \theta_L) \in \mathbb{R}^L$, we denote $\overline{\theta} = \frac{1}{L}\sum_{i=1}^{L} \theta_i$.

B Appendix: Bernoulli–Gaussian Distributions

We define the notion of the Bernoulli–Gaussian distribution, and the symmetric version thereof. For that, we first define the notion of a Gaussian distribution indexed by a subset of $\mathbb{Z}_L$.

Definition 38 (Subset-indexed Gaussian distributions) Let $A \subset \mathbb{Z}_L$, $\mu : \mathbb{Z}_L \to \mathbb{R}$ a function supported on A and Σ be a positive definite $|A| \times |A|$ matrix. Then the Gaussian distribution indexed by A with mean μ and covariance Σ, denoted $N_A(0, \Sigma)$, is the random vector $(\eta_k)_{k\in\mathbb{Z}_L}$, with $\eta_k = 0$ for $k \in A^{\complement}$, and $(\eta_k)_{k\in A}$ is the $|A|$-dimensional Gaussian random vector with mean μ and covariance Σ.

This allows us to define the Bernoulli–Gaussian distribution, a key property of which is that the support is chosen at random according to a Bernoulli sampling scheme.

Definition 39 (*Bernoulli–Gaussian distribution*) Let $s \in [L]$ and $\Xi \subset \mathbb{Z}_L$ be a random subset obtained by selecting each member of $\mathbb{Z}_L$ independently with probability s/L. The Bernoulli–Gaussian distribution on Z_L with variance ζ^2 and sparsity s is then defined as the Gaussian distribution indexed by Ξ with mean $\mathbf{0}$ and covariance $\zeta^2 I$; in other words the random variable $N_\Xi(\mathbf{0}, \zeta^2 I)$, with the Gaussian entries being statistically independent of the support Ξ.

Next, we introduce the concept of a standard symmetric Gaussian random variable indexed by a subset of $\mathbb{Z}_L$. To introduce the notion of a symmetric signal, we first recall the notion of the *standard parametrization* of $\mathbb{Z}_L$ (1.7).

We are now ready to define

Definition 40 (*Symmetric subset-indexed Gaussian distributions*) Let $\mathbb{Z}_L$ be in the standard enumeration (1.7), let $A \subset \mathbb{Z}_L$ be symmetric, i.e. $A = -A$ and let $\rho > 0$. Let $A_+ := \{0, \ldots, \lfloor (L-1)/2 \rfloor\} \cap A$, and let $(X_k)_{k\in\mathbb{Z}_L}$ denote the random variable $N_{A_+}(0, \zeta^2 I)$. Then the symmetric Gaussian distribution indexed by A with mean 0 and variance ζ^2, denoted $N_A^{\mathrm{symm}}(0, \zeta^2 I)$, is the random vector $(\eta_k)_{k\in\mathbb{Z}_L}$ with $\eta_k = X_{|k|}$.

Finally, all of the above taken together allows us to define

Definition 41 (*Symmetric Bernoulli–Gaussian distribution*) Let $\mathbb{Z}_L$ be in the standard enumeration (1.7). Let $\Xi_0 \subset \mathbb{Z}_L^+ = \{0, \ldots, \lfloor (L-1)/2 \rfloor\}$ be a random subset obtained by selecting each member of $\mathbb{Z}_L^+$ independently with probability s/L, and consider the symmetric subset $\Xi := \Xi_0 \cup (-\Xi_0)$. Then the symmetric Bernoulli–Gaussian distribution with mean zero, variance ζ^2 and sparsity parameter s is the distribution $N_\Xi^{\mathrm{symm}}(0, \zeta^2 I)$, with the Gaussian entries being statistically independent of the support Ξ.

Heuristically, the symmetric Bernoulli–Gaussian distribution is obtained by taking a Bernoulli–Gaussian random variable on the positive part of $\mathbb{Z}_L$ and extending it to all of $\mathbb{Z}_L$ by making it symmetric about the origin.

C Appendix: Generic Sparse Signals

We introduce the notions of signal support sets that are *typically s-sparse* and Γ-cosine generic.

Definition 42 Let $\alpha, \beta > 0$ be fixed numbers and $s \in [L]$ be a parameter that possibly depends on L. A probability distribution over subsets $\Xi \subset \mathbb{Z}_L$ is said to be *typically s-sparse* with sparsity constants (α, β) if we have $\alpha \cdot s \leq |\Xi| \leq \beta \cdot s$ with probability $1 - o_L(1)$.

To introduce the concept of cosine-genericity of a set, we first define the cosine functional of a set $\Xi \subset \mathbb{Z}_L$ for an element $a \in \mathbb{Z}_L$.

For $\Xi \subset \mathbb{Z}_L$ and $a \in \mathbb{Z}_L$, define

$$\mathcal{V}(\Xi, a) = \mathbb{1}_{\{0 \in \Xi\}} + 2 \sum_{k \in \Xi \setminus \{0\}} \cos^2(2\pi ak/L), \tag{C.1}$$

where $\mathbb{1}_A$ denotes the indicator function of the event A.

Then we are ready to introduce

Definition 43 Let $\Gamma > 0$ be a parameter, possibly depending on L. A probability distribution over subsets $\Xi \subset \mathbb{Z}_L$ is said to be Γ-*cosine generic* if, with probability $1 - o_L(1)$, we have $\min_{a \in \mathbb{Z}_L} \mathcal{V}(\Xi, a) \geq \Gamma(1 - o_L(1))$.

Equivalently, we say that the random variable Ξ is cosine generic with parameter Γ. Cosine genericity of a (random) set is a condition that aims to ensure that, with high probability, the set under consideration is sufficiently generic, in the sense that there are no specialized algebraic or arithmetic relations satisfied by the elements of the set which would make $\min_{a \in \mathbb{Z}_L} \mathcal{V}(\Xi, a)$ small.

Putting all of the above together, we may introduce the generic s-sparse symmetric signals.

Definition 44 Let $s \in [L]$ be a parameter, possibly depending on L, and $\alpha, \beta, \zeta, \tau > 0$ be fixed. We call a random signal $\theta : \mathbb{Z}_L \to \mathbb{R}$ to be a generic s-sparse symmetric signal with dispersion ζ^2, sparsity constants α, β and index τ if the following hold:

- The support Ξ of θ is typically s-sparse with sparsity constants $(\alpha, \beta)$and s^τ-cosine generic.
- $\theta \sim N_\Xi^{\text{symm}}(0, \zeta^2 I)$, with the non-zero entries of θ being statistically independent of Ξ.

D Appendix: On the Size of Collision Free Sets

In this section, we provide detailed arguments for the assertions that the size of a collision-free subset $A \subset \mathbb{Z}_L$ is maximally $O(L^{1/2})$ and typically $O(L^{1/3})$.

To this end, we let $1 \le k \le L$, and we consider a subset $B \subset \mathbb{Z}_L$ of size $|B| = k$. If B is collision-free, then B entails $k(k-1)$ distinct differences between its points; we call this set of differences D. For $x \in \mathbb{Z}_L \setminus B$, we want to understand size restrictions on $|B|$ that enable $B \cup \{x\}$ to be a collision-free set. If $B \cup \{x\}$ has to be collision-free, we note that for any fixed $u \in B$, the difference $x - u$ needs to be $\notin D$. This rules out $k(k-1)$ choices for x. Thus, such a point x can be found only if $k(k-1) < L - k$, which gives us an upper bound of $k = O(L^{1/2})$, as desired.

We note in passing that the probability of a randomly selected x in the above setting to yield a collision-free subset $B \cup \{x\}$ is bounded above by $(L - k - k(k-1))/L$, for *any* set B.

Now we examine the largest value of m for which a random subset drawn of size m drawn from $\mathbb{Z}_L$ collision free with positive probability. For concreteness, we consider m samples without replacement from $\mathbb{Z}_L$.

For $1 \le k \le m$, we denote by $\mathfrak{S}_k$ the set of first k random samples without replacement. Then we may write

$$
\begin{aligned}
&\mathbb{P}[\mathfrak{S}_m \text{ is collision-free}] \\
&\quad = \mathbb{P}[\mathfrak{S}_m \text{ is collision-free} \mid \mathfrak{S}_{m-1} \text{ is collision-free}] \cdot \mathbb{P}[\mathfrak{S}_{m-1} \text{ is collision-free}] \\
&\quad = \prod_{k=1}^{m-1} \mathbb{P}[\mathfrak{S}_{k+1} \text{ is collision-free} \mid \mathfrak{S}_k \text{ is collision-free}] \\
&\quad = \prod_{k=1}^{m-1} \mathbb{P}_{x \sim \mathrm{Unif}(\mathbb{Z}_L \setminus \mathfrak{S}_k)}\big[\mathfrak{S}_k \cup \{x\} \text{ is collision-free} \mid \mathfrak{S}_k \text{ is collision-free}\big] \\
&\quad \le \prod_{k=1}^{m-1} \frac{L - k - k(k-1)}{L} \quad \text{[using the analysis for the set } B \text{ above]} \\
&\quad = \prod_{k=1}^{m-1} \left(1 - \frac{k^2}{L}\right) \ \le \prod_{k=1}^{m-1} \exp(-\frac{k^2}{L}) \ \le \exp(-cm^3/L).
\end{aligned}
$$

Thus, if $m^3/L \to \infty$, $\mathbb{P}[\mathfrak{S}_k \text{ is collision-free}] \to 0$. Therefore, for a random subset of size m to be collision-free with positive probability, we must have $m = O(L^{1/3})$, and to have the same property with high probability, we must have $m = o(L^{1/3})$.

References

1. Emmanuel Abbe, Tamir Bendory, William Leeb, João M Pereira, Nir Sharon, and Amit Singer. Multireference alignment is easier with an aperiodic translation distribution. *IEEE Transactions on Information Theory*, 65(6):3565–3584, 2018.
2. Emmanuel Abbe, João M Pereira, and Amit Singer. Estimation in the group action channel. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 561–565. IEEE, 2018.
3. Afonso S Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 459–470. ACM, 2014.

4. Afonso S Bandeira, Philippe Rigollet, and Jonathan Weed. Optimal rates of estimation for multi-reference alignment. arXiv preprint arXiv:1702.08546, 2017.
5. LD Barron. Symmetry and molecular chirality. *Chemical Society Reviews*, 15(2):189–223, 1986.
6. Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline LS Milne, and Sriram Subramaniam. 2.2 å resolution cryo-em structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.
7. Robert Beinert and Gerlind Plonka. Sparse phase retrieval of one-dimensional signals by prony's method. *Frontiers in Applied Mathematics and Statistics*, 3:5, 2017.
8. Ahmad Bekir. *On the nonexistence of additional counterexamples to Sophie Piccard's theorem*. University of Southern California, 2004.
9. Ahmad Bekir and Solomon W Golomb. There are no further counterexamples to s. piccard's theorem. *IEEE transactions on information theory*, 53(8):2864–2867, 2007.
10. Tamir Bendory, Robert Beinert, and Yonina C Eldar. Fourier phase retrieval: Uniqueness and algorithms. In *Compressed Sensing and its Applications*, pages 55–91. Springer, 2017.
11. Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on Signal Processing*, 66(4):1037–1050, 2017.
12. Tamir Bendory and Dan Edidin. Toward a mathematical theory of the crystallographic phase retrieval problem. *SIAM Journal on Mathematics of Data Science*, 2(3):809–839, 2020.
13. Tamir Bendory, Dan Edidin, William Leeb, and Nir Sharon. Dihedral multi-reference alignment. *IEEE Transactions on Information Theory*, 2022.
14. Gary S Bloom. A counterexample to a theorem of s. piccard. *Journal of Combinatorial Theory, Series A*, 22(3):378–379, 1977.
15. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
16. Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer. Heterogeneous multireference alignment: A single pass approach. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.
17. Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.
18. Victor-Emmanuel Brunel. Learning rates for gaussian mixtures under group action. In *Conference on Learning Theory*, pages 471–491. PMLR, 2019.
19. A David Buckingham. Chirality in nmr spectroscopy. *Chemical physics letters*, 398(1–3):1–5, 2004.
20. Philip R Bunker and Per Jensen. *Molecular symmetry and spectroscopy*, volume 46853. NRC Research Press, 2006.
21. Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.
22. Robert Diamond. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Science*, 1(10):1279–1287, 1992.
23. Ian L. Dryden and Kanti V. Mardia. *Statistical shape analysis*. Wiley series in probability and statistics. Wiley, Chichester [u.a.], 1998.
24. Veit Elser, Ti-Yen Lan, and Tamir Bendory. Benchmark problems for phase retrieval. *SIAM Journal on Imaging Sciences*, 11(4):2429–2455, 2018.
25. Zhou Fan, Roy R Lederman, Yi Sun, Tianhao Wang, and Sheng Xu. Maximum likelihood for high-noise group orbit estimation and single-particle cryo-em. arXiv preprint arXiv:2107.01305, 2021.
26. Zhou Fan, Yi Sun, Tianhao Wang, and Yihong Wu. Likelihood landscape and maximum likelihood estimation for the discrete orbit recovery model. arXiv preprint arXiv:2004.00041, 2020.
27. Charles L Fefferman. The uncertainty principle. *Bulletin of the American Mathematical Society*, 9(2):129–206, 1983.
28. James R Fienup. Phase retrieval algorithms: a personal tour. *Applied optics*, 52(1):45–56, 2013.
29. Gerald B Folland and Alladi Sitaram. The uncertainty principle: a mathematical survey. *Journal of Fourier analysis and applications*, 3(3):207–238, 1997.
30. Hassan Foroosh, Josiane B Zerubia, and Marc Berthod. Extension of phase correlation to subpixel registration. *IEEE transactions on image processing*, 11(3):188–200, 2002.
31. Roberto Gil-Pita, Manuel Rosa-Zurera, P Jarabo-Amores, and Francisco López-Ferreras. Using multilayer perceptrons to align high range resolution radar signals. In *International Conference on Artificial Neural Networks*, pages 911–916. Springer, 2005.

32. Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 1473–1477. IEEE, 2012.
33. Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Transactions on Signal Processing*, 65(9):2402–2410, 2017.
34. Anya Katsevich and Afonso Bandeira. Likelihood maximization and moment matching in low snr gaussian mixture models. arXiv preprint arXiv:2006.15202, 2020.
35. J Kormylo and J Mendel. Maximum likelihood detection and estimation of bernoulli-gaussian processes. *IEEE transactions on information theory*, 28(3):482–488, 1982.
36. Rick P Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.
37. Henrik Ohlsson and Yonina C Eldar. On conditions for uniqueness in sparse phase retrieval. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1841–1845. IEEE, 2014.
38. Wooram Park and Gregory S Chirikjian. An assembly automation approach to alignment of noncircular projections in electron microscopy. *IEEE Transactions on Automation Science and Engineering*, 11(3):668–679, 2014.
39. Wooram Park, Charles R Midgett, Dean R Madden, and Gregory S Chirikjian. A stochastic kinematic model of class averaging in single-particle electron microscopy. *The International journal of robotics research*, 30(6):730–754, 2011.
40. Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019.
41. Sophie Piccard. *Sur les ensembles de distances des ensembles de points d'un espace Euclidien*. Paris, 1939.
42. Juri Ranieri, Amina Chebira, Yue M Lu, and Martin Vetterli. Phase retrieval for sparse signals: Uniqueness conditions. arXiv preprint arXiv:1308.3058, 2013.
43. K Veera Reddy. *Symmetry and Spectroscopy of Molecules*. New Age International, 1998.
44. Ya'Acov Ritov. Estimating a signal with noisy nuisance parameters. *Biometrika*, 76(1):31–37, 1989.
45. Dirk Robinson, Sina Farsiu, and Peyman Milanfar. Optimal registration of aliased images using variable projection with applications to super-resolution. *The Computer Journal*, 52(1):31–42, 2007.
46. Elad Romanov, Tamir Bendory, and Or Ordentlich. Multi-reference alignment in high dimensions: sample complexity and phase transition. *SIAM Journal on Mathematics of Data Science*, 3(2):494–523, 2021.
47. David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *The International Journal of Robotics Research*, 38(2-3):95–125, 2019.
48. Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.
49. B.m. Sadler. Shift and rotation invariant object reconstruction using the bispectrum. *Workshop on Higher-Order Spectral Analysis*, 1989.
50. Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.
51. Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.
52. Fred J Sigworth. A maximum-likelihood approach to single-particle image refinement. *Journal of structural biology*, 122(3):328–339, 1998.
53. Devika Sirohi, Zhenguo Chen, Lei Sun, Thomas Klose, Theodore C Pierson, Michael G Rossmann, and Richard J Kuhn. The 3.8 å resolution cryo-em structure of zika virus. *Science*, 352(6284):467–470, 2016.
54. Charles Soussen, Jérôme Idier, David Brie, and Junbo Duan. From bernoulli–gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, 2011.
55. Terence Tao. *Structure and randomness: pages from year one of a mathematical blog*. American Mathematical Society, Providence, RI, 2008.

56. Douglas L Theobald and Phillip A Steindel. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics*, 28(15):1972–1979, 2012.
57. Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.
58. Eugene Wigner. *Group theory: and its application to the quantum mechanics of atomic spectra*, volume 5. Elsevier, 2012.
59. J Portegies Zwart, René van der Heiden, Sjoerd Gelsema, and Frans Groen. Fast translation invariant classification of hrr range profiles in a zero phase representation. *IEE Proceedings-Radar, Sonar and Navigation*, 150(6):411–418, 2003.