*Information and Inference: A Journal of the IMA* (2022) **00**, 1–39
https://doi.org/10.1093/imaiai/iaac016

# Compressive independent component analysis: theory and algorithms

Michael P. Sheehan[†] and Mike E. Davies
*Institute of Digital Communications, University of Edinburgh, Edinburgh EH8 9YL, UK*
[†]Corresponding author. Email: michael.sheehan@ed.ac.uk

[Received on 15 October 2021; revised on 11 April 2022; accepted on 21 April 2022]

Compressive learning forms the exciting intersection between compressed sensing and statistical learning where one exploits sparsity of the learning model to reduce the memory and/or computational complexity of the algorithms used to solve the learning task. In this paper, we look at the independent component analysis (ICA) model through the compressive learning lens. In particular, we show that solutions to the cumulant-based ICA model have a particular structure that induces a low-dimensional model set that resides in the cumulant tensor space. By showing that a restricted isometry property holds for random cumulants e.g. Gaussian ensembles, we prove the existence of a compressive ICA scheme. Thereafter, we propose two algorithms of the form of an iterative projection gradient and an alternating steepest descent algorithm for compressive ICA, where the order of compression asserted from the restricted isometry property is realized through empirical results. We provide analysis of the CICA algorithms including the effects of finite samples. The effects of compression are characterized by a trade-off between the sketch size and the statistical efficiency of the ICA estimates. By considering synthetic and real datasets, we show the substantial memory gains achieved over well-known ICA algorithms by using one of the proposed CICA algorithms.

*Keywords*: independent component analysis; compressive learning; sketching; compressive sensing; summary statistics; cumulants.

## 1. Introduction

In recent years, the size of datasets has grown exponentially as a result of advances in technology, signal acquisition and the sophistication of modern day mobile phones and devices. This has enabled researchers, statisticians and machine learning practitioners to build increasingly accurate models as a consequence of larger sample sizes and feature dimensions. Nevertheless, this poses a fundamental challenge to large-scale learning as (i) traditional algorithms that act on the non-compressed full data have computational complexity that scales with the order of the dataset dimensions[1] (ii) the whole dataset has to be stored centrally or transferred to a local computer as optimization methods need to return to the data (or a random subset of the data) at subsequent iterations and (iii) one is vulnerable to malicious attacks of potentially sensitive and personal information as the data need to be stored or transferred locally. Compressive learning (CL) [33, 34] partially addresses these fundamental challenges by severely compressing the whole dataset into a random representation of fixed size, named a so-called sketch, in a single (or limited) pass of the data prior to learning. Once the sketch is formed, the parameters of the model are inferred solely from the sketch, hence a CL algorithm, for a given task or model, needs never to return to the original dataset, and the latter can be deleted from memory as a result.

[1] for instance the number of samples and features of the data.

Fundamental to the CL framework [33, 44], the size of the sketch does not scale with the dimensions of the dataset, or indeed the data's underlying dimensionality, but instead is driven by the complexity or dimensionality of the task or model of interest. In theory, one can work with datasets of arbitrary length (number of samples/data points), as the dimension of the sketch is fixed constant throughout, making CL especially amenable to large-scale learning. Inferring the parameters of a model solely from the sketch is an underdetermined inverse problem. As a result, we need regularity assumptions to make the problem well-posed. These assumptions come in the form of a low-dimensional model set that the solution to the inference problem lies on or close to. The reader may notice this is reminiscent of compressive sensing where one assumes the signal of interest is $k$ sparse in some domain, and therefore the solution lies on or close to the union of $k$-dimensional subspaces representing a low-dimensional model set. The sparse regularity assumption allows one to take a limited number of measurements to recover the signal of interest and reduce the complexity and cost of acquisition. In later sections, we take inspiration from compressive sensing to develop and analyse our CL algorithms.

In this paper, we develop a CL framework, including theory and practical algorithms, for independent component analysis (ICA). ICA is an unsupervised learning task that attempts to find the linear transformation that separates some given data into components of maximal independence. It is used extensively in the machine learning and signal processing communities, for example, as a dimensionality reduction tool [42], to uncover underlying factors that effect the price movements of a collection of stocks [48] and to detect independent sources in the brain through EEG signals [62]. As will be discussed in Section 2, the ICA problem can be solved directly from the data or through some higher order statistics of the data, such as the kurtosis. Denoting the number of independent sources by $n$ and the signal or data length by $N$, the memory complexity typically scales either with $\mathscr{O}(nN + n^2)$ or $\mathscr{O}(n^4)$ depending on the method of choice. As one can see, this becomes infeasible for large-scale datasets. In this paper, we show theoretically and empirically that it is possible to design a CL ICA algorithm where the sketch dimension, and therefore the memory complexity, scales at $\mathscr{O}(n^2)$ which can be orders of magnitudes smaller than current approaches.

### 1.1 *Contributions and Outline*

This paper is an extension of the conference paper in [55] that further includes theoretical results as well as algorithmic improvements. Next, we highlight the main contributions of the paper:

- Focusing on the cumulant-based ICA approach, we establish a low-dimensional model set that resides in the larger cumulant space. We show that an optimal sketch of size $m \gtrsim 2n(n + 1)$, computed using sub-Gaussian measurements, satisfies the well-known restricted isometry property[2] (RIP) on the model set with high probability. Furthermore, the RIP induces information preservation guarantees on the recovered cumulant tensor that shows the error between an arbitrary cumulant tensor, and the recovered cumulant tensor is bounded linearly by modelling error and sampling noise. This establishes the existence of a robust decoder that, coupled with the sketching operator, forms a tractable compressive ICA scheme.
- In general, we do not have access to the true expected cumulant tensor but instead an approximation of the cumulant tensor formed by the finite samples. We establish an upper bound on the finite sampling error between the sketch of the expected cumulant tensor and the sketch of the

[2] The RIP states that the distance between points in the model set is approximately preserved under the action of the sketching operator (see Section 3)

approximated cumulant tensor. It is shown that the sampling error reduces as a function of the number of samples.

- Two inherently different compressive ICA algorithms are proposed. The first algorithm is inspired by a greedy approach, where we design a projection operator that projects the updated tensor onto the model set at each iteration. The second algorithm is an alternative steepest descent scheme that employs Riemannian optimization to optimize directly on the model set.
- As part of the empirical results, we show that in practice a sharp phase transition, between successful and unsuccessful parameter estimation, occurs as the sketch size $m$ grows. The region at which the transition transpires provides a pragmatic lower bound on the size of the sketch which one can use in practice. Furthermore, it is shown that this pragmatic lower bound coincides with the size of the sketch required to satisfy the RIP result. The loss of information incurred by taking a sketch of the ICA cumulants is demonstrated by comparing the statistical efficiency between the ICA estimates inferred by our compressive ICA algorithms and by existing algorithms that use the full data available.

## 1.2 *Related Works*

1.2.1 *Existing Compressive Learning Models.* The framework of CL has been successfully applied to a host of learning tasks and models with the desired outcome of reducing the complexities associated with signal acquisition, computation and memory storage. In [44], Keriven *et al.* proposed a CL framework for mixture models, in particular the mixture of Gaussians distribution and $k$-means models. In both cases, a sketch is constructed by randomly sampling the (empirical) characteristic function of the mixture model which can be equivalently seen as taking and averaging random Fourier features of the data [50]. The compact representational sketch of each mixture model scales both theoretically and empirically as $\mathscr{O}(k^2d)$, where $k$ is the number of mixtures in the model and $d$ is the feature space dimensions of the data. In [44], a compressive mixture model algorithm was proposed that minimized the $\ell_2$ distance between the characteristic function and its empirical counterpart was calculated over the data for $m$ randomly sampled frequencies. Fundamentally, the compressive mixture model algorithm have both computational and space complexities that scale independently of the number of data points $N$. In [33], a compressive principal component analysis (PCA) framework was proposed. As will be discussed in Section 2.2, the compressive PCA methodology is aligned closely to our compressive ICA framework. Distinct from compressive mixture models, the compressive PCA method is distribution free and it is assumed that the data live on, or can be approximately modelled, by a $k$-dimensional subspace related to the top $k$ eigenvectors of the covariance matrix. As a result, a sketch of size $\mathscr{O}(kd)$ can be computed by taking a random projection of the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ of the data, hence reducing the memory complexities of storing either the data of size $Nd$ or the covariance matrix of size $d^2$.

1.2.2 *Generalized Method of Moments.* Compressive learning is similar to the technique of Generalized Method of Moments (GeMM) [36, 37, 65], where the parameters of interest $\theta$ are estimated by matching a collection of generalized moments of the distribution with the empirical counterparts calculated through the data. In most cases it is used instead of maximum likelihood estimation when calculating the likelihood is not tractable. CL differs from much of the GeMM literature as the goal is fundamentally different: in compressive learning one attempts to construct a compact representation of data with the aim of reducing complexity constraints (computation, memory, acquisition), while in

GeMM the goal is to primarily estimate $\theta$ when the model is either partially specified or the likelihood does not have a closed-form solution. Moreover, the selected generalized moments may be a function of the parameter being estimated, hence not providing a one-off sketch. In [4], the authors employ a generalized method of moments technique to estimate the parameters for a range of latent variable models, including ICA. For the case of ICA, the decomposition of a fourth-order cumulant tensor is used; however, as it will be shown in Section 2.3.2, the fourth-order cumulant tensor has a size that scales with $n^4$ compared with our proposed sketch which has a size that scales with $n^2$.

1.2.3 *Streaming Methods.* Closely related to CL is the collection of streaming methods [25, 60], where data items are seen and queried only once by the user and then discarded. This is of particular interest when the summary statistic of choice is updated and maintained in real time, for example, in the online learning setting [35], to reduce space complexities. Notably, the count-min-sketch [26] was developed to query data in an online fashion with the application of maintaining histograms of quantiles. However, these methods in general focus on discrete collections of objects and database queries, while in CL the framework and method is applied to machine learning tasks where typically the signal is question is continuous. Tropp *et al.* [59] proposed a streaming framework for large-scale PCA. In particular, in [60], the authors design random sketches for on-the-fly compression of data matrices associated with large-scale scientific simulations. Here, the data matrix $\mathbf{A}$ of interest can be decomposed into a sequence

$$\mathbf{A} = \mathbf{H}_1 + \mathbf{H}_2 + \mathbf{H}_3 + \dots, \tag{1.1}$$

where it is assumed each $\mathbf{H}_i$ has some structural redundancies for example sparsity or low rank. These methods have a subtle yet fundamental difference from CL, as in CL the structural assumptions which are exploited to form the CL sketch arise from the model or distribution itself, while in these streaming methods the structural assumptions come directly from the data. Moreover, several passes of the data may be required to reduce the low-rank approximation error [59]. Compressive learning, which will be formally introduced in Section 2.1, falls under the category of sketching. However, there is a subtle yet important difference compared with the (linear) sketching techniques discussed in this section. In compressive learning, the sketch is linear over the space of distributions $\mathbb{P}$ but typically non-linear over the data. In contrast, the sketches defined in this section are linear over the dataset.

1.2.4 *Other Compression Techniques.* Coresets are a popular method used to compress a database into a summary statistic used for inferring the parameters of a given model and have been used widely in subspace clustering based tasks [31, 38]. In a similar vein to CL, the compact data representation has size that is a fraction of the original dataset dimensions. However, the coresets are constructed in a hierarchical manner, possibly resulting in multiple passes of the data and are therefore not naturally amenable to online or distributed learning. Projections that include both random projections and feature selections [11] are used widely to reduce the dimensionality of the data. In [11], datasets were randomly projected into a compressed domain using both random Gaussian and Bernoulli matrices. In a similar vein to compressive sensing [30], the data were assumed to be $k$-sparse therefore the dependency of the feature space dimension $d$ was removed within the space and acquisition complexities. In contrast to random projections, more structural based projections are proposed. In [57], different feature selection techniques for classification are reviewed including structured graph methods and the use of embedded models. The well-known PCA method is a popular preprocessing technique that projects the dataset onto a $k$-dimensional subspace of maximal variance [43]. In both random and structured projections, the methods discussed only tackle the dependency of the feature space dimension $d$ and do not address the challenges posed by a large data size $N$. Subsampling methods are also a popular method for dimensionality reduction whereby a subset of the original dataset is used for learning. As discussed previously, the method of coresets [31, 38] is a subsampling technique that attempts to sub-select dominant items that well approximate the structure of the dataset. Other subsampling techniques include random and adaptive subsampling [25]. The disadvantage of subsampling techniques is that there is a risk of discarding important information relating to non-sampled data items. Moreover, these techniques only tackle the constraint on the number of data items $N$ and do not combat the complexity issues posed by the feature space dimensional $d$.

Specifically to ICA compression, Sela *et al.* [53] used kernel approximation techniques to reduce the dimensions of the Kernel ICA method proposed by Bach [5]. Random Fourier features are used to approximate the kernel, reducing the memory complexity from $\mathscr{O}(d^2N^2)$ to $\mathscr{O}(MN)$, where $M$ is the number of random Fourier weights used. Despite the reduction in memory complexity, the algorithm still has storage demands which scale linearly with $N$. In comparison, we remove the dependency of the data length $N$ completely, within our framework, when estimating the ICA mixing matrix.

## 2. Background

### 2.1 *Compressive Learning*

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ be independent and identically distributed samples from an unknown probability distribution $\pi$ on $(X, B\ )$, where $X \subset \mathbb{R}^d$ is some Euclidean space and $B$ is a Borel $\sigma$-field. Classically, $\pi$ is parametrized by some parameters denoted by $\theta \in \Theta (\subset \mathbb{R}^k)$. A statistical learning problem can be formalized as follows: find a hypothesis $h^*$ from a hypothesis class $\mathscr{H}$ that best matches the probability distribution $\pi$ over the training collection $\{\mathbf{x}_i\}_{i=1}^N$, given some data fidelity term. Given a loss function $l : X \times \mathscr{H} \longmapsto \mathbb{R}$, this is equivalent to minimizing the risk defined as

$$h^* = \arg\min_{h\in\mathscr{H}} \mathscr{R}\,(\pi, h) = \arg\min_{h\in\mathscr{H}} \mathbb{E}_{\mathbf{x}\sim\pi}\, l(\mathbf{x}, h). \tag{2.1}$$

Formally, the model set associated with the hypothesis class can be defined as

$$\mathfrak{S}_{\mathscr{H}} := \{\pi \in \mathscr{P}(X) : \exists h \in \mathscr{H},\ \mathscr{R}\,(\pi, h) = 0\}. \tag{2.2}$$

In other words, the set containing all distributions for which zero risk is achievable. As a result, the model set has a dimension which is intrinsic to the hypothesis class of the model. In practice, one cannot minimize the true risk as we generally do not have access to the true distribution $\pi$, so instead, one can minimize the empirical risk with respect to the finite samples of the true distribution and as a result this may mean all the data are required to be stored in memory.

In CL [33, 34, 44], we find a compact representation, or a so-called sketch, that encodes some statistical properties of the data. Its size is ideally chosen relative to the intrinsic complexity of the problem, making it possible to work with arbitrarily large datasets while storing in memory an object of fixed size. Given a feature function $\Phi : X \longmapsto \mathbb{C}^m$, such that $\Phi$ is integrable with respect to any $\pi \in \mathscr{P}(X\ )$, define the linear operator $\mathscr{A} : \mathscr{P}(X\ ) \longmapsto \mathbb{R}^m$ by

$$\mathscr{A}(\pi\ ) := \mathbb{E}_{\mathbf{x}\sim\pi}\, \Phi(\mathbf{x}). \tag{2.3}$$

The sketch defined in (2.3) can be seen as taking the expectation of some particular features of the distribution $\pi$, which is similar to the field of kernel mean embedding [47] where one uses feature maps to embed probability distributions. Therefore, we would like to construct $\mathscr{A}$ so that $\mathscr{A}(\pi)$ captures sufficiently relevant information of the data to allow us to infer the parameters of the model directly from the sketch. As a trivial example, if we seek to infer only the mean of a normal distribution $\pi = \mathscr{N}(\mu, \sigma)$, the construction $\mathscr{A}(\pi)$ where $\Phi(\mathbf{x}) = \mathbf{x}$ would constitute a trivial yet sufficient sketch. In reality, CL is applicable to much more complex models where the feature function is non-trivial and the model may not necessarily possess a finite-dimensional sufficient statistic. The goal of CL is to therefore construct a sketch of size $m \ll Nd$ that captures enough information to recover an estimated risk which is *close* to the true risk with high probability [33]. In practice, as in the kernel mean embedding literature [47], the empirical distribution is used to form an empirical sketch defined as

$$\hat{\mathbf{y}} = \mathscr{A}(\pi_N), \quad \text{where} \quad \pi_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}_i} \tag{2.4}$$

denoting by $\delta_x$ the Dirac distribution on $x$, and therefore the empirical sketch can be formed directly from the data. Due to the law of large numbers, $\lim_{N\to\infty} \mathscr{A}(\pi_N) = \mathscr{A}(\pi)$. Once the sketch has been computed, one can discard the dataset $\{\mathbf{x}_i\}_{i=1}^N$ from memory. As a result, CL reduces down to solving an inverse problem of the form

$$\hat{\theta} = \arg\min_{\theta\in\Theta} C(\theta \mid \hat{\mathbf{y}}), \tag{2.5}$$

where $C(\cdot \mid \hat{\mathbf{y}})$ is a cost function designed for the specific learning task at hand. In a compressive sensing light, we can exploit structural assumptions of the model set and the associated parameter space $\Theta$, e.g sparsity, low rankness, low-dimensional manifold properties, to make (2.5) well-posed and finding a solution tractable. As such, one can design a decoder $\Delta$ that exploits the structural assumptions of the model set $\mathfrak{S}_{\mathscr{H}}$ to recover the parameters of the model from the sketch while minimizing the risk. The sketching operator $\mathscr{A}$ and the decoder $\Delta$ form the pair $(\Delta, \mathscr{A})$ that define the CL algorithm for a specific learning problem. It should be noted that minimizing (2.5) plays the role of a proxy for minimizing the empirical risk.

### 2.2 *Compressive PCA*

In Section 2.1, the framework of CL was discussed in a general manner without specific consideration of the distributional form of the model. As will be discussed in Section 2.3, the PCA and ICA models are similar in nature in that the model is often left distribution free. In other words, the distribution of the sampled data is left unspecified. In Table 1, it is shown that the compressive PCA model set [33] is defined as

$$\mathfrak{S}_{\mathscr{H}} = \left\{\pi \mid \operatorname{rank}(\Sigma_\pi) \le k\right\}. \tag{2.6}$$

Due to the distribution free assumption of the PCA model, we seek structural assumptions that are manifested within some *intermediary* statistic space $\mathbb{S}$ to make computing a sketch possible [55]. In the case of compressive PCA, the space of $d \times d$ covariance matrices is leveraged as an intermediary statistic space $\mathbb{S}$ where the rank of the covariance matrices is exploited. Figure 1 depicts a geometric viewpoint of both compressive parametric learning (e.g. $k$-means, GMM) and distribution free compressive learning

TABLE 1 *Summary of existing methods in the CL framework. For the compressive k-means and compressive GMM scheme, the random Fourier feature function is used where* $\mathbf{w}$ *are randomly sampled weights. For compressive PCA and our proposed ICA scheme,* $\mathbf{A}_j$ *denotes the jth row of a random Gaussian matrix* $\mathbf{A}$ *(see Section 4). For the compressive PCA case, the cost function reduces to a low-rank matrix recovery problem by minimizing the nuclear norm* $\|\cdot\|_\star$*. For more details on the existing compressive learning schemes see [33].*

| **Model** | **Model Set** $\mathfrak{S}_{\mathscr{H}}$ | **Feat. Func.** $\Phi(\mathbf{x})$ | **Cost** $C(\theta \mid \mathbf{y})$ | **Sketch size** $m$ |
|---|---|---|---|---|
| $k$- Means | $\{\pi \mid \text{mix. of } k \text{ Diracs}\}$ | $\left(e^{\mathrm{i}\omega_j^T\mathbf{x}}/w(\omega_j)\right)_{j=1}^m$ | $\min_{\pi\in\mathfrak{S}_{\mathscr{H}}} \|\mathbf{y}-\mathscr{A}(\pi)\|_2$ | $\mathscr{O}(k^2d)$ |
| GMM | $\{\pi \,\text{mix. of } k \text{ Gauss.}\}$ | $\left(e^{\mathrm{i}\omega_j^T\mathbf{x}}\right)_{j=1}^m$ | $\min_{\pi\in\mathfrak{S}_{\mathscr{H}}} \|\mathbf{y}-\mathscr{A}(\pi)\|_2$ | $\mathscr{O}(k^2d)$ |
| PCA | $\{\pi \text{ Rank } k \text{ cov. mat.} \Sigma_\pi\}$ | $(\langle\mathbf{A}_j, \mathbf{x}\mathbf{x}^T\rangle)_{j=1}^m$ | $\min\|\Sigma_\pi\|_* \text{ s.t } \mathscr{A}(\Sigma_\pi)=\mathbf{y}$ | $\mathscr{O}(kd)$ |
| ICA | $\{\pi \text{ sparse tensor decomp. } Z_\pi\}$ | $(\langle\mathbf{A}_j, \mathbf{z}^{\otimes^4}\rangle)_{j=1}^m$ | $\min_{\mathscr{Z}\in\mathfrak{S}_{\mathscr{H}}} \|\mathbf{y}-\mathscr{A}(\mathscr{Z})\|_2$ | $\mathscr{O}(n^2)$ |

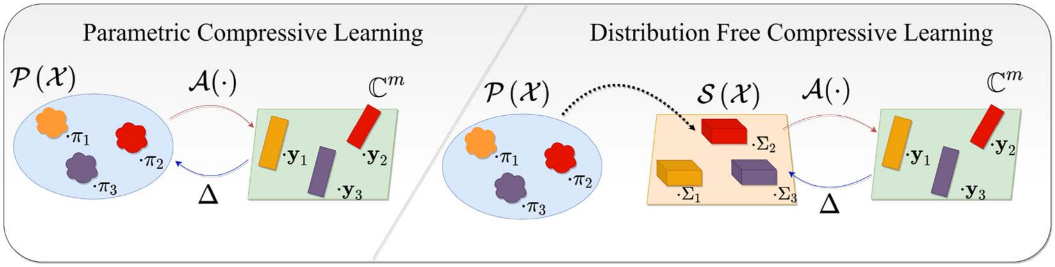

FIG. 1. A schematic diagram of parametric compressive learning (left) and distribution free compressive learning (right).

(e.g. PCA, ICA). In general, distribution free CL poses distinct challenges and advantages from the typical parametric CL framework [54]. Challenges arise when choosing an intermediary statistic space $\mathbb{S}$, for instance (1) what set of intermediate statistics can we use? (2) How do the structural assumptions of the model set manifest within the intermediate statistic? Equivalently, there are many advantages. Specifically, by leveraging some set of intermediate statistics we have implicitly mapped the problem from an infinite dimensional probability space to a typically finite dimensional statistic space. As a result, we can utilize a host of existing techniques within the compressive sensing literature to design encoder and decoder pairs $(\mathscr{A}, \Delta)$. Moreover, it also allows us to use a more flexible semi-parametric model that is only partially specified. As will be discussed in Section 4, the compressive ICA framework follows a similar convention where the space of fourth-order cumulant tensors $\mathbb{S} = \mathfrak{C}$ is used as an intermediary statistic space to exploit structural assumptions of the model set $\mathfrak{S}_{\mathscr{H}}$to form a sketch.

### 2.3 *Independent Component Analysis*

ICA is used frequently in the machine learning and signal processing communities to identify latent variables that are mutually independent to one another. It can be seen as an extension to PCA due to the assumption of independence between the latent variables, which is stronger than the uncorrelated constraint in PCA. To formulate the ICA problem, consider a data point $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$, then the problem of ICA concerns finding a mixing matrix $\mathbf{M} \in \mathbb{R}^{d \times n}$ (here we assume $d \geq n$) such that

$$\mathbf{x} = \mathbf{M}\mathbf{s}, \tag{2.7}$$

where $\mathbf{s} = (s_1, s_2, \dots, s_n)^T$ and the components $s_i$ are statistically independent:

$$p(s_1, s_2, \dots, s_n) = \prod_{i=1}^{n} p_i(s_i). \tag{2.8}$$

Here, $p$ denotes the joint probability distribution of the independent components and $p_i$ denotes the probability distribution of the $i$th component. The ICA model in (2.7) has the following ambiguities:

- As both $\mathbf{s}$ and $\mathbf{M}$ are unknown, any scalar multiplier in one of the independent components $s_j$ can always be cancelled by dividing the corresponding column in $\mathbf{M}$ by the same scalar.
- The order of the independent components and the corresponding columns in $\mathbf{M}$ can freely change.

As a result, the mixing matrix $\mathbf{M}$ and the independent components $\mathbf{s}$ are only identifiable up to scaling and permutation ambiguities.

The data point $\mathbf{x}$ is only one realization of a data matrix or signal $\mathbf{X} \in \mathbb{R}^{N \times d}$ of length $N$, therefore we attempt to infer $\mathbf{M}$ with the collective set of linear equations $\mathbf{X} = \mathbf{M}\mathbf{S}$, where $\mathbf{s}$ is a realization of $\mathbf{S} \in \mathbb{R}^{N \times n}$. There are many techniques and methods in the literature to solve the ICA problem. The simplest method is to assume the distributional form of each of the independent components $p_i(s_i)$ and then solve the ICA problem through a maximum likelihood approach [6]. In practice, the distributions are not known a priori therefore in most methods the distributions are left unspecified. As a result, practitioners and researchers often resort to minimizing a given contrast function (see Section 2.3.3) to solve the ICA problem.

2.3.1 *Prewhitening.* A useful preprocessing strategy in ICA is to first whiten the observed variables $\mathbf{x}$ via a linear transformation. This involves the process of finding the matrix $\mathbf{V}$ such that

$$\mathbf{z} = \mathbf{V}\mathbf{x}, \tag{2.9}$$

where $\mathbf{z}$ has identity covariance matrix. One popular method of whitening is to use the eigendecomposition of the covariance matrix and retain the $n$ principal eigenvectors. Whitening has two main advantages: (1) it handles the scenario when there are more mixing components than independent components $d > n$ as one can discard the $d-n$ smallest eigenvalues, (2) the matrix $\mathbf{Q} := \mathbf{V}\mathbf{M} \in \mathbb{R}^{n \times n}$ is necessarily orthogonal and contains $n(n-1)/2$ degrees of freedom. The whitening matrix $\mathbf{V}$ is not unique and any orthogonal rotation of $\mathbf{V}$ will also define a whitening matrix [42]. For the sake of presentation, we will subsequently consider the whitened version of the data for the remainder of this section and the corresponding whitened ICA equation

$$\mathbf{z} = \mathbf{Q}\mathbf{s}. \tag{2.10}$$

In Section 4.3, we propose two equivalent sketching frameworks that can either incorporate prewhitened and unwhitened data.

2.3.2 *Cumulant-based ICA.* Tensorial or cumulant-based methods are a group of techniques used to solve the ICA problem and are of particular interest in this paper. Statistical properties of the data instance **z** can be described by its cumulants $\mathscr{Z}^K_{i_1 i_2 \dots i_K}$. In the multivariate setting, cumulants give rise to tensors, denoted $\mathscr{Z}^K$ for a cumulant tensor of order $K$. Assuming the data have zero mean, the first four cumulants are defined [28] as

$$\begin{aligned}
\mathscr{Z}^1_i =& 0\\
\mathscr{Z}^2_{ij} =& \mathbb{E}[z_i z_j]\\
\mathscr{Z}^3_{ijk} =& \mathbb{E}[z_i z_j z_k]\\
\mathscr{Z}^4_{ijkl} =& \mathbb{E}[z_i z_j z_k z_l] - \mathbb{E}[z_i z_j]\mathbb{E}[z_k z_l] - \mathbb{E}[z_i z_k]\mathbb{E}[z_j z_l] - \mathbb{E}[z_i z_l]\mathbb{E}[z_j z_k],
\end{aligned} \tag{2.11}$$

where $\mathbb{E}$ is the expectation operator. Given the model in (2.10) that equates **z** to **s**, the following multilinear property holds for their associated cumulant tensors:

$$\mathscr{Z}^K = \mathscr{S}^K \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \cdots \times_K \mathbf{Q}, \tag{2.12}$$

where $\times_j$ represents the $j$-mode tensor-matrix product and $\mathscr{S}^K$ represents the $K^{th}$ order cumulant tensor of the independent source signals [28]. In this paper, we will only consider fourth-order cumulant tensors (e.g. $K = 4$) and for the sake of simplified notation we shall drop the superscript in (2.12) for the rest of the discussion. We denote by $\mathfrak{C} \subset \mathbb{R}^{d\times d\times d\times d}$ the space of fourth-order cumulant tensors which account for the symmetry in (2.11), where each cumulant tensor $\mathscr{Z} \in \mathfrak{C}$ has a maximum of $\left({}_{n+34}\right)$ unique entries (degrees of freedom) [23]. The diagonal entries $\mathscr{Z}_{ijkl}(ijkl = iiii)$ are the auto-cumulants of **z**, while the off-diagonal entries $\mathscr{Z}_{ijkl}(ijkl \neq iiii)$ are the cross-cumulants. If the variables $(z_1, z_2, \dots, z_n)$ are statistically independent then, as seen by (2.11), the cross-cumulants vanish to 0 resulting in a strictly diagonal cumulant tensor. In other words, independence implies diagonality. It is shown in [24] that under mild conditions[3] the converse is also true, i.e. diagonality implies independence. Once the data are whitened, the cumulant-based ICA problem reduces to finding a linear transformation $\mathbf{Q}^T$ such that the resulting cumulant tensor

$$\mathscr{S} := \mathscr{Z} \times_1 \mathbf{Q}^T \times_2 \mathbf{Q}^T \times_3 \mathbf{Q}^T \times_4 \mathbf{Q}^T \tag{2.13}$$

is strictly diagonal. We can define the following ICA model set:

$$\mathfrak{S}_{\mathscr{H}} := \{\pi \mid \mathscr{Z}_\pi = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q},\ \mathscr{S} \in \mathfrak{D},\, \mathbf{Q} \in O(n)\}, \tag{2.14}$$

where $O(n)$ denotes the group of $n \times n$ orthogonal matrices and $\mathfrak{D} \in \mathfrak{C}$ is the set of diagonal cumulant tensors, defined formally as

$$\mathfrak{D} := \left\{\mathscr{S} \mid \mathscr{S}_{ijkl} = 0\ \forall ijkl \neq iiii \text{ and} \mathscr{S}_{iiii} \geq \epsilon_{\mathscr{S}}\right\}, \tag{2.15}$$

[3] For instance, at most one independent component is Gaussian distributed.

Here, we have the additional requirement[4] that each diagonal cumulant is greater than or equal to a small constant $\epsilon_{\mathscr{S}} > 0$. The expected cumulant tensor $\mathscr{Z}$ is typically not known owing to finite data length approximations and non-Gaussian additive noise [29] and so in general $\mathscr{Z}$ cannot be *fully* diagonalized by a linear transform. As a result, contrast functions are used to approximately diagonalize $\mathscr{Z}$ and maximize the independence of the system.

2.3.3 *Contrast Functions* Comon [22] proposed the use of contrast functions as a solution to tractably measure independence even when the independent components are left distribution-free. A contrast function $\varrho : \mathbb{P}(X) \mapsto \mathbb{R}$ is a mapping from the space of distributions to the real line and can be thought of as a tractable approximation of mutual information. For a function $\varrho$ to be a contrast function it must be both permutation and scale invariant, due to the ICA ambiguities, as well as being maximum if and only if components are statistically independent. Comon [22] proposed various cumulant-based contrast functions that are Edgeworth expansions of information theoretic measures such as negative mutual information, maximum likelihood and negentropy. In Section 3.4, we utilize contrast functions as a measure of independence as part of our compressive ICA algorithms. For further details, comprehensive reviews of cumulants and tensors can be found in [22, 28].

2.3.4 *Existing ICA Algorithms* The Fast ICA algorithm developed by Hyvarinen in [40] is a computationally efficient algorithm that iteratively estimates the column vectors of the mixing matrix in a projection pursuit manner with respect to some measure of independence (e.g. non-Gaussianity). In general, the Fast ICA algorithm requires an initial prewhitening step as discussed in Section 2.3.1. A limitation of the Fast ICA algorithm requires access to the whole dataset for each iteration of the algorithm resulting in a memory and computational complexity that is dependent on the number of samples $N$. The algorithm proposed by Comon in [22] and the joint approximation diagonalization of Eigen-matrices (JADE) algorithm proposed by Cardoso et al. [18] estimate the mixing matrix directly from the fourth-order cumulant tensor associated with the data. In principle, the two algorithms attempt to approximately diagonalize the fourth-order cumulant tensor of the data with respect to a mixing matrix estimate by maximizing a contrast function (see 2.3.3). The JADE algorithm [18] estimates the mixing matrix by jointly diagonalizing the Eigen-matrices of the cumulant tensor with respect to a mixing matrix. Utilizing a Given's rotation scheme, an optimal mixing matrix can be estimated that best diagonalizes the set of Eigen-matrices. Similarly, Comon proposed an ICA algorithm that recursively employs a rotation scheme on pairwise cumulants to approximately diagonalize the fourth-order cumulant tensor with respect to a contrast function. Comon showed that simple functions based on the cumulant tensors are tractable approximations to information theoretic measures like negentropy and mutual information [22]. In contrast to the Fast ICA algorithm, both the Comon and JADE algorithm have a computational complexity that is independent to the number of samples $N$. We have detailed three of the most well-known algorithms in the ICA literature and we will compare these algorithms with our proposed compressive ICA scheme in Section 6. See [42] for a thorough exposition on other existing ICA algorithms.

As discussed in Section 2.2, the fourth-order cumulant tensor will act as an intermediary statistic space ($\mathbb{S} = \mathfrak{C}$). It is well documented through identifiability results in the cumulant-based ICA literature

[4] A standard requirement in ICA is that at maximum one diagonal cumulant $\mathscr{S}_{iiii}$ can be zero which arises from the ICA assumption that at maximum one source signal $s_i$ is Gaussian [42]. Here, we have the slightly stronger assumption that all source signals are non-Gaussian.

[22, 29] that the parameters of the ICA model, namely the mixing matrix $\mathbf{Q}$, can be estimated solely from the fourth-order cumulant tensor. As such, the fourth-order cumulant tensor can be seen in its own right as a sketch, albeit inefficient with respect to compression being $\mathscr{O}(n^4)$. In the next section, we motivate the principles behind sketching the fourth-order cumulant tensor to form a compact representational sketch that has size $\mathscr{O}(n^2)$.

## 3. Compressive Learning Principles for Cumulant ICA

It was discussed in Section 2.3.2 that the model set $\mathfrak{S}_{\mathscr{H}}$ of the ICA problem defined in (2.14) maximizes any given cumulant-based contrast function [22]. The model set $\mathfrak{S}_{\mathscr{H}}$ is itself a low-dimensional space residing in the space of cumulant tensors $\mathfrak{C}$. Specifically, $\mathfrak{S}_{\mathscr{H}}$ can be described as the product set of the set of $n \times n$ orthogonal matrices, denoted $O(n)$, and the set of diagonal cumulant tensors $\mathfrak{D}$ that was defined in (2.13). We can therefore initially count the degrees of freedom of the model set $\mathfrak{S}_{\mathscr{H}}$:

- $\mathfrak{D}$ - A maximum of $n$ degrees of freedom on the leading diagonal.
- $O(n)$ - A maximum of $\frac{n(n-1)}{2}$ degrees of freedom [56].

In total, the model set has a maximum of $\frac{n(n+1)}{2}$ degrees of freedom. In comparison, the space of fourth-order cumulant tensors $\mathfrak{C}$, in which the model set resides, has $p := \binom{n+3}{4} \approx \mathscr{O}(n^4)$ degrees of freedom. As the model set is of low complexity, in principle we could form a sketch of the fourth-order cumulant tensor $\mathscr{Z}$ and estimate the parameters of the ICA model, namely the mixing matrix $\mathbf{Q}$, solely from the sketch. The sketch of the fourth-order cumulant tensor $\mathscr{Z}$ is defined by

$$\mathbf{y}^{\mathbf{w}} = \mathscr{A}(\mathscr{Z}), \tag{3.1}$$

where $\mathbf{w}$ denotes that the sketch is acting on the whitened data $\mathbf{z}$. The computation of the sketch is very related to the sketching method of compressive PCA highlighted in Table 1. Akin to compressive PCA, the sketching operator $\mathscr{A}$ acts on the finite-dimensional space of fourth-order cumulant tensors instead of the infinite-dimensional probability space which is left unspecified due to the nature of the ICA model. The ICA sketch defined in (3.1) draws strong connections to finite-dimensional compressive sensing [15, 30] where limited (random) measurements of a finite-dimensional sparse vector are taken to reduce the complexities associated with signal acquisition. Throughout the compressive sensing literature [14, 15, 30], the RIP is a fundamental tool that is extensively used to show that a sketching operator $\mathscr{A}$ stably embeds all elements of the model set into a compressive domain $\mathbb{R}^m$, provided that the sketch dimension $m$ is of sufficient size. In other words, given a sketching operator $\mathscr{A}$, it proves that the distance between every pair of signals in the model set are approximately preserved under the action of the sketch therefore providing a near isometry. In the case of compressive ICA, if $\forall \mathscr{Z}_1, \mathscr{Z}_2 \in \mathfrak{S}_{\mathscr{H}}$ and an RIP constant $\delta \in (0, 1)$, then

$$(1-\delta)\|\mathscr{Z}_1 - \mathscr{Z}_2\|^2 \leq \|\mathscr{A}(\mathscr{Z}_1 - \mathscr{Z}_2)\|^2 \leq (1+\delta)\|\mathscr{Z}_1 - \mathscr{Z}_2\|^2, \tag{3.2}$$

provided that the sketch size $m$ is of sufficient dimension. In many cases, the sketch size $m$ is sufficient to be of the order of the degrees of freedom of the model set. In [7, 10], it is proved that if the lower RIP (LRIP) holds for a given sketching operator $\mathscr{A}$, e.g. the left of (3.2), then there exists a robust decoder $\Delta$ that recovers a signal from the model set in a stable manner with respect to noise and signals that lie close to the model set. Moreover, it is proved in [10] that if the LRIP holds for the sketching operator $\mathscr{A}$ on the model set $\mathfrak{S}_{\mathscr{H}}$, then the decoder $\Delta$ is robust and can be the constrained $\ell_2$ optimization, for instance

$$\Delta\left(\mathbf{y}^{\mathbf{w}}, \mathscr{A}\right) \in \min_{\mathscr{Z} \in \mathfrak{S}_{\mathscr{H}}} \|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z})\|_2. \tag{3.3}$$

In principle, if the RIP can be proved for a sketching operator $\mathscr{A}$ on the ICA model set $\mathfrak{S}_{\mathscr{H}}$, then we have an optimization strategy for solving the compressive ICA problem.

## 4. Compressive ICA Theory

We begin by explicitly defining the sketching operator $\mathscr{A}: \mathfrak{C} \mapsto \mathbb{R}^m$ as

$$\mathscr{A}(\mathscr{Z}) = \mathbf{A}\ \mathrm{vec}(\mathscr{Z}), \tag{4.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times p}$ and vec denotes the vectorization operator. Here, we assume $\mathbf{A}$ is some random measurement matrix where the entries $\mathbf{A}_{ij}$ are sampled according to some distributing law, $\mathbf{A}_{ij} \sim \Lambda$. In this paper, we consider two randomized linear dimension reduction maps, namely the Gaussian map and the subsampled randomized Hadamard transform (SRHT) stated below. The CICA RIP, our main result stated in Theorem 4.1, is proved using the Gaussian map; however fast Johnson-Lindenstrauss transforms (FJLT), for instance the SRHT, still work in practice as will be discussed in Section 6.

4.0.1 *Gaussian Maps* The most traditional randomized linear dimension reduction map is the sub-Gaussian matrix which has been used extensively in the CS literature [15, 30]. The sub-Gaussian matrix $\mathbf{A} \in \mathbb{R}^{m \times p}$ has entries that follow

$$\mathbf{A}_{ij} \sim \mathscr{N}\left(0, m^{-\frac{1}{2}}\right). \tag{4.2}$$

Gaussian maps typically require $\mathscr{O}(mp)$ in memory as well as exhibiting a computational complexity of $\mathscr{O}(mp)$.

4.0.2 *Subsampled Randomized Hadamard Transform* The SRHT is an instance of an FJLT that approximates the properties of the full Gaussian map [46]. Here, $\mathbf{A} \in \mathbb{R}^{m \times p}$ is defined as

$$\mathbf{A} = \sqrt{\frac{p}{m}}\mathbf{RHD}, \tag{4.3}$$

where

- $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix whose elements are independent random signs $\{1, -1\}$;
- $\mathbf{H} \in \mathbb{R}^{p \times p}$ is a normalized Walsh–Hadamard matrix that is scaled by $p^{-\frac{1}{2}}$ so it is an orthogonal matrix;
- $\mathbf{R} \in \mathbb{R}^{m \times p}$ is a matrix consisting of a subset of $m$ randomly sampled rows from the $p \times p$ identity matrix.

The SRHT is particularly cheaper to compute and store in comparison to the Gaussian map. As we do not explicitly store $\mathbf{H}$, the SRHT only requires $\mathscr{O}(m+p)$ in memory [58]. In addition, the computational complexity of computing the sketch reduces to $\mathscr{O}(p\log(m))$ in comparison to using the Gaussian map [1, 58]. Next, we state our main result of the paper.

Theorem 4.1. (Compressive ICA RIP) Denote by $\mathscr{A}$ the Gaussian map sketching operator defined in (4.2). Then $\forall \mathscr{X}_1, \mathscr{X}_2 \in \mathfrak{S}_{\mathscr{H}}$, the sketching operator $\mathscr{A}$ satisfies the RIP in (3.2) with constant $\delta \in (0,1)$ and probability $1-\xi$ provided that

$$m \geq \frac{C}{\delta^2} \max\left\{2n(n+1)\log(C_0), \log\left(\frac{6}{\xi}\right)\right\}, \tag{4.4}$$

where $C > 0$ is an absolute constant and $C_0 = C_0(\epsilon_{\mathscr{S}})$ is a constant that is dependent on $\epsilon_{\mathscr{S}}$ defined in Lemma 4.1.

The proof of Theorem 4.1 is detailed in Section 4.1.

Corollary 4.1. (Information Preservation) Let $\mathscr{X}^* \in \mathfrak{C}$ be an arbitrary fourth-order cumulant tensor and denote $\mathbf{y}^{\mathbf{w}} = \mathscr{A}(\mathscr{X}^*) + \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^m$ is some additive noise. Furthermore, let $\tilde{\mathscr{X}} := \Delta\left(\mathbf{y}^{\mathbf{w}}, \mathscr{A}\right)$ denote the solution to (3.3). Given that $\mathscr{A}$ satisfies the RIP in Theorem 4.1, then with probability $1-\xi$

$$\|\mathscr{X}^* - \tilde{\mathscr{X}}\|_F \leq \min_{\mathscr{X}\in\mathfrak{S}_{\mathscr{H}}} \left(2\|\mathscr{X}^* - \mathscr{X}\|_F + \frac{2}{\sqrt{1-\delta}}\|\mathbf{A}\mathrm{vec}\left(\mathscr{X}^* - \mathscr{X}\right)\|_2\right) + \frac{2}{\sqrt{1-\delta}}\|\mathbf{e}\|_2 + \nu, \tag{4.5}$$

where $0 < \nu \leq 1$ is a small positive constant.

*Proof.* Given the LRIP in Theorem 4.1, we use Theorem 7 in [10] to obtain our result. □

The proof of Theorem 4.1 uses covering numbers and $\epsilon$-nets of the normalized secant set of $\mathfrak{S}_{\mathscr{H}}$. In particular, the size of the sketch $m$ required to satisfy the RIP will scale with the upper box counting dimension of the normalized secant set. So that the proof is self-contained, we summarize the following definitions below.

Definition 4.2. (Secant Set) The secant set of a set $\mathfrak{S}_{\mathscr{H}}$ is defined as

$$\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}} := \left\{\mathscr{Y} = \mathscr{X}_1 - \mathscr{X}_2 \mid \mathscr{X}_1, \mathscr{X}_2 \in \mathfrak{S}_{\mathscr{H}}\right\}. \tag{4.6}$$

Definition 4.3. (Normalized Secant Set) The normalized secant set $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$ of a set $\mathfrak{S}_{\mathscr{H}}$ is defined as

$$\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}}\right) := \left\{\mathscr{Y}/\|\mathscr{Y}\|_F \mid \mathscr{Y} \in \left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right) \setminus \{\mathbf{0}\}\right\}, \tag{4.7}$$

where $\mathbf{0}$ defines the zero tensor.

Definition 4.4. (Covering number) Let . The covering number $\mathrm{CN}(\mathfrak{S}_{\mathscr{H}}, \|\cdot\|, \epsilon)$ of a set $\mathfrak{S}_{\mathscr{H}}$ is the *minimum number* of closed balls of radius $\epsilon$, with respect to the norm $\|\cdot\|$, with centres in $\mathfrak{S}_{\mathscr{H}}$ needed to cover $\mathfrak{S}_{\mathscr{H}}$. The set of centres of these balls is a minimal $\epsilon$-net for $\mathfrak{S}_{\mathscr{H}}$.

LEMMA 4.1. (Covering number of $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$) The covering number of $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$ with respect to the Frobenius norm $\|\cdot\|_F$ is

$$\mathrm{CN}\left(\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right), \|\cdot\|_F, \epsilon\right) \leq \left(\frac{C_0}{\epsilon}\right)^{2n(n+1)}, \tag{4.8}$$

where $C_0 = C_0(\epsilon_{\mathscr{S}}) > 0$ is some constant.

*Proof.* See Appendix A.1. □

DEFINITION 4.5. (Upper box counting dimension) The upper box counting dimension of a set $S$ is defined as

$$\dim_{\mathrm{B}}(S) := \limsup_{\epsilon \to 0} \log[\mathrm{CN}\left(S, \|\cdot\|, \epsilon\right)]/\log[1/\epsilon]. \tag{4.9}$$

## 4.1 *Proof of Theorem 4.1*

*Proof.* To prove an RIP exists for the ICA model set $\mathfrak{S}_{\mathscr{H}}$ using the sketching operator $\mathscr{A}$ defined in (4.1), we follow a similar line of argument to [14, 52] by using an $\epsilon$-covering of $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$ to extend the concentration results of the random Gaussian matrix $\mathbf{A}$ uniformly over the whole low-dimensional set. Specifically, we use the *Recipe* framework proposed by Puy *et al.* [49], to formulate the compressive ICA RIP proof. The proof is separated by showing that the following assumptions hold:

**(A1)** The normalized secant set, denoted $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$, has finite upper box counting dimension $\dim_{\mathrm{B}}\left(\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)\right)$ which is strictly bounded by $s \geq 1$, $\dim_{\mathrm{B}}\left(\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)\right) < s$

**(A2)** The sketching operator $\mathscr{A}$ satisfies the following concentration inequalities:

$$\mathbb{E}_{\mathscr{A}\sim\Lambda}\left(\|\mathscr{A}(\mathscr{Z})\|_2^2\right) \;=\; \|\mathscr{Z}\|_F^2, \tag{4.10}$$

and

$$\Pr\left(\left|\|\mathscr{A}(\mathscr{Z})\|_2^2 - \|\mathscr{Z}\|_F^2\right| \geq \delta\|\mathscr{Z}\|_F^2\right) \;\leq\; 2e^{-pc_0}, \tag{4.11}$$

for a constant $c_0$ depending on $\delta$ [49].

We begin with Assumption **(A1)**. Using Lemma 4.1 and the definition of the upper box counting dimension in Definition 4.4, it can be seen that $\dim_{\mathrm{B}}\left(\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)\right) \leq 2n(n+1)$, so for any $s > 2n(n+1)$ we satisfy Assumption **(A1)**. To prove Assumption **(A2)**, we have the following definition.

DEFINITION 4.6. (Subguassian random variable) A sub-Gaussian random variable $X$ is a random variable that satisfies

$$(\mathbb{E}|X|^q)^{1/q} \leq C_1\sqrt{q} \text{ for all } q \geq 1,$$

with $C_1 > 0$. The sub-Gaussian norm of $X$, denoted by $\|X\|_{\Psi_2}$, is the smallest $C_1$ for which the last property holds, i.e.

$$\|X\|_{\Psi_2} := \sup_{q\geq 1}\left\{q^{-1/2}(\mathbb{E}|X|^q)^{1/q}\right\}.$$

Let $\mathbf{A}_i$ denote the $i$th row of the random Gaussian matrix $\mathbf{A}$. Then we use the fact [49, 61] that

$$\|\mathbf{A}_i^T \operatorname{vec}(\mathscr{X})\|_{\Psi_2} \leq D\|\mathscr{X}\|_F \tag{4.12}$$

for all $\mathscr{X} \in \mathfrak{C}$, where $D > 0$ is an absolute constant. Therefore, Assumption **A2** is satisfied. Finally, using Theorem 8 of [49], we get the desired RIP result in Theorem 4.1. □

## 4.2 *Finite Sample Effects*

In practice, the sketch is constructed from a finite set of data $\{\mathbf{z}_i\}_{i=1}^N$ such that

$$\hat{\mathbf{y}}^{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \Phi^{\mathbf{w}}\left(\mathbf{z}_i\right), \tag{4.13}$$

where $\Phi^{\mathbf{w}}(\cdot)$ is the feature function discussed in Section 2.1 acting on the whitened data $\mathbf{z}$. For compressive ICA we can explicitly define the feature function, acting on the whitened data, as

$$\Phi^{\mathbf{w}}(\mathbf{z}) = \langle \mathbf{A}_j, \mathbf{z}^{\otimes^4}\rangle_F, \tag{4.14}$$

for $j = 1, \ldots, m$, where $\mathbf{A}_j \in \mathbb{R}^p$ are the rows of a Gaussian matrix $\mathbf{A}$ and $\langle\cdot\rangle$ denotes the Frobenius inner product. Furthermore, for shorthand we denote $\mathbf{z}^{\otimes^4} = \mathbf{z}\otimes\mathbf{z}\otimes\mathbf{z}\otimes\mathbf{z}$, where $\otimes$ denotes the Kronecker product. In other words, the feature function is taking random quartics of the data point $\mathbf{z}$. Note that the empirical sketch $\hat{\mathbf{y}}^{\mathbf{w}}$ is equivalent to $\hat{\mathbf{y}}^{\mathbf{w}} = \mathscr{A}\left(\hat{\mathscr{X}}\right)$, as specified in (2.4), where $\hat{\mathscr{X}}$ is the finite data approximation of the fourth-order cumulant tensor $\mathscr{X}$ defined by

$$\begin{aligned}\hat{\mathscr{X}}^4_{ijkl} = \frac{1}{N}\sum_{i,j,k,l=1}^{N} z_i z_j z_k z_l - \frac{1}{N^2}\sum_{i,j=1}^{N} z_i z_j \sum_{k,l=1}^{N} z_k z_l - \frac{1}{N^2}\sum_{i,k=1}^{N} z_i z_k \sum_{j,l=1}^{N} z_j z_l \\ - \frac{1}{N^2}\sum_{i,l=1}^{N} z_i z_l \sum_{j,k=1}^{N} z_j z_k.\end{aligned} \tag{4.15}$$

In this case, the error $\mathbf{e}$ defined in Theorem 4.1 can be attributed to the finite sample effects of approximating the true fourth-order cumulant tensor $\mathscr{X}$ from finite data. We now state our final result of this section.

Theorem 4.7. (Finite Sample Effects) Let $\mathscr{A}(\mathscr{X}) = \mathbf{A}\operatorname{vec}(\mathscr{X})$ denote the sketching operator where $\mathbf{A}_{ij} \sim \mathscr{N}\left(0, m^{-\frac{1}{2}}\right)$. Furthermore, let the independent components $\mathbf{s}$ have bounded support such that

$\|\mathscr{S}\|_F \le R$. Given that $\hat{\mathscr{Z}}$ is the finite approximation fourth-order cumulant tensors computed from the random draw of finite samples $\mathbf{z}_1, \dots, \mathbf{z}_N$, then with probability at least $1-\rho-\xi$

$$\|\mathscr{A}(\mathscr{Z}) - \mathscr{A}(\hat{\mathscr{Z}})\|_2 \le \frac{CR\left(1+\sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{4.16}$$

*Proof.* See Appendix B. □

### 4.3 *Discussion*

The results in this section are all based on proving an RIP on the model set $\mathfrak{S}_{\mathscr{H}}$ defined in (2.14), where it is assumed the data $\mathbf{x}$ have been prewhitened to reduce the ICA model to $\mathbf{z} = \mathbf{Q}\mathbf{s}$ as discussed in Section 2.3. The prewhitening stage removes some of the degrees of freedom within the ICA inference task as it is necessary to estimate an orthogonal mixing matrix $\mathbf{Q}$. In some sketching cases, we may only see the data once, for example in the streaming context [60], and therefore prewhitening may not be possible. The fact that we are now estimating an arbitrary mixing matrix $\mathbf{M}$ instead of an orthogonal mixing matrix $\mathbf{Q}$ increases the degrees of freedom from $\frac{n(n+1)}{2}$ to $n(n+1)$. As a result, we must sketch the unwhitened moment tensor $\mathscr{X}$ such that

$$\mathbf{y}^{\mathbf{u}} = \mathscr{A}(\mathscr{X}), \tag{4.17}$$

where $\mathscr{A}(\cdot) = \mathbf{A}\text{vec}(\cdot)$ and $\mathbf{A} \in \mathbb{R}^{m \times p}$ is a random matrix as defined in (4.1). Here, $\mathbf{u}$ denotes that the sketch is acting on the unwhitened data $\mathbf{x}$. In addition, the feature function $\Phi^{\mathbf{u}}(\cdot)$ for the unwhitened data can be defined as

$$\Phi^{\mathbf{u}}(\mathbf{x}) = \begin{bmatrix} \langle \mathbf{A}_j, \mathbf{x}^{\otimes^4} \rangle_F \\ \mathbf{x}^{\otimes^2} \end{bmatrix}, \tag{4.18}$$

for $j = 1, \dots, m$, where $\mathbf{A}_j \in \mathbb{R}^p$ are the rows of the matrix $\mathbf{A}$. Note that the feature function for the unwhitened data now includes quadratic moments[5] , as well as random quartic moments, that are needed to estimate the mixing matrix $\mathbf{M}$ which has extra degrees of freedom. Recall from (2.9) that the mixing matrix $\mathbf{M}$ has the following decomposition: [29]

$$\mathbf{M} = \mathbf{V}^{-1}\mathbf{Q}, \tag{4.19}$$

where $\mathbf{V} := \Pi^{-\frac{1}{2}}\mathbf{P}^T$ is computed via the eigendecomposition of the covariance matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$, where $\mathbf{P} \in \mathbb{R}^{d \times n}$ and $\Pi \in \mathbb{R}^{n \times n}$ are the orthogonal and diagonal matrix, respectively.

## 5. CICA Algorithms

In this section, we propose two distinct compressive ICA algorithms to estimate the mixing matrix $\mathbf{M}$ for both the whitened and unwhitened case.

[5] One could further reduce the size of the unwhitened sketch by instead computing random quadratic moments, however the reduction in complexity is minimal and therefore we leave this for future work.

### 5.1 *Iterative Projection Gradient*

Iterative projection gradient (IPG) descent is a popular optimization scheme which enforces low-dimensional structure e.g. sparsity, rank, etc., by projecting the object of interest onto the model set $\mathfrak{S}_{\mathscr{H}}$after each subsequent gradient step. An iterative hard thresholding scheme was proposed in sparsity based compressive sensing [8, 9], where the smallest $n - k$ absolute entries are thresholded to zero to enforce the sparsity constraint and project the object onto the $k$-sparse model set. Blumensath [7] shows that the thresholding operator is an orthogonal projection onto the $k$-sparse set thereby projecting to an element on the model set that is of minimal distance. For the case of compressive ICA, we also seek an orthogonal projection on to the ICA model set $\mathfrak{S}_{\mathscr{H}}$. Formally, we can define an orthogonal projection operator $\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}} : \mathfrak{C} \mapsto \mathfrak{S}_{\mathscr{H}}$of a fourth-order cumulant tensor $\mathscr{Z}^*$ as

$$\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}\left(\mathscr{Z}^*\right) \in \operatorname{arg\,min}_{\mathscr{Z}\in\mathfrak{S}_{\mathscr{H}}} \|\mathscr{Z}^* - \mathscr{Z}\|_F. \tag{5.1}$$

In other words, $\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}$projects the object $\mathscr{Z}^* \in \mathfrak{C}$ onto the element in the model set that is of minimum distance w.r.t the Frobenius norm. In practice it is often difficult to find a projection operator that is both orthogonal and tractable in terms of computation. In [16, 17], Cardoso showed that the ICA model set $\mathfrak{S}_{\mathscr{H}} \subseteq \mathfrak{R} \cap \mathfrak{L}$, where $\mathfrak{R}$ is the set of rank-$n$ tensors defined as

$$\mathfrak{R} := \{\mathscr{Z} \in \mathfrak{R} \mid \operatorname{rank}(\bar{\mathbf{Z}}) = n\}, \tag{5.2}$$

where $\bar{\mathbf{Z}} \in \mathbb{R}^{n^2\times n^2}$ is the matrix formed by rearranging the elements of the tensor $\mathscr{Z}$ into an $n^2 \times n^2$ Hermitian matrix and where rank defines the standard matrix rank [17], and $\mathfrak{L}$ is the set of supersymmetric tensors defined by

$$\mathfrak{L} := \{\mathscr{Z} \in \mathfrak{L} \mid \mathscr{Z}_{q(ijkl)} = \mathscr{Z}_{ijkl}\}, \tag{5.3}$$

where $q$ defines all permutations of the index $ijkl$. In fact, Cardoso proved in [16] that locally the converse is true, for instance let $\mathscr{Z}'$ be within some neighbourhood of $\mathscr{Z}$ then the following holds:

$$\mathscr{Z}' \in \mathfrak{R} \cap \mathfrak{L} \implies \mathscr{Z}' \in \mathfrak{S}_{\mathscr{H}}. \tag{5.4}$$

Therefore, within some neighbourhood of $\mathscr{Z}^*$, projecting onto the ICA model set $\mathfrak{S}_{\mathscr{H}}$is equivalent to projecting onto $\mathfrak{R} \cap \mathfrak{L}$ (i.e. $\mathfrak{R} \cap \mathfrak{L} \subseteq \mathfrak{S}_{\mathscr{H}}$). Moreover, in [13], Cadzow proved that alternate projections onto $\mathfrak{R}$ and $\mathfrak{L}$ is guaranteed to converge onto the intersection[6] $\mathfrak{R}\cap\mathfrak{L}$. Fundamentally, the projections onto $\mathfrak{R}$ (rank-$n$ approximation) and $\mathfrak{L}$ (averaging over permutations), denoted by $\mathscr{P}_{\mathfrak{R}}$ and $\mathscr{P}_{\mathfrak{L}}$, respectively, are both simple to compute and are orthogonal. Alternate orthogonal projections onto $\mathfrak{R}$ and $\mathfrak{L}$ ensure a stable projection onto $\mathfrak{R} \cap \mathfrak{L}$ [13], which locally, results in an orthogonal projection onto the ICA model set $\mathfrak{S}_{\mathscr{H}}$. Formally, we define the orthogonal projection $\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}$in Algorithm 1. In practice, Algorithm 1 converges to below a small tolerance in very few iterations ($\sim$ 10 iterations). We can now state our full

[6] In general, rank forcing destroys symmetry while symmetrization destroys the rank-$n$ property, therefore alternate projections are needed until convergence.

CICA IPG algorithm detailed in Algorithm 2. Here, the step size $\mu_j$ is computed optimally to guarantee convergence [7, 9], $\mathscr{A}^*$ denotes the adjoint sketching operator and $\beta$ is a fixed shrinking step size parameter.

```
Algorithm 1 $\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}$ : Orthogonal Projection onto ICA Model Set
Require: Cumulant tensor $\mathscr{Z}^* \in \mathfrak{C}$
  while Not Converged do
    Project onto $\mathfrak{R}$: $\mathscr{Z}^1 = \mathscr{P}_{\mathfrak{R}}(\mathscr{Z})$ (Matricize $\mathscr{Z}$ into a $n^2 \times n^2$ Hermitian matrix and take a rank-$n$ approximation using truncated SVD)
    Project onto $\mathfrak{L}$: $\mathscr{Z}^2 = \mathscr{P}_{\mathfrak{L}}(\mathscr{Z}^1)$ (Average across all permutations of $q(ijkl)$ for all indices $ijkl$)
  end while
```

```
Algorithm 2 CICA_IPG : Iterative Projection Gradient Descent Compressive ICA
Require: Initialisation $\mathscr{Z}^0$, tolerance $\varepsilon$ and shrinking parameter $\beta$.
  while $\|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z}^j)\|_2^2 > \varepsilon$ do
    Compute $\mu_j = \frac{\|\mathscr{A}^*(\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z}^j))\|_F^2}{\|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z}^j)\|_2^2}$
    while $\|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z}^{j+1})\|_2^2 > \|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{Z}^j)\|_2^2$ do
      $\mu_j \leftarrow \beta\mu_j$
      $\mathscr{Z}^{j+\frac{1}{2}} \leftarrow \mathscr{Z}^j + \mu_j\mathscr{A}^*\left(\mathbf{y}^{\mathbf{w}} - \mathscr{A}\left(\mathscr{Z}^j\right)\right)$
      $\mathscr{Z}^{j+1} \leftarrow \mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}\left(\mathscr{Z}^{j+\frac{1}{2}}\right)$
    end while
  end while
```

5.1.1 *Unwhitened IPG* It was discussed in Section 4.3 that it is often convenient, from an online processing point of view, to directly sketch the unwhitened data $\mathbf{x}$. Using the properties of the matrix-tensor product [28], it can be seen that

$$\mathbf{A}\mathrm{vec}\left(\mathscr{X}\right) = \mathbf{A}\bar{\mathbf{V}}^{-1}\mathrm{vec}\left(\mathscr{Z}\right), \tag{5.5}$$

where $\bar{\mathbf{V}} := \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V}$. As defined in (4.18), the unwhitened feature function $\Phi^{\mathbf{u}}$ includes the second-order moment of $\mathbf{x}$, namely $\mathbf{x}^{\otimes^2}$. The empirical sketch $\hat{\mathbf{y}}^{\mathbf{u}}$ therefore includes the sample covariance $\hat{\Sigma} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\otimes^2}$, which can be used to estimate an approximation of $\mathbf{V}$, denoted $\hat{\mathbf{V}}$, by using the eigenvalue decomposition of $\hat{\Sigma}$ [22] at the beginning of Algorithm 2. By denoting $\hat{\bar{\mathbf{V}}} := \hat{\mathbf{V}} \otimes \hat{\mathbf{V}} \otimes \hat{\mathbf{V}} \otimes \hat{\mathbf{V}}$, the gradient step in Algorithm 2 can be replaced by

$$\mathscr{Z}^{j+\frac{1}{2}} = \mathscr{Z}^j + \mu_j\mathbf{A}^T\big(\mathbf{y}^{\mathbf{u}} - \mathbf{A}\hat{\bar{\mathbf{V}}}^{-1}\mathrm{vec}\big(\mathscr{Z}^j\big)\big), \tag{5.6}$$

as well as the associated step size $\mu_j$ and stopping criteria. As a result, the CICA IPG algorithm proceeds as normal by employing the original orthogonal projection $\mathscr{P}_{\mathfrak{S}_{\mathscr{H}}}$

### 5.2 *Alternating Steepest Descent*

The second proposed algorithm in the way of alternating steepest descent (ASD) is inherently different from the IPG scheme previously discussed. To see why, it is insightful to rewrite (3.3) in terms of the elements of the product set $\mathfrak{D}$ and $\mathrm{O}(n)$:

$$\min_{\substack{\mathbf{Q}^T\mathbf{Q}=I \\ \mathscr{S}\in\mathfrak{D}}} F\,(\mathscr{S},\mathbf{Q}) = \|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q})\|_2^2, \tag{5.7}$$

where we have used the multilinear property discussed in (2.12). As the optimization problem is now explicitly defined by the mixing matrix $\mathbf{Q}$ and a sparse diagonal tensor $\mathscr{S}$, it is sufficient to optimize with respect to these parameters in an ASD scheme. This approach contrasts the IPG scheme, as once we initialize the mixing matrix $\mathbf{Q}$ and the diagonal cumulant tensor $\mathscr{S}$ appropriately, then we can optimize directly on the model set $\mathfrak{S}_{\mathscr{H}}$. We can initially state the ASD steps:

(1) $\mathscr{S}^* = \min_{\mathscr{S}\in\mathfrak{D}} F(\mathscr{S},\mathbf{Q})$

(2) $\mathbf{Q}^* = \min_{\mathbf{Q}^T\mathbf{Q}=I} F(\mathscr{S}^*,\mathbf{Q})$

Note that the diagonal cumulant tensor $\mathscr{S} \in \mathfrak{D}$ can be simply reformulated as an $n$ sparse vector with known support, therefore one can perform element-wise differentiation on the $n$ entries $\mathscr{S}_{iiii}$ for $i = 1 : n$. The second step requires more attention as we have the constraint $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ (i.e. $\mathbf{Q} \in \mathrm{O}(n)$). The set of $n \times n$ orthogonal matrices is an instance of a Stiefel manifold [63], therefore $F$ is minimized directly on the Stiefel manifold.

#### 5.2.1 *Stiefel Manifold Optimization*

Given a feasible matrix $\mathbf{Q}$ and the gradient $\nabla_{\mathbf{Q}} F = \left(\frac{\partial F(\mathscr{S},\mathbf{Q})}{\partial \mathbf{Q}_{ij}}\right)$, define a skew-symmetric matrix $\mathbf{B}$ as

$$\mathbf{B} = \nabla_{\mathbf{Q}} F \mathbf{Q}^T - \mathbf{Q}(\nabla_{\mathbf{Q}} F)^T. \tag{5.8}$$

The update on the Stiefel manifold is determined by the Crank–Nicholson scheme [27] denoted

$$Y(\tau) = \mathbf{Q} - \frac{\tau}{2}\mathbf{B}(\mathbf{Q} + Y(\tau)), \tag{5.9}$$

where $Y(\tau) = (I - \frac{\tau}{2}\mathbf{B})^{-1}(I + \frac{\tau}{2}\mathbf{B})\mathbf{Q}$. The matrix $(I - \frac{\tau}{2}\mathbf{B})^{-1}(I + \frac{\tau}{2}\mathbf{B})$ is referred to as the Cayley transform [63] of $\mathbf{B}$. The descent curve $Y(\tau)$ has the following useful features:

- $Y(\tau)$ is smooth on $\tau$
- $Y(0) = \mathbf{Q}$
- $Y(\tau)^T Y(\tau) = \mathbf{Q}^T\mathbf{Q}$ for all $\tau \in \mathbb{R}$.

As a result, we perform a steepest descent on $\mathbf{Q}$ with line search along the descent curve $Y(\tau)$ with respect to $\tau$. For more details on optimization methods constrained to the Stiefel manifold refer to [63].

We can now state our second proposed CICA algorithm in Algorithm 3.

**Algorithm 3** $\texttt{CICA}_{\texttt{ASD}}$ : Alternating Steepest Descent Compressive ICA

```
Require: Initialisation $\mathscr{X}^0 = \mathscr{S}^0 \times_1 \mathbf{Q}^0 \times_2 \mathbf{Q}^0 \times_3 \mathbf{Q}^0 \times_4 \mathbf{Q}^0$, tolerance $\varepsilon$ and step size $\mu$.
  while $\|\mathbf{y}^{\mathbf{w}} - \mathscr{A}(\mathscr{X}_j)\|_2^2 > \varepsilon$ do
      $\mathscr{S}^{j+1} = \mathscr{S}^j + \mu \nabla_{\mathscr{S}} F(\mathscr{S}^j, \mathbf{Q}^j)$
      while Perform line search do
          $Y(\tau) = \mathbf{Q} - \frac{\tau}{2}\mathbf{B}(\mathbf{Q} + Y(\tau))$
          $\mathbf{Q}^{t+1} \leftarrow Y(\tau^*)$
      end while
      $\mathscr{X}^{j+1} \leftarrow \mathscr{S}^{j+1} \times_1 \mathbf{Q}^{j+1} \times_2 \mathbf{Q}^{j+1} \times_3 \mathbf{Q}^{j+1} \times_4 \mathbf{Q}^{j+1}$
  end while
```

5.2.2 *Practicalities* We start by stating the computational complexity of each proposed CICA algorithm. Here, we assume that a fast SRHT, as discussed in 4.0.2, is used to compute the sketch. For the IPG scheme, the symmetry projection $\mathscr{P}_{\mathfrak{L}}$ costs $\mathscr{O}(n^4)$ flops through averaging along all index permutations. A rank-$r$ approximation of a general matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ costs $\mathscr{O}(r^2(n+m))$ flops [64], therefore the rank projection operator $\mathscr{P}_{\mathfrak{R}}$ costs a total of $\mathscr{O}(n^4)$ flops. The gradient step in Algorithm 2 costs a total of $\mathscr{O}(p\log(m))$ flops due to the use of the sketching operator $\mathscr{A}(\mathscr{X}^j)$ at each iteration which results in the IPG algorithm therefore having a total cost of $\mathscr{O}(p\log(m) + n^4))$ flops. In the second proposed ASD algorithm, the gradient step in terms of the diagonal tensor in Algorithm 3, again has a cost of $\mathscr{O}(p\log(m))$ flops. The line search $Y(\tau)$ costs a total of $\mathscr{O}(n^3)$ flops [63] resulting in the ASD algorithm having a computational complexity of $\mathscr{O}(p\log(m) + n^3)$. Note that both proposed CICA algorithms have computational complexity that is independent of the length of the data $N$ which can be extremely large for modern day applications.

As is the case for the general ICA problem, the compressive ICA optimization problem is non-convex and both algorithms proposed may be prone to converging to local minima. As a result, we consider the option of possible restarts at random initializations to obtain a good solution. We also consider a proxy projection operator that uses a Given's rotation scheme, popular in many ICA algorithms (see [18, 22]), that approximately diagonalizes the cumulant tensor $\mathscr{X}$ with respect to some contrast function, followed by thresholding the cross cumulants of that approximately diagonalized tensor to zero [55]. We have observed in practice that this proxy projection operator is less sensitive to the non-convex landscape of the optimization problem, which could be explained by the robustness of Given's rotations [22], hence multiple restarts are rarely required. The proxy projection operator, which we denote by $\hat{\mathscr{P}}_{\mathfrak{S}_{\mathscr{H}}}$ costs $\mathscr{O}(n^4)$ flops for the Given's rotation scheme to approximately diagonalize the cumulant tensor [22], and $\mathscr{O}(n^4 - n)$ flops for the thresholding of the cross-cumulants. Therefore, in total the proxy IPG algorithm has approximately the same computational complexity as our previous IPG algorithm.

## 6. Empirical Results

### 6.1 *Phase Transition*

Phase transitions are an integral part of analysis that are used frequently in the compressive sensing literature [3] to show a sharp change in the probability of successful reconstruction of the low-

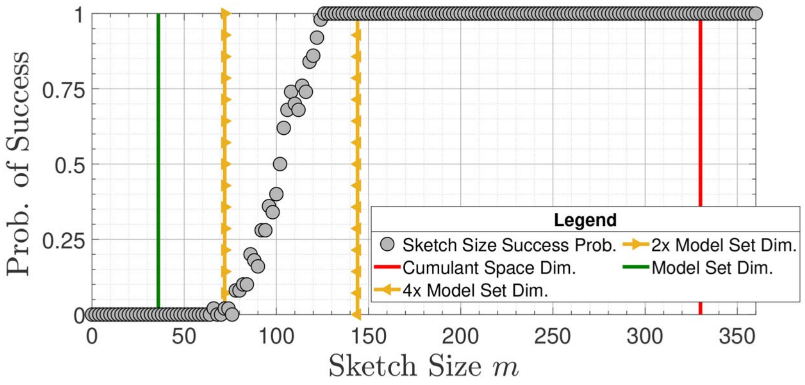

FIG. 2. A phase transition between unsuccessful and successful mixing matrix inference as the sketch size $m$ increases and the number of independent components is fixed at $n = 8$.

dimensional object as the sketch size $m$ increases. The location at which the phase transition occurs can provide a tight bound on the required sketch size needed given the number of independent components $n$ and further consolidates the theoretical bound of the RIP derived in Section 4. To set up the phase transition experiment, we constructed the expected cumulant tensor $\mathscr{S}$ of $n$ Laplacian sources and transformed the tensor with an orthogonal mixing matrix $\mathbf{M}$ using the multilinear property in (2.12), resulting in an expected cumulant tensor $\mathscr{X}$. For each number of independent components $n$, 250 Monte Carlo simulations on the mixing matrix $\mathbf{M}$ were executed for increasing sketch size $m$ between 2 and 700. A successful reconstruction was determined if the Amari error[7] [2] between the true mixing matrix $\mathbf{M}$ and the estimated mixing matrix $\hat{\mathbf{M}}$, defined by

$$d(\mathbf{M}, \hat{\mathbf{M}}) = \frac{1}{2n} \sum_{i=1}^{n} \left( \frac{\sum_{j=1}^{n} |b_{ij}|}{\max_j |b_{ij}|} - 1 \right) + \frac{1}{2n} \sum_{j=1}^{n} \left( \frac{\sum_{i=1}^{n} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right), \tag{6.1}$$

was smaller than $d(\mathbf{M}, \hat{\mathbf{M}}) \leq 10^{-6}$, where $b_{ij} = (\mathbf{M}\hat{\mathbf{M}}^{-1})_{ij}$. The probability of successful reconstruction was given by the number of successful reconstructions within the 250 Monte-Carlo tests. We use the IPG version of the CICA algorithm for these results, although the ASD version provides nearly exactly the same results. It is insightful to begin by fixing the number of sources, here $n = 8$, to highlight the sharp transition as shown in Fig. 2. We highlight some important bounds including the multiples of two and four times the dimension of the model set $\mathfrak{S}_{\mathscr{H}}$ depicted by the orange lines. For comparison, the dimension of the space of cumulant tensors $\mathfrak{C}$, in other words the size of the cumulant tensor, is shown by the red line. The phase transition occurs in between two and four times the model set dimension indicating that choosing $m \geq 2n(n + 1)$ would be sufficient in successfully inferring the mixing matrix with high probability.

Figure 3 generalizes the single phase transition result for the number of independent components varying between $n = 2$ and $n = 10$. Once again, the important bounds of the model set dimension (green), two and four multiples of the model set dimension (orange) and the dimension of the space

[7] The Amari error is used widely in the ICA literature as it is both scale and permutation invariant, which are the two inherent ambiguities of ICA inference.

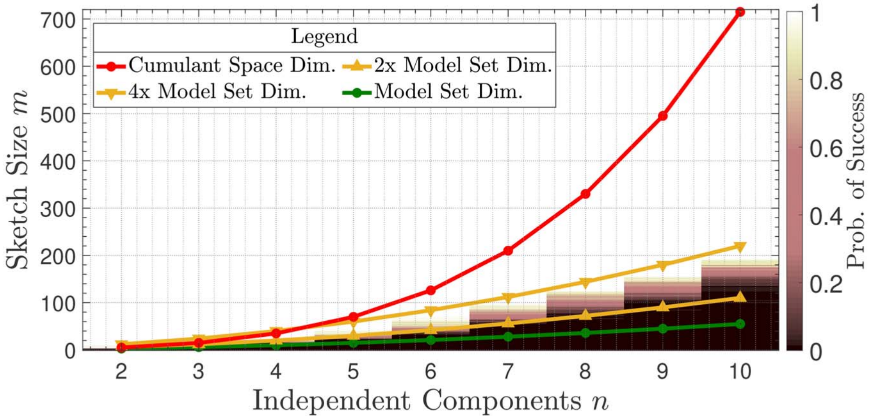

FIG. 3. A phase transition between unsuccessful and successful mixing matrix inference as the sketch size $m$ and the number of independent components $n$ increases.

of cumulant tensors (red) are shown. Figure 3 explicitly shows that the phase transition empirically occurs within the location of $m = n(n+1)$ and $m = 2n(n+1)$ and provides us with a tight practical lower bound of $m \geq 2n(n+1)$ on the sketch size for successful inference of the mixing matrix with high probability. Recall that in Theorem 4.1, the RIP holds when $m \gtrsim 2n(n+1)$. The location of the phase transition in the empirical results therefore further consolidates the theoretical result. For a given number of independent components $n$, the ratio between the upper orange line (four times the model set dimension) and the red line (space of cumulant tensor dimension) provides a realistic compression rate in comparison to using the whole cumulant tensor of which many ICA techniques use. Importantly, as the number of independent components increases the ratio between these two lines decreases, resulting in further compression.

### 6.2 *Statistical Efficiency*

As was shown in Section 6.1, the potential compression rates of sketching the cumulant tensor are high, which can lead to a significantly reduced memory requirement. In this section, we numerically analyse the trade-off between the sketch size and the loss of information. Statistical efficiency is a measure of the variability or quality of an unbiased estimator [32]. To quantify the statistical efficiency of the compressive ICA algorithms and compare them to the well-known ICA algorithms in the literature (e.g. Comon, JADE and Fast ICA - see Section 2.3.4), we compute the root mean squared error (RMSE) with respect to the Amari distance in (6.1). By considering different sketch sizes, we can quantify the loss of efficiency (or information) incurred by different compression rates compared with the other techniques in the literature that do not compress the data. We perform our efficiency test on $n = 6$ independent components of signal length $N = 1000$. For each of the 250 Monte-Carlo simulations, the $n = 6$ independent components are randomly sampled [5] from a range of distributions with unique characteristics that are shown in Fig. 4. The true mixing matrix $\mathbf{M}_\theta$ was sampled once and fixed throughout. The smallest sketch size considered was $m = 65$ coinciding with approximately three times the model set dimension and, as it was established in Section 6.1, this sketch size achieves successful estimation with high probability. Figure 5 shows the RMSE for both the unwhitened and whitened CICA

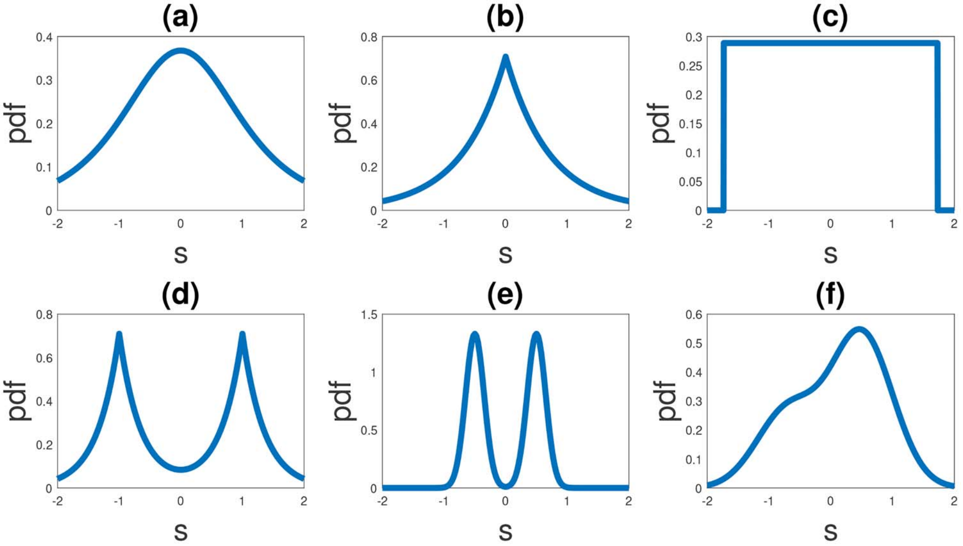

Fig. 4. (a) Student $t$ distribution ($\nu = 3$) (b) Laplace distribution ($\mu = 0, b = 1$) (c) continuous uniform distribution ($a = -\sqrt{3}, b = \sqrt{3}$) (d) mixture of two Laplaces ($\mu_1, \mu_2 = -1, 1$ $b_1 = b_2 = 1$) (e) symmetric bimodal mixture of Gaussians ($\mu_1, \mu_2 = -1, 1$ $\sigma_1 = \sigma_2 = 0.15$) (f) asymmetric unimodal mixture of Gaussians ($\mu_1, \mu_2 = -0.7, 0.5$ $\sigma_1 = \sigma_2 = 0.5$).

algorithms[8] as well as the benchmark algorithms of JADE, Comon and Fast ICA. For each sketch size, the 95% confidence interval is plotted as illustrated by the error bars. For both the whitened and unwhitened versions of the CICA algorithm, the RMSE converges quickly towards the RMSE of the full data algorithms as the sketch size $m$ increases. However, even at a sizeable compression of $m = 70$, the whitened CICA algorithm achieves an RMSE that is less than double that of the full data approaches showing that there is a controlled trade-off between compression and loss of efficiency. The unwhitened version of the CICA algorithm achieves a larger RMSE than its whitened counterpart, however this can be attributed to the whitening errors propagating throughout each iteration of the algorithm.

### 6.3 *Cylinder Velocity Field*

We next analyse and compare the proposed CICA scheme on a dataset consisting of a flow field around a cylinder obstruction as depicted in Fig. 6. Using ICA, one can obtain a model that describes the fluctuations of the streamwise velocity field around its mean value as a function of time. Details of the experimental set-up can be seen in [12, 39]. The dataset is of size $\mathbf{X} \in \mathbb{R}^{100\times 14400}$ consisting of 14400 spatial locations over 100 time intervals. Here, we compare our proposed CICA scheme with the well-known fast ICA algorithm [41], as well the JADE [18] and Comon algorithm [22] which, like the proposed CICA scheme, are cumulant based. An initial prewhitening stage inferred the prewhiten matrix $\mathbf{V} \in \mathbb{R}^{8\times 14400}$. Each algorithm then estimated the $\mathbf{Q} \in \mathbb{R}^{8\times 8}$, resulting in a mixing matrix estimate $\mathbf{M} = \mathbf{V}^{-1}\mathbf{Q}$. For the proposed CICA scheme, the IPG version was used with an SRHT matrix $\mathbf{A}$, however ASD version produces similar reconstructions. Figure 7 shows the eight independent components which describe the fluctuations of the streamwise velocity around the cylinder obtained by

[8] For the sake of simplicity, we only consider the IPG version of the CICA algorithm from Section 5.

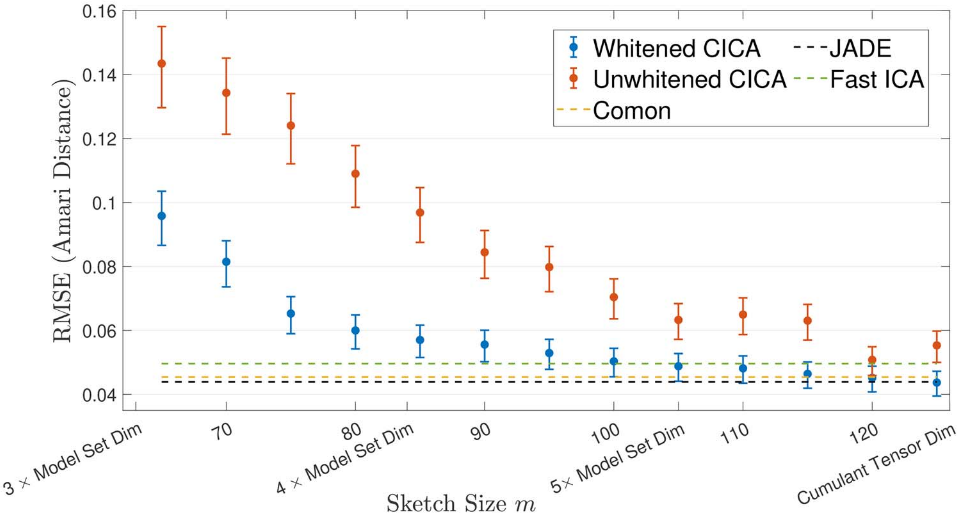

FIG. 5. The relative efficiency of the full data cumulant tensor (Comon's ICA) and sketch mixing matrix estimates for increasing sketch size *m*.

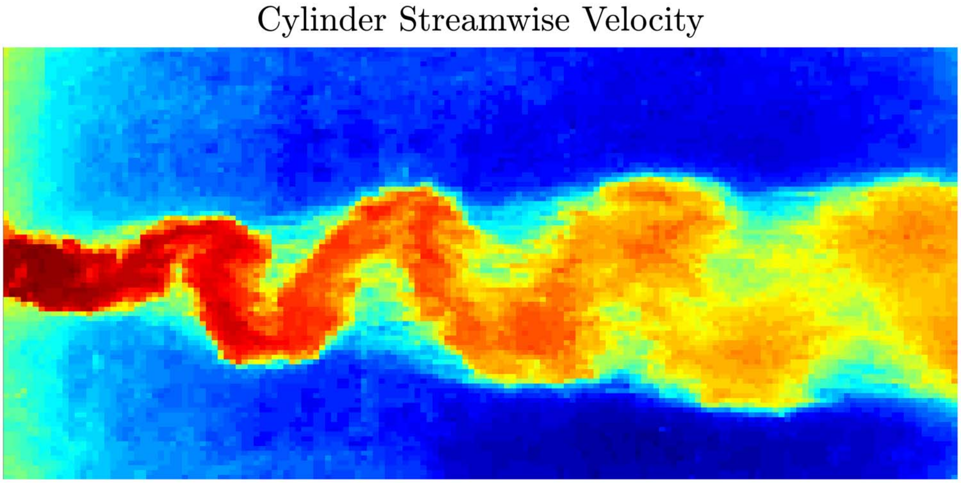

FIG. 6. The figure shows the velocity field around a cylinder for a fixed point in time.

Fast ICA, JADE, Comon and CICA, respectively. For our proposed CICA algorithm, a sketch of size $m = 114$ is used. Visually comparing the reconstructions, one can see that the CICA algorithm performs competitively with negligible artefacts present. In addition, the CICA scheme achieves a compression rate of approximately 3 in comparison to the other cumulant-based ICA methods discussed.

Next, we compare the effect of the sketch size on the resulting reconstructions. A sketch size of $m = 72, 108$ and 144 are considered with the reconstructions shown in Fig. 8. For $m = 108$, the sketch is of sufficient size to successfully identify the unique fluctuations of the velocity field; however, due

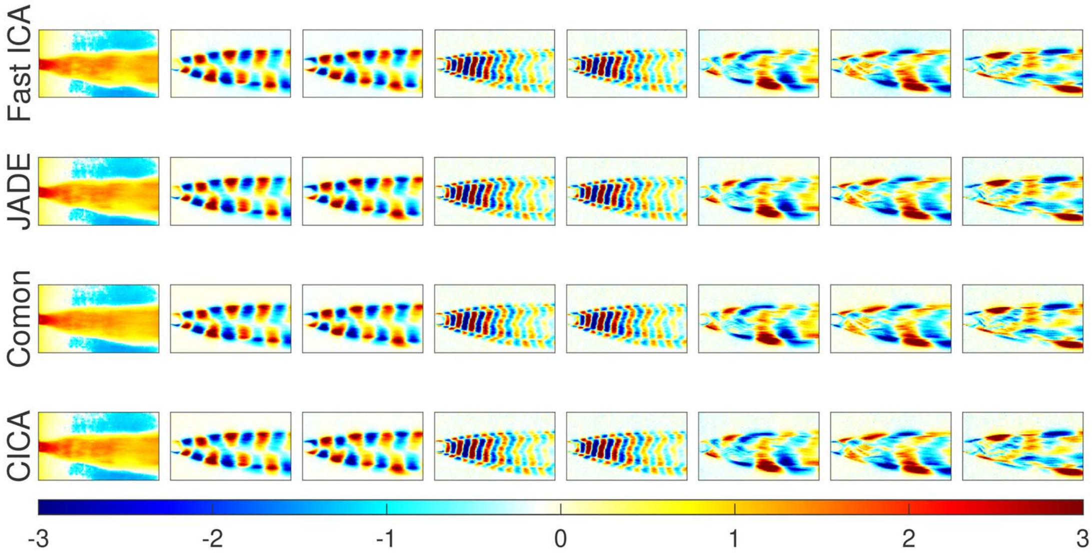

FIG. 7. From left to right the dominant fluctuations of the streamwise velocity field. From top to bottom the Fast ICA, JADE, Comon and CICA reconstructions.

to the harsher compression rate some notable artefacts are present. For example, in the first and third fluctuations there are some oscillating type artefacts which can be attributed to the higher frequencies in the system. Furthermore, the sketch of size $m = 72$ fails to identity the main fluctuations of the velocity field.

## 7. Conclusion

In this paper, we initially showed that a low-dimensional model set exists for the ICA problem. It was demonstrated theoretically that an RIP exists for the ICA model using Gaussian ensembles provided the sketch size was set proportionally to the model set dimensions, which in turn induced the existence of an instance optimal decoder. The theoretical results were empirically validated by showing the location of a sharp phase transition between a state of unsuccessful inference to a state of successful inference of the ICA mixing matrix as the sketch size increased. Using both synthetic and real data, we analysed the robustness of the proposed CICA algorithms and highlighted the effect of choosing the sketch size $m$. Furthermore, the particular branch of compressive learning was discussed that consists of sketching distribution free models (e.g. PCA, ICA) that leverage some intermediary statistic space, here the space of cumulant tensors, to form the sketch. This poses some interesting open questions on how to design a sketch given other distribution free models and how the low-dimension nature of the model set manifests itself structurally, in terms of sparsity, low rank, etc. to construct a practical sketching decoder. It can often be challenging to find a statistic of the data that permit identifiability of the model parameters. In addition, the identifiable statistic associated with the learning model might scale exponentially with the dimensions of the underlying model, resulting in a intractable compression scheme.

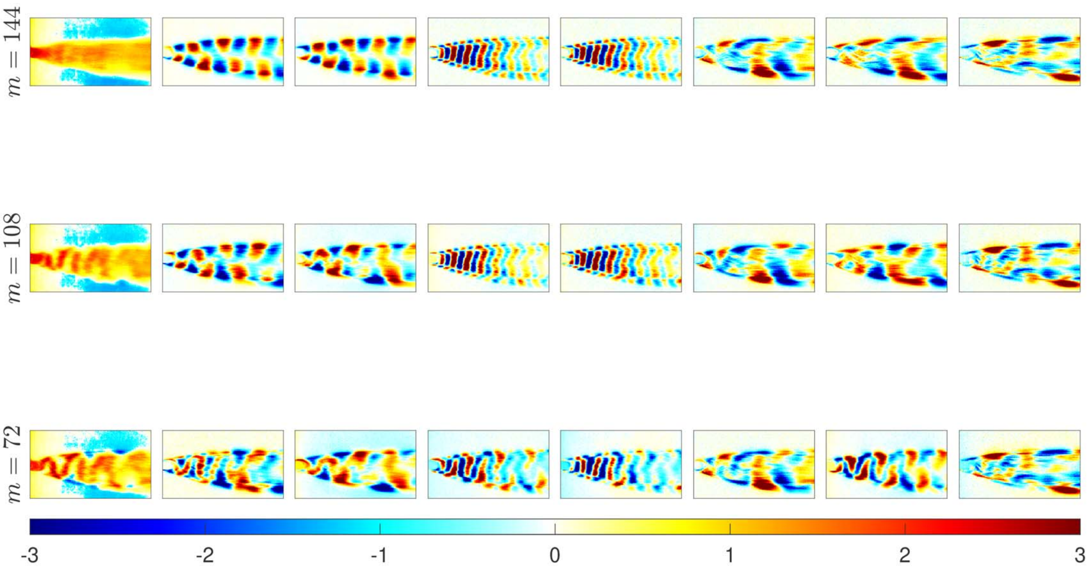

FIG. 8. The figure shows the effect of the sketch size on the reconstruction of the fluctuations. From top to bottom a sketch size of $m = 144, 108$ and 72.

## Data Availability Statement

The data used in Section 6.3 is available at the repository https://github.com/jonnyhigham/POD_DMD.

## Code Availability

A MATLAB implementation of the proposed CICA algorithms are available at the repository https://gitlab.com/mpsheehan1995/CICA.

## Acknowledgements

This work was supported by the ERC Advanced grant, project C-SENSE (ERC-ADG-2015-694888). Mike E. Davies is also supported by a Royal Society Wolfson Research Merit Award.

## REFERENCES

1. AILON, N. & CHAZELLE, B. (2006) Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of Computing (STOC'06)*, New York, NY, USA: Association for Computing Machinery, pp. 557–563. https://doi.org/10.1145/1132516.1132597.
2. AMARI, S., CICHOCKI, A. & YANG, H. (1996) A new learning algorithm for blind signal separation. *Advances in neural information processing systems 8*, David S. Touretzky and Michael Mozer and Michael E. Hasselmo eds. NIPS, Denver, CO, USA: MIT Press, pp. 757–763.
3. AMELUNXEN, D., LOTZ, M., MCCOY, M. B. & TROPP, J. A. (2014) Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, **3**, 224–294.
4. ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. & TELGARSKY, M. (2014) Tensor decompositions for learning latent variable models. *Journal of machine learning research*, **15**, 2773–2832.

5. Bach, F. & Jordan, M. (2002) Kernel independent component analysis. *Journal of machine learning research*, **3**, 1–48.
6. Bell, A. J. & Sejnowski, T. J. (1995) An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.*, **7**, 1129–1159.
7. Blumensath, T. (2011) Sampling and Reconstructing Signals From a Union of Linear Subspaces. *IEEE Trans. Information Theory*, **57**, 4660–4671.
8. Blumensath, T. & Davies, M. E. (2008) Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, **14**, 629–654.
9. Blumensath, T. & Davies, M. E. (2009) Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, **27**, 265–274.
10. Bourrier, A., Davies, M. E., Peleg, T., Pérez, P. & Gribonval, R. (2014) Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, **60**, 7928–7946.
11. Boutsidis, C., Zouzias, A. & Drineas, P. (2010) Random projections for $k$-means clustering. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, John D. Lafferty and Christopher K. I. Williams and John Shawe-Taylor and Richard S. Zemel and Aron Culotta eds. Vancouver, British Columbia, Canada: Curran Associates, Inc., pp. 298–306.
12. Brevis, W. & García-Villalba, M. (2011) Shallow-flow visualization analysis by proper orthogonal decomposition. *Journal of Hydraulic Research*, **49**, 586–594.
13. Cadzow, J. (1988) Signal enhancement-a composite property mapping algorithm. *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 49–62.
14. Candes, E. J. & Plan, Y. (2011) Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, **57**, 2342–2359.
15. Candès, E. J. & Wakin, M. B. (2008) An introduction to compressive sampling. *IEEE signal processing magazine*, **25**, 21–30.
16. Cardoso, J. F. (1992) Fourth-order cumulant structure forcing: application to blind array processing. *[1992] IEEE Sixth SP Workshop on Statistical Signal and Array Processing*, pp. 136–139, Victoria, BC, Canada.
17. Cardoso, J. F. (1995) 37 - A Tetradic Decomposition of 4th-Order Tensors: Application to the Source Separation Problem. *SVD and Signal Processing III*, Marc Moonen and Bart De Moor eds. Amsterdam: Elsevier Science B.V., pp. 375–382.
18. Cardoso, J. F. & Souloumiac, A. (1993) Blind beamforming for non-Gaussian signals. *IEE proceedings F (radar and signal processing)*, United Kingdom, vol. **140**. IET, pp. 362–370.
19. Clarkson, K. L. (2008) Tighter bounds for random projections of manifolds. *Proceedings of the 24th ACM Symposium on Computational Geometry*, Monique Teillaud ed. College Park, MD, USA: ACM, pp. 39–48.
20. Cohen, A., Dahmen, W. & DeVore, R. (2009) Compressed sensing and best $k$-term approximation. *J. Amer. Math. Soc.*, **22**, 211–231.
21. Coleman, R. (2012) *Calculus on Normed Vector Spaces*. New York: Springer.
22. Comon, P. (1994a) Independent component analysis, a new concept? *Signal processing*, **36**, 287–314.
23. Comon, P. (1994b) Tensor Diagonalization, A Useful Tool in Signal Processing. *IFAC Proceedings Volumes*, **27**, 77–82 IFAC Symposium on System Identification (SYSID'94), Copenhagen, Denmark, 4-6 July.
24. Comon, P. (2009) Tensor decompositions, state of the art and applications. J. G. McWhirter and I. K. Proudler. Mathematics in Signal Processing V, Clarendon Press, Oxford, pp. 1–24, 2002. ffhal00347139.
25. Cormode, G., Garofalakis, M., Haas, P. J. & Jermaine, C. (2012) Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, **4**, 1–294.
26. Cormodea, G. & Muthukrishnanb, S. (2005) An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, **55**, 58–75.
27. Crank, J. & Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, Crank, J. and Nicolson, P. eds. vol. **43**. Cambridge University Press, UK, pp. 50–67.

28. De Lathauwer, L. (1997) *Signal processing based on multilinear algebra*. Katholieke Universiteit Leuven, Leuven, Belgium.
29. De Lathauwer, L., De Moor, B. & Vandewalle, J. (2000) An introduction to independent component analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **14**, 123–149.
30. Donoho, D. L. (2006) Compressed sensing. *IEEE Transactions on Information Theory*, **52**, 1289–1306.
31. Feldman, D., Schmidt, M. & Sohler, C. (2020) Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM J. Comput.*, **49**, 601–657.
32. Fisher, R. A. (1922) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **222**, 309–368.
33. Gribonval, R., Blanchard, G., Keriven, N. & Traonmilin, Y. (2021) Compressive statistical learning with random feature moments. *Mathematical Statistics and Learning*, **3**, 113–164.
34. Gribonval, R., Chatalic, A., Keriven, N., Schellekens, V., Jacques, L. & Schniter, P. (2021) Sketching Data Sets for Large-Scale Learning: Keeping only what you need. *IEEE Signal Processing Magazine*, **38**, 12–36.
35. Guha, S., Meyerson, A., Mishra, N., Motwani, R. & O'Callaghan, L. (2003) Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, **15**, 515–528.
36. Hall, A. R. (2003) Generalized Method of Moments (Advanced Texts in Econometrics Series, Oxford University Press). *A Companion to Theoretical Econometrics*, Hall, A R. ed. United Kingdom: Oxford University Press, 230–255.
37. Hansen, L. P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**, 1029–1054.
38. Har-Peled, S. & Mazumdar, S. (2004) On coresets for k-means and k-median clustering. *Proceedings of the 36th Annual {ACM} Symposium on Theory of Computing*, László Babai ed. Chicago, IL, USA: ACM, pp. 291–300.
39. Higham, J., Brevis, W. & Keylock, C. (2018) Implications of the selection of a particular modal decomposition technique for the analysis of shallow flows. *Journal of Hydraulic Research*, **56**, 796–805.
40. Hyvarinen, A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, **10**, 626–634.
41. Hyvarinen, A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, **10**, 626–634.
42. Hyvärinen, A. & Oja, E. (2000) Independent component analysis: algorithms and applications. *Neural networks*, **13**, 411–430.
43. Jolliffe, I. T. & Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**, 20150202.
44. Keriven, N., Bourrier, A., Gribonval, R. & Pérez, P. (2018) Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, **7**, 447–508.
45. Kolda, T. G. & Bader, B. W. (2009) Tensor Decompositions and Applications. *SIAM Rev.*, **51**, 455–500.
46. Krahmer, F. & Ward, R. (2011) New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, **43**, 1269–1281.
47. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017) Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends®. Machine Learning*, **10**, 1–141.
48. Oja, E., Kiviluoto, K. & Malaroiu, S. (2000) Independent component analysis for financial time series. *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 111–116.
49. Puy, G., Davies, M. E. & Gribonval, R. (2017) Recipes for Stable Linear Embeddings From Hilbert Spaces to $\mathrm{R}^{\mathrm{m}}$. *IEEE Transactions on Information Theory*, **63**, 2171–2187.
50. Rahimi, A. & Recht, B. (2008) Random features for large-scale kernel machines. *Advances in neural information processing systems*, 1177–1184.

51. Rahimi, A. & Recht, B. (2009) Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 1313–1320.

52. Rauhut, H., Schneider, R. & Stojanac, Ž. (2017) Low rank tensor recovery via iterative hard thresholding. *Linear Algebra Appl.*, **523**, 220–262.

53. Sela, M. & Kimmel, R. (2016) Randomized independent component analysis. *Science of Electrical Engineering (ICSEE), IEEE International Conference on the Science of Electrical Engineering*.IEEE, pp. 1–5.

54. Sheehan, M. P., Gonon, A. & Davies, M. E. (2019) Compressive Learning for Semi-Parametric Models. arXiv preprint arXiv:1910.10024.

55. Sheehan, M. P., Kotzagiannidis, M. S. & Davies, M. E. (2019) Compressive Independent Component Analysis. *2019 27th European Signal Processing Conference (EUSIPCO)*.IEEE, pp. 1–5.

56. Szarek, S. (1981) Nets of Grassmann manifold and orthogonal group. *Proceedings of Research Workshop on Banach Space Theory*, pp. 169–186.

57. Tang, J., Alelyani, S. & Liu, H. (2014) Feature selection for classification: A review. *Data classification: Algorithms and applications*, p. 37.

58. Tropp, J. A. (2011) Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, **3**, 115–126.

59. Tropp, J. A., Yurtsever, A., Udell, M. & Cevher, V. (2017) Practical sketching algorithms for low-rank matrix approximation. *SIAM J. Matrix Anal. Appl.*, **38**, 1454–1485.

60. Tropp, J. A., Yurtsever, A., Udell, M. & Cevher, V. (2019) Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM J. Sci. Comput.*, **41**, A2430–A2463.

61. Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*.Cambridge University Press, pp. 210–268.

62. Vigario, R., Sarela, J., Jousmiki, V., Hamalainen, M. & Oja, E. (2000) Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, **47**, 589–593.

63. Wen, Z. & Yin, W. (2013) A Feasible Method for Optimization with Orthogonality Constraints. *Math. Programming*, **142**, 397–434.

64. Woolfe, F., Liberty, E., Rokhlin, V. & Tygert, M. (2008) A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.*, **25**, 335–366.

65. Wu, Y. & Yang, P. (2020) Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.*, **48**, 1981–2007.

## A. Proof of Lemma 4.1

To prove Lemma 4.1, we use a similar line of argument to Clarkson in [19] by splitting the normalized secant set into the set of short and long secants parametrized by a distance $\eta$. First, we state an important lemma on covering the model set intersected with the unit sphere in $\mathbb{R}^{\bar{n}}$, where $\bar{n} = n^4$, denoted by $\bar{\bar{\mathfrak{S}}}_{\mathscr{H}} := \mathfrak{S}_{\mathscr{H}} \cap \mathbb{S}^{\bar{n}-1}$ (e.g. $\|\mathscr{X}\|_F = 1$), which will used later in the proof.

LEMMA A.1. (Covering number of $\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$) The covering number of $\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ with respect to the Frobenius norm $\|\cdot\|_F$ is

$$\mathrm{CN}\left(\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}, \|\cdot\|_F, \epsilon\right) \leq \left(\frac{6}{\epsilon}\right)^{n(n+1)} \tag{A.1}$$

*Proof.* Recall that $\mathscr{X} \in \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ has the decomposition $\mathscr{X} = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}$ such that $\|\mathscr{X}\|_F = 1$, where $\mathscr{S} \in \mathfrak{D}$ and $\mathbf{Q} \in \mathrm{O}(n)$. As the Frobenius norm is rotationally invariant [9] the following holds: $\|\mathscr{X}\|_F = \|\mathscr{S}\|_F = 1$ for all $\mathscr{X} \in \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$. Our argument constructs an $\epsilon$-net for $\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ by covering the sets $\mathfrak{D}$ and $\mathrm{O}(n)$, respectively. As $\|\mathscr{X}\|_F = 1 \implies \|\mathscr{S}\|_F = 1$, it is sufficient to consider $\bar{\bar{\mathfrak{D}}} := \mathfrak{D} \cap \mathbb{S}^{n-1}$. Then we take $\underline{\bar{\bar{\mathfrak{D}}}}$ to be an $\epsilon/2$- net for $\bar{\bar{\mathfrak{D}}}$. As $\bar{\bar{\mathfrak{D}}}$ is an $n$-dimensional subspace, then

$$\mathrm{CN}\left(\bar{\bar{\mathfrak{D}}}, \|\cdot\|_F, \epsilon/2\right) \leq \left(\frac{6}{\epsilon}\right)^n.$$

Next, we cover the set of $n \times n$ orthogonal matrices denoted $\mathrm{O}(n)$. We follow a similar argument to [14, 52] by letting $\mathrm{Q}(n) := \{\mathbf{X} \in \mathbb{R}^{n\times n} : \|\mathbf{X}\|_{1,2} \leq 1\}$, where

$$\|\mathbf{X}\|_{1,2} = \max_i \|X(:, i)\|_2$$

is the maximum column norm of a matrix $\mathbf{X}$. It is straightforward to see that $\mathrm{O}(n) \subset \mathrm{Q}(n)$ since the columns of an orthogonal matrix are unit normed. It can be seen in [14] that an $\epsilon/2$-net $\mathrm{O}(n)$, denoted by $\underline{\mathrm{O}}(n)$, has a covering number

$$\mathrm{CN}(\mathrm{O}(n), \|\cdot\|_{1,2}, \epsilon/2) \leq \left(\frac{6}{\epsilon}\right)^{n^2}.$$

Now let $\underline{\bar{\bar{\mathfrak{S}}}}_{\mathscr{H}} := \{\underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} : \underline{\mathscr{S}} \in \underline{\bar{\bar{\mathfrak{D}}}}, \underline{\mathbf{Q}} \in \underline{\mathrm{O}}(n)\}$, and remark that

$$\begin{aligned}\mathrm{CN}\big(\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}, \|\cdot\|_F, \epsilon\big) &\leq \mathrm{CN}\big(\bar{\bar{\mathfrak{D}}}, \|\cdot\|_F, \epsilon/2\big) \;\; \mathrm{CN}\big(\mathrm{O}(n), \|\cdot\|_{1,2}, \epsilon/2\big)\\ &\leq \left(\frac{6}{\epsilon}\right)^{n(n+1)}.\end{aligned}$$

It remains to show that for all $\mathscr{X} \in \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ there exists $\underline{\mathscr{X}} \in \underline{\bar{\bar{\mathfrak{S}}}}_{\mathscr{H}}$ such that $\|\mathscr{X} - \underline{\mathscr{X}}\|_F \leq \epsilon$.

Fix $\mathscr{X} \in \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ and note the decomposition $\mathscr{X} = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}$. Then there exists $\underline{\mathscr{X}} = \underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} \in \underline{\bar{\bar{\mathfrak{S}}}}_{\mathscr{H}}$ with $\underline{\mathscr{S}} \in \underline{\bar{\bar{\mathfrak{D}}}}$ and $\underline{\mathbf{Q}} \in \underline{O}(n)$ obeying $\|\mathscr{S} - \underline{\mathscr{S}}\|_F \leq \epsilon/2$ and

$\|\mathbf{Q}-\underline{\mathbf{Q}}\|_{1,2} \leq \epsilon/2$. This gives

$$\begin{aligned}
\|\mathscr{X} - \underline{\mathscr{X}}\|_F &= \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q} - \underline{\mathscr{L}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\
&= \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q} + (\mathscr{S} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} - \mathscr{S} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}) \\
&\qquad\qquad - \underline{\mathscr{L}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\
&= \|\mathscr{S} \times_1 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_2 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_3 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_4 (\mathbf{Q}-\underline{\mathbf{Q}}) + (\mathscr{S} - \underline{\mathscr{D}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\
&\leq \|\mathscr{S} \times_1 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_2 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_3 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_4 (\mathbf{Q}-\underline{\mathbf{Q}})\|_F + \|(\mathscr{S} - \underline{\mathscr{D}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F
\end{aligned}$$

The first part of the last line gives

$$\begin{aligned}
\|\mathscr{S} \times_1 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_2 \cdots \times_4 (\mathbf{Q}-\underline{\mathbf{Q}})\|_F &= \|\mathrm{vec}\left(\mathscr{S} \times_1 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_2 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_3 (\mathbf{Q}-\underline{\mathbf{Q}}) \times_4 (\mathbf{Q}-\underline{\mathbf{Q}})\right)\|_2 \\
&= \|(\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}})\mathrm{vec}(\mathscr{S})\|_2 \\
&\leq \|(\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}}) \otimes (\mathbf{Q}-\underline{\mathbf{Q}})\|_2 \|\mathscr{S}\|_F \\
&= \|(\mathbf{Q}-\underline{\mathbf{Q}})\|_2^4 \\
&\leq \|(\mathbf{Q}-\underline{\mathbf{Q}})\|_{1,2}^4 \\
&\leq (\epsilon/2)^4 \\
&\leq \epsilon/2
\end{aligned}$$

From line 1 to 2, the identity on pages [477–478] of [45] was used. From line 2 to 3 we have used the Cauchy–Schwarz inequality, from line 3 to 4 we have used the equality $\|\mathbf{A} \otimes \mathbf{B}\| = \|\mathbf{A}\|\|\mathbf{B}\|$ and from line 4 to 5 we have used the identity in [52]. Finally, notice that as $\mathbf{Q}$ is orthogonal

$$\|(\mathscr{S} - \underline{\mathscr{D}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F = \|(\mathscr{S} - \underline{\mathscr{D}})\|_F = \epsilon/2.$$

Therefore,

$$\|\mathscr{X} - \underline{\mathscr{X}}\|_F \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

□

Continuing, we let $\Omega := O(n) \times \mathfrak{D}$ define the product set between the set of $n \times n$ orthogonal matrices $O(n)$ and the set of super symmetric cumulant tensors defined in (2.15) and define the map $f : \Omega \mapsto \mathfrak{S}_{\mathscr{H}}$by

$$f(u) = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}, \tag{A.2}$$

for all $u := (\mathbf{Q}, \mathscr{S}) \in \Omega$. Let $\mathscr{X} = f(u)$ be the tensor corresponding to the image of the map $f$. It is insightful to decompose the normalized secant set $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right)$ into the set of long and short secants parametrized by some distance $\eta$ [19]. The set of long secants of $\mathfrak{S}_{\mathscr{H}}$is defined as

$$\mathfrak{N}_\eta\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right) := \left\{ \frac{\mathscr{X}_1 - \mathscr{X}_2}{\|\mathscr{X}_1 - \mathscr{X}_2\|_F} \;\middle|\; \mathscr{X}_1, \mathscr{X}_2 \in \mathfrak{S}_{\mathscr{H}},\ \|\mathscr{X}_1 - \mathscr{X}_2\|_F > \eta \right\}. \tag{A.3}$$

Furthermore, the set of short secants $\mathfrak{N}_\eta^c\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right)=\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right)\setminus\mathfrak{N}_\eta\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right)$ is the complement to the set of long secants defined by

$$\mathfrak{N}_\eta^c\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right):=\left\{\frac{\mathscr{Z}_1-\mathscr{Z}_2}{\left\|\mathscr{Z}_1-\mathscr{Z}_2\right\|_F}\;\middle|\;\mathscr{Z}_1\neq\mathscr{Z}_2\in\mathfrak{S}_{\mathscr{H}},\ \left\|\mathscr{Z}_1-\mathscr{Z}_2\right\|_F\leq\eta\right\}. \tag{A.4}$$

Remark A.1. As the model set $\mathfrak{S}_{\mathscr{H}}$ is conic, it is sufficient to cover the normalized secant set of $\bar{\mathfrak{S}}_{\mathscr{H}}:=\mathfrak{S}_{\mathscr{H}}\cap\mathfrak{B}_1(0)$, where $\mathfrak{B}_1(0)$ denotes the unit Frobenius ball centred at 0, since we have $\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right)=\mathfrak{N}\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)$.

As the model set is conic, we can decompose the normalized secant set as follows:

$$\begin{aligned}\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}}-\mathfrak{S}_{\mathscr{H}}\right)&=\mathfrak{N}\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)\\&=\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)\cup\mathfrak{N}_\eta^c\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)\\&=\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)\cup\mathfrak{N}_\eta^c\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}\right),\end{aligned} \tag{A.5}$$

We begin by covering the set of long secants $\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)$.

Lemma A.2. (Long Secants Covering Number) Let $\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}$ be an $\epsilon\gamma$-cover for $\bar{\mathfrak{S}}_{\mathscr{H}}$. Then $\mathfrak{N}\left(\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}-\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}\right)$ is an $\epsilon$-cover for $\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)$ with associated covering number of

$$\mathrm{CN}\left(\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right),\epsilon\gamma\right)\leq\left(\frac{6}{\epsilon\gamma}\right)^{2n(n+1)}. \tag{A.6}$$

*Proof.* Lemma 4.1 in [19] states that if $\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}$ is a generalized $\epsilon\gamma$- cover of $\bar{\mathfrak{S}}_{\mathscr{H}}$, then $\mathfrak{N}\left(\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}-\underline{\bar{\mathfrak{S}}}_{\mathscr{H}}\right)$ is a generalized $\epsilon$-cover for $\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_{\mathscr{H}}-\bar{\mathfrak{S}}_{\mathscr{H}}\right)$. Using the covering number of $\bar{\bar{\mathfrak{S}}}_{\mathscr{H}}$ from Lemma A.1 we get the result. □

Continuing, we cover the set of short secants. We begin by stating some preliminary lemmas.

Lemma A.3. (Taylor Approximation Error) Let $f\,:\,\varOmega\mapsto\mathfrak{S}_{\mathscr{H}}$ be defined as in (A.2) and let $Df_u$ define the first-order differential of $f$ evaluated at the point $u$. Further assume that $\|\mathscr{S}\|_F\leq R$. Then $\forall u,u'\in\varOmega$, $\|u-u'\|\leq2\epsilon_0$, we have

$$\left\|f(u)-f(u')-Df_{u'}^T(u-u')\right\|_F\leq C_1\left\|u-u'\right\|_2^2, \tag{A.7}$$

where $C_1=n^2(n+1)^2\max\{3R,1\}$

*Proof.* w.l.o.g consider the vectorized function $\tilde{f}(u):=\mathrm{vec}(f(u))$ such that

$$\begin{aligned}\tilde{f}(u)=&\mathrm{vec}\left(\mathscr{S}\times_1\mathbf{Q}\times_2\mathbf{Q}\times_3\mathbf{Q}\times_4\mathbf{Q}\right)\\=&\mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{Q}\ \mathrm{vec}\left(\mathscr{S}\right).\end{aligned}$$

Using Taylor's theorem [21, p. 110] of $\tilde{f}$ evaluated at the point $u'\in\varOmega$, we get

$$\left\|\tilde{f}(u)-\tilde{f}(u')-D\tilde{f}_{u'}^T(u-u')\right\|_2\leq\frac{1}{2}\left\|(u-u')^TH\tilde{f}_\xi(u-u')\right\|_2. \tag{A.8}$$

where $D\tilde{f}_u$ and $H\tilde{f}_u$ denote the Jacobian and Hessian of $\tilde{f}$ evaluated at $u$ and $\xi=\lambda u+(1-\lambda)u'\in\varOmega$, for $\lambda\in(0,1)$, denotes a point on the line segment between $u$ and $u'$. For shorthand let $h=u-u'$, and denote the integer $T := \frac{n(n+1)}{2}$, we then have

$$
\begin{aligned}
\left\| h^T H\tilde{f}_{\xi} h \right\|_2 &= \left\| \sum_{i=1}^{T} \sum_{j=1}^{T} h_i h_j \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \max_{i,j} \left\| h_i h_j \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \left( \max_i |h_i| \right)^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&= T^2 \|h\|_{\infty}^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \|h\|_2^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 ,
\end{aligned}
$$

where $h_i = (u_i - u_i')$. w.l.o.g let $\xi = (\mathbf{Q}, \mathscr{S})$, we have that

$$
\max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 = \max \left\{ \overset{1}{\max_{i,j,k,\ell} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{kl}}(\xi) \right\|_2} , \overset{2}{\max_{i,j,k} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathscr{S}_{kkkk}}(\xi) \right\|_2} , \overset{3}{\max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii} \partial \mathscr{S}_{jjjj}}(\xi) \right\|_2} \right\}
$$

1 It can be shown that

$$
\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{k\ell}}(\xi) = \Pi_{ijk\ell} \operatorname{vec}(\mathscr{S}), \tag{A.9}
$$

where

$$
\begin{aligned}
\Pi_{ijk\ell} = \; & \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \; + \; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \; + \; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \; + \; \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \; + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij}
\end{aligned}
$$

and the matrix $\mathbf{E}^{ij} = \mathbf{e}_i\mathbf{e}_j^T$, where $\mathbf{e}_i$ is the $i$th unit basis vector. Using the properties of the Kronecker product and the triangle inequality we get

$$\begin{aligned}\left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{kl}}(\xi) \right\|_2 &\leq 12 \left\| \mathbf{E}^{ij} \right\|_2 \left\| \mathbf{E}^{kl} \right\|_2 \|\mathbf{Q}\|_2^2 \|\mathscr{S}\|_F \\ &= 12 \|\mathscr{S}\|_F .\end{aligned}$$

Assuming that the diagonal tensor has bounded support $\|\mathscr{S}\|_2 \leq R$, then it follows that

$$\max_{i,j,k,\ell} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{kl}}(\xi) \right\|_2 \leq 12R. \tag{A.10}$$

2 It can be shown that

$$\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathscr{S}_{kkkk}}(\xi) = \Gamma_{ij}\mathbf{e}_k, \tag{A.11}$$

where and

$$\begin{aligned}\Gamma_{ij} = \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \\ + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij}.\end{aligned}$$

Similarly to [1], we get

$$\max_{i,j,k} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathscr{S}_{kkkk}}(\xi) \right\|_2 \leq 4$$

3 It can be easily shown that

$$\frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii} \partial \mathscr{S}_{jjjj}}(\xi) = \mathbf{0}, \tag{A.12}$$

therefore

$$\max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii} \partial \mathscr{S}_{jjjj}}(\xi) \right\|_2 = 0. \tag{A.13}$$

It therefore follows that

$$\max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 = \max\{12R, 4\}, \tag{A.14}$$

and

$$\left\| \tilde{f}(u) - \tilde{f}(u') - D\tilde{f}_{u'}^T(u - u') \right\|_2 \leq n^2(n+1)^2 \max\{3R, 1\} \left\| u - u' \right\|_2^2. \tag{A.15}$$

□

Lemma A.4. (Bounded Curvature) Let $f : \Omega \mapsto \mathfrak{S}_{\mathscr{H}}$ be defined as in (A.2) and let $Df_u$ define the first-order differential of $f$ evaluated at the point $u$. Further assume that $\|\mathscr{S}\|_F \leq R$. Then $\forall u, u' \in \Omega$,

$\|u - u'\| \le 2\epsilon_0$, we have

$$\left\| Df_u - Df_{u'} \right\|_F \le C_2 \left\| u - u' \right\|_2 , \tag{A.16}$$

where $C_2 = 2C_1$.

*Proof.* Using the mean value theorem [21], it can be shown that

$$\left\| D\tilde{f}_u \; - \; D\tilde{f}_{u'} \right\|_2 \le \left\| H\tilde{f}_\xi^T (u - u') \right\|_2 \tag{A.17}$$

for some $\xi = \lambda u + (1 - \lambda)u' \in \Omega$, for $\lambda \in (0, 1)$. Then using the same argument as in the proof of Lemma A.3, it can easily shown that

$$\left\| D\tilde{f}_u \; - \; D\tilde{f}_{u'} \right\|_2 \le 2C_1 \left\| u - u' \right\|_2 , \tag{A.18}$$

giving $C_2 = 2C_1$. □

LEMMA A.5. (Bounded Gradient) Let $f : \Omega \mapsto \mathfrak{S}_{\mathscr{H}}$ be defined as in (A.2) and let $Df_u$ define the first-order differential of $f$ evaluated at the point $u$. Further assume, as in (2.15), that $\mathscr{S}_{iiii} \ge \epsilon_{\mathscr{S}} (> 0) \forall i$. Then $\forall u \in \Omega$

$$\left\| Df_u^{\dagger} \right\|_F \le C_3, \tag{A.19}$$

where $C_3 = 2\epsilon_{\mathscr{S}}$.

*Proof.* As in A.3, we consider the vectorized function $\tilde{f}(u) := \mathrm{vec}\,(f(u))$ w.l.o.g. It can be shown that the first-order differential has the following decomposition:

$$D\tilde{f}(u) = \left[ \frac{\partial \tilde{f}}{\partial \mathbf{Q}}(u), \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u) \right], \tag{A.20}$$

where

$$\frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u) \;=\; \Gamma_{ij} \mathrm{vec}\,(\mathscr{S}) \,. \tag{A.21}$$

Furthermore, the partial derivative with respect to the super symmetric cumulant tensor $\mathscr{S}$ is defined as

$$\frac{\partial \tilde{f}}{\partial \mathscr{S}}(u) \;=\; \mathbf{B},$$

where $\mathbf{B} := \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q}$. Equivalently, (A.19) can be rewritten as

$$\min_{\|\Delta u\|=1} \left\| D\tilde{f}\,(u)^T \, \Delta u \right\|_2 \ge C_3, \tag{A.22}$$

where $\Delta u = (\Delta\mathbf{Q}, \Delta\mathscr{S})$. We therefore have

$$\begin{aligned}\left\| D\tilde{f}(u)^T \Delta u \right\|_2^2 &= \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}}(u)^T \Delta\mathbf{Q} \right\|_F^2 + \left\| \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u)^T \Delta\mathscr{S} \right\|_2^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u)^T \Delta\mathbf{Q}_{ij} \right\|_F^2 + \left\| \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u)^T \Delta\mathscr{S} \right\|_2^2 \\ &= (\star).\end{aligned}$$

As $f$ is equivariant in $\mathbf{Q}$, we can set $\mathbf{Q} = \mathbf{I}_n$ w.l.o.g. As a result $\mathbf{B} = \mathbf{I}$ and $\Gamma_{ij}$ reduces to

$$\begin{aligned}\Gamma_{ij} \;=\; & \mathbf{E}^{ij} \otimes \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n \;+\; \mathbf{I}_n \otimes \mathbf{E}^{ij} \otimes \mathbf{I}_n \otimes \mathbf{I}_n \\ & + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{E}^{ij} \otimes \mathbf{I}_n \;+\; \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{E}^{ij}.\end{aligned}$$

For shorthand, let $\mathscr{T} = \Gamma_{ab}\mathrm{vec}\,(\mathscr{S})$ and noting that $\mathbf{E}^{ab} = \mathbf{e}_a\mathbf{e}_b^T$, we have

$$\begin{aligned}\mathscr{T}_{ijk\ell} &= \sum_{p=1}^{n} \left( \mathbf{E}^{ab}_{ip}\mathbf{I}_{jp}\mathbf{I}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{E}^{ab}_{jp}\mathbf{I}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{I}_{jp}\mathbf{E}^{ab}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{I}_{jp}\mathbf{I}_{kp}\mathbf{E}^{ab}_{\ell p} \right) \mathscr{S}_{pppp} \\ &= \sum_{p=1}^{n} \left( \delta_{ai}\delta_{bp}\delta_{jp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{aj}\delta_{bp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{ak}\delta_{bp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{kp}\delta_{a\ell}\delta_{bp} \right) \mathscr{S}_{pppp} \\ &= \sum_{p=1}^{n} \left( \delta_{ai}\delta_{jp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{aj}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{ak}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{kp}\delta_{a\ell} \right) \delta_{bp}\mathscr{S}_{pppp} \\ &= \left( \delta_{ai}\delta_{jb}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{aj}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{ak}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{kb}\delta_{a\ell} \right) \mathscr{S}_{bbbb}.\end{aligned}$$

As a result, we have that

$$\left\| \Gamma^{ab}\mathrm{vec}\,(\mathscr{S})\,\Delta\mathbf{Q}_{ab} \right\|_F^2 = \sum_{i,j,k,\ell=1}^{n} \left| \left( \delta_{ai}\delta_{jb}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{aj}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{ak}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{kb}\delta_{a\ell} \right) \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{ab} \right|^2.$$

It can be easily shown that for $a = b$

$$\left\| \Gamma^{bb}\mathrm{vec}\,(\mathscr{S})\,\Delta\mathbf{Q}_{bb} \right\|_F^2 = 16\left| \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{bb} \right|^2,$$

and for $a \neq b$

$$\left\| \Gamma^{ab}\mathrm{vec}\,(\mathscr{S})\,\Delta\mathbf{Q}_{ab} \right\|_F^2 = 4\left| \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{ab} \right|^2.$$

We therefore have

$$
\begin{aligned}
(\star) &= \sum_{i=j} \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ii}}(u)^T \Delta \mathbf{Q}_{ii} \right\|_F^2 + \sum_{i\neq j} \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u)^T \Delta \mathbf{Q}_{ij} \right\|_F^2 + \|\Delta \mathscr{S}\|_2^2 \\
&= 16 \sum_{i=j} \left|\mathscr{S}_{iiii}\right|^2 \left|\Delta \mathbf{Q}_{ii}\right|^2 + 4 \sum_{i\neq j} \left|\mathscr{S}_{iiii}\right|^2 \left|\Delta \mathbf{Q}_{ij}\right|^2 + \|\Delta \mathscr{S}\|_2^2 \\
&\geq 4 \sum_{i,j} \left|\mathscr{S}_{iiii}\right|^2 \left|\Delta \mathbf{Q}_{ij}\right|^2 + \|\Delta \mathscr{S}\|_2^2 \\
&= (\star)\,.
\end{aligned}
$$

Now assume that $\left|\mathscr{S}_{iiii}\right| \geq \epsilon_{\mathscr{S}}$ for all $i$, therefore

$$
\begin{aligned}
(\star) &\geq 4\epsilon_{\mathscr{S}}^2 \|\Delta \mathbf{Q}\|_F^2 + \|\Delta \mathscr{S}\|_2^2 \\
&\geq 4\epsilon_{\mathscr{S}}^2 \|\Delta u\|_2^2\,.
\end{aligned}
$$

In the last line, we assume w.l.o.g that $4\epsilon_{\mathscr{S}}^2 \leq 1$. We have therefore proved that

$$
\min_{\|\Delta u\|=1} \left\| D\tilde{f}\,(u)^T\, \Delta u \right\|_2 \geq 2\epsilon_{\mathscr{S}}, \tag{A.23}
$$

yielding $C_3 := 2\epsilon_{\mathscr{S}}$. □

We have the following lemma to cover the set of short secants.

Lemma A.6. (Short Secants Covering Number) Let $\underline{\Omega'} = \{u_i\}$ be an $\epsilon$- cover for $\Omega' = O(n) \times \left(\mathfrak{D} \cap \mathfrak{B}_1(0)\right)$ and considering the following:

1. $\left\|f(u) - f(u') - Df_{u'}^T(u - u')\right\| \leq C_1 \left\|u - u'\right\|^2$ (Taylor approximation Lemma A.3)
2. $\left\|Df_u - Df_{u'}\right\| \leq C_2 \left\|u - u'\right\|$ (bounded curvature Lemma A.4)
3. $\left\|Df_u^{\dagger}\right\| \leq C_3$ (bounded gradient Lemma A.5),

where $f : \Omega \mapsto \mathfrak{S}_{\mathscr{H}}$is defined in Eqn. A.2 and $Df_u$ defines the first order differential of $f$ evaluated at the point $u$. Then given $u_i \in \Omega$, $\forall u, u' \in \mathfrak{B}_{\epsilon_0}(u_i)$ and $\left\|\mathscr{X} - \mathscr{X}'\right\| \leq \eta$, where $\mathscr{X} = f(u)$ and $\mathscr{X}' = f(u')$, we have

$$
\left\| \frac{\mathscr{X} - \mathscr{X}'}{\|\mathscr{X} - \mathscr{X}'\|} - Df_{u_i}^T \frac{u - u'}{\|\mathscr{X} - \mathscr{X}'\|} \right\| \leq C_4 \epsilon_0. \tag{A.24}
$$

where $C_4 := C_3(2C_1 + C_2)$.

*Proof.*

$$
\begin{aligned}
\left\| \mathscr{X} - \mathscr{X}' - Df_{u_i}^T(u-u') \right\| &= \left\| f(u) - f(u') - Df_u^T(u-u') + \left(Df_u - Df_{u_i}\right)^T \left(u-u'\right) \right\| \\
&\leq \left\| f(u) - f(u') - Df_u^T(u-u') \right\| + \left\| \left(Df_u - Df_{u_i}\right)^T \left(u-u'\right) \right\| \\
&\leq C_1 \left\| u - u' \right\|^2 + C_2 \left\| u - u_i \right\| \left\| u - u' \right\| \\
&= (\star)
\end{aligned}
$$

Given that $u, u' \in \mathfrak{B}_{\epsilon_0}(u_i)$, we have that $\left\| u - u_i \right\| \leq \epsilon_0$ and $\| u - u' \| \leq 2\epsilon_0$. Therefore

$$
\begin{aligned}
(\star) &\leq 2C_1\epsilon_0 \left\| u - u' \right\| + C_2\epsilon_0 \left\| u - u' \right\| \\
&= (2C_1 + C_2)\epsilon_0 \left\| u - u' \right\|.
\end{aligned}
$$

Now dividing by $\left\| \mathscr{X} - \mathscr{X}' \right\|$ gives:

$$
\begin{aligned}
\left\| \frac{\mathscr{X} - \mathscr{X}'}{\|\mathscr{X} - \mathscr{X}'\|} - Df_{u_i}^T \frac{u-u'}{\|\mathscr{X} - \mathscr{X}'\|} \right\| &\leq (2C_1 + C_2) \frac{\left\| u - u' \right\|}{\|\mathscr{X} - \mathscr{X}'\|} \\
&\leq C_3(2C_1 + C_2).
\end{aligned}
$$

In the last line, we have used the fact that bounded (inverse) gradient implies Lipschitzness. □

As a result, the set of bounded tangent vectors, defined by

$$
\mathscr{V} := \left\{ Df_{u_i}^T \frac{u-u'}{\|\mathscr{X} - \mathscr{X}'\|} \mid \forall u_i \in \Omega \right\},
$$

forms a generalized $\epsilon$-cover for $\mathfrak{N}_\eta^c \left( \bar{\mathfrak{S}}_{\mathscr{H}} - \bar{\bar{\mathfrak{S}}}_{\mathscr{H}} \right)$ with covering number (see Lemma 4.3 of [19])

$$
\begin{aligned}
\mathrm{CN}\left(\mathscr{V}, \|\cdot\|, \epsilon\right) &\leq C_4 \, \mathrm{CN}\left( \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}, \|\cdot\|, \epsilon_0 \right) \left( \frac{3}{\epsilon} \right)^{\frac{n(n+1)}{2}} \\
&\leq C_4 \left( \frac{6}{\epsilon_0} \right)^{n(n+1)} \left( \frac{3}{\epsilon} \right)^{\frac{n(n+1)}{2}}.
\end{aligned}
$$

From (A.5), we can bound the covering number of the normalized secant set:

$$
\begin{aligned}
\mathrm{CN}\left(\mathfrak{N}\left(\mathfrak{S}_{\mathscr{H}} - \mathfrak{S}_{\mathscr{H}}\right), \|\cdot\|, \epsilon\right) &\leq \mathrm{CN}\left(\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_{\mathscr{H}} - \bar{\mathfrak{S}}_{\mathscr{H}}\right), \|\cdot\|, \epsilon\right) + \mathrm{CN}\left(\mathfrak{N}_\eta^c\left(\bar{\mathfrak{S}}_{\mathscr{H}} - \bar{\bar{\mathfrak{S}}}_{\mathscr{H}}\right), \|\cdot\|, \epsilon\right) \\
&\leq \left( \frac{6}{\gamma\epsilon} \right)^{2n(n+1)} + C_4 \left( \frac{6}{\epsilon_0} \right)^{n(n+1)} \left( \frac{3}{\epsilon} \right)^{\frac{n(n+1)}{2}} \\
&\leq \left( \frac{6}{\gamma\epsilon} \right)^{2n(n+1)} + C_4 \left( \frac{6}{\epsilon_0} \right)^{n(n+1)} \left( \frac{3}{\epsilon} \right)^{n(n+1)} \\
&= \left( \frac{6}{\gamma\epsilon} \right)^{2n(n+1)} + C_4 \left( \frac{18}{\epsilon_0\epsilon} \right)^{n(n+1)} \\
&= (\star).
\end{aligned}
$$

Note that by definition $\epsilon_0 \le \eta\ (= 4\gamma)$, therefore $\gamma \ge \frac{\epsilon_0}{4}$. As a result

$$
\begin{aligned}
(\star) &\le C_4\left(\left(\frac{24}{\epsilon_0\epsilon}\right)^{2n(n+1)} + \left(\frac{24}{\epsilon_0\epsilon}\right)^{n(n+1)}\right)\\
&\le \left(\frac{48C_4}{\epsilon_0\epsilon}\right)^{2n(n+1)}\\
&\le \left(\frac{C_0}{\epsilon}\right)^{2n(n+1)},
\end{aligned}
$$

where $C_0 = \frac{48C_4}{\epsilon_0}$.

## B. Proof of Theorem 4.2

*Proof.* First, note that as the Frobenius norm is rotationally invariant we have that

$$\|\mathscr{X}\|_F = \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \cdots \times_4 \mathbf{Q}\|_F = \|\mathscr{S}\|_F \le R.$$

As $\hat{\mathscr{X}}$ is an empirical average of the true expected cumulant tensor $\mathscr{X}$, we can use a version of the vectorial Hoeffding's inequality in Lemma 4 of [51] that states with probability at least $1-\rho$ on the random draw of $\mathbf{z}_1, \ldots, \mathbf{z}_N$ that

$$\|\hat{\mathscr{X}} - \mathscr{X}\|_F \le \frac{R\left(1+\sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{B.1}$$

Next, we can use the boundedness property of random Gaussian measurements [20]. First, let $\mathbf{e} = \operatorname{vec}\left(\hat{\mathscr{X}} - \mathscr{X}\right)$ denote the finite approximation error from above. Then the boundedness property of sub-Gaussian matrices (see Definition 6.2 in [20]) states with probability at least $1-\xi$ on the sampling of $\mathbf{A}$ that

$$\|\mathbf{A}\mathbf{e}\|_2 \le C\|\mathbf{e}\|_2 \tag{B.2}$$

for some constant $C > 0$.

Combining the two equations, we get with probability at least $(1-\rho)(1-\xi) \ge 1-\rho-\xi$ on the drawing of both $\mathbf{A}$ and $\mathbf{z}_1, \ldots, \mathbf{z}_N$ that

$$\|\mathscr{A}(\mathscr{X}) - \mathscr{A}(\hat{\mathscr{X}})\|_2 \le \frac{CR\left(1+\sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{B.3}$$

□