The Annals of Statistics
2024, Vol. 52, No. 1, 131–156
https://doi.org/10.1214/23-AOS2331

STATISTICAL-COMPUTATIONAL TRADE-OFFS IN TENSOR PCA AND RELATED PROBLEMS VIA COMMUNICATION COMPLEXITY

BY RISHABH DUDEJA[1,a] AND DANIEL HSU[2,b]

[1]*Department of Statistics, University of Wisconsin–Madison,* [a]*rdudeja@wisc.edu*
[2]*Department of Computer Science, Columbia University,* [b]*djhsu@cs.columbia.edu*

Tensor PCA is a stylized statistical inference problem introduced by Montanari and Richard to study the computational difficulty of estimating an unknown parameter from higher-order moment tensors. Unlike its matrix counterpart, Tensor PCA exhibits a statistical-computational gap, that is, a sample size regime where the problem is information-theoretically solvable but conjectured to be computationally hard. This paper derives computational lower bounds on the run-time of memory bounded algorithms for Tensor PCA using communication complexity. These lower bounds specify a trade-off among the number of passes through the data sample, the sample size and the memory required by any algorithm that successfully solves Tensor PCA. While the lower bounds do not rule out polynomial-time algorithms, they do imply that many commonly-used algorithms, such as gradient descent and power method, must have a higher iteration count when the sample size is not large enough. Similar lower bounds are obtained for non-Gaussian component analysis, a family of statistical estimation problems in which low-order moment tensors carry no information about the unknown parameter. Finally, stronger lower bounds are obtained for an asymmetric variant of Tensor PCA and related statistical estimation problems. These results explain why many estimators for these problems use a memory state that is significantly larger than the effective dimensionality of the parameter of interest.

1. Introduction. Many statistical inference problems exhibit a range of sample sizes or signal-to-noise ratios in which it is information-theoretically possible to infer the unknown parameter of interest, but all known (computationally) efficient estimators fail to give accurate inferences. It is widely conjectured that, for many such problems, no efficient algorithm can produce accurate inferences in these so-called (conjectured) "hard" phases, even though there may be efficient algorithms that work if the sample size or signal-to-noise ratio is sufficiently high (i.e., in the "easy" phase of the problem). The existence of such a hard phase is known as a *statistical-to-computational gap*. Since proofs of such gaps are currently out-of-reach, a popular way to give evidence for the gaps is to prove that certain restricted classes of estimators fail to solve the inference problems in the conjectured hard phases. These restrictions are often chosen to capture the techniques used by the best efficient estimators known to date, for example, sum-of-squares relaxations [54], belief propagation and message passing [4, 67], general first-order methods [18] and low-degree polynomial functions [44, 47, 57].

Another way to constrain estimators is to require additional desirable properties, such as:

1. robustness to deviations from model assumptions,
2. low memory footprint,
3. low communication cost in a distributed computing environment.

If all estimators with these properties were proved to fail in the conjectured hard phases of inference problems, then we would have a satisfying practical theory of statistical-computational gaps. That is, even if efficient estimators exist in the conjectured hard phase

Received June 2022; revised May 2023.
MSC2020 subject classifications. 62F30, 68Q17.
Key words and phrases. Statistical-computational trade-offs, tensors.

of inference problems, their use in practice would be limited since they would use too much memory, be nonrobust to slight model mismatches, etc.

Steinhardt, Valiant and Wager [60] provide another motivation for studying inference problems under such constraints. They hypothesize that computationally easy problems remain solvable even in the face of constraints, such as those related to robustness, memory and communication. In contrast, hard problems have brittle solutions, which are unable to endure such constraints. Hence, hard problems should exhibit certain hallmarks such as the nonrobustness, high memory footprint or high communication cost of efficient estimators. Studying inference problems under such constraints enriches our understanding of the computational complexity of these problems.

The hypothesis of Steinhardt, Valiant and Wager [60] is supported by the power of Kearns' Statistical Query (SQ) model for explaining known statistical-computational gaps [33, 34, 45, 64]. In the SQ model, estimators can only access the data set by querying summary statistics of the data set, and they must be tolerant to adversarial perturbations in query responses of magnitude similar to the random fluctuations of these statistics. For many inference problems believed to exhibit a hard phase, it is known that all efficient estimators that are robust in the SQ-sense will fail to solve these inference problems in that phase (e.g., [33, 34, 64]).

In this paper, we further investigate the hypothesis of Steinhardt, Valiant and Wager by studying Tensor PCA and related problems that exhibit a similar statistical-computational gap under *memory constraints*. Our results are, in fact, obtained by studying the effect of *communication constraints*, and then leveraging a reduction from communication-bounded estimation to memory-bounded estimation.

2. Statistical inference and computational constraints.

In this section, we introduce notation used throughout this paper, the setup for general statistical inference problems and the computational model for memory-bounded estimators under which we derive our runtime lower bounds.

2.1. *Notation.*

Important sets. $\mathbb{N}$ and $\mathbb{R}$ denote the set of positive integers and the set of real numbers, respectively. $\mathbb{N}_0 \overset{\text{def}}{=} \mathbb{N} \cup \{0\}$ is the set of nonnegative integers. For each $k, d \in \mathbb{N}$, $[k]$ denotes the set $\{1, 2, 3, \dots, k\}$, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space, $\mathbb{R}^{d\times k}$ denotes the set of all $d \times k$ matrices, and $\bigotimes^k \mathbb{R}^d$ denotes the set of all $d \times d \times \cdots \times d$ (k times) tensors with $\mathbb{R}$-valued entries.

Linear algebra. We denote the d-dimensional vectors $(1, 1, \dots, 1)$, $(0, 0, \dots, 0)$ and the $d \times d$ identity matrix using $\mathbf{1}_d$, $\mathbf{0}_d$ and $\boldsymbol{I}_d$, respectively. We will omit the subscript d when the dimension is clear from the context. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\|$ denotes the ℓ_2 norm of $\boldsymbol{v}$. For two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$ denotes the standard inner product on $\mathbb{R}^d$: $\langle \boldsymbol{u}, \boldsymbol{v}\rangle \overset{\text{def}}{=} \sum_{i=1}^d u_i v_i$. For two matrices or tensors $\boldsymbol{U}$ and $\boldsymbol{V}$, we analogously define $\|\boldsymbol{U}\|$, and $\langle \boldsymbol{U}, \boldsymbol{V}\rangle$ by stacking their entries to form a vector. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathsf{T}}$ denotes the transpose of $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_{\mathsf{op}}$ denotes the operator (or spectral) norm of $\boldsymbol{A}$. For a square matrix $\boldsymbol{A}$, $\mathsf{Tr}(\boldsymbol{A})$ denotes the trace of $\boldsymbol{A}$. Finally, for vectors $\boldsymbol{v}_{1:k} \in \mathbb{R}^d$, $\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k$ denotes the k-tensor with entries $(\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k)_{i_1, i_2, \dots, i_k} = (\boldsymbol{v}_1)_{i_1} \cdot (\boldsymbol{v}_2)_{i_2} \cdots (\boldsymbol{v}_k)_{i_k}$ for $i_{1:k} \in [d]$. When $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \cdots = \boldsymbol{v}_k = \boldsymbol{v}$, we shorthand $\boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}$ as $\boldsymbol{v}^{\otimes k}$.

Asymptotic notation. Given a two nonnegative sequences a_d and b_d indexed by $d \in \mathbb{N}$, we use the following notation to describe their relative magnitudes for large d. We say that $a_d \lesssim b_d$ or $a_d = O(b_d)$ or $b_d = \Omega(a_d)$ if $\limsup_{d\to\infty}(a_d/b_d) < \infty$. If $a_d \lesssim b_d$ and $b_d \lesssim a_d$, then we say that $a_d \asymp b_d$. If there exists a constant $\epsilon > 0$ such that $a_d \cdot d^{\epsilon} \lesssim b_d$ we say that $a_d \ll b_d$. We use polylog(d) to denote any sequence a_d such that $a_d \asymp \log^t(d)$ for some fixed constant $t \geq 0$.

```
Memory bounded estimation algorithm with resource profile $(N, T, s)$.
Input: $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)$, a data set of $N$ samples.
Output: An estimator $\hat{\boldsymbol{V}} \in \widehat{\mathcal{V}}$.
Variables: Memory state $\texttt{state} \in \{0,1\}^s$, initially all zeros.
• For iteration $t \in \{1, 2, \dots, T\}$:
  – For each sample $i \in \{1, 2, \dots, N\}$:
        $\texttt{state} \leftarrow f_{t,i}(\texttt{state}, \boldsymbol{x}_i)$
• Return estimator $\hat{\boldsymbol{V}} = g(\texttt{state})$.
```

FIG. 1. *Template for memory bounded estimation algorithms with resource profile* (N, T, s).

Important distributions. $\mathcal{N}(0, 1)$ denotes the standard Gaussian measure on $\mathbb{R}$, and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ denotes the standard Gaussian measure on $\mathbb{R}^d$. For any finite set A, $\mathsf{Unif}(A)$ denotes the uniform distribution on the elements of A.

Hermite polynomials. We will make extensive use of the Hermite polynomials $\{H_i : i \in \mathbb{N}_0\}$, which are the orthonormal polynomials for the Gaussian measure $\mathcal{N}(0, 1)$. We provide the necessary background regarding Hermite polynomials and analysis on the Gaussian Hilbert space in the Supplementary Material [32], Appendix I.2.

Miscellaneous. For an event $\mathcal{E}$, $\mathbb{I}_{\mathcal{E}}$ denotes the indicator random variable for $\mathcal{E}$. For $x, y \in \mathbb{R}$, $x \vee y$ and $x \wedge y$ denote $\max(x, y)$ and $\min(x, y)$, respectively. For $x > 0$, $\log(x)$ denotes the natural logarithm (base e) of x.

2.2. *Statistical inference problems.* A general statistical inference problem is specified by a *model* $\mathcal{P}$, which is a collection of probability distributions on a space $\mathcal{X}$. Elements of $\mathcal{P}$ are indexed by a *parameter* $\boldsymbol{V} \in \mathcal{V}$, so $\mathcal{P} = \{\mu_{\boldsymbol{V}} : \boldsymbol{V} \in \mathcal{V}\}$, where $\mathcal{V}$ is the *parameter set*. A statistical inference problem can be thought of as a game between nature and a statistician. First, nature picks a parameter $\boldsymbol{V} \in \mathcal{V}$, which is not revealed to the statistician. Then the N samples $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)$ are drawn i.i.d. from $\mu_{\boldsymbol{V}}$ and revealed to the statistician. The statistician constructs an *estimator* $\hat{\boldsymbol{V}}(\boldsymbol{x}_{1:N}) \in \widehat{\mathcal{V}}$ using the data set $\boldsymbol{x}_{1:N}$ and incurs a loss $\ell(\boldsymbol{V}, \hat{\boldsymbol{V}})$, where $\ell : \mathcal{V} \times \widehat{\mathcal{V}} \to [0, \infty)$ is the *loss function*. An estimator $\hat{\boldsymbol{V}} : \mathcal{X}^N \to \widehat{\mathcal{V}}$ is (ϵ, δ)-*accurate* if

$$\sup_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\big(\ell\big(\boldsymbol{V}, \hat{\boldsymbol{V}}(\boldsymbol{x}_{1:N})\big) \geq \epsilon\big) < \delta. \tag{1}$$

The statistician's goal is to construct an (ϵ, δ)-*accurate* estimator with the smallest sample size N.

2.3. *Memory bounded estimators.* We study iterative estimation algorithms that maintain and update an internal memory state of s bits in the course of T passes (iterations) over a data set of N samples. A general template of such an iterative algorithm is given in Figure 1, and a formal definition appears below.

DEFINITION 1 (Memory bounded estimation algorithm with resource profile (N, T, s)). A *memory bounded estimation algorithm* with *resource profile* (N, T, s) computes an estimator by making T passes through a data set of N samples using a memory state of s bits (initially all zeros). Such an algorithm is specified by the *update functions* $f_{t,i} : \{0, 1\}^s \times \mathcal{X} \to \{0, 1\}^s$ and an *estimator function* $g : \{0, 1\}^s \to \widehat{\mathcal{V}}$, which are used as follows. In the tth pass through the data set, the algorithm considers each sample $\boldsymbol{x}_i$ for $i \in [N]$ in sequence, and it updates the memory state by applying the update function $f_{t,i}$ to the current memory state and the sample $\boldsymbol{x}_i$ under consideration. After all T passes are complete, the estimator is computed by applying the estimator function g to the final memory state. A template for a general memory bounded algorithm is given in Figure 1.

The above class of iterative algorithms is well suited for modeling commonly-used estimators, for example, spectral estimators (using the power method) and empirical risk minimizers (using gradient descent). We measure the computational cost of an estimation algorithm by T, the number of passes it makes through the data set. This cost measure is not sensitive to the size of the data set. Furthermore, the update and estimator functions are permitted to be arbitrary functions, and we do not consider their computational cost in our lower bounds. This means that the lower bounds are conservative, in that a more detailed accounting of their costs in a concrete computational model would only improve our lower bounds.

3. Our contributions.

We study several statistical inference problems exhibiting a statistical-computational gap and prove lower bounds on the total number of resources, as measured by the product $N \cdot T \cdot s$ of the sample size N, number of iterations (or passes) T and the size of the memory state s that all iterative algorithms must use to solve these problems. In the following paragraphs, we introduce the problems we study at a high level and highlight our main results.

Tensor principal components analysis. In the order-k Tensor Principal Components Analysis (k-TPCA) problem introduced by Montanari and Richard [50], one observes N noisy independent realizations of an unknown rank-1 symmetric k-tensor (the signal) corrupted by Gaussian noise. The unknown signal tensor can be specified using a d-dimensional vector, which is the parameter of interest for this problem, and the goal is to estimate it. This problem is believed to exhibit a sizeable computational-statistical gap. Our main result for this problem (Theorem 1 in Section 5) provides a lower bound on the total number of resources $N \cdot T \cdot s$ used by any iterative algorithm for Tensor PCA. Many natural algorithms for this problem (such as the tensor power method or the maximum likelihood estimator computed using gradient descent) use a memory state of size $s \asymp d$ proportional to the dimension of the parameter of interest (i.e., use linear memory). By instantiating our lower bound for algorithms with this property, we obtain unconditional lower bounds on their run-time. While these lower bounds do not rule out polynomial-time linear-memory algorithms for Tensor PCA, we are not aware of any other approach that yields an unconditional lower bound for linear memory iterative algorithms that are comparable to our results. In particular, the popular low-degree likelihood ratio framework [44, 47] yields weaker run-time lower bounds.

Non-Gaussian component analysis. Montanari and Richard intended Tensor PCA as a stylized statistical inference problem that captures computational difficulties in extracting information about a parameter of interest from the empirical higher-order moment tensor of a data set. Taking a cue from this motivation, we study the order-k Non-Gaussian Component Analysis (k-NGCA) problem [11], defined as follows. The goal is to estimate an unknown unit vector $\boldsymbol{v}$ from N i.i.d. realizations of a d-dimensional non-Gaussian vector $\boldsymbol{x}$ in which: (1) the order-k moment tensor differs from the moment tensor of a standard Gaussian vector $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ along precisely one direction, given by $\boldsymbol{v}$; and (2) for any $\ell \leq k-1$, the order-ℓ moment tensor of $\boldsymbol{x}$ is identical to that of the Gaussian vector $\boldsymbol{z}$, and hence it reveals no information about $\boldsymbol{v}$. We show (Theorem 3 in Section 7) that a resource lower bound identical to our result for k-TPCA holds for k-NGCA when the signal-to-noise ratio is sufficiently small as a function of d. Since our lower bound applies to a broad family of constructions of the non-Gaussian vector $\boldsymbol{x}$, we obtain as corollaries to Theorem 3, similar results for the estimation problems in specific statistical models, including Gaussian mixture models and certain generalized linear models.

Asymmetric tensor PCA. We also study an *asymmetric* version of k-TPCA (k-ATPCA), where the signal tensor is allowed to be an arbitrary (possibly asymmetric) rank-1 tensor. The conjectured hard phase for this problem is identical to that for (symmetric) k-TPCA. However, our main result for this problem k-ATPCA (Theorem 2 in Section 6) shows that this problem requires significantly more total resources $N \cdot T \cdot s$ than k-TPCA. The state-of-the-art efficient estimators for k-ATPCA (such as the Montanari and Richard spectral estimator) are, in a sense, overparameterized: they require a memory state size that is significantly larger than the effective-dimension of the parameter of interest. A key consequence of our results is that this overparameterization is necessary: estimators that use a smaller memory state have a strictly worse run-time veersus sample size trade-off (shown in Figure 3b) compared to sufficiently overparametrized estimators such as as the Montanari and Richard spectral estimator. This explains why estimators for k-ATPCA "lift" the problem to higher dimensions.

Canonical correlation analysis. Since k-TPCA captures the computational difficulties in extracting information about a parameter of interest from the empirical k-moment tensor of a data set, it is natural to expect that k-ATPCA should capture the computational difficulties of the same but for the empirical *k-cross-moment tensor* of a data set. To develop this analogy, we study the order-k Canonical Correlation Analysis problem (k-CCA), in which one observes N i.i.d. realizations of a kd-dimensional random vector $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(k)})$ consisting of k separate d-dimensional "views." The parameter of interest is the order-k cross-moment tensor $\mathbb{E}[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \cdots \otimes \boldsymbol{x}^{(k)}]$, and hard instances of this problem have the property that no other moment tensor of order-ℓ with $\ell \leq k$ carries information regarding the parameter of interest. For the k-CCA problem, our main result (Theorem H.1 in the Supplementary Material [32], Appendix H.3) shows that a resource lower bound identical to our result for k-ATPCA holds for the k-CCA problem in the regime when signal-to-noise ratio is sufficiently small as a function of d. Furthermore, since the problem of *learning parity functions* can be reduced to the k-CCA instance used to prove the resource lower bound for k-CCA, we also obtain interesting resource lower bounds for the problem of learning parities. This is discussed further in the Supplementary Material [32], Appendix H.4.

Organization. The remainder of this paper is organized as follows. Section 4 discusses several strands of related works relevant to this paper. Sections 5–7 introduce the various inference problems we study and state and discuss our computational lower bounds for each of them: Section 5 is devoted to (symmetric) tensor PCA, Section 6 to asymmetric tensor PCA, and Section 7 to non-Gaussian component analysis. The results for canonical correlation analysis are presented in the Supplementary Material [32], Appendix H. Section 8 presents the proof framework that underlies each of the computational lower bounds presented in this paper. As an illustrative application of the proof framework presented in Section 8, we provide the proof for the computational lower bound for (symmetric) tensor PCA in Section 9. The detailed proofs for the remaining inference problems are provided in the Supplementary Material [32].

4. Related work.

Since the work of Montanari and Richard, which introduced k-TPCA, a number of subsequent works have proposed and analyzed various estimators and proved different kinds of computational lower bounds for this and related problems.

Hardness of symmetric and asymmetric tensor PCA. Many works have designed computationally efficient estimators for k-TPCA that attain the conjectured optimal sample complexity for polynomial-time estimators ($N\lambda^2 \gtrsim d^{k/2}$). This includes spectral estimators [10, 41–43, 50, 69], sum-of-squares relaxations [41–43], tensor power method with well-designed initializations [3, 10] and higher-order generalizations of belief propagation [65]. The spectral estimator of Hopkins et al. [42] for k-TPCA and the spectral estimator of Montanari and

Richard [50] for k-ATPCA (discussed in detail in Section 5.4 and Section 6.4) are particularly relevant for our work. The resources used by these estimators (as measured by the product $N \cdot T \cdot s$) nearly match our resource lower bounds for k-TPCA and k-ATPCA, respectively. Several works have shown that many natural classes of estimators fail in the conjectured hard phase for k-TPCA ($d \lesssim N\lambda^2 \ll d^{k/2}$). This includes sum-of-squares relaxations [9, 41–43], estimators that compute a low degree polynomial of the data set [44, 47] and SQ algorithms [15, 30]. A more detailed comparison with the low-degree lower bounds appears in Section 5.4. The landscape of the maximum likelihood objective for this problem has been shown to have numerous spurious critical points [8, 56] and it is known that Langevin dynamics on the maximum likelihood objective fails to solve k-TPCA in the conjectured hard phase [7]. Finally, using average-case reductions, it has been shown that the hardness of hypergraph planted clique implies the hardness of k-TPCA [14, 68].

Hardness of non-Gaussian component analysis. The k-NGCA problem was formally introduced by Blanchard et al. [11], and various computationally efficient estimators have been proposed and analyzed [21, 39, 48, 61, 62]. These estimators have a sample size requirement, which is significantly more than the information-theoretic sample size requirement. In the special case when the distribution of the non-Gaussian direction is discrete, Zadik et al. [66] and Diakonikolas and Kane [24] have designed computationally efficient algorithms that recover the non-Gaussian direction with the information-theoretically optimal sample complexity. However, these algorithms are brittle and break down when the distribution of the non-Gaussian component is sufficiently nice (e.g., absolutely continuous with respect to the standard Gaussian distribution; see Remark 2 for additional details). In this situation, Diakonikolas, Kane and Stewart [28] have identified a sample size regime where SQ algorithms fail to identify the non-Gaussian direction with polynomially many queries. This suggests that this problem is computationally hard in this regime. A problem closely related to k-NGCA problem is the continuous learning with errors problem [16]. Bruna et al. [16] show that this problem is computationally hard provided that a plausible conjecture from cryptography is true [49], Conjecture 1.2. Since k-NGCA is connected to many other inference problems, the SQ lower bounds of Diakonikolas, Kane and Stewart are at the heart of SQ lower bounds for many other robust estimation and learning problems [17, 22, 23, 25–29].

Memory and communication lower bounds for statistical inference. The computational lower bounds obtained in our work rely on a reduction (Fact 1) of Alon, Matias and Szegedy [2], which was more recently used in the context of statistical inference problems in the works of Shamir [58] and Dagan and Shamir [20]. This reduction shows that any iterative algorithm that solves a statistical inference problem using few resources (as measured by the product $N \cdot T \cdot s$) can be used to solve the statistical inference problem in a distributed setting with a limited amount of communication between the machines holding the data samples. Consequently, the claimed resource lower bounds follow from communication lower bounds for the distributed versions of these inference problems. Recent works by Han, Özgür and Weissman [40], Barnes, Han and Ozgur [5], Acharya et al. [1] have developed general frameworks to prove communication lower bounds distributed statistical inference problems. However, for k-TPCA and k-NGCA, we were unable to obtain the desired communication lower bounds using these frameworks (see Section 8.7 for more details). Hence, building on these works, we develop a different approach to obtain communication lower bounds for distributed inference problems, which yields stronger communication lower bounds for k-TPCA and k-NGCA than those obtained using the prior works [1, 5, 40]. On the other hand, for k-ATPCA and k-CCA problems, the desired communication lower bounds can be obtained from existing communication lower bounds for sparse Gaussian mean estimation [1, 13] and correlation detection problems [20]. We show that the approach developed in this paper also yields alternative proofs for the desired communication lower bounds for k-ATPCA and the k-CCA in a unified manner. We refer the reader to the Supplementary Material [32], Remark E.1, for a detailed discussion. A different line of work [6, 35–37, 46, 51–53, 55, 59, 60] initiated by Steinhardt, Valiant and Wager [60] and Raz [55] provides another approach to obtain memory lower bounds without relying on the connection with distributed inference problems. We provide a comparison with lower bounds obtained using this approach in the Supplementary Material [32], Section H.4.1.

5. Symmetric tensor PCA.

5.1. *Problem formulation.* In the symmetric order-k Tensor PCA (k-TPCA) problem introduced by Montanari and Richard [50], one observes N i.i.d. tensors $\boldsymbol{X}_{1:m} \in \bigotimes^k \mathbb{R}^d$ sampled as follows:

$$\boldsymbol{X}_i = \frac{\lambda \boldsymbol{V}^{\otimes k}}{\sqrt{d^k}} + \boldsymbol{W}_i, \qquad (W_i)_{j_1,j_2,\dots j_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \forall j_1, j_2, \dots, j_k \in [d]. \tag{2}$$

In the above display, $\lambda > 0$ is the signal-to-noise ratio, and $\boldsymbol{V} \in \mathcal{V}$ is the unknown parameter one seeks to estimate. The parameter space for this problem is $\mathcal{V} = \{\boldsymbol{V} \in \mathbb{R}^d : \|\boldsymbol{V}\| = \sqrt{d}\}$. We let the probability measure $\mu_{\boldsymbol{V}}$ denote the distribution of a single sample $\boldsymbol{X}_i$ in (2).

5.2. *Statistical-computational gap in k-TPCA.* Depending on the scaling of the effective sample size $N\lambda^2$, k-TPCA exhibits three phases.

Impossible phase. If $N\lambda^2 \ll d$, recovering $\boldsymbol{V}$ is information-theoretically impossible [50].

Conjectured hard phase. In the regime $d \lesssim N\lambda^2 \ll d^{k/2}$, the maximum likelihood estimator succeeds in recovering $\boldsymbol{V}$ [50]. However, it is not known how to compute the maximum likelihood estimator using a polynomial-time algorithm. No known polynomial-time estimation algorithm has a nontrivial performance in this phase. Based on evidence from the low degree likelihood ratio framework [47], the statistical query framework [15, 30], the sum-of-squares hierarchy framework [41] and the average-case reductions framework [14, 68], it is believed that no polynomial-time algorithm can have nontrivial performance in this phase.

Easy phase. In the regime $N\lambda^2 \gtrsim d^{k/2}$, there are polynomial-time algorithms that accurately estimate $\boldsymbol{V}$ [3, 10, 42, 43, 50, 65, 69].

5.3. *Computational lower bound.* The following is our lower bound for k-TPCA.

THEOREM 1. *Let $\hat{\boldsymbol{V}}$ denote any estimator for k-TPCA with $k \geq 2$ and $\lambda \asymp 1$ (as $d \to \infty$) that can be computed using a memory bounded estimation algorithm with resource profile (N, T, s) scaling with d as*

$$N \asymp d^\eta/\lambda^2, \qquad T \asymp d^\tau, \qquad s \asymp d^\mu$$

for any constants $\eta \geq 1$, $\tau \geq 0$, $\mu \geq 0$. If

$$\eta + \tau + \mu < \left\lceil \frac{k+1}{2} \right\rceil,$$

then, for any $t \in \mathbb{R}$,

$$\limsup_{d\to\infty} \inf_{\boldsymbol{V}\in\mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d} \right) \leq 2\exp\left(-\frac{t^2}{2}\right).$$

The above result shows that if the total resources used by a memory-bounded estimator $\hat{\boldsymbol{V}}$ (as measured by the product $N \cdot T \cdot s$) is too small, there is a worst-case choice of $\boldsymbol{V} \in \mathcal{V}$ such that, on an event of probability arbitrarily close to 1, we have

$$\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \lesssim \frac{1}{d}.$$

On the other hand, for any $\boldsymbol{V} \in \mathcal{V}$, the trivial estimator $\hat{\boldsymbol{V}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ achieves

$$\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \gtrsim \frac{1}{d},$$

with probability arbitrarily close to 1. Hence, memory bounded estimation algorithms using too few total resources perform no better than a random guess.

5.4. *Discussion of Theorem* 1. We now discuss some key implications of Theorem 1. Recall that we consider the scaling regime where $d \to \infty$ and $\lambda \asymp 1$, $N \asymp d^\eta/\lambda^2$, $T \asymp d^\tau$, $s \asymp d^\mu$; the exponents $\eta \geq 1$, $\tau \geq 0$ and $\mu \geq 0$ are fixed constants. We additionally restrict our discussion to the case where k *is even*, because our lower bounds appear to be deficient by a factor of $\sqrt{d}$ when k is odd (additional details are provided in the Supplementary Material [32], Appendix D.1, regarding the odd case).

5.4.1. *Consequences for linear memory algorithms.* Theorem 1 has some interesting consequences for memory-bounded estimation algorithms with a memory state of size $s \asymp d \operatorname{polylog}(d)$ bits. We call such algorithms nearly linear memory algorithms. For such algorithms to have a nontrivial performance for k-TPCA, the sample size exponent $\eta = \log(N\lambda^2)/\log(d)$ and the run-time exponent $\tau = \log(T)/\log(d)$ must satisfy

$$\tau + \eta \geq \frac{k}{2}. \tag{3}$$

This gives a lower bound on the run-time exponent as a function of the sample-size exponent $\tau \geq k/2 - \eta$, which rules out certain run-time exponents in the conjectured hard phase for Tensor PCA ($1 < \eta < k/2$). The run-time exponents ruled out by Theorem 1 is the triangular subregion of the conjectured hard phase shaded in red and gray in Figure 2.

Many natural algorithms for k-TPCA are nearly linear memory algorithms, and hence, the run-time versus sample size trade-off obtained in (3) also applies to them. This includes algorithms like the tensor power method [3, 10, 50], Langevin dynamics or gradient descent on the maximum likelihood objective [7] and partial trace spectral estimator of Hopkins et al. [42]. These algorithms are discussed in more detail in the Supplementary Material [32], Appendix D.2.

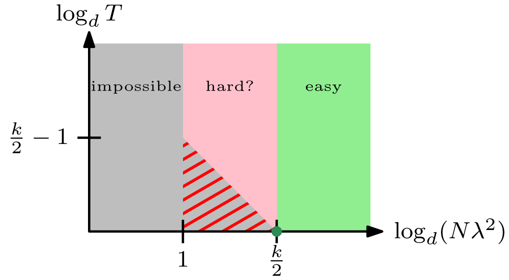


FIG. 2. *Consequences of Theorem* 1 *for linear memory* k*-TPCA algorithms* (k *even*).

5.4.2. *Tightness of Theorem* 1. The partial trace spectral estimator of Hopkins et al. [42] is a memory bounded estimation algorithm with resource profile:

$$\left(N \asymp d^{\frac{k}{2}}/\lambda^2, T = \text{polylog}(d), s = d \cdot \text{polylog}(d)\right).$$

This is a (nearly) linear memory estimation algorithm for Tensor PCA whose sample size exponent η and run-time exponent τ satisfy $\eta + \tau \leq k/2 + \epsilon$ for arbitrary $\epsilon > 0$. This shows that the run-time versus sample size trade-offs implied for linear-memory algorithms by Theorem 1 are tight. Furthermore, this shows that Theorem 1 provides a weak separation between the easy and the conjectured hard phases:

1. In the easy phase, when the sample size exponent $\eta > k/2$, there are (nearly) linear memory algorithms whose run-time exponent is arbitrarily close to zero ($\tau \leq \epsilon$ for any $\epsilon > 0$).
2. In contrast, in the hard phase, when the sample size exponent $\eta < k/2$, Theorem 1 shows that any linear memory algorithm must have a strictly positive run-time exponent $\tau \geq k/2 - \eta$.

5.4.3. *Comparison with low-degree lower bounds.* Lastly, it is interesting to compare the lower bounds implied by Theorem 1 with the lower bounds obtained using the low-degree likelihood framework. Kunisky, Wein and Bandeira [47], Theorem 3.3, show that when

$$N\lambda^2 \lesssim \frac{d^{\frac{k}{2}}}{D^{\frac{k-2}{2}}},$$

any procedure that computes a degree-D polynomial of the data set $\boldsymbol{X}_{1:N}$ fails to solve k-TPCA [47], Theorem 4. In general, the lower bounds obtained from Theorem 1 are incomparable to those obtained from the low-degree framework for the following reasons:

1. The low-degree polynomial makes no restrictions on the amount of memory used to compute the polynomial.
2. There are no degree restrictions placed on memory bounded estimation algorithms.

However, one can still make interesting comparisons between lower bounds obtained for algorithms that can be implemented in both computational models. An important example is the tensor power method (a general class of examples is discussed in the Supplementary Material [32], Appendix D.2). The Tth iterate of the tensor power method is a polynomial in $\boldsymbol{X}_{1:N}$ of degree $D = (k-1)^T$. Hence, low-degree lower bounds only show the failure of iterative schemes like the tensor power method in the conjectured hard phase of k-TPCA for $T \lesssim \log(d)$ iterations. The failure of the low-degree framework to give iteration lower bounds of the form $T \gtrsim d^\delta$ for any $\delta > 0$ because of the following reasons:

1. The low-degree framework measures the computational cost of computing a polynomial only using its degree. Hence, in order to show an iteration lower bound of $T \gtrsim d^\delta$, one would have to show that polynomials of degree $D = \exp(O(d^\delta))$ fail to solve k-TPCA.
2. However, it is known that for every $\epsilon \in (0, 1)$, there is a computationally inefficient estimator based on a degree $D \lesssim d^\epsilon$ polynomial that solves k-TPCA in a part of the hard phase with sample-size exponent $\eta = \epsilon + k(1-\epsilon)/2 < k/2$ (see discussion in [47], page 16, and references therein).

In contrast, since tensor power method can be implemented with a memory state of size $s \asymp d$ polylog (d) bits (see [32], Appendix D.2), stronger iteration lower bounds (recall (3) and Figure 2) are obtained via Theorem 1 by exploiting the fact that the output of the tensor power method has an additional structural property not shared by arbitrary polynomials of comparable degree: it can be computed by T iterations of a linear memory algorithm.

6. Asymmetric tensor PCA.

6.1. *Problem formulation.* In the asymmetric order-k Tensor PCA (k-ATPCA) problem, one observes N i.i.d. tensors $\boldsymbol{X}_{1:N} \in \bigotimes^k \mathbb{R}^d$ sampled as follows:

$$\boldsymbol{X}_i = \frac{\lambda \boldsymbol{V}_1 \otimes \boldsymbol{V}_2 \otimes \cdots \otimes \boldsymbol{V}_k}{\sqrt{d^k}} + \boldsymbol{W}_i, \qquad (W_i)_{j_1, j_2, \ldots, j_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \forall j_1, j_2, \ldots, j_k \in [d]. \tag{4}$$

In the above display, $\lambda > 0$ is the signal-to-noise ratio, and $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_k$ are unknown vectors in $\mathbb{R}^d$ with $\|\boldsymbol{V}_i\| = \sqrt{d}$. The goal is to estimate the rank-1 signal tensor $\boldsymbol{V} \overset{\text{def}}{=} \boldsymbol{V}_1 \otimes \boldsymbol{V}_2 \cdots \otimes \boldsymbol{V}_k$. The parameter space for this problem is $\mathcal{V} = \{\boldsymbol{V}_1 \otimes \boldsymbol{V}_2 \otimes \cdots \otimes \boldsymbol{V}_k : \boldsymbol{V}_i \in \mathbb{R}^d, \|\boldsymbol{V}_i\| = \sqrt{d} \ \forall i \in [k]\}$. We let the probability measure $\mu_{\boldsymbol{V}}$ denote the distribution of a single sample $\boldsymbol{X}_i$ in (4).

6.2. *Statistical-computational gap in k-ATPCA.* The delineations between the impossible phase, conjectured hard phase and easy phase for k-ATPCA are the same as in (symmetric) k-TPCA. In particular, the known polynomial-time algorithms that accurately estimate $\boldsymbol{V}$ require $N\lambda^2 \gtrsim d^{k/2}$ [50, 69].

6.3. *Computational lower bound.* The following is our lower bound for k-ATPCA.

THEOREM 2. *Let $\hat{\boldsymbol{V}} \in \bigotimes^k \mathbb{R}^d$ denote any estimator for k-ATPCA with $k \geq 2$ and signal-to-noise ratio $\lambda \asymp 1$ (as $d \to \infty$) that can be computed using a memory bounded estimation algorithm with resource profile (N, T, s) scaling with d as*

$$N \asymp d^\eta / \lambda^2, \qquad T \asymp d^\tau, \qquad s \asymp d^b$$

for any constants $\eta > 0$, $\tau \geq 0$, $b \geq 0$. If

$$\eta + \tau + b < k,$$

then, for any $t \in \mathbb{R}$,

$$\limsup_{d \to \infty} \inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}} \left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d^k} \right) \leq \frac{1}{t^2}.$$

Just as in the case of (symmetric) k-TPCA, Theorem 2 shows that memory bounded estimation algorithms for k-ATPCA using too few total resources (as measured by the product $N \cdot T \cdot s$) perform no better than a random guess.

6.4. *Discussion of Theorem 2.* We now discuss some implications of Theorem 2. We restrict attention to the situation where $\lambda \asymp 1$ and $k = 2\ell$ is even.

6.4.1. *Price of asymmetry.* A comparison of the computational lower bound for k-ATPCA (Theorem 2 and k-TPCA (Theorem 1) reveals that k-ATPCA is a more resource-intensive inference problem. The minimum amount of resources (as measured by the product $N \cdot T \cdot s$) needed to solve k-ATPCA is strictly more than the minimum amount of resources required to solve k-TPCA.

6.4.2. *Tightness of Theorem 2.* Let $\overline{\boldsymbol{X}} \in \bigotimes^k \mathbb{R}^d$ denote the empirical average of $\boldsymbol{X}_{1:N}$. Montanari and Richard [50] proposed estimating $\boldsymbol{V}$ by the best rank-1 approximation of the matrix obtained by flattening $\overline{\boldsymbol{X}}$ into a $d^\ell \times d^\ell$ matrix. The estimator is based on the *matricization operation* $\mathrm{Mat} : \bigotimes^k \mathbb{R}^d \to \mathbb{R}^{d^\ell} \times \mathbb{R}^{d^\ell}$, which reshapes a tensor into a matrix. To define $\mathrm{Mat}(\boldsymbol{T})$ for a tensor $\boldsymbol{T} \in \bigotimes^k \mathbb{R}^d$, we index the rows and columns of $\mathrm{Mat}(\boldsymbol{T})$ by ℓ-tuples of indices $(i_1, i_2, \dots, i_\ell) \in [d]^\ell$, so the entries of $\mathrm{Mat}(\boldsymbol{T})$ are given by

$$\mathrm{Mat}(\boldsymbol{T})_{(i_1,i_2,\dots,i_\ell);(j_1,j_2,j_3,\dots,j_\ell)} \overset{\text{def}}{=} T_{i_1,i_2,\dots,i_\ell,j_1,j_2,\dots,j_\ell}. \tag{5}$$

The estimator $\hat{\boldsymbol{V}}_{\mathrm{MR}}$ of Montanari and Richard is defined by $\hat{\boldsymbol{V}}_{\mathrm{MR}} = \mathrm{Mat}^{-1}(\hat{\boldsymbol{M}})$, where $\hat{\boldsymbol{M}}$ is the best rank-1 approximation (or the rank-1 SVD) of $\mathrm{Mat}(\overline{\boldsymbol{X}})$. This estimator was analyzed by Zheng and Tomioka [69] for k-ATPCA. Their analysis shows that in the regime $\lambda \asymp 1$, when $N\lambda^2 \gtrsim d^{\frac{k}{2}}$, $\hat{\boldsymbol{V}}_{\mathrm{MR}}$ is a consistent estimator for $\boldsymbol{V}$. Moreover, in this regime, the matrix $\hat{\boldsymbol{M}}$ has a spectral gap of size $\Delta \gtrsim 1$. Consequently, $\hat{\boldsymbol{V}}_{\mathrm{MR}}$ can be computed using $\mathrm{polylog}(d)$ iterations of the power method. Since $\mathrm{Mat}(\overline{\boldsymbol{X}}) \in \mathbb{R}^{d^\ell \times d^\ell}$ with $\ell = k/2$, in order to implement the power method using a memory bounded algorithm, one requires a memory state of size $s \asymp d^\ell\, \mathrm{polylog}(d)$ bits. Consequently, this estimator can be computed using a memory bounded estimation algorithm with resource profile

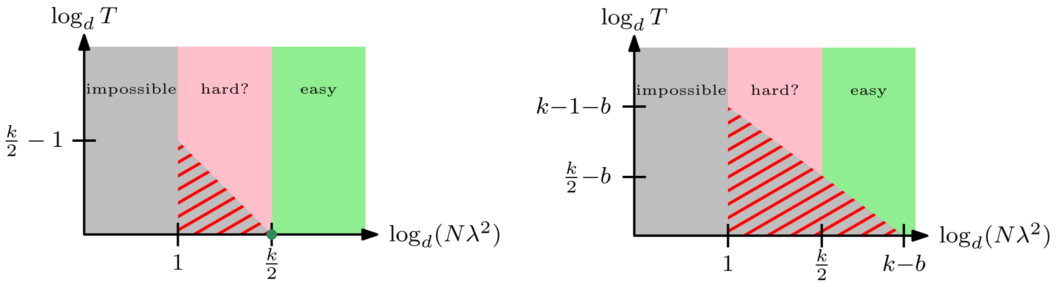
$$\big(N \asymp d^{\frac{k}{2}}/\lambda^2, T \asymp \mathrm{polylog}(d), s \asymp d^{\frac{k}{2}}\, \mathrm{polylog}(d)\big).$$

The total resources consumed by this estimation algorithm satisfies $N \cdot T \cdot s \ll d^{k+\epsilon}$ for any $\epsilon > 0$. This shows that the resource lower bound in Theorem 2 is nearly tight.

6.4.3. *A separation between easy and hard phases.* Theorem 2 has interesting consequences for memory bounded estimation algorithms that have a memory requirement comparable to the spectral estimator of Montanari and Richard, that is, $s \asymp d^{\frac{k}{2}}$. For such algorithms to have a nontrivial performance for k-ATPCA, the sample size exponent $\eta = \log(N\lambda^2)/\log(d)$ and the run-time exponent $\tau = \log(T)/\log(d)$ must satisfy

$$\tau + \eta \geq \frac{k}{2}. \tag{6}$$

This gives a lower bound on the run-time exponent as a function of the sample-size exponent: $\tau \geq k/2 - \eta$. This rules out certain run-time exponents in the conjectured hard phase for k-ATPCA ($1 < \eta < k/2$), specifically those in the striped triangular region in Figure 3a. (The spectral estimator of Montanari and Richard is depicted by the green dot at $(\log_d(N\lambda^2) =$



(a) $s \asymp d^{k/2}$ bits.

(b) $s \asymp d^b$ bits, $b < k/2$.

FIG. 3. *Consequences of Theorem 2 for k-ATPCA algorithms with memory size of* $s \asymp d^{\frac{k}{2}}$ *bits (left) and* $s \asymp d^b$ *bits for* $b < k/2$ *(right). The striped triangular region represents the run-time versus sample size trade-offs ruled out by Theorem 2. The green dot at* $(\log_d(N\lambda^2) = k/2, \log_d(T) = 0)$ *in Figure 3a represents the Montanari and Richard estimator.*

$k/2, \log_d(T) = 0$) in Figure 3a.) Hence, Theorem 2 provides a weak separation between the easy and the conjectured hard phases, similar to that provided by Theorem 1 for linear memory algorithms and k-TPCA.

6.4.4. *Necessity of overparameterization.* The Montanari–Richard spectral estimator is *overparameterized* in the sense that it uses a memory state of $s \gtrsim d^{\frac{k}{2}}$ bits, which is significantly larger than the effective dimension of the parameter of interest $\boldsymbol{V}$, namely kd, whenever $k \geq 3$. Theorem 2 shows that this amount of overparameterization is necessary. To see this, we instantiate Theorem 2 for memory state sizes of $s \asymp d^b$ bits for some $b < k/2$. For such memory bounded estimation algorithms to have a nontrivial performance for k-ATPCA, the sample size exponent $\eta = \log(N\lambda^2)/\log(d)$ and the run-time exponent $\tau = \log(T)/\log(d)$ must satisfy

$$\tau + \eta \geq \frac{k}{2} + \left(\frac{k}{2} - b\right). \tag{7}$$

The trade-off in (7) is strictly worse than the trade-off obtained from (6); compare the phase diagram in Figure 3b to that in Figure 3a. Hence, one cannot significantly reduce the overparameterization level (as measured by the size of the memory state) of the Montanari and Richard spectral estimator without increasing its run-time or sample-size exponents.

7. Non-Gaussian component analysis.

7.1. *Problem formulation.* In the non-Gaussian Component Analysis (NGCA) problem, one seeks to estimate an unknown vector $\boldsymbol{V} \in \mathbb{R}^d$ with $\|\boldsymbol{V}\| = \sqrt{d}$ from an i.i.d. sample $\boldsymbol{x}_{1:N}$ generated as follows:

$$\boldsymbol{x}_i = \eta_i \frac{1}{\sqrt{d}} \boldsymbol{V} + \left(\boldsymbol{I}_d - \frac{1}{d} \boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right) \boldsymbol{z}_i, \tag{8a}$$

where $\eta_i \in \mathbb{R}$ and $\boldsymbol{z}_i \in \mathbb{R}^d$ are independent random variables with distributions

$$\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad \eta_i \sim \nu. \tag{8b}$$

In the above display, ν is a non-Gaussian distribution on $\mathbb{R}$. Let $\mu_{\boldsymbol{V}}$ denote the distribution of $\boldsymbol{x}_i$ described by the above generating process (8). The likelihood ratio (with respect to the standard Gaussian distribution μ_0) of a single sample $\boldsymbol{x} \in \mathbb{R}^d$ from the model (8) is

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) = \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(\eta) \quad \text{where } \eta \overset{\text{def}}{=} \left\langle \boldsymbol{x}, \frac{1}{\sqrt{d}} \boldsymbol{V} \right\rangle. \tag{9}$$

REMARK 1. We overload the symbol μ_0 to mean $\mu_0 = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ on the left-hand side of (9), and $\mu_0 = \mathcal{N}(0, 1)$ on the right-hand side. We will use this overloaded notation throughout our analysis of NGCA, but the meaning of μ_0 should be clear from the context.

7.1.1. *Degree of non-Gaussianity.* The statistical and computational difficulty of estimating $\boldsymbol{V}$ depends on how non-Gaussian ν is. For positive integer $k \geq 2$, order-k NGCA (k-NGCA) refers to instances of NGCA in which the first $k - 1$ moments of ν are identical to a standard Gaussian random variable,

$$\int x^i \nu(\mathrm{d}x) = \mathbb{E} Z^i, \quad Z \sim \mathcal{N}(0, 1),\ \forall i \in [k-1],$$

but the kth moment differs from the corresponding standard Gaussian moment,

$$\left| \int x^k \nu(\mathrm{d}x) - \mathbb{E}[Z^k] \right| = \lambda > 0, \quad Z \sim \mathcal{N}(0, 1).$$

The parameter $\lambda > 0$ is the signal-to-noise ratio for this problem.

7.1.2. *Assumptions on the non-Gaussian component.* The computational lower bounds we prove holds for a broad class of non-Gaussian distributions ν that have a density with respect to the standard Gaussian measure $\mu_0 = \mathcal{N}(0, 1)$ on $\mathbb{R}$, and that satisfy some additional assumptions. Before stating these assumptions, for any probability measure ν on $\mathbb{R}$, we define the coefficients $\hat{\nu}_i$ for any $i \in \mathbb{N}_0$ as follows:

$$\hat{\nu}_i \overset{\text{def}}{=} \mathbb{E}\big[H_i(\eta)\big], \quad \eta \sim \nu.$$

(Recall that $\{H_i\}_{i\in\mathbb{N}_0}$ are the orthonormalized Hermite polynomials.) Note that since $H_0(z) = 1$, we have $\hat{\nu}_0 = 1$. Since we always assume that ν has a density with respect to $\mu_0 = \mathcal{N}(0, 1)$, we can equivalently write

$$\hat{\nu}_i = \mathbb{E}_0\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)H_i(Z)\right], \quad Z \sim \mu_0 = \mathcal{N}(0, 1).$$

Hence, $\hat{\nu}_i$ is the ith Hermite coefficient of the likelihood ratio function $\mathrm{d}\nu/\mathrm{d}\mu_0$. By Plancheral's identity,

$$\mathbb{E}_0\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z) - 1\right)^2\right] = \sum_{i=1}^{\infty} \hat{\nu}_i^2.$$

We now state our assumptions below.

ASSUMPTION 1. Distribution ν satisfies the *moment matching assumption* with parameter $k \in \mathbb{N}$, $k \geq 2$ if

$$\int z^i \nu(\mathrm{d}z) = \int z^i \mu_0(\mathrm{d}z) \quad \forall i \in [k-1].$$

Equivalently, $\hat{\nu}_i = 0$ for any $i \in [k-1]$.

ASSUMPTION 2. Distribution ν satisfies the *bounded signal strength assumption* with parameters (λ, K) for some $\lambda \geq 0$ and $K \geq 0$ if

$$\sum_{i=1}^{\infty} \hat{\nu}_i^2 \leq K^2\lambda^2, \quad Z \sim \mathcal{N}(0, 1).$$

ASSUMPTION 3. Distribution ν satisfies the *locally bounded likelihood ratio assumption* with parameters (λ, K, κ) for some $\lambda \geq 0$, $K \geq 0$ and $\kappa \geq 0$ if

$$\left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) - 1\right| \leq K\lambda\big(1 + |z|\big)^{\kappa} \quad \forall z \in \mathbb{R} \text{ such that } K\lambda\big(1 + |z|\big)^{\kappa} \leq 1.$$

ASSUMPTION 4. Distribution ν satisfies the *minimum signal strength assumption* with parameters (λ, k) for some $\lambda > 0$ and $k \in \mathbb{N}$, $k \geq 2$ if

$$\left|\mathbb{E}_0 Z^k - \int x^k \nu(\mathrm{d}x)\right| = \lambda, \quad Z \sim \mu_0 = \mathcal{N}(0, 1).$$

ASSUMPTION 5. The random vector $\boldsymbol{x} \sim \mu_{\boldsymbol{V}}$ is *sub-Gaussian* with variance proxy ϑ for some $\vartheta \geq 1$.[1]

[1] A random vector $\boldsymbol{w} \in \mathbb{R}^d$ is *sub-Gaussian with variance proxy* v (a.k.a. v *sub-Gaussian*) if $\mathbb{E}[\boldsymbol{w}] = 0$ and $\mathbb{E}[\exp(\langle \boldsymbol{u}, \boldsymbol{w}\rangle)] \leq \exp(v\|\boldsymbol{u}\|^2/2)$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Note that $\boldsymbol{x} \sim \mu_{\boldsymbol{V}}$ is sub-Gaussian with variance proxy ϑ if $\eta \sim \nu$ is sub-Gaussian with variance proxy ϑ.

7.2. *Statistical-computational gap in k-NGCA.* Similar to k-TPCA, the k-NGCA problem exhibits three phases depending on the effective sample size $N\lambda^2$.

Impossible phase. When $N\lambda^2 \ll d$, there is no consistent estimator for the non-Gaussian direction $\boldsymbol{V}$. This follows from standard lower bounds based on Fano's inequality. We refer the reader to the arXiv version [31], Appendix F.2, of this paper for the proof of the information-theoretic lower bound.

Conjectured hard phase. When $d \lesssim N\lambda^2 \ll d^{k/2}$ and $\lambda \lesssim 1$, there is a consistent, but computationally inefficient estimator for the non-Gaussian direction $\boldsymbol{V}$ (provided Assumptions 4 and 5 hold). This estimator is described and analyzed in the arXiv version of this paper [31], Appendix F.3. The lower bounds of Diakonikolas, Kane and Stewart [28] show that SQ algorithms fail to estimate the non-Gaussian direction with polynomially many queries in this regime. This suggests that this regime is the conjectured hard phase for k-NGCA. We provide additional evidence for this using the low-degree likelihood ratio framework of Hopkins [44] in the arXiv version of this paper [31], Appendix F.4. In the situation when the non-Gaussian measure ν is a mixture of Gaussians, similar lower bounds appear in the work of Mao and Wein [48]. Alternatively, low-degree lower bounds for this problem can also be derived from the SQ lower bounds of Diakonikolas, Kane and Stewart [28] by verifying the general conditions proposed by Brennan et al. [15], which ensure equivalence between the low-degree computational model and the SQ model.

Easy phase. When $N\lambda^2 \gtrsim d^{k/2}$, there are polynomial-time estimators for k-NGCA. In the arXiv version of this paper [31], Appendix F.5, we study a spectral estimator for k-NGCA (with even k) that estimates the non-Gaussian direction $\boldsymbol{V}$ by the leading eigenvector $\hat{\boldsymbol{V}}$ (in the magnitude) of a data-dependent matrix $\hat{\boldsymbol{M}}$:

$$\hat{\boldsymbol{M}} \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}(\|\boldsymbol{x}_i\|^2 - d)^{\frac{k-2}{2}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}} - \mathbb{E}[(\|\boldsymbol{z}\|^2 - d)^{\frac{k-2}{2}}\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}], \tag{10a}$$

$$\hat{\boldsymbol{V}} \overset{\text{def}}{=} \max_{\|\boldsymbol{u}\|=1} |\boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{M}}\boldsymbol{u}|. \tag{10b}$$

In the above display, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. When $N\lambda^2 \gg d^{\frac{k}{2}}$, we show that $\hat{\boldsymbol{V}}$ is a consistent estimator for the non-Gaussian direction (provided Assumptions 4 and 5 hold). This estimator generalizes spectral estimators proposed in prior work of Mao and Wein [48] and Davis, Diaz and Wang [21] for the special case $k = 4$.

REMARK 2 (Lattice and sum-of-squares algorithms for k-NGCA). When the non-Gaussian measure is discrete or close to discrete, Diakonikolas and Kane [24] and Zadik et al. [66] have designed estimators for the non-Gaussian direction, which use $N = d + 1$ samples and run in polynomial-time. In contrast, Davis, Diaz and Wang [21] leverage the results of Ghosh et al. [38] show that estimators based on sum-of-squares relaxations fail to solve these instances when $N \ll d^{3/2}$. Since we assume that the non-Gaussian distribution ν has a density with respect to $\mathcal{N}(0, 1)$ and the signal-to-noise ratio $\lambda \lesssim 1$ as $d \to \infty$, these estimators are not applicable to the instances of k-NGCA studied in this paper.

7.3. *Connections to other inference problems.* By considering particular families of the non-Gaussian distribution ν, we can relate k-NGCA to other inference problems. In the Supplementary Material [32], Appendix G, we provide two constructions of the non-Gaussian distribution ν and leverage them to obtain computational lower bounds for learning Gaussian mixture models and generalized linear models with binary responses as corollaries of Theorem 3.

7.4. *Computational lower bound.* The following is our lower bound for k-NGCA.

THEOREM 3. *Consider the k-NGCA problem with non-Gaussian distribution ν satisfying*:

1. *the moment matching assumption* (*Assumption* 1) *with parameter $k \geq 2$ and $k \asymp 1$*;
2. *the bounded signal strength assumption* (*Assumption* 2) *with parameters* $(\lambda, K \asymp 1)$;
3. *the locally bounded likelihood ratio assumption* (*Assumption* 3) *with parameters* $(\lambda, K \asymp 1, \kappa \asymp 1)$.

Suppose that $\lambda \asymp d^{-\gamma}$ (as $d \to \infty$) for any constant $\gamma > 2\lceil (k+1)/2 \rceil + \kappa$. Let $\hat{\boldsymbol{V}} \in \mathbb{R}^d$ denote any estimator for this k-NGCA problem that can be computed using a memory bounded estimation algorithm with resource profile (N, T, s) scaling with d as

$$N\lambda^2 \asymp d^\eta, \qquad T \asymp d^\tau, \qquad s \asymp d^\mu$$

for any constants $\eta \geq 1$, $\tau \geq 0$, $\mu \geq 0$. If

$$\eta + \tau + \mu < \left\lceil \frac{k+1}{2} \right\rceil,$$

then, for any $t \in \mathbb{R}$,

$$\limsup_{d \to \infty} \inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d} \right) \leq 2\exp\left(-\frac{t^2}{2}\right).$$

Theorem 3 shows that if the signal-to-noise ratio λ is sufficiently small, then memory bounded estimation algorithms using too few total resources (as measured by the product $N\lambda^2 \cdot T \cdot s$) perform no better than a random guess.

7.5. *Discussion of Theorem* 3. Theorem 3 is quantitatively similar to the computational lower bound obtained for k-TPCA (modulo the condition on the signal-to-noise ratio), so most of the implications discussed in Section 5.4 continue to hold. This includes the following (again, just considering even k).

1. Theorem 3 gives a nearly tight lower bound on the total resources, as evidenced by the existence of the spectral estimator from (10) that can be implemented by a memory bounded estimation algorithm with resource profile $(N \asymp d^{k/2} \cdot \mathrm{polylog}(d)/\lambda^2, T \asymp \mathrm{polylog}(d), s \asymp d \cdot \mathrm{polylog}(d))$.
2. The run-time versus sample-size trade-offs for (nearly) linear memory estimators, shown in Figure 2 for k-TPCA, also applies to k-NGCA. Nearly linear memory estimators for k-NGCA include the spectral estimator from (10), the tensor power method on the empirical order-k moment tensor and gradient descent on natural nonconvex objectives [21, 63].
3. For many nearly linear memory algorithms, stronger iteration lower bounds can be obtained using Theorem 3 in the low signal-to-noise regime as compared to the low-degree likelihood ratio framework [44, 47], which only yields lower bounds of the form $T \gtrsim \log(d)$.

REMARK 3. The computational lower bound of Theorem 3 applies only when the signal-to-noise ratio λ^2 is sufficiently small. This requirement is an inherent limitation of the proof technique, which derives a lower bound for memory bounded estimation algorithms from a communication lower bound for distributed estimation algorithms. The Supplementary Material [32], Appendix F, Remark F.1, discusses a simple distributed estimation algorithm that rules out the required communication lower bound in the high signal-to-noise ratio regime.

8. Proof framework.

In this section, we present a general framework used to obtain the computational lower bounds presented in Theorems 1, 2 and 3, as well as the computational lower bounds for the canonical correlation analysis problem in the Supplementary Material [32], Appendix H.

8.1. *Reduction to distributed estimation.*

Although our primary focus is on proving lower bounds for memory-bounded estimation algorithms, these lower bounds are consequences of communication lower bounds for distributed estimation protocols in the "blackboard" model of communication, introduced next.

DEFINITION 2 (Distributed estimation protocol with parameters (m, n, b)). A *distributed estimation protocol* with parameters (m, n, b) computes an estimator based on a data set $\{\boldsymbol{x}_{i,j} \in \mathcal{X} : i \in [m], j \in [n]\}$ of $N = mn$ samples that are distributed across m machines, with n samples $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,j} : j \in [n]\}$ per machine, after each machine writes at b bits to a (public) blackboard. The execution of the protocol occurs in a sequence of mb rounds; a single bit is written on the blackboard per round. In round t: (1) a machine $\ell_t \in [m]$ is chosen as a function of the current contents of the blackboard $\boldsymbol{Y}_{<t} = (Y_1, Y_2, \dots, Y_{t-1}) \in \{0, 1\}^{t-1}$; then, (2) machine ℓ_t computes a Boolean function of the local data set $\boldsymbol{X}_{\ell_t}$ stored on machine ℓ_t, as well as the current contents of the blackboard $\boldsymbol{Y}_{<t}$; and finally, (3) the output $Y_t \in \{0, 1\}$ of the function computed by machine ℓ_t is then written on the blackboard. Each machine is chosen in b rounds. At the end of the mb rounds, the estimator is computed as a function of the final contents of the blackboard $\boldsymbol{Y} \in \{0, 1\}^{mb}$. A general template for a distributed estimation protocol is shown in Figure 4.

The connection between the memory bounded computational model and the distributed computational model is encapsulated in Fact 1, below, formalized by Shamir [58] and Dagan and Shamir [20]. It is a consequence of a simple reduction of Alon, Matias and Szegedy [2] that simulates a memory bounded estimation algorithm using a distributed estimation protocol: the machines take turns to simulate the algorithm's passes over the data set, with one machine concluding its turn by writing the memory state on the blackboard so the next machine can continue the simulation.

FACT 1 ([2, 20, 58]). *A memory bounded estimation algorithm with resource profile (N, T, s) can be simulated using a distributed estimation protocol with parameters $(N/n, n, sT)$ for any $n \in \mathbb{N}$ such that $N/n \in \mathbb{N}$.*

We rely on Fact 1 to convert lower bounds for distributed estimation protocols to lower bounds for memory bounded estimation algorithms. Note that in the reduction, there is some

```
Distributed estimation protocol with parameters $(m, n, b)$.
Input: $\{\boldsymbol{x}_{i,j} : i \in [m],\ j \in [n]\}$, a data set of $N = mn$ samples distributed across $m$ machines, with machine $i$ receiving $n$ samples $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,j} : j \in [n]\}$.
Output: An estimator $\hat{\boldsymbol{V}} \in \widehat{\mathcal{V}}$.
Variables: Contents of the blackboard $\boldsymbol{Y} \in \{0,1\}^{mb}$.

• For round $t \in \{1, 2, \dots, mb\}$
    – Select machine $\ell_t \in [m]$, a function of $\boldsymbol{Y}_{<t}$.
    – Machine $\ell_t$ writes bit $Y_t$, a function of $(\boldsymbol{X}_{\ell_t}, \boldsymbol{Y}_{<t})$, on the blackboard.
• Return estimator $\hat{\boldsymbol{V}}$, a function of $\boldsymbol{Y}$.
```

FIG. 4. *Template for distributed estimation protocols with parameters (m, n, b).*

flexibility in the choice of n, the number of samples per machine. In our use of Fact 1, we will set n in a way that gives us the most interesting lower bounds.

REMARK 4 (Deterministic versus randomized distributed estimation protocols). In Definition 2, we have defined distributed estimation protocols to be deterministic, so they do not use any additional randomness apart from the data set. However, all of our lower bounds for deterministic protocols also apply to randomized protocols, in which all computations (of the ℓ_t's, Y_t's and $\hat{\boldsymbol{V}}$) are permitted to additionally depend on a (shared) uniformly random bit vector $\boldsymbol{q} \in \{0,1\}^R$. This is because, to rule out (ϵ, δ)-accurate distributed estimators, we study the Bayesian version of the inference problem, in which the parameter is drawn from a prior $\boldsymbol{V} \sim \pi$. In the Bayesian problem, there is no advantage of using a randomized protocol: one can always use the deterministic protocol corresponding to the bit vector $\boldsymbol{q}$ that achieves the lowest Bayes risk (averaged over the realization of $\boldsymbol{V} \sim \pi$). This deterministic protocol is guaranteed to perform as well as the original randomized protocol.

8.2. *Lower bounds for distributed estimation protocols.* As a consequence of the reduction from memory bounded estimation to communication bounded estimation, we focus our attention on proving lower bounds for distributed estimation protocols. We introduce a general lower bound technique for showing that if an estimator $\hat{\boldsymbol{V}}$ is computed by a distributed estimation protocol using insufficiently-many resource (as measured by the parameters (m, n, b)), then it is not (ϵ, δ)-accurate (for suitable choices of ϵ and δ):

$$\sup_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\big(\ell(\boldsymbol{V}, \hat{\boldsymbol{V}}) \geq \epsilon\big) \geq \delta.$$

To show this, we consider the Bayesian (a.k.a. average-case) version of the statistical inference problem, in which nature draws the parameter $\boldsymbol{V}$ from a prior π on the parameter space $\mathcal{V}$. Since

$$\sup_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\big(\ell(\boldsymbol{V}, \hat{\boldsymbol{V}}) \geq \epsilon\big) \geq \int \mathbb{P}_{\boldsymbol{V}}\big(\ell(\boldsymbol{V}, \hat{\boldsymbol{V}}) \geq \epsilon\big)\pi(\mathrm{d}\boldsymbol{V}),$$

it is enough to show that the RHS of the above display is at least δ. In order to do so, we will rely on Fano's inequality for Hellinger Information [19], which we introduce next.

8.3. *Hellinger information and Fano's inequality.* Recall that in a statistical inference problem, the N samples $\{\boldsymbol{x}_{i,j} : i \in [m], j \in [n]\} \subset \mathcal{X}$ are drawn i.i.d. from $\mu_{\boldsymbol{V}}$. In the present distributed setting, the samples are distributed across $m = N/n$ machines, with n samples per machine. The data set at machine i is denoted by $\boldsymbol{X}_i \in \mathcal{X}^n$. The machines then communicate via a distributed estimation protocol to write a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ on the blackboard; the final estimator $\hat{\boldsymbol{V}}$ is only a function of $\boldsymbol{Y}$. Let $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m})$ denote the conditional probability that the final transcript is $\boldsymbol{y} \in \{0,1\}^{mb}$ given the data sets $\boldsymbol{X}_{1:m}$. Now define

$$\mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_i) \stackrel{\mathrm{def}}{=} \prod_{j=1}^{n} \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{x}_{i,j}), \tag{11}$$

$$\mathbb{P}_{\boldsymbol{V}}(\boldsymbol{Y} = \boldsymbol{y}) \stackrel{\mathrm{def}}{=} \int \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m}) \cdot \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_1) \cdot \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_2) \cdots \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_m). \tag{12}$$

In words, $\mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_i)$ and $\mathbb{P}_{\boldsymbol{V}}(\boldsymbol{Y} = \boldsymbol{y})$ are respectively the marginal laws of $\boldsymbol{X}_i$ (the data set at machine i) and the blackboard transcript $\boldsymbol{Y}$ when the parameter picked by nature is $\boldsymbol{V}$. We compare two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\{0,1\}^{mb}$ using the *squared Hellinger distance*, defined by

$$d_{\mathsf{hel}}^2(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2} \sum_{\boldsymbol{y} \in \{0,1\}^{mb}} \big(\sqrt{\mathbb{P}_1(\boldsymbol{Y} = \boldsymbol{y})} - \sqrt{\mathbb{P}_2(\boldsymbol{Y} = \boldsymbol{y})}\big)^2.$$

With these preliminary definitions in place, we now define the *Hellinger Information between the parameter* $\boldsymbol{V}$ *and the blackboard transcript* $\boldsymbol{Y}$ by

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y}) \overset{\text{def}}{=} \inf_{\mathbb{Q}} \int d^2_{\mathsf{hel}}(\mathbb{P}_{\boldsymbol{V}},\mathbb{Q})\pi(\mathrm{d}\boldsymbol{V}), \tag{13}$$

where the infimum is taken over all probability measures on $\{0,1\}^{mb}$.

Fano's inequality for Hellinger Information (due to Chen, Guntuboyina and Zhang [19]) provides a lower bound on the error of any estimator for $\boldsymbol{V}$ based on the transcript $\boldsymbol{Y}$ in terms of the Hellinger Information $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$ between $\boldsymbol{V}$ and $\boldsymbol{Y}$.

FACT 2 (Fano's inequality for Hellinger Information [19]). *Let* $\ell : \mathcal{V}\times\widehat{\mathcal{V}} \to \{0,1\}$ *be an arbitrary* 0-1 *loss. Let* π *be an arbitrary prior on* $\mathcal{V}$. *Define*

$$R_0(\pi) \overset{\text{def}}{=} \min_{\boldsymbol{u}\in\widehat{\mathcal{V}}}\left(\int_{\mathcal{V}} \ell(\boldsymbol{V},\boldsymbol{u})\pi(\mathrm{d}\boldsymbol{V})\right).$$

Then, for any estimator $\hat{\boldsymbol{V}} : \{0,1\}^{mb} \to \widehat{\mathcal{V}}$, *we have*

$$\int_{\mathcal{V}} \mathbb{E}_{\boldsymbol{V}}\big[\ell(\boldsymbol{V},\hat{\boldsymbol{V}}(\boldsymbol{Y}))\big]\,\pi(\mathrm{d}\boldsymbol{V}) \geq R_0(\pi) - \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}.$$

In the above display, $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$ *denotes the Hellinger information between the random variables*: $\boldsymbol{V}\sim\pi$ *and* $\boldsymbol{Y}\sim\mathbb{P}_{\boldsymbol{V}}$.

PROOF. The above claim is a minor modification of a result proved by Chen, Guntuboyina and Zhang [19], Corollary 7, item (ii). We provide a derivation in the Supplementary Material [32], Appendix I.3, for completeness. □

Note that $R_0(\pi)$ is the lowest possible estimation error when the transcript $\boldsymbol{Y}$ is not observed. The above fact says that $\sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}$ is an upper bound on the reduction in estimation error possible by leveraging information contained in the transcript $\boldsymbol{Y}$. Since we wish to lower bound

$$\int \mathbb{P}_{\boldsymbol{V}}(\ell(\boldsymbol{V},\hat{\boldsymbol{V}}) \geq \epsilon)\pi(\mathrm{d}\boldsymbol{V}),$$

we will apply Fano's inequality with the 0-1 loss $\widetilde{\ell}$ defined as follows:

$$\widetilde{\ell}(v,\hat{v}) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } \ell(v,\hat{v}) < \epsilon, \\ 1 & \text{if } \ell(v,\hat{v}) \geq \epsilon. \end{cases}$$

8.4. *Information bound for distributed estimation protocols.* Next, we present a general upper bound on $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$ for distributed estimation protocols.

PROPOSITION 1. *Let*:

1. π *be a prior distribution on the parameter space* $\mathcal{V}$;
2. μ_0 *be a reference probability measure on* $\mathcal{X}^n$ *such that* $\mu_{\boldsymbol{V}} \ll \mu_0$ *for all* $\boldsymbol{V}\in\mathcal{V}$;
3. $\overline{\mu}$ *be a null probability measure on* $\mathcal{X}^n$ *such that* $\overline{\mu}$ *and* μ_0 *are mutually absolutely continuous*;
4. $\mathcal{Z}\subset\mathcal{X}^n$ *be an event such that*

$$\left\{\boldsymbol{X}\in\mathcal{X}^n : \left|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - 1\right| \leq \frac{1}{2}\right\} \subset \mathcal{Z},$$

and let Z_i *for* $i\in[m]$ *be the indicator random variables defined by* $Z_i \overset{\text{def}}{=} \mathbb{I}_{\boldsymbol{X}_i\in\mathcal{Z}}$.

Consider a hypothetical setup in which machine $i \in [m]$ is exceptional, and the data $\boldsymbol{X}_{1:m}$ are sampled independently as follows:

$$(\boldsymbol{X}_j)_{j\neq i} \overset{\text{i.i.d.}}{\sim} \overline{\mu}, \qquad \boldsymbol{X}_i \sim \mu_0.$$

Let $\overline{\mathbb{P}}_0^{(i)}$ and $\overline{\mathbb{E}}_0^{(i)}$ denote the probabilities and expectations in this setup:

$$\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\,\mu_0(\mathrm{d}\boldsymbol{X}_i)\cdot\prod_{j\neq i}\overline{\mu}(\mathrm{d}\boldsymbol{X}_j),$$

$$\overline{\mathbb{E}}_0^{(i)} f(\boldsymbol{X}_{1:m},\boldsymbol{Y}) \overset{\text{def}}{=} \int \sum_{\boldsymbol{y}\in\{0,1\}^{mb}} f(\boldsymbol{X}_{1:m},\boldsymbol{y})\,\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\,\mu_0(\mathrm{d}\boldsymbol{X}_i)\cdot\prod_{j\neq i}\overline{\mu}(\mathrm{d}\boldsymbol{X}_j).$$

Also, let $\overline{\mathbb{E}}_0^{(i)}[\cdot|\cdot]$ denote conditional expectations in this setup.

There is a universal constant K such that, if $\boldsymbol{V}\sim\pi$, $\boldsymbol{X}_{1:m} \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}$ and $\boldsymbol{Y}$ is the transcript produced by a distributed estimation protocol with parameters (m,n,b), then

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y}) \leq K\sum_{i=1}^m \overline{\mathbb{E}}_0^{(i)}\left[Z_i\int\left(\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)-\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle|\boldsymbol{Y},Z_i,(\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V})\right]$$
$$+\frac{mK}{2}\left(\int\mu_{\boldsymbol{V}}(\mathcal{Z}^c)\pi(\mathrm{d}\boldsymbol{V})+\overline{\mu}(\mathcal{Z}^c)\right).$$

PROOF. The proof of this result is presented in the Supplementary Material [32], Appendix B.1. □

In order to apply Proposition 1, one needs to suitably choose the reference measure μ_0, null measure $\overline{\mu}$ and the event $\mathcal{Z}$. The considerations involved in these choices are as follows:

1. The reference measure μ_0 is chosen so that it is easy to analyze the concentration behavior of the following likelihood ratios when $\boldsymbol{X}\sim\mu_0$:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}), \qquad \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}).$$

Typically, μ_0 will be the standard Gaussian measure over $\mathcal{X}^n$.

2. We will often set the null measure $\overline{\mu}=\mu_0$. However, in some cases, we will be able to obtain improved lower bounds with the following choice:

$$\overline{\mu}(\cdot)=\int\mu_{\boldsymbol{V}}(\cdot)\,\pi(\mathrm{d}\boldsymbol{V}),$$

which is the marginal law of the data set in a single machine after integrating out $\boldsymbol{V}\sim\pi$.

3. Finally, we will typically set $\mathcal{Z}$ minimally as

$$\mathcal{Z}=\left\{\boldsymbol{X}\in\mathcal{X}^n:\left|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})-1\right|\leq\frac{1}{2}\right\}.$$

However, in some cases, we will find it helpful to enrich $\mathcal{Z}$ with other high probability events that facilitate the analysis of (our upper bound on) Hellinger information.

8.5. *Linearization.* To use our upper bound on Hellinger information, we develop upper bounds on

$$\Psi_i^2(\boldsymbol{y},z_i,(\boldsymbol{X}_j)_{j\neq i}) \overset{\text{def}}{=} \int\left(\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)-\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y},Z_i=z_i,(\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V}).$$

A useful technique to control $\Psi_i^2(\boldsymbol{y},z_i,(\boldsymbol{x}_j)_{j\neq i})$ is linearization, described below.

LEMMA 1 (Linearization). *We have*

$$\Psi\big(\boldsymbol{y}, z_i, (\boldsymbol{X}_j)_{j\neq i}\big) = \sup_{\substack{S:\mathcal{V}\to\mathbb{R}\\ \|S\|_\pi\leq 1}} \overline{\mathbb{E}}_0^{(i)}\left[\left\langle \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i), S\right\rangle_\pi \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right],$$

where $\|\cdot\|_\pi$ *and* $\langle\cdot,\cdot\rangle_\pi$ *denote the* L_2 *norm and inner product with respect to the prior* π:

$$\|S\|_\pi^2 = \int S^2(\boldsymbol{V})\pi(\mathrm{d}\boldsymbol{V}),$$

$$\left\langle \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i), S\right\rangle_\pi = \int S(\boldsymbol{V})\cdot\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right)\pi(\mathrm{d}\boldsymbol{V}).$$

PROOF. The proof follows from the following identities:

$$\begin{aligned}\Psi\big(\boldsymbol{y}, z_i, (\boldsymbol{X}_j)_{j\neq i}\big) &\overset{\text{(a)}}{=} \left\| \overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|_\pi \\ &\overset{\text{(b)}}{=} \sup_{\substack{S:\mathcal{V}\to\mathbb{R}\\ \|S\|_\pi\leq 1}} \left\langle S, \overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\rangle_\pi \\ &\overset{\text{(c)}}{=} \sup_{\substack{S:\mathcal{V}\to\mathbb{R}\\ \|S\|_\pi\leq 1}} \overline{\mathbb{E}}_0^{(i)}\left[\left\langle S, \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right\rangle_\pi \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right].\end{aligned}$$

In the step marked (a), we used the definition of Ψ and $\|\cdot\|_\pi$. In the step marked (b), we used the Cauchy–Schwarz inequality (and its tightness condition). In the step marked (c), we used Fubini's theorem to move the inner product $\langle\cdot,\cdot\rangle_\pi$ inside the conditional expectation. □

8.6. *Geometric inequalities.* In order to upper bound

$$\left(\overline{\mathbb{E}}_0^{(i)}\left[\left\langle S, \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right\rangle_\pi \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2, \tag{14}$$

we will use the framework of geometric inequalities introduced by Han, Özgür and Weissman [40], which shows that the task of upper bounding (14) can be reduced to the task of understanding the concentration properties of the following function when $\boldsymbol{X} \sim \mu_0$:

$$f(\boldsymbol{X}_i) = \left\langle S, \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right\rangle_\pi.$$

Similar results were known in the concentration of measure literature prior to the work of Han, Özgür and Weissman under the name 'transportation lemma" (see, e.g., [12], Lemma 4.18). This result has also been used in other works studying communication lower bounds for distributed estimation [1, 5]. The following proposition summarizes this technique.

PROPOSITION 2 (Boucheron, Lugosi and Massart [12], Han, Özgür and Weissman [40]). *Let* $f : \mathcal{X}^n \to \mathbb{R}$ *be given, and consider* $\boldsymbol{X} \sim \mu_0$.

1. *For any* $\xi > 0$,

$$\begin{aligned}&\big|\overline{\mathbb{E}}_0^{(i)}[f(\boldsymbol{X}_i)|\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}]\big| \\ &\quad\leq \frac{\log(\mathbb{E}_0[e^{\xi f(\boldsymbol{X})}] \vee \mathbb{E}_0[e^{-\xi f(\boldsymbol{X})}])}{\xi} + \frac{1}{\xi}\log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i})}.\end{aligned}$$

2. *For any* $q \geq 1$,

$$\left|\overline{\mathbb{E}}_0^{(i)}[f(\boldsymbol{X}_i)|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z_i, (\boldsymbol{X}_j)_{j\neq i}]\right| \leq \left(\frac{\mathbb{E}_0|f(\boldsymbol{X})|^q}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z_i|(\boldsymbol{X}_j)_{j\neq i})}\right)^{\frac{1}{q}}.$$

PROOF. The proof of this result is presented in the Supplementary Material [32], Appendix B.2. □

We have now introduced all the key elements of our lower bound framework.

8.7. *Comparison to prior works.* Recent works by Han, Özgür and Weissman [40], Barnes, Han and Ozgur [5], Acharya et al. [1] have developed general frameworks to obtain communication lower bounds for distributed statistical inference problems. The general information bounds developed in these works yield lower bounds for the simpler "hide-and-seek" variant of the inference problems [58]. In the hide-and-seek variant, the statistician knows the entire parameter vector $\boldsymbol{V} \in \{\pm 1\}^d$ except for a single coordinate, hidden at an unknown index $i \in [d]$. The goal is to infer the sign of the hidden coordinate.

A hide-and-seek version inference problem can always be solved in the distributed setting with the information-theoretic sample complexity as long as each machine is allowed to communicate at least $\Omega(d \cdot \text{polylog}(d))$ bits. To see this, note that because the statistician knows the entire parameter vector except for a single coordinate hidden at an unknown index $i \in [d]$, the possible parameter space for the inference problem is a discrete set of size $2d$—there are d possibilities for the index of the unknown coordinate, and two possibilities for the sign of the unknown coordinate. Hence, each machine $j \in [m]$ can transmit the likelihoods of all $2d$ elements of this discrete set given its own data set $\boldsymbol{X}_j$, using $O(d \cdot \text{polylog}(d))$ bits of communication. These likelihoods can be aggregated (by taking their product) to obtain the likelihoods given *all* of the data; this is a sufficient statistic for any inference problem.

Since the information bounds developed in the previously mentioned works [1, 5, 40] apply to the hide-and-seek variant of inference problems, we are unable to use them directly to obtain nontrivial lower bounds for k-TPCA and k-NGCA in the regime where the problems are information-theoretically solvable ($N = m \cdot n \gtrsim d$) and each machine is allowed at least $b \gtrsim d \cdot \text{polylog}(d)$ bits of communication. The information bound in Proposition 1 builds on these works to address this limitation.

9. Proof of computational lower bound for tensor PCA (Theorem 1).

As an illustration, we instantiate the framework introduced in Section 8 to obtain the computational lower bound for Symmetric Tensor PCA (k-TPCA) claimed in (Theorem 1). The computational lower bounds for the other inference problems studied in this paper are obtained by following the same recipe and their detailed proofs appear in the Supplementary Material [32].

The computational lower bound for k-TPCA (Theorem 1) is obtained by transferring a communication lower bound for distributed estimation protocols for k-TPCA to memory bounded estimators for the same problem using the reduction in Fact 1.

In the (Bayesian) distributed setup for k-TPCA, the parameter $\boldsymbol{V}$ is drawn from the prior $\pi \stackrel{\text{def}}{=} \mathsf{Unif}(\{\pm 1\}^d)$, and then $\boldsymbol{X}_{1:N}$ are sampled i.i.d. from $\mu_{\boldsymbol{V}}$; these tensors are distributed across $m = N$ machines with $n = 1$ sample/machine. The execution of a distributed estimation protocol with parameters $(m, n = 1, b)$ results in a transcript $\boldsymbol{Y} \in \{0, 1\}^{mb}$ written on the blackboard.

Instantiating Fano's inequality (Fact 2) in the context of k-TPCA (the Supplementary Material [32], Appendix C.2, contains a detailed derivation) shows that for any distributed estimator $\hat{\boldsymbol{V}}(\boldsymbol{Y})$:

$$\inf_{\boldsymbol{V}\in\mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}}(\boldsymbol{Y})\rangle|^2}{\|\boldsymbol{V}\|^2\|\hat{\boldsymbol{V}}(\boldsymbol{Y})\|^2} \geq \frac{t^2}{d}\right) \leq 2\exp\left(-\frac{t^2}{2}\right) + \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}. \tag{15}$$

The main technical result needed to prove Theorem 1 is the following bound on $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$ for k-TPCA in Proposition 3. We obtain this result by instantiating our general information bound (Proposition 1) and controlling the resulting upper bound using the linearization technique (Lemma 1) and the geometric inequalities stated in Proposition 2. Applying the geometric inequalities, in turn, requires sharp variance and concentration estimates for nonlinear functions of Gaussian random variables derived from the likelihood ratio for this model, which we obtain by exploiting the Hermite decomposition of the likelihood ratio.

PROPOSITION 3 (Information bound for k-TPCA). *Let $\boldsymbol{Y} \in \{0,1\}^{mb}$ be the transcript generated by a distributed estimation protocol for k-TPCA with parameters $(m, 1, b)$. Then*

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$$
$$\leq C_k\left(\sigma^2 \cdot m \cdot b + \frac{1}{d} + \lambda^2 \cdot b \cdot \left(\frac{\lambda^2 \vee \log(m\cdot d)}{d}\right)^{\frac{k}{2}} + \inf_{\alpha\geq 2} \frac{\lambda^2\alpha}{d} + m \cdot \left(\frac{C_k\alpha\lambda^2}{\sqrt{d^k}} + e^{-d}\right)^{\frac{\alpha}{2}}\right),$$

where

$$\sigma^2 \overset{\text{def}}{=} \begin{cases} C_k \cdot \lambda^2 \cdot d^{-\frac{k+2}{2}} & \text{if } k \text{ is even}, \\ C_k \cdot \lambda^2 \cdot d^{-\frac{k+1}{2}} & \text{if } k \text{ is odd}; \end{cases}$$

and $C_k > 0$ is a positive constant that depends only on k. In particular, in the scaling regime (as $d \to \infty$)

$$\lambda \asymp 1, \qquad m \asymp d^\eta, \qquad b \asymp d^\beta$$

for any constants $\eta \geq 1$ and $\beta \geq 0$ that satisfy

$$\eta + \beta < \left\lceil \frac{k+1}{2} \right\rceil,$$

we have $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y}) \to 0$ as $d \to \infty$.

Proposition 3 is proved in the Supplementary Material [32], Appendix C. With this information bound in hand, we can complete the proof of Theorem 1.

PROOF OF THEOREM 1. Appealing to the reduction in Fact 1 with the choice $n = 1$, we note that any memory bounded estimator $\hat{\boldsymbol{V}}$ with resource profile (N, T, s) can be implemented using a distributed estimation protocol with parameters $(N, 1, sT)$. Applying Fano's inequality (15) and Proposition 3 to the distributed implementation of the memory-bounded estimator immediately yields Theorem 1. □

Acknowledgments. This project was done when RD was a graduate student at Columbia University and a postdoctoral scholar at Harvard University.

Funding. DH acknowledges support from NSF awards CCF-1740833 and IIS-1563785 and a JP Morgan Faculty Award.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical-computational trade-offs in tensor PCA and related problems via communication complexity" (DOI: 10.1214/23-AOS2331SUPP; .pdf). This supplement provides computational lower bounds for the higher-order canonical correlation analysis problem, along with the complete proofs and some additional discussion of the results presented in the paper.

REFERENCES

[1] ACHARYA, J., CANONNE, C. L., SUN, Z. and TYAGI, H. (2020). Unified lower bounds for interactive high-dimensional estimation under information constraints. Preprint. Available at arXiv:2010.06562.

[2] ALON, N., MATIAS, Y. and SZEGEDY, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** 137–147. MR1688610 https://doi.org/10.1006/jcss.1997.1545

[3] ANANDKUMAR, A., DENG, Y., GE, R. and MOBAHI, H. (2017). Homotopy analysis for tensor PCA. In *Conference on Learning Theory* 79–104.

[4] BANDEIRA, A. S., PERRY, A. and WEIN, A. S. (2018). Notes on computational-to-statistical gaps: Predictions using statistical physics. *Port. Math.* **75** 159–186. MR3892753 https://doi.org/10.4171/PM/2014

[5] BARNES, L. P., HAN, Y. and ÖZGÜR, A. (2020). Lower bounds for learning distributions under communication constraints via Fisher information. *J. Mach. Learn. Res.* **21** Paper No. 236, 30 pp. MR4209522

[6] BEAME, P., GHARAN, S. O. and YANG, X. (2018). Time-space tradeoffs for learning finite functions from random evaluations, with applications to polynomials. In *Conference on Learning Theory* 843–856. PMLR.

[7] BEN AROUS, G., GHEISSARI, R. and JAGANNATH, A. (2020). Algorithmic thresholds for tensor PCA. *Ann. Probab.* **48** 2052–2087. MR4124533 https://doi.org/10.1214/19-AOP1415

[8] BEN AROUS, G., MEI, S., MONTANARI, A. and NICA, M. (2019). The landscape of the spiked tensor model. *Comm. Pure Appl. Math.* **72** 2282–2330. MR4011861 https://doi.org/10.1002/cpa.21861

[9] BHATTIPROLU, V., GURUSWAMI, V. and LEE, E. (2017). Sum-of-squares certificates for maxima of random tensors on the sphere. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. LIPIcs. Leibniz Int. Proc. Inform.* **81** Art. No. 31, 20 pp. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3695598

[10] BIROLI, G., CAMMAROTA, C. and RICCI-TERSENGHI, F. (2019). How to iron out rough landscapes and get optimal performances: Replicated gradient descent and its application to tensor PCA. Preprint. Available at arXiv:1905.12294.

[11] BLANCHARD, G., KAWANABE, M., SUGIYAMA, M., SPOKOINY, V. and MÜLLER, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *J. Mach. Learn. Res.* **7** 247–282. MR2274368

[12] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[13] BRAVERMAN, M., GARG, A., MA, T., NGUYEN, H. L. and WOODRUFF, D. P. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing* 1011–1020. ACM, New York. MR3536632 https://doi.org/10.1145/2897518.2897582

[14] BRENNAN, M. and BRESLER, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory* 648–847. PMLR.

[15] BRENNAN, M., BRESLER, G., HOPKINS, S. B., LI, J. and SCHRAMM, T. (2020). Statistical query algorithms and low-degree tests are almost equivalent. Preprint. Available at arXiv:2009.06107.

[16] BRUNA, J., REGEV, O., SONG, M. J. and TANG, Y. (2021). Continuous LWE. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* 694–707. ACM, New York. MR4398874 https://doi.org/10.1145/3406325.3451000

[17] BUBECK, S., LEE, Y. T., PRICE, E. and RAZENSHTEYN, I. (2019). Adversarial examples from computational constraints. In *International Conference on Machine Learning* 831–840. PMLR.

[18] CELENTANO, M., MONTANARI, A. and WU, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory* 1078–1141. PMLR.

[19] CHEN, X., GUNTUBOYINA, A. and ZHANG, Y. (2016). On Bayes risk lower bounds. *J. Mach. Learn. Res.* **17** Paper No. 219, 58 pp. MR3595153

[20] DAGAN, Y. and SHAMIR, O. (2018). Detecting correlations with little memory and communication. In *Conference on Learning Theory* 1145–1198. PMLR.

[21] DAVIS, D., DIAZ, M. and WANG, K. (2021). Clustering a mixture of Gaussians with unknown covariance. Preprint. Available at arXiv:2110.01602.

[22] DIAKONIKOLAS, I., KANE, D. and ZARIFIS, N. (2020). Near-optimal sq lower bounds for agnostically learning halfspaces and relus under Gaussian marginals. *Adv. Neural Inf. Process. Syst.* **33** 13586–13596.

[23] DIAKONIKOLAS, I. and KANE, D. M. (2020). Hardness of learning halfspaces with Massart noise. Preprint. Available at arXiv:2012.09720.

[24] DIAKONIKOLAS, I. and KANE, D. M. (2021). Non-Gaussian component analysis via lattice basis reduction. Preprint. Available at arXiv:2112.09104.

[25] DIAKONIKOLAS, I., KANE, D. M., KONTONIS, V., TZAMOS, C. and ZARIFIS, N. (2022). Learning general halfspaces with general Massart noise under the Gaussian distribution. In *STOC '22—Proceedings of the* 54*th Annual ACM SIGACT Symposium on Theory of Computing* 874–885. ACM, New York. MR4490047

[26] DIAKONIKOLAS, I., KANE, D. M., KONTONIS, V. and ZARIFIS, N. (2020). Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory* 1514–1539. PMLR.

[27] DIAKONIKOLAS, I., KANE, D. M., PITTAS, T. and ZARIFIS, N. (2021). The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model. In *Conference on Learning Theory* 1552–1584. PMLR.

[28] DIAKONIKOLAS, I., KANE, D. M. and STEWART, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract). In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 73–84. IEEE Computer Soc., Los Alamitos, CA. MR3734219 https://doi.org/10.1109/FOCS.2017.16

[29] DIAKONIKOLAS, I., KONG, W. and STEWART, A. (2019). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2745–2754. SIAM, Philadelphia, PA. MR3909639 https://doi.org/10.1137/1.9781611975482.170

[30] DUDEJA, R. and HSU, D. (2021). Statistical query lower bounds for tensor PCA. *J. Mach. Learn. Res.* **22** Paper No. 83, 51 pp. MR4253776

[31] DUDEJA, R. and HSU, D. (2022). Statistical-computational trade-offs in tensor PCA and related problems via communication complexity. Preprint. Available at arXiv:2204.07526.

[32] DUDEJA, R. and HSU, D. (2024). Supplement to "Statistical-computational trade-offs in tensor PCA and related problems via communication complexity." https://doi.org/10.1214/23-AOS2331SUPP

[33] FELDMAN, V., GRIGORESCU, E., REYZIN, L., VEMPALA, S. S. and XIAO, Y. (2017). Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM* **64** Art. 8, 37 pp. MR3664576 https://doi.org/10.1145/3046674

[34] FELDMAN, V., PERKINS, W. and VEMPALA, S. (2018). On the complexity of random satisfiability problems with planted solutions. *SIAM J. Comput.* **47** 1294–1338. MR3827195 https://doi.org/10.1137/16M1078471

[35] GARG, S., KOTHARI, P. K., LIU, P. and RAZ, R. (2021). Memory-sample lower bounds for learning parity with noise. In 24*th International Conference on Approximation Algorithms for Combinatorial Optimization Problems*, *APPROX* 2021 *and* 25*th International Conference on Randomization and Computation*, *RANDOM* 2021 60. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

[36] GARG, S., RAZ, R. and TAL, A. (2018). Extractor-based time-space lower bounds for learning. In *STOC*'18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 990–1002. ACM, New York. MR3826311

[37] GARG, S., RAZ, R. and TAL, A. (2019). Time-space lower bounds for two-pass learning. In 34*th Computational Complexity Conference*. *LIPIcs. Leibniz Int. Proc. Inform.* **137** Art. No. 22, 39 pp. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3984627

[38] GHOSH, M., JERONIMO, F. G., JONES, C., POTECHIN, A. and RAJENDRAN, G. (2020). Sum-of-squares lower bounds for Sherrington–Kirkpatrick via planted affine planes. In 2020 *IEEE* 61*st Annual Symposium on Foundations of Computer Science* 954–965. IEEE Computer Soc., Los Alamitos, CA. MR4232101

[39] GOYAL, N. and SHETTY, A. (2019). Non-Gaussian component analysis using entropy methods. In *STOC*'19*—Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 840–851. ACM, New York. MR4003388

[40] HAN, Y., ÖZGÜR, A. and WEISSMAN, T. (2021). Geometric lower bounds for distributed parameter estimation under communication constraints. *IEEE Trans. Inf. Theory* **67** 8248–8263. MR4346086 https://doi.org/10.1109/TIT.2021.3108952

[41] HOPKINS, S. B., KOTHARI, P. K., POTECHIN, A., RAGHAVENDRA, P., SCHRAMM, T. and STEURER, D. (2017). The power of sum-of-squares for detecting hidden structures. In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 720–731. IEEE Computer Soc., Los Alamitos, CA. MR3734275 https://doi.org/10.1109/FOCS.2017.72

[42] HOPKINS, S. B., SHI, J., SCHRAMM, T. and STEURER, D. (2016). Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 178–191. ACM, New York. MR3536564 https://doi.org/10.1145/2897518.2897529

[43] HOPKINS, S. B., SHI, J. and STEURER, D. (2015). Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory* 956–1006.

[44] HOPKINS, S. B. K. (2018). Statistical inference and the sum of squares method. PhD thesis, Cornell University. MR3864930

[45] KEARNS, M. (1998). Efficient noise-tolerant learning from statistical queries. *J. ACM* **45** 983–1006. MR1678849 https://doi.org/10.1145/293347.293351

[46] KOL, G., RAZ, R. and TAL, A. (2017). Time-space hardness of learning sparse parities. In *STOC'*17*—Proceedings of the* 49*th Annual ACM SIGACT Symposium on Theory of Computing* 1067–1080. ACM, New York. MR3678252 https://doi.org/10.1145/3055399.3055430

[47] KUNISKY, D., WEIN, A. S. and BANDEIRA, A. S. (2022). Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *Mathematical Analysis*, *Its Applications and Computation*. *Springer Proc*. *Math*. *Stat*. **385** 1–50. Springer, Cham. MR4461037 https://doi.org/10.1007/978-3-030-97127-4_1

[48] MAO, C. and WEIN, A. S. (2021). Optimal spectral recovery of a planted vector in a subspace. Preprint. Available at arXiv:2105.15081.

[49] MICCIANCIO, D. and REGEV, O. (2009). Lattice-based cryptography. In *Post-Quantum Cryptography* 147–191. Springer, Berlin. MR2590647 https://doi.org/10.1007/978-3-540-88702-7_5

[50] MONTANARI, A. and RICHARD, E. (2014). A statistical model for tensor PCA. Preprint. Available at arXiv:1411.1076.

[51] MOSHKOVITZ, D. and MOSHKOVITZ, M. (2017). Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory* 1516–1566. PMLR.

[52] MOSHKOVITZ, D. and MOSHKOVITZ, M. (2018). Entropy samplers and strong generic lower bounds for space bounded learning. In 9*th Innovations in Theoretical Computer Science*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **94** Art. No. 28, 20 pp. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3761764

[53] MOSHKOVITZ, M. and TISHBY, N. (2017). Mixing complexity and its applications to neural networks. Preprint. Available at arXiv:1703.00729.

[54] RAGHAVENDRA, P., SCHRAMM, T. and STEURER, D. (2018). High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro* 2018. *Vol*. *IV*. *Invited Lectures* 3389–3423. World Sci. Publ., Hackensack, NJ. MR3966537

[55] RAZ, R. (2019). Fast learning requires good memory: A time-space lower bound for parity learning. *J. ACM* **66** Art. 3, 18 pp. MR3892562 https://doi.org/10.1145/3186563

[56] ROS, V., AROUS, G. B., BIROLI, G. and CAMMAROTA, C. (2019). Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Phys*. *Rev*. *X* **9** 011003.

[57] SCHRAMM, T. and WEIN, A. S. (2022). Computational barriers to estimation from low-degree polynomials. *Ann*. *Statist*. **50** 1833–1858. MR4441142 https://doi.org/10.1214/22-aos2179

[58] SHAMIR, O. (2014). Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Adv*. *Neural Inf*. *Process*. *Syst*. **27**.

[59] SHARAN, V., SIDFORD, A. and VALIANT, G. (2019). Memory-sample tradeoffs for linear regression with small error. In *STOC'*19*—Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 890–901. ACM, New York. MR4003393 https://doi.org/10.1145/3313276.3316403

[60] STEINHARDT, J., VALIANT, G. and WAGER, S. (2016). Memory, communication, and statistical queries. In *Conference on Learning Theory* 1490–1516. PMLR.

[61] TAN, Y. S. and VERSHYNIN, R. (2018). Polynomial time and sample complexity for non-Gaussian component analysis: Spectral methods. In *Conference on Learning Theory* 498–534. PMLR.

[62] VEMPALA, S. S. and XIAO, Y. (2011). Structure from local optima: Learning subspace juntas via higher order PCA. Preprint. Available at arXiv:1108.3329.

[63] WANG, K., YAN, Y. and DIAZ, M. (2020). Efficient clustering for stretched mixtures: Landscape and optimality. *Adv*. *Neural Inf*. *Process*. *Syst*. **33** 21309–21320.

[64] WANG, Z., GU, Q. and LIU, H. (2015). Sharp computational-statistical phase transitions via oracle computational model. Preprint. Available at arXiv:1512.08861.

[65] WEIN, A. S., EL ALAOUI, A. and MOORE, C. (2019). The Kikuchi hierarchy and tensor PCA. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 1446–1468. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228236

[66] ZADIK, I., SONG, M. J., WEIN, A. S. and BRUNA, J. (2021). Lattice-based methods surpass sum-of-squares in clustering. Preprint. Available at arXiv:2112.03898.

[67] ZDEBOROVÁ, L. and KRZAKALA, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Adv. Phys.* **65** 453–552.

[68] ZHANG, A. and XIA, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Trans. Inf. Theory* **64** 7311–7338. MR3876445 https://doi.org/10.1109/TIT.2018.2841377

[69] ZHENG, Q. and TOMIOKA, R. (2015). Interpolating convex and non-convex tensor decompositions via the subspace norm. In *Advances in Neural Information Processing Systems* 3106–3113.

Submitted to the Annals of Statistics

SUPPLEMENT TO STATISTICAL-COMPUTATIONAL TRADE-OFFS IN TENSOR PCA AND RELATED PROBLEMS VIA COMMUNICATION COMPLEXITY

BY RISHABH DUDEJA[1,a], DANIEL HSU[2,b]

[1]Department of Statistics, University of Wisconsin–Madison, [a]rdudeja@wisc.edu

[2]Department of Computer Science, Columbia University, [b]djhsu@cs.columbia.edu

CONTENTS

Organization . . . 2
A Notation . . . 3
B Proofs of the Information Bound and Geometric Inequalities . . . 4
B.1 Proof of Proposition 1 . . . 4
B.1.1 Additional Notation . . . 4
B.1.2 Properties of Distributed Algorithms . . . 5
B.1.3 Proof of Proposition 1 . . . 6
B.2 Proof of Proposition 2 . . . 9
C Proofs for Symmetric Tensor PCA . . . 10
C.1 Setup . . . 10
C.2 Fano's Inequality for Symmetric Tensor PCA . . . 11
C.3 The Likelihood Ratio for Symmetric Tensor PCA . . . 11
C.4 Proof of Information Bound (Proposition 3) . . . 13
C.4.1 Analysis of Low Degree Part . . . 19
C.4.2 Analysis of High Degree Part . . . 21
C.4.3 Analysis of Bad Event . . . 21
C.5 Omitted Proofs from Appendix C.3 . . . 23
C.5.1 Proof of Lemma C.1 . . . 23
C.5.2 Proof of Lemma C.2 . . . 24
C.5.3 Proof of Lemma C.3 . . . 24
C.5.4 Proof of Lemma C.4 . . . 26
D Additional Discussions for Symmetric Tensor PCA . . . 27
D.1 Discussion for k-Tensor PCA with odd k . . . 27
D.2 Linear Memory Algorithms for Symmetric Tensor PCA . . . 29
E Proofs for Asymmetric Tensor PCA . . . 30
E.1 Proof of Computational Lower Bound (Theorem 2) . . . 30
E.2 Proof of the Information Bound (Proposition E.1) . . . 32
E.3 Concentration of Likelihood Ratio . . . 36
F Proofs for Non-Gaussian Component Analysis . . . 37
F.1 Proof of Computational Lower Bound (Theorem 3) . . . 37
F.2 Proof Overview of the Information Bound (Proposition F.1) . . . 39
F.3 The Likelihood Ratio for Non-Gaussian Component Analysis . . . 40
F.4 Proof of the Information Bound (Proposition F.1) . . . 43
F.4.1 Proof of Lemma F.6 . . . 48
F.4.2 Proof of Lemma F.8 . . . 51
F.4.3 Proof of Lemma F.7 . . . 55
F.5 Omitted Proofs from Section F.3 . . . 56
F.5.1 Proof of Lemma F.3 . . . 56

F.5.2 Proof of Lemma F.4 57
F.5.3 Proof of Lemma F.5 58
G Further Results for Non-Gaussian Component Analysis 59
G.1 Connections to Learning Mixtures of Gaussians 59
G.2 Connections to Learning Binary Generalized Linear Models 60
G.3 Constructions of Non-Gaussian Distributions 62
G.3.1 Proof of Lemma G.1 62
G.3.2 Proof of Lemma G.2 65
H Canonical Correlation Analysis 66
H.1 Problem Formulation 66
H.2 Statistical-Computational Gap in k-CCA 67
H.3 Computational Lower Bound for k-CCA 68
H.4 Connections to Learning Parities 68
H.4.1 Comparison to Prior Works 71
H.5 Proof of Computational Lower Bound (Theorem H.1) 72
H.6 Proof of Information Bound (Proposition H.1) 73
I Miscellaneous Results 79
I.1 Additional Technical Facts and Lemmas 79
I.2 Analysis on Gaussian Space 83
I.3 Fano's Inequality for Hellinger Information 85
References 85

Organization. This supplement proves the results presented in the main paper. Throughout the supplement, intermediate results introduced in the supplement are numbered according to the section in which they appear. Results numbered without their section number refer to results from the main paper. For e.g. Fact B.1, Lemma C.1 and Proposition H.1 refer to results introduced in Appendix B, Appendix C, and Appendix H in the supplement. On the other hand, Proposition 1 refers to Proposition 1 from the main paper. The supplement is organized as follows:

1. Appendix A contains a glossary of notations used through out the supplement.
2. Appendix B proves the general information bound stated in Proposition 1 and the Geometric Inequalities from Proposition 2.
3. Appendix C is devoted to the proof of the information bound for Symmetric Tensor PCA (Proposition 3).
4. Appendix D provides come additional discussion for the Symmetric Tensor PCA problem. Appendix D.1 discusses the apparent deficiencies in our lower bounds for odd order Tensor PCA. Appendix D.2 complements Section 5.4.1 in the main paper by providing additional details to explain how many natural algorithms for Symmetric k-Tensor PCA (k-TPCA) fit into the template of (nearly) linear memory iterative algorithms and discusses the consequences of the computational lower bound for k-TPCA (Theorem 1) for these algorithms.
5. Appendix E presents the proof of the computational lower bound Asymmetric k-Tensor PCA (Theorem 2).
6. Appendix F provides the proof of the computational lower bound for the order-k Non-Gaussian Component Analysis problem (Theorem 3)
7. Appendix G shows that computational lower bounds for k-NGCA imply lower bounds for learning Gaussian mixture models and binary generalized linear models. This appendix also provides two constructions for non-Gaussian distributions that satisfy our assumptions.

8. Appendix H introduces the order-k Canonical Correlation Analysis problem (k-CCA), states our computational lower bound for this problem and provides its complete proof.
9. Finally, Appendix I contains some background on Hermite polynomials and the Gaussian Hilbert space along with some additional technical facts and results used in this paper.

APPENDIX A: NOTATION

Important sets. $\mathbb{N}$ and $\mathbb{R}$ denote the set of positive integers and the set of real numbers, respectively. $\mathbb{N}_0 \overset{\text{def}}{=} \mathbb{N} \cup \{0\}$ is the set of non-negative integers. For each $k, d \in \mathbb{N}$, $[k]$ denotes the set $\{1, 2, 3, \dots, k\}$, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space, $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, $\mathbb{R}^{d\times k}$ denotes the set of all $d \times k$ matrices, $\bigotimes^k \mathbb{R}^d$ denotes the set of all $d \times d \times \cdots \times d$ (k times) tensors with $\mathbb{R}$-valued entries, and $\bigotimes^k \mathbb{N}_0^d$ denotes the set of all $d \times d \times \cdots \times d$ (k times) tensors with $\mathbb{N}_0$-valued entries.

Linear Algebra. We denote the d-dimensional vectors $(1, 1, \dots, 1)$, $(0, 0, \dots, 0)$ and the $d \times d$ identity matrix using $\mathbf{1}_d$, $\mathbf{0}_d$, and $\boldsymbol{I}_d$ respectively. We will omit the subscript d when the dimension is clear from the context. The vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_d$ denote the standard basis vectors of $\mathbb{R}^d$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\|, \|\boldsymbol{v}\|_1, \|\boldsymbol{v}\|_\infty$ denote the $\ell_2, \ell_1, \ell_\infty$ norms of $\boldsymbol{v}$, and $\|\boldsymbol{v}\|_0$ denotes the sparsity (number of non-zero entries) of $\boldsymbol{v}$. For two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ denotes the standard inner product on $\mathbb{R}^d$: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \overset{\text{def}}{=} \sum_{i=1}^d u_i v_i$. For two matrices or tensors $\boldsymbol{U}$ and $\boldsymbol{V}$, we analogously define $\|\boldsymbol{U}\|, \|\boldsymbol{U}\|_1, \|\boldsymbol{U}\|_\infty, \|\boldsymbol{U}\|_0$, and $\langle \boldsymbol{U}, \boldsymbol{V} \rangle$ by stacking their entries to form a vector. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^\mathsf{T}$ denotes the transpose of $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_{\mathsf{op}}$ denotes the operator (or spectral) norm of $\boldsymbol{A}$. For a square matrix $\boldsymbol{A}$, $\mathsf{Tr}(\boldsymbol{A})$ denotes the trace of $\boldsymbol{A}$. Finally, for vectors $\boldsymbol{v}_{1:k} \in \mathbb{R}^d$, $\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k$ denotes the k-tensor with entries $(\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k)_{i_1, i_2, \dots, i_k} = (\boldsymbol{v}_1)_{i_1} \cdot (\boldsymbol{v}_2)_{i_2} \cdots (\boldsymbol{v}_k)_{i_k}$ for $i_{1:k} \in [d]$. When $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \cdots = \boldsymbol{v}_k = \boldsymbol{v}$, we shorthand $\boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}$ as $\boldsymbol{v}^{\otimes k}$. Analogously, given two tensors $\boldsymbol{U} \in \bigotimes^\ell \mathbb{R}^d$ and $\boldsymbol{V} \in \bigotimes^m \mathbb{R}^d$, $\boldsymbol{U} \otimes \boldsymbol{V}$ is the $(\ell + m)$-tensor with entries $(\boldsymbol{U} \otimes \boldsymbol{V})_{i_1, i_2, \dots, i_\ell, j_1, j_2, \dots, j_m} = (\boldsymbol{U})_{i_1, i_2, \dots, i_\ell} \cdot (\boldsymbol{V})_{j_1, j_2, \dots, j_m}$ for $i_{1:\ell} \in [d], j_{1:m} \in [d]$. This definition is naturally extended to define the $(\ell_1 + \ell_2 + \cdots + \ell_k)$-tensor $\boldsymbol{U}_1 \otimes \boldsymbol{U}_2 \otimes \cdots \otimes \boldsymbol{U}_k$ for tensors $\boldsymbol{U}_{1:k}$ with $\boldsymbol{U}_i \in \bigotimes^{\ell_i} \mathbb{R}^d$ for each $i \in [k]$.

Asymptotic notation. Given a two non-negative sequences a_d and b_d indexed by $d \in \mathbb{N}$, we use the following notations to describe their relative magnitudes for large d. We say that $a_d \lesssim b_d$ or $a_d = O(b_d)$ or $b_d = \Omega(a_d)$ if $\limsup_{d\to\infty}(a_d/b_d) < \infty$. If $a_d \lesssim b_d$ and $b_d \lesssim a_d$, then we say that $a_d \asymp b_d$. If there exists a constant $\epsilon > 0$ such that $a_d \cdot d^\epsilon \lesssim b_d$ we say that $a_d \ll b_d$. We use $\mathrm{polylog}\,(d)$ to denote any sequence a_d such that $a_d \asymp \log^t(d)$ for some fixed constant $t \geq 0$.

Important distributions. $\mathcal{N}(0, 1)$ denotes the standard Gaussian measure on $\mathbb{R}$, and $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ denotes the standard Gaussian measure on $\mathbb{R}^d$. For any finite set A, $\mathsf{Unif}\,(A)$ denotes the uniform distribution on the elements of A.

Hermite polynomials. We will make extensive use of the Hermite polynomials $\{H_i : i \in \mathbb{N}_0\}$ which are the orthonormal polynomials for the Gaussian measure $\mathcal{N}(0, 1)$ and their multivariate analogs $\{H_{\boldsymbol{c}} : \boldsymbol{c} \in \mathbb{N}_0^d\}$, which are the orthornormal polynomials for the d-dimensional Gaussian measure $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. We provide the necessary background regarding Hermite polynomials and analysis on the Gaussian Hilbert space in Appendix I.2.

Miscellaneous. For an event $\mathcal{E}$, $\mathbb{I}_{\mathcal{E}}$ denotes the indicator random variable for $\mathcal{E}$. For $x, y \in \mathbb{R}$, $x \vee y$ and $x \wedge y$ denote $\max(x, y)$ and $\min(x, y)$, respectively; and $\mathsf{sign}(x)$ denotes the sign function ($\mathsf{sign}(x) = 1$ iff $x > 0$, $\mathsf{sign}(x) = -1$ iff $x < 0$ and $\mathsf{sign}(0) = 0$). For $x > 0$, $\log(x)$ denotes the natural logarithm (base e) of x.

APPENDIX B: PROOFS OF THE INFORMATION BOUND AND GEOMETRIC INEQUALITIES

This appendix presents the proofs of our general information bound (Proposition 1) and the Geometric Inequalities (Proposition 2).

B.1. Proof of Proposition 1. In this section, we present the proof of Proposition 1. This section is organized as follows:

1. In Section B.1.1, we introduce some additional notation used in the proof.
2. In Section B.1.2, we collect some well-known properties of distributed estimation algorithms.
3. In Section B.1.3, we present the actual proof of Proposition 1.

B.1.1. *Additional Notation.* Recall that in the distributed learning setup, the data $\boldsymbol{X}_{1:m} \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}$. We use $\mathbb{P}_{\boldsymbol{V}}$ and $\mathbb{E}_{\boldsymbol{V}}$ to denote probabilities and expectations, respectively, when the dataset of each machine is generated i.i.d. from $\mu_{\boldsymbol{V}}$. For instance, the marginal distribution of the transcript in this setup is given by

$$\mathbb{P}_{\boldsymbol{V}}(\boldsymbol{Y}=\boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_1)\mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_2)\cdots\mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_m). \tag{B.1}$$

Similarly, the expectation of any function f of the data $\boldsymbol{X}_{1:m}$ and the transcript $\boldsymbol{Y}$ in this setup is

$$\mathbb{E}_{\boldsymbol{V}} f(\boldsymbol{X}_{1:m},\boldsymbol{Y}) \overset{\text{def}}{=} \int \sum_{\boldsymbol{y}\in\{0,1\}^{mb}} f(\boldsymbol{X}_{1:m},\boldsymbol{y})\, \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_1)\cdots\mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_m). \tag{B.2}$$

For our analysis, it will be helpful to consider additional hypothetical setups in which the datasets for some (or all) of the machines are generated from a distribution other than $\mu_{\boldsymbol{V}}$ (such as the null measure $\overline{\mu}$ or the reference measure μ_0 introduced in Proposition 1). We introduce the following three hypothetical setups:

Setup 1: Here, the data samples $\boldsymbol{X}_{1:m} \overset{\text{i.i.d.}}{\sim} \overline{\mu}$. We use $\overline{\mathbb{P}}$ and $\overline{\mathbb{E}}$ to denote the probabilities and expectations in this setup:

$$\overline{\mathbb{P}}(\boldsymbol{Y}=\boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \overline{\mu}^{\otimes m}(\mathrm{d}\boldsymbol{X}_{1:m}), \tag{B.3a}$$

$$\overline{\mathbb{E}} f(\boldsymbol{X}_{1:m},\boldsymbol{Y}) \overset{\text{def}}{=} \int_{\mathcal{X}^m} \sum_{\boldsymbol{y}\in\{0,1\}^{mb}} f(\boldsymbol{X}_{1:m},\boldsymbol{y})\, \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \overline{\mu}^{\otimes m}(\mathrm{d}\boldsymbol{X}_{1:m}). \tag{B.3b}$$

We also use $\overline{\mathbb{E}}[g(\boldsymbol{X}_{1:m})|\boldsymbol{Y}=\boldsymbol{y}]$ to denote conditional expectations in this setup.

Setup 2: Here, data samples $\boldsymbol{X}_{1:m} \overset{\text{i.i.d.}}{\sim} \mu_0$. We use $\mathbb{P}_0$ and $\mathbb{E}_0$ to denote the probabilities and expectations in this setup:

$$\mathbb{P}_0(\boldsymbol{Y}=\boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \mu_0^{\otimes m}(\mathrm{d}\boldsymbol{X}_{1:m}), \tag{B.4a}$$

$$\mathbb{E}_0 f(\boldsymbol{X}_{1:m},\boldsymbol{Y}) \overset{\text{def}}{=} \int_{\mathcal{X}^m} \sum_{\boldsymbol{y}\in\{0,1\}^{mb}} f(\boldsymbol{X}_{1:m},\boldsymbol{y})\, \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}_{1:m})\, \mu_0^{\otimes m}(\mathrm{d}\boldsymbol{X}_{1:m}). \tag{B.4b}$$

We also use $\mathbb{E}_0[g(\boldsymbol{X}_{1:m})|\boldsymbol{Y}=\boldsymbol{y}]$ to denote conditional expectations in this setup.

Setup 3: Here, a fixed machine $i \in [m]$ is exceptional, and the data $\boldsymbol{X}_{1:m}$ are sampled independently as follows:

$$(\boldsymbol{X}_j)_{j \neq i} \overset{\text{i.i.d.}}{\sim} \overline{\mu}, \quad \boldsymbol{X}_i \sim \mu_{\boldsymbol{V}}.$$

We use $\overline{\mathbb{P}}^{(i)}_{\boldsymbol{V}}$ and $\overline{\mathbb{E}}^{(i)}_{\boldsymbol{V}}$ to denote the probabilities and expectations in this setup:

(B.5a)
$$\overline{\mathbb{P}}^{(i)}_{\boldsymbol{V}}(\boldsymbol{Y} = \boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m})\, \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_i) \cdot \prod_{j \neq i} \overline{\mu}(\mathrm{d}\boldsymbol{X}_j),$$

(B.5b)
$$\overline{\mathbb{E}}^{(i)}_{\boldsymbol{V}} f(\boldsymbol{X}_{1:m}, \boldsymbol{Y}) \overset{\text{def}}{=} \int_{\mathcal{X}^m} \sum_{\boldsymbol{y} \in \{0,1\}^{mb}} f(\boldsymbol{X}_{1:m}, \boldsymbol{y})\, \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m})\, \mu_{\boldsymbol{V}}(\mathrm{d}\boldsymbol{X}_i) \cdot \prod_{j \neq i} \overline{\mu}(\mathrm{d}\boldsymbol{X}_j).$$

We also use $\overline{\mathbb{E}}^{(i)}_{\boldsymbol{V}}[g(\boldsymbol{X}_{1:m}) | \boldsymbol{Y} = \boldsymbol{y}]$ to denote conditional expectations in this setup.

Setup 4: Here, a fixed machine $i \in [m]$ is exceptional, and the data $\boldsymbol{X}_{1:m}$ are sampled independently as follows:

$$(\boldsymbol{X}_j)_{j \neq i} \overset{\text{i.i.d.}}{\sim} \overline{\mu}, \quad \boldsymbol{X}_i \sim \mu_0.$$

We use $\overline{\mathbb{P}}^{(i)}_0$ and $\overline{\mathbb{E}}^{(i)}_0$ to denote the probabilities and expectations in this setup:

(B.6a)
$$\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y} = \boldsymbol{y}) \overset{\text{def}}{=} \int \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m})\, \mu_0(\mathrm{d}\boldsymbol{X}_i) \cdot \prod_{j \neq i} \overline{\mu}(\mathrm{d}\boldsymbol{X}_j),$$

(B.6b)
$$\overline{\mathbb{E}}^{(i)}_0 f(\boldsymbol{X}_{1:m}, \boldsymbol{Y}) \overset{\text{def}}{=} \int_{\mathcal{X}^m} \sum_{\boldsymbol{y} \in \{0,1\}^{mb}} f(\boldsymbol{X}_{1:m}, \boldsymbol{y})\, \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m})\, \mu_0(\mathrm{d}\boldsymbol{X}_i) \cdot \prod_{j \neq i} \overline{\mu}(\mathrm{d}\boldsymbol{X}_j).$$

We also use $\overline{\mathbb{E}}^{(i)}_0[g(\boldsymbol{X}_{1:m}) | \boldsymbol{Y} = \boldsymbol{y}]$ to denote conditional expectations in this setup.

(Note that Setup 4 is the hypothetical setup defined in Proposition 1.)

B.1.2. *Properties of Distributed Algorithms.* We recall two well-known properties of distributed estimation protocols in the blackboard model of communication (Definition 2), taken from Bar-Yossef et al. [3] and Jayram [22].

FACT B.1 (Bar-Yossef et al. [3]). Suppose datasets $\boldsymbol{X}_{1:m}$ are distributed across m machines. Let $\boldsymbol{Y} \in \{0,1\}^{mb}$ be the transcript produced by a distributed estimation protocol.

1. The likelihood of the transcript given the data factorizes as follows:
$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m}) = \prod_{i=1}^{m} F_i(\boldsymbol{y} | \boldsymbol{X}_i),$$
where each $F_i(\boldsymbol{y} | \boldsymbol{X}_i)$ takes values in $[0,1]$.
2. Suppose that the datasets $\boldsymbol{X}_{1:m}$ are drawn from a product measure,
$$\boldsymbol{X}_{1:m} \sim \bigotimes_{i=1}^{m} \nu_i,$$

then the conditional distribution of $\boldsymbol{X}_{1:m}$ given $\boldsymbol{Y}=\boldsymbol{y}$ is also a product measure:

$$\boldsymbol{X}_{1:m}|\boldsymbol{Y}=\boldsymbol{y}\sim\bigotimes_{i=1}^{m}\nu_i^{\boldsymbol{y}},$$

where, for each $i\in[m]$,

$$\nu_i^{\boldsymbol{y}}(\mathrm{d}\boldsymbol{X})=\frac{F_i(\boldsymbol{y}|\boldsymbol{X})\nu_i(\mathrm{d}\boldsymbol{X})}{\int_{\mathcal{X}}F_i(\boldsymbol{y}|\boldsymbol{X})\nu_i(\mathrm{d}\boldsymbol{X})}.$$

We also use the following bound on the Hellinger distance, which is a consequence of the "cut-and-paste" property of distributed estimation protocols [3, 22]. This result has been used in several prior works that prove lower bounds for such protocols [e.g., 10, 1].

FACT B.2 (Jayram [22]). Recall the definitions of $\mathbb{P}_{\boldsymbol{V}}$ from (B.1), $\overline{\mathbb{P}}$ from (B.3) and $\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)}$ from (B.5). There exists a universal constant K such that

$$d_{\mathsf{hel}}^2\left(\mathbb{P}_{\boldsymbol{V}},\overline{\mathbb{P}}\right)\leq K\cdot\sum_{i=1}^{m}d_{\mathsf{hel}}^2\left(\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)},\overline{\mathbb{P}}\right).$$

We are now ready to present the proof of Proposition 1.

B.1.3. *Proof of Proposition 1.*

PROOF OF PROPOSITION 1. Recall that:

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\overset{\text{def}}{=}\inf_{\mathbb{Q}}\int d_{\mathsf{hel}}^2(\mathbb{P}_{\boldsymbol{V}},\mathbb{Q})\,\pi(\mathrm{d}\boldsymbol{V}),$$

We choose $\mathbb{Q}=\overline{\mathbb{P}}$ to obtain the bound

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\leq\int d_{\mathsf{hel}}^2\left(\mathbb{P}_{\boldsymbol{V}},\overline{\mathbb{P}}\right)\pi(\mathrm{d}\boldsymbol{V}).$$

By Fact B.2, we have

$$d_{\mathsf{hel}}^2\left(\mathbb{P}_{\boldsymbol{V}},\overline{\mathbb{P}}\right)\leq K\cdot\int\sum_{i=1}^{m}d_{\mathsf{hel}}^2\left(\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)},\overline{\mathbb{P}}\right)\pi(\mathrm{d}\boldsymbol{V}).$$

Recall that

$$\begin{aligned}d_{\mathsf{hel}}^2\left(\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)},\overline{\mathbb{P}}\right)&=\frac{1}{2}\sum_{\boldsymbol{y}\in\{0,1\}^{mb}}\left(\sqrt{\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}-\sqrt{\overline{\mathbb{P}}(\boldsymbol{Y}=\boldsymbol{y})}\right)^2\\&=\frac{1}{2}\sum_{\boldsymbol{y}\in\{0,1\}^{mb}}\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})\cdot\left(\sqrt{\frac{\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}}-\sqrt{\frac{\overline{\mathbb{P}}(\boldsymbol{Y}=\boldsymbol{y})}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}}\right)^2.\end{aligned}$$

Next we observe that, by Fact B.1,

$$\overline{\mathbb{P}}(\boldsymbol{Y}=\boldsymbol{y})=\prod_{j=1}^{m}\overline{\mathbb{E}}F_j(\boldsymbol{y}|\boldsymbol{X}_j),$$

$$\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})=\mathbb{E}_0F_i(\boldsymbol{y}|\boldsymbol{X}_i)\cdot\prod_{\substack{j=1,\\j\neq i}}^{m}\overline{\mathbb{E}}F_j(\boldsymbol{y}|\boldsymbol{X}_j),$$

$$\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)}(\boldsymbol{Y}=\boldsymbol{y})=\mathbb{E}_{\boldsymbol{V}}F_i(\boldsymbol{y}|\boldsymbol{X}_i)\cdot\prod_{\substack{j=1,\\ j\neq i}}^{m}\overline{\mathbb{E}}F_j(\boldsymbol{y}|\boldsymbol{X}_j).$$

Hence,

$$\begin{aligned}\frac{\overline{\mathbb{P}}_{\boldsymbol{V}}^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}&=\frac{\mathbb{E}_{\boldsymbol{V}}F_i(\boldsymbol{y}|\boldsymbol{X}_i)}{\mathbb{E}_0F_i(\boldsymbol{y}|\boldsymbol{X}_i)}=\frac{1}{\mathbb{E}_0F_i(\boldsymbol{y}|\boldsymbol{X}_i)}\mathbb{E}_0\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)F_i(\boldsymbol{y}|\boldsymbol{X}_i)\right]\\&\overset{\text{(a)}}{=}\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}\right],\\\frac{\overline{\mathbb{P}}(\boldsymbol{Y}=\boldsymbol{y})}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})}&=\frac{\overline{\mathbb{E}}F_i(\boldsymbol{y}|\boldsymbol{X}_i)}{\mathbb{E}_0F_i(\boldsymbol{y}|\boldsymbol{X}_i)}=\frac{1}{\mathbb{E}_0F_i(\boldsymbol{y}|\boldsymbol{X}_i)}\mathbb{E}_0\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)F_i(\boldsymbol{y}|\boldsymbol{X}_i)\right]\\&\overset{\text{(a)}}{=}\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}\right].\end{aligned}$$

In the step marked (a) above, we used the characterization on the conditional distribution of $\boldsymbol{X}_i$ given $\boldsymbol{Y}=\boldsymbol{y}$. Hence, we have obtained

$$\begin{aligned}\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\leq\frac{K}{2}\sum_{i=1}^{m}\sum_{\boldsymbol{y}\in\{0,1\}^{mb}}&\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})\times\\&\int\left(\sqrt{\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)|\boldsymbol{Y}=\boldsymbol{y}\right]}-\sqrt{\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)|\boldsymbol{Y}=\boldsymbol{y}\right]}\right)^2\pi(\mathrm{d}\boldsymbol{V}).\end{aligned}$$

We can write

$$\begin{aligned}&\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}\right]\\&\quad=\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\mathbb{I}_{\boldsymbol{X}_i\in\mathcal{Z}}\middle|\boldsymbol{Y}=\boldsymbol{y}\right]+\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\mathbb{I}_{\boldsymbol{X}_i\notin\mathcal{Z}}\middle|\boldsymbol{Y}=\boldsymbol{y}\right],\end{aligned}$$

and analogously for the term involving the likelihood ratio $\mathrm{d}\overline{\mu}/\,\mathrm{d}\mu_0$. For any $a_1,a_2,\epsilon_1,\epsilon_2\geq 0$, we have the scalar inequality

$$\begin{aligned}(\sqrt{a_1+\epsilon_1}-\sqrt{a_2+\epsilon_2})^2&=\epsilon_1+\epsilon_2+(\sqrt{a_1}-\sqrt{a_2})^2+2\sqrt{a_1a_2}-2\sqrt{(a_1+\epsilon_1)(a_2+\epsilon_2)}\\&\leq\epsilon_1+\epsilon_2+(\sqrt{a_1}-\sqrt{a_2})^2.\end{aligned}$$

This gives us

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\leq\frac{K}{2}\cdot(\mathsf{I}+\mathsf{II}),$$

where

$$\begin{aligned}\mathsf{I}\overset{\text{def}}{=}\int\sum_{i=1}^{m}\sum_{\boldsymbol{y}}&\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y})\times\\&\left(\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\mathbb{I}_{\boldsymbol{X}_i\notin\mathcal{Z}}\middle|\boldsymbol{Y}=\boldsymbol{y}\right]+\overline{\mathbb{E}}_0^{(i)}\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\mathbb{I}_{\boldsymbol{X}_i\notin\mathcal{Z}}\middle|\boldsymbol{Y}=\boldsymbol{y}\right]\right)\pi(\mathrm{d}\boldsymbol{V})\end{aligned}$$

and

$$\mathsf{II} \overset{\text{def}}{=} \sum_{i=1}^{m} \sum_{\boldsymbol{y}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}) \times$$

$$\int \left(\sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right]} - \sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right]} \right)^2 \pi(\mathrm{d}\boldsymbol{V}).$$

We simplify I and II separately below.

Analysis of I. By the tower property of conditional expectations,

$$\begin{aligned} \mathsf{I} &= \int \sum_{i=1}^{m} \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \notin \mathcal{Z}} \right] + \overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \notin \mathcal{Z}} \right] \right) \pi(\mathrm{d}\boldsymbol{V}) \\ &= \int \sum_{i=1}^{m} \left(\mathbb{E}_0 \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \notin \mathcal{Z}} \right] + \mathbb{E}_0 \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \notin \mathcal{Z}} \right] \right) \pi(\mathrm{d}\boldsymbol{V}) \\ &= m \cdot \left(\int \mu_{\boldsymbol{V}}(\mathcal{Z}^c) \pi(\mathrm{d}\boldsymbol{V}) + \overline{\mu}(\mathcal{Z}^c) \right). \end{aligned}$$

Analysis of II. Note that

$$\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right] = \overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right] \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{X}_i \in \mathcal{Z} | \boldsymbol{Y}=\boldsymbol{y}).$$

Analogously,

$$\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right] = \overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right] \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{X}_i \in \mathcal{Z} | \boldsymbol{Y}=\boldsymbol{y}).$$

And hence,

$$\left(\sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right]} - \sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}} \middle| \boldsymbol{Y}=\boldsymbol{y} \right]} \right)^2 =$$

$$\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{X}_i \in \mathcal{Z} | \boldsymbol{Y}=\boldsymbol{y}) \times$$

$$\left(\sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right]} - \sqrt{\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right]} \right)^2.$$

Note that by definition of $\mathcal{Z}$,

$$\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right] \geq \frac{1}{2}.$$

Note the scalar inequality for any $a_1 \geq 0, a_2 \geq 1/2$,

$$(\sqrt{a_1} - \sqrt{a_2})^2 = \frac{(a_1 - a_2)^2}{(\sqrt{a_1} + \sqrt{a_2})^2} \leq \frac{(a_1 - a_2)^2}{a_2} \leq 2(a_1 - a_2)^2.$$

This gives us

$$\tfrac{1}{2}(\mathsf{II}) \leq$$

$$\sum_{i=1}^{m} \sum_{\boldsymbol{y} \in \{0,1\}^{mb}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z}) \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_i \in \mathcal{Z} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}).$$

Recall that $Z_i = \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}}$, so $\frac{1}{2}(\mathsf{II})$ can be bounded by:

$$\sum_{i=1}^{m} \sum_{\boldsymbol{y} \in \{0,1\}^{mb}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1) \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V})$$

$$\leq \sum_{i=1}^{m} \sum_{(\boldsymbol{y},z) \in \{0,1\}^{mb+1}} z \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z) \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V})$$

$$= \sum_{i=1}^{m} \overline{\mathbb{E}}_0^{(i)} \left[Z_i \cdot \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}, Z_i \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right].$$

Note that, due to the conditional independence property given in Fact B.1 (item 2), we have

$$\boldsymbol{X}_i | \boldsymbol{Y} \overset{\mathrm{d}}{=} \boldsymbol{X}_i | (\boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i}),$$

where $\overset{\mathrm{d}}{=}$ denotes equality of distributions. Since Z_i is a function of $\boldsymbol{X}_i$, we have

$$(\boldsymbol{X}_i, Z_i) | \boldsymbol{Y} \overset{\mathrm{d}}{=} (\boldsymbol{X}_i, Z_i) | \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \implies \boldsymbol{X}_i | Z_i, \boldsymbol{Y} \overset{\mathrm{d}}{=} \boldsymbol{X}_i | Z_i, \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i}.$$

Hence

$$\mathsf{II} \leq 2 \sum_{i=1}^{m} \overline{\mathbb{E}}_0^{(i)} \left[Z_i \cdot \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right]. \quad \square$$

B.2. Proof of Proposition 2.

In this section, we present the proof of Proposition 2.

PROOF OF PROPOSITION 2. The proof follows the argument from Han, Özgür and Weissman [20]. We prove each item separately. Fix any $z \in \{0, 1\}$, and define $\mathcal{Z}^{(1)} = \mathcal{Z}$ and $\mathcal{Z}^{(0)} = \mathcal{Z}^c$.

1. Consider the following sequence of inequalities:

$$\begin{aligned}
\left| \overline{\mathbb{E}}_0^{(i)} \left[f(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z, (\boldsymbol{X}_j)_{j \neq i} \right] \right|^q &\overset{\text{(a)}}{\leq} \overline{\mathbb{E}}_0^{(i)} \left[|f(\boldsymbol{X}_i)|^q \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z, (\boldsymbol{X}_j)_{j \neq i} \right] \\
&= \frac{\int_{\mathcal{Z}^{(z)}} |f(\boldsymbol{X}_i)|^q \cdot \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}_{1:m}) \, \mu_0(\mathrm{d}\boldsymbol{X}_i)}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z | (\boldsymbol{X}_j)_{j \neq i})} \\
&\overset{\text{(b)}}{\leq} \frac{\int_{\mathcal{X}} |f(\boldsymbol{X}_i)|^q \mu_0(\mathrm{d}\boldsymbol{X}_i)}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z | (\boldsymbol{X}_j)_{j \neq i})} \\
&= \frac{\mathbb{E}_0[|f(\boldsymbol{X})|^q]}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z | (\boldsymbol{X}_j)_{j \neq i})}.
\end{aligned}$$

   In the step marked (a) above, we used Jensen's Inequality; in the step marked (b), we used $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \in [m]}) \leq 1$. Hence,

$$\left| \overline{\mathbb{E}}_0^{(i)} \left[f(\boldsymbol{X}_i) \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z, (\boldsymbol{X}_j)_{j \neq i} \right] \right| \leq \left(\frac{\mathbb{E}_0 |f(\boldsymbol{X})|^q}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z | (\boldsymbol{X}_j)_{j \neq i})} \right)^{\frac{1}{q}},$$

   as claimed.

2. For any $\eta \in \mathbb{R}$,

$$\begin{aligned}\eta \cdot \overline{\mathbb{E}}_0^{(i)}\left[f(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z, (\boldsymbol{X}_j)_{j\neq i}\right] &\overset{\text{(c)}}{\leq} \log \mathbb{E}\left[e^{\eta f(\boldsymbol{X}_i)}\middle|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z, (\boldsymbol{X}_j)_{j\neq i}\right] \\ &\overset{\text{(d)}}{\leq} \log\left(\frac{\mathbb{E}_0[e^{\eta f(\boldsymbol{X}_i)}]}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z|(\boldsymbol{X}_j)_{j\neq i})}\right) \\ &= \log \mathbb{E}_0[e^{\eta f(\boldsymbol{X})}] + \log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z|(\boldsymbol{X}_j)_{j\neq i})}.\end{aligned}$$

The step marked (c) above uses Jensen's inequality, and the step marked (d) relies on the fact that $\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|(\boldsymbol{X}_j)_{j\in[m]}) \leq 1$. Hence, for any $\eta \in \mathbb{R}$,

$$\begin{aligned}&\eta \cdot \overline{\mathbb{E}}_0^{(i)}\left[f(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z, (\boldsymbol{X}_j)_{j\neq i}\right] \\ &\qquad\leq \log \mathbb{E}_0[e^{\eta f(\boldsymbol{X})}] + \log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z|(\boldsymbol{X}_j)_{j\neq i})}.\end{aligned} \tag{B.7}$$

Now, fix $\xi > 0$, and set η as follows:

$$\eta = \xi \cdot \mathsf{sign}\left(\overline{\mathbb{E}}_0^{(i)}\left[f(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z, (\boldsymbol{X}_j)_{j\neq i}\right]\right),$$

so (B.7) with this choice of η yields

$$\begin{aligned}&\left|\overline{\mathbb{E}}_0^{(i)}\left[f(\boldsymbol{X}_i)\middle|\boldsymbol{Y}=\boldsymbol{y}, Z_i=z, (\boldsymbol{X}_j)_{j\neq i}\right]\right| \\ &\qquad\leq \frac{\log \mathbb{E}_0[e^{\eta f(\boldsymbol{X})}]}{\xi} + \frac{1}{\xi}\log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z|(\boldsymbol{X}_j)_{j\neq i})} \\ &\qquad\leq \frac{\log(\mathbb{E}_0[e^{\xi f(\boldsymbol{X})}] \vee \mathbb{E}_0[e^{-\xi f(\boldsymbol{X})}])}{\xi} + \frac{1}{\xi}\log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i=z|(\boldsymbol{X}_j)_{j\neq i})}. \square\end{aligned}$$

APPENDIX C: PROOFS FOR SYMMETRIC TENSOR PCA

C.1. Setup. This appendix is devoted to the proof Proposition 3, the information bound for the distributed k-TPCA problem. Recall that in the distributed k-TPCA problem:

1. An unknown parameter $\boldsymbol{V} \sim \pi$ is drawn from the prior $\pi = \mathsf{Unif}\left(\{\pm 1\}^d\right)$.
2. A dataset consisting of m tensors $\boldsymbol{X}_{1:m}$ is drawn i.i.d. from $\mu_{\boldsymbol{V}}$, where $\mu_{\boldsymbol{V}}$ is the distribution of a single tensor from the k-TPCA problem:

$$\boldsymbol{X}_i = \frac{\lambda \boldsymbol{V}^{\otimes k}}{\sqrt{d^k}} + \boldsymbol{W}_i, \quad (W_i)_{j_1,j_2,\dots j_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad \forall\, j_1, j_2, \dots, j_k \in [d]. \tag{C.1}$$

   This dataset is divided among m machines with 1 tensor per machine.
3. The execution of a distributed estimation protocol with parameters $(m, n=1, b)$ results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

The information bound stated in Proposition 3 is obtained using the general information bound given in Proposition 1 with the following choices:

Choice of μ_0: Under the reference measure, $\boldsymbol{X} \sim \mu_0$ is a k-tensor with i.i.d. $\mathcal{N}(0,1)$ coordinates.

Choice of $\overline{\mu}$: Under the measure $\overline{\mu}$, the sample in each machine is sampled i.i.d. from:

$$\overline{\mu}(\cdot) \overset{\text{def}}{=} \int \mu_{\boldsymbol{V}}(\cdot)\, \pi(\mathrm{d}\boldsymbol{V}).$$

Choice of $\mathcal{Z}$: We choose the event $\mathcal{Z}$ as follows:

$$\mathcal{Z} \overset{\text{def}}{=} \left\{ \boldsymbol{X} \in \bigotimes^k \mathbb{R}^d : \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - 1 \right| \leq \frac{1}{2} \right\}.$$

This appendix is organized into subsections as follows.

1. Appendix C.2 instantiates Fano's inequality in the context of k-TPCA.
2. To prove the information bound for k-TPCA stated in Proposition 3, we rely on certain analytic properties of the likelihood ratio for the k-TPCA problem. These properties are stated (without proofs) in Appendix C.3.
3. Using these properties, Proposition 3 is proved in Appendix C.4.
4. Finally, the proofs of the analytic properties of the likelihood ratio are given in Appendix C.5.

C.2. Fano's Inequality for Symmetric Tensor PCA. Instantiating Fano's inequality (Fact 2) in the context of k-TPCA yields the following corollary.

COROLLARY C.1 (Fano's Inequality for k-TPCA). *For any estimator $\hat{\boldsymbol{V}}(\boldsymbol{Y})$ for k-TPCA computed by a distributed estimation protocol, and for any $t \in \mathbb{R}$, we have*

$$\inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}} \left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d} \right) \leq 2\exp\left(-\frac{t^2}{2}\right) + \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}.$$

PROOF. We apply Fano's Inequality (Fact 2) with the following loss function:

$$\ell(\boldsymbol{V}, \boldsymbol{u}) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } \frac{|\langle \boldsymbol{V}, \boldsymbol{u} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\boldsymbol{u}\|^2} < \frac{t^2}{d}, \\ 0 & \text{otherwise.} \end{cases}$$

To do so, we need to compute a lower bound on $R_0(\pi)$. By Hoeffding's inequality, for any fixed unit vector $\boldsymbol{u}$, we have

$$\mathbb{P}\left(\frac{|\langle \boldsymbol{V}, \boldsymbol{u} \rangle|^2}{\|\boldsymbol{V}\|^2} \geq \frac{t^2}{d} \right) \leq 2\exp\left(-\frac{t^2}{2}\right).$$

Consequently $R_0(\pi) \geq 1 - 2e^{-t^2/2}$. The claim is now immediate from Fact 2. □

C.3. The Likelihood Ratio for Symmetric Tensor PCA. In this section, we collect some important properties of the likelihood ratio for the Tensor PCA problem without proofs. The proofs of these properties are provided in Appendix C.5. This section requires familiarity with Hermite polynomials and their some of their properties, which are reviewed in Appendix I.2.

In order to prove our desired information bound (Proposition 3) we will find it useful to decompose the likelihood ratio for Tensor PCA in the orthogonal basis given by the Hermite polynomials. This decomposition is given in the lemma stated below.

LEMMA C.1 (Hermite Decomposition for Tensor PCA). *For any $\boldsymbol{X} \in \bigotimes^k \mathbb{R}^d$, we have*

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \sum_{i=0}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right).$$

PROOF. See Appendix C.5.1. □

Next, we introduce the following family of functions derived from the Hermite polynomials.

DEFINITION C.1 (Integrated Hermite Polynomials). Let $S : \{\pm 1\}^d \to \mathbb{R}$ be a function with $\|S\|_\pi = 1$. For any $i \in \mathbb{N}_0$, the *integrated Hermite polynomials* are defined as

$$\overline{H}_i(\boldsymbol{X}; S) \overset{\text{def}}{=} \int H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) \cdot S(\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V}).$$

Our rationale for introducing this definition is that proving the communication lower bounds using Proposition 1 requires understanding the following quantities derived from the likelihood ratio:

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \overset{\text{def}}{=} \int \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X})\, \pi(\mathrm{d}\boldsymbol{V}),$$

$$\left\langle \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}), S \right\rangle_\pi \overset{\text{def}}{=} \int \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \cdot S(\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V}).$$

Using Lemma C.1, these quantities are naturally expressed in terms of the integrated Hermite polynomials:

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) = \sum_{i=0}^{\infty} \frac{\lambda^i}{\sqrt{i}} \cdot \overline{H}_i(\boldsymbol{X}; 1),$$

$$\left\langle \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}), S \right\rangle_\pi = \sum_{i=0}^{\infty} \frac{\lambda^i}{\sqrt{i}} \cdot \overline{H}_i(\boldsymbol{X}; S).$$

The following lemma shows that the integrated Hermite polynomials inherit the orthogonality property of the standard Hermite polynomials.

LEMMA C.2. *For any $i, j \in \mathbb{N}_0$ such that $i \neq j$, we have*

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S) \cdot \overline{H}_j(\boldsymbol{X}; S)] = 0,$$

where $\boldsymbol{X} \sim \mu_0$.

PROOF. See Appendix C.5.2. □

Though the integrated Hermite polynomials are orthogonal, they do not have unit norm. In general, the norm of these polynomials depends on the choice of the function S in Definition C.1. The following lemma provides bounds on the norm of the integrated Hermite polynomials.

LEMMA C.3. *There is a universal constant C (independent of d) such that, for any $i \in \mathbb{N}_0$, we have the following.*

1. *For any $S : \{\pm 1\}^d \to \mathbb{R}$ with $\|S\|_\pi \leq 1$, we have $\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2] \leq (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki}{2} \rceil}$.*
2. *For any $S : \{\pm 1\}^d \to \mathbb{R}$ with $\|S\|_\pi \leq 1$, $\langle S, 1\rangle_\pi = 0$, we have $\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2] \leq (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil}$,*

where $\boldsymbol{X} \sim \mu_0$.

PROOF. See Appendix C.5.3. □

As a consequence of the orthogonality property of integrated Hermite polynomials (Lemma C.2) and the estimates obtained in Lemma C.3, one can easily estimate the second moment of functions constructed by linear combinations of the integrated Hermite polynomials:

$$\left\| \sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right\|_2^2 \overset{\text{def}}{=} \mathbb{E}_0 \left(\sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right)^2 = \sum_{i=0}^{\infty} \alpha_i^2 \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2].$$

In our analysis, we will also find it useful to estimate the q-norms of linear combinations of integrated Hermite polynomials for $q \geq 2$:

$$\left\| \sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right\|_q^q \overset{\text{def}}{=} \mathbb{E}_0 \left| \sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right|^q.$$

The following lemma uses Gaussian Hypercontractivity (Fact I.7) to provide an estimate for the above quantity.

LEMMA C.4. *Let $\{\alpha_i : i \in \mathbb{N}_0\}$ be an arbitrary collection of real-valued coefficients. For any $q \geq 2$, we have*

$$\left\| \sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right\|_q^2 \leq \sum_{i=0}^{\infty} (q-1)^i \cdot \alpha_{\boldsymbol{i}}^2 \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2]$$

Furthermore, the inequality holds as an equality when $q = 2$.

PROOF. See Appendix C.5.4. □

C.4. Proof of Information Bound (Proposition 3). In this subsection, we present a proof of the information bound for distributed Tensor PCA (Proposition 3). We begin by recalling the general information bound from Proposition 1:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}{K} \leq \sum_{i=1}^{m} \overline{\mathbb{E}}_0^{(i)} \left[\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right] + m\overline{\mu}(\mathcal{Z}^c).$$

In order to analyze the conditional expectation of the centered likelihood ratio, we will approximate it by a low-degree polynomial. Recall that in Lemma C.1, we computed the following expansion of the likelihood ratio in terms of the Hermite polynomials:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \sum_{i=0}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot H_i \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right).$$

Recalling the definition of integrated Hermite polynomials (Definition C.1), and also that

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0} = \int \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0} \, \pi(\mathrm{d}\boldsymbol{V}),$$

we can express the integrated likelihood ratio in terms of the integrated Hermite polynomials:

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0} = \sum_{i=0}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot \overline{H}_i(\boldsymbol{X}; 1).$$

For any $t \in \mathbb{N}$, we define the degree t-approximation to the centered likelihood ratio:

$$\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t} \overset{\text{def}}{=} \sum_{i=0}^{t} \frac{\lambda^i}{\sqrt{i!}} \cdot \left(H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) - \overline{H}_i(\boldsymbol{X};1)\right)$$

and the corresponding truncation error:

$$\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{> t} \overset{\text{def}}{=} \sum_{i=t+1}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot \left(H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) - \overline{H}_i(\boldsymbol{X};1)\right).$$

By choosing t large enough, we hope that:

$$\begin{aligned}\overline{\mathbb{E}}_0^{(i)}&\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right] \\ &\approx \overline{\mathbb{E}}_0^{(i)}\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right)_{\leq t}\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right].\end{aligned}$$

We estimate the approximation error in the above equation using the following lemma.

LEMMA C.5. *Let $\boldsymbol{X} \sim \mu_0$. Suppose that:*

$$t \geq (\lambda^2 e^2) \vee \log\frac{4}{\epsilon} \vee 1.$$

Then

$$\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{>t}^2\right] \leq \epsilon.$$

PROOF. The proof of this result appears at the end of this subsection (Appendix C.4.2). □

Finally to analyze the conditional expectation of the low degree approximation using the Geometric Inequality (Proposition 2), we need to understand the concentration properties of the low-degree approximation of the likelihood ratio. This is done using the moment estimates provided in the following lemma.

LEMMA C.6. *Let $\boldsymbol{X} \sim \mu_0$. There exists a finite constant C_k depending only on k such that for any $q \geq 2$ which satisfies:*

$$\lambda^2(q-1) \leq \frac{1}{C_k} \cdot \frac{d^{\frac{k}{2}}}{t^{\frac{k-2}{2}}},$$

we have

$$\sup_{\substack{S:\{\pm 1\}^d \to \mathbb{R} \\ \|S\|_\pi \leq 1}} \left(\mathbb{E}_0\left[\left|\left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t}, S\right\rangle_\pi\right|^q\right]\right)^{\frac{2}{q}} \leq (q-1)\cdot\sigma^2,$$

where

$$\sigma^2 \overset{\text{def}}{=} \begin{cases} C_k \cdot \lambda^2 \cdot d^{-\frac{k+2}{2}} & \textit{if } k \textit{ is even;} \\ C_k \cdot \lambda^2 \cdot d^{-\frac{k+1}{2}} & \textit{if } k \textit{ is odd.}\end{cases}$$

PROOF. The proof of this result appears at the end of this subsection (Appendix C.4.1). □

Finally, we also need to estimate $\overline{\mu}(\mathcal{Z}^c)$ to upper bound the Hellinger Information using Proposition 1. This is the content of the following lemma.

LEMMA C.7. *Consider the event $\mathcal{Z}$:*

$$\mathcal{Z} \overset{def}{=} \left\{ \boldsymbol{x} \in \mathcal{X} : \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right| \leq \frac{1}{2} \right\},$$

There exists a universal constant C_k (depending only on k) such that, for any $2 \leq q \leq d/(C_k \lambda^2)$, we have

$$\overline{\mu}(\mathcal{Z}^c) \leq \left(\frac{C_k q \lambda^2}{\sqrt{d^k}} + e^{-d} \right)^{\frac{q}{2}}.$$

PROOF. The proof of this lemma appears at the end of this subsection (Appendix C.4.3). □

With these results, we are now ready to provide a proof of Proposition 3.

PROOF OF PROPOSITION 3. Recall that in Proposition 1 we showed:

$$\frac{\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)}{K} \leq \sum_{i=1}^{m} \overline{\mathbb{E}}_0^{(i)} \left[\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right] + m\overline{\mu}(\mathcal{Z}^c)$$

The centered likelihood ratio can be decomposed as:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t} + \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{> t}.$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and Cauchy Schwarz Inequality:

$$\frac{1}{2}\left(\overline{\mathbb{E}}_0^{(i)} \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i} \right] \right)^2 \leq \left(\overline{\mathbb{E}}_0^{(i)} \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right)_{\leq t} \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i} \right] \right)^2 + \overline{\mathbb{E}}_0^{(i)} \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right)_{> t}^2 \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i} \right].$$

Hence,

$$\frac{\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)}{2K} \leq$$

(C.2)

$$\sum_{i=1}^{m} \overline{\mathbb{E}}_0^{(i)} \left[\Psi_i^2(\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}) \right] + \frac{m\overline{\mu}(\mathcal{Z}^c)}{2} + m \cdot \int \mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{> t}^2 \right] \pi(\mathrm{d}\boldsymbol{V}),$$

where:

$$\Psi_i^2(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i}) \overset{\text{def}}{=}$$

$$\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right)_{\leq t} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}).$$

Our goal is to show that for any $\alpha \geq 2$, we have

$$\frac{\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V}; \boldsymbol{Y}\right)}{2K} \leq \underbrace{\frac{C_k \lambda^2 \alpha}{d} + m \cdot \left(\frac{C_k \alpha \lambda^2}{\sqrt{d^k}} + e^{-d} \right)^{\frac{\alpha}{2}}}_{\text{Step 1}} + \underbrace{\frac{1}{d}}_{\text{Step 2}}$$

$$\underbrace{+ 3C_k \cdot \lambda^2 \cdot b \left(\frac{(\lambda^2 e^2) \vee \log(m \cdot d)}{d} \right)^{\frac{k}{2}} + 16\sigma^2 \cdot m \cdot b}_{\text{Step 3}} \tag{C.3}$$

The information bound in the statement of the proposition follows by choosing α optimally. The proof proceeds in several steps. In the above display, we have grouped the terms in the information bound according to the step they arise in.

Step 1: Controlling $\overline{\mu}(\mathcal{Z}^c)$. Note that if $\lambda^2 \alpha > d/C_k$, then the claimed upper bound (C.3) on $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{Y}; \boldsymbol{V}\right)$ is trivial since $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{Y}; \boldsymbol{V}\right) \leq 1$. Hence we assume $\lambda^2 \alpha \leq d/C_k$. Applying Lemma C.7 with $q = \alpha$, we have

$$m \cdot \overline{\mu}(\mathcal{Z}^c) \leq m \cdot \left(\frac{C_k \alpha \lambda^2}{\sqrt{d^k}} + e^{-d} \right)^{\frac{\alpha}{2}} \leq \frac{C_k \lambda^2 \alpha}{d} + m \cdot \left(\frac{C_k \alpha \lambda^2}{\sqrt{d^k}} + e^{-d} \right)^{\frac{\alpha}{2}}. \tag{C.4}$$

Step 2: Controlling High Degree Term. We set:

$$t = (\lambda^2 e^2) \vee \log(m \cdot d).$$

Applying Lemma C.5, we obtain,

$$\mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{>t}^2 \right] \leq \frac{1}{m \cdot d}. \tag{C.5}$$

Step 3: Controlling Low Degree Term. Next we control $\Psi_i^2(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})$. By linearization (Lemma 1) we have:

$$\Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i}) =$$

$$\sup_{\substack{S: \mathcal{V} \to \mathbb{R} \\ \|S\|_\pi \leq 1}} \overline{\mathbb{E}}_0^{(i)} \left[\left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right)_{\leq t}, S \right\rangle_\pi \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i} \right].$$

Using the Geometric Inequality framework (Proposition 2) we can bound $|\Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})|$ if we can understand the concentration properties of:

$$f_S(\boldsymbol{X}) \overset{\text{def}}{=} \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right)_{\leq t}, S \right\rangle_\pi, \ \boldsymbol{X} \sim \mu_0$$

for any $S : \mathcal{V} \to \mathbb{R}$, $\|S\|_\pi \leq 1$. The concentration properties of $f_S(\boldsymbol{X})$ are studied in Lemma C.6 which shows that for any q such that:

$$1 \leq q - 1 \leq \frac{1}{C_k \cdot \lambda^2} \left(\frac{d}{t} \right)^{\frac{k}{2}}, \tag{C.6}$$

we have

$$\sup_{\substack{S:\{\pm 1\}^d \to \mathbb{R} \\ \|S\|_\pi \le 1}} \left(\mathbb{E}_0\left[|f_S(\boldsymbol{X})|^q\right]\right)^{\frac{2}{q}} \le \sigma^2(q-1),$$

where:

$$\sigma^2 \overset{\text{def}}{=} \begin{cases} C_k \cdot \lambda^2 \cdot d^{-\frac{k+2}{2}} : & k \text{ is even} \\ C_k \cdot \lambda^2 \cdot d^{-\frac{k+1}{2}} : & k \text{ is odd} \end{cases}.$$

In order to apply Proposition 2 we need to choose q appropriately. The choice of q depends on

$$\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}).$$

We define the set of rare and frequent realizations of $\boldsymbol{Y}, Z_i$:

$$\mathcal{R}_{\mathsf{freq}}^{(i)} \overset{\text{def}}{=} \left\{(\boldsymbol{y}, \boldsymbol{z}_i) \in \{0,1\}^{mb+1} : \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) > \frac{1}{e}\right\},$$

$$\mathcal{R}_{\mathsf{rare}}^{(i)} \overset{\text{def}}{=} \left\{(\boldsymbol{y}, \boldsymbol{z}_i) \in \{0,1\}^{mb+1} : 0 < \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) \le 4^{-b}\right\}.$$

By the tower property,

$$\begin{aligned}\overline{\mathbb{E}}_0^{(i)}\left[\Psi_i^2(\boldsymbol{Y}, X_i, (\boldsymbol{X}_j)_{j\neq i})\right] &= \overline{\mathbb{E}}_0^{(i)}\overline{\mathbb{E}}_0^{(i)}\left[\Psi_i^2(\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}) \mid (\boldsymbol{X}_j)_{j\neq i})\right] \\ &= \overline{\mathbb{E}}_0^{(i)} F_i((\boldsymbol{X}_j)_{j\neq i}) + \overline{\mathbb{E}}_0^{(i)} R_i((\boldsymbol{X}_j)_{j\neq i}) + \overline{\mathbb{E}}_0^{(i)} O_i((\boldsymbol{X}_j)_{j\neq i}).\end{aligned}$$

where:

$$F_i((\boldsymbol{x}_j)_{j\neq i}) \overset{\text{def}}{=} \sum_{(\boldsymbol{y}, z_i) \in \mathcal{R}_{\mathsf{freq}}^{(i)}} \Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})^2 \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}),$$

$$R_i((\boldsymbol{x}_j)_{j\neq i}) \overset{\text{def}}{=} \sum_{(\boldsymbol{y}, z_i) \in \mathcal{R}_{\mathsf{rare}}^{(i)}} \Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})^2 \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}),$$

$$O_i((\boldsymbol{x}_j)_{j\neq i}) \overset{\text{def}}{=} \sum_{(\boldsymbol{y}, z_i) \notin \mathcal{R}_{\mathsf{freq}}^{(i)} \cup \mathcal{R}_{\mathsf{rare}}^{(i)}} \Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})^2 \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}).$$

We bound each of the terms separately.

Case 1: Frequent realizations. Consider the case when $(\boldsymbol{y}, z_i) \in \mathcal{R}_{\mathsf{freq}}^{(i)}$. In this case we set $q = 2$. We need to check that this choice obeys (C.6). Indeed if (C.6) is violated for $q = 2$, then the upper bound on $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})$ in (C.3) is trivial since the term:

$$3C_k \cdot \lambda^2 \cdot b \left(\frac{(\lambda^2 e^2) \vee \log(m \cdot d)}{d}\right)^{\frac{k}{2}} > 1.$$

Hence we may assume that $q = 2$ obeys (C.6) without loss of generality and we obtain by Proposition 2,

$$|\Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})| \le \sigma \cdot \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i})^{-\frac{1}{2}}, \forall\, (\boldsymbol{y}, z_i) \in \mathcal{R}_{\mathsf{freq}}^{(i)}.$$

Note that $|\mathcal{R}_{\mathsf{freq}}^{(i)}| \le e$, and hence,

$$F_i((\boldsymbol{x}_j)_{j\neq i}) \le 2\sigma^2.$$

Case 2: Rare realizations. Consider the case when $(\boldsymbol{y}, z_i) \in \mathcal{R}^{(i)}_{\mathsf{rare}}$. In this case we set $q = 4$. It is straightforward to check that if q doesn't satisfy (C.6),then the claimed bound on $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)$ in (C.3) is vacuous and hence we assume $q = 4$ satisfies (C.6). Applying Proposition 2 gives us $\forall\, (\boldsymbol{y}, z_i) \in \mathcal{R}^{(i)}_{\mathsf{rare}}$:

$$|\Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})| \leq \sqrt{3}\cdot\sigma\cdot\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i})^{-\frac{1}{4}}.$$

Hence we can upper bound R_i:

$$\begin{aligned} R_i((\boldsymbol{x}_j)_{j\neq i}) &\overset{\text{def}}{=} \sum_{(\boldsymbol{y},z_i)\in\mathcal{R}^{(i)}_{\mathsf{rare}}} \Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})^2\cdot\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) \\ &\leq 3\sigma^2 \sum_{(\boldsymbol{y},z_i)\in\mathcal{R}^{(i)}_{\mathsf{rare}}} \overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i})^{\frac{1}{2}} \\ &\leq 3\sigma^2 2^{-b}|\mathcal{R}^{(i)}_{\mathsf{rare}}|. \end{aligned}$$

Recall that we assume that the communication protocol is deterministic, i.e. the bit written by a machine is a deterministic function its local dataset and the bits written on the blackboard so far. Hence, conditional on $(\boldsymbol{X}_j)_{j\neq i}$ there are only 2^{b+1} possible realizations of $(\boldsymbol{Y}, Z_i)$ with non-zero probability. And hence, $|\mathcal{R}^{(i)}_{\mathsf{rare}}| \leq 2^{b+1}$. Hence,

$$R_i((\boldsymbol{x}_j)_{j\neq i}) \leq 6\sigma^2.$$

Case 3: All other realizations. Now consider any realization $(\boldsymbol{y}, z_i) \notin \mathcal{R}^{(i)}_{\mathsf{rare}} \cup \mathcal{R}^{(i)}_{\mathsf{freq}}$. In this case, we set q as:

$$q = -2\log\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}).$$

Since $(\boldsymbol{y}, z_i) \notin \mathcal{R}^{(i)}_{\mathsf{rare}} \cup \mathcal{R}^{(i)}_{\mathsf{freq}}$, we have

$$2 \leq q \leq b\log(4) \leq 2b.$$

In particular, if

$$b\lambda^2 \leq \frac{1}{2C_k}\left(\frac{d}{t}\right)^{\frac{k}{2}},$$

then (C.6) holds for this choice of q. On the other hand, if this is not the case, then the claimed upper bound (C.3) on $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)$ is trivial since

$$3C_k\cdot\lambda^2\cdot b\left(\frac{(\lambda^2 e^2)\vee\log(m\cdot d)}{d}\right)^{\frac{k}{2}} > 1.$$

Hence, we have for any $(\boldsymbol{y}, z_i) \notin \mathcal{R}^{(i)}_{\mathsf{rare}} \cup \mathcal{R}^{(i)}_{\mathsf{freq}}$:

$$|\Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})| \leq \sqrt{2e}\cdot\sigma\cdot(-\log\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}))^{-\frac{1}{2}}.$$

Hence,

$$\begin{aligned} &O_i((\boldsymbol{x}_j)_{j\neq i}) \\ &\quad\overset{\text{def}}{=} \sum_{(\boldsymbol{y},z_i)\notin\mathcal{R}^{(i)}_{\mathsf{freq}}\cup\mathcal{R}^{(i)}_{\mathsf{rare}}} \Psi(\boldsymbol{y}, z_i, (\boldsymbol{x}_j)_{j\neq i})^2\cdot\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) \\ &\quad\leq 2e\sigma^2 \sum_{(\boldsymbol{y},z_i)\in\{0,1\}^{mb}} h(\overline{\mathbb{P}}^{(i)}_0(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i})), \end{aligned}$$

where $h(\cdot)$ is the entropy function $h(x) \overset{\text{def}}{=} -x\log(x)$. We note that the expression appearing in the above equation is the entropy of $(\boldsymbol{Y}, Z_i)$ conditional on $(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}$. Since the protocol is deterministic (cf. Remark 4) there are at most 2^{b+1} realizations $(\boldsymbol{y}, z_i)$ such that $\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i | (\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) > 0$. Since the entropy is maximized by the uniform distribution:

$$O_i((\boldsymbol{x}_j)_{j\neq i}) \leq 2 \cdot e \cdot \log(2) \cdot \sigma^2 \cdot (b+1).$$

Using the bounds from the above 3 cases, we have:

$$\overline{\mathbb{E}}_0^{(i)}\left[\Psi_i^2(\boldsymbol{Y}, X_i, (\boldsymbol{X}_j)_{j\neq i})\right] = \overline{\mathbb{E}}_0^{(i)}\overline{\mathbb{E}}_0^{(i)}\left[\Psi_i^2(\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}) \mid (\boldsymbol{X}_j)_{j\neq i})\right] \tag{C.7}$$

$$\leq 8\sigma^2 + 2e\log(2)\cdot\sigma^2\cdot(b+1) \leq 16\sigma^2 b. \tag{C.8}$$

Substituting the estimates (C.4), (C.5) and (C.7) in (C.2) we obtain (C.3). This proves the first claim made in the statement of the proposition. Lastly, we consider scaling regime:

$$\lambda = \Theta(1),\ m = \Theta(d^\eta),\ b = \Theta(d^\beta)$$

for some constants $\eta \geq 1, \beta \geq 0$, which satisfy:

$$\eta + \beta < \left\lceil \frac{k+1}{2} \right\rceil.$$

We set α to be any constant strictly more than $\max(2, 4\eta/k)$ and observe that $\beta < k/2 + 1 - \eta \leq k/2$. Consequently, each term in the upper bound in (C.3) is $o_d(1)$. This finishes the proof of the proposition. □

The remainder of this subsection is devoted to the proofs of Lemma C.6 (in Appendix C.4.1), Lemma C.5 (in Appendix C.4.2) and Lemma C.7 (in Appendix C.4.3).

C.4.1. *Analysis of Low Degree Part.*

In this section, we provide a proof of Lemma C.6.

PROOF OF LEMMA C.6. Recall that,

$$\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t} \overset{\text{def}}{=} \sum_{i=0}^{t} \frac{\lambda^i}{\sqrt{i!}} \cdot \left(H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) - \overline{H}_i(\boldsymbol{X}; 1)\right),$$

and in particular,

$$\int \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t} \pi(\mathrm{d}\boldsymbol{V}) = 0.$$

Hence,

$$\begin{aligned}\left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t}, S\right\rangle_\pi &= \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t}, S - \langle S, 1\rangle_\pi\right\rangle_\pi \\ &= \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t}, S - \langle S, 1\rangle_\pi\right\rangle_\pi,\end{aligned}$$

where,

$$\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X})\right)_{\leq t} \overset{\text{def}}{=} \sum_{i=0}^{t} \frac{\lambda^i}{\sqrt{i!}} \cdot H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right).$$

Consequently,

$$\sup_{\substack{S:\{\pm 1\}^d \to \mathbb{R} \\ \|S\|_\pi \leq 1}} \mathbb{E}_0 \left[\left| \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t}, S \right\rangle_\pi \right|^q \right]$$

$$= \sup_{\substack{S:\{\pm 1\}^d \to \mathbb{R} \\ \|S\|_\pi \leq 1 \\ \langle S, 1 \rangle_\pi = 0}} \mathbb{E}_0 \left[\left| \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t}, S \right\rangle_\pi \right|^q \right].$$

We can compute:

$$\left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t}, S \right\rangle_\pi = \sum_{i=1}^t \frac{\lambda^i}{\sqrt{i!}} \cdot \overline{H}_i(\boldsymbol{X}; S).$$

Note that when $\boldsymbol{X} \sim \mu_0$, $\boldsymbol{X}$ is a Gaussian tensor with i.i.d. $\mathcal{N}(0,1)$ entries. Hence by Gaussian Hypercontractivity (Lemma C.4):

$$\left(\mathbb{E}_0 \left[\left| \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t}, S \right\rangle_\pi \right|^q \right] \right)^{\frac{2}{q}} = \sum_{i=1}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2]$$

$$\overset{\text{(a)}}{\leq} \sum_{i=1}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil},$$

where in the step marked (a) we used the estimate on $\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2]$ from Lemma C.3. We split the above sum into two parts:

$$\sum_{i=1}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil} = \sum_{\substack{i=1 \\ i \text{ is odd}}}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil}$$

$$+ \sum_{\substack{i=1 \\ i \text{ is even}}}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil}.$$

We first analyze the sum corresponding to the odd terms. By estimating the ratio of two consecutive terms in the sum, one can obtain a constant C_k such that if,

$$\text{(C.9)} \qquad \frac{\lambda^4 \cdot (q-1)^2 \cdot t^{k-2}}{d^k} \leq \frac{1}{C_k},$$

then the sum decays geometrically by a factor of $1/2$. Hence,

$$\sum_{\substack{i=1 \\ i \text{ is odd}}}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil} \leq \frac{C_k \cdot \lambda^2 \cdot (q-1)}{d^{\lceil \frac{k+1}{2} \rceil}} \left(1 + \frac{1}{2} + \frac{1}{4} + \dots \right)$$

$$\leq \frac{C_k \cdot \lambda^2 \cdot (q-1)}{d^{\lceil \frac{k+1}{2} \rceil}}.$$

The same argument can be used to estimate the sum corresponding to the even terms. Under the same assumption (C.9), we have

$$\sum_{\substack{i=1 \\ i \text{ is even}}}^t \frac{\lambda^{2i}}{i!} \cdot (q-1)^i \cdot (Cki)^{\frac{ki}{2}} \cdot d^{-\lceil \frac{ki+1}{2} \rceil} \leq \frac{C_k \cdot \lambda^4 \cdot (q-1)^2}{d^{k+1}} \left(1 + \frac{1}{2} + \dots \right)$$

$$\leq \frac{C_k \cdot \lambda^4 \cdot (q-1)^2}{d^{k+1}}.$$

Hence,

$$\sup_{\substack{S:\{\pm 1\}^d \to \mathbb{R} \\ \|S\|_\pi \leq 1}} \left(\mathbb{E}_0 \left[\left| \left\langle \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{\leq t}, S \right\rangle_\pi \right|^q \right] \right)^{\frac{2}{q}} \leq \frac{C_k \lambda^2 (q-1)}{d^{\lceil \frac{k+1}{2} \rceil}} + \frac{C_k \lambda^4 (q-1)^2}{d^{k+1}}$$

$$\leq \frac{C_k \cdot \lambda^2 \cdot (q-1)}{d^{\lceil \frac{k+1}{2} \rceil}}.$$

In the above display, in order to obtain the final inequality, we again used assumption (C.9). This concludes the proof. □

C.4.2. *Analysis of High Degree Part.* In this section, we provide a proof of Lemma C.5.

PROOF OF LEMMA C.5. Recall that,

$$\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{>t} \overset{\text{def}}{=} \sum_{i=t+1}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot \left(H_i \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right) - \overline{H}_i(\boldsymbol{X}; 1) \right).$$

Hence,

$$\mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{>t}^2 \right] \leq$$

$$2\mathbb{E}_0 \left[\left(\sum_{i=t+1}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \cdot H_i \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right) \right)^2 \right] + 2\mathbb{E}_0 \left[\left(\sum_{i=t+1}^{\infty} \overline{H}_i(\boldsymbol{X}; 1) \right)^2 \right].$$

By the orthogonality of Hermite and integrated Hermite polynomials (see Lemma C.2), we obtain,

$$\mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{>t}^2 \right] \leq 2 \left(\sum_{i=t+1}^{\infty} \mathbb{E}_0 \left[H_i^2 \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right) \right] + \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \right).$$

By Jensen's Inequality,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \leq \int \mathbb{E}_0 \left[H_i^2 \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{\sqrt{d^k}} \right) \right] \pi(\mathrm{d}\boldsymbol{V}) = 1.$$

Hence,

$$\mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \right)_{>t}^2 \right] \leq 4 \sum_{i=t+1}^{\infty} \frac{\lambda^{2i}}{i!} \overset{\text{(a)}}{\leq} \epsilon.$$

In the last step, we used the hypothesis on t and Fact I.1 given in Appendix I. □

C.4.3. *Analysis of Bad Event.* In this section, we provide a proof of Lemma C.7.

PROOF OF LEMMA C.7. For any $q \geq 2$, we have, by Chebychev's Inequality:

$$\overline{\mathbb{P}}(\mathcal{Z}^c) \leq 2^q \cdot \overline{\mathbb{E}} \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^q$$

$$= 2^q \cdot \mathbb{E}_0 \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0} \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^q$$

$$\leq 2^q \left(\mathbb{E}_0(\boldsymbol{x}) \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^{q+1} + \mathbb{E}_0 \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^q \right)$$

Recalling Lemma C.1 and Definition C.1, we have

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = 1 + \sum_{t=1}^{\infty} \frac{\lambda^i}{\sqrt{i!}} \overline{H}_i(\boldsymbol{X}; 1).$$

By Gaussian Hypercontractivity (Lemma C.4) we have, for any $q \geq 2$:

$$\left(\mathbb{E}_0 \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^q \right)^{\frac{2}{q}} \leq \sum_{i=1}^{\infty} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2].$$

We split the above sum into a high-degree part and a low degree part. Let $t \in \mathbb{N}$ be arbitrary. Then

$$\left(\mathbb{E}_0 \left| \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|^q \right)^{\frac{2}{q}} \leq \sum_{i=1}^{t} \frac{\lambda^{2i}(q-1)^i}{i!} \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] + \sum_{i=t+1}^{\infty} \frac{\lambda^{2i}(q-1)^i}{i!} \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2].$$

Analysis of the low degree part. Using the bound on $\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2]$ obtained in Lemma C.3, we have

$$\sum_{i=1}^{t} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \leq \sum_{i=1}^{t} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \left(\frac{Cki}{d} \right)^{\frac{ki}{2}}.$$

By analyzing the ratio of consecutive terms in the sum, one can find a constant C_k depending only on k such that, if,

$$\lambda^2 \cdot (q-1) \leq \frac{1}{C_k} \cdot \frac{d^{\frac{k}{2}}}{t^{\frac{k-2}{2}}}, \tag{C.10}$$

then the sum decays geometrically with a factor of atleast $1/2$ and hence,

$$\sum_{i=1}^{t} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \leq \frac{C_k \cdot \lambda^2 \cdot (q-1)}{d^{\frac{k}{2}}} \cdot \left(1 + \frac{1}{2} + \cdots \right) \leq \frac{C_k \lambda^2 (q-1)}{d^{\frac{k}{2}}}.$$

Analysis of high degree part. Recall the definition of integrated Hermite polynomials (Definition C.1):

$$\overline{H}_i(\boldsymbol{X}; 1) = \int H_i \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{d^{\frac{k}{2}}} \right) \pi(\mathrm{d}\boldsymbol{V}).$$

Hence, by Jensen's Inequality,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \leq \int \mathbb{E}_0 \left[H_i \left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k} \rangle}{d^{\frac{k}{2}}} \right)^2 \right] \pi(\mathrm{d}\boldsymbol{V}) = 1.$$

Hence,

$$\sum_{i=t+1}^{\infty} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; 1)^2] \leq \sum_{i=t+1}^{\infty} \frac{\lambda^{2i} \cdot (q-1)^i}{i!}.$$

Appealing to Fact I.1, we set,

$$t = (e^2\lambda^2(q-1)) \vee d \vee 1, \tag{C.11}$$

and obtain,

$$\sum_{i=t+1}^{\infty} \frac{\lambda^{2i} \cdot (q-1)^i}{i!} \cdot \mathbb{E}_0[\overline{H}_i(\boldsymbol{X};1)^2] \leq e^{-d}.$$

Note that the hypothesis assumed in the statememt of the lemma $\lambda^2 \cdot q \leq d/C_k$ guarantees that the choice of t in (C.11) satisfies (C.10). Hence, we have shown that for any $q \geq 2$

$$\left(\mathbb{E}_0 \left|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1\right|^q\right)^{\frac{2}{q}} \leq \frac{C_k(q-1)\lambda^2}{\sqrt{d^k}} + e^{-d},$$

for a universal constant C_k depending only on k provided $\lambda^2(q-1) \leq d/C_k$. Applying this with $q, q+1$ we obtain:

$$\overline{\mu}(\mathcal{Z}^c) \leq \left(\frac{C_k q\lambda^2}{\sqrt{d^k}} + \frac{1}{d}\right)^{\frac{q}{2}} + \left(\frac{C_k q\lambda^2}{\sqrt{d^k}} + e^{-d}\right)^{\frac{q+1}{2}}$$

provided $\lambda^2 q \leq d/C_k$. Note that under this assumption, since $k \geq 2$, we have $C_k\lambda^2 q/\sqrt{d^k} \leq C_k\lambda^2 q/d \leq 1$. Hence the above bound can be simplified to:

$$\overline{\mu}(\mathcal{Z}^c) \leq \left(\frac{C_k q\lambda^2}{\sqrt{d^k}} + e^{-d}\right)^{\frac{q}{2}},$$

for a suitably large constant C_k. □

C.5. Omitted Proofs from Appendix C.3. This section contains the proofs of the various analytic properties of the likelihood ratio for Tensor PCA, which were stated in Appendix C.3.

C.5.1. *Proof of Lemma C.1.*

PROOF OF LEMMA C.1. We observe that,

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \exp\left(\lambda \cdot \frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}} - \frac{\lambda^2}{2}\right).$$

In particular, the likelihood ratio depends on $\boldsymbol{X}$ only through the projection $\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle$. Observe that when $\boldsymbol{X} \sim \mu_0$,

$$\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}} \sim \mathcal{N}(0,1).$$

Since Hermite polynomials form a complete orthonormal basis for $L^2(\mathcal{N}(0,1))$, the likelihood ratio admits an expansion of the form:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \sum_{i=0}^{\infty} c_i \cdot H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right).$$

The coefficients c_i are given by:

$$c_i \stackrel{\text{def}}{=} \mathbb{E}_0\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \cdot H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right],$$

where $\boldsymbol{X} \sim \mu_0$. We can simplify c_i as follows:

$$\begin{aligned} c_i &\overset{\text{def}}{=} \mathbb{E}_0\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) \cdot H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] \\ &\overset{\text{(a)}}{=} \mathbb{E}_{\boldsymbol{V}}\left[H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] \\ &\overset{\text{(b)}}{=} \mathbb{E}_0\left[H_i(\lambda + Z)\right] \\ &\overset{\text{(c)}}{=} \frac{\lambda^i}{\sqrt{i!}}. \end{aligned}$$

In the above display, in the step marked (a), applied a change of measure to change the distribution of $\boldsymbol{X}$ to $\boldsymbol{X} \sim \mu_{\boldsymbol{V}}$. In the step marked (b), we used the fact that when $\boldsymbol{X} \sim \mu_{\boldsymbol{V}}$,

$$\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}} \sim \lambda + Z, \quad Z \sim \mathcal{N}(0,1).$$

In the step marked (c), we appealed to Fact I.4. □

C.5.2. *Proof of Lemma C.2.*

PROOF OF LEMMA C.2. Using Definition C.1 and Fubini's theorem, we obtain,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S) \cdot \overline{H}_j(\boldsymbol{X};S)] =$$

$$\int \mathbb{E}_0\left[H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) H_j\left(\frac{\langle \boldsymbol{X}, \widetilde{\boldsymbol{V}}^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] \cdot S(\boldsymbol{V}) \cdot S(\widetilde{\boldsymbol{V}})\, \pi(\mathrm{d}\boldsymbol{V})\, \pi(\mathrm{d}\widetilde{\boldsymbol{V}}).$$

Since $i \neq j$, Fact I.6 gives us,

$$\mathbb{E}_0\left[H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) H_j\left(\frac{\langle \boldsymbol{X}, \widetilde{\boldsymbol{V}}^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] = 0.$$

Hence, we obtain the claim of the lemma. □

C.5.3. *Proof of Lemma C.3.*

PROOF OF LEMMA C.3. Using Definition C.1 and Fubini's theorem, we obtain,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] =$$

$$\int \mathbb{E}_0\left[H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}_1^{\otimes k}\rangle}{\sqrt{d^k}}\right) H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}_2^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] \cdot S(\boldsymbol{V}_1) \cdot S(\boldsymbol{V}_2)\, \pi(\mathrm{d}\boldsymbol{V}_1)\, \pi(\mathrm{d}\boldsymbol{V}_2).$$

Fact I.6 gives us,

$$\mathbb{E}_0\left[H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}_1^{\otimes k}\rangle}{\sqrt{d^k}}\right) H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}_2^{\otimes k}\rangle}{\sqrt{d^k}}\right)\right] = \left(\frac{\langle \boldsymbol{V}_1, \boldsymbol{V}_2\rangle}{d}\right)^{ki}.$$

We define $\boldsymbol{V} = \boldsymbol{V}_1 \odot \boldsymbol{V}_2$, where $\odot$ denotes entry-wise product of vectors and,

$$\overline{V} = \frac{1}{d}\sum_{i=1}^{d} V_i.$$

Hence,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \int \overline{V}^{ki} \cdot S(\boldsymbol{V}_1) \cdot S(\boldsymbol{V}_2)\, \pi(\mathrm{d}\boldsymbol{V}_1)\, \pi(\mathrm{d}\boldsymbol{V}_2).$$

Since $\boldsymbol{V}_1, \boldsymbol{V}_2$ are independently sampled from the prior π and $\boldsymbol{V} = \boldsymbol{V}_1 \odot \boldsymbol{V}_2$, it is straightforward to check that $\boldsymbol{V}_1, \boldsymbol{V}$ are independent, uniformly random $\{\pm 1\}^d$ vectors and $\boldsymbol{V}_2 = \boldsymbol{V}_1 \odot \boldsymbol{V}$. Hence,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \int \overline{V}^{ki} \cdot S(\boldsymbol{V}_1) \cdot S(\boldsymbol{V} \odot \boldsymbol{V}_1)\, \pi(\mathrm{d}\boldsymbol{V}_1)\, \pi(\mathrm{d}\boldsymbol{V}). \tag{C.12}$$

Recall that, the collection of polynomials:

$$\left\{ \boldsymbol{V}^{\boldsymbol{r}} \stackrel{\text{def}}{=} \prod_{i=1}^{d} V_i^{r_i} : \boldsymbol{r} \in \{0,1\}^d \right\},$$

form an orthonormal basis for functions on the Boolean hypercube $\{\pm 1\}^d$ with respect to the uniform distribution $\pi = \mathsf{Unif}\left(\{\pm 1\}^d\right)$. Hence, we can expand $\boldsymbol{S}$ in this basis:

$$\boldsymbol{S}(\boldsymbol{V}) = \sum_{\boldsymbol{r} \in \{0,1\}^d} \hat{S}_{\boldsymbol{r}} \cdot \boldsymbol{V}^{\boldsymbol{r}},\ \hat{S}_{\boldsymbol{r}} \stackrel{\text{def}}{=} \int S(\boldsymbol{V}) \cdot \boldsymbol{V}^{\boldsymbol{r}} \pi(\mathrm{d}\boldsymbol{V}).$$

Substituting this in (C.12) gives us:

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \sum_{\boldsymbol{r},\boldsymbol{s} \in \{0,1\}^d} \hat{S}_{\boldsymbol{r}} \hat{S}_{\boldsymbol{s}} \int \overline{V}^{ki} \cdot \boldsymbol{V}_1^{\boldsymbol{r}+\boldsymbol{s}} \cdot \boldsymbol{V}^{\boldsymbol{s}}\, \pi(\mathrm{d}\boldsymbol{V}_1)\, \pi(\mathrm{d}\boldsymbol{V}).$$

Noting that, if $\boldsymbol{r} \neq \boldsymbol{s}$,

$$\int \boldsymbol{V}_1^{\boldsymbol{r}+\boldsymbol{s}}\, \pi(\mathrm{d}\boldsymbol{V}_1) = 0,$$

we obtain,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \sum_{\boldsymbol{r} \in \{0,1\}^d} \hat{S}_{\boldsymbol{r}}^2 \int \overline{V}^{ki} \cdot \boldsymbol{V}^{\boldsymbol{r}}\, \pi(\mathrm{d}\boldsymbol{V}).$$

Since $\|S\|_\pi \leq 1$, we know that $\sum_{\boldsymbol{r}} \hat{S}_{\boldsymbol{r}}^2 \leq 1$. When $\langle S, 1 \rangle_\pi = 0$, one additionally has $\hat{S}_{\boldsymbol{0}} = 0$. Hence,

$$\sup_{S: \|S\|_\pi \leq 1} \mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \max_{\boldsymbol{r} \in \{0,1\}^d} \int \overline{V}^{ki} \cdot \boldsymbol{V}^{\boldsymbol{r}}\, \pi(\mathrm{d}\boldsymbol{V}),$$

$$\sup_{\substack{S: \|S\|_\pi \leq 1 \\ \langle S,1 \rangle_\pi = 0}} \mathbb{E}_0[\overline{H}_i(\boldsymbol{X};S)^2] = \max_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \int \overline{V}^{ki} \cdot \boldsymbol{V}^{\boldsymbol{r}}\, \pi(\mathrm{d}\boldsymbol{V}).$$

The right hand sides of the above equations have been analyzed in Lemma I.1. Appealing to this result immediately yields the claims of this lemma. □

C.5.4. *Proof of Lemma C.4.*

PROOF OF LEMMA C.4. Note that the result for $q = 2$ follows from the discussion preceding this lemma. Hence we focus on proving the inequality when $q \geq 2$. Recalling Definition C.1, we have

$$\overline{H}_i(\boldsymbol{X}; S) \stackrel{\text{def}}{=} \int H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) \cdot S(\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V}).$$

Observe that for any fixed $\boldsymbol{V}$, the quantity

$$H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right)$$

can be expressed as a polynomial in $\boldsymbol{X}$ of degree i (see Fact I.5). Since,

$$\overline{H}_i(\boldsymbol{X}; S) \stackrel{\text{def}}{=} \int H_i\left(\frac{\langle \boldsymbol{X}, \boldsymbol{V}^{\otimes k}\rangle}{\sqrt{d^k}}\right) \cdot S(\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V}),$$

is a weighted linear combination of such polynomials, $\overline{H}_i(\boldsymbol{X}; S)$ is also a homogeneous polynomial in $\boldsymbol{X}$ of degree i. Hence, by the completeness of the Hermite polynomial basis, $\overline{H}_i(\boldsymbol{X}; S)$ must have a representation of the form:

$$\overline{H}_i(\boldsymbol{X}; S) = \sum_{\substack{\boldsymbol{c} \in \bigotimes^k \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \beta(\boldsymbol{c}; S) \cdot H_{\boldsymbol{c}}(\boldsymbol{X}),$$

for some coefficients $\beta(\boldsymbol{c}; S)$. While these coefficients can be computed, we will not need their exact formula for our discussion. Hence,

$$\sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) = \sum_{i=0}^{\infty} \sum_{\substack{\boldsymbol{c} \in \bigotimes^k \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \alpha_i \cdot \beta(\boldsymbol{c}; S) \cdot H_{\boldsymbol{c}}(\boldsymbol{X})$$

By Gaussian Hypercontractivity (Fact I.7),

$$\begin{aligned} \left\| \sum_{i=0}^{\infty} \alpha_i \cdot \overline{H}_i(\boldsymbol{X}; S) \right\|_q^2 &\leq \sum_{i=0}^{\infty} \sum_{\substack{\boldsymbol{c} \in \bigotimes^k \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \alpha_i^2 \cdot \beta(\boldsymbol{c}; S)^2 \cdot (q-1)^i \\ &= \sum_{i=0}^{\infty} \alpha_i^2 \cdot (q-1)^i \cdot \sum_{\substack{\boldsymbol{c} \in \bigotimes^k \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \beta(\boldsymbol{c}; S)^2 \end{aligned}$$

Observing that,

$$\mathbb{E}_0[\overline{H}_i(\boldsymbol{X}; S)^2] = \sum_{\substack{\boldsymbol{c} \in \bigotimes^k \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \beta(\boldsymbol{c}; S)^2,$$

yields the claim of the lemma. □

APPENDIX D: ADDITIONAL DISCUSSIONS FOR SYMMETRIC TENSOR PCA

D.1. Discussion for k-Tensor PCA with odd k. When k is odd, our computational lower bound for k-TPCA (Theorem 1) shows that any iterative algorithm which uses N samples, makes T passes through the dataset, and has a memory state of size s bits fails to solve k-TPCA if:

$$(N\lambda^2)\cdot T\cdot s \ll \sqrt{d^{k+1}}. \tag{D.1}$$

On the other hand, there are iterative algorithms [2, 5] for k-TPCA with odd k with a resource profile:

$$N\lambda^2 \asymp \sqrt{d^k}\cdot \mathrm{polylog}(d),\ T = \mathrm{polylog}(d),\ s \asymp d\cdot \mathrm{polylog}(d), \tag{D.2}$$

which succeed in estimating the unknown signal vector $\boldsymbol{V}$ *consistently* in the sense that the estimator $\hat{\boldsymbol{V}}$ computed by these algorithms satisfies:

$$\frac{\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle^2}{\|\boldsymbol{V}\|^2\|\hat{\boldsymbol{V}}\|^2} \to 1 \text{ as } d\to\infty. \tag{D.3}$$

We believe that the $\sqrt{d}$ gap between the resource lower bound in (D.1) and the upper bound in (D.2) arises due to our use of the Hellinger information. Specifically, our approach relies on showing that the Hellinger Information $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})$ between the signal vector $\boldsymbol{V}\sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$ and the transcript $\boldsymbol{Y}$ generated by an iterative algorithm which uses too few resources (when run in a distributed setting via the reduction in Fact 1) satisfies:

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y}) \to 0 \text{ as } d\to\infty. \tag{D.4}$$

Due to Fano's Inequality for Hellinger Information (Fact 2, see also Corollary C.1 for its instantiation for k-TPCA), showing (D.4) not only rules out consistent estimation (cf. (D.3)) but yields a stronger-than-desired result that any estimator $\hat{\boldsymbol{V}}$ computed via an iterative algorithm which uses too few resources fails to achieve *better-than-random estimation*:

$$\forall\, \epsilon > 0,\ \mathbb{P}\left(\frac{\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle^2}{\|\boldsymbol{V}\|^2\|\hat{\boldsymbol{V}}\|^2} \geq \frac{d^\epsilon}{d}\right) \to 0 \text{ as } d\to\infty. \tag{D.5}$$

For better-than-random estimation, the resource lower bound in (D.1) is, in fact, optimal. Specifically, for any arbitrarily small constant $\epsilon\in(0,1)$, there is an iterative algorithm with the following properties:

1. The algorithm has a resource profile of:

   $$N\lambda^2 \asymp \sqrt{d^{k+\epsilon}},\ T=1,\ s\asymp \sqrt{d^{1+\epsilon}}. \tag{D.6}$$

   In particular, it uses $N\cdot T\cdot s \asymp \sqrt{d^{k+1+2\epsilon}}$ resources, which matches the lower bound in (D.1) upto an arbitrarily small polynomial factor of d^ϵ.
2. The estimator $\hat{\boldsymbol{V}}_\epsilon$ computed by the algorithm satisfies:

   $$\frac{\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle^2}{\|\boldsymbol{V}\|^2\|\hat{\boldsymbol{V}}\|^2} \gtrsim \frac{d^\epsilon}{d} \text{ with high probability (say, 0.9).} \tag{D.7}$$

   In particular, this estimator works better-than-random and consequently, the Hellinger information between the signal $\boldsymbol{V}\sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$ the transcript $\boldsymbol{Y}$ generated by this algorithm in a distributed setting must satisfy $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\gtrsim 1$ to avoid contradicting Fano's Inequality (Corollary C.1).

The existence of an algorithm with the above properties (described below) shows that the resource bound in (D.1) is the best possible lower bound that can be obtained using the approach based on Hellinger information used in our work. In order to improve the lower bound, one would need to use other information measures. A natural approach would be to show that the mutual information $\mathbf{I}_{\mathrm{KL}}(\boldsymbol{V};\boldsymbol{Y})$ (based on KL divergence) satisfies $\mathbf{I}_{\mathrm{KL}}(\boldsymbol{V};\boldsymbol{Y}) \leq cd$ for a suitably constant c, which would rule out consistent estimation. However, the key challenge in bounding the mutual information $\mathbf{I}_{\mathrm{KL}}(\boldsymbol{V};\boldsymbol{Y})$ is that the analog of the "cut-and-paste property" [22] (see Fact B.2), which plays a crucial role in proving the general information bound that underlies all our results (Proposition 1), is not known for KL divergence. This is why we chose to use Hellinger information in our analysis.

The algorithm that satisfies the properties (D.6) and (D.7) is as follows. The idea is that since one is only allowed a memory state of size $s \asymp \sqrt{d^{1+\epsilon}} \ll d$, one tries to estimate only the first s coordinates of the unknown signal $\boldsymbol{V}$. To do so, one uses $N \asymp \sqrt{d^{k+\epsilon}}/\lambda^2$ samples $\boldsymbol{T}_{1:N}$ (where each $\boldsymbol{T}_i = \lambda \cdot d^{-\frac{k}{2}} \cdot \boldsymbol{V}^{\otimes k} + \boldsymbol{W}_i$ where $\boldsymbol{V} \sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$ is the unknown signal vector and $\boldsymbol{W}_i$ is the i.i.d. Gaussian noise tensor) to compute the $s \asymp \sqrt{d^{1+\epsilon}}$-dimensional statistic $\boldsymbol{t}$ whose entries are given by:

$$\forall\, \ell \in [s],\ t_\ell = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{T}_i(\underbrace{\boldsymbol{I}_d, \boldsymbol{I}_d, \dots, \boldsymbol{I}_d}_{q \overset{\text{def}}{=} (k-1)/2 \text{ times}}, \boldsymbol{e}_\ell) \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \sum_{j_1, j_2, \dots, j_q = 1}^{d} (\boldsymbol{T}_i)_{j_1, j_1, j_2, j_2, \dots, j_q, j_q, \ell}$$

In the above display $q = (k-1)/2$ and $\boldsymbol{e}_{1:s}$ are the standard basis vectors of $\mathbb{R}^s$. The above statistic can be computed in a single pass over the data set, so the algorithm satisfies the resource requirement in (D.6). The final estimator for $\boldsymbol{V}$ is obtained by appending $d - s$ zeros to $\boldsymbol{t}$ to obtain a d-dimensional vector:

$$\hat{\boldsymbol{V}}_\epsilon = (\boldsymbol{t}^\top, \underbrace{0, \dots, 0}_{d-s \text{ times}})^\top$$

To see why this algorithm yields a better-than-random estimate, we observe that the distribution of the statistic $\boldsymbol{t}$ is given by:

$$\boldsymbol{t} \overset{\mathrm{d}}{=} \frac{\lambda}{\sqrt{d}} \cdot \boldsymbol{V}_{[s]} + \sqrt{\frac{d^{\frac{k-1}{2}}}{N}} \cdot \boldsymbol{g}, \quad \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_s)\,.$$

In the above display $\boldsymbol{V}_{[s]} \in \{\pm 1\}^s$ represents the vector obtained by the first s coordinates of $\boldsymbol{V}$. As a result, we have the following lower bound on the dot product:

$$\langle \hat{\boldsymbol{V}}_\epsilon, \boldsymbol{V} \rangle \geq \frac{\lambda s}{\sqrt{d}} - \sqrt{\frac{d^{\frac{k-1}{2}}}{N}} \cdot |\langle \boldsymbol{g}, \boldsymbol{V}_{[s]} \rangle| \overset{\text{(a)}}{\gtrsim} \frac{\lambda s}{\sqrt{d}} - \sqrt{\frac{d^{\frac{k-1}{2}}}{N}} \cdot \sqrt{s} \overset{\text{(b)}}{\gtrsim} \lambda \sqrt{d^\epsilon}.$$

In the above display, step (a) follows from the fact that $\langle \boldsymbol{g}, \boldsymbol{V}_{[s]} \rangle \sim \mathcal{N}\left(0, \|\boldsymbol{V}_{[s]}\|^2\right)$ and hence satisfies $|\langle \boldsymbol{g}, \boldsymbol{V}_{[s]} \rangle| \lesssim \|\boldsymbol{V}_{[s]}\| = \sqrt{s}$ with high probability. Step (b) uses the fact that $N\lambda^2 \asymp \sqrt{d^{k+\epsilon}}$, $s \asymp \sqrt{d^{1+\epsilon}}$. We can also upper bound the norm:

$$\|\hat{\boldsymbol{V}}_\epsilon\| = \|\boldsymbol{t}\| \leq \frac{\lambda \|\boldsymbol{V}_{[s]}\|}{\sqrt{d}} + \sqrt{\frac{d^{\frac{k-1}{2}}}{N}} \cdot \|\boldsymbol{g}\| \overset{\text{(a)}}{\lesssim} \frac{\lambda\sqrt{s}}{\sqrt{d}} + \sqrt{\frac{d^{\frac{k-1}{2}}}{N}} \cdot \sqrt{s} \overset{\text{(b)}}{\lesssim} \frac{\lambda \cdot d^{\epsilon/4}}{d^{1/4}} + \lambda \lesssim \lambda.$$

In the above display, step (a) follows by observing that $\|\boldsymbol{V}_{[s]}\| = \sqrt{s}$ and $\|\boldsymbol{g}\| \lesssim \sqrt{s}$ (with high probability). Step (b) uses the fact that $N\lambda^2 \asymp \sqrt{d^{k+\epsilon}}$, $s \asymp \sqrt{d^{1+\epsilon}}$. Recalling that $\|\boldsymbol{V}\|^2 = d$, the above estimates yield the claim in (D.7). While the above discussion focused on k-TPCA, analogous considerations also apply to k-NGCA.

D.2. Linear Memory Algorithms for Symmetric Tensor PCA. As discussed in Section 5.4.1, an important consequence of the computational lower bound for k-TPCA (Theorem 1) for even k is that *(nearly) linear memory* iterative algorithms which use a memory state of size $s \asymp d\,\mathrm{polylog}(d)$ fail to solve k-TPCA using N samples and T iterations if:

$$(N\lambda^2) \cdot T \ll \sqrt{d^k}. \tag{D.8}$$

This appendix provides additional details to explain how many natural algorithms for symmetric Tensor PCA fit into the template of *(nearly) linear memory* iterative algorithms which use a memory state of size $s \asymp d\,\mathrm{polylog}(d)$ bits. Consequently, the run-time v.s. sample size trade-off implied by (D.8) applies to these algorithms. Consider a broad class of iterative algorithms that maintain a sequence of iterates $\boldsymbol{u}_t \in \mathbb{R}^d$ and run for a total of T iterations. At iteration t, the iterate $\boldsymbol{u}_t$ is generated using the dataset $\boldsymbol{X}_{1:N}$ and the previous iterate $\boldsymbol{u}_{t-1}$ as follows:

$$\boldsymbol{u}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i\{\psi_t(\boldsymbol{u}_{t-1}), \cdot\}, \tag{D.9}$$

where $\psi_t : \mathbb{R}^d \to \bigotimes^{k-1} \mathbb{R}^d$ maps the previous iterate $\boldsymbol{u}_{t-1}$ to an order-$(k-1)$ tensor; and for tensors $\boldsymbol{X} \in \bigotimes^{k} \mathbb{R}^d$ and $\boldsymbol{\Psi} \in \bigotimes^{k-1} \mathbb{R}^d$, the tensor contraction $\boldsymbol{X}\{\boldsymbol{\Psi}, \cdot\}$ yields a vector in $\mathbb{R}^d$ defined by

$$\boldsymbol{X}\{\boldsymbol{\Psi}, \cdot\}_i = \sum_{j_1, j_2, \dots, j_{k-1}} X_{j_1, j_2, \dots, j_{k-1}, i} \Psi_{j_1, j_2, \dots, j_{k-1}}.$$

One can implement T iterations of the above scheme using a memory bounded algorithm (recall Definition 1) with a memory state of size $s = d \cdot \mathrm{polylog}(d)$ bits and T passes through the data. In order to see this, let us first consider the situation when the memory bounded algorithm is allowed a real-valued memory state (instead of a Boolean memory state). In this situation, the update in (D.9) can be implemented using a memory bounded algorithm that maintains two d-dimensional state variables `PartialSum` $\in \mathbb{R}^d$ and `iterate` $\in \mathbb{R}^d$. This implementation is shown in Figure 1. By using $\mathrm{polylog}(d)$ bits to represent a real number, one can approximate the real-valued state variables `PartialSum` and `iterate` using a Boolean vector of size $d \cdot \mathrm{polylog}(d)$, while ensuring that the quantization error is negligible. Consequently, T iterations of (D.9) can be implemented using a memory bounded estimation algorithm with resource profile $(N, T, s = d \cdot \mathrm{polylog}(d))$. Hence, the run-time vs. sample size tradeoffs implied by (D.8) of the preceding paragraph also apply to iterative algorithms with update rules as in (D.9).

By suitably defining the function ψ_t in (D.9) one can obtain many algorithms for k-TPCA that have been proposed in prior works. This means that the run-time vs. sample size trade-off obtained in (D.8) also applies to such algorithms. Two examples include:

Tensor power method. The tensor power method is given by the iterations:

$$\boldsymbol{u}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \left\{\boldsymbol{u}_{t-1}^{\otimes k-1}, \cdot\right\}$$

Hence, the tensor power method can be obtained from the general iteration (D.9) by choosing:

$$\psi_t(\boldsymbol{u}) = \boldsymbol{u}^{\otimes k-1}.$$

```
Memory bounded implementation of the iterative algorithm with update rule (D.9).
Input: $\boldsymbol{X}_{1:N}$, a dataset of $N$ tensors.
Output: An estimator $\hat{\boldsymbol{V}} \in \mathbb{R}^d$.
Variables: iterate $\in \mathbb{R}^d$ and PartialSum $\in \mathbb{R}^d$

• For iteration $t \in \{1, 2, \dots, T\}$
  – Set PartialSum $= \boldsymbol{0}_d$.
  – For sample $i \in \{1, 2, \dots, N\}$
        PartialSum ← PartialSum + $\frac{\boldsymbol{X}_i\{\psi_t(\texttt{iterate}); \cdot\}}{N}$
  – Update iterate ← PartialSum.
• Return Estimator $\hat{\boldsymbol{V}} =$ iterate.
```

Fig 1: Memory bounded implementation of the iterative algorithm with update rule (D.9).

Spectral method with partial trace. This estimator is given by the leading eigenvector of the matrix $\boldsymbol{M}$ whose entries are constructed as follows:

$$M_{\alpha,\beta} \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \sum_{\gamma_1, \gamma_2, \dots, \gamma_\ell \in [d]} (\boldsymbol{X}_i)_{\gamma_1, \gamma_1, \gamma_2, \gamma_2, \dots, \gamma_\ell, \gamma_\ell, \alpha, \beta}.$$

In the above display, we defined $\ell \overset{\text{def}}{=} k/2 - 1$. This estimator is due to Hopkins et al. [21]; see Biroli, Cammarota and Ricci-Tersenghi [5] for a simple analysis. It can be verified that

$$\boldsymbol{M} \overset{\text{d}}{=} \frac{\lambda}{d} \boldsymbol{V}\boldsymbol{V}^\intercal + \sqrt{\frac{d^\ell}{N}} \cdot \boldsymbol{Z},$$

where $\boldsymbol{Z} \in \mathbb{R}^{d \times d}$ is a random $d \times d$ random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. In the regime when $\lambda \asymp 1$ and $N\lambda^2 \gg d^{\frac{k}{2}}$, the largest eigenvector of $\boldsymbol{M}$ yields a consistent estimator for $\boldsymbol{V}$. Furthermore, in this regime $\boldsymbol{M}$ exhibits a spectral gap of size $\Delta \gtrsim 1$ [see, e.g., 5, for detailed arguments]. Hence, the largest eigenvector of $\boldsymbol{M}$ can be computed by running $T \asymp \log(d)$ iterations of the power method beginning from a random initialization. The power iterations are given by the update rule

$$\boldsymbol{u}_t = \boldsymbol{M} \cdot \frac{\boldsymbol{u}_{t-1}}{\|\boldsymbol{u}_{t-1}\|}.$$

Recalling the formula for $\boldsymbol{M}$, we can express the update rule for the power iteration as

$$\boldsymbol{u}_t = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \left\{ \underbrace{\boldsymbol{I}_d \otimes \boldsymbol{I}_d \otimes \cdots \otimes \boldsymbol{I}_d}_{\ell \text{ times}} \otimes \frac{\boldsymbol{u}_{t-1}}{\|\boldsymbol{u}_{t-1}\|}, \cdot \right\}.$$

This is an instantiation of the general iteration in (D.9) with

$$\psi_t(\boldsymbol{u}) = \underbrace{\boldsymbol{I}_d \otimes \boldsymbol{I}_d \otimes \cdots \otimes \boldsymbol{I}_d}_{\ell \text{ times}} \otimes \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}.$$

APPENDIX E: PROOFS FOR ASYMMETRIC TENSOR PCA

E.1. Proof of Computational Lower Bound (Theorem 2). Similar to Theorem 1, we prove Theorem 2 by transferring a communication lower bound for distributed estimation protocols for k-ATPCA to memory bounded estimators for the same problem using the reduction in Fact 1.

In the (Bayesian) distributed setup for k-ATPCA, the parameter $\boldsymbol{V}$ is drawn from the prior

$$\pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{\sqrt{d^k}\cdot \boldsymbol{e}_{i_1}\otimes \boldsymbol{e}_{i_2}\cdots\otimes \boldsymbol{e}_{i_k} : i_1,i_2,\ldots,i_k\in[d]\}\right). \tag{E.1}$$

Here, $\boldsymbol{e}_i$ denotes the i-th standard basis vector in $\mathbb{R}^d$, so $\boldsymbol{V}\sim\pi$ is a uniformly random 1-sparse tensor. The tensors $\boldsymbol{X}_{1:N}$ are sampled i.i.d. from $\mu_{\boldsymbol{V}}$, and are distributed across $m=N$ machines with $n=1$ sample/machine. The execution of a distributed estimation protocol with parameters $(m,n=1,b)$ results in a transcript $\boldsymbol{Y}\in\{0,1\}^{mb}$ written on the blackboard.

We obtain the following corollary of Fano's Inequality for Hellinger Information (Fact 2).

COROLLARY E.1 (Fano's Inequality for k-ATPCA). *For any estimator $\hat{\boldsymbol{V}}(\boldsymbol{Y})$ for k-ATPCA computed by a distributed estimation protocol, and for any $t\in\mathbb{R}$, we have*

$$\inf_{\boldsymbol{V}\in\mathcal{V}}\mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle\boldsymbol{V},\hat{\boldsymbol{V}}\rangle|^2}{\|\boldsymbol{V}\|^2\|\hat{\boldsymbol{V}}\|^2}\geq\frac{t^2}{d^k}\right)\leq\frac{1}{t^2}+\sqrt{2\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)}.$$

PROOF. As in the proof of Corollary C.1, we apply Fact 2 with the following loss function:

$$\ell(\boldsymbol{V},\boldsymbol{U})\overset{\text{def}}{=}\begin{cases}1 & \text{if } \frac{|\langle\boldsymbol{V},\boldsymbol{U}\rangle|^2}{\|\boldsymbol{V}\|^2\|\boldsymbol{U}\|^2}<t^2/d^k,\\ 0 & \text{otherwise.}\end{cases}$$

We also compute a lower bound on $R_0(\pi)$: by Markov's inequality, for any fixed tensor $\boldsymbol{U}\in\bigotimes^k\mathbb{R}^d$ with $\|\boldsymbol{U}\|=1$, we have

$$\mathbb{P}\left(\frac{|\langle\boldsymbol{V},\boldsymbol{U}\rangle|^2}{\|\boldsymbol{V}\|^2}\geq\frac{t^2}{d^k}\right)\leq\frac{d^k}{t^2}\cdot\int_{\mathcal{V}}\frac{|\langle\boldsymbol{V},\boldsymbol{U}\rangle|^2}{\|\boldsymbol{V}\|^2}\pi(\mathrm{d}\boldsymbol{V})=\frac{\|\boldsymbol{U}\|^2}{t^2}=\frac{1}{t^2}.$$

Consequently $R_0(\pi)\geq 1-1/t^2$. The claim is now immediate from Fact 2. □

The main technical result is the following information bound for k-ATPCA.

PROPOSITION E.1. *Let $\boldsymbol{Y}\in\{0,1\}^{mb}$ be the transcript generated by a distributed estimation protocol for k-ATPCA with parameters $(m,1,b)$. Then*

$$\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)\leq C\left(\frac{\delta^2 mb}{d^k}+\frac{1}{m}\right),$$

where

$$\delta\overset{\text{def}}{=}\exp\left(\frac{3\lambda^2}{2}+2\lambda\sqrt{\log(d^k)+\log(m)}\right)-1.$$

In the above display C is a universal constant (independent of m,b,d,λ). In particular, in the scaling regime (as $d\to\infty$)

$$\lambda\asymp 1,\quad m\asymp d^\eta,\quad b\asymp d^\beta$$

for any constants $\eta\geq1$ and $\beta\geq0$ that satisfy

$$\eta+\beta<k,$$

we have $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)\to 0$ as $d\to\infty$.

With this information bound in hand, we first present the proof of the computational lower bound for k-ATPCA (Theorem 2) and defer the proof of the above information bound to the following section in this appendix (Appendix E.2).

PROOF OF THEOREM 2. Appealing to the reduction in Fact 1 with the choice $n = 1$, we note that any memory-bounded estimator $\hat{\boldsymbol{V}}$ with resource profile (N, T, s) can be implemented using a distributed algorithm with parameters $(N, 1, sT)$. Applying Corollary E.1 and Proposition E.1 to the distributed implementation of the memory-bounded estimator immediately yields Theorem 2. □

We end this section with the following remark, which discusses the connection between k-ATPCA and the sparse Gaussian mean estimation problem studied in prior work [10, 1].

REMARK E.1 (Connection with Sparse Gaussian Mean Estimation). Observe that due to the choice of the prior in (E.1), the instance of k-ATPCA used to obtain the communication lower bound is also an instance of the 1-sparse Gaussian mean estimation problem in dimension $D = d^k$. Communication lower bounds for this problem in the blackboard model (cf. Definition 2) were obtained in prior work by Braverman et al. [10]. This result is sufficient to obtain Theorem 2. Recent work by Acharya et al. [1] also provides an alternate proof for the communication lower bounds for sparse Gaussian mean estimation. We present another proof of these results using the information bound in Proposition 1, which is used to derive all communication lower bounds presented in this paper.

E.2. Proof of the Information Bound (Proposition E.1). This section is devoted to the proof of the information bound for distributed k-ATPCA (Proposition E.1). Recall that in the distributed k-ATPCA problem:

1. An unknown rank-1 signal tensor $\boldsymbol{V} \in \bigotimes^k \mathbb{R}^d$ is drawn from the prior:
$$\pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{\sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \cdots \otimes \boldsymbol{e}_{i_k} : i_1, i_2, \dots, i_k \in [d]\}\right).$$
2. A dataset consisting of m tensors $\boldsymbol{X}_{1:m}$ is drawn i.i.d. from $\mu_{\boldsymbol{V}}$, where $\mu_{\boldsymbol{V}}$ is the distribution of a single tensor from the k-ATPCA problem:

   (E.2)
$$\boldsymbol{X}_i = \frac{\lambda \boldsymbol{V}_1 \otimes \boldsymbol{V}_2 \cdots \otimes \boldsymbol{V}_k}{\sqrt{d^k}} + \boldsymbol{W}_i, \quad (W_i)_{j_1, j_2, \cdots j_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad \forall j_1, j_2, \dots, j_k \in [d].$$
   This dataset is divided among m machines with 1 tensor per machine.
3. The execution of a distributed estimation protocol with parameters $(m, n = 1, b)$ results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

We will obtain Proposition E.1 by instantiating the general information bound provided in Proposition 1 with the following choices:

Choice of μ_0: Under the reference measure, $\boldsymbol{X} \sim \mu_0$ is a k-tensor with i.i.d. $\mathcal{N}(0,1)$ coordinates.
Choice of $\overline{\mu}$: The null measure is set to $\overline{\mu} \overset{\text{def}}{=} \mu_0$.
Choice of $\mathcal{Z}$: We choose the event $\mathcal{Z}$ as follows:
$$\mathcal{Z} \overset{\text{def}}{=} \left\{\boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k} : \|\boldsymbol{X}\|_\infty \leq \lambda + \sqrt{2(k\log(d) + \epsilon)}\right\}.$$

With these choices, we can write down the formula for the relevant likelihood ratio that appears Proposition 1. Note that if $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{e}_{j_1} \otimes \boldsymbol{e}_{j_2} \cdots \otimes \boldsymbol{e}_{j_k}$, then the likelihood ratio has the following formula:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \exp\left(\lambda (X)_{j_1, j_2 \dots, j_k} - \frac{\lambda^2}{2}\right). \tag{E.3}$$

Consequently, we define:

$$\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(x) \overset{\text{def}}{=} \exp\left(\lambda x - \frac{\lambda^2}{2}\right).$$

Note that this is exactly the likelihood ratio between $\mu_\lambda = \mathcal{N}(\lambda, 1)$ and $\mu_0 = \mathcal{N}(0,1)$. Hence, when $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{e}_{j_1} \otimes \boldsymbol{e}_{j_2} \cdots \otimes \boldsymbol{e}_{j_k}$, we have:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) = \frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(X_{j_1, j_2 \dots, j_k}).$$

The proof of Proposition E.1 relies on two intermediate results. First, Lemma E.1 analyzes the concentration properties of a suitably truncated version of the Likelihood ratio.

LEMMA E.1. *Let $\boldsymbol{S} \in (\mathbb{R}^d)^{\otimes k}$ be an arbitrary tensor with $\|\boldsymbol{S}\| \leq 1$. Let $\epsilon > \lambda$ be arbitrary scalar. Let $\boldsymbol{X} \sim \mu_0$ be Gaussian tensor with $i.i.d.$ $\mathcal{N}(0,1)$ entries. Define the tensor $\boldsymbol{T}^{\leq \epsilon} \in (\mathbb{R}^d)^{\otimes k}$ with entries:*

$$(T^{\leq \epsilon})_{j_1, j_2, \dots, j_k} \overset{\text{def}}{=} \left(\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(X_{j_1, j_2 \dots, j_k}) - 1\right) \cdot \mathbb{I}_{|X_{j_1, j_2 \dots, j_k}| \leq \epsilon}$$

Then,

1. *There exists a universal constant C such that for any $q \in \mathbb{N}$,*

$$(\mathbb{E}_0 |\left\langle \boldsymbol{S}, \boldsymbol{T}^{\leq \epsilon} - \mathbb{E}_0 \boldsymbol{T}^{\leq \epsilon} \right\rangle|^q)^{\frac{2}{q}} \leq C \cdot q \cdot \delta^2.$$

In the above display $\langle \cdot, \cdot \rangle$ denotes the standard inner product:

$$\left\langle \boldsymbol{S}, \boldsymbol{T}^{\leq \epsilon} - \mathbb{E}_0 \boldsymbol{T}^{\leq \epsilon} \right\rangle = \sum_{j_1, j_2, \dots, j_k \in [d]} (T^{\leq \epsilon} - \mathbb{E}_0 T^{\leq \epsilon})_{j_1, j_2, \dots, j_k} (S)_{j_1, j_2, \dots, j_k}.$$

and,

$$\delta \overset{\text{def}}{=} \max\left(e^{\lambda \epsilon - \frac{\lambda^2}{2}} - 1, \lambda \epsilon + \frac{\lambda^2}{2}\right).$$

2. *Furthermore,*

$$\|\mathbb{E}\boldsymbol{T}^{\leq \epsilon}\|^2 \leq 4 \cdot d^k \cdot e^{-(\epsilon - \lambda)^2}.$$

PROOF. The proof of this result appears at the end of this appendix (Appendix E.3). □

The second result required to complete the proof of Proposition E.1 is Lemma E.2. This result provides an estimate on $\mu_{\boldsymbol{V}}(\mathcal{Z}^c)$ which appears in the information bound in Proposition 1.

LEMMA E.2. *Let $\epsilon \geq 0$ be arbitrary. Let $\mathcal{Z}$ denote the set:*

$$\mathcal{Z} = \left\{ \boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k} : \|\boldsymbol{X}\|_\infty \leq \lambda + \sqrt{2(k\log(d) + \epsilon)} \right\}.$$

Then, $\mu_{\boldsymbol{V}}(\mathcal{Z}^c) \leq e^{-\epsilon}$. In the above display, $\|\boldsymbol{X}\|_\infty$ denotes the entry-wise ∞-norm:

$$\|\boldsymbol{X}\|_\infty \overset{def}{=} \max_{j_{1:k} \in [d]} |X_{j_1, j_2, \dots, j_k}|.$$

PROOF. Recall that under $\mu_{\boldsymbol{V}}$, we have:

$$\boldsymbol{X} = \frac{\lambda \boldsymbol{V}_1 \otimes \boldsymbol{V}_2 \cdots \otimes \boldsymbol{V}_k}{\sqrt{d^k}} + \boldsymbol{W}, \ (W)_{j_1, j_2, \cdots j_k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \ \forall\, j_1, j_2 \cdots, j_k \in [d].$$

The claim follows by observing $\|\boldsymbol{V}_i\|_\infty \leq \|\boldsymbol{V}_i\| = \sqrt{d}$ along with the standard tail bound for maximum of Gaussian random variables:

$$\mathbb{P}\left(\|\boldsymbol{W}\|_\infty \geq \sqrt{2(k\log(d) + \epsilon)} \right) \leq e^{-\epsilon}.$$

□

With these results in hand, we now present the proof of Proposition E.1.

PROOF OF PROPOSITION E.1. Recall that in Proposition 1 we showed:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}{K} \leq$$

$$\sum_{i=1}^m \mathbb{E}_0 \left[\mathbb{I}_{Z_i = 1} \cdot \int \left(\mathbb{E}_0 \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1 \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right] + m \int \mu_{\boldsymbol{V}}(\mathcal{Z}^c) \pi(\mathrm{d}\boldsymbol{V}).$$

We will choose:

$$\mathcal{Z} \overset{\text{def}}{=} \left\{ \boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k} : \|\boldsymbol{X}\|_\infty \leq \epsilon \right\}, \ \epsilon \overset{\text{def}}{=} \lambda + 2\sqrt{(k\log(d) + \log(m))}.$$

By Lemma E.2, we have, $\mu_{\boldsymbol{V}}(\mathcal{Z}^c) \leq 1/m^2$. Hence,

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}{K} \leq \sum_{i=1}^m \mathbb{E}_0 \left[\mathbb{I}_{Z_i = 1} \cdot \int \left(\mathbb{E}_0 \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1 \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right] + \frac{1}{m}.$$

Next we recall that we chose

$$\pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{ \sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \cdots \otimes \boldsymbol{e}_{i_k} : \ i_{1:k} \in [d] \} \right).$$

And when $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{e}_{j_1} \otimes \boldsymbol{e}_{j_2} \cdots \otimes \boldsymbol{e}_{j_k}$, we simplified the likelihood ratio:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1 = \frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}((X_i)_{j_1, j_2, \dots, j_k}) - 1, \ \frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(x) \overset{\text{def}}{=} \exp\left(\lambda x - \frac{\lambda^2}{2} \right).$$

We define the tensors $\boldsymbol{T}_i \in (\mathbb{R}^d)^{\otimes k}$ we entries:

$$T_{j_1, j_2, \dots, j_k} = \frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}((X_i)_{j_1, j_2, \dots, j_k}) - 1.$$

With this notation, the upper bound on Hellinger Information can be written as:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}{K} \leq \frac{1}{d^k} \sum_{i=1}^m \mathbb{E}_0 \left[\mathbb{I}_{Z_i = 1} \cdot \left\| \mathbb{E}_0 \left[\boldsymbol{T}_i \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right\|^2 \right] + \frac{1}{m}.$$

Truncation of Likelihood Ratio. Recall that $Z_i = \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}}$. Consequently, we only need to analyze the likelihood ratio on $\mathcal{Z}$. Note that on $\mathcal{Z}$, we have, $\boldsymbol{T}_i = \boldsymbol{T}_i^{\leq \epsilon}$, where:

$$(T^{\leq \epsilon})_{j_1,j_2,\dots,j_k} \overset{\text{def}}{=} \left(\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(X_{j_1,j_2\dots,j_k}) - 1\right) \cdot \mathbb{I}_{|X_{j_1,j_2\dots,j_k}| \leq \epsilon},$$

$$\epsilon \overset{\text{def}}{=} \lambda + 2\sqrt{(k\log(d) + \log(m))}.$$

Note that:

$$\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2 \leq 2\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2 + 2\|\mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\|^2.$$

Using the bound on $\|\mathbb{E}_0\boldsymbol{T}^{\leq\epsilon}\|^2$ obtained in Lemma E.1, we obtain the following bound on Hellinger information:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}{K} \leq \frac{2}{d^k}\sum_{i=1}^m \mathbb{E}_0\left[\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2\right] + \frac{4}{d^{4k}m^3} + \frac{1}{m}.$$

Linearization and Geometric Inequality. We observe that:

$$\begin{aligned}&\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\| \\ &\qquad\qquad \leq \sup_{\boldsymbol{S}\in(\mathbb{R}^d)^{\otimes k}:\|\boldsymbol{S}\|\leq 1} \mathbb{E}_0\left[\left\langle \boldsymbol{S}, \boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\right\rangle\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right].\end{aligned}$$

Using the Proposition 2 along with the moment bounds in Lemma E.1, we obtain:

$$\begin{aligned}&\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2 \\ &\qquad\qquad\qquad \leq \inf_{q\geq 1} \frac{C\cdot\delta^2\cdot q}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i})^{\frac{2}{q}}},\end{aligned}$$

where

$$\begin{aligned}\delta &= \max\left(\exp\left(2\lambda\sqrt{(k\log(d)+\log(m))} + \frac{\lambda^2}{2}\right) - 1, \frac{3\lambda^2}{2} + 2\lambda\sqrt{(k\log(d)+\log(m))}\right) \\ &\leq \exp\left(2\lambda\sqrt{(k\log(d)+\log(m))} + \frac{3\lambda^2}{2}\right) - 1.\end{aligned}$$

We define:

$$\mathcal{R}^{(i)}_{\mathsf{freq}} \overset{\text{def}}{=} \left\{(\boldsymbol{y}, \boldsymbol{z}_i) \in \{0,1\}^{mb+1} : \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) > \frac{1}{e}\right\}.$$

Note that $|\mathcal{R}^{(i)}_{\mathsf{freq}}| \leq e$. We set q as:

$$q = \begin{cases} 1 & \text{if } (\boldsymbol{y}, \boldsymbol{z}_i) \in \mathcal{R}^{(i)}_{\mathsf{freq}}; \\ -2\log\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) & \text{if } (\boldsymbol{y}, \boldsymbol{z}_i) \notin \mathcal{R}^{(i)}_{\mathsf{freq}}.\end{cases}$$

This ensures $q \geq 1$. Setting q as above yields:

$$\begin{aligned}&\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon} - \mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2 \\ &\qquad\qquad \leq \begin{cases} C\delta^2 & \text{if } (\boldsymbol{y}, \boldsymbol{z}_i) \in \mathcal{R}^{(i)}_{\mathsf{freq}}; \\ -2C\delta^2\log\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y}, Z_i = z_i|(\boldsymbol{X}_j)_{j\neq i} = (\boldsymbol{x}_j)_{j\neq i}) & \text{if } (\boldsymbol{y}, \boldsymbol{z}_i) \notin \mathcal{R}^{(i)}_{\mathsf{freq}}.\end{cases}\end{aligned}$$

Hence we obtain:

$$\mathbb{E}\left[\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon}-\mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}=\boldsymbol{y},Z_i=z_i,(\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2\right] \\ \leq C|\mathcal{R}^{(i)}_{\mathsf{freq}}|\delta^2+C\delta^2\mathsf{H}(\boldsymbol{Y},Z_i|(\boldsymbol{X}_j)_{j\neq i}).$$

In the above display $\mathsf{H}(\boldsymbol{Y},Z_i|(\boldsymbol{X}_j)_{j\neq i})$ denotes the conditional entropy of the $(\boldsymbol{Y},Z_i)$ given $(\boldsymbol{X}_j)_{j\neq i}$. Since we assumed that the communication protocol used to generate $\boldsymbol{Y}$ is deterministic, conditioning on $(\boldsymbol{X}_j)_{j\neq i}$ determines all but $b+1$ bits of the vector $(\boldsymbol{Y},Z_i)$. Hence $\mathsf{H}(\boldsymbol{Y},Z_i|(\boldsymbol{X}_j)_{j\neq i})\leq C(b+1)$. This gives us,

$$\mathbb{E}\left[\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon}-\mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y}=\boldsymbol{y},Z_i=z_i,(\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2\right]\leq C\delta^2 b,$$

which in turn yields:

$$\begin{aligned}\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})&\leq K\left(\frac{2}{d^k}\sum_{i=1}^m\mathbb{E}_0\left[\left\|\mathbb{E}_0\left[\boldsymbol{T}_i^{\leq\epsilon}-\mathbb{E}_0\boldsymbol{T}_i^{\leq\epsilon}\middle|\boldsymbol{Y},Z_i,(\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2\right]+\frac{4}{d^{4k}m^3}+\frac{1}{m}\right)\\&\leq C\left(\frac{\delta^2 mb}{d^k}+\frac{1}{d^{4k}m^3}+\frac{1}{m}\right)\\&\leq C\left(\frac{\delta^2 mb}{d^k}+\frac{1}{m}\right).\end{aligned}$$

Finally we observe that in the scaling regime: $\lambda=\Theta(1),\ m=\Theta(d^\eta),\ b=\Theta(d^\beta)$, for some constants $\eta\geq 1,\beta\geq 0$, which satisfy $\eta+\beta<k$, the above upper bound on $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})\to 0$ as $d\to\infty$. □

E.3. Concentration of Likelihood Ratio. In this section, we provide a proof for Lemma E.1.

PROOF OF LEMMA E.1. We prove the two parts separately.

1. Recall that,

$$\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(x)\overset{\text{def}}{=}\exp\left(\lambda x-\frac{\lambda^2}{2}\right).$$

Observe that,

$$(T^{\leq\epsilon})_{j_1,j_2,\dots,j_k}\overset{\text{def}}{=}\left(\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(X_{j_1,j_2\dots,j_k})-1\right)\cdot\mathbb{I}_{|X_{j_1,j_2\dots,j_k}|\leq\epsilon}\leq e^{\lambda\epsilon-\frac{\lambda^2}{2}}-1.$$

Furthermore,

$$(T^{\leq\epsilon})_{j_1,j_2,\dots,j_k}\geq\left(e^{-\lambda\epsilon-\frac{\lambda^2}{2}}-1\right)\cdot\mathbb{I}_{|X_{j_1,j_2\dots,j_k}|\leq\epsilon}\geq-\lambda\epsilon-\frac{\lambda^2}{2}.$$

Hence,

$$\left|(T^{\leq\epsilon})_{j_1,j_2,\dots,j_k}\right|\leq\max\left(e^{\lambda\epsilon-\frac{\lambda^2}{2}}-1,\lambda\epsilon+\frac{\lambda^2}{2}\right)\overset{\text{def}}{=}\delta.$$

Hence, $(T^{\leq\epsilon})_{j_1,j_2,\dots,j_k}$ are independent random variables in $[-\delta,\delta]$. Consequently, by Hoeffding's Inequality $\left\langle\boldsymbol{S},\boldsymbol{T}^{\leq\epsilon}-\mathbb{E}_0\boldsymbol{T}^{\leq\epsilon}\right\rangle$ is sub-Gaussian with variance proxy δ^2.The claim follows by standard estimates on the moments of sub-Gaussian random variables.

2. Since the entries of $\boldsymbol{T}^{\leq \epsilon}$ are identically distributed, we have:

$$\|\mathbb{E}\boldsymbol{T}^{\leq \epsilon}\|^2 = d^k (\mathbb{E}(T^{\leq \epsilon})_{1,1,1,\dots,1})^2.$$

Recall that:

$$(T^{\leq \epsilon})_{1,1,1,\dots,1} \overset{\mathrm{d}}{=} \left(\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(Z) - 1\right) \cdot \mathbb{I}_{|Z| \leq \epsilon},\ Z \sim \mathcal{N}(0,1)\,.$$

Hence,

$$\begin{aligned}
|\mathbb{E}(T^{\leq \epsilon})_{1,1,1,\dots,1}| &= \left|\mathbb{E}\left[\left(\frac{\mathrm{d}\mu_\lambda}{\mathrm{d}\mu_0}(Z) - 1\right) \cdot \mathbb{I}_{|Z| \leq \epsilon}\right]\right| \\
&= |\mathbb{P}(|Z| \leq \epsilon) - \mathbb{P}(|Z+\lambda| \leq \epsilon)| \\
&= |\mathbb{P}(|Z+\lambda| \geq \epsilon) - \mathbb{P}(|Z| \geq \epsilon)| \\
&\leq \max(\mathbb{P}(|Z+\lambda| \geq \epsilon), \mathbb{P}(|Z| \geq \epsilon)) \\
&\leq \mathbb{P}(|Z| \geq \epsilon - \lambda) \\
&\leq 2e^{-\frac{(\epsilon-\lambda)^2}{2}}.
\end{aligned}$$

Hence,

$$\|\mathbb{E}\boldsymbol{T}^{\leq \epsilon}\|^2 \leq 4 \cdot d^k \cdot e^{-(\epsilon-\lambda)^2}.$$

□

APPENDIX F: PROOFS FOR NON-GAUSSIAN COMPONENT ANALYSIS

F.1. Proof of Computational Lower Bound (Theorem 3). This appendix is devoted to the proof of the computational lower bound for the order-k Non-Gaussian Component Analysis (k-NGCA) problem. Similar to Theorems 1 and 2, we prove Theorem 3 by transferring a communication lower bound for distributed estimation protocols for k-NGCA to memory bounded estimators for the same problem using the reduction in Fact 1.

In the (Bayesian) distributed setup for k-NGCA, the parameter $\boldsymbol{V}$ is drawn from the prior $\pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{\pm 1\}^d\right)$, and then $\boldsymbol{x}_{1:N}$ are sampled i.i.d. from $\mu_{\boldsymbol{V}}$; these samples are distributed across $m = N/n \in \mathbb{N}$ machines with n samples/machine. We will obtain Theorem 3 with a suitable choice of n. As usual, the execution of a distributed estimation protocol with parameters (m,n,b) results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

We have the following corollary of Fano's Inequality for Hellinger Information (Fact 2), proved in exactly the same way as Corollary C.1.

COROLLARY F.1 (Fano's Inequality for k-NGCA). *For any estimator $\hat{\boldsymbol{V}}(\boldsymbol{Y})$ for k-NGCA computed by a distributed estimation protocol, and for any $t \in \mathbb{R}$, we have*

$$\inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d}\right) \leq 2\exp\left(-\frac{t^2}{2}\right) + \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}.$$

The main technical result is the following information bound for k-NGCA.

PROPOSITION F.1. *Consider the k-NGCA problem with non-Gaussian distribution ν satisfying*

1. *the Moment Matching Assumption (Assumption 1) with parameter $k \geq 2$;*

2. *the Bounded Signal Strength Assumption (Assumption 2) with parameters* (λ, K)*;*
3. *the Locally Bounded Likelihood Ratio Assumption (Assumption 3) with parameters* (λ, K, κ).

Let $\boldsymbol{Y} \in \{0,1\}^{mb}$ *be the transcript generated by a distributed estimation protocol for this* k*-NGCA problem with parameters* (m,n,b)*. Let* $q \geq 2$ *be arbitrary but fixed constant. Then, there is a finite constant* $C_{k,K,\kappa,q}$ *depending only on* (k, K, κ, q) *such that if*

$$n \geq C_{k,K,\kappa,q} \cdot b \cdot d^{\kappa + \lceil \frac{k+1}{2} \rceil} \quad \textit{and} \quad n\lambda^2 \leq \frac{1}{C_{k,K,\kappa,q}}, \tag{F.1}$$

then

$$\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right) \leq C_{k,K,\kappa,q} \cdot \left(b \cdot mn\lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil} + m \cdot (n\lambda^2)^2 + \frac{m}{d^{\frac{q}{2}}} + m \cdot n \cdot (m+n) \cdot e^{-d/C_{k,K,\kappa,q}}\right).$$

With this information bound in hand, we first present the proof of the computational lower bound for k-NGCA (Theorem 3) and defer the proof of the above information bound to a later section in this appendix (Appendix F.4).

PROOF OF THEOREM 3. Appealing to the reduction in Fact 1, we note that any memory-bounded estimator $\hat{\boldsymbol{V}}$ with resource profile (N, T, s) can be implemented using a distributed estimation protocol with parameters $(N/n, n, sT)$ for any $n \in \mathbb{N}$ such that $m := N/n \in \mathbb{N}$. As assumed in Theorem 3, we consider the situation when

$$\eta + \tau + \mu < \left\lceil \frac{k+1}{2} \right\rceil, \quad \gamma > 2\left\lceil \frac{k+1}{2} \right\rceil + \kappa. \tag{F.2}$$

We set $n = d^{\xi}$ with

$$\xi \overset{\text{def}}{=} \tau + \mu + \kappa + \left\lceil \frac{k+1}{2} \right\rceil + \frac{1}{2}\underbrace{\left(\left\lceil \frac{k+1}{2} \right\rceil - (\eta + \tau + \mu)\right)}_{>0} > \tau + \mu + \kappa + \left\lceil \frac{k+1}{2} \right\rceil. \tag{F.3}$$

With this choice, we verify that the information bound in Proposition F.1 shows that $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right) \to 0$; combining this with Corollary F.1 proves the theorem. We begin by observing

$$\begin{aligned}
\gamma &> \left\lceil \frac{k+1}{2} \right\rceil + \kappa + \eta + \tau + \mu + \left(\left\lceil \frac{k+1}{2} \right\rceil - (\eta + \tau + \mu)\right) \\
&= \eta + \xi + \frac{1}{2}\left(\left\lceil \frac{k+1}{2} \right\rceil - (\eta + \tau + \mu)\right) \\
&> \eta + \xi.
\end{aligned}$$

Next, we verify the conditions required for Proposition F.1:

1. Since $\eta > \tau + \mu + \kappa + \lceil (k+1)/2 \rceil$ we have $n \gg b \cdot d^{\kappa + \lceil \frac{k+1}{2} \rceil}$ as required.
2. Since $\gamma > \eta + \xi > \xi$ we have $n\lambda^2 \ll 1$ as required.

Now, from the information bound of Proposition F.1, for any $q \geq 2$, we have:

$$\begin{aligned}
&\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right) \\
&\leq C_{k,K,\kappa,q} \cdot \left(b \cdot mn\lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil} + m \cdot (n\lambda^2)^2 + \frac{m}{d^{\frac{q}{2}}} + m \cdot n \cdot (m+n) \cdot e^{-d/C_{k,K,\kappa,q}}\right)
\end{aligned}$$

$$= C_{k,K,\kappa,q} \cdot \left(b \cdot N\lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil} + (N\lambda^2) \cdot (n\lambda^2) + \frac{m}{d^{\frac{q}{2}}} + m \cdot n \cdot (m+n) \cdot e^{-d/C_{k,K,\kappa,q}} \right).$$

We now check that this bound on $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right)$ vanishes with $d \to \infty$ with a suitable choice of q:

1. The assumption $\eta + \tau + \mu < \lceil (k+1)/2 \rceil$ guarantees $b \cdot N\lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil} \to 0$.
2. Since $\gamma > \eta + \xi$, we have $(N\lambda^2) \cdot (n\lambda^2) \to 0$.
3. Observe that $m = (N\lambda^2)/(n\lambda^2) = d^{\gamma+\eta-\xi} \geq d^{2\eta}$. Hence, choosing $q = 2(\gamma+\eta)+\eta$ ensures $m/d^{q/2} \to 0$.
4. Since n, m scale polynomially with d, we have $m \cdot n \cdot (m+n) \cdot e^{-d/C_{k,K,\kappa,q}} \to 0$.

This concludes the proof. □

REMARK F.1. The computational lower bound in Theorem 3 requires that λ^2 is sufficiently small because the information bound in Proposition F.1 holds when n is sufficiently large and $n\lambda^2$ is sufficiently small (F.1). In light of this, a natural question is whether an information bound of the form:

$$\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right) \overset{??}{\lesssim} b \cdot mn\lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil}, \tag{F.4}$$

holds without assuming λ^2 is small. Unfortunately, an information of the form (F.4) is ruled out by a simple distributed estimator unless $\lambda^2 \lesssim d^3 \operatorname{polylog}(d)/d^{\lceil \frac{k+1}{2} \rceil}$. This distributed estimator uses the information theoretically optimal sample size of $N = mn \asymp d/\lambda^2$ distributed across $m = N/n$ machines with n samples per machine. The estimator simply writes the entire dataset on the blackboard as the transcript and computes the Maximum Likelihood Estimator using the transcript. Since each sample is a d-dimensional real-valued vector, which can be quantized to a $d\operatorname{polylog}(d)$ bit vector, the total bits written on the black board are $mb = Nd\operatorname{polylog}(d) = d^2\operatorname{polylog}(d)/\lambda^2$. Since the maximum likelihood estimator is consistent, we must have $\mathbf{I}_{\mathsf{hel}}\left(\boldsymbol{V};\boldsymbol{Y}\right) \gtrsim 1$. This leads to a contradiction to (F.4) unless $n \gtrsim d^{\lceil \frac{k+1}{2} \rceil - 2}/\operatorname{polylog}(d)$. Since $n \leq N \asymp d/\lambda^2$, this means that the information bound (F.4) cannot hold unless $\lambda^2 \lesssim d^3 \operatorname{polylog}(d)/d^{\lceil \frac{k+1}{2} \rceil}$. For sufficiently large k ($k \geq 5$), this means that (F.4) cannot hold unless $\lambda^2 \ll 1$.

F.2. Proof Overview of the Information Bound (Proposition F.1).

The remainder of this appendix is devoted to the proof Proposition F.1, the information bound for the distributed NGCA problem. Recall that in the distributed k-NGCA problem:

1. An unknown d dimensional parameter $\boldsymbol{V} \sim \pi$ is drawn from the prior $\pi = \mathsf{Unif}\left(\{\pm 1\}^d\right)$.
2. A dataset $\{\boldsymbol{x}_{ij} : i \in [m], j \in [n]\}$ consisting of $N = mn$ samples is drawn i.i.d. from $\mu_{\boldsymbol{V}}$, where $\mu_{\boldsymbol{V}}$ is the distribution of a single sample from the Non-Gaussian Component Analysis problem. Recall that this means that:

$$\boldsymbol{x}_{ij} = \eta_{ij} \frac{1}{\sqrt{d}} \boldsymbol{V} + \left(\boldsymbol{I}_d - \frac{1}{d} \boldsymbol{V}\boldsymbol{V}^{\intercal} \right) \boldsymbol{z}_{ij}, \tag{F.5a}$$

where $\eta_{ij} \in \mathbb{R}$ and $\boldsymbol{z}_{ij} \in \mathbb{R}^d$ are independent random variables with distributions

$$\boldsymbol{z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_d\right), \quad \eta_{ij} \overset{\text{i.i.d.}}{\sim} \nu. \tag{F.5b}$$

In the above display, ν is a non-Gaussian distribution on $\mathbb{R}$ and $\mu_{\boldsymbol{V}}$ denotes the distribution of $\boldsymbol{x}_i$ described by the above generating process (F.5).

3. This dataset is divided among m machines with n samples per machine. We denote the dataset in one machine by $\boldsymbol{X}_i \in \mathbb{R}^{d\times n}$, where,
$$\boldsymbol{X}_i = \left[\boldsymbol{x}_{i1}\ \boldsymbol{x}_{i2}\ \dots\ \boldsymbol{x}_{in}\right].$$
Since $\boldsymbol{x}_{ij} \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}$, $\boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}^{\otimes n}$.
4. The execution of a distributed estimation protocol with parameters (m, n, b) results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

The information bound stated in Proposition F.1 is obtained using the general information bound given in Proposition 1 with the following choices:

Choice of μ_0: Under the measure μ_0, $\boldsymbol{x}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ for any $i \in [m]$, $j \in [n]$.
Choice of $\overline{\mu}$: Under the measure $\overline{\mu}$, the dataset of each machine is sampled i.i.d. from:
$$\mu_0(\cdot) = \int \mu_{\boldsymbol{V}}^{\otimes n}(\cdot)\, \pi(\mathrm{d}\boldsymbol{V}).$$

Note that the data across machines is independent, but the n samples within a machine are dependent since the were sampled from the same $\mu_{\boldsymbol{V}}$.
Choice of $\mathcal{Z}$: We choose the event $\mathcal{Z}$ as follows:

$$\mathcal{Z} \overset{\text{def}}{=} \mathcal{Z}_1 \cap \mathcal{Z}_2, \tag{F.6a}$$

$$\mathcal{Z}_1 \overset{\text{def}}{=} \{\boldsymbol{x}_{1:n} \in \mathbb{R}^d : \|\boldsymbol{x}_i\| \leq \sqrt{2d}\, \forall\, i\ \in\ [n]\}, \tag{F.6b}$$

$$\mathcal{Z}_2 \overset{\text{def}}{=} \left\{\boldsymbol{x}_{1:n} \in \mathbb{R}^d : \left|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right| \leq \frac{1}{2}\right\}. \tag{F.6c}$$

The proof of the k-NGCA information bound (Proposition F.1) is organized into subsections as follows.

1. To prove Proposition F.1, we rely on certain analytic properties of the likelihood ratio for this problem (similar to k-TPCA). These properties are stated (without proofs) in Appendix F.3.
2. Using these properties, Proposition F.1 is proved in Appendix F.4.
3. Finally, the proofs of the analytic properties of the likelihood ratio appear in Appendix F.5.

F.3. The Likelihood Ratio for Non-Gaussian Component Analysis. In this section, we collect some important properties of the likelihood ratio for the Non-Gaussian Component Analysis problem without proofs. The proofs of these properties are provided in Appendix F.5.

Recalling the data-generating process for k-NGCA problem (F.5), we can compute the likelihood ratio (with respect to the standard Gaussian distribution μ_0) of a single sample $\boldsymbol{x} \in \mathbb{R}^d$ from the model (F.5) as:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) = \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(\eta), \tag{F.7a}$$

where $\eta \overset{\text{def}}{=} \left\langle \boldsymbol{x}, \frac{1}{\sqrt{d}}\boldsymbol{V}\right\rangle$. Consequently the N-sample likelihood ratio is given by:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) = \prod_{i=1}^{N} \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(\eta_i),\ \eta_i = \frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}. \tag{F.7b}$$

Next we compute the Hermite decomposition of the N-sample likelihood ratio. Due to the product structure of the N-sample likelihood ratio, it is sufficient to compute the Hermite decomposition of the one-sample likelihood ratio. We have

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(\eta) = \sum_{t=0}^{\infty} \mathbb{E}_0\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)H_t(Z)\right] \cdot H_t(\eta) = \sum_{t=0}^{\infty} \hat{\nu}_t H_t(\eta).$$

In the last step, we defined,

$$\hat{\nu}_t \overset{\text{def}}{=} \mathbb{E}_0\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)H_t(Z)\right] = \mathbb{E}_{\eta\sim\nu}[H_t(\eta)].$$

Hence, we obtain the expression for the Hermite decomposition of the likelihood ratio summarized in the following lemma.

LEMMA F.1 (Hermite Decomposition for Non-Gaussian Component Analysis). *We have*

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) = \sum_{\boldsymbol{t}\in\mathbb{N}_0^N} \hat{\nu}_{\boldsymbol{t}} \cdot H_{\boldsymbol{t}}\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{V}\rangle}{\sqrt{d}}, \frac{\langle \boldsymbol{x}_2, \boldsymbol{V}\rangle}{\sqrt{d}}, \dots, \frac{\langle \boldsymbol{x}_N, \boldsymbol{V}\rangle}{\sqrt{d}}\right),$$

where, for any $\boldsymbol{t} \in \mathbb{N}_0^N$,

$$H_{\boldsymbol{t}}\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{V}\rangle}{\sqrt{d}}, \frac{\langle \boldsymbol{x}_2, \boldsymbol{V}\rangle}{\sqrt{d}}, \dots, \frac{\langle \boldsymbol{x}_N, \boldsymbol{V}\rangle}{\sqrt{d}}\right) \overset{\text{def}}{=} \prod_{i=1}^{N} H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right),$$

$$\hat{\nu}_{\boldsymbol{t}} \overset{\text{def}}{=} \prod_{i=1}^{N} \hat{\nu}_{t_i}.$$

In the above display, $\hat{\nu}_t$, $t \in \mathbb{N}$ are the Hermite coefficients of the one-sample likelihood ratio:

$$\hat{\nu}_t = \mathbb{E}_{Z\sim\mu_0}\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z) \cdot H_t(Z)\right] = \mathbb{E}_{\eta\sim\nu}\left[H_t(\eta)\right].$$

We now introduce an N-sample generalization of the integrated Hermite polynomials (Definition C.1).

DEFINITION F.1 (N-sample Integrated Hermite Polynomials). Let $S : \{\pm 1\}^d \to \mathbb{R}$ be a function with $\|S\|_\pi = 1$. For any $\boldsymbol{t} \in \mathbb{N}_0^N$, we define the N-sample integrated Hermite polynomials as:

$$\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \overset{\text{def}}{=} \int \left(\prod_{i=1}^{N} H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right)\right) \cdot S(\boldsymbol{V})\,\pi(\mathrm{d}\boldsymbol{V}).$$

The rationale for introducing this definition is analogous to that for their single-sample counterparts. We use Lemma F.1 to express important quantities derived from the likelihood ratio in terms of the N-sample integrated Hermite polynomials:

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) \overset{\text{def}}{=} \int \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N})\,\pi(\mathrm{d}\boldsymbol{V}) = \sum_{\boldsymbol{t}\in\mathbb{N}_0^N} \hat{\nu}_{\boldsymbol{t}} \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; 1),$$

$$\left\langle \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}), S\right\rangle_\pi \overset{\text{def}}{=} \int \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) \cdot S(\boldsymbol{V})\,\pi(\mathrm{d}\boldsymbol{V}) = \sum_{\boldsymbol{t}\in\mathbb{N}_0^N} \hat{\nu}_{\boldsymbol{t}} \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S).$$

Just like their single-sample counterparts, the N-sample integrated Hermite polynomials inherit the orthogonality property of the standard Hermite polynomials.

LEMMA F.2. *For any $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{N}_0^N$ such that $\boldsymbol{s} \neq \boldsymbol{t}$, we have*

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{s}}(\boldsymbol{x}_{1:N}; S) \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)] = 0.$$

PROOF. Using Definition F.1 and Fubini's theorem, we obtain,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{s}}(\boldsymbol{x}_{1:N}; S) \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)] =$$

$$\int\int \prod_{i=1}^{N} \mathbb{E}_0 \left[H_{s_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right) H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}'\rangle}{\sqrt{d}}\right)\right] \cdot S(\boldsymbol{V}) \cdot S(\boldsymbol{V}')\, \pi(\mathrm{d}\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V}').$$

Since $\boldsymbol{s} \neq \boldsymbol{t}$, there must be an $i \in \mathbb{N}$ such that $s_i \neq t_i$, and for this i, Fact I.6 gives us,

$$\mathbb{E}_0 \left[H_{s_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right) H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}'\rangle}{\sqrt{d}}\right)\right] = 0.$$

Hence, we obtain the claim of the lemma. □

We have the following N-sample generalization of Lemma C.3. Note that these worst-case bounds can be much smaller than 1.

LEMMA F.3. *There is a universal constant C (independent of d) such that, for any $\boldsymbol{t} \in \mathbb{N}_0^N$ with $\|\boldsymbol{t}\|_1 = t$, we have*

1. *When t is odd, $\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; 1) = 0$.*
2. *When t is even, $\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; 1)^2] \leq (Ct)^{\frac{t}{2}} \cdot d^{-\frac{t}{2}}$.*
3. *For even $t \leq d$, $\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; 1)^2] \geq (t/C)^{\frac{t}{2}} \cdot d^{-\frac{t}{2}}$.*
4. *For any $S : \{\pm 1\}^d \to \mathbb{R}$ with $\|S\|_\pi \leq 1$, we have $\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] \leq (Ct)^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil}$.*
5. *For any $S : \{\pm 1\}^d \to \mathbb{R}$ with $\|S\|_\pi \leq 1$, $\langle S, 1\rangle_\pi = 0$, we have $\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] \leq (Ct)^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}$.*

PROOF. See Appendix F.5.1. □

A limitation of the bound obtained in Lemma F.3 is that it is vacuous when $\|\boldsymbol{t}\|_1 \gg d$. The following lemma provides a bound on the norm of integrated Hermite polynomials with degree $\|\boldsymbol{t}\|_1 \gg d$.

LEMMA F.4. *For any $\boldsymbol{t} \in \mathbb{N}_0^N$ with $\|\boldsymbol{t}\|_1 = t$, we have*

$$\sup_{S : \|S\|_\pi \leq 1} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] \leq 2 \exp\left(-\frac{(1 - e^{-2t/d})}{2} \cdot d\right).$$

PROOF. See Appendix F.5.3. □

Using the orthogonality of the N-sample integrated Hermite polynomials (Lemma F.2) and the estimates obtained in Lemma F.3 and Lemma F.4, one can easily estimate the second moment of functions constructed by linear combinations of these polynomials:

$$\left\| \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right\|_2^2 \overset{\text{def}}{=} \mathbb{E}_0 \left(\sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right)^2 = \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}}^2 \cdot \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2].$$

We will also need to estimate q-norms of these linear combinations, for $q \geq 2$:

$$\left\| \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right\|_q^q \overset{\text{def}}{=} \mathbb{E}_0 \left| \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right|^q .$$

The following lemma, generalizing Lemma C.4, again uses Gaussian Hypercontractivity (Fact I.7) to provide an estimate for the above quantity.

LEMMA F.5. *Let $\{\alpha_{\boldsymbol{t}} : \boldsymbol{t} \in \mathbb{N}_0^N\}$ be an arbitrary collection of real-valued coefficients. For any $q \geq 2$, we have*

$$\left\| \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right\|_q^2 \leq \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} (q-1)^{\|\boldsymbol{t}\|_1} \cdot \alpha_{\boldsymbol{t}}^2 \cdot \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2]$$

Furthermore, the inequality holds as an equality when $q = 2$.

PROOF. See Appendix F.5.3. □

In the following section, we present a proof of the information bound for distributed Non-Gaussian Component Analysis (Proposition F.1) using the results of this section.

F.4. Proof of the Information Bound (Proposition F.1). This section provides a proof for Proposition F.1, the main information bound for the Non-Gaussian Component Analysis problem. Recall that the information bound of Proposition 1 is:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}{K} \leq$$

(F.8)

$$\sum_{i=1}^m \overline{\mathbb{E}}_0^{(i)} \left[\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right) \cdot \mathbb{I}_{Z_i=1} \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right] + m\overline{\mu}(\mathcal{Z}^c),$$

where $Z_i = \mathbb{I}_{\boldsymbol{X}_i \in \mathcal{Z}}$. The following lemma analyzes the failure probability $\overline{\mu}(\mathcal{Z}^c)$.

LEMMA F.6. *Suppose that ν satisfies the Moment Matching Assumption (Assumption 1) with constant $k \geq 2$ and the Bounded Signal Strength Assumption (Assumption 2) with constants (λ, K). Then, for any $q \geq 2$, there is exists a finite constant $C_{q,k,K}$ depending only on (q, k, K) such that, if,*

$$n\lambda^2 \leq \frac{d}{C_{q,k,K}},$$

then,

$$\overline{\mu}(\mathcal{Z}^c) \leq C_{q,k,K} \cdot \left((1+\lambda^2) \cdot n \cdot e^{-\frac{d}{C_{q,k,K}}} + \left(\frac{n\lambda^2}{d} \right)^{\frac{q}{2}} \right).$$

PROOF. The proof of this result appears at the end of this section (Appendix F.4.1). □

We also need to analyze:

$$\overline{\mathbb{E}}_0^{(i)} \left[\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) \right) \cdot \mathbb{I}_{Z_i=1} \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right],$$

For any $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$, $S \subset [n]$, we introduce the notation,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S) \overset{\text{def}}{=} \prod_{i\in S}\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1\right),$$

$$\overline{\mathscr{L}}(\boldsymbol{X}_S) \overset{\text{def}}{=} \int \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S)\,\pi(\mathrm{d}\boldsymbol{V}).$$

In the special case when $S = \{i\}$, we will use the simplified notation $\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_i)$, $\overline{\mathscr{L}}(\boldsymbol{x}_i)$. We consider the following decomposition: For any $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$,

$$\begin{aligned}\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}) &= \prod_{\ell=1}^{n}\left(1 + \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_\ell) - 1\right) - \int \prod_{\ell=1}^{n}\left(1 + \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_\ell) - 1\right)\,\pi(\mathrm{d}\boldsymbol{V}) \\ &= \underbrace{\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell))}_{\text{Additive Term}} + \underbrace{\sum_{S\subset[n],\,|S|\geq 2}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S) - \overline{\mathscr{L}}(\boldsymbol{X}_S))}_{\text{Non Additive Term}}.\end{aligned}$$

With this decomposition, using the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain,

(F.9)

$$\overline{\mathbb{E}}_0^{(i)}\left[\int\left(\overline{\mathbb{E}}_0^{(i)}\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right)\cdot \mathbb{I}_{Z_i=1}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V})\right] \leq 2\cdot(\mathsf{I} + \mathsf{II}),$$

where,

$$\mathsf{I} \overset{\text{def}}{=} \overline{\mathbb{E}}_0^{(i)}\left[\int\left(\overline{\mathbb{E}}_0^{(i)}\left[\left(\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell}))\right)\cdot \mathbb{I}_{Z_i=1}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V})\right],$$

$$\mathsf{II} \overset{\text{def}}{=} \overline{\mathbb{E}}_0^{(i)}\left[\int\left(\overline{\mathbb{E}}_0^{(i)}\left[\left|\sum_{S\subset[n],\,|S|\geq 2}(\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S) - \overline{\mathscr{L}}((\boldsymbol{X}_i)_S))\right|\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V})\right].$$

In order to control the term (II), we apply Jensen's Inequality:

$$\begin{aligned}\mathsf{II} &\leq \int \mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}(\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S) - \overline{\mathscr{L}}((\boldsymbol{X}_i)_S))\right|^2\right]\pi(\mathrm{d}\boldsymbol{V}) \\ &= \int \mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S)\right|^2\right]\pi(\mathrm{d}\boldsymbol{V}) - \mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\overline{\mathscr{L}}((\boldsymbol{X}_i)_S)\right|^2\right] \\ &\leq \int \mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S)\right|^2\right]\pi(\mathrm{d}\boldsymbol{V}).\end{aligned}$$

The following lemma analyzes the above upper bound on (II).

LEMMA F.7. *Suppose that ν satisfies the Bounded Signal Strength Assumption (Assumption 2) with constants (λ, K). Suppose that $K^2 n\lambda^2 \leq 1/2$. Let $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_n]$ where $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then,*

$$\mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)\right|^2\right] \leq 2\cdot(K^2 n\lambda^2)^2.$$

PROOF. The proof of this result appears at the end of this section (Appendix F.4.3). □

In order to control the term (I), we will rewrite it as follows:

$$\begin{aligned}\mathsf{I} &\overset{\text{def}}{=} \overline{\mathbb{E}}_0^{(i)}\left[\int \left(\overline{\mathbb{E}}_0^{(i)}\left[\left(\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell}))\right) \cdot \mathbb{I}_{Z_i=1}\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right] \\ &\overset{\text{(a)}}{=} \overline{\mathbb{E}}_0^{(i)}\left[\int \left(\overline{\mathbb{E}}_0^{(i)}\left[\left(\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}\right) \cdot \mathbb{I}_{Z_i=1}\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right] \\ &\overset{\text{(b)}}{=} \overline{\mathbb{E}}_0^{(i)}\left[\mathbb{I}_{Z_i=1} \cdot \int \left(\overline{\mathbb{E}}_0^{(i)}\left[\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right].\end{aligned}$$

In the step marked (a), we used the identity $\mathbb{I}_{Z_i=1} = \mathbb{I}_{\boldsymbol{X}_i\in\mathcal{Z}} = \mathbb{I}_{\boldsymbol{X}_i\in\mathcal{Z}} \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}$ (cf. (F.6)). In the step marked (b), we observed that $\mathbb{I}_{Z_i=1}$ is measurable with respect to the conditioning σ-algebra. Next, we linearize the integral with respect to the prior π (Lemma 1):

$$\begin{aligned}&\int \left(\overline{\mathbb{E}}_0^{(i)}\left[\sum_{\ell=1}^{n}(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}\middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V}) \\ &= \sup_{S:\|S\|_\pi\leq 1}\left(\overline{\mathbb{E}}_0^{(i)}\left[\sum_{\ell=1}^{n}\left\langle(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}, S\right\rangle_\pi \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2.\end{aligned}$$

We will apply the Geometric Inequality framework (Proposition 2) to control the above conditional expectation. In order to do so, we need to understand the concentration behavior of the random variable:

$$\sum_{\ell=1}^{n}\left\langle(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\|\leq\sqrt{2d}}, S\right\rangle_\pi.$$

This is the subject of the following lemma.

LEMMA F.8. *Suppose that ν satisfies:*

1. *the Moment Matching Assumption (Assumption 1) with parameter k,*
2. *the Bounded Signal Strength Assumption (Assumption 2) with parameters (λ, K),*
3. *the Locally Bounded Likelihood Ratio Assumption (Assumption 3) with parameters (λ, K, κ).*

Then, there is a constant $C_{k,\kappa}$ that depends only on (k,κ) such that if the parameters (λ, K, κ) satisfy $K\lambda \leq d^{-\frac{\kappa}{2}}/C_{k,\kappa}$, we have, for any $S : \{\pm 1\}^d \to \mathbb{R}$ with $\|S\|_\pi \leq 1$ and any $\zeta \in \mathbb{R}$ with $|\zeta| \leq 1/2L$,

$$\log \mathbb{E} \exp\left(\zeta \sum_{\ell=1}^{n}\left\langle(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\|\leq\sqrt{2d}}, S\right\rangle_\pi\right) \leq \zeta n \sigma e^{-\frac{d}{16}} + n\zeta^2\sigma^2.$$

In the above display the parameters L, σ^2 are defined as follows:

$$\begin{aligned}L &\overset{\text{def}}{=} C_{\kappa,k} \cdot K\lambda \cdot d^{\frac{\kappa}{2}}, \\ \sigma^2 &\overset{\text{def}}{=} C_{k,\kappa} \cdot K^2\lambda^2 \cdot d^{-\lceil\frac{k+1}{2}\rceil}.\end{aligned}$$

Furthermore,

$$\left\| \sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right\|_4 \le n \cdot \sigma \cdot e^{-\frac{d}{16}} + \sqrt{L\sigma} \cdot n^{\frac{1}{4}} + \sqrt{n\sigma},$$

where,

$$\left\| \sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right\|_4^4 \overset{\mathit{def}}{=} \mathbb{E}_0 \left[\left(\sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right)^4 \right]$$

PROOF. The proof of this result appears at the end of this section (Appendix F.4.2). □

We can now use Geometric Inequalities (Proposition 2) to control:

$$\left| \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1, (\boldsymbol{X}_j)_{j \neq i} \right] \right|^2.$$

We consider two cases depending upon whether $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}$ or $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}$, where,

$$\mathcal{R}^{(i)}_{\mathsf{rare}} \overset{\text{def}}{=} \left\{ \boldsymbol{y} \in \{0,1\}^{mb} : 0 < \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i}) \le 4^{-b} \right\},$$

$$\mathcal{R}^{(i)}_{\mathsf{freq}} \overset{\text{def}}{=} \left\{ \boldsymbol{y} \in \{0,1\}^{mb} : \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i}) > 4^{-b} \right\}.$$

Case 1: $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}$. In this situation we apply the moment version of the Geometric Inequality (Proposition 2, item (1)) with $q = 4$. Using the moment estimate in Lemma F.8, we obtain,

$$\left| \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1, (\boldsymbol{X}_j)_{j \neq i} \right] \right|$$

$$\le \frac{n \cdot \sigma \cdot e^{-\frac{d}{16}} + \sqrt{L\sigma} \cdot n^{\frac{1}{4}} + \sigma\sqrt{n}}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})^{\frac{1}{4}}}, \tag{F.10}$$

where L, σ are as defined in Lemma F.8.

Case 2: $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}$. In this situation we apply the m.g.f. version of the Geometric Inequality (Proposition 2, item (2)). Using the m.g.f. estimate in Lemma F.8, we obtain, for any $0 < \zeta \le 1/2L$,

$$\left| \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1, (\boldsymbol{X}_j)_{j \neq i} \right] \right|$$

$$\le n\sigma e^{-\frac{d}{16}} + n\zeta\sigma^2 + \frac{1}{\zeta} \log \frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})},$$

where L, σ are as defined in Lemma F.8. We set:

$$\zeta^2 = \frac{1}{n\sigma^2} \cdot \log \frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})} \le \frac{b \cdot \log(4)}{n\sigma^2}.$$

If,

$$n \geq \frac{4\log(4) \cdot b \cdot L^2}{\sigma^2}, \tag{F.11}$$

then this choice is valid, i.e. $\zeta \leq 1/2L$. With this choice, we obtain,

$$\left| \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \leq \sqrt{2d}}, S \right\rangle_{\pi} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1, (\boldsymbol{X}_j)_{j \neq i} \right] \right|$$

$$\leq n\sigma e^{-\frac{d}{16}} + 2 \cdot \sigma \cdot \sqrt{n} \cdot \log^{\frac{1}{2}} \left(\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})} \right). \tag{F.12}$$

With these estimates, we can control the term I, which we decompose as follows:

$$\overline{\mathbb{E}}_0^{(i)} \left[\mathbb{I}_{Z_i=1} \cdot \int \left(\overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \leq \sqrt{2d}} \middle| \boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right]$$

$$= (\mathsf{Ia}) + (\mathsf{Ib}),$$

$$(\mathsf{Ia}) \overset{\text{def}}{=} \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\boldsymbol{y} \in \mathcal{R}_{\mathsf{rare}}^{(i)}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i}) \cdot \Psi_i^2(\boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i}) \right],$$

$$(\mathsf{Ib}) \overset{\text{def}}{=} \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\boldsymbol{y} \in \mathcal{R}_{\mathsf{freq}}^{(i)}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i}) \cdot \Psi_i^2(\boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i}) \right].$$

In the above display, we defined,

$$\Psi_i^2(\boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i}) \overset{\text{def}}{=}$$

$$\int \left(\overline{\mathbb{E}}_0^{(i)} \left[\sum_{\ell=1}^{n} (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) - \overline{\mathscr{L}}(\boldsymbol{x}_{i\ell})) \cdot \mathbb{I}_{\|\boldsymbol{x}_{i\ell}\| \leq \sqrt{2d}} \middle| \boldsymbol{Y} = \boldsymbol{y}, Z_i = 1, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}).$$

In order to control (Ia), we rely on the estimate (F.10):

$$(\mathsf{Ia}) \leq (n \cdot \sigma \cdot e^{-\frac{d}{16}} + \sqrt{L\sigma} \cdot n^{\frac{1}{4}} + \sigma\sqrt{n})^2 \cdot \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\boldsymbol{y} \in \mathcal{R}_{\mathsf{rare}}^{(i)}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})^{\frac{1}{2}} \right]$$

$$\leq (n \cdot \sigma \cdot e^{-\frac{d}{16}} + \sqrt{L\sigma} \cdot n^{\frac{1}{4}} + \sigma\sqrt{n})^2 \cdot 2^{-b} \cdot \overline{\mathbb{E}}_0^{(i)}[|\mathcal{R}_{\mathsf{rare}}^{(i)}|].$$

Since we assume the communication protocol to be deterministic conditioned on $(\boldsymbol{X}_j)_{j \neq i}$, all but b bits of $\boldsymbol{Y}$ are fixed. Consequently, $|\mathcal{R}_{\mathsf{rare}}| \leq 2^b$. Hence,

$$(\mathsf{Ia}) \leq 3 \cdot \left(n^2 \cdot \sigma^2 \cdot e^{-\frac{d}{8}} + L\sigma\sqrt{n} + n\sigma^2 \right) \overset{(c)}{\leq} 3 \cdot \left(n^2 \cdot \sigma^2 \cdot e^{-\frac{d}{8}} + 2n\sigma^2 \right).$$

In the above display, in the step marked (c) we observed that the assumption (F.11) guarantees $L\sigma\sqrt{n} \leq n\sigma^2$. In order to control (Ib), we rely on the estimate (F.12):

$$(\mathsf{Ib}) \leq 2n^2\sigma^2 e^{-\frac{d}{8}} + 8\sigma^2 n \cdot \overline{\mathbb{E}}_0^{(i)} \left[\sum_{\boldsymbol{y} \in \mathcal{R}_{\mathsf{freq}}^{(i)}} h(\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y} = \boldsymbol{y}, Z_i = 1 | (\boldsymbol{X}_j)_{j \neq i})) \right]$$

$$\leq 2n^2\sigma^2 e^{-\frac{d}{8}} + 8\sigma^2 n \cdot \overline{\mathbb{E}}_0^{(i)}\left[\sum_{(\boldsymbol{y},z)\in\{0,1\}^{b+1}} h(\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z|(\boldsymbol{X}_j)_{j\neq i}))\right],$$

where $h(x) \overset{\text{def}}{=} -x\log(x)$ is the entropy function. Since we assume the communication protocol to be deterministic (cf. Remark 4), conditioned on $(\boldsymbol{X}_j)_{j\neq i}$, all but $b+1$ bits of $(\boldsymbol{Y}, Z_i)$ are fixed. Hence conditioned on $(\boldsymbol{X}_j)_{j\neq i}$, the random vector $(\boldsymbol{Y}, Z_i)$ has a support size of at most 2^{b+1}. The maximum entropy distribution on a given set S is the uniform distribution, which attains an entropy of $\log|S|$. Hence,

$$\sum_{(\boldsymbol{y},z)\in\{0,1\}^{b+1}} \overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z|(\boldsymbol{X}_j)_{j\neq i}) \log\frac{1}{\overline{\mathbb{P}}_0^{(i)}(\boldsymbol{Y}=\boldsymbol{y}, Z_i = z|(\boldsymbol{X}_j)_{j\neq i})} \leq (b+1)\log(2)$$

This yields the estimate,

$$(\mathsf{Ib}) \leq 2n^2\sigma^2 e^{-\frac{d}{8}} + 8\log(2)\cdot\sigma^2 n\cdot(b+1).$$

Combining the estimates on the terms Ia, Ib we obtain, $(\mathsf{I}) \leq 5n^2\sigma^2 e^{-\frac{d}{8}} + 18\sigma^2 n\cdot b$. Substituting this estimate on I and the estimate on II obtained in Lemma F.7 in (F.9), we obtain,

$$\overline{\mathbb{E}}_0^{(i)}\left[\int\left(\overline{\mathbb{E}}_0^{(i)}\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - \frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)\right)\cdot\mathbb{I}_{Z_i=1}\middle|\boldsymbol{Y}, Z_i, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right]$$
$$\leq 10n^2\sigma^2 e^{-\frac{d}{8}} + 36\sigma^2 n\cdot b + 4\cdot(K^2 n\lambda^2)^2.$$

Plugging the above bound in (F.8) we obtain,

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}{K} \leq 10mn^2\sigma^2 e^{-\frac{d}{8}} + 36\sigma^2\cdot m\cdot n\cdot b + 4\cdot m\cdot(K^2 n\lambda^2)^2 + m\cdot\overline{\mu}(\mathcal{Z}^c).$$

Finally, by Lemma F.6, for any $q \geq 2$, we have

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}{K} \leq 10mn^2\sigma^2 e^{-\frac{d}{8}} + 36\sigma^2\cdot m\cdot n\cdot b + 4\cdot m\cdot(K^2 n\lambda^2)^2$$
$$+ C_{q,k,K}\cdot m\cdot\left((1+\lambda^2)\cdot m\cdot n\cdot e^{-\frac{d}{C_{q,k,K}}} + \left(\frac{n\lambda^2}{d}\right)^{\frac{q}{2}}\right).$$

This is precisely the information bound claimed in Proposition F.1. This concludes the proof of Proposition F.1. The remainder of this section is devoted to the proof of the various intermediate results used in the above proof.

F.4.1. *Proof of Lemma F.6.*

PROOF OF LEMMA F.6. We begin by observing that by a union bound,

$$\overline{\mu}(\mathcal{Z}_1^c) \leq \sum_{i=1}^n \overline{\mu}(\{\|\boldsymbol{x}_i\| > \sqrt{2d}\})$$
$$= \sum_{i=1}^n \int \mu_{\boldsymbol{V}}(\{\|\boldsymbol{x}_i\| > \sqrt{2d}\})\,\pi(\mathrm{d}\boldsymbol{V}).$$

When $\boldsymbol{x} \sim \mu_0 = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, standard χ^2-concentration (see for e.g. [37, Example 2.11]) gives us:

$$\mu_0(\{\|\boldsymbol{x}\| > \sqrt{2d}\}) \leq e^{-d/8}.$$

$$\begin{aligned}\mu_{\boldsymbol{V}}(\{\|\boldsymbol{x}\| > \sqrt{2d}\}) &= \mathbb{E}_0\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x})\mathbb{I}_{\|\boldsymbol{x}\|>\sqrt{2d}}\right] \\ &\overset{\text{(a)}}{=} \mathbb{E}_0\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)\,\mathbb{I}_{\|\boldsymbol{x}\|>\sqrt{2d}}\right],\ Z \overset{\text{def}}{=} \left\langle \boldsymbol{x}, \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}\right\rangle \\ &\overset{\text{(b)}}{\leq} \left(\mu_0\left(\{\|\boldsymbol{x}\| > \sqrt{2d}\}\right)\cdot \mathbb{E}_0\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)\right)^2\right]\right)^{\frac{1}{2}} \\ &\overset{\text{(c)}}{\leq} (1+K^2\lambda^2)\cdot e^{-\frac{d}{16}}.\end{aligned}$$

In the above display, in the step marked (a), we used the formula for the likelihood ratio in (F.7), in step (b) we used Cauchy-Schwarz inequality and in step (c) we appealed to Bounded Signal Strength Assumption (Assumption 2). Hence, we conclude that,

$$\overline{\mu}(\mathcal{Z}_1^c) \leq (1+K^2\lambda^2)\cdot n \cdot e^{-\frac{d}{16}}.$$

In order to analyze $\overline{\mu}(\mathcal{Z}_2^c)$, we recall that,

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 = \sum_{\substack{\boldsymbol{t}\in\mathbb{N}_0^n \\ \|\boldsymbol{t}\|_1\geq 1}} \hat{\nu}_{\boldsymbol{t}}\cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:n};1).$$

We decompose the centered likelihood ratio into the low degree part and the high degree part:

$$\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 = \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t} + \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t},$$

where,

$$\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t} \overset{\text{def}}{=} \sum_{\substack{\boldsymbol{t}\in\mathbb{N}_0^n \\ 1\leq\|\boldsymbol{t}\|_1\leq t}} \hat{\nu}_{\boldsymbol{t}}\cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:n};1),$$

$$\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t} \overset{\text{def}}{=} \sum_{\substack{\boldsymbol{t}\in\mathbb{N}_0^n \\ \|\boldsymbol{t}\|_1> t}} \hat{\nu}_{\boldsymbol{t}}\cdot \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:n};1).$$

With this decomposition, for any $q \geq 1$, we have, by Markov's Inequality,

$$\begin{aligned}&\overline{\mu}(\mathcal{Z}_2^c) \leq \overline{\mu}\left(\left|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t}\right| > \frac{1}{4}\right) + \overline{\mu}\left(\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t} > \frac{1}{4}\right) \\ &\leq 4^q\overline{\mathbb{E}}\left[\left|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t}\right|^q\right] + 4\overline{\mathbb{E}}\left[\left|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t}\right|\right] \\ &= 4^q\mathbb{E}_0\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\left|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t}\right|^q\right] + 4\mathbb{E}_0\left[\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\left|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t}\right|\right] \\ &\leq 4^q\left\|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\right\|_2\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{\leq t}\right\|_{2q}^q + 4\left\|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\right\|_2\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1\right)_{>t}\right\|_2.\end{aligned}$$

In order to obtain the last inequality, we applied Cauchy-Schwarz inequality. We also note that all the norms $\|\cdot\|_q$ are defined with respect to μ_0. We now estimate each of the norms in the above display. The quantity:

$$\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) - 1 \right)_{\leq t} \right\|_q^2,$$

with the choice $q = 2$ is a central object in the low-degree likelihood ratio framework. In the arXiv version of this paper [16, Appendix F.4, Proposition 8] we analyze the Non-Gaussian Component Analysis problem in the low-degree likelihood ratio framework and show that there is a constant $C_{k,K} > 0$ depending only on k, K such that, for any $q \geq 2$, if,

$$t \leq \frac{1}{C_{k,K}} \cdot \frac{d}{(q-1)^2}, \; N\lambda^2 \leq \frac{1}{C_{k,K}} \cdot \frac{1}{(q-1)^k} \cdot \frac{d^{\frac{k}{2}}}{t^{\frac{k-2}{2}}}, \tag{F.13}$$

then,

$$\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:N}) - 1 \right)_{\leq t} \right\|_q^2 \leq \frac{C_{k,K} \cdot (q-1)^k \cdot N\lambda^2 \cdot t^{\frac{k-2}{2}}}{d^{\frac{k}{2}}} \leq 1.$$

We set,

$$t = \frac{1}{C_{k,K}} \cdot \frac{d}{(2q-1)^2}.$$

The hypothesis on the effective sample size $n\lambda^2 \leq d/C_{q,k,K}$ ensures the requirement (F.13) is met, and we obtain,

$$\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 \right)_{\leq t} \right\|_{2q}^2 \leq \frac{C_{k,K} \cdot (2q-1)^2 \cdot n\lambda^2}{d},$$

$$\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 \right)_{\leq t} \right\|_2^2 \leq \frac{C_{k,K} \cdot n\lambda^2}{d} \leq 1.$$

On the other hand, by Lemma F.5, we have

$$\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 \right)_{> t} \right\|_2^2 = \sum_{\substack{\boldsymbol{t} \in \mathbb{N}_0^n \\ \|\boldsymbol{t}\|_1 > t}} \hat{\nu}_{\boldsymbol{t}}^2 \cdot \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:n}; 1)^2]$$

In Lemma F.4, we showed that, for any $\|\boldsymbol{t}\|_1 > t$,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:n}; 1)^2] \leq 2 \exp\left(-\frac{d}{C_{q,k,K}} \right), \; C_{q,k,K} = 2\left(1 - e^{-\frac{1}{C_{k,K}(q-1)^2}} \right)^{-1}.$$

Hence,

$$\begin{aligned}
\left\| \left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n}) - 1 \right)_{> t} \right\|_2^2 &\leq 2 \cdot e^{-\frac{d}{C_{q,k,K}}} \cdot \sum_{\substack{\boldsymbol{t} \in \mathbb{N}_0^n \\ \|\boldsymbol{t}\|_1 > t}} \hat{\nu}_{\boldsymbol{t}}^2 \\
&\leq 2 \cdot e^{-\frac{d}{C_{q,k,K}}} \cdot \sum_{\boldsymbol{t} \in \mathbb{N}_0^n} \hat{\nu}_{\boldsymbol{t}}^2 \\
&\overset{\text{(e)}}{\leq} 2 \cdot e^{-\frac{d}{C_{q,k,K}}} \cdot (1 + K^2\lambda^2)^n \\
&\leq 2 \cdot e^{-\frac{d}{C_{q,k,K}} + K^2\lambda^2 n} \\
&\overset{\text{(f)}}{\leq} 2 \cdot e^{-\frac{d}{2C_{q,k,K}}}.
\end{aligned}$$

In the above display, the step (e) relies on the Bounded Signal Strength Assumption and in the step marked (f) we used the effective sample size assumption $n\lambda^2 \leq d/(2C_{q,k,K})$. Due to the orthogonality of integrated Hermite polynomials (Lemma F.2), one can compute:

$$\begin{aligned}\left\|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})-1\right\|_2^2 &= 1+\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})-1\right)_{\leq t}\right\|_2^2+\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})-1\right)_{>t}\right\|_2^2 \\ &\leq 1+1+2=4.\end{aligned}$$

Hence,

$$\begin{aligned}\overline{\mu}(\mathscr{Z}_2^c) &\leq 4^q\left\|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\right\|_2\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})-1\right)_{\leq t}\right\|_{2q}^q \\ &\qquad + 4\left\|\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})\right\|_2\left\|\left(\frac{\mathrm{d}\overline{\mu}}{\mathrm{d}\mu_0}(\boldsymbol{x}_{1:n})-1\right)_{>t}\right\|_2 \\ &\leq \left(\frac{C_{q,k,K}\cdot n\lambda^2}{d}\right)^{\frac{q}{2}}+16e^{-d/C_{q,k,K}}.\end{aligned}$$

Finally, by suitably redefining constants, we obtain, by a union bound,

$$\overline{\mu}(\mathscr{Z}^c) \leq C_{q,k,K}\cdot\left((1+\lambda^2)\cdot n\cdot e^{-\frac{d}{C_{q,k,K}}}+\left(\frac{n\lambda^2}{d}\right)^{\frac{q}{2}}\right),$$

as claimed. □

F.4.2. *Proof of Lemma F.8.* Let $\boldsymbol{x}_{1:N}$ be generated as: $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$. Let $S:\{\pm 1\}^d \to \mathbb{R}$ be a real-valued function defined on the Boolean hypercube with $\|S\|_\pi \leq 1$. In this section, we wish to understand the concentration behavior of the random variable:

$$\sum_{\ell=1}^n\left\langle(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell)-\overline{\mathscr{L}}(\boldsymbol{x}_\ell))\cdot\mathbb{I}_{\|\boldsymbol{x}_\ell\|\leq\sqrt{2d}},S\right\rangle_\pi. \tag{F.14}$$

Since this is a sum of i.i.d. random variables, we will find the Bernstein Inequality useful in our analysis, and we reproduce the statement of this inequality below for convenience. This result is attributed to Bernstein. The statement below has been reproduced from Wainwright [37, Proposition 2.10]

FACT F.1 (Bernstein's Inequality). Let $U_1,U_2\cdots,U_n$ be i.i.d. random variables which satisfy:

1. $\mathbb{E}U_i \leq u$
2. $\mathsf{Var}(U_i)\leq\sigma^2$
3. $|U_i|\leq L$ with probability 1.

Then, for any $|\zeta|\leq 1/L$,

$$\log\mathbb{E}\exp\left(\zeta\sum_{i=1}^n U_i\right)\leq\zeta nu+\frac{n\zeta^2\sigma^2}{2(1-L|\zeta|)}.$$

We can now provide the proof of Lemma F.8.

PROOF OF LEMMA F.8. The proof involves an application of Bernstein Inequality (Fact F.1) after the computation of relevant quantities, which we compute in the following paragraphs. Let $\boldsymbol{x} \sim \mu_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. We recall that,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) \overset{\text{def}}{=} \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1,\ \overline{\mathscr{L}}(\boldsymbol{x}) \overset{\text{def}}{=} \int \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x})\, \pi(\mathrm{d}\boldsymbol{V}).$$

Worst-case upper bound: We begin by computing:

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^d} &\left| \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right| \\ &\le \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|S\|_{\pi} \cdot \|(\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\|_{\pi} \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}} \\ &\le 2 \cdot \sup_{\boldsymbol{V}\in\{\pm 1\}^d} \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|^2 \le 2d} \left| \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right|.\end{aligned}$$

Since,

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 = \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}\left(\frac{\langle \boldsymbol{x}, \boldsymbol{V}\rangle}{\sqrt{d}}\right) - 1,$$

we obtain,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right| \le 2 \cdot \sup_{z:|z|\le\sqrt{2d}} \left| \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) - 1 \right|.$$

Recall that ν satisfies the Locally Bounded Likelihood Ratio Assumption (Assumption 3) with parameters (λ, K, κ). Furthermore if,

$$3^{\kappa} \cdot K \cdot \lambda \cdot d^{\frac{\kappa}{2}} \le 1, \tag{F.15}$$

Assumption 3 yields,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi} \right| &\le 2 \cdot 3^{\kappa} \cdot K \cdot \lambda \cdot d^{\frac{\kappa}{2}} \\ &\overset{\text{def}}{=} L/2.\end{aligned} \tag{F.16}$$

Upper Bound on Variance: Observe that:

$$\begin{aligned}\mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi}\right) &\le \mathbb{E}_0 \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi}^2 \\ &\le \mathbb{E}_0 \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S \right\rangle_{\pi}^2\end{aligned}$$

Observe that,

$$\int \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})\, \pi(\mathrm{d}\boldsymbol{V}) = 0.$$

Hence,

$$\begin{aligned}\sup_{\substack{S:\{\pm1\}\to\mathbb{R}\\ \|S\|_{\pi}\le 1}} &\mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_{\pi}\right) \\ &\le \sup_{\substack{S:\{\pm1\}\to\mathbb{R}\\ \|S\|_{\pi}\le 1}} \mathbb{E}_0 \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S \right\rangle_{\pi}^2\end{aligned}$$

$$= \sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi \le 1,\, \langle S,1\rangle_\pi = 0}} \mathbb{E}_0 \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S \right\rangle_\pi^2$$

$$= \sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi \le 1,\, \langle S,1\rangle_\pi = 0}} \mathbb{E}_0 \left\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S \right\rangle_\pi^2 .$$

Next we recall the Hermite decomposition of the $\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x})$ computed in Lemma F.1:

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) = \sum_{t=1}^{\infty} \hat{\nu}_t \cdot H_t\left(\frac{\langle \boldsymbol{x}, \boldsymbol{V}\rangle}{\sqrt{d}}\right).$$

Recalling the definition of Integrated Hermite Polynomials (Definition F.1), we can write:

$$\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S\rangle_\pi = \sum_{t=1}^{\infty} \hat{\nu}_t \cdot \overline{H}_t(\boldsymbol{x}; S).$$

Since the integrated Hermite polynomials are orthogonal,

$$\mathbb{E}_0 \left\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S \right\rangle_\pi^2 = \sum_{t=1}^{\infty} \hat{\nu}_t^2 \cdot \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2].$$

Since ν satisfies the Moment Matching Assumption (Assumption 1) with parameter k, we have $\hat{\nu}_t = 0$ for any $t \le k-1$. Hence,

$$\begin{aligned}
\sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi \le 1}} \mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_\pi\right) &\le \sum_{t=k}^{\infty} \hat{\nu}_t^2 \cdot \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2] \\
&\le \left(\sup_{t \ge k} \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2]\right) \cdot \sum_{t=k}^{\infty} \hat{\nu}_t^2 \\
&\overset{\text{(a)}}{\le} K^2 \lambda^2 \cdot \left(\sup_{t \ge k} \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2]\right).
\end{aligned}$$

In the step marked (a), we appealed to the Bounded Signal Strength Assumption (Assumption 2). In Lemma F.3, we showed that, $\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] \le (Ct)^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}$. Hence,

$$\max_{k \le t \le d/C} \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2] \le (Ck)^{\frac{k}{2}} \cdot d^{-\lceil \frac{k+1}{2} \rceil}.$$

On the other hand, when $t \ge d/C$, Lemma F.4 gives us,

$$\max_{t \ge d/C} \mathbb{E}_0[\overline{H}_t(\boldsymbol{x}; S)^2] \le 2e^{-d/C'},\ C' = \frac{2}{(1 - e^{-2/C})}.$$

By suitably defining the constant C_k (depending on k), we arrive at the following estimate of the variance:

$$\begin{aligned}
&\text{(F.17a)} \\
&\sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi \le 1}} \mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \le \sqrt{2d}}, S \right\rangle_\pi\right) \le \sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi \le 1,\, \langle S,1\rangle_\pi = 0}} \mathbb{E}_0 \left\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S \right\rangle_\pi^2 \\
&\text{(F.17b)} \qquad \le C_k \cdot K^2 \lambda^2 \cdot d^{-\lceil \frac{k+1}{2} \rceil} \\
&\text{(F.17c)} \qquad \overset{\text{def}}{=} \sigma^2.
\end{aligned}$$

Upper Bound on Expectation: As before (by centering S) we can argue,

$$\sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi\le 1}} \mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}, S\right\rangle_\pi\right]$$

$$= \sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi\le 1\\ \langle S,1\rangle_\pi=0}} \mathbb{E}_0\left[\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S\rangle_\pi\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}\right].$$

Next we observe that since $\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x})] = \mathbb{E}_0[\overline{\mathscr{L}}(\boldsymbol{x})] = 0$, we have

$$\mathbb{E}_0[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S\right\rangle_\pi] = 0.$$

Hence, we can write:

$$\mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S\right\rangle_\pi\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}\right] = -\mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S\right\rangle_\pi\cdot \mathbb{I}_{\|\boldsymbol{x}\|>\sqrt{2d}}\right].$$

Consequently, by Cauchy-Schwarz Inequality,

$$\left(\mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})), S\right\rangle_\pi\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}\right]\right)^2 \le \mu_0(\|\boldsymbol{x}\|^2 \ge 2d)\cdot \mathbb{E}_0[\langle \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}), S\rangle_\pi^2].$$

Standard χ^2-concentration (see for e.g. [37, Example 2.11]) gives us $\mu_0(\{\|\boldsymbol{x}\| > \sqrt{2d}\}) \le e^{-d/8}$. Combining this with the estimate in (F.17), we obtain,

$$.\sup_{\substack{S:\{\pm 1\}\to\mathbb{R}\\ \|S\|_\pi\le 1}} \mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}, S\right\rangle_\pi\right] \le \sigma e^{-\frac{d}{16}}. \tag{F.18}$$

Combining the estimates obtained in (F.16), (F.17) and (F.18) with the Bernstein Inequality immediately gives the claim of the first claim of the lemma. In order to obtain the second claim, we first define:

$$\alpha(S) \overset{\text{def}}{=} \mathbb{E}_0\left[\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}, S\right\rangle_\pi\right].$$

We have

$$\left\|\sum_{\ell=1}^n \left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell))\cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\|\le\sqrt{2d}}, S\right\rangle_\pi\right\|_4 \le$$

$$n\alpha(S) + \left\|\sum_{\ell=1}^n \left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell))\cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\|\le\sqrt{2d}}, S\right\rangle_\pi - \alpha(S)\right)\right\|_4.$$

We can compute:

$$\left\|\sum_{\ell=1}^n \left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell))\cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\|\le\sqrt{2d}}, S\right\rangle_\pi - \alpha(S)\right)\right\|_4^4$$

$$= n\cdot\mathbb{E}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell))\cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\|\le\sqrt{2d}}, S\right\rangle_\pi - \alpha(S)\right)^4$$

$$+ 3n(n-1)\cdot \mathsf{Var}^2\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}, S\right\rangle_\pi\right).$$

We recall that,

$$\alpha(S) \le \sigma e^{-\frac{d}{16}},$$

$$\mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x}))\cdot \mathbb{I}_{\|\boldsymbol{x}\|\le\sqrt{2d}}, S\right\rangle_\pi\right) \le \sigma^2,$$

and that,

$$\mathbb{E}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\| \leq \sqrt{2d}}, S\right\rangle_\pi - \alpha(S)\right)^4 \\ \leq L^2 \cdot \mathsf{Var}\left(\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}) - \overline{\mathscr{L}}(\boldsymbol{x})) \cdot \mathbb{I}_{\|\boldsymbol{x}\| \leq \sqrt{2d}}, S\right\rangle_\pi\right) \leq L^2\sigma^2.$$

Hence,

$$\left\|\sum_{\ell=1}^{n}\left\langle (\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell) - \overline{\mathscr{L}}(\boldsymbol{x}_\ell)) \cdot \mathbb{I}_{\|\boldsymbol{x}_\ell\| \leq \sqrt{2d}}, S\right\rangle_\pi\right\|_4 \leq n \cdot \sigma \cdot e^{-\frac{d}{16}} + \sqrt{L\sigma} \cdot n^{\frac{1}{4}} + \sigma\sqrt{n}.$$

This concludes the proof of this lemma. □

F.4.3. *Proof of Lemma F.7.*

PROOF OF LEMMA F.7. We have

$$\mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)\right|^2\right] = \sum_{\substack{S_1,S_2\subset[n]\\|S_1|\geq 2,|S_2|\geq 2}}\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_1})\cdot\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_2})]$$

Recall that,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S) \stackrel{\text{def}}{=} \prod_{i\in S}\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1\right).$$

We observe that, $\boldsymbol{x}_{1:n}$ are independent and,

$$\mathbb{E}_0\left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1\right] = 0.$$

Hence if $S_1 \neq S_2$, $\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_1})\cdot\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_2})] = 0$. This gives us:

$$\mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)\right|^2\right] = \sum_{S\subset[n]\ |S|\geq 2}\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)^2].$$

Recall the formula for the likelihood ratio for the Non-Gaussian Component Analysis problem (F.7), we obtain,

$$\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)^2] = \left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z) - 1\right)^2\right]\right)^{|S|},\ Z \sim \mathcal{N}(0,1).$$

Since ν satisfies the Bounded Signal Strength Assumption, we have

$$\begin{aligned}\mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)\right|^2\right] &\leq \sum_{S\subset[n]\ |S|\geq 2}(K^2\lambda^2)^{|S|}\\ &= \sum_{s=2}^{n}\binom{n}{s}(K^2\lambda^2)^s\\ &\leq \sum_{s=2}^{n}(K^2n\lambda^2)^s.\end{aligned}$$

The assumption $K^2n\lambda^2 \leq 1/2$ guarantees that the above sum is dominated by a Geometric series, which immediately yields the claim of the lemma. □

F.5. Omitted Proofs from Section F.3. This section contains the proofs of the various analytic properties (Lemma F.3, Lemma F.4 and Lemma F.5) of the likelihood ratio for the Non-Gaussian Component Analysis problem, which were stated in Appendix F.3.

F.5.1. *Proof of Lemma F.3.*

PROOF OF LEMMA F.3. Using Definition F.1 and Fubini's theorem, we obtain,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] =$$

$$\int\int\prod_{i=1}^{N}\mathbb{E}_0\left[H_{t_i}\left(\frac{\langle \boldsymbol{x}_i,\boldsymbol{V}_1\rangle}{\sqrt{d}}\right)H_{t_i}\left(\frac{\langle \boldsymbol{x}_i,\boldsymbol{V}_2\rangle}{\sqrt{d}}\right)\right]\cdot S(\boldsymbol{V}_1)\cdot S(\boldsymbol{V}_2)\,\pi(\mathrm{d}\boldsymbol{V}_1)\,\pi(\mathrm{d}\boldsymbol{V}_2).$$

Fact I.6 gives us,

$$\mathbb{E}_0\left[H_{t_i}\left(\frac{\langle \boldsymbol{x}_i,\boldsymbol{V}\rangle}{\sqrt{d}}\right)H_{t_i}\left(\frac{\langle \boldsymbol{x}_i,\boldsymbol{V}'\rangle}{\sqrt{d}}\right)\right] = \left(\frac{\langle \boldsymbol{V}_1,\boldsymbol{V}_2\rangle}{d}\right)^t.$$

We define $\boldsymbol{V} = \boldsymbol{V}_1 \odot \boldsymbol{V}_2$, where $\odot$ denotes entry-wise product of vectors and,

$$\overline{V} = \frac{1}{d}\sum_{i=1}^{d} V_i.$$

Hence,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \int\int \overline{V}^t\cdot S(\boldsymbol{V}_1)\cdot S(\boldsymbol{V}_2)\,\pi(\mathrm{d}\boldsymbol{V}_1)\,\pi(\mathrm{d}\boldsymbol{V}_2).$$

Since $\boldsymbol{V}_1,\boldsymbol{V}_2$ are independently sampled from the prior π and $\boldsymbol{V} = \boldsymbol{V}_1 \odot \boldsymbol{V}_2$, it is straightforward to check that $\boldsymbol{V}_1,\boldsymbol{V}$ are independent, uniformly random $\{\pm 1\}^d$ vectors and $\boldsymbol{V}_2 = \boldsymbol{V}_1 \odot \boldsymbol{V}$. Hence,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \int\int \overline{V}^t\cdot S(\boldsymbol{V}_1)\cdot S(\boldsymbol{V}\odot\boldsymbol{V}_1)\,\pi(\mathrm{d}\boldsymbol{V}_1)\,\pi(\mathrm{d}\boldsymbol{V}). \tag{F.19}$$

Recall that, the collection of polynomials:

$$\left\{\boldsymbol{V}^{\boldsymbol{r}} \overset{\text{def}}{=} \prod_{i=1}^{d} V_i^{r_i} : \boldsymbol{r}\in\{0,1\}^d\right\},$$

form an orthonormal basis for functions on the Boolean hypercube $\{\pm 1\}^d$ with respect to the uniform distribution $\pi = \mathsf{Unif}\left(\{\pm 1\}^d\right)$. Hence, we can expand $\boldsymbol{S}$ in this basis:

$$\boldsymbol{S}(\boldsymbol{V}) = \sum_{\boldsymbol{r}\in\{0,1\}^d}\hat{S}_{\boldsymbol{r}}\cdot\boldsymbol{V}^{\boldsymbol{r}},\ \hat{S}_{\boldsymbol{r}} \overset{\text{def}}{=} \int S(\boldsymbol{V})\cdot\boldsymbol{V}^{\boldsymbol{r}}\pi(\mathrm{d}\boldsymbol{V}).$$

Substituting this in (F.19) gives us:

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \sum_{\boldsymbol{r},\boldsymbol{s}\in\{0,1\}^d}\hat{S}_{\boldsymbol{r}}\hat{S}_{\boldsymbol{s}}\int\int\overline{V}^t\cdot\boldsymbol{V}_1^{\boldsymbol{r}+\boldsymbol{s}}\cdot\boldsymbol{V}^{\boldsymbol{s}}\,\pi(\mathrm{d}\boldsymbol{V}_1)\,\pi(\mathrm{d}\boldsymbol{V}).$$

Noting that, if $\boldsymbol{r}\neq\boldsymbol{s}$,

$$\int\boldsymbol{V}_1^{\boldsymbol{r}+\boldsymbol{s}}\,\pi(\mathrm{d}\boldsymbol{V}_1) = 0,$$

we obtain,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \sum_{\boldsymbol{r}\in\{0,1\}^d} \hat{S}_{\boldsymbol{r}}^2 \int \overline{V}^t \cdot \boldsymbol{V}^{\boldsymbol{r}} \, \pi(\mathrm{d}\boldsymbol{V}).$$

With this formula, we can prove each claim of the lemma.

When $S=1$: When $S=1$, $\hat{S}_{\boldsymbol{0}}=1$ and $\hat{\boldsymbol{S}}_{\boldsymbol{r}}=0$ for any $\boldsymbol{r}\neq 0$. Hence,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \int \overline{V}^t \, \pi(\mathrm{d}\boldsymbol{V}).$$

When t is odd, this is indeed zero due to symmetry. This proves item (1). For even t, we recall that $\sqrt{d}\overline{V}$ is sub-Gaussian with variance proxy 1, and standard moment bounds on sub-Gaussian random variables (see e.g. [33, Lemma 1.4])

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \int \overline{V}^t \, \pi(\mathrm{d}\boldsymbol{V}) \leq (4t)^{\frac{t}{2}} \cdot d^{-\frac{i}{2}}.$$

This proves item (2). The lower bound in item (3) is obtained by appealing to Fact I.3 (due to Kunisky, Wein and Bandeira [27]):

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] = \int \overline{V}^t \, \pi(\mathrm{d}\boldsymbol{V}) \geq (t/e^2)^{\frac{t}{2}} \cdot d^{-\frac{t}{2}}.$$

General S with $\|S\|_\pi \leq 1$:

Since $\|S\|_\pi \leq 1$, we know that $\sum_{\boldsymbol{r}} \hat{S}_{\boldsymbol{r}}^2 \leq 1$. When $\langle S, 1\rangle_\pi = 0$, one additionally has $\hat{S}_{\boldsymbol{0}} = 0$. Hence,

$$\sup_{S:\|S\|_\pi\leq 1} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] \leq \max_{\boldsymbol{r}\in\{0,1\}^d} \int \overline{V}^t \cdot \boldsymbol{V}^{\boldsymbol{r}} \, \pi(\mathrm{d}\boldsymbol{V}),$$

$$\sup_{\substack{S:\|S\|_\pi\leq 1\\ \langle S,1\rangle_\pi = 0}} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] \leq \max_{\substack{\boldsymbol{r}\in\{0,1\}^d\\ \|\boldsymbol{r}\|_1\geq 1}} \int \overline{V}^t \cdot \boldsymbol{V}^{\boldsymbol{r}} \, \pi(\mathrm{d}\boldsymbol{V}).$$

The right hand sides of the above equations have been analyzed in Lemma I.1. Appealing to this result immediately yield claim (5) and (6). □

F.5.2. *Proof of Lemma F.4.*

PROOF OF LEMMA F.4. Let $\|\boldsymbol{t}\|_1 = t$. Recall that in the proof of Lemma F.3, we showed:

$$\sup_{S:\|S\|_\pi\leq 1} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] \leq \max_{\boldsymbol{r}\in\{0,1\}^d} \int \overline{V}^t \cdot \boldsymbol{V}^{\boldsymbol{r}} \, \pi(\mathrm{d}\boldsymbol{V}).$$

Hence, using the triangle inequality and the fact that $|\boldsymbol{V}^{\boldsymbol{r}}| \leq 1$ we obtain,

$$\sup_{S:\|S\|_\pi\leq 1} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N};S)^2] \leq \int |\overline{V}|^t \, \pi(\mathrm{d}\boldsymbol{V}).$$

Let $D_+(\boldsymbol{V})$ denote the number of positive coordinates of $\boldsymbol{V}$. Let $D_-(\boldsymbol{V})$ denote the number of negative coordinates of $\boldsymbol{V}$. We observe that,

$$|\overline{V}| = 1 - \frac{2}{d} \cdot D_+(\boldsymbol{V}) \wedge D_-(\boldsymbol{V}).$$

Hence,

$$
\begin{aligned}
|\overline{V}|^t &= \left(1 - \frac{2}{d} \cdot D_+(\boldsymbol{V}) \wedge D_-(\boldsymbol{V})\right)^t \\
&\leq \exp\left(-\frac{2t}{d} \cdot D_+(\boldsymbol{V}) \wedge D_-(\boldsymbol{V})\right) \\
&\leq \exp\left(-\frac{2t}{d} \cdot D_+(\boldsymbol{V})\right) + \exp\left(-\frac{2t}{d} \cdot D_-(\boldsymbol{V})\right).
\end{aligned}
$$

Observing that,

$$
\begin{aligned}
\int \exp\left(-\frac{2t}{d} \cdot D_-(\boldsymbol{V})\right) \pi(\mathrm{d}\boldsymbol{V}) = \int \exp\left(-\frac{2t}{d} \cdot D_+(\boldsymbol{V})\right) \pi(\mathrm{d}\boldsymbol{V}) &= \left(\frac{1 + e^{-2t/d}}{2}\right)^d \\
&= \left(1 - \frac{(1 - e^{-2t/d})}{2}\right)^d \\
&\leq \exp\left(-\frac{(1 - e^{-2t/d})}{2} \cdot d\right).
\end{aligned}
$$

Hence,

$$
\sup_{S : \|S\|_\pi \leq 1} \mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] \leq 2 \exp\left(-\frac{(1 - e^{-2t/d})}{2} \cdot d\right),
$$

as claimed. □

F.5.3. *Proof of Lemma F.5.*

PROOF OF LEMMA F.5. Note that the result for $q = 2$ follows from the discussion preceding this lemma. Hence we focus on proving the inequality when $q \geq 2$. Recalling Definition F.1, we have

$$
\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \overset{\text{def}}{=} \int \left(\prod_{i=1}^{N} H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right)\right) \cdot S(\boldsymbol{V})\, \pi(\mathrm{d}\boldsymbol{V})
$$

Observe that for any fixed $\boldsymbol{V}$ and any $i \in [N]$,

$$
H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right),
$$

can be written as a homogeneous polynomial in $\boldsymbol{x}_i$ (see Fact I.5) with degree t_i. Since,

$$
\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) = \int \left(\prod_{i=1}^{N} H_{t_i}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right)\right) \cdot S(\boldsymbol{V})
$$

is a weighted linear combination of such polynomials, it must have a representation of the form:

$$
\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) = \sum_{\substack{\boldsymbol{c}_{1:N} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}_i\|_1 = t_i}} \beta(\boldsymbol{c}_{1:N}; S) \cdot H_{\boldsymbol{c}_1}(\boldsymbol{x}_1) \cdot H_{\boldsymbol{c}_2}(\boldsymbol{x}_2) \cdots H_{\boldsymbol{c}_N}(\boldsymbol{x}_N),
$$

for some coefficients $\beta(\boldsymbol{c}_{1:N}; S)$. While these coefficients can be computed, we will not need their exact formula for our discussion. Hence,

$$\sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) = \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \sum_{\substack{\boldsymbol{c}_{1:N} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}_i\|_1 = t_i}} \alpha_{\boldsymbol{t}} \cdot \beta(\boldsymbol{c}_{1:N}; S) \cdot H_{\boldsymbol{c}_1}(\boldsymbol{x}_1) \cdot H_{\boldsymbol{c}_2}(\boldsymbol{x}_2) \cdots \cdot H_{\boldsymbol{c}_N}(\boldsymbol{x}_N)$$

By Gaussian Hypercontractivity (Fact I.7),

$$\begin{aligned} \left\| \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}} \overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S) \right\|_q^2 &\leq \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} \alpha_{\boldsymbol{t}}^2 \cdot \sum_{\substack{\boldsymbol{c}_{1:N} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}_i\|_1 = t_i}} (q-1)^{\|\boldsymbol{c}_1\|_1 + \|\boldsymbol{c}_2\|_1 + \cdots + \|\boldsymbol{c}_N\|_1} \cdot \beta(\boldsymbol{c}_{1:N}; S)^2 \\ &= \sum_{\boldsymbol{t} \in \mathbb{N}_0^N} (q-1)^{\|\boldsymbol{t}\|_1} \cdot \alpha_{\boldsymbol{t}}^2 \cdot \sum_{\substack{\boldsymbol{c}_{1:N} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}_i\|_1 = t_i}} \beta(\boldsymbol{c}_{1:N}; S)^2 \end{aligned}$$

Observing that,

$$\mathbb{E}_0[\overline{H}_{\boldsymbol{t}}(\boldsymbol{x}_{1:N}; S)^2] = \sum_{\substack{\boldsymbol{c}_{1:N} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}_i\|_1 = t_i}} \beta(\boldsymbol{c}_{1:N})^2,$$

yields the claim of the lemma. □

APPENDIX G: FURTHER RESULTS FOR NON-GAUSSIAN COMPONENT ANALYSIS

This appendix provides some additional results for Non-Gaussian Component Analysis. It is organized as follows:

1. Appendix G.1 describes a connection between the k-NGCA problem and the problem of learning Gaussian mixture models.
2. Appendix G.2 describes a connection between the k-NGCA problem and the problem of learning binary generalized linear models.
3. Appendix G.3 provides two constructions for the non-Gaussian distributions.

G.1. Connections to Learning Mixtures of Gaussians. In this section, we describe a connection between the problem of learning mixtures of Gaussians and the k-NGCA problem. The following lemma provides a construction where ν is a mixture of Gaussian distributions. The k-NGCA problem with this particular non-Gaussian measure provides a hard instance for the problem of learning a Gaussian mixture model.

LEMMA G.1. *For each even $k = 2\ell \in \mathbb{N}$, there are two positive constants $\lambda_k > 0$ and $K_k < \infty$ (depending only on k) such that for any $\lambda \in (0, \lambda_k/2]$, there is a probability measure ν has the following properties:*

1. *ν is a mixture of Gaussians on $\mathbb{R}$ with ℓ components with equal variances:*

$$\nu = \sum_{i=1}^{\ell} p_i \cdot \mathcal{N}\left(w_i, \sigma^2\right),$$

for some probability weights $p_1, p_2, \ldots, p_\ell$, mean parameters $w_1, w_2, \ldots, w_\ell \in \mathbb{R}$, and variance parameter $0 < \sigma^2 < 1$.

2. *ν satisfies the Moment Matching Assumption (Assumption 1) with parameter k.*

3. *ν satisfies the Bounded Signal Strenth Assumption (Assumption 2) with parameters (λ, K_k).*
4. *ν satisfies the Locally Bounded Likelihood Ratio Assumption (Assumption 3) with parameters $(\lambda, K_k, \kappa = k)$.*
5. *ν satisfies the Minimum Signal Strength Assumption (Assumption 4) with parameters (λ, k).*
6. *ν is sub-Gaussian (Assumption 5) with variance proxy $\vartheta = 1$.*
7. *Furthermore we have $\lambda^{1/k}/K_k \leq \min_{i \neq j} |w_i - w_j| \leq \max_{i \neq j} |w_i - w_j| \leq K_k \cdot \lambda^{1/k}$.*

The proof of this result is provided in Appendix G.3.1. This construction also appears in the work of Diakonikolas, Kane and Stewart [14], who use it to prove computational lower bounds for estimating Gaussian Mixture Models in the SQ model. We use the above construction to relate k-NGCA to the problem of estimating Gaussian mixture models.

Mixtures of Gaussians and k-NGCA. Consider the problem of fitting a Gaussian mixture model, where the mixture components have identical but unknown covariance matrices. Formally, one is given a dataset $\boldsymbol{x}_{1:N} \in \mathbb{R}^d$ generated i.i.d. from the Gaussian mixture model:

$$\boldsymbol{x}_{1:N} \overset{\text{i.i.d.}}{\sim} \sum_{i=1}^{\ell} p_i \cdot \mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right), \tag{G.1}$$

where the mean vectors $\boldsymbol{\mu}_{1:\ell}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown. The goal is to estimate the mean vectors $\boldsymbol{\mu}_{1:\ell}$. Observe that when the dataset $\boldsymbol{x}_{1:N}$ is generated by the k-NGCA model with the non-Gaussian measure ν from Lemma G.1 and non-Gaussian direction $\boldsymbol{V}$, then, recalling (F.5), we obtain

$$\boldsymbol{x}_{1:N} \overset{\text{i.i.d.}}{\sim} \sum_{i=1}^{\ell} p_i \cdot \mathcal{N}\left(\frac{w_i}{\sqrt{d}}\boldsymbol{V}, \boldsymbol{I}_d - \frac{(1-\sigma^2)}{d}\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}\right).$$

This is an instance of a Gaussian mixture model (G.1). The parameter λ from the Bounded Signal Strength Assumption (Lemma G.1), which determines the statistical difficulty of the k-NGCA problem, can be reinterpreted as the minimum separation between the component means:

$$\lambda^{1/k} \asymp \min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|.$$

This is a natural notion of the signal strength for the fitting Gaussian mixture models. The parameter k from the Moment Matching Assumption (Lemma G.1), which determines the computational difficulty of the k-NGCA problem, relates to the number of mixing components in the Gaussian mixture model instance via the relation $k = 2\ell$. In summary, the lower bounds we prove for the k-NGCA problem automatically yield lower bounds for estimating Gaussian mixture models.

G.2. Connections to Learning Binary Generalized Linear Models. In this section, we describe a connection between the problem of learning binary Generalized Linear Models (GLMs) and the k-NGCA problem. The following lemma provides a construction of a non-Gaussian measure ν designed such that the likelihood ratio of ν with respect to the standard Gaussian measure $\mathcal{N}(0,1)$ is uniformly bounded. We will show that the k-NGCA problem with this particular non-Gaussian measure can be reduced to the problem of learning a binary GLM with a particular link function.

LEMMA G.2. *For every $k \in \mathbb{N}$, there is a positive constant $\lambda_k > 0$ that depends only on k such that, for any $\lambda \in (0, \lambda_k]$, there is a probability measure ν with the following properties:*

1. *ν satisfies the Moment Matching Assumption (Assumption 1) with parameter k.*
2. *ν has a bounded density with respect to $\mu_0 = \mathcal{N}(0,1)$:*

$$\left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) - 1\right| \leq \frac{\lambda}{\lambda_k} \leq 1.$$

   In particular, ν satisfies the Bounded Signal Strength Assumption (Assumption 2) with parameters $(\lambda, K = 1/\lambda_k)$ and the Locally Bounded Ratio Assumption (Assumption 3) with parameters $(\lambda, K = 1/\lambda_k, \kappa = 0)$.
3. *ν satisfies the Minimum Signal Strength Assumption (Assumption 4) with parameters (λ, k).*
4. *ν is sub-Gaussian (Assumption 5) with variance proxy $\vartheta \leq C$ for some universal constant C.*
5. *When k is odd, ν satisfies*

$$\frac{1}{2} \cdot \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) + \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(-x)\right) = 1.$$

The proof of this result is provided in Appendix G.3.2. We use the above construction to relate the k-NGCA problem to the problem of learning binary generalized linear models, which we introduce below. The connection described below is also implicit in the work of Diakonikolas et al. [15], which studies SQ lower bounds for agnostic learning of half-spaces.

Generalized linear models and k-NGCA. Consider the problem of fitting a binary generalized linear model (GLM) with Gaussian covariates. One observes a data set consisting of N feature-response pairs $\{(\boldsymbol{f}_i, r_i) : i \in [N]\} \subset \mathbb{R}^d \times \{0,1\}$ sampled i.i.d. as follows:

$$\boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad r_i \mid \boldsymbol{f}_i \sim \text{Bernoulli}\left(\rho\left(\frac{\langle \boldsymbol{f}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right)\right). \tag{G.2}$$

In the above display, $\boldsymbol{V} \in \mathbb{R}^d$ is the unknown parameter of interest with $\|\boldsymbol{V}\| = \sqrt{d}$, and $\rho : \mathbb{R} \to [0,1]$ is a known but arbitrary regression function. The goal is to estimate the vector $\boldsymbol{V}$.

The GLM learning problem is closely related to the k-NGCA problem, because for certain non-Gaussian measures ν (including those coming from Lemma G.2 with k odd), it is possible to transform a dataset $\boldsymbol{x}_{1:N}$ sampled from the k-NGCA problem with non-Gaussian direction $\boldsymbol{V}$ into a dataset $\{(\boldsymbol{f}_i, r_i) : i \in [N]\}$ sampled from the GLM (G.2). Consequently, estimators designed for learning GLMs can be used to solve the k-NGCA problem. Hence, the lower bounds we prove for k-NGCA immediately yield lower bounds for the GLM learning problem.

We now describe the transformation that converts a dataset $\boldsymbol{x}_{1:N}$ for the k-NGCA problem to a dataset $\{(\boldsymbol{f}_i, r_i) : i \in [N]\}$ for the GLM learning problem:

$$r_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad \boldsymbol{f}_i = (2r_i - 1) \cdot \boldsymbol{x}_i. \tag{G.3}$$

We verify that $(\boldsymbol{f}_i, r_i)$ are samples from (G.2). First we observe that conditioned on r_i, we can compute the distribution of $\boldsymbol{f}_i$:

$$\boldsymbol{f}_i \mid r_i = 1 \sim \mu_{\boldsymbol{V}}, \quad \boldsymbol{f}_i | r_i = 0 \sim \mu_{\boldsymbol{V}}^-, \tag{G.4}$$

where $\mu_{\boldsymbol{V}}$ is the measure from (F.5), and $\mu_{\boldsymbol{V}}^-$ is the measure defined as follows by its likelihood ratio with respect to $\mu_0 = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$:

$$\frac{\mu_{\boldsymbol{V}}^-}{\mathrm{d}\mu_0}(\boldsymbol{x}) \overset{\text{def}}{=} \frac{\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(-\boldsymbol{x}) \overset{\text{(F.7)}}{=} \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}\left(-\frac{\langle \boldsymbol{x}, \boldsymbol{V}\rangle}{\sqrt{d}}\right).$$

If the likelihood ratio of the non-Gaussian distribution ν with respect to $\mathcal{N}(0,1)$ satisfies

$$\text{(G.5)} \qquad \frac{1}{2}\cdot\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x)+\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(-x)\right)=1,$$

then computing the marginal distribution of $\boldsymbol{f}_i$ from (G.4) yields $\boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. The requirement in (G.5) is satisfied, for instance, when ν is the non-Gaussian measure constructed in Lemma G.2 for odd k. Under this condition, an application of Bayes' rule to (G.4) gives the conditional distribution of $r_i \mid \boldsymbol{f}_i$:

$$r_i \mid \boldsymbol{f}_i \sim \text{Bernoulli}\left(\frac{1}{2}\cdot\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}\left(\frac{\langle \boldsymbol{f}_i, \boldsymbol{V}\rangle}{\sqrt{d}}\right)\right).$$

This verifies the transformation in (G.3) produces an instance of the GLM learning problem with the regression function

$$\text{(G.6)} \qquad \rho(\xi)=\frac{1}{2}\cdot\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(\xi).$$

Finally, we estimate the parameters λ and k, which respectively determine the statistical and computational difficulty of the k-NGCA problem in terms of the regression function ρ. Recall that when the non-Gaussian measure satisfies the Minimum Signal Strength Assumption (Assumption 4) and the Bounded Signal Strength Assumption (Assumption 2), we have

$$\lambda^2 \asymp \text{Var}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)\right), \quad k=\min\left\{\ell\in\mathbb{N}:\mathbb{E}\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)\cdot H_\ell(Z)\right]\neq 0\right\}, \quad Z\sim\mathcal{N}(0,1).$$

Hence, (G.6) shows that the statistical difficulty of the GLM learning problem is determined by

$$\text{(G.7)} \qquad \lambda^2 \asymp \text{Var}\left(\rho(Z)\right),$$

where as the computational difficulty is determined by

$$\text{(G.8)} \qquad k=\min\left\{\ell\in\mathbb{N}:\mathbb{E}\left[\rho(Z)\cdot H_\ell(Z)\right]\neq 0\right\}.$$

We note that $\text{Var}(\rho(Z))$ appears to be a natural notion of signal strength for the GLM learning problem, since if $\text{Var}(\rho(Z))=0$, we have $\rho(\xi)=1/2$ almost everywhere. This means that the feature and response are independent and carry no information about the parameter $\boldsymbol{V}$. An analog of (G.8) (for the case when $k=2$) appears in the work of Mondelli and Montanari [29, Theorem 2], who show that if $\mathbb{E}[\rho(Z)H_2(Z)]=0$, then a broad class of spectral estimators fail to have a non-trivial performance in the regime $N\asymp d$ and $\lambda\asymp 1$. Furthermore, the hard instance constructed in the work of Diakonikolas et al. [15, Proposition 2.1] to prove SQ lower bounds for agnostic learning of half-spaces has the property that the parameter k (as defined in (G.8)) is large.

G.3. Constructions of Non-Gaussian Distributions. In this section we provide the proofs for Lemma G.1 and Lemma G.2.

G.3.1. *Proof of Lemma G.1.* The proof of Lemma G.1 relies on the following fact.

FACT G.1 (38, Section 2.7). Let $k=2\ell$ be even. There is a discrete random variable W with support size ℓ such that

$$\mathbb{E}H_i(W)=0 \quad \forall\, 1\leq i\leq k-1,$$

and, $\mathbb{E}H_k(W)=-\ell!/\sqrt{k!}$. Furthermore, W is a bounded random variable $|W|\leq\sqrt{2k+2}$ and is sub-Gaussian with variance proxy 1.

With this fact, we can now provide a proof for Lemma G.1.

PROOF OF LEMMA G.1. Let W be any bounded random variable from Fact G.1 with the property that $\mathbb{E}H_i(W) = 0$ for any $1 \leq i \leq k-1$ and $\mathbb{E}H_k(W) \neq 0$. Let $Z \sim \mathcal{N}(0,1)$ be independent of W. Define:

$$\lambda_k \overset{\text{def}}{=} |\mathbb{E}Z^k H_k(Z)| \cdot |\mathbb{E}H_k(W)| > 0.$$

For any $\lambda \leq \lambda_k$, we claim that the law ν of the random variable W_λ defined by

$$W_\lambda \overset{\text{def}}{=} \gamma(\lambda) \cdot W + \sqrt{1-\gamma^2(\lambda)} \cdot Z, \quad \gamma(\lambda) \overset{\text{def}}{=} \left(\frac{\lambda}{\lambda_k}\right)^{\frac{1}{k}}$$

is a non-Gaussian measure with the desired properties.

We begin by computing $\mathbb{E}H_t(W_\lambda)$. Recall the generating function for the Hermite Polynomials: for any $x, w \in \mathbb{R}$, we have

$$e^{xw - \frac{x^2}{2}} = \sum_{t=0}^{\infty} \frac{x^t}{\sqrt{t!}} H_t(w).$$

In particular:

$$H_t(w) = \frac{1}{\sqrt{t!}} \frac{\mathrm{d}^t}{\mathrm{d}x^t} e^{xw - \frac{x^2}{2}} \bigg|_{x=0}. \tag{G.9}$$

Hence,

$$\begin{aligned}
\mathbb{E}H_t(W_\lambda) &= \frac{1}{\sqrt{t!}} \frac{\mathrm{d}^t}{\mathrm{d}x^t} \mathbb{E}e^{xW_\lambda - \frac{x^2}{2}} \bigg|_{x=0} \\
&= \frac{1}{\sqrt{t!}} \frac{\mathrm{d}^t}{\mathrm{d}x^t} \mathbb{E}e^{x\gamma W + x\sqrt{1-\gamma^2} Z - \frac{x^2}{2}} \bigg|_{x=0} \\
&= \frac{1}{\sqrt{t!}} \frac{\mathrm{d}^t}{\mathrm{d}x^t} \mathbb{E}e^{x\gamma W - \frac{x^2\gamma^2}{2}} \bigg|_{x=0} \\
&= \frac{1}{\sqrt{t!}} \mathbb{E} \frac{\mathrm{d}^t}{\mathrm{d}x^t} e^{x\gamma W - \frac{x^2\gamma^2}{2}} \bigg|_{x=0}.
\end{aligned}$$

Applying the differential identity in (G.9) after making the change of variables $z = \gamma x$, we obtain,

$$\mathbb{E}H_t(W_\lambda) = \frac{1}{\sqrt{t!}} \mathbb{E} \frac{\mathrm{d}^t}{\mathrm{d}x^t} e^{x\gamma W - \frac{x^2\gamma^2}{2}} \bigg|_{x=0} = \gamma^t(\lambda) \cdot \mathbb{E}H_t(W).$$

Recalling the properties of W stated in Fact G.1, we obtain,

$$\mathbb{E}H_t(W_\lambda) = \begin{cases} 0 & \text{if } t \leq k-1; \\ \frac{\lambda}{\lambda_k} \cdot \mathbb{E}H_k(W) & \text{if } t = k. \end{cases} \tag{G.10}$$

Furthermore, Bonan and Clark [9, Theorem 1] have shown:

$$C \overset{\text{def}}{=} \sup_{t \in \mathbb{N}_0} \sup_{w \in \mathbb{R}} \left\{ |H_t(w)| \cdot e^{-\frac{w^2}{2}} \right\} < \infty.$$

Consequently, we have

$$|\mathbb{E}H_t(W_\lambda)| \leq C e^{\frac{\|W\|_\infty^2}{2}} \cdot \left(\frac{\lambda}{\lambda_k}\right)^{\frac{t}{k}} \tag{G.11}$$

Using (G.10) and (G.11), we can now establish the desired properties of ν:

1. By (G.10), we see that $\mathbb{E}H_t(Z) = \mathbb{E}H_t(W_\lambda)$ for any $t \leq k-1$. This immediately yields $\mathbb{E}Z^t = \mathbb{E}W_\lambda^t$ for any $t \leq k-1$. Hence, ν satisfies the Moment Matching Assumption with parameter k.
2. We expand the likelihood ratio in the Hermite basis:

$$\begin{aligned} \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) &= \sum_{t=0}^{\infty} \left(\mathbb{E}H_t(Z) \cdot \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z)\right) H_t(z) \\ &= \sum_{t=0}^{\infty} \mathbb{E}H_t(W_\lambda) \cdot H_t(z) \\ &= 1 + \sum_{t=k}^{\infty} \mathbb{E}H_t(W_\lambda) \cdot H_t(z). \end{aligned}$$

   In the above display, in the last step, we used the fact that $\mathbb{E}H_t(W_\lambda) = 0$ for any $1 \leq t \leq k-1$. To verify that the second moment of the likelihood ratio is bounded, we note that,

$$\mathbb{E}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z) - 1\right)^2 = \sum_{t=k}^{\infty} |\mathbb{E}H_t(W_\lambda)|^2.$$

   Using the estimates in (G.10) and (G.11) and the assumption $\lambda/\lambda_k \leq 1/2$, we obtain,

$$\text{(G.12)} \qquad \mathbb{E}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(Z) - 1\right)^2 \leq \frac{C^2 \cdot e^{\|W\|_\infty^2}}{\lambda_k^2} \cdot \frac{2^{2/k}}{2^{2/k}-1} \cdot \lambda^2.$$

   This verifies the Bounded Signal Strength Assumption.
3. In order to verify that the likelihood ratio is locally bounded, we begin with the estimate:

$$\left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) - 1\right| \leq \sum_{t=k}^{\infty} |\mathbb{E}H_t(W_\lambda)| \cdot |H_t(z)| \leq Ce^{\frac{\|W\|_\infty^2}{2}} \sum_{t=k}^{\infty} \left(\frac{\lambda}{\lambda_k}\right)^{\frac{t}{k}} \cdot |H_t(z)|.$$

   Fact I.2 shows that $|H_t(z)| \leq (1+|z|)^t$. Hence,

$$\left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) - 1\right| \leq Ce^{\frac{\|W\|_\infty^2}{2}} \sum_{t=k}^{\infty} \left(\frac{\lambda}{\lambda_k}\right)^{\frac{t}{k}} \cdot (1+|z|)^t.$$

   Under the assumption:

$$\text{(G.13)} \qquad \frac{\lambda}{\lambda_k} \cdot (1+|z|)^k \leq \frac{1}{2},$$

   we obtain,

$$\text{(G.14)} \qquad \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(z) - 1\right| \leq \frac{Ce^{\frac{\|W\|_\infty^2}{2}}}{\lambda_k} \cdot \frac{2^{1/k}}{2^{1/k}-1} \cdot \lambda \cdot (1+|z|)^k.$$

   Inspecting (G.13) and (G.14), we obtain that the Locally Bounded Likelihood Ratio Assumption holds with

$$\kappa = k, \quad K = \frac{Ce^{\frac{\|W\|_\infty^2}{2}}}{\lambda_k} \cdot \frac{2^{1/k}}{2^{1/k}-1}.$$

4. Recall that the monomial w^k can be written as a linear combination of $\{H_t(w)\}_{t \leq k}$:

$$w^k = \sum_{t=0}^{k} a_t H_t(w), \quad a_t = \mathbb{E}Z^t H_t(Z).$$

Hence,

$$|\mathbb{E}Z^k - \mathbb{E}W_\lambda^k| = |a_k \cdot \mathbb{E}H_k(W_\lambda)| = \lambda \cdot \frac{|\mathbb{E}Z^k H_k(Z)| \cdot |\mathbb{E}H_k(W)|}{\lambda_k} = \lambda.$$

This verifies the Minimum Signal Strength Assumption (Assumption 4).

5. Observe that $\mathbb{E}e^{tW_\lambda} = \mathbb{E}[e^{t\gamma(\lambda)W}] \cdot e^{t^2(1-\gamma^2(\lambda)/2}$. Since W is 1 sub-Gaussian, $\mathbb{E}e^{tW_\lambda} \leq e^{t^2/2}$, which verifies that ν is 1 sub-Gaussian.

This concludes the proof of this lemma. □

G.3.2. *Proof of Lemma G.2.*

PROOF OF LEMMA G.2. Consider the vector space of polynomials on $\mathbb{R}$. On this vector space, define the inner product:

$$\langle f, g \rangle_\Delta \overset{\text{def}}{=} \int_{\mathbb{R}} f(x)g(x)\Delta(x)\mu_0(\mathrm{d}x),$$

where the weight function $\Delta(x)$ is defined as:

$$\Delta(x) = \begin{cases} 1 & \text{if } |x| \leq 1; \\ 0 & \text{if } |x| > 1. \end{cases}.$$

Let $(H_i^\Delta)_{i \in \mathbb{N}_0}$ denote the orthonormal polynomials obtained by the Gram-Schmidt orthogonalization of the ordered linearly independent collection $(x^i)_{i \in \mathbb{N}_0}$. In particular, for all $i, j \in \mathbb{N}_0$,

- $\left\langle H_i^\Delta, H_j^\Delta \right\rangle_\Delta = \delta_{ij}$,
- $\mathrm{Span}(\{1, x, \dots, x^i\}) = \mathrm{Span}(\{H_0^\Delta, H_1^\Delta, \dots, H_i^\Delta\})$,
- The degree of H_i^Δ is exactly i.

Define

$$\|H_k^\Delta \cdot \Delta\|_\infty \overset{\text{def}}{=} \sup_{x \in \mathbb{R}} |H_k^\Delta(x)\Delta(x)|, \qquad \lambda_k \overset{\text{def}}{=} \frac{|\langle x^k, H_k^\Delta \rangle_\Delta|}{\|H_k^\Delta \cdot \Delta\|_\infty}.$$

Since polynomials are uniformly bounded on compact sets, we have $\|H_k^\Delta \cdot \Delta\|_\infty = \sup_{|x| \leq 1} |H_k^\Delta(x)| < \infty$. Furthemore, we observe that $\lambda_k \neq 0$ (otherwise x^k lies in the span of $H_{0:k-1}^\Delta$, which is not possible since x^k has degree k). With these definitions, we are ready to construct the measure ν as follows:

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) \overset{\text{def}}{=} 1 + \frac{\lambda}{\lambda_k} \frac{H_k^\Delta(x) \cdot \Delta(x)}{\|H_k^\Delta \cdot \Delta\|_\infty}. \tag{G.15}$$

We first check that the above ν is a valid probability measure. The density defined above is non-negative for any $0 \leq \lambda \leq \lambda_k$. Furthermore,

$$\int \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) = 1 + \frac{\lambda}{\lambda_k} \cdot \frac{\langle H_k^\Delta, 1 \rangle_\Delta}{\|H_k^\Delta \cdot \Delta\|_\infty} = 1 + \frac{\lambda}{\lambda_k} \cdot \frac{\cdot\sqrt{\langle 1, 1 \rangle_\Delta} \cdot \langle H_k^\Delta, H_0^\Delta \rangle_\Delta}{\|H_k^\Delta \cdot \Delta\|_\infty} \overset{\text{(a)}}{=} 1.$$

In the step marked (a), we used the orthogonality property $\langle H_k^\Delta, H_0^\Delta \rangle_\Delta = 0$ for any $k \in \mathbb{N}$. Hence ν defines a valid probability measure. Next, we verify each of the claims in the statement of the lemma.

1. For any $i \le k-1$, since x^i lies in the span of $H^{\Delta}_{1:k-1}$, we have $\langle x^i, H^{\Delta}_k \rangle_{\Delta} = 0$. Consequently,
$$\int x^i \nu(\mathrm{d}x) = \int x^i \cdot \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) \cdot \mu_0(\mathrm{d}x) = \mathbb{E}Z^i + \frac{\lambda}{\lambda_k} \cdot \frac{\langle x^i, H^{\Delta}_k \rangle_{\Delta}}{\|H^{\Delta}_k \cdot \Delta\|_{\infty}} = \mathbb{E}Z^i.$$
2. This claim is immediate from the formula in (G.15).
3. Following the same steps as in the proof of item (1), we obtain
$$\left| \int x^k \nu(\mathrm{d}x) - \mathbb{E}Z^k \right| = \frac{\lambda}{\lambda_k} \cdot \frac{|\langle x^k, H^{\Delta}_k \rangle_{\Delta}|}{\|H^{\Delta}_k \cdot \Delta\|_{\infty}} = \lambda.$$
4. Observe that
$$\int |x|^i \, \nu(\mathrm{d}x) = \int |x|^i \cdot \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) \, \mu_0(\mathrm{d}x) \le 2\mathbb{E}|Z|^i, \; Z \sim \mathcal{N}(0,1).$$
Hence, ν is sub-Gaussian with variance proxy $\vartheta \le C$ for some universal constant C.
5. An inductive argument shows that H^{Δ}_i is an odd function for odd i and an even function for even i. Hence when k is odd, (G.15) gives:
$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(x) + \frac{\mathrm{d}\nu}{\mathrm{d}\mu_0}(-x) = 1,$$
as claimed.

This concludes the proof of this lemma. □

APPENDIX H: CANONICAL CORRELATION ANALYSIS

This appendix presents introduces the order-k Canonical Correlation Analysis (k-CCA) problem and presents a computational lower bound for this problem. The appendix is organized as follows:

1. Appendix H.1 formally defines the k-CCA problem inference problem. The computational-statistical gap in k-CCA is discussed in Appendix H.2
2. A formal statement of the k-CCA computational lower bound appears in Appendix H.3.
3. Appendix H.4 formalizes a connection between k-CCA and the well-studied problem problem of learning parities with noise. This allows us to obtain a computational lower bound for learning parities with noise.
4. The proof of the k-CCA computational lower bound is provided in Appendix H.5. The proof relies on an information bound for the distributed k-CCA problem, which is proved in Appendix H.6.

H.1. Problem Formulation. In the order-k Canonical Correlation Analysis (k-CCA) problem, one observes a dataset of N i.i.d. samples $\boldsymbol{x}_{1:N}$, in which each $\boldsymbol{x}_i \in \mathbb{R}^{kd}$ consists of k "views" (or "modes"):
$$\boldsymbol{x}_i = (\boldsymbol{x}_i{}^{(1)}, \boldsymbol{x}_i{}^{(2)}, \ldots, \boldsymbol{x}_i{}^{(k)})^{\intercal}.$$
The correlation structure between the different views is such that
$$\mathbb{E}\left[\boldsymbol{x}_i{}^{(1)} \otimes \boldsymbol{x}_i{}^{(2)} \otimes \cdots \otimes \boldsymbol{x}_i{}^{(k)}\right] = \frac{\lambda}{\sqrt{d^k}} \boldsymbol{V}, \tag{H.1}$$
where $\boldsymbol{V} \in \bigotimes^k \mathbb{R}^d$ is the rank-1 cross-moment tensor:
$$\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k,$$

for some unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k \in \mathbb{R}^d$. Note that $\|\boldsymbol{V}\| = \sqrt{d^k}$. The parameter $\lambda > 0$ is the signal-to-noise ratio parameter. The goal is to estimate the cross-moment tensor $\boldsymbol{V}$.

Note that we have not explicitly specified the probability measure $\mu_{\boldsymbol{V}}$ of $\boldsymbol{x}_i$, as the goal of estimating correlation structure is often considered in a non-parametric setting. However, our lower bounds will consider a particular measure $\mu_{\boldsymbol{V}}$ specified by its likelihood ratio with respect to $\mu_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{kd})$:

$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) \overset{\text{def}}{=} 1 + \frac{\lambda}{\lambda_k} \cdot \mathsf{sign}\left(\frac{\langle \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(k)}, \boldsymbol{V} \rangle}{\sqrt{d^k}}\right), \tag{H.2a}$$

where

$$\lambda_k \overset{\text{def}}{=} \left(\frac{2}{\pi}\right)^{\frac{k}{2}} = (\mathbb{E}|Z|)^{\frac{k}{2}}, \quad Z \sim \mathcal{N}(0,1). \tag{H.2b}$$

This is a valid probability distribution that satisfies (H.1) as long as $0 \leq \lambda \leq \lambda_k$.

Finally, our computational lower bound for k-CCA will hold even under further restrictions on $\boldsymbol{V}$, namely

$$\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \cdots \otimes \boldsymbol{e}_{i_k} \tag{H.3}$$

for some $\{i_1, i_2, \dots, i_k\} \subseteq [d]$, where $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_d$ are the standard basis vectors in $\mathbb{R}^d$. This restriction is relevant to the connection between k-CCA and the parity learning problem.

H.2. Statistical-Computational Gap in k-CCA. The k-CCA problem exhibits the same computational gap as the other inference problems studied in this paper. Depending upon the effective sample size $N\lambda^2$, the k-CCA problem exhibits the following three phases:

Impossible phase. When the effective sample size $N\lambda^2 \ll d$, there is no consistent estimator for $\boldsymbol{V}$. This follows from standard lower bounds based on Fano's Inequality.[1]

Conjectured hard phase. In the regime $d \lesssim N\lambda^2 \ll d^{k/2}$, there is a consistent, but computationally inefficient estimator for the cross-moment tensor $\boldsymbol{V}$; see [16, Theorem 7, Appendix G.1] for details. We believe that no polynomial-time estimator can recover $\boldsymbol{V}$ in this phase. In the arXiv version of this paper [16, Proposition 10, Appendix G.2], we give evidence for this conjecture using the low-degree likelihood ratio framework.

Easy phase. In the regime $N\lambda^2 \gg d^{k/2}$, there are polynomial-time estimators for the k-CCA problem. The correlation structure in (H.1) suggests that $\boldsymbol{V}$ can be estimated by the rank-1 approximation to the empirical cross-moment tensor:

$$\hat{\boldsymbol{T}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i^{(1)} \otimes \boldsymbol{x}_i^{(2)} \otimes \cdots \otimes \boldsymbol{x}_i^{(k)}.$$

However, computing a rank-1 approximation to an order-k tensor is non-trivial for $k \geq 3$. For even $k = 2\ell$, we can reshape $\hat{\boldsymbol{T}}$ to a $d^{\frac{k}{2}} \times d^{\frac{k}{2}}$ matrix $\mathrm{Mat}(\hat{\boldsymbol{T}})$, as was done for k-ATPCA, where the matricization operation $\mathrm{Mat}(\cdot)$ was defined as follows:

$$\mathrm{Mat}(\boldsymbol{T})_{(i_1,i_2,\dots,i_\ell);(j_1,j_2,j_3,\dots,j_\ell)} \overset{\text{def}}{=} T_{i_1,i_2,\dots,i_\ell,j_1,j_2,\dots,j_\ell} \quad i_{1:\ell} \in [d],\ j_{1:\ell} \in [d]. \tag{H.4}$$

[1] This follows from similar arguments as those used to prove the corresponding result for k-NGCA in arXiv version of this paper [16, Proposition 7, Appendix F.2].

To estimate $\boldsymbol{V}$, we first estimate $\mathsf{Mat}(\boldsymbol{V})$ by computing the best rank-1 approximation to $\mathsf{Mat}(\hat{\boldsymbol{T}})$ using SVD:

$$(\hat{\boldsymbol{U}}^{(L)}, \hat{\boldsymbol{U}}^{(R)}) \overset{\text{def}}{=} \arg \max_{\substack{\|\boldsymbol{U}^{(L)}\|=1 \\ \|\boldsymbol{U}^{(R)}\|=1}} \left\langle \boldsymbol{U}^{(L)}, \mathsf{Mat}(\hat{\boldsymbol{T}}) \cdot \boldsymbol{U}^{(R)} \right\rangle. \tag{H.5a}$$

We then construct an estimate $\hat{\boldsymbol{V}}$ of $\boldsymbol{V}$ by reshaping $\hat{\boldsymbol{U}}^{(L)} \otimes \hat{\boldsymbol{U}}^{(R)}$ into a tensor:

$$\hat{\boldsymbol{V}} \overset{\text{def}}{=} \mathsf{Mat}^{-1}(\hat{\boldsymbol{U}}^{(L)} \otimes \hat{\boldsymbol{U}}^{(R)}). \tag{H.5b}$$

Under an additional concentration assumption, we analyze this spectral estimator in the longer arXiv version of this paper [16, Theorem 8, Appendix G.3] and show that when $N\lambda^2 \gg d^{k/2}$, $\hat{\boldsymbol{V}}$ is a consistent estimator for $\boldsymbol{V}$.

H.3. Computational Lower Bound for k-CCA.

The following is our computational lower bound for k-CCA.

THEOREM H.1. *Consider the k-CCA problem for $k \geq 2$ with signal-to-noise ratio $\lambda^2 \asymp d^{-\gamma}$ (as $d \to \infty$) for any constant $\gamma > 3k/2$. Let $\hat{\boldsymbol{V}} \in \bigotimes^k \mathbb{R}^d$ denote any estimator for this k-CCA problem that can be computed using a memory bounded estimation algorithm with resource profile (N, T, s) scaling with d as*

$$N\lambda^2 \asymp d^{\eta}, \quad T \asymp d^{\tau}, \quad s \asymp d^{\mu}$$

for any constants $\eta \geq 1$, $\tau \geq 0$, $\mu \geq 0$. If

$$\eta + \tau + \mu < k,$$

then, for any $t \in \mathbb{R}$,

$$\limsup_{d \to \infty} \inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}} \left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}} \rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d^k} \right) \leq \frac{1}{t^2}.$$

These results hold even when $\boldsymbol{V}$ and $\mu_{\boldsymbol{V}}$ are promised to satisfy (H.3) *and* (H.2).

Theorem H.1 shows that if the signal-to-noise ratio λ is sufficiently small, then memory bounded estimation algorithms using too few total resources (as measured by the product $N\lambda^2 \cdot T \cdot s$) perform no better than a random guess. Given the close relationship between k-CCA and k-ATPCA (analogous to that between k-NGCA and k-TPCA), it is not surprising that Theorem H.1 and Theorem 2 are quantitatively similiar (modulo the condition on the signal-to-noise ratio). So, most of the implications discussed in Section 6.4 regarding k-ATPCA continue to hold for k-CCA.

H.4. Connections to Learning Parities.

Learning parity functions from labeled examples is a well-studied problem in computational learning theory with numerous connections to cryptography and coding theory [4, 7, 8, 6, 28, 17, 36, 26, 24, 32, 19, 18]. In our generalization of this problem, one observes a data set consisting of N feature-response pairs $\{(\boldsymbol{f}_i, r_i) : i \in [N]\} \subset \mathbb{R}^D \times \{0,1\}$ sampled i.i.d. as follows:

$$\boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D), \quad r_i \mid \boldsymbol{f}_i \sim \text{Bernoulli}\left(\frac{1}{2} + \frac{\Lambda}{2} \cdot \prod_{j=1}^{k} \mathsf{sign}(\langle \boldsymbol{v}_j, \boldsymbol{f}_i \rangle) \right). \tag{H.6}$$

In the above display, $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k \in \mathbb{R}^D$ are unknown parameters with $\|\boldsymbol{v}_i\| = 1$, and $\Lambda \in [0,1]$ controls the signal-to-noise ratio of the problem. The goal is to estimate the parameter $\boldsymbol{V}$:

$$\boldsymbol{V} = \sqrt{D^k} \cdot \boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k.$$

Depending on the assumptions made on k and $\boldsymbol{v}_{1:k}$, one obtains the following different variants of the original parity learning problem:

1. If $\boldsymbol{v}_1 = \boldsymbol{e}_{i_1}, \boldsymbol{v}_2 = \boldsymbol{e}_{i_2}, \dots, \boldsymbol{v}_k = \boldsymbol{e}_{i_k}$ for some unknown subset $\{i_1, i_2, \dots, i_k\} \subset [D]$, then this is the problem of *learning k-sparse parities with noise* (k-LPN). Here, $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_D$ are the standard basis vectors in $\mathbb{R}^D$, and we typically consider $k \asymp 1$.
2. The generalization of k-LPN where $k \in [D]$ is arbitrary (possibly growing with D, and also possibly unknown) is called the problem of *learning (non-sparse) parities with noise* (LPN).
3. If $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k$ is an unknown collection of mutually orthogonal unit vectors, then this is the problem of *learning k-sparse parities with noise in an unknown basis*.

The computational lower bounds for k-CCA derived in this paper have interesting implications for each of the three variants of the parity problem introduced above. This is because it is possible the hard instance of k-CCA used to prove the computational lower bounds in this paper can be transformed into an instance of k-LPN (for odd k). Since k-LPN is the simplest of the three variants of the parity learning problem introduced above, an estimator for any of the three variants can be used to solve a k-LPN instance. This means that the lower bounds for k-CCA derived in this paper immediately yield computational lower bounds for each of the variants of parity learning problem mentioned above. To make this connection precise, we give a reduction from the hard instance of k-CCA studied in this paper to k-LPN.

Reduction to k-LPN. In the hard instances of k-CCA considered in Theorem H.1, the cross-moment tensor has the form $\boldsymbol{V} = \sqrt{d^k}\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_k$ where $\boldsymbol{v}_j = \boldsymbol{e}_{i_j}$ for some $i_1, i_2, \dots, i_k \in [d]$, as per (H.3). The dataset $\boldsymbol{x}_{1:N} \in \mathbb{R}^{kd}$ is sampled i.i.d. from the probability distribution $\mu_{\boldsymbol{V}}$ defined via the likliehood ratio in (H.2)

We transform a k-CCA dataset $\boldsymbol{x}_{1:N}$ into the k-LPN dataset $\{(\boldsymbol{f}_i, r_i) : i \in [N]\} \subset \mathbb{R}^{kd} \times \{0,1\}$ as follows:

$$r_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad \boldsymbol{f}_i = (2r_i - 1)\boldsymbol{x}_i.$$

Since this specifies the joint distribution of $(r_i, \boldsymbol{f}_i)$, one can compute the marginal distribution of $\boldsymbol{f}_i$ and the conditional distribution of r_i given $\boldsymbol{f}_i$ using this information. When k is odd and if $\boldsymbol{x}_{1:N} \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}$ for $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \cdots \otimes \boldsymbol{e}_{i_k} : i_{1:k} \in [d]$, we find that

$$\boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{kd}), \quad r_i | \boldsymbol{f}_i \sim \text{Bernoulli}\left(\frac{1}{2} + \frac{\lambda}{2\lambda_k} \cdot \prod_{j=1}^{k} \mathsf{sign}(\langle \boldsymbol{v}_j, \boldsymbol{f}_j \rangle)\right),$$

where

$$\boldsymbol{v}_j = \Bigg(\underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{j-1 \text{ times}}, \boldsymbol{e}_{i_j}, \underbrace{\boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}}_{k-j \text{ times}}\Bigg) \ \forall\, j \in [k].$$

This verifies that $\{(\boldsymbol{f}_i, r_i) : i \in [N]\} \subset \mathbb{R}^{kd} \times \{0,1\}$ is an instance of the k-LPN in dimension $D = kd$ with signal-to-noise ratio $\Lambda = \lambda/\lambda_k$.

Implications for Learning (Non-Sparse) Parities with Noise. To discuss the implications of the computational lower bound in Theorem H.1, we focus on the problem of learning non-sparse parities. Recall that in this problem, one is given a data set consisting of N feature-response pairs $\{(\boldsymbol{f}_i, r_i) : i \in [N]\} \subset \mathbb{R}^D \times \{0,1\}$ sampled i.i.d. as follows:

$$\text{(H.7)} \qquad \boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D), \quad r_i \sim \text{Bernoulli}\left(\frac{1}{2} + \frac{\Lambda}{2} \cdot \prod_{j \in S} \mathsf{sign}(f_{ij})\right),$$

where $S \subset [D]$ is the unknown parameter of interest. While this problem can be solved efficiently with $N = D$ samples using Gaussian elimination when $\Lambda = 1$ (the noiseless setting), this problem is believed to exhibit a large computational gap when $\Lambda < 1$ (the noisy setting). The MLE for this problem consistently estimates S with a sample size $N \gtrsim D/\Lambda^2$, but requires an exhaustive search over all 2^D possible subsets of $[D]$. No estimator with a $\mathrm{poly}(D, 1/\Lambda)$ sample complexity and $\mathrm{poly}(D, 1/\Lambda)$ run-time is currently known. Some notable algorithms[2] that improve over the run-time of exhaustive search include the following.

1. An algorithm due to Blum, Kalai and Wasserman [6] that solves LPN with $N = 2^{O(D/\log(D))}$ samples and run-time in the regime $\Lambda \geq 2^{-O(D^\delta)}$ for any $\delta < 1$.[3]
2. An algorithm due to Lyubashevsky [28] that solves LPN using $N \lesssim D^{1+\epsilon}$ and run-time $2^{O(D/\log\log(D))}$ in the regime $\Lambda \geq 2^{-O(\log^\delta(D))}$ for any $\epsilon > 0$ and $\delta < 1$.
3. An algorithm due to Valiant [36] that solves k-LPN using $N \lesssim D^{(1+\epsilon)2k/3}/\Lambda^{2+\epsilon}$ and run-time $O((D^{(1+\epsilon)k/3}/\Lambda^{2+\epsilon})^\omega)$ for any $\epsilon > 0$, where $\omega < 2.372$ is the matrix multiplication exponent. Note that the exponent on D in the run-time is less than $0.8k$.

The SQ framework has been used to provide evidence for the hardness of learning parities in the work of Kearns [23] and Blum et al. [8]. The latter work shows that any SQ algorithm which learns noisy parities with a sample size $N \leq 2^{D/3}$ must make at least $2^{D/3}/2$ queries. Using the reduction between k-CCA and k-LPN outlined previously, we can obtain the following corollary for learning (non-sparse) parities.

COROLLARY H.1. *Consider the problem of learning non-sparse parities in dimension D with signal-to-noise ratio $\Lambda^2 \asymp D^{-\gamma}$ (as $D \to \infty$). Let $\hat{S}$ be any estimator of S computed using a memory bounded estimation algorithm with resource profile (N, T, s) scaling with D as*

$$N\Lambda^2 \asymp D^\eta, \quad T \asymp D^\tau, \quad s \asymp \frac{D^\mu}{\Lambda^\alpha}$$

for any constants $\eta \geq 1,\ \tau \geq 0,\ \mu \geq 0,\ \alpha < 4/3$. If

$$\gamma > \frac{2(\eta + \tau + \mu + 2)}{4/3 - \alpha},$$

then

$$\lim_{D \to \infty} \inf_{S \subset [D]} \mathbb{P}_S\left(S = \hat{S}\right) = 0.$$

[2] These works in fact study the Boolean version of the (non-sparse) parity problem where the features are drawn from $\mathsf{Unif}\left(\{\pm 1\}^D\right)$. However, the Gaussian and Boolean parity problems are statistically and computationally equivalent. Given a sample $(\boldsymbol{f}, y)$ from the Gaussian parity problem, $(\mathsf{sign}(\boldsymbol{f}), y)$ is a sample from the Boolean parity problem where $\mathsf{sign}(\cdot)$ acts entry-wise on $\boldsymbol{f}$. Likewise, given a sample $(\boldsymbol{b}, y)$ from the Boolean parity problem, $(\boldsymbol{b} \odot |\boldsymbol{g}|, y)$ is a sample from the Gaussian parity problem where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ and $|\boldsymbol{g}|$ is the entry-wise absolute value of $\boldsymbol{g}$ and $\boldsymbol{b} \odot |\boldsymbol{g}|$ is the entry-wise product of $\boldsymbol{b}$ and $|\boldsymbol{g}|$.

[3] Though Blum, Kalai and Wasserman only state their result in the regime $\Lambda \asymp 1$, their algorithm works in the regime $\Lambda \geq 2^{-O(D^\delta)}$ for any $\delta < 1$, as stated in Lyubashevsky [28, Propsition 4].

Informally, the above corollary shows that for any $\alpha < 4/3$, there is no memory-bounded estimation algorithm which solves the parity problem with an effective sample size $N\Lambda^2 = \mathrm{poly}(D)$, a memory state of size $s = \mathrm{poly}(D)/\Lambda^\alpha$ after making $T = \mathrm{poly}(D)$ passes through the data set, provided the signal-to-noise ratio Λ is sufficiently small.

PROOF OF COROLLARY H.1. Let $k \in \mathbb{N}$ be a parameter to be determined. Consider an arbitrary memory bounded estimation algorithm for LPN with signal-to-noise ratio Λ which has a resource profile (N, T, s) where,

$$\Lambda^2 \asymp D^{-\gamma},\ N\Lambda^2 \asymp D^\eta,\ T \asymp D^\tau,\ s \asymp \frac{D^\mu}{\Lambda^\alpha},$$

for arbitrary constants $\eta \geq 1,\ \tau \geq 0,\ \mu \geq 0, \alpha < 4/3$. As a consequence of the reduction from k-CCA to k-LPN, we obtain using Theorem H.1 that, if we choose k odd such that

$$k \in \left(\eta + \tau + \mu + \frac{\alpha\gamma}{2}, \frac{2\gamma}{3}\right), \tag{H.8}$$

then

$$\lim_{D\to\infty} \inf_{\substack{S\subset[D]\\|S|=k}} \mathbb{P}_S\left(S = \hat{S}\right) = 0.$$

Under the assumptions on γ, α stated in the corollary, the interval in (H.8) is non-empty and has a width > 2. Hence, one can indeed find an odd $k \in \mathbb{N}$ which satisfies (H.8). Hence, the claim of the corollary follows. □

H.4.1. *Comparison to Prior Works.* A recent line of work initiated by Steinhardt, Valiant and Wager [34] and Raz [32] has obtained memory vs. sample-size lower bounds for *single-pass* memory-bounded estimation algorithms for learning parities:

1. Raz [32] showed that 1-pass ($T = 1$) memory-bounded estimation algorithms for learning noiseless ($\Lambda = 1$) parities require either a memory state of size $s \gtrsim D^2$ or an exponential sample size $N \geq 2^{\Omega(D)}$, proving a conjecture of Steinhardt, Valiant and Wager [34].
2. Garg et al. [19] studied the problem of learning noisy parities (i.e., $\Lambda \in (0, 1)$) using the techniques of Raz and showed that 1-pass ($T = 1$) memory-bounded estimation algorithms for learning noisy parities require either a memory state of size $s \gtrsim D^2/\Lambda$ or an exponential sample size $N \geq 2^{\Omega(D)}$.
3. Garg et al. conjectured that 1-pass ($T = 1$) memory-bounded estimation algorithms for learning noisy parities require either a memory state of size $s \gtrsim D^2/\Lambda^2$ or an exponential sample size $N \geq 2^{\Omega(D)}$. The information-theoretic sample complexity of learning noisy parities scales as $N \asymp D/\Lambda^2$. Hence, an interpretation of this conjecture is that any estimation algorithm which learns noisy parities with $N = \mathrm{poly}(D)$ sample complexity must have the capacity to memorize a dataset of size $N \asymp D/\Lambda^2$ (the information-theoretic sample complexity).

In comparison to the results discussed above, a key weakness of the lower bound in Corollary H.1 is that it requires the signal-to-noise ratio Λ to decay as a sufficiently large power of D. In contrast the results of Raz and Garg et al. can allow any $\Lambda \in (0, 1]$. This is a limitation of the proof approach which relies on the connection between estimation with limited memory and estimation with limited communication in a distributed setting (recall Fact 1). The techniques used by Raz and Garg et al. are very different and do not rely on this connection. On the other hand, an advantage of the lower bounds obtained using communication complexity is that they apply to multi-pass estimation algorithms whereas it seems challenging to extend the approach of Raz to the multi-pass setting. The work of Garg, Raz and Tal [18] is the current state-of-the-art result in this direction and shows that 2-pass ($T = 2$) memory bounded estimation algorithms for noiseless parity ($\Lambda = 1$) require a memory state of size $s \gtrsim D^{3/2}$ or a sample size of $N \geq 2^{\Omega(\sqrt{D})}$.

H.5. Proof of Computational Lower Bound (Theorem H.1). As with the other main theorems of this paper, we prove Theorem H.1 by transferring a communication lower bound for distributed estimation protocols for k-CCA to memory bounded estimators for the same problem using the reduction in Fact 1.

In the (Bayesian) distributed setup for k-CCA, the cross-moment tensor $\boldsymbol{V}$ is drawn from the prior

$$\pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{\sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \cdots \otimes \boldsymbol{e}_{i_k} : i_1, i_2, \dots, i_k \in [d]\}\right). \tag{H.9}$$

Here, $\boldsymbol{e}_i$ denotes the i-th standard basis vector in $\mathbb{R}^d$, so $\boldsymbol{V} \sim \pi$ is a uniformly random 1-sparse tensor. Then, $\boldsymbol{x}_{1:N}$ are sampled i.i.d. from the distribution $\mu_{\boldsymbol{V}}$ specified in (H.2) and then distributed across $m = N/n \in \mathbb{N}$ machines with n samples/machine; n will be suitably chosen to yield Theorem H.1. The execution of a distributed estimation protocol with parameters (m, n, b) results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

The following corollary is proved in exactly the same way as Corollary E.1.

COROLLARY H.2 (Fano's Inequality for k-CCA). *For any estimator $\hat{\boldsymbol{V}}(\boldsymbol{Y})$ for k-CCA computed by a distributed estimation protocol, and for any $t \in \mathbb{R}$, we have*

$$\inf_{\boldsymbol{V} \in \mathcal{V}} \mathbb{P}_{\boldsymbol{V}}\left(\frac{|\langle \boldsymbol{V}, \hat{\boldsymbol{V}}\rangle|^2}{\|\boldsymbol{V}\|^2 \|\hat{\boldsymbol{V}}\|^2} \geq \frac{t^2}{d^k}\right) \leq \frac{1}{t^2} + \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}.$$

The main technical result is the following information bound for k-CCA.

PROPOSITION H.1. *Consider the k-CCA problem with $\mu_{\boldsymbol{V}}$ as defined in (H.2). Let $\boldsymbol{Y} \in \{0,1\}^{mb}$ be the transcript generated by a distributed estimation protocol for this k-CCA problem with parameters (m, n, b). There is a finite constant C_k depending only on k, such that if*

$$n \geq C_k \cdot b \cdot d^{\frac{k}{2}} \quad \textit{and} \quad n\lambda^2 \leq \frac{1}{C_k},$$

then

$$\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y}) \leq C_k \cdot \left(\frac{b \cdot m \cdot n \cdot \lambda^2}{d^k} + m \cdot n^2 \cdot \lambda^4\right).$$

Proposition H.1 is proved in Appendix H.6. We can now complete the proof of Theorem H.1.

PROOF OF THEOREM H.1. Appealing to the reduction in Fact 1, we note that any memory-bounded estimator $\hat{\boldsymbol{V}}$ with resource profile (N, T, s) can be implemented using a distributed estimation protocol with parameters $(N/n, n, sT)$ for any $n \in \mathbb{N}$ such that $m := N/n \in \mathbb{N}$. As assumed in Theorem H.1, we consider the situation when:

$$\eta + \tau + \mu < k, \; \gamma > \frac{3k}{2}. \tag{H.10}$$

We set $n = d^{\xi}$ with

$$\xi \overset{\text{def}}{=} \tau + \mu + \frac{k}{2} + \frac{1}{2}\underbrace{(k - (\eta + \tau + \mu))}_{>0} > \tau + \mu + \frac{k}{2}. \tag{H.11}$$

With this choice, we verify that the information bound in Proposition H.1 shows that $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y}) \to 0$. This will yield the claim of the theorem. We begin by observing

$$\gamma > \frac{3k}{2} = \frac{k}{2} + \eta + \tau + \mu + (k - \eta - \tau - \mu) = \eta + \xi + \frac{(k - \eta - \tau - \mu)}{2} > \eta + \xi. \tag{H.12}$$

Next, we verify the conditions required for Proposition H.1:

1. Since $\eta > \tau + \mu + k/2$ (cf. (H.11)) we have $n \gg b \cdot d^{k/2}$ as required.
2. Since $\gamma > \eta + \xi > \xi$ (cf. (H.12)) we have $n\lambda^2 \ll 1$ as required.

Now, from the information bound of Proposition H.1,

$$\begin{aligned}\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y}) &\leq C_k \cdot \left(\frac{b \cdot m \cdot n \cdot \lambda^2}{d^k} + m \cdot n^2 \cdot \lambda^4\right). \\ &= C_k \cdot \left(b \cdot N\lambda^2 \cdot d^{-k} + (n\lambda^2) \cdot (N\lambda^2)\right).\end{aligned}$$

We now check that this bound on $\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})$ vanishes as $d \to \infty$:

1. The assumption $\eta + \tau + \mu < k$ (cf. (H.10)) guarantees $b \cdot N\lambda^2 \cdot d^{-k} \to 0$.
2. Since $\gamma > \eta + \xi$, we have $(N\lambda^2) \cdot (n\lambda^2) \to 0$.

This concludes the proof. □

REMARK H.1 (Connection with Correlation Detection). Observe that due to the choice of the prior in (H.9), the instance of k-CCA used to obtain the communication lower bound is an instance of the correlation detection problem. In this problem, the goal is to find a k-tuple of coordinates $(i_1, i_2, \dots, i_k) \subset [d]^k$ in the k vectors $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(k)})$ such that $x^{(1)}_{i_1}, x^{(2)}_{i_2}, \dots, x^{(k)}_{i_k}$ are k-wise correlated using N i.i.d. realizations of $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(k)})$. Communication lower bounds for this problem in the blackboard model (cf. Definition 2) were obtained in prior work by Dagan and Shamir [13]. This result is sufficient to obtain Theorem H.1. In this paper, we present another proof of this result using the information bound in Proposition 1, which is used to derive all communication lower bounds presented in this paper.

H.6. Proof of Information Bound (Proposition H.1).

We now present the proof of Proposition H.1, the information bound for the distributed k-CCA problem. Recall that in the distributed k-CCA problem:

1. The unknown rank-1 cross moment tensor $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \dots \otimes \boldsymbol{v}_k$ (the parameter of interest) is drawn from the prior π:
$$\boldsymbol{V} \sim \pi \overset{\text{def}}{=} \mathsf{Unif}\left(\{\sqrt{d^k} \cdot \boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \dots \otimes \boldsymbol{e}_{i_k} : i_{1:k} \in [d]\}\right).$$
2. A dataset consisting of $N = mn$ samples is drawn i.i.d. from $\mu_{\boldsymbol{V}}$, where $\mu_{\boldsymbol{V}}$ is the distribution of a single sample from the k-CCA problem. Recall that for $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(k)}) \in \mathbb{R}^{kd}$ and $\boldsymbol{V} = \sqrt{d^k} \cdot \boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \dots \otimes \boldsymbol{v}_k$, $\mu_{\boldsymbol{V}}$ was defined using its likelihood ratio with respect to the Gaussian measure $\mu_0 = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{kd})$:

(H.13a)
$$\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) \overset{\text{def}}{=} 1 + \frac{\lambda}{\lambda_k} \cdot \mathsf{sign}\left(\left\langle \boldsymbol{x}^{(1)}, \boldsymbol{v}_1 \right\rangle\right) \cdot \mathsf{sign}\left(\left\langle \boldsymbol{x}^{(2)}, \boldsymbol{v}_2 \right\rangle\right) \cdots \mathsf{sign}\left(\left\langle \boldsymbol{x}^{(k)}, \boldsymbol{v}_k \right\rangle\right),$$

where,

$$\lambda_k \overset{\text{def}}{=} \left(\frac{2}{\pi}\right)^{\frac{k}{2}} = (\mathbb{E}|Z|)^{\frac{k}{2}}, \; Z \sim \mathcal{N}(0,1)\,. \tag{H.13b}$$

3. This dataset is divided among m machines with n samples per machine. We denote the dataset in one machine by $\boldsymbol{X}_i \in \mathbb{R}^{d\times n}$, where,

$$\boldsymbol{X}_i = \left[\boldsymbol{x}_{i1}\, \boldsymbol{x}_{i2} \ldots \boldsymbol{x}_{in}\right],$$

   with $\boldsymbol{x}_{ij} \overset{\text{i.i.d.}}{\sim} \mu_{\boldsymbol{V}}$.
4. The execution of a distributed estimation protocol with parameters (m,n,b) results in a transcript $\boldsymbol{Y} \in \{0,1\}^{mb}$ written on the blackboard.

The information bound stated in Proposition H.1 is obtained using the general information bound given in Proposition 1 with the following choices:

Choice of μ_0 *and* $\overline{\mu}$*:* We set $\mu_0 = \overline{\mu} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{kd})$. That is, under $\mu_0 = \overline{\mu}$, $\boldsymbol{x}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ for any $i \in [m]$, $j \in [n]$.
Choice of $\mathcal{Z}$*:* We choose the event $\mathcal{Z}$ as the unrestricted sample space $\mathcal{Z} = \mathbb{R}^{kd\times n}$. Since $\mu_0 = \overline{\mu}$, this choice of $\mathcal{Z}$ satisfies the requirements of Proposition 1.

With these choices, appealing to the information bound provided in Proposition 1, we obtain:

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}{K} \le \sum_{i=1}^{m} \mathbb{E}_0\left[\int \left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1\right)\middle|\boldsymbol{Y}, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right]. \tag{H.14}$$

Hence, we need to analyze:

$$\mathbb{E}_0\left[\int \left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1\right)\middle|\boldsymbol{Y}, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right],$$

For any $\boldsymbol{X} \in \mathbb{R}^{kd\times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1\, \boldsymbol{x}_2\, \cdots\, \boldsymbol{x}_n]$, $S \subset [n]$, we introduce the notation,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S) \overset{\text{def}}{=} \prod_{i\in S}\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1\right).$$

In the special case when $S = \{i\}$, we will use the simplified notation $\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_i)$. We consider the following decomposition: For any $\boldsymbol{X} \in \mathbb{R}^{kd\times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1\, \boldsymbol{x}_2\, \cdots\, \boldsymbol{x}_n]$,

$$\begin{aligned}\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}) - 1 &= \prod_{\ell=1}^{n}\left(1 + \frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_\ell) - 1\right) - 1 \\ &= \underbrace{\sum_{\ell=1}^{n} \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_\ell)}_{\text{Additive Term}} + \underbrace{\sum_{S\subset[n],\,|S|\ge 2} \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S)}_{\text{Non Additive Term}}.\end{aligned}$$

With this decomposition, using the elementary inequality $(a+b)^2 \le 2a^2 + 2b^2$, we obtain,

$$\mathbb{E}_0\left[\int \left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i) - 1\right)\middle|\boldsymbol{Y}, (\boldsymbol{X}_j)_{j\neq i}\right]\right)^2 \pi(\mathrm{d}\boldsymbol{V})\right] \le 2\cdot(\mathsf{I}) + 2\cdot(\mathsf{II}), \tag{H.15}$$

where,

$$\mathsf{I} \overset{\text{def}}{=} \mathbb{E}_0 \left[\int \left(\mathbb{E}_0 \left[\sum_{\ell=1}^{n} \mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x}_{i\ell}) \middle| \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right],$$

$$\mathsf{II} \overset{\text{def}}{=} \mathbb{E}_0 \left[\int \left(\mathbb{E}_0 \left[\sum_{S \subset [n],\, |S| \geq 2} \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S) \middle| \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right)^2 \pi(\mathrm{d}\boldsymbol{V}) \right].$$

In order to control the term (II), we apply Jensen's Inequality:

$$\mathsf{II} \leq \int \mathbb{E}_0 \left[\left| \sum_{S \subset [n],\, |S| \geq 2} \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X}_i)_S) \right|^2 \right] \pi(\mathrm{d}\boldsymbol{V}).$$

The following lemma analyzes the above upper bound on (II).

LEMMA H.1. *Let $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \boldsymbol{x}_n]$ where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{kd})$. Suppose that $n\lambda^2/\lambda_k^2 \leq 1/2$. Then,*

$$\mathbb{E}_0 \left[\left| \sum_{S \subset [n],\, |S| \geq 2} \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S) \right|^2 \right] \leq 2 \left(\frac{n\lambda^2}{\lambda_k^2} \right)^2,$$

where λ_k is as defined in (H.13).

PROOF OF LEMMA H.1. We have

$$\mathbb{E}_0 \left[\left| \sum_{S \subset [n],\, |S| \geq 2} \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S) \right|^2 \right] = \sum_{\substack{S_1, S_2 \subset [n] \\ |S_1| \geq 2, |S_2| \geq 2}} \mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_1}) \cdot \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_2})]$$

Recall that,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{X}_S) \overset{\text{def}}{=} \prod_{i \in S} \left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1 \right).$$

We observe that, $\boldsymbol{x}_{1:n}$ are independent and,

$$\mathbb{E}_0 \left[\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}_i) - 1 \right] = 0.$$

Hence if $S_1 \neq S_2$, $\mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_1}) \cdot \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_{S_2})] = 0$. This gives us:

$$\mathbb{E}_0 \left[\left| \sum_{S \subset [n],\, |S| \geq 2} \mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S) \right|^2 \right] = \sum_{S \subset [n]\ |S| \geq 2} \mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)^2].$$

We can compute:

$$\begin{aligned} \mathbb{E}_0[\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)^2] &= \left(\mathbb{E}_0 \left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{x}) - 1 \right)^2 \right] \right)^{|S|}, \ \boldsymbol{x} \sim \mathcal{N}(0,1), \\ &\overset{\text{(a)}}{=} \left(\frac{\lambda}{\lambda_k} \right)^{2|S|}. \end{aligned}$$

In step (a), we recalled the formula for the likelihood ratio from (H.13). Hence, we have

$$\begin{aligned}\mathbb{E}_0\left[\left|\sum_{S\subset[n],\,|S|\geq 2}\mathscr{L}_{\boldsymbol{V}}((\boldsymbol{X})_S)\right|^2\right] &\leq \sum_{S\subset[n]\;|S|\geq 2}\left(\frac{\lambda^2}{\lambda_k^2}\right)^{|S|}\\ &=\sum_{s=2}^{n}\binom{n}{s}\left(\frac{\lambda^2}{\lambda_k^2}\right)^{s}\\ &\leq\sum_{s=2}^{n}\left(\frac{n\lambda^2}{\lambda_k^2}\right)^{s}.\end{aligned}$$

The assumption $n\lambda^2/\lambda_k^2 \leq 1/2$ guarantees that the above sum is dominated by a Geometric series, which immediately yields the claim of the lemma. □

In order to control the term (I), we recall that when $\boldsymbol{V}\sim\pi$, we have

$$\boldsymbol{V}=\sqrt{d^k}\cdot\boldsymbol{e}_{j_1}\otimes\boldsymbol{e}_{j_2}\cdots\otimes\boldsymbol{e}_{j_k},\; j_{1:k}\overset{\text{i.i.d.}}{\sim}\mathsf{Unif}\left([d]\right).$$

Consequently, for any $\boldsymbol{x}=(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\dots,\boldsymbol{x}^{(k)})\in\mathbb{R}^{kd}$,

$$\mathscr{L}_{\boldsymbol{V}}(\boldsymbol{x})=\frac{\lambda}{\lambda_k}\cdot\mathsf{sign}(x_{j_1}^{(1)})\cdot\mathsf{sign}(x_{j_2}^{(2)})\cdot\dots\cdot\mathsf{sign}(x_{j_k}^{(k)}).$$

For each machine $i\in[m]$ we can define n i.i.d. tensors $\boldsymbol{T}_{i1},\boldsymbol{T}_{i2},\dots,\boldsymbol{T}_{in}$ as:

$$\boldsymbol{T}_{i\ell}\overset{\text{def}}{=}\mathsf{sign}(\boldsymbol{x}_{i\ell}^{(1)})\otimes\mathsf{sign}(\boldsymbol{x}_{i\ell}^{(2)})\cdots\otimes\mathsf{sign}(\boldsymbol{x}_{i\ell}^{(k)}),$$

where the $\mathsf{sign}(\cdot)$ operation is understood to act entry-wise on a vector $\boldsymbol{v}\in\mathbb{R}^d$ to produce another vector $\mathsf{sign}(\boldsymbol{v})\in\{\pm1\}^d$. With this notation in place, we observe that we can rewrite (I) as:

$$(\mathsf{I})=\frac{\lambda^2}{\lambda_k^2}\cdot\frac{1}{d^k}\cdot\mathbb{E}_0\left[\left\|\mathbb{E}_0\left[\sum_{\ell=1}^{n}\boldsymbol{T}_{i\ell}\middle|\boldsymbol{Y},(\boldsymbol{X}_j)_{j\neq i}\right]\right\|^2\right],$$

Linearizing $\|\cdot\|$ we obtain (c.f. Lemma 1):

$$\left\|\mathbb{E}_0\left[\sum_{\ell=1}^{n}\boldsymbol{T}_{i\ell}\middle|\boldsymbol{Y},(\boldsymbol{X}_j)_{j\neq i}\right]\right\|=\sup_{\substack{\boldsymbol{S}\in\bigotimes^k\mathbb{R}^d\\\|\boldsymbol{S}\|\leq1}}\left(\mathbb{E}_0\left[\sum_{\ell=1}^{n}\langle\boldsymbol{T}_{i\ell},\boldsymbol{S}\rangle\middle|\boldsymbol{Y},(\boldsymbol{X}_j)_{j\neq i}\right]\right).$$

We will apply the Geometric Inequality framework (Proposition 2) to control the above conditional expectation. In order to do so, we need to understand the concentration behavior of the random variable:

$$\sum_{\ell=1}^{n}\langle\boldsymbol{T}_{i\ell},\boldsymbol{S}\rangle\,.$$

This is the subject of the following lemma.

LEMMA H.2. *Let $\boldsymbol{T},\boldsymbol{T}_1,\dots,\boldsymbol{T}_n$ be i.i.d. random tensors distributed as:*

$$\boldsymbol{T}=\mathsf{sign}(\boldsymbol{x}^{(1)})\otimes\mathsf{sign}(\boldsymbol{x}^{(2)})\cdots\otimes\mathsf{sign}(\boldsymbol{x}^{(k)}),$$

where $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \dots, \boldsymbol{x}^{(k)}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{kd})$. *Then, we have, for any* $S \in \bigotimes^k \mathbb{R}^d$ *with* $\|\boldsymbol{S}\| \leq 1$ *and any* $\zeta \in \mathbb{R}$ *with* $|\zeta| \leq d^{-\frac{k}{2}}/2$,

$$\log \mathbb{E}_0 \exp\left(\zeta \sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle\right) \leq n\zeta^2.$$

Furthermore,

$$\left\| \sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle \right\|_4 \leq \sqrt{3^k n}.$$

where,

$$\left\| \sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle \right\|_4^4 \overset{def}{=} \mathbb{E}_0 \left[\left(\sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle \right)^4 \right]$$

PROOF. The first claim follows from Bernstein's Inequality (Fact F.1) by observing that $\langle \boldsymbol{T}_i, \boldsymbol{S} \rangle \leq \|\boldsymbol{T}_i\| \|\boldsymbol{S}\| = \sqrt{d^k}$ and that $\mathbb{E}_0 \langle \boldsymbol{T}_i, \boldsymbol{S} \rangle = 0$, $\mathbb{E}_0 \langle \boldsymbol{T}_i, \boldsymbol{S} \rangle^2 = 1$. In order to obtain the moment bound, we observe that:

$$\sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle,$$

is a polynomial of degree k in the kdn i.i.d. $\mathsf{Unif}(\{\pm 1\})$ random variables $({x_\ell}^{(j)})_i$ where $j \in [k]$, $\ell \in [n]$, $i \in [d]$. Hence by Boolean Hypercontractivity (see for e.g. O'Donnell [31, Theorem 9.21]) we have

$$\left\| \sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle \right\|_4^2 \leq 3^k \cdot \mathbb{E}_0 \left[\left(\sum_{\ell=1}^{n} \langle \boldsymbol{T}_\ell, \boldsymbol{S} \rangle \right)^2 \right] = 3^k n.$$

□

We can now use Geometric Inequalities (Proposition 2) to control:

$$\left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y} = \boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\| = \sup_{\substack{\boldsymbol{S} \in \bigotimes^k \mathbb{R}^d \\ \|\boldsymbol{S}\| \leq 1}} \left(\mathbb{E}_0 \left[\sum_{\ell=1}^{n} \langle \boldsymbol{T}_{i\ell}, \boldsymbol{S} \rangle \middle| \boldsymbol{Y} = \boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right).$$

We consider two cases depending upon whether $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}$ or $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}$, where,

$$\mathcal{R}^{(i)}_{\mathsf{rare}} \overset{\text{def}}{=} \left\{ \boldsymbol{y} \in \{0,1\}^{mb} : 0 < \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i}) \leq 4^{-b} \right\},$$

$$\mathcal{R}^{(i)}_{\mathsf{freq}} \overset{\text{def}}{=} \left\{ \boldsymbol{y} \in \{0,1\}^{mb} : \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i}) > 4^{-b} \right\}.$$

Case 1: $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}$. In this situation we apply the moment version of the Geometric Inequality (Proposition 2, item (1)) with $q = 4$. Using the moment estimate in Lemma H.2, we obtain,

$$\left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y} = \boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\| \leq \frac{\sqrt{3^k \cdot n}}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})^{\frac{1}{4}}}. \tag{H.16}$$

Case 2: $\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}$. In this situation we apply the m.g.f. version of the Geometric Inequality (Proposition 2, item (2)). Using the m.g.f. estimate in Lemma H.2, we obtain, for any $0 < \zeta \leq d^{-\frac{k}{2}}/2$,

$$\left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y} = \boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\| \leq n\zeta + \frac{1}{\zeta} \log \frac{1}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})},$$

We set:

$$\zeta^2 = \frac{1}{n} \cdot \log \frac{1}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})} \leq \frac{b \cdot \log(4)}{n}.$$

If,

$$n \geq 2 \log(4) \cdot b \cdot d^{\frac{k}{2}}, \tag{H.17}$$

then this choice is valid, i.e. $\zeta \leq d^{-\frac{k}{2}}/2$. Hence,

$$\left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y} = \boldsymbol{y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\|^2 \leq 4 \cdot n \cdot \log \frac{1}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})}. \tag{H.18}$$

With these estimates, we can control the term (I), which we decompose as follows:

$$\begin{aligned}
(\mathsf{I}) &= \frac{\lambda^2}{\lambda_k^2} \cdot \frac{1}{d^k} \cdot \mathbb{E}_0 \left[\left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\|^2 \right] \\
&= \frac{\lambda^2}{\lambda_k^2} \cdot \frac{1}{d^k} \cdot ((\mathsf{Ia}) + (\mathsf{Ib})), \\
(\mathsf{Ia}) &\overset{\text{def}}{=} \mathbb{E}_0 \left[\sum_{\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}} \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i}) \cdot \left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\|^2 \right], \\
(\mathsf{Ib}) &\overset{\text{def}}{=} \mathbb{E}_0 \left[\sum_{\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}} \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i}) \cdot \left\| \mathbb{E}_0 \left[\sum_{\ell=1}^{n} \boldsymbol{T}_{i\ell} \middle| \boldsymbol{Y}, (\boldsymbol{X}_j)_{j \neq i} \right] \right\|^2 \right].
\end{aligned}$$

In order to control (Ia), we rely on the estimate (H.16):

$$\begin{aligned}
(\mathsf{Ia}) &\leq 3^k \cdot n \cdot \mathbb{E}_0 \left[\sum_{\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{rare}}} \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})^{\frac{1}{2}} \right] \\
&\leq 3^k \cdot n \cdot 2^{-b} \cdot \mathbb{E}_0[|\mathcal{R}^{(i)}_{\mathsf{rare}}|].
\end{aligned}$$

Since we assume the communication protocol to be deterministic conditioned on $(\boldsymbol{X}_j)_{j \neq i}$, all but b bits of $\boldsymbol{Y}$ are fixed. Consequently, $|\mathcal{R}_{\mathsf{rare}}| \leq 2^b$. Hence,

$$(\mathsf{Ia}) \leq 3^k \cdot n.$$

In order to control (Ib), we rely on the estimate (H.18):

$$(\mathsf{Ib}) \leq 4n \cdot \mathbb{E}_0 \left[\sum_{\boldsymbol{y} \in \mathcal{R}^{(i)}_{\mathsf{freq}}} \mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i}) \cdot \log \frac{1}{\mathbb{P}_0(\boldsymbol{Y} = \boldsymbol{y} | (\boldsymbol{X}_j)_{j \neq i})} \right].$$

Since we assume the communication protocol to be deterministic conditioned on $(\boldsymbol{X}_j)_{j\neq i}$, all but b bits of $\boldsymbol{Y}$ are fixed. Hence conditioned on $(\boldsymbol{X}_j)_{j\neq i}$, the random vector $(\boldsymbol{Y})$ has a support size of at most 2^b. The maximum entropy distribution on a given set S is the uniform distribution, which attains an entropy of $\log|S|$. Hence,

$$\sum_{(\boldsymbol{y},z)\in\{0,1\}^{b+1}} \mathbb{P}_0(\boldsymbol{Y}=\boldsymbol{y}|(\boldsymbol{X}_j)_{j\neq i})\cdot\log\frac{1}{\mathbb{P}_0(\boldsymbol{Y}=\boldsymbol{y}|(\boldsymbol{X}_j)_{j\neq i})}\leq b\cdot\log(2)$$

This yields the estimate,

$$(\mathsf{Ib})\leq 4\log(2)\cdot b\cdot n.$$

Combining the estimates on the terms Ia,Ib we obtain, $(\mathsf{I})\leq C_k\cdot b\cdot n\cdot\lambda^2/d^k$, where C_k is a constant depending only on k. Substituting this estimate on (I) and the estimate on (II) obtained in Lemma H.1 in (H.15), we obtain,

$$\mathbb{E}_0\left[\int\left(\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mu_{\boldsymbol{V}}}{\mathrm{d}\mu_0}(\boldsymbol{X}_i)-1\right)\middle|\boldsymbol{Y},(\boldsymbol{X}_j)_{j\neq i}\right]\right)^2\pi(\mathrm{d}\boldsymbol{V})\right]\leq C_k\cdot\left(\frac{b\cdot n\cdot\lambda^2}{d^k}+n^2\cdot\lambda^4\right).$$

Plugging the above bound in (H.14) we obtain,

$$\frac{\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V};\boldsymbol{Y})}{K}\leq C_k\cdot\left(\frac{b\cdot m\cdot n\cdot\lambda^2}{d^k}+m\cdot n^2\cdot\lambda^4\right).$$

This is exactly the information bound claimed in Proposition H.1.

APPENDIX I: MISCELLANEOUS RESULTS

I.1. Additional Technical Facts and Lemmas.

FACT I.1 (Estimates on Partial Exponential Series [25]). We have, for any $\lambda\geq 0$ and for any $t\in\mathbb{N}_0$ such that $t+1\geq\lambda$, we have

$$\frac{\lambda^t}{t!}\leq\sum_{i=t}^{\infty}\frac{\lambda^i}{i!}\leq\frac{1}{1-\frac{\lambda}{t+1}}\cdot\frac{\lambda^t}{t!}$$

In particular if $t\geq(e^2\lambda)\vee\log(1/\epsilon)\vee 1$, by Stirling's approximation,

$$\sum_{i=t}^{\infty}\frac{\lambda^i}{i!}\leq\epsilon.$$

FACT I.2 (A Bound on Hermite Polynomials). For any $k\in\mathbb{N}_0$, we have

$$|H_k(z)|\leq(1+|z|)^k.$$

PROOF. H_k has the following Taylor series expansion around $z=0$ (see for e.g. [12, Section 2.4]):

$$H_k(z)=\sum_{i=0}^{k}\binom{k}{i}\cdot\frac{\sqrt{i!}}{\sqrt{k!}}\cdot H_i(0)\cdot z^i.$$

The values $H_i(0)$ are known explicitly (see for e.g. [12, Section 2.10]):

$$|H_k(z)|\leq\sum_{i=0}^{k}\binom{k}{i}\cdot|z|^i=(1+|z|)^k.$$

□

FACT I.3 ([27], Equation 12). Let $\boldsymbol{V} \sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$. Define,

$$\overline{V} = \frac{1}{d}\sum_{i=1}^{d} V_i.$$

We have, for any $t \in \mathbb{N}_0,\ t \leq d$,

$$\mathbb{E}\overline{V}^{2t} \geq \frac{(2t)!}{2^t d^{2t}} \cdot \binom{d}{t} \geq \left(\frac{2}{e^2} \cdot \frac{t}{d}\right)^t.$$

LEMMA I.1. *Let $\boldsymbol{V} \sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$. Define,*

$$\overline{V} = \frac{1}{d}\sum_{i=1}^{d} V_i$$

Then for any $t \in \mathbb{N}$,

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] \leq 4^t \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil},$$

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] \leq 2 \cdot 5^t \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}.$$

Furthermore if $t \leq 2(d-1)$ and $d \geq 3$,

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] \geq 5^{-t} \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil}/2,$$

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] \geq 5^{-t} \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}/2.$$

PROOF. Due to coordinate symmetry, degree, and parity considerations, we have

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] = \sup_{\substack{\ell \in \{0,1,2\dots,t\} \\ t+\ell \text{ is even}}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right],$$

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] = \sup_{\substack{\ell \in \{1,2\dots,t\} \\ t+\ell \text{ is even}}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right].$$

Hence, we focus on proving upper and lower bounds on:

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right].$$

We decompose $\overline{V}$ as

$$\overline{V} = \frac{S_1}{d} + \frac{S_2}{d},$$

where $S_1 = V_1 + V_2 + \cdots + V_\ell$, $S_2 = V_{\ell+1} + V_{\ell+2} + \cdots + V_d$. By the Binomial Theorem,

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \le i \le \ell} V_i\right] = \sum_{i=0}^{t} \binom{t}{i} \cdot \frac{\mathbb{E}S_2^i}{d^t} \cdot \mathbb{E}\left[S_1^{t-i} \prod_{1 \le i \le \ell} V_i\right].$$

Observing that when $t - i < \ell$, we have

$$\mathbb{E}\left[S_1^{t-i} \prod_{1 \le i \le \ell} V_i\right] = 0,$$

and thus

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \le i \le \ell} V_i\right] = \sum_{i=0}^{t-\ell} \binom{t}{i} \cdot \frac{\mathbb{E}S_2^i}{d^t} \cdot \mathbb{E}\left[S_1^{t-i} \prod_{1 \le i \le \ell} V_i\right].$$

We now prove an upper bound and lower bound on the above expression.

Upper Bound: Since S_2 is sub-Gaussian with variance proxy $d - \ell$, we have (see, e.g.,, 33, Lemma 1.4)

$$\mathbb{E}S_2^i \le 2^i \cdot i^{\frac{i}{2}} \cdot (d-\ell)^{\frac{i}{2}}.$$

By an analogous argument,

$$\mathbb{E}\left[S_1^{t-i} \prod_{1 \le i \le \ell} V_i\right] \le \mathbb{E}[|S_1|^{t-i}] \le 2^{t-i} \cdot (t-i)^{\frac{t-i}{2}} \cdot \ell^{\frac{t-i}{2}}.$$

Hence,

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \le i \le \ell} V_i\right] \le \frac{2^t}{d^t} \sum_{i=0}^{t-\ell} \binom{t}{i} \cdot (t-i)^{\frac{t-i}{2}} \cdot i^{\frac{i}{2}} \cdot (d-\ell)^{\frac{i}{2}} \cdot \ell^{\frac{t-i}{2}}$$

Using the AM-GM Inequality,

$$(t-i)^{t-i} i^i \le \left(\frac{(t-i)^2 + i^2}{t}\right)^t \le t^t.$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \le i \le \ell} V_i\right] &\le \left(\frac{4t}{d}\right)^{\frac{t}{2}} \cdot \sum_{i=0}^{t-\ell} \binom{t}{i} \cdot \frac{(d-\ell)^{\frac{i}{2}}}{d^{\frac{i}{2}}} \cdot \frac{\ell^{\frac{t-i}{2}}}{d^{\frac{t-i}{2}}} \\ &\le \left(\frac{4t}{d}\right)^{\frac{t}{2}} \cdot \left(\frac{\ell}{d}\right)^{\frac{\ell}{2}} \cdot \sum_{i=0}^{t-\ell} \binom{t}{i} \\ &\le \left(\frac{16t}{d}\right)^{\frac{t}{2}} \cdot \left(\frac{\ell}{d}\right)^{\frac{\ell}{2}}.\end{aligned}$$

Hence,

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d] : r_i = 1} V_i\right] = \sup_{\substack{\ell \in \{0,1,2\ldots,t\} \\ t+\ell \text{ is even}}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \le i \le \ell} V_i\right]$$

$$\leq \left(\frac{16t}{d}\right)^{\frac{t}{2}} \cdot \left(\sup_{\substack{\ell \in \{0,1,2\ldots,t\} \\ t+\ell \text{ is even}}} \left(\frac{\ell}{d}\right)^{\frac{\ell}{2}} \right).$$

If $\ell \leq t \leq d/e$, the function $(\ell/d)^{\ell}$ is decreasing, and hence,

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] \leq 4^t \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil}.$$

On the other hand, when $t \geq d/e$, the same upper bound holds since,

$$4^t \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil} \geq \frac{4^t e^{-\frac{t}{2}}}{\sqrt{te}} \geq 4^t e^{-t} \geq 1,$$

whereas, since $|\overline{V}| \leq 1$, we always have the trivial upper bound,

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] \leq 1.$$

With an analogous argument, we also obtain,

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] \leq 2 \cdot 5^t \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}.$$

Lower Bound: Recall that,

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right] = \sum_{i=0}^{t-\ell} \binom{t}{i} \cdot \frac{\mathbb{E} S_2^i}{d^t} \cdot \mathbb{E}\left[S_1^{t-i} \prod_{1 \leq i \leq \ell} V_i\right].$$

For proving the claim of the lemma, it will be sufficient to lower bound the above expression under the assumption $t + \ell$ is even and $\ell \in \{0, 1, 2\}$. We observe that each of the terms in the above sum is non-negative. This is because, $\mathbb{E} S_2^i = 0$ when i is odd and, by expanding S_1^{t-i} using the Multinomial Theorem, one sees that:

$$\mathbb{E}\left[S_1^{t-i} \prod_{1 \leq i \leq \ell} V_i\right] \geq 0.$$

Hence, retaining the term corresponding to $i = (t - \ell)$ we obtain,

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right] \geq \binom{t}{\ell} \cdot \frac{\mathbb{E} S_2^{t-\ell}}{d^t} \cdot \mathbb{E}\left[S_1^{\ell} \prod_{1 \leq i \leq \ell} V_i\right].$$

Expanding S_1^{ℓ} using the Multinomial Theorem and comparing the coefficient of $V_1 \cdot V_2 \cdots V_{\ell}$, we observe,

$$\mathbb{E}\left[S_1^{\ell} \prod_{1 \leq i \leq \ell} V_i\right] = 1.$$

Hence,

$$\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right] \geq \binom{t}{\ell} \cdot \frac{\mathbb{E}S_2^{t-\ell}}{d^t}.$$

Since $t+\ell$ is assumed to be even, so is $t-\ell$. Furthermore since we assume that $\ell \in \{0,1,2\}$ and $t \leq 2(d-1)$ we have $t-\ell \leq 2(d-\ell)$. Hence using by Fact I.3, we have

$$\mathbb{E}S_2^{t-\ell} \geq \binom{d-\ell}{\frac{t-\ell}{2}} \cdot \frac{(t-\ell)!}{2^{\frac{t-\ell}{2}}}.$$

This give us,

$$\begin{aligned}
\mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right] &\geq \frac{1}{2^{\frac{t-\ell}{2}}} \cdot \frac{t!}{\ell!} \cdot \binom{d-\ell}{\frac{t-\ell}{2}} \cdot \frac{1}{d^t} \\
&\overset{\text{(a)}}{\geq} \frac{1}{2^{\frac{t-\ell}{2}}} \cdot \frac{t^t e^{-t}}{\ell!} \cdot \left(\frac{2(d-\ell)}{t-\ell}\right)^{\frac{t-\ell}{2}} \cdot \frac{1}{d^t} \\
&\geq \frac{t^{\frac{t}{2}} e^{-t}}{\ell!} \cdot \left(1 - \frac{\ell}{d}\right)^{\frac{t}{2}} \cdot \frac{1}{d^{\frac{t+\ell}{2}}} \\
&\overset{\text{(b)}}{\geq} \frac{t^{\frac{t}{2}} e^{-t}}{2} \cdot 3^{-\frac{t}{2}} \cdot \frac{1}{d^{\frac{t+\ell}{2}}}.
\end{aligned}$$

In the step marked (a), we used the standard lower bounds for the Binomial coefficient $\binom{n}{k} \geq (n/k)^k$ and factorial $n! \geq n^n e^{-n}$. In the step marked (b), we used the fact that $\ell \leq 2$ and $d \geq 3$. Hence,

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] = \sup_{\substack{\ell \in \{0,1,2\ldots,t\} \\ t+\ell \text{ is even}}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{1 \leq i \leq \ell} V_i\right] \geq \sup_{\substack{\ell \in \{0,1,2\} \\ t+\ell \text{ is even}}} \frac{5^{-t}}{2} \cdot t^{\frac{t}{2}} \cdot d^{-\frac{t+\ell}{2}}.$$

Choosing $\ell = 0$ if t is even and $\ell = 1$ if t is odd gives us:

$$\sup_{\boldsymbol{r} \in \{0,1\}^d} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] \geq 5^{-t} \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t}{2} \rceil}/2.$$

Choosing $\ell = 2$ if t is even and $\ell = 1$ if t is odd gives us:

$$\sup_{\substack{\boldsymbol{r} \in \{0,1\}^d \\ \|\boldsymbol{r}\|_1 \geq 1}} \mathbb{E}\left[\overline{V}^t \cdot \prod_{i \in [d]: r_i = 1} V_i\right] \geq 5^{-t} \cdot t^{\frac{t}{2}} \cdot d^{-\lceil \frac{t+1}{2} \rceil}/2.$$

This concludes the proof of this lemma. □

I.2. Analysis on Gaussian Space. Consider the functional space $\mathcal{L}_2\left(\mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right)\right)$ defined as follows:

$$\mathcal{L}_2\left(\mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right)\right) \overset{\text{def}}{=} \left\{f \colon \mathbb{R}^d \to \mathbb{R} : \mathbb{E}_{\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)} f^2(\boldsymbol{Z}) < \infty\right\}.$$

The multivariate Hermite polynomials for a complete orthonormal basis for $\mathcal{L}_2\left(\mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right)\right)$. These are defined as follows: for any $\boldsymbol{c} \in \mathbb{N}_0^d$, define

$$H_{\boldsymbol{c}}(\boldsymbol{z}) \overset{\text{def}}{=} \prod_{i=1}^d H_{c_i}(z_i),$$

where, for any $k \in \mathbb{N}_0$, the H_k are the (probabilist's) orthonormal Hermite polynomials with the property

$$\mathbb{E}_{\mathcal{N}(0,1)} H_k(Z) H_l(Z) = \begin{cases} 0 & \text{if } k \neq l; \\ 1 & \text{if } k = l. \end{cases}$$

The orthonormality property is inherited by the multivariate Hermite polynomials:

$$\mathbb{E}_{\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} H_{\boldsymbol{c}}(\boldsymbol{Z}) H_{\boldsymbol{d}}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } \boldsymbol{c} \neq \boldsymbol{d}; \\ 1 & \text{if } \boldsymbol{c} = \boldsymbol{d}. \end{cases}$$

Since these polynomials form an orthonormal basis of $\mathcal{L}_2\left(\mathcal{N}\left(\mathbf{0},\boldsymbol{I}_d\right)\right)$ any $f \in \mathcal{L}_2\left(\mathcal{N}\left(\mathbf{0},\boldsymbol{I}_d\right)\right)$ admits an expansion of the form:

$$f(\boldsymbol{z}) = \sum_{\boldsymbol{c} \in (\mathbb{N} \cup \{0\})^d} \hat{f}(\boldsymbol{c}) H_{\boldsymbol{c}}(\boldsymbol{z}).$$

In the above display, $\hat{f}(\boldsymbol{c}) \in \mathbb{R}$ are the Hermite (or Fourier) coefficients of f. They satisfy the usual Parseval's relation:

$$\sum_{\boldsymbol{c} \in \mathbb{N}_0^d} \hat{f}^2(\boldsymbol{c}) = \mathbb{E}_{\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} f^2(\boldsymbol{Z}).$$

A particular desirable property of the univariate Hermite polynomials is the following: for any $\mu \in \mathbb{R}, k \in \mathbb{N}_0$ we have

$$\mathbb{E}_{\mathcal{N}(0,1)} H_k(\mu + Z) = \frac{\mu^k}{\sqrt{k!}}.$$

This implies the following property of multivariate Hermite polynomials which will be particularly useful for us.

FACT I.4. For any $\boldsymbol{\mu} \in \mathbb{R}^d$ and any $\boldsymbol{c} \in \mathbb{N}_0^d$, we have,

$$\mathbb{E}_{\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} H_k(\boldsymbol{\mu} + \boldsymbol{Z}) = \frac{\boldsymbol{\mu}^{\boldsymbol{c}}}{\sqrt{\boldsymbol{c}!}}.$$

In the above display, we are using the following notation:

$$\boldsymbol{\mu}^{\boldsymbol{c}} \overset{\text{def}}{=} \prod_{i=1}^{m} \mu_i^{c_i}, \quad \boldsymbol{c}! \overset{\text{def}}{=} \prod_{i=1}^{m} (c_i!). \tag{I.1}$$

FACT I.5. For any vector $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\| = 1$ we have,

$$H_i(\langle \boldsymbol{u}, \boldsymbol{x} \rangle) = \sum_{\substack{\boldsymbol{c} \in \mathbb{N}_0^d \\ \|\boldsymbol{c}\|_1 = i}} \frac{\boldsymbol{u}^{\boldsymbol{c}}}{\sqrt{\boldsymbol{c}!}} H_{\boldsymbol{c}}(\boldsymbol{x}),$$

for any $\boldsymbol{x} \in \mathbb{R}^d$. In the above display, the notations $\boldsymbol{u}^{\boldsymbol{c}}$ and $\boldsymbol{c}!$ are as defined in (I.1).

FACT I.6. Let Z, Z' be ρ-correlated standard Gaussian random variables:

$$\begin{bmatrix} Z \\ Z' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$

Then, for any $i, j \in \mathbb{N}_0$,

$$\mathbb{E} H_i(Z) H_j(Z') = \begin{cases} \rho^i & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases}$$

We will also rely on the Gaussian Hypercontractivity theorem which is usually attributed to Nelson [30]. Our reference for this result was the book of O'Donnell [31].

FACT I.7 (Gaussian Hypercontractivity [30]). Let $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Then, for any $q \geq 2$,

$$\mathbb{E}\left|\sum_{\boldsymbol{\alpha} \in \mathbb{N}_0^d} c_{\boldsymbol{\alpha}} H_{\boldsymbol{\alpha}}(\boldsymbol{Z})\right|^q \leq \left(\sum_{\boldsymbol{\alpha} \in W^d} (q-1)^{\|\boldsymbol{\alpha}\|_1} \cdot c_{\boldsymbol{\alpha}}^2\right)^{\frac{q}{2}}$$

The inequality is tight for $q = 2$.

I.3. Fano's Inequality for Hellinger Information. In this section, we provide a derivation of the Fano's Inequality for Hellinger Information quoted in Fact 2. This result is a minor modification of a result due to Chen, Guntuboyina and Zhang [11]. Although these authors derive a version of Fano's Inequality for Hellinger Information [11, Corollary 7, item (iii)], it has a slightly more complicated form than the claim of Fact 2. The simpler form stated in Fact 2 (which suffices for our results) is derived by combining Fano's Inequality for the Total Variation (TV) Information proved by Chen, Guntuboyina and Zhang [11, Corollary 7, item (ii)] with standard a comparison between Hellinger and total variation distances. Specifically, Chen, Guntuboyina and Zhang [11, Corollary 7, item (ii)] show that for any estimator $\hat{\boldsymbol{V}} : \{0,1\}^{mb} \rightarrow \widehat{\mathcal{V}}$, we have

$$\int_{\mathcal{V}} \mathbb{E}_{\boldsymbol{V}}[\ell(\boldsymbol{V}, \hat{\boldsymbol{V}}(\boldsymbol{Y}))]\, \pi(\mathrm{d}\boldsymbol{V}) \geq R_0(\pi) - \mathbf{I}_{\mathsf{TV}}(\boldsymbol{V}; \boldsymbol{Y}), \tag{I.2}$$

where $\mathbf{I}_{\mathsf{TV}}(\boldsymbol{V}; \boldsymbol{Y})$ denotes the Total Variation Information which is defined as:

$$\mathbf{I}_{\mathsf{TV}}(\boldsymbol{V}; \boldsymbol{Y}) \overset{\text{def}}{=} \inf_{\mathbb{Q}} \int d_{\mathsf{TV}}(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q})\, \pi(\mathrm{d}\boldsymbol{V}).$$

In the above display, $d_{\mathsf{TV}}(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q})$ denotes the total variation distance between the probability measures $\mathbb{P}_{\boldsymbol{V}}$ and $\mathbb{Q}$. Since $d_{\mathsf{TV}}(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q}) \leq (2d_{\mathsf{hel}}^2(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q}))^{1/2}$ (see for e.g., [35, Lemma 2.3]) the Total Variation Information can be bounded in terms of the Hellinger Information:

$$\begin{aligned}\mathbf{I}_{\mathsf{TV}}(\boldsymbol{V}; \boldsymbol{Y}) \leq \inf_{\mathbb{Q}} \int (2d_{\mathsf{hel}}^2(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q}))^{1/2} \pi(\mathrm{d}\boldsymbol{V}) &\overset{\text{(a)}}{\leq} \left(2 \inf_{\mathbb{Q}} \int d_{\mathsf{hel}}^2(\mathbb{P}_{\boldsymbol{V}}, \mathbb{Q})\, \pi(\mathrm{d}\boldsymbol{V})\right)^{1/2} \\ &\overset{\text{(b)}}{=} \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}.\end{aligned}$$

In the above display step (a) follows from Jensen's Inequality and step (b) follows from the definition of Hellinger Information. Substituting the bound $\mathbf{I}_{\mathsf{TV}}(\boldsymbol{V}; \boldsymbol{Y}) \leq \sqrt{2\mathbf{I}_{\mathsf{hel}}(\boldsymbol{V}; \boldsymbol{Y})}$ in (I.2) immediately yields the claim of Fact 2.

REFERENCES

[1] ACHARYA, J., CANONNE, C. L., SUN, Z. and TYAGI, H. (2020). Unified lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*.

[2] ANANDKUMAR, A., DENG, Y., GE, R. and MOBAHI, H. (2017). Homotopy Analysis for Tensor PCA. In *Conference on Learning Theory* 79–104.

[3] BAR-YOSSEF, Z., JAYRAM, T. S., KUMAR, R. and SIVAKUMAR, D. (2004). An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences* **68** 702–732.

[4] BERLEKAMP, E., MCELIECE, R. and VAN TILBORG, H. (1978). On the inherent intractability of certain coding problems. *IEEE Transactions on Information Theory* **24** 384–386.

[5] BIROLI, G., CAMMAROTA, C. and RICCI-TERSENGHI, F. (2019). How to iron out rough landscapes and get optimal performances: Replicated Gradient Descent and its application to tensor PCA. *arXiv preprint arXiv:1905.12294*.

[6] Blum, A., Kalai, A. and Wasserman, H. (2003). Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)* **50** 506–519.

[7] Blum, A., Furst, M., Kearns, M. and Lipton, R. J. (1993). Cryptographic primitives based on hard learning problems. In *Annual International Cryptology Conference* 278–291. Springer.

[8] Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y. and Rudich, S. (1994). Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing* 253–262.

[9] Bonan, S. S. and Clark, D. S. (1990). Estimates of the Hermite and the Freud polynomials. *Journal of Approximation Theory* **63** 210–224.

[10] Braverman, M., Garg, A., Ma, T., Nguyen, H. L. and Woodruff, D. P. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* 1011–1020.

[11] Chen, X., Guntuboyina, A. and Zhang, Y. (2016). On Bayes risk lower bounds. *The Journal of Machine Learning Research* **17** 7687–7744.

[12] Wikipedia contributors (2021). Hermite polynomials — Wikipedia, The Free Encyclopedia. [Online; accessed 14-May-2021].

[13] Dagan, Y. and Shamir, O. (2018). Detecting correlations with little memory and communication. In *Conference On Learning Theory* 1145–1198. PMLR.

[14] Diakonikolas, I., Kane, D. M. and Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* 73–84. IEEE.

[15] Diakonikolas, I., Kane, D. M., Pittas, T. and Zarifis, N. (2021). The Optimality of Polynomial Regression for Agnostic Learning under Gaussian Marginals in the SQ Model. In *Conference on Learning Theory* 1552–1584. PMLR.

[16] Dudeja, R. and Hsu, D. (2022). Statistical-Computational Trade-offs in Tensor PCA and Related Problems via Communication Complexity. *arXiv preprint arXiv:2204.07526*.

[17] Feldman, V., Gopalan, P., Khot, S. and Ponnuswami, A. K. (2009). On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing* **39** 606–645.

[18] Garg, S., Raz, R. and Tal, A. (2019). Time-space lower bounds for two-pass learning. In *34th Computational Complexity Conference (CCC)*.

[19] Garg, S., Kothari, P. K., Liu, P. and Raz, R. (2021). Memory-sample lower bounds for learning parity with noise. In *24th International Conference on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2021 and 25th International Conference on Randomization and Computation, RANDOM 2021* 60. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

[20] Han, Y., Özgür, A. and Weissman, T. (2018). Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*.

[21] Hopkins, S. B., Schramm, T., Shi, J. and Steurer, D. (2016). Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* 178–191. ACM.

[22] Jayram, T. (2009). Hellinger strikes back: A note on the multi-party information complexity of AND. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* 562–573. Springer.

[23] Kearns, M. (1998). Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* **45** 983–1006.

[24] Kiltz, E., Pietrzak, K., Venturi, D., Cash, D. and Jain, A. (2017). Efficient authentication from hard learning problems. *Journal of Cryptology* **30** 1238–1275.

[25] Klar, B. (2000). Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences* **14** 161–171.

[26] Klivans, A. and Kothari, P. (2014). Embedding hard learning problems into Gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*.

[27] Kunisky, D., Wein, A. S. and Bandeira, A. S. (2019). Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*.

[28] Lyubashevsky, V. (2005). The parity problem in the presence of noise, decoding random linear codes, and the subset sum problem. In *Approximation, randomization and combinatorial optimization. Algorithms and techniques* 378–389. Springer.

[29] Mondelli, M. and Montanari, A. (2019). Fundamental Limits of Weak Recovery with Applications to Phase Retrieval. *Foundations of Computational Mathematics* **19**.

[30] NELSON, E. (1966). A quartic interaction in two dimensions. In *Mathematical Theory of Elementary Particles, Proc. Conf., Dedham, Mass., 1965* 69–73. MIT Press.

[31] O'DONNELL, R. (2014). *Analysis of boolean functions*. Cambridge University Press.

[32] RAZ, R. (2018). Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)* **66** 1–18.

[33] RIGOLLET, P. and HÜTTER, J.-C. (2015). High dimensional statistics. *Lecture notes for course 18S997* **813** 814.

[34] STEINHARDT, J., VALIANT, G. and WAGER, S. (2016). Memory, communication, and statistical queries. In *Conference on Learning Theory* 1490–1516. PMLR.

[35] TSYBAKOV, A. B. (2009). Introduction to Nonparametric Estimation.

[36] VALIANT, G. (2015). Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM* **62** 1–45.

[37] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. https://doi.org/10.1017/9781108627771

[38] WU, Y. and YANG, P. (2021). Polynomial methods in statistical inference: theory and practice. *arXiv preprint arXiv:2104.07317*.