The Complexity of Sparse Tensor PCA

Davin Choo[*]
Department of Computer Science
National University of Singapore
davin@u.nus.edu

Tommaso d'Orsi
Department of Computer Science
ETH Zürich
tommaso.dorsi@inf.ethz.ch

Abstract

We study the problem of sparse tensor principal component analysis: given a tensor $\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p}$ with $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ having i.i.d. Gaussian entries, the goal is to recover the k-sparse unit vector $x \in \mathbb{R}^n$. The model captures both sparse PCA (in its Wigner form) and tensor PCA. For the highly sparse regime of $k \leq \sqrt{n}$, we present a family of algorithms that smoothly interpolates between a simple polynomial-time algorithm and the exponential-time exhaustive search algorithm. For any $1 \leq t \leq k$, our algorithms recovers the sparse vector for signal-to-noise ratio $\lambda \geq \widetilde{\mathcal{O}}(\sqrt{t} \cdot (k/t)^{p/2})$ in time $\widetilde{\mathcal{O}}(n^{p+t})$, capturing the state-of-the-art guarantees for the matrix settings (in both the polynomial-time and sub-exponential time regimes). Our results naturally extend to the case of r distinct k-sparse signals with disjoint supports, with guarantees that are independent of the number of spikes. Even in the restricted case of sparse PCA, known algorithms only recover the sparse vectors for $\lambda \geq \widetilde{\mathcal{O}}(k \cdot r)$ while our algorithms require $\lambda \geq \widetilde{\mathcal{O}}(k)$. Finally, by analyzing the low-degree likelihood ratio, we complement these algorithmic results with rigorous evidence illustrating the trade-offs between signal-to-noise ratio and running time. This lower bound captures the known lower bounds for both sparse PCA and tensor PCA. In this general model, we observe a more intricate three-way trade-off between the number of samples n, the sparsity k, and the tensor power p.

1 Introduction

Sparse tensor principal component analysis is a statistical primitive generalizing both sparse PCA[2] and tensor PCA[3]. We are given multi-linear measurements in the form of a tensor

$$\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p} \in \otimes^p \mathbb{R}^n \tag{SSTM}$$

for a Gaussian noise tensor $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ containing i.i.d. $N(0,1)$ entries[4] and signal-to-noise ratio $\lambda > 0$. Our goal is to estimate the "structured" unit vector $x \in \mathbb{R}^n$. The structure we enforce on x is sparsity: $|\text{supp}(x)| \leq k$. The model can be extended to include multiple spikes in a natural way: $\boldsymbol{Y} = \boldsymbol{W} + \sum_{q=1}^{r} \lambda_q x_{(q)}^{\otimes p}$, and even general order-p tensors: $\boldsymbol{Y} = \boldsymbol{W} + \sum_{q=1}^{r} \lambda_q \mathcal{X}_{(q)}$ for $\mathcal{X}_{(q)} = x_{(q,1)} \otimes \cdots \otimes x_{(q,p)} \in \otimes^p \mathbb{R}^n$. In this introduction, we focus on the simplest single spike setting of SSTM.

It is easy to see that sparse PCA corresponds to the setting with tensor order $p = 2$. On the other hand, tensor PCA is captured by effectively removing the sparsity constraint: $|\text{supp}(x)| \leq n$. In

[*] Part of the work was done while the author was in ETH Zürich.

[2] Often in the literature, the terms sparse PCA and spiked covariance model refer to the sparse spiked *Wishart* model. However, here we consider the sparse spiked *Wigner* matrix model.

[3] Tensor PCA is also known as the spiked *Wigner* tensor model, or simply the spiked tensor model.

[4] Throughout the paper, we will write random variables in boldface.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

recent years, two parallel lines of work focused respectively on sparse PCA [JL09, AW08, BR13a, DM16, HKP$^+$17, DKWB19, HSV20, dKNS20] and tensor PCA [MR14, HSS15, MSS16, HKP$^+$17, KWB19, AMMN19], however no result captures both settings. The appeal of the *sparse spiked tensor model* (henceforth SSTM) is that it allows one to study the computational and statistical aspects of these other fundamental statistical primitives in a unified framework, understanding the computational phenomena at play from a more general perspective.

In this work, we investigate SSTM from both algorithmic and computational hardness perspectives. Our algorithm improves over known tensor algorithms whenever the signal vector is highly sparse. We also present a lower bound against low-degree polynomials which extends the known lower bounds for both sparse PCA and tensor PCA, leading to a more intricate understanding of how all three parameters (n, k and p) interact.

1.1 Related work

Disregarding computational efficiency, it is easy to see that optimal statistical guarantees can be achieved with a simple exhaustive search (corresponding to the maximum likelihood estimator): find a k-sparse unit vector maximizing $\langle \boldsymbol{Y}, x^{\otimes p}\rangle$. This algorithm returns a k-sparse unit vector $\widehat{x}$ achieving constant squared correlation[5] with the signal x as soon as $\lambda \gtrsim \sqrt{k \cdot \log(np/k)}$. That is, whenever $\lambda \gtrsim \max_{\|x\|=1, \|x\|_0=k} \langle \boldsymbol{W}, x^{\otimes p}\rangle$. Unfortunately, this approach runs in time exponential in k and takes super-polynomial time when $p \lesssim k$.[6] *As such, we assume $p \leq k$ from now on.*

Taking into account computational aspects, the picture changes. A good starting point to draw intuition for SSTM is the literature on sparse PCA and tensor PCA. We briefly outline some known results here. To simplify the discussion, we hide absolute constant multiplicative factors using $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\lesssim$, and $\gtrsim$, and hide multiplicative factors logarithmic in n using $\widetilde{\mathcal{O}}(\cdot)$.

1.1.1 Sparse PCA (Wigner noise)

Sparse PCA with Wigner noise exhibits a sharp phase transition in the top eigenvalue of $\boldsymbol{Y}$ for $\lambda \geq \sqrt{n}$ [FP07]. In this strong signal regime, the top eigenvector[7] v of $\boldsymbol{Y}$ correlates[8] with x with high probability, thus the following spectral method achieves the same guarantees as the exhaustive search suggested above: compute a leading eigenvector of $\boldsymbol{Y}$ and restrict it to the top k largest entries in absolute value. Conversely, when $\lambda < \sqrt{n}$, the top eigenvector of $\boldsymbol{Y}$ does not correlate with the signal x. In this weak signal regime, [JL09] proposed a simple algorithm known as diagonal thresholding: compute the top eigenvector of the principal submatrix defined by the k largest diagonal entries of $\boldsymbol{Y}$. This algorithm recovers the sparse direction when $\lambda \gtrsim \widetilde{\mathcal{O}}(k)$, thus requiring almost an additional $\sqrt{k}$ factor when compared to inefficient algorithms. More refined polynomial-time algorithms (low-degree polynomials [dKNS20], covariance thresholding [DM16] and the basic SDP relaxation [dGJL07, dKNS20]) only improve over diagonal thresholding by a logarithmic factor in the regime $n^{1-o(1)} \lesssim k^2 \lesssim n$. Interestingly, multiple results suggest that this *information-computation gap* is inherent to the sparse PCA problem [BR13a, BR13b, DKWB19, dKNS20]. Subexponential time algorithms and lower bounds have also been shown. For instance, [DKWB19, HSV20] presented smooth trade-offs between signal strength and running time.[9]

[5] One could also aim to find a unit vector with correlation approaching one or, in the restricted setting of $x \in \{0, \pm 1/\sqrt{k}\}$, aim to recover the support of x. At the coarseness of our discussion here, these goals could be considered mostly equivalent.

[6] Note that the problem input is of size n^p. So when $p \gtrsim k$, exhaustive search takes $n^{\mathcal{O}(p)}$ time which is polynomial in n^p. Thus, the interesting parameter regimes occur when $p \lesssim k$.

[7] By "top eigenvector" or "leading eigenvector", we mean the eigenvector corresponding to "largest (in absolute value) eigenvalue".

[8] More precisely, the vector consisting of the k largest (in absolute value) entries of v.

[9] Both works studied the single spike matrix setting. [HSV20] only considers the Wishart noise model and thus its guarantees cannot be compared to ours. [DKWB19] studied both the Wishart and Wigner noise models. In the Wishart noise model setting, both [HSV20] and [DKWB19] observe the same tradeoff between running time and signal-to-noise ratio. In the Wigner noise model setting, our algorithm and the algorithm of [DKWB19] offer the same smooth-trade off between running time and signal strength, up to universal constants.

1.1.2 Tensor PCA

In tensor settings, computing $\max_{\|x\|=1}\langle \boldsymbol{Y}, x^{\otimes p}\rangle$ is NP-hard already for $p = 3$ [HL13]. For even tensor powers p, one can unfold the tensor $\boldsymbol{Y}$ into a $n^{p/2}$-by-$n^{p/2}$ matrix and solve for the top eigenvector [MR14]. However, this approach is sub-optimal for odd tensor powers. For general tensor powers p, a successful strategy to tackle tensor PCA has been the use of semidefinite programming [HSS15, BGL16, HSS19]. Spectral algorithms inspired by the insight of these convex relaxations have also been successfully applied to the problem [SS17]. These methods succeed in recovering the single-spike x when $\lambda \gtrsim \widetilde{O}\left(n^{p/4}\right)$, thus exhibiting a large gap when compared to exhaustive search algorithms. Matching lower bounds have been shown for constant degrees in the Sum-of-Squares hierarchy [BGL16, HKP+17] and through average case reductions [BB20].

1.1.3 Sparsity-exploiting algorithms and tensor algorithms

It is natural to ask how do the characteristics of sparse PCA and tensor PCA extend to the more general setting of SSTM. In particular, there are two main observations to be made.

The first observation concerns the sharp computational transition that we see for $k \lesssim \sqrt{n}$ in sparse PCA. In these highly sparse settings, the top eigenvector of $\boldsymbol{Y}$ does not correlate with the signal x and so algorithms primarily based on spectral methods fail to recover it. Indeed, the best known guarantees are achieved through algorithms that crucially exploit the sparsity of the hidden signal. These algorithms require the signal strength to satisfy $\lambda \geq \widetilde{O}(\sqrt{k})$, with only logarithmic dependency on the ambient dimension. To exemplify this to an extreme, notice how the following algorithm can recover the support of xx^{T} with the same guarantees as diagonal thresholding, essentially disregarding the matrix structure of the data: zero all but the k^2 largest (in absolute value) entries of $\boldsymbol{Y}$. A natural question to ask is whether a similar phenomenon may happen for higher order tensors. In the highly sparse settings where $k \lesssim \sqrt{n}$, *can we obtain better algorithms exploiting the sparsity of the hidden vector?* Recently, a partial answer appeared in [LZ20] with a polynomial time algorithm recovering the hidden signal for $\lambda \geq \widetilde{O}(p \cdot k^{p/2})$, albeit with suboptimal dependency on the tensor order p.

The second observation concerns the computational-statistical gap in the spiked tensor model. As p grows, the gap between efficient algorithms and exhaustive search widens with the polynomial time algorithms requiring signal strength $\lambda \geq \widetilde{O}\left(n^{p/4}\right)$ while exhaustive search succeeds when $\lambda \geq \widetilde{O}(\sqrt{n})$ [MR14]. The question here is: *how strong is the dependency on p for efficient algorithms in sparse signal settings?*

In this work, we investigate these questions in the high order tensors regime $p \in \omega(1)$. We present a family of algorithms with a smooth trade-off between running time and signal-to-noise ratio. Even restricting to polynomial-time settings, our algorithms improve over previous results. Furthermore, through the lens of low-degree polynomials, we provide rigorous evidence of an *exponential gap* in the tensor order p between algorithms and lower bounds.

Remark. *The planted sparse densest sub-hypergraph model [CPMB19, BCPS20, CPSB20] is closely related to SSTM. We discuss this model in Appendix E.*

1.2 Results

1.2.1 Single spike setting

Consider first the restricted, but representative, case where the planted signal is a (k, A)-sparse unit vector with k non-zero entries having magnitudes in the range $\left[\frac{1}{A\sqrt{k}}, \frac{A}{\sqrt{k}}\right]$ for some constant $A \geq 1$. We say that the signal is *flat* when $A = 1$ and *approximately flat* when $A \geq 1$.

Our first result is a limited brute force algorithm – informally, an algorithm that smoothly interpolates between some brute force approach and some "simple" polynomial time algorithm – that *exactly* recovers the signal support of the planted signal[10].

Theorem 1 (Algorithm for single spike sparse tensor PCA, Informal)**.** *Let $A \geq 1$ be a constant. Consider the observation tensor*

$$\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p}$$

[10] A similar algorithm was analyzed by [DKWB19] for the special case of $p = 2$ and $r = 1$.

where the additive noise tensor $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ contains i.i.d. $N(0,1)$ entries and the signal $x \in \mathbb{R}^n$ is a (k, A)-sparse unit vector with signal strength $\lambda > 0$. Let $1 \leq t \leq k$ be an integer. Suppose that

$$\lambda \gtrsim \sqrt{t \left(\frac{2A^2 k}{t}\right)^p \ln n}\,.$$

Then, there exists an algorithm that runs in $\mathcal{O}(pn^{p+t})$ time and, with probability 0.99, outputs the support of x.

Let's first consider Theorem 1 in its simplest setting where $A = 1$ and t is a fixed constant. For $k \lesssim \sqrt{n}$, the theorem succeed when $\lambda \geq \widetilde{\mathcal{O}}(k^{p/2})$, thus improving over the guarantees of known tensor PCA methods which require $\lambda \geq \widetilde{\mathcal{O}}(n^{p/4})$. In addition, since support recovery is *exact*, one can obtain a good estimate[11] of the planted signal by running any known tensor PCA algorithm on the subtensor corresponding to its support. Indeed, the resulting subtensor will be of significantly smaller dimension and the requirement needed on the signal strength by single-spike tensor PCA algorithms are weaker than the requirement we impose on λ (see Remark 9 for details). As a result, our algorithm recovers the guarantees of diagonal thresholding in the matrix ($p = 2$) setting. Our polynomial-time algorithm also improves over the result of [LZ20], which required $\lambda \gtrsim \sqrt{pk^p \log n}$, by removing the polynomial dependency of the tensor order p in the signal strength λ.[12]

Consider now the limited brute force parameter t. From the introductory exposition, we know that one can obtain a statistically optimal algorithm by performing a brute force search over the space of k-sparse flat vectors in $\mathbb{R}^n$. The *limited brute force* algorithm is a natural extension that takes into account computational constraints by searching over the smaller set of t-sparse flat vectors, for $1 \leq t \leq k$, to maximize $\langle \boldsymbol{Y}, u^{\otimes p} \rangle$. The parametric nature of the algorithm captures both the brute force search algorithm (when $t = k$) and the idea of diagonal thresholding (when $t = 1$ and $p = 2$). As long as $t \leq k$, using a larger t represents a direct trade-off between running time and the signal-to-noise ratio. Extending the result to approximately flat vectors, the dependency on A in the term $\left(2A^2\right)^p$ can be removed by increasing the computational budget to some value $t' \geq 2A^2 t$.

1.2.2 Multiple spikes

Theorem 2 (Algorithm for multi-spike sparse tensor PCA, Informal)**.** *Let $A \geq 1$ be a constant. Consider the observation tensor*

$$\boldsymbol{Y} = \boldsymbol{W} + \sum_{q=1}^{r} \lambda_q x_{(q)}^{\otimes p}$$

where the additive noise tensor $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ contains i.i.d. $N(0,1)$ entries and the signals $x_{(1)}, \ldots, x_{(r)} \in \mathbb{R}^n$ are (k, A)-sparse unit vectors with disjoint supports and corresponding signal strengths $\lambda_1 \geq \ldots \geq \lambda_r > 0$. Let $1 \leq t \leq k$ be an integer and $0 < \epsilon \leq 1/2$. Suppose that

$$\lambda_r \gtrsim \frac{1}{\epsilon} \cdot \sqrt{t \left(\frac{2A^2 k}{t}\right)^p \ln n} \quad \textit{and} \quad \lambda_r \gtrsim A^{2p} \cdot (2\epsilon)^{p-1} \cdot \lambda_1\,.$$

Then, there exists an algorithm that runs in $\mathcal{O}(rpn^{p+t})$ time and, with probability 0.99, outputs the individual signal supports of $x_{(\pi(1))}, \ldots, x_{(\pi(r))}$ for some unknown bijection $\pi : [r] \to [r]$.

Theorem 2 requires two assumptions on the signals: (1) signals have disjoint support; (2) there is a bounded signal strength gap of $\lambda_r \gtrsim A^{2p} \cdot (2\epsilon)^{p-1} \cdot \lambda_1$. In the context of sparse PCA, algorithms that recover multiple spikes (e.g. [JL09, DM16]) only require the sparse vectors to be orthogonal. However, their guarantees are of the form $\lambda_r \geq \widetilde{\mathcal{O}}\left(\left|\bigcup_{q \in [r]} \operatorname{supp}\left(x_{(q)}\right)\right|\right)$. That is, when the r signals have disjoint supports, they require the smallest signal to satisfy $\lambda_r \geq \widetilde{\mathcal{O}}\left(k \cdot r\right)$. In comparison, already for constant t, Theorem 2 successfully recovers the supports when $\lambda_r \geq \widetilde{\mathcal{O}}(k)$,

[11] Recovery is up to a global sign flip since $\langle u, v \rangle^p = \langle u, -v \rangle^p$ for even tensor powers p.

[12] The result of [LZ20] extends to the settings where $\boldsymbol{Y} = \boldsymbol{W} + \lambda \mathcal{X}$ for an approximately flat tensor $\mathcal{X} \in \otimes^p \mathbb{R}^n$. Both Theorem 1 and Theorem 2 can also be extended to these settings (see Appendix B.2).

thus removing the dependency on the number of signals and improving the bound by a $1/r$ factor[13]. Meanwhile, the bounded signal strength gap assumption is a common identifiability assumption (e.g. see [CMW13, DM16]). We remark that Theorem 2 provides a tradeoff between this signal strength gap assumption and the signal strengths: we can recover the supports with a smaller gap if the signal strengths are increased proportionally – increasing λ_r by a multiplicative factor α enables the algorithm to succeed with gap that is smaller by a multiplicative factor of $1/\alpha$. As an immediate consequence, we also obtain a tradeoff between gap assumption and running time: every time we double t (while ensuring $1 \leq t \leq k$), λ_r increases by a factor of $(1/\sqrt{2})^{p-1}$ and thus the algorithm can succeed with a smaller gap. Finally, as in the single spike case, the exact support recovery allow us to obtain good estimate of each signal by running known tensor PCA algorithms.

Remark We remark that these results can be extended to the general tensor settings

$$\boldsymbol{Y} = \boldsymbol{W} + \sum_{q=1}^{r} \lambda_q \mathcal{X}_{(q)}$$

where for $q \in [r]$, $\mathcal{X}_{(q)} = x_{(q,1)} \otimes \cdots \otimes x_{(q,p)} \in \otimes^p \mathbb{R}^n$ in a natural way. See Appendix B.2.

1.2.3 An exponential gap between lower bounds and algorithms

SSTM generalizes both sparse PCA and tensor PCA. Hence, a tight hardness result for the model is interesting as it may combine and generalize the known bounds for these special cases. Here, we give a lower bound for the restricted computational model captured by *low-degree polynomials*. Originally developed in the context of the sum of squares hierarchy, this computational model appears to accurately predict the current best-known guarantees for problems such as sparse PCA, tensor PCA, community detection, and planted clique (e.g. see [HS17, HKP+17, Hop18, BHK+19, DKWB19, KWB19, dKNS20]).

Theorem 3 (Lower bound for low-degree polynomials, Informal)**.** *Let $1 \leq D \leq 2n/p$ and ν be the distribution of $\boldsymbol{Z} \in \otimes^p \mathbb{R}^n$ with i.i.d. entries from $N(0,1)$. Then, there exists a distribution μ over tensors $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ of the form*

$$\boldsymbol{Y} = \boldsymbol{W} + \lambda \boldsymbol{x}^{\otimes p}$$

where $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ is a noise tensor with i.i.d. $N(0,1)$ entries, the marginal distribution of $\boldsymbol{x}$ is supported on vectors with entries $\left\{\pm 1/\sqrt{k}, 0\right\}^n$, and $\boldsymbol{x}$ and $\boldsymbol{W}$ are distributionally independent, such that whenever

$$\lambda \lesssim \frac{\sqrt{D}}{2^p} \min\left\{ \left(\frac{n}{pD}\right)^{p/4}, \left(\frac{k}{pD}\left(1 + \left|\ln\left(\frac{npD}{ek^2}\right)\right|\right)\right)^{p/2} \right\},$$

μ is indistinguishable[14] from ν with respect to all polynomials of degree at most D.

Theorem 3 states that for certain values of λ, low-degree polynomials cannot be used to distinguish between the distribution of $\boldsymbol{Y}$ and $\boldsymbol{W}$ as typical values of low-degree polynomials are the same (up to a vanishing difference) under both distributions. The theorem captures known results in both sparse and tensor PCA settings. When $p = 2$, our bound reduces to $\lambda \lesssim \min\left\{\sqrt{n}, \frac{k}{\sqrt{D}}\left(1 + \left|\ln\left(\frac{2nD}{ek^2}\right)\right|\right)\right\}$, matching known low-degree bounds of [DKWB19] in the sparse PCA setting. Meanwhile, in the tensor PCA settings ($p \geq 2$, $k = n$), Theorem 3 implies a bound of the form $\lambda \lesssim \sqrt{D}\left(\frac{n}{pD}\right)^{p/4}$, thus recovering the results of [KWB19].

[13] It is an intriguing question whether an improvement of $1/r$ can be achieved in the more general settings of orthogonal spikes. Our approach relies on the signals having disjoint support and we expect it to *not* be generalizable to orthogonal signals. This can be noticed in the simplest settings with brute-force parameter $t = 1$ and $p = 2$ where the criteria of Algorithm 3 for finding an entry of a signal vector is to look at the diagonal entries of the data matrix. In this case, the algorithm may be fooled since the largest diagonal entry can depend on more than one spike. Nevertheless, we are unaware of any fundamental barrier suggesting that such guarantees are computationally hard to achieve.

[14] In the sense that for any low-degree polynomial $p(\boldsymbol{Y})$ we have $\frac{\mathbb{E}_\mu p(\boldsymbol{Y}) - \mathbb{E}_\nu p(\boldsymbol{Y})}{\sqrt{\mathbb{V} p(\boldsymbol{Y})}} \in o(1)$. See Appendix A.4.2.

For constant power p and $k \lesssim \sqrt{n}$, our lower bound suggests that no estimator captured by polynomials of degree $D \lesssim t \log n$ can improve over our algorithmic guarantees by more than a logarithmic factor. However, for $p \in \omega(1)$, an exponential gap appears between the bounds of Theorem 3 and state-of-the-art algorithms (both in the sparse settings as well as in the dense settings).[15] As a concrete example, let us consider the setting where $p = n^{0.1} < k$. The polynomial time algorithm of Theorem 1 requires $\lambda \geq \tilde{\mathcal{O}}(k^{p/2})$ while according to Theorem 3 it may be enough to have $\lambda \geq \tilde{\mathcal{O}}\left(k/n^{0.1}\right)^{p/2}$. Similarly, for $k \gtrsim \sqrt{np}$, known tensor algorithms recovers the signal for $\lambda \geq \tilde{\mathcal{O}}(n^{p/4})$ while our lower bound only rules out algorithms for $\lambda \leq \tilde{\mathcal{O}}\left(n^{0.9 \cdot p/4}\right)$.

Surprisingly, for the distinguishing problem considered in Theorem 3, these bounds appear to be tight. For a wide range of parameters (in both the dense and sparse settings) there exists polynomial time algorithms that can distinguish the distributions ν and μ right at the threshold considered in Theorem 3 (see Appendix C). It remains a fascinating open question whether sharper recovering algorithms can be designed or stronger lower bounds are required.

Finally, we would like to highlight that this non-trivial dependency on p is a purely computational phenomenon as it does not appear in information-theoretic bounds (see Appendix D).

Remark Note that Theorem 3 is *not* in itself a lower bound for the recovery problem. However, any algorithm which obtains a good estimation of the signal vector x for signal strength $\lambda \geq \sqrt{k \log n}$ can be used to design a probabilistic algorithm which solve the distinguishing problem for signal strength $\mathcal{O}_p(\lambda)$. Let us elaborate. Consider an algorithm that given $\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p}$ outputs a vector $\hat{x}$ such that $|\langle \hat{x}, x\rangle| \geq 0.9$. With high probability, $\max_{|z|_2=1, |z|_0=k} |\langle \boldsymbol{W}, z^{\otimes p}\rangle| \leq \tilde{\mathcal{O}}(\sqrt{k})$ and thus $|\langle \boldsymbol{Y}, \hat{x}^{\otimes p}\rangle| \geq \lambda \cdot (0.9)^p - \tilde{\mathcal{O}}(\sqrt{k})$. Therefore, one can solve the distinguishing problem as follows: output "planted" if $|\langle \boldsymbol{Y}, \hat{x}^{\otimes p}\rangle| \gtrsim \sqrt{k \log n}$ and "null" otherwise.

1.3 Notation and outline of paper

We write random variables in boldface and the set $\{1, \ldots, n\}$ as $[n]$. We hide absolute constant multiplicative factors and multiplicative factors logarithmic in n using standard notations: $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\lesssim$, $\gtrsim$, and $\tilde{\mathcal{O}}(\cdot)$. We denote by $e_1, \ldots, e_n \in \mathbb{R}^n$ the standard basis vectors. For $x \in \mathbb{R}^n$, we use $\operatorname{supp}(x) \subseteq [n]$ to denote the set of support coordinates. We say that x is a (k, A)-*sparse vector* if $k \in [n]$, constant $A \geq 1$, $|\operatorname{supp}(x)| = k$, and $\frac{1}{A\sqrt{k}} \leq |x_\ell| \leq \frac{A}{\sqrt{k}}$ for $\ell \in \operatorname{supp}(x)$. When $A = 1$, we say that x is a k-*sparse flat vector* and may omit the parameter A. For general $A \geq 1$, we say that x is *approximately flat*. For an integer $t \geq 1$, we define $U_t = \left\{u \in \left\{-\frac{1}{\sqrt{t}}, 0, \frac{1}{\sqrt{t}}\right\}^n : |\operatorname{supp}(u)| = t\right\}$ as the set of t-sparse flat vectors. For a tensor $T \in \otimes^p \mathbb{R}^n$ and a vector $u \in \mathbb{R}^n$, their inner product is defined as $\langle T, u^{\otimes p}\rangle = \sum_{i_1, \ldots, i_p \in [n]} T_{i_1, \ldots, i_p} u_{i_1} \ldots u_{i_p}$.

The rest of the paper is organized as follows: In Section 2, we introduce the main ideas behind Theorem 1 and Theorem 2. In Section 3, we flesh out some concrete unresolved research questions. Appendix A contains preliminary notions. We formally prove Theorem 1 and Theorem 2 in Appendix B. The lower bound Theorem 3 is given in Appendix C. We present an information theoretic bound in Appendix D. Appendix E discusses the planted sparse densest sub-hypergraph model. Finally, Appendix F contains any deferred technical proofs required throughout the paper.

2 Recovering signal supports via limited brute force searches

We describe here the main ideas behind our limited brute force algorithm. We consider the model

Model 4 (Sparse spiked tensor model)**.** *For $A \geq 1, r \geq 1, k \leq n$ we observe a tensor of the form*

$$\boldsymbol{Y} = \boldsymbol{W} + \sum_{q=1}^{r} \lambda_q x_{(q)}^{\otimes p} \in \otimes^p \mathbb{R}^n$$

[15] In particular, in the sparse settings $k \leq \sqrt{np}$, the $p^{-p/2}$ factor could not be seen in the restricted case of sparse PCA (as this factor is a constant when $p = 2$).

where $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ is a noise tensor with i.i.d. $N(0,1)$ entries, $\lambda_1 \geq \ldots \geq \lambda_r > 0$ are the signal strengths, and $x_{(1)}, \ldots, x_{(r)}$ are k-sparse flat unit length signal vectors.

We first look at the simplest setting of a single flat signal (i.e. $A = 1$ and $r = 1$). Then, we explain how to extend the analysis to multiple flat signals. For a cleaner discussion, we assume here that all the non-zero entries of the sparse vector x and vectors in the set U_t have positive sign. Our techniques also easily extend to approximately flat vectors (where $A \geq 1$) and general signal tensors $x_{(1)} \otimes \cdots \otimes x_{(p)} \in \otimes^p \mathbb{R}^n$. We provide details for these extensions in Appendix B.

2.1 Single flat signal

Limited-brute force As already mentioned in the introduction, a brute force search over U_k for the vector maximizing $\langle \boldsymbol{Y}, u^{\otimes p} \rangle$ returns the signal vector x (up to a global sign flip) with high probability whenever $\lambda \gtrsim \sqrt{k \log n}$. This algorithm provides provably optimal guarantees but requires exponential time (see Appendix D for an information-theoretic lower bound). The idea of a *limited brute force search* is to search over a smaller set U_t ($1 \leq t \leq k$) instead, and use the maximizer $\boldsymbol{v}_*$ to determine the signal support $\operatorname{supp}(x)$. The hope is that for a sufficiently large signal-to-noise ratio, this t-sparse vector $\boldsymbol{v}_*$ will still be non-trivially correlated with the hidden vector x. Indeed as t grows, the requirement on λ decreases towards the information-theoretic bound, at the expense of increased running time.

As a concrete example, consider the matrix settings ($p = 2$). It is easy to generalize the classic diagonal thresholding algorithm ([JL09]) into a limited brute-force algorithm. Recall that diagonal thresholding identifies the support of x by picking the indices of the largest k diagonal entries of $\boldsymbol{Y}$. In other words, the algorithm simply computes $\langle \boldsymbol{Y}, e_i^{\otimes 2} \rangle$ for all $i \in [n]$ and returns the largest k indices. From this perspective, the algorithm can be naturally extended to $t > 1$ by computing the $\binom{k}{t}$ vectors $u \in U_t$ maximizing $\langle \boldsymbol{Y}, u^{\otimes 2} \rangle$ and reconstructing the signal from them. For $t = k$, the algorithm corresponds to exhaustive search.

With this intuition in mind, we now introduce our family of algorithms, heavily inspired[16] by [DKWB19]. We first apply a preprocessing step to obtain two independent copies of the data.

```
Algorithm 1 Preprocessing
  Input: $\boldsymbol{Y}$.
  Sample a Gaussian tensor $\boldsymbol{Z} \in \otimes^p \mathbb{R}^n$ where each entry is an i.i.d. standard Gaussian $N(0,1)$.
  Return two independent copies $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ of $\boldsymbol{Y}$ as follows:
      $\boldsymbol{Y}^{(1)} = \frac{1}{\sqrt{2}}(\boldsymbol{Y} + \boldsymbol{Z})$   and   $\boldsymbol{Y}^{(2)} = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - \boldsymbol{Z})$
```

Algorithm 1 effectively creates two independent copies of the observation tensor $\boldsymbol{Y}$. To handle the noise variance, the signal-to-noise ratio is only decreased by the constant factor $1/\sqrt{2}$. For simplicity, we will ignore this constant factor in the remainder of the section.

```
Algorithm 2 Single spike limited brute force
  Input: $k, t$ and $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ obtained from Algorithm 1.
  Compute $\boldsymbol{v}_* := \operatorname{argmax}_{u \in U_t} \langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$.
  Compute the vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ with entries $\boldsymbol{\alpha}_\ell := \langle \boldsymbol{Y}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle$ for every $\ell \in [n]$.
  Return the indices of the largest $k$ entries of $\boldsymbol{\alpha}$.
```

[16] The algorithm in [DKWB19] is a specialization of ours (with comparable guarantees) in the simplest setting of $p = 2$ and a single spike. However, looking at [DKWB19], it is a priori unclear how to generalize the result to the settings of our interest. This is especially true in the tensor settings ($p \geq 3$) with multiple spikes, where signals may interfere with each other.

The signal support recovery process outlined in Algorithm 2 has two phases. In the first phase, we search over U_t to obtain a vector $\boldsymbol{v}_*$ that is correlated with the signal x. In the second phase, we use $\boldsymbol{v}_*$ to identify $\operatorname{supp}(x)$. The correctness of the algorithm follows from these two claims:

(i) The t-sparse maximizer $\boldsymbol{v}_*$ shares a large fraction of its support coordinates with signal x.

(ii) The k largest entries of $\boldsymbol{\alpha}$ belong to the support $\operatorname{supp}(x)$ of signal x.

Crucial to our analysis is the following standard concentration bound on Gaussian tensors. We directly use Lemma 5 in our exposition here, and formally prove a more general form in Appendix F.1.

Lemma 5. *Let $p \leq n$, $t > 0$ be an integer, and $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ be a tensor with i.i.d. $N(0,1)$ entries. Then, with high probability, for any $u \in U_t$,*

$$\langle \boldsymbol{W}, u^{\otimes p} \rangle \lesssim \sqrt{t \log n} \,.$$

For some constant $0 < \epsilon \leq 1/2$, suppose that

$$\lambda \gtrsim \frac{1}{\epsilon \cdot (1-\epsilon)^{p-1}} \cdot \sqrt{t \left(\frac{k}{t}\right)^p \log n} \,. \tag{1}$$

For any $u \in U_t$ with support $\operatorname{supp}(u) \subseteq \operatorname{supp}(x)$, we have

$$\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle = \lambda \langle x, u \rangle^p + \langle \boldsymbol{W}^{(1)}, u^{\otimes p} \rangle \geq \lambda \cdot \left(\frac{t}{k}\right)^{\frac{p}{2}} - \mathcal{O}\left(\sqrt{t \log n}\right) \,.$$

On the other hand, any $u \in U_t$ with support satisfying $|\operatorname{supp}(u) \cap \operatorname{supp}(x)| \leq (1-\epsilon) \cdot t$ has small correlation with $\boldsymbol{Y}^{(1)}$ in the sense that

$$\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle = \lambda \langle x, u \rangle^p + \langle \boldsymbol{W}^{(1)}, u^{\otimes p} \rangle \leq \lambda \cdot (1-\epsilon)^p \cdot \left(\frac{t}{k}\right)^{\frac{p}{2}} + \mathcal{O}\left(\sqrt{t \log n}\right) \,.$$

By Eq. (1), with high probability, $\boldsymbol{v}_*$ will have at least a fraction $(1-\epsilon)$ of its support contained in $\operatorname{supp}(x)$, yielding the first claim. Observe that $\boldsymbol{v}_*$ does not completely overlap with x. A priori, this might seem to be an issue. However, it turns out that we can still use $\boldsymbol{v}_*$ to exactly reconstruct the support of x. Indeed, for all $\ell \in \operatorname{supp}(x)$,

$$\begin{aligned}
\boldsymbol{\alpha}_\ell &= \lambda \cdot x_\ell \cdot \langle x, \boldsymbol{v}_* \rangle^{p-1} + \langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle \\
&\geq \lambda \cdot \frac{(1-\epsilon)^{p-1}}{\sqrt{k}} \cdot \left(\frac{t}{k}\right)^{\frac{p-1}{2}} + \langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle \\
&\gtrsim \frac{1}{\epsilon} \cdot \sqrt{\log n} + \langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle \,.
\end{aligned}$$

Now, by independence of $\boldsymbol{W}^{(2)}$ and $\boldsymbol{v}_*$, $\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle$ behaves like a standard Gaussian. Thus, with high probability, $\left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle\right| \lesssim \sqrt{\log n}$ and $\boldsymbol{\alpha}_\ell \gtrsim \sqrt{\log n}$. Conversely, if ℓ is *not* in the support of the signal, then $\boldsymbol{\alpha}_\ell \lesssim \sqrt{\log n}$. So, the vector $\boldsymbol{\alpha}$ acts as indicator of the support of x!

Remark 6. In its simplest form of $t = 1$, Algorithm 2 does *not* exploit the tensor structure of the data: it performs entry-wise search for the largest (in magnitude) over a subset of $\boldsymbol{Y}$. However, this is no longer true as t grows. For $t = k$, the algorithm computes the k-sparse flat unit vector u maximizing $\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$.

2.2 Multiple flat signals with disjoint signal supports

Consider now the setting with $r > 1$ spikes. Recall that we assumed the vectors $x_{(1)}, \ldots, x_{(r)}$ to have non-intersecting supports. We also assumed that for any $q, q' \in [r]$ and some fixed scalar $0 \leq \kappa \leq 1$, if $\lambda_q \geq \lambda_{q'}$, then $\lambda_{q'} \geq \kappa \cdot \lambda_q$. We remark that we may *not* recover the signal supports in a known order, but we are guaranteed to recover *all of them* exactly. For simplicity of discussion, let us assume here that we recover the vector $x_{(i)}$ at iteration i.

The idea to recover the r spikes is essentially to run Algorithm 2 r times. At first, we compute the t-sparse vector $\boldsymbol{v}_*$ by maximizing the product $\langle \boldsymbol{Y}^{(1)}, \boldsymbol{v}_*^{\otimes p} \rangle$. Then, using $\boldsymbol{v}_*$, we compute the vector $\boldsymbol{\alpha}$ to obtain a set $\mathcal{I}_1 \subseteq [n]$. With high probability, we will have $\mathcal{I}_1 = \operatorname{supp}\left(x_{(1)}\right)$ and so we will exactly recover the support of $x_{(1)}$. In the second iteration of the loop, we repeat the same procedure with the additional constraint of searching only over the $n-k$ dimensional subset of U_t containing vectors with disjoint support from $\mathcal{I}_1$. Similarly, at iteration i, we search over the subset of U_t containing vectors with disjoint support from $\bigcup\limits_{1 \leq j < i} \mathcal{I}_j$. As before, we first preprocess the data to create two independent copies $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$. Concretely:

Algorithm 3 Multi-spike limited brute force

```
Input: $k, t, r$ and $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ obtained from Algorithm 1.
Repeat for $i = 1$ to $r$:
    Compute $\boldsymbol{v}_* := \operatorname{argmax}_{u \in U_t} \langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$ subject to $\operatorname{supp}(\boldsymbol{v}_*) \cap \left( \bigcup_{1 \leq j < i} \mathcal{I}_j \right) = \emptyset$.
    Compute the vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ with entries $\boldsymbol{\alpha}_\ell := \langle \boldsymbol{Y}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle$ for every $\ell \in [n]$.
    Let $\mathcal{I}_i$ be the set of indices of the largest $k$ entries of $\boldsymbol{\alpha}$.
Return $\mathcal{I}_1, \ldots, \mathcal{I}_r$.
```

The proof structure is similar to that of Algorithm 2 and essentially amounts to showing that the claims (i) and (ii) described in Section 2.1 hold in each iteration.

Let $\lambda_{\min} = \min_{q \in [r]} \lambda_q$ and $\lambda_{\max} = \max_{q \in [r]} \lambda_q$. For some $0 < \epsilon \leq 1/2$, let $\kappa \gtrsim \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}$ such that $\lambda_{\min} \geq \kappa \cdot \lambda_{\max}$. Suppose that

$$\lambda_{\min} \gtrsim \frac{1}{\epsilon \cdot (1-\epsilon)^p} \cdot \sqrt{t \left(\frac{k}{t}\right)^p \log n} \quad \text{and} \quad \lambda_{\min} \gtrsim \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1} \cdot \lambda_{\max} \,. \tag{2}$$

Consider an arbitrary iteration i and suppose that we exactly recovered the support of one signal in each of the previous iterations. Without loss of generality, assume that $\lambda_{\max}$ is the largest signal strength among the yet to be recovered signals, and let $x_{(\max)}$ be one such corresponding signal.

For $u \in U_t$ satisfying $\operatorname{supp}(u) \subseteq \operatorname{supp}(x_{(\max)})$, we have

$$\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle = \lambda_{\max} \langle x_{(\max)}, u \rangle^p + \langle \boldsymbol{W}^{(1)}, u^{\otimes p} \rangle \geq \lambda_{\max} \cdot \left(\frac{t}{k}\right)^{\frac{p}{2}} - \mathcal{O}\left(\sqrt{t \log n}\right) \,.$$

On the other hand, for any $u \in U_t$ such that $\left|\operatorname{supp}(u) \cap \operatorname{supp}\left(x_{(q)}\right)\right| \leq (1-\epsilon) \cdot t$ for all $q \in [r]$,

$$\begin{aligned}
\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle &= \sum_{q \in [r]} \lambda_q \langle x_{(q)}, u \rangle^p + \langle \boldsymbol{W}^{(1)}, u^{\otimes p} \rangle \\
&\leq \lambda_{\max} \cdot \left(\frac{t}{k}\right)^{\frac{p}{2}} \cdot \left((1-\epsilon)^p + \epsilon^p\right) + \mathcal{O}\left(\sqrt{t \log n}\right) \\
&\leq \lambda_{\max} \cdot \left(\frac{t}{k}\right)^{\frac{p}{2}} \cdot (1-\epsilon)^{p-1} + \mathcal{O}\left(\sqrt{t \log n}\right) \,.
\end{aligned}$$

Thus, as in Section 2.1, it follows that $\boldsymbol{v}_*$ satisfies $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(i)}\right)\right| \geq (1-\epsilon) \cdot t$ for some signal $x_{(i)}$. Note that $x_{(i)}$ may not be $x_{(\max)}$. Even though $\boldsymbol{v}_*$ does not exactly overlap with any of the signal vectors, we will not accumulate an error at each iteration. This is because, analogous to the single spike setting, we can exactly identify the support of a signal through $\boldsymbol{\alpha}$. For any

$\ell \in \operatorname{supp}\left(x_{(i)}\right)$, it holds that $\boldsymbol{\alpha}_\ell \gtrsim \sqrt{\log n}$ as before because $\left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle\right| \lesssim \sqrt{\log n}$. Conversely, since signal supports are disjoint, we see that for $\ell \notin \operatorname{supp}\left(x_{(i)}\right)$,

$$\begin{aligned}
\boldsymbol{\alpha}_\ell &= \sum_{q \in [r]} \lambda_q \cdot x_{(q),\ell} \cdot \langle x_{(q)}, \boldsymbol{v}_* \rangle^{p-1} + \langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle \\
&\leq \lambda_{\max} \cdot \frac{\epsilon^{p-1}}{\sqrt{k}} \cdot \left(\frac{t}{k}\right)^{\frac{p-1}{2}} + \mathcal{O}\left(\sqrt{\log n}\right) \\
&\leq \frac{\lambda_{\min}}{\kappa} \cdot \frac{\epsilon^{p-1}}{\sqrt{k}} \cdot \left(\frac{t}{k}\right)^{\frac{p-1}{2}} + \mathcal{O}\left(\sqrt{\log n}\right) \\
&\lesssim \sqrt{\log n}\,.
\end{aligned}$$

So, once again, $\boldsymbol{\alpha}$ exactly identifies the support of $x_{(i)}$ with high probability.

Remark 7 (On the strength of the assumption on κ)**.** As already briefly discussed in Section 1.2, the algorithm provides a three-way trade-off between signal gap κ, signal-to-noise ratio λ and running time. By appropriately choosing the constant $\epsilon > 0$, the algorithm can tolerate different values of κ. Indeed, the above analysis holds as long as $\kappa \gtrsim \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}$. This suggests two ways in which we can loosen the requirement $\lambda_{\min} \geq \kappa \cdot \lambda_{\max}$ and still successfully recover the signals through Algorithm 3. One is increase the running time, so that we can decrease ϵ without increasing the signal-to-noise ratio $\lambda_{\min}$. The other is to decrease ϵ and increase the value of $\lambda_{\min}$ accordingly.

Remark 8 (On independent copies of $\boldsymbol{Y}$)**.** To clarify why it suffices to have 2 independent copies of $\boldsymbol{Y}$ even for multiple iterations, observe that at each iteration i, the choice of the set $\mathcal{I}_i$ depends only on the vector $\boldsymbol{v}_*$ with high probability. Consider the following thought experiment where we are given a fresh copy $\boldsymbol{Y}^{(i)}$ of $\boldsymbol{Y}$ in the second phase of each iteration i of the algorithm (while still using only a single copy $\boldsymbol{Y}^{(1)}$for *all* the first phases). Even with fresh randomness, the result is the same as Algorithm 3 with high probability because at each iteration the choice of maximizer $\boldsymbol{v}_*$ causes the same output.

Remark 9 (Reconstructing the signals from their supports)**.** After recovering individual signal supports, one can reconstruct signals using known tensor PCA algorithms (e.g. [MR14, HSS15]) on the subtensor defined by each recovered support. The signal strength required for this new subproblem is weaker and is satisfied by our recovery assumptions. For instance, by concatenating our algorithm with [HSS15, Theorem 7.1], one obtains vectors $\widehat{x}_{(1)}, \ldots, \widehat{x}_{(r)}$ such that $\left|\langle \widehat{x}_{(i)}, x_{(i)} \rangle\right| \geq 0.99$, for any $i \in [r]$, with probability 0.99.

3 Open questions

Open question 1. *Theorem 2 improves over existing sparse PCA multi-spike recovery algorithms (which only assumed orthogonal spikes) by a factor of $1/r$ in the case where these planted signals have disjoint support. Can one still obtain an improvement of $1/r$ if we only assume orthogonality?*

Open question 2. *For $p \in \omega(1)$, there is an exponential gap between the bounds of Theorem 3 and state-of-the-art algorithms. A natural question is whether one can design better recover algorithms or prove stronger lower bounds for this range of tensor power $p \in \omega(1)$?*

Remark (Societal impact)**.** *This work does not present any foreseeable negative societal consequence.*

Acknowledgments and Disclosure of Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464). This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-08-013). We thank David Steurer for several helpful conversations. We thank Luca Corinzia and Paolo Penna for useful discussions about the planted sparse densest sub-hypergraph model.

References

[AMMN19] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019. 2

[AW08] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008. 2

[BB20] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, pages 648–847, 2020. 3

[BCPS20] Joachim M Buhmann, Luca Corinzia, Paolo Penna, and Wojciech Szpankowski. Recovery of a Planted k-Densest Sub-Hypergraph, 2020. Available at: https://www.cs.purdue.edu/homes/spa/papers/isit20-clique.pdf. 3, 27

[BGL16] Vijay Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee. Sum-of-squares certificates for maxima of random tensors on the sphere. *arXiv preprint arXiv:1605.00903*, 2016. 3

[BHK$^+$19] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019. 5, 16, 17

[BKW20] Afonso S. Bandeira, Dmitriy Kunisky, and Alexander S. Wein. Computational hardness of certifying bounds on constrained PCA problems. In *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, pages 78:1–78:29, 2020. 16

[BR13a] Quentin Berthet and Philippe Rigollet. Complexity Theoretic Lower Bounds for Sparse Principal Component Detection. In *Conference on Learning Theory*, pages 1046–1066, 2013. 2

[BR13b] Quentin Berthet and Philippe Rigollet. OPTIMAL DETECTION OF SPARSE PRINCIPAL COMPONENTS IN HIGH DIMENSION. *The Annals of Statistics*, 41(4):1780–1815, 2013. 2

[Cam60] Lucien Le Cam. Locally asymptotically normal families. *Univ. California Publ. Statist.*, 1960. 17

[CMW13] T Tony Cai, Zongming Ma, and Yihong Wu. SPARSE PCA: OPTIMAL RATES AND ADAPTIVE ESTIMATION. *The Annals of Statistics*, 41(6):3074–3110, 2013. 5

[CPMB19] Luca Corinzia, Paolo Penna, Luca Mondada, and Joachim M Buhmann. Exact Recovery for a Family of Community-Detection Generative Models. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 415–419. IEEE, 2019. 3, 27

[CPSB20] Luca Corinzia, Paolo Penna, Wojciech Szpankowski, and Joachim M. Buhmann. Statistical and computational thresholds for the planted k-densest sub-hypergraph problem, 2020. 3, 27

[dGJL07] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007. 2

[dKNS20] Tommaso d'Orsi, Pravesh K. Kothari, Gleb Novikov, and David Steurer. Sparse PCA: Algorithms, Adversarial Perturbations and Certificates, 2020. To appear in FOCS 2020. 2, 5, 18

[DKWB19] Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-Time Algorithms for Sparse PCA. *arXiv preprint arXiv:1907.11635*, 2019. 2, 3, 5, 7, 22

[DM16] Yash Deshpande and Andrea Montanari. Sparse PCA via Covariance Thresholding. *The Journal of Machine Learning Research*, 17(1):4913–4953, 2016. 2, 4, 5

[Duc16] John Duchi. Lecture Notes for Statistics 311/Electrical engineering 377, 2016. Available at: `https://stanford.edu/class/stats311/Lectures/full_notes.pdf`. 15, 16

[FP07] Delphine Féral and Sandrine Péché. The Largest Eigenvalue of Rank One Deformation of Large Wigner Matrices. *Communications in mathematical physics*, 272(1):185–228, 2007. 2

[HKP$^+$17] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017. 2, 3, 5, 16, 17, 22

[HL13] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013. 3

[Hop18] Samuel Brink Klevit Hopkins. *STATISTICAL INFERENCE AND THE SUM OF SQUARES METHOD*. PhD thesis, Cornell University, 2018. 5, 16, 17

[HS17] Samuel B Hopkins and David Steurer. Efficient Bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017. 5, 16, 17, 22

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-squares proofs. In *Conference on Learning Theory*, pages 956–1006, 2015. 2, 3, 10

[HSS19] Samuel B Hopkins, Tselil Schramm, and Jonathan Shi. A robust spectral algorithm for overcomplete tensor decomposition. In *Conference on Learning Theory*, pages 1683–1722, 2019. 3

[HSV20] Guy Holtzman, Adam Soffer, and Dan Vilenchik. A greedy anytime algorithm for sparse PCA. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, pages 1939–1956, 2020. 2

[HW20] Justin Holmgren and Alexander S Wein. Counterexamples to the low-degree conjecture. *arXiv preprint arXiv:2004.08454*, 2020. 17

[JL09] Iain M Johnstone and Arthur Yu Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. 2, 4, 7

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on Computational Hardness of Hypothesis Testing: Predictions using the Low-Degree Likelihood Ratio. *arXiv preprint arXiv:1907.11636*, 2019. 2, 5

[LZ20] Yuetian Luo and Anru Zhang. Tensor clustering with planted structures: Statistical optimality and computational limits. *CoRR*, abs/2005.10743, 2020. 3, 4, 21

[MR14] Andrea Montanari and Emile Richard. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014. 2, 3, 10

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016. 2

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. 17

[NWZ20] Jonathan Niles-Weed and Ilias Zadik. The all-or-nothing phenomenon in sparse tensor pca. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17674–17684. Curran Associates, Inc., 2020. 26

[NZ20] Jonathan Niles-Weed and Ilias Zadik. The all-or-nothing phenomenon in sparse tensor PCA. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 25

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014. 15

[PWB16] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. *CoRR*, abs/1612.07728, 2016. 25

[PWB+20] Amelia Perry, Alexander S Wein, Afonso S Bandeira, et al. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 230–264. Institut Henri Poincaré, 2020. 26

[SC19] Jonathan Scarlett and Volkan Cevher. An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation. *arXiv preprint arXiv:1901.00555*, 2019. 16

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. *arXiv preprint arXiv:1706.08672*, 2017. 3

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012. 18

[Tao14] Terence Tao. Metric entropy analogues of sum set theory, 2014. Available at: `https://terrytao.wordpress.com/2014/03/19/metric-entropy-analogues-of-sum-set-theory/`. 15

[Tro15] Joel A Tropp. An Introduction to Matrix Concentration Inequalities. *arXiv preprint arXiv:1501.01571*, 2015. 18

[TS14] Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014. 18

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018. 15, 18, 28

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.2 for an overview of our contributions.
   (b) Did you describe the limitations of your work? [Yes] We explicit state any assumptions made.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work does not present any foreseeable negative societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state and discuss any assumptions made in our statements.
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to lack of space, we focus on giving the key ideas and proof techniques in the main paper and defer formal proofs to the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Background

A.1 Packings and nets

Packings and nets are useful in helping us discretize a possibly infinite metric space. Let $\mathcal{X}$ be a set of points and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ be a (pseudo)metric[17] for $\epsilon > 0$. That is, $(\mathcal{X}, d)$ is a (pseudo)metric space. An ϵ-packing $\mathcal{X}' \subseteq \mathcal{X}$ is a subset where any two distinct points $x, y \in \mathcal{X}'$ have distance $d(x, x') > \epsilon$. An ϵ-net[18] $\mathcal{X}' \subseteq \mathcal{X}$ is a subset such that for any point $x \in \mathcal{X}$, there exists some point $x' \in \mathcal{X}'$ (possibly itself) where $d(x, x') \leq \epsilon$. Under these notions, the covering number $N(\mathcal{X}, d, \epsilon)$ and packing number $P(\mathcal{X}, d, \epsilon)$ are defined as the size of the *smallest* ϵ-net of $\mathcal{X}$ and *largest* ϵ-packing of $\mathcal{X}$ respectively. It is known[19] that

$$P(\mathcal{X}, d, 2\epsilon) \leq N(\mathcal{X}, d, \epsilon) \leq P(\mathcal{X}, d, \epsilon)$$

A.2 Hermite polynomials

In this section, we introduce Hermite polynomials and state some properties used in our low-degree analysis. For further details, see [O'D14, Section 11.2].

Definition 10 (Inner product of functions)**.** *For a pair of functions f and g operating on the same domain $\mathcal{D}$, their inner product is defined by $\langle f, g \rangle = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}}[f(\boldsymbol{z})g(\boldsymbol{z})]$.*

The set $\{1, \boldsymbol{z}, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^D\}$ is a basis for the set of polynomials with maximum degree D on Gaussian variable $\boldsymbol{z} \sim N(\mu, 1)$. By applying the Gram-Schmidt process and noting that odd functions in $\boldsymbol{z}$ have expectation 0, we can diagonalize this set to obtain an orthogonal basis: $H_{e_0}(\boldsymbol{z}) = 1$, $H_{e_1}(\boldsymbol{z}) = \boldsymbol{z}$, $H_{e_2}(\boldsymbol{z}) = \boldsymbol{z}^2 - 1$, $H_{e_3}(\boldsymbol{z}) = \boldsymbol{z}^3 - 3\boldsymbol{z}$, etc.

The orthogonal basis for polynomials of maximum degree D $\{H_{e_n}\}_{n \in [D]}$ is also called the *probabilists' Hermite polynomials*. It is known that

$$\mathbb{E}_{\boldsymbol{z} \sim N(\mu,1)}[H_{e_n}(\boldsymbol{z})] = \mu^n \quad \text{and} \quad \mathbb{E}_{\boldsymbol{z} \sim N(\mu,1)}[(H_{e_n}(\boldsymbol{z}))^2] = n!$$

As we are interested in *orthonormal* bases, we use the *normalized* probabilists' Hermite polynomials $\{h_n\}_{n \in \mathbb{N}}$ where $h_n = \frac{1}{\sqrt{n!}} H_{e_n}$. One can check that

$$\mathbb{E}_{\boldsymbol{z} \sim N(\mu,1)}[h_n(\boldsymbol{z})] = \frac{1}{\sqrt{n!}} \mu^n \quad \text{and} \quad \mathbb{E}_{\boldsymbol{z} \sim N(\mu,1)}[(h_n(\boldsymbol{z}))^2] = 1$$

A.3 Information theory

Techniques from the statistical minimax theory, such as the Fano method, allow us to *lower bound* the worst case behavior of *any* estimator. In the following discourse, we borrow some notation from [Duc16]. Given a single tensor observation $\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p}$ generated from an underlying signal $x \in U_t$ (i.e. the parameter of the observation is $\theta(\boldsymbol{Y}) = x$), an estimator $\widehat{\theta}(\boldsymbol{Y})$ outputs some unit vector in $\widehat{x} \in U_k$. For two vectors x and x', we use the pseudometric[20] $\rho(x, x') = \min\{\|x - x'\|_2, \|x + x'\|_2\}$ and the loss function $\Phi(t) = t^2/2$. Thus, $\Phi(\rho(x, x')) = 1 - |\langle x, x' \rangle|$ with the corresponding minimax risk being

$$\inf_{\widehat{\theta}} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[\Phi \left(\rho \left(\widehat{\theta}(\boldsymbol{Y}), \theta(\boldsymbol{Y}) \right) \right) \right] = \inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle| \right]$$

A common way to *lower bound* the minimax risk function is to look at it from the lens of a *finite* testing problem. The *canonical hypothesis testing problem*[21] (in our context) is as follows. Let $\mathcal{X} \subseteq U_k$ be

[17] A metric satisfies 3 properties: (1) $d(x, y) = 0$ if and only if $x = y$; (2) $d(x, y) = d(y, x)$; (3) $d(x, y) \leq d(x, z) + d(z, y)$. A pseudometric may violate (1) by allowing $d(x, y) = 0$ for distinct $x \neq y$. Pseudometrics are also sometimes called semimetrics.

[18] Nets are also referred to as coverings.

[19] e.g. See resources such as [Tao14] and [Ver18, Section 4.2].

[20] Instead of the "standard" $\|x - x'\|_2$ loss, we want a loss function that captures the "symmetry" that $\langle x, x' \rangle^p = \langle x, -x' \rangle^p$ for even tensor powers p. Clearly, $\rho(x, y) = \rho(y, x)$ and one can check that $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$. Observe that ρ is a pseudometric (and not a metric) because $\rho(x, y) = 0$ holds for $x = -y$.

[21] See Section 13.2.1 in [Duc16].

an ϵ-packing of U_k of size $|\mathcal{X}| = m \geq P(U_k, \rho, \epsilon)$. That is, $\min_{x_i, x_j \in \mathcal{X}, i \neq j} \rho(x_i, x_j) > \epsilon$. Then, (1) Nature chooses a unit vector $\boldsymbol{x} \in \mathcal{X}$ *uniformly at random*; (2) We observe tensor $\boldsymbol{Y} = \boldsymbol{W} + \lambda \boldsymbol{x}^{\otimes p}$; (3) A test $\Psi : \mathcal{Y} \to \mathcal{X}$ determines what is the planted unit vector. Applying Fano's inequality (Lemma 11) and the data processing inequality (Lemma 12) to the above-mentioned canonical hypothesis testing problem, one can show[22] that

$$\inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle|\right] \geq \inf_{\widehat{x} \in U_k} \sup_{x \in \mathcal{X}} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle|\right] \geq \frac{\epsilon^2}{4} \cdot \left(1 - \frac{I(x; \boldsymbol{Y}) + 1}{\log m}\right)$$

where $I(x; \boldsymbol{Y})$ is the mutual information between x and $\boldsymbol{Y}$. Since it is known[23] that for $x \in \mathcal{X}$, one can upper bound $I(x; \boldsymbol{Y})$ by $I(x; \boldsymbol{Y}) \leq \max_{u,v \in \mathcal{X}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right)$, we have

$$\inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle|\right] \geq \frac{\epsilon^2}{4} \cdot \left(1 - \frac{\max_{u,v \in \mathcal{X}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) + 1}{\log m}\right) \tag{3}$$

where $D_{KL}(\cdot \| \cdot)$ is the KL-divergence function and $\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u}$ is the probability distribution of observing $\boldsymbol{Y}$ from signal u with additive standard Gaussian noise tensor $\boldsymbol{W}$.

We now state standard facts regarding Fano's inequality without proof. For an introductory exposition on Fano's inequality and its applications, we refer readers to [SC19].

Lemma 11 (Fano's inequality (Uniform input distribution))**.** *Let $X, Y \in \mathcal{X}$ denote the (hidden) input and (observed) output. Given Y, let $\widehat{X} \in \mathcal{X}$ be the estimated version of X by* any *estimator, and $P_e = \mathbb{P}[X \neq \widehat{X}]$ be the event that the estimation is wrong. If X is uniformly distributed over $\mathcal{X}$, then*

$$P_e \geq 1 - \frac{I(X; \widehat{X}) + 1}{\log |\mathcal{X}|}$$

where $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$ is the mutual information function.

The following inequality makes the Fano's inequality more user-friendly since it replaces the $I(X; \widehat{X})$ term with $I(X; Y)$. In statistical learning, it is typically easier to bound $I(X; Y)$ as we know how Y is generated given X.

Lemma 12 (Data processing inequality)**.** *Suppose variables X, Y and $\widehat{X}$ form a Markov chain relation $X \to Y \to \widehat{X}$. That is, X and $\widehat{X}$ are independent given Y. Then, $I(X; Y) \geq I(X; \widehat{X})$.*

A.4 Low-degree method

The low-degree likelihood ratio is a proxy to model efficiently computable functions. It is closely related to the pseudo-calibration technique and it has been developed in a recent line of work on the Sum-of-Squares hierarchy ([BHK+19, HS17, HKP+17, Hop18]). In this section, we will only introduce the basic idea and encourage interested readers to see [Hop18, BKW20] for further details.

The objects of study are distinguishing versions of planted problems: given two distributions and an instance, the goal is to decide from which distribution the instance was sampled. For us, the distinguishing formulation takes the form of deciding whether the tensor $\boldsymbol{Y}$ was sampled according to the (planted) distribution as described in Model 4, or if it was sampled from the (null) distribution where $\boldsymbol{W} \in \otimes^p \mathbb{R}^n$ has i.i.d. entries sampled from $N(0, 1)$. In general, we denote with ν the null distribution and with μ the planted distribution with the hidden structure.

A.4.1 Background on Classical Decision Theory

From the point of view of classical Decision Theory, the optimal algorithm to distinguish between two distribution is well-understood. Given distributions ν and μ on a measurable space $\mathcal{S}$, the likelihood ratio $L(\boldsymbol{Y}) := d\mathbb{P}_\mu(\boldsymbol{Y}) / d\mathbb{P}_\nu(\boldsymbol{Y})$[24] is the optimal function to distinguish whether $\boldsymbol{Y} \sim \nu$ or $\boldsymbol{Y} \sim \mu$ in the following sense.

[22] E.g. see [Duc16, Proposition 13.10], adapted to our context. Recall that we had $\Phi(t) = t^2/2$.

[23] E.g. see [SC19, Lemma 4].

[24] The Radon-Nikodym derivative.

Proposition 13 ([NP33]). *If ν is absolutely continuous with respect to μ, then the unique solution of the optimization problem*

$$\max \mathbb{E}_\mu \left[f(\boldsymbol{Y})\right] \qquad \textit{subject to } \mathbb{E}_\nu \left[f^2(\boldsymbol{Y})\right] = 1$$

is the normalized likelihood ratio $L(\boldsymbol{Y})/\mathbb{E}_\nu \left[L(\boldsymbol{Y})^2\right]$ and the optimum value is $\mathbb{E}_\nu \left[L(\boldsymbol{Y})^2\right]$.

Arguments about statistical distinguishability are also well-understood. Unsurprisingly, the likelihood ratio plays a major role here as well and a key concept is Le Cam's contiguity.

Definition 14 ([Cam60]). *Let $\underline{\mu} = (\mu_n)_{n\in\mathbb{N}}$ and $\underline{\nu} = (\nu_n)_{n\in\mathbb{N}}$ be sequences of probability measures on a common probability space $\mathcal{S}_n$. Then $\underline{\mu}$ and $\underline{\nu}$ are contiguous, written $\underline{\mu} \triangleleft \underline{\nu}$, if as $n \to \infty$, whenever for $A_n \in \mathcal{S}_n$, $\mathbb{P}_{\underline{\mu}}(A_n) \to 0$ then $\mathbb{P}_{\underline{\nu}}(A_n) \to 0$.*

Contiguity allows us to capture the idea of indistinguishability of probability measures. Two contiguous sequences $\underline{\mu}, \underline{\nu}$ of probability measures are said to be indistinguishable if there is no function $f : \mathcal{S}_n \to \{0, 1\}$ such that $f(\boldsymbol{Y}) = 1$ with high probability whenever $\boldsymbol{Y} \sim \underline{\mu}$ and $f(\boldsymbol{Y}) = 0$ with high probability whenever $\boldsymbol{Y} \sim \underline{\nu}$. The *second moment method* allows us to establish contiguity through the likelihood ratio.

Proposition 15. *If $\mathbb{E}_\nu \left[L_n(\boldsymbol{Y})^2\right]$ remains bounded as $n \to \infty$, then $\underline{\mu} \triangleleft \underline{\nu}$.*

This discussion allows us to argue whether a given function can be used to distinguish between our planted and null distributions. In particular, for probability measures μ and ν over $\mathcal{S}$, and a given function $f : \mathcal{S} \to \mathbb{R}$, we can say that f cannot distinguish between μ and ν if it satisfies the following bound on the χ^2-divergence:

$$\frac{|\mathbb{E}_\mu(f(\boldsymbol{Y})) - \mathbb{E}_\nu(f(\boldsymbol{Y}))|}{\sqrt{\mathbb{V}_\nu(f(\boldsymbol{Y}))}} \leq o(1).$$

A.4.2 Background on the Low-degree Method

The main problem with the likelihood ratio is that it is hard to compute in general, thus the analysis has to be restricted to the space of efficiently computable functions. Concretely, we use low-degree multivariate polynomials in the entries of the observation $\boldsymbol{Y}$ as a proxy for efficiently computable functions. By denoting the space of degree $\leq D$ polynomials in $\boldsymbol{Y}$ with $\mathbb{R}_{\leq D}[\boldsymbol{Y}]$, we can establish a low-degree version of the Neyman-Pearson lemma.

Proposition 16 (e.g. [Hop18]). *The unique solution of the optimization problem*

$$\max_{f\in\mathbb{R}_{\leq D}[\boldsymbol{Y}]} \mathbb{E}_\mu \left[f(\boldsymbol{Y})\right] \qquad \textit{subject to } \mathbb{E}_\nu \left[F(\boldsymbol{Y})\right] = 1$$

is the normalized orthogonal projection $L^{\leq D}(\boldsymbol{Y})/\mathbb{E}_\nu \left[L^{\leq D}(\boldsymbol{Y})^2\right]$ of the likelihood ratio $L(\boldsymbol{Y})$ onto $\mathbb{R}_{\leq D}[\boldsymbol{Y}]$ and the value of the optimization problem is $\mathbb{E}_\nu \left[L^{\leq D}(\boldsymbol{Y})^2\right]$.

With the reasoning above in mind, it is then natural to argue that a polynomial $p(\boldsymbol{Y}) \in \mathbb{R}_{\leq D}[\boldsymbol{Y}]$ cannot distinguish between μ and ν if

$$\frac{|\mathbb{E}_\mu(p(\boldsymbol{Y})) - \mathbb{E}_\nu(p(\boldsymbol{Y}))|}{\sqrt{\mathbb{V}_\nu(p(\boldsymbol{Y}))}} \leq o(1). \tag{4}$$

It is important to remark that, at the heart of our discussion, there is the belief that in the study of planted problems, low-degree polynomials capture the computational power of efficiently computable functions. This can be phrased as the following conjecture.

Conjecture 17 (Informal[25]). *For "nice" sequences of probability measures $\underline{\mu}$ and $\underline{\nu}$, if there exists $D = D(d) \geq O(\log d)$ for which $\mathbb{E}_\nu \left[L^{\leq D}(\boldsymbol{Y})^2\right]$ remains bounded as $d \to \infty$, then there is no polynomial-time algorithm that distinguishes in the sense described in Appendix A.4.1*[26].

A large body of work support this conjecture (see citations mentioned) by providing evidence of an intimate relation between polynomials, sum of squares algorithms, and lower bounds.

[25] See [BHK$^+$19, HS17, HKP$^+$17, Hop18].

[26] We do not explain what "nice" means (e.g. see [Hop18]) and remark that the most general formulation of the conjecture above (i.e. a broad definition of "nice" distributions) has been rejected ([HW20]).

A.4.3 Chi-squared divergence and orthogonal polynomials

From a technical point of view, the key observation used to prove bounds for low-degree polynomials is the fact that the polynomial which maximizes the ratio in Eq. (4) has a convenient characterization in terms of orthogonal polynomials with respect to the null distribution.

Formally, for any linear subspace of polynomials $\mathcal{S}_{\leq D} \subseteq \mathbb{R}[\boldsymbol{Y}]_{\leq D}$ and any absolutely continuous probability distribution ν such that all polynomials of degree at most $2D$ are ν-integrable, one can define an inner product in the space $\mathcal{S}_{\leq D}$ as follows

$$\forall p, q \in \mathcal{S}_{\leq D} \quad \langle p, q\rangle = \mathbb{E}_{\boldsymbol{Y}\sim\nu} p(\boldsymbol{Y}) q(\boldsymbol{Y})\,.$$

Hence we can talk about orthonormal basis in $\mathcal{S}_{\leq D}$ with respect to this inner product.

Proposition 18 (See [dKNS20] for a proof)**.** *Let $\mathcal{S}_{\leq D} \subseteq \mathbb{R}[\boldsymbol{Y}]_{\leq D}$ be a linear subspace of polynomials of dimension N. Suppose that ν and μ are probability distributions over $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ such that any polynomial of degree at most D is μ-integrable and any polynomial of degree at most $2D$ is ν-integrable. Suppose also that ν is absolutely continuous. Let $\{\psi_i(\boldsymbol{Y})\}_{i=1}^{N}$ be an orthonormal basis in $\mathcal{S}_{\leq D}[\boldsymbol{Y}]$ with respect to ν. Then*

$$\max_{p\in\mathcal{S}_{\leq D}} \frac{(\mathbb{E}_\mu p(\boldsymbol{Y}))^2}{\mathbb{E}_\nu p^2(\boldsymbol{Y})} = \sum_{i=1}^{N} (\mathbb{E}_\mu \psi_i)^2\,.$$

In the case of Gaussian noise, a useful orthonormal basis in $\mathbb{R}[\boldsymbol{Y}]_{\leq D}$ is the system of Hermite polynomials $\{H_\alpha(\boldsymbol{Y})\}_{|\alpha|\leq D}$ (see Appendix A.2). By applying Proposition 18 to the subspace of polynomials such that $\mathbb{E}_\nu p(\boldsymbol{Y}) = 0$, we get

Corollary 19. *Let ν be Gaussian. Suppose that the distribution μ is so that any polynomial of degree at most D is μ-integrable. Then*

$$\max_{p\in\mathbb{R}[\boldsymbol{Y}]_{\leq D}} \frac{(\mathbb{E}_\mu p(\boldsymbol{Y}) - \mathbb{E}_\nu p(\boldsymbol{Y}))^2}{\mathbb{V}_\nu p(\boldsymbol{Y})} = \sum_{0<|\alpha|\leq D} (\mathbb{E}_\mu H_\alpha(Y))^2\,.$$

A.5 Sparse norm bounds

Denote $\mathcal{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ as the n-dimensional unit sphere and $A \in \mathbb{R}^{m\times n}$ be a matrix. Then, the matrix norm of A is defined as

$$\|A\| = \max_{x\in\mathcal{S}^{n-1}} \|Ax\|_2 = \max_{x\in\mathcal{S}^{m-1}, y\in\mathcal{S}^{n-1}} x^\top A y = \max_{x\in\mathcal{S}^{m-1}, y\in\mathcal{S}^{n-1}} \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j} x_i y_j$$

More generally, the tensor norm of an order $p \geq 2$ tensor $T \in \mathbb{R}^{n_1\times\cdots\times n_p}$ is defined as

$$\begin{aligned}\|T\| &= \max_{x_{(1)}\in\mathcal{S}^{n_1-1},\ldots,x_{(p)}\in\mathcal{S}^{n_p-1}} T\left(x_{(1)}, x_{(2)}, \ldots, x_{(p)}\right)\\ &= \max_{x_{(1)}\in\mathcal{S}^{n_1-1},\ldots,x_{(p)}\in\mathcal{S}^{n_p-1}} \sum_{i_1=1}^{n_1}\cdots\sum_{i_p=1}^{n_p} T_{i_1,\ldots,i_p} x_{(1),i_1} x_{(2),i_2}\ldots x_{(p),i_p}\end{aligned}$$

In the following, let all dimensions be equal (i.e. $n = m = n_1 = \ldots = n_p$). Without any sparsity conditions, it is known[27] that $\|\boldsymbol{A}\| \leq \mathcal{O}(\sqrt{n} + t)$ with high probability for matrix $\boldsymbol{A}$ with i.i.d. standard Gaussian entries. For tensors, [TS14] proved that $\|\boldsymbol{T}\| \leq \mathcal{O}\left(\sqrt{np\log(p) + \log(1/\gamma)}\right)$ with probability at least $1-\gamma$. These results are typically proven using ϵ-net arguments over the unit sphere $\mathcal{S}^{n-1}$. However, to the best of our knowledge, there is no known result for bounding the norm of $\boldsymbol{A}$ or $\boldsymbol{T}$ when interacting with r distinct (at most) s-sparse unit vectors from the set $\mathcal{S}_s^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1,\ |\mathcal{I}_x| \leq s \leq n\}$.

Using ϵ-net arguments, we bound $\left|\boldsymbol{T}\left(x_{(1)}, \ldots, x_{(p)}\right)\right|$ when $x_{(1)}, \ldots, x_{(p)}$ are r distinct vectors from $\mathcal{S}_s^{n-1}$. Our result recovers known bounds, up to constant factors, when $s = n$ and $r = p$.

[27] e.g. See [Ver18, Section 4.4.2], [Tro15, Section 4.2.2] and [Tao12, Section 2.3.1].

Lemma 20 (Tensor sparse bound)**.** *Let $\boldsymbol{T} \in \otimes^p \mathbb{R}^n$ be an order $p \geq 2$ tensor with i.i.d. standard Gaussian entries and $x_{(1)}, \ldots, x_{(p)} \in \mathcal{S}_s^{n-1}$ be r distinct (at most) s-sparse unit vectors. Then, for $1 \leq s \leq n$, $1 \leq r \leq p$, and $\gamma \in (0,1)$,*

$$\mathbb{P}\left[\left|\boldsymbol{T}(x_{(1)}, \ldots, x_{(p)})\right| \geq \sqrt{8 \cdot \left(4rs \ln\left(\frac{np}{s}\right) + \ln\left(\frac{1}{\gamma}\right)\right)}\right] \leq 2\gamma$$

Proof sketch of Lemma 20 (described for $p = 2$). We fix an ϵ-net $\mathcal{N}(\mathcal{S}_s^{n-1})$ of size $N(\mathcal{S}_s^{n-1}, \epsilon)$ over sparse vectors, bound the norm for an arbitrary point in $\mathcal{N}(\mathcal{S}_s^{n-1}) \times \mathcal{N}(\mathcal{S}_s^{n-1})$, and apply union bound over all points in $\mathcal{N}(\mathcal{S}_s^{n-1}) \times \mathcal{N}(\mathcal{S}_s^{n-1})$. Since $\mathcal{S}_s^{n-1} \times \mathcal{S}_s^{n-1}$ is compact, there is a maximizing point that attains the norm. We complete the proof by relating the maximizer (and hence the norm) to its closest point in $\mathcal{N}(\mathcal{S}_s^{n-1}) \times \mathcal{N}(\mathcal{S}_s^{n-1})$. See Appendix F.1 for the formal proof. □

B Limited Brute Force Recovery Algorithm

We first discuss how to extend the discussions from Section 2 to the case where signals are approximately flat. Then, we will prove the most general form of our algorithmic result[28] (Theorem 21) that our limited brute force search algorithm that recovers exact individual signal supports under some algorithmic assumptions. Finally, we explain how to extend our techniques to handle single-spike general tensors where the signal takes the form $x_{(1)} \otimes \ldots \otimes x_{(p)}$ involving $1 \leq \ell \leq p$ distinct k-sparse tensors in Appendix B.2. As in Section 2, our exposition will ignore the constant factors introduced by Algorithm 1.

B.1 Approximately flat signals

By factoring $A \geq 1$ into our assumptions on minimal signal strength $\lambda_{\min}$ and relative strength ratio κ, we can extend the above analyses (using the same proof outline as in Section 2.2) so that Algorithm 3 recovers the individual supports of multiple approximately flat (k, A)-sparse signals. Besides accounting for A factors, the only significant change in the analysis is in how we lower bound $\langle \boldsymbol{Y}^{(1)}, \boldsymbol{v}_* \rangle$. Consider the first iteration (by our discussion in Section 2.2, other iterations are similar) and let $u_* \in U_t$ be the vector satisfying

$$\lambda \langle x, u_* \rangle^p = \max_{q \in [r], u \in U_t} \lambda_q \langle x_{(q)}, u \rangle^p \,.$$

This choice allows us to account for skewed signals. Let $\lambda_{\min} = \min_{q \in [r]} \lambda_q$ and $\lambda_{\max} = \max_{q \in [r]} \lambda_q$. For some $0 < \epsilon \leq 1/2$. Suppose that

$$\lambda_{\min} \gtrsim \frac{A^p}{\epsilon \cdot (1-\epsilon)^p} \cdot \sqrt{t \left(\frac{k}{t}\right)^p \log n} \quad \text{and} \quad \lambda_{\min} \gtrsim A^{2p} \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1} \cdot \lambda_{\max} \,. \tag{5}$$

By definition of $\boldsymbol{v}_*$, we have

$$\langle \boldsymbol{Y}^{(1)}, \boldsymbol{v}_* \rangle \geq \langle \boldsymbol{Y}^{(1)}, u_* \rangle = \lambda \langle x, u_* \rangle^p + \langle \boldsymbol{W}^{(1)}, u_*^{\otimes p} \rangle \,. \tag{6}$$

Conversely, consider an arbitrary $u \in U_t$ such that $\left|\operatorname{supp}(u) \cap \operatorname{supp}\left(x_{(q)}\right)\right| < (1-\epsilon) \cdot t$ for all $q \in [r]$. Intuitively, the largest attainable value for $\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$ is obtained removing $\epsilon \cdot t$ entries from u_* and placing them on some highly skewed signal. Using Eq. (5) and Eq. (6), it is possible to show[29]

$$\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle \leq \lambda \langle x, u_* \rangle^p \cdot \left(1 - \frac{\epsilon}{A^2}\right)^{p-1} + \langle \boldsymbol{W}^{(1)}, u^{\otimes p} \rangle \,.$$

Thus, it follows that $\boldsymbol{v}_*$ satisfies $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(i)}\right)\right| \geq (1-\epsilon)t$ for some $i \in [r]$. We can now repeat the same analysis as Section 2.2 to argue that $\boldsymbol{\alpha}$ behaves as an indicator vector for $\operatorname{supp}\left(x_{(i)}\right)$, taking account of A factors.

[28] Theorem 1 and Theorem 2 from the main text are direct consequences of Theorem 21.

[29] For the full derivation, see the proof of Lemma 23 in Appendix F.2. A $\sqrt{2}$ factor appears due to Algorithm 1.

Theorem 21 (Multi-spike recovery for approximately flat signals)**.** *Consider Model 4. Suppose*

$$\lambda_r \gtrsim \frac{\kappa}{(A\epsilon)^p} \sqrt{t \left(\frac{k}{t}\right)^p \ln\left(\frac{n}{\delta}\right)}, \quad \lambda_r \geq \kappa \cdot \lambda_1, \quad \textit{and} \quad \kappa \geq 5A^{2p} \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}.$$

Then, Algorithm 3 that runs in $\mathcal{O}(rpn^{p+t})$ time and, with probability at least $1-\delta$, outputs the individual signal supports $\operatorname{supp}\left(x_{(\pi(1))}\right), \ldots, \operatorname{supp}\left(x_{(\pi(r))}\right)$ *with respect to some unknown bijection* $\pi : [r] \to [r]$.

We remark that the constant factors are chosen to make the analysis clean; smaller factors are possible. As discussed in Remark 9, after recovering the support coordinates, one can then run known tensor PCA recovery methods on the appropriate sub-tensor to obtain a good approximation of each signal.

For convenience, let us restate Algorithm 1 and Algorithm 3.

Algorithm 1 Preprocessing

```
Input: $\boldsymbol{Y}$.
Sample a Gaussian tensor $\boldsymbol{Z} \in \otimes^p \mathbb{R}^n$ where each entry is an i.i.d. standard Gaussian $N(0,1)$.
Return two independent copies $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ of $\boldsymbol{Y}$ as follows:
    $\boldsymbol{Y}^{(1)} = \frac{1}{\sqrt{2}}(\boldsymbol{Y} + \boldsymbol{Z})$   and   $\boldsymbol{Y}^{(2)} = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - \boldsymbol{Z})$
```

Algorithm 3 Multi-spike limited brute force

```
Input: $k, t, r$ and $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ obtained from Algorithm 1.
Repeat for $i = 1$ to $r$:
    Compute $\boldsymbol{v}_* := \operatorname{argmax}_{u \in U_t} \langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$ subject to $\operatorname{supp}(\boldsymbol{v}_*) \cap \left( \bigcup_{1 \leq j < i} \mathcal{I}_j \right) = \emptyset$.
    Compute the vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ with entries $\boldsymbol{\alpha}_\ell := \langle \boldsymbol{Y}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle$ for every $\ell \in [n]$.
    Let $\mathcal{I}_i$ be the set of indices of the largest $k$ entries of $\boldsymbol{\alpha}$.
Return $\mathcal{I}_1, \ldots, \mathcal{I}_r$.
```

Let us begin with a simple running time analysis.

Lemma 22. *Algorithm 3 runs in $\mathcal{O}(rpn^{p+t})$ time.*

Proof of Lemma 22. Sampling $\boldsymbol{Z}$ and creating copies $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ take $\mathcal{O}(n^p)$ time. Fix an arbitrary round. Observe that $|U_t| = \binom{n}{t} 2^t \leq ((2e)/t)^t n^t \leq e^2 n^t$. Each computation of $\langle \boldsymbol{Y}^{(1)}, u^{\otimes p} \rangle$ can be naively performed in $\mathcal{O}(pn^p)$ time while checking whether for disjoint support can be done naively in additional $\mathcal{O}(n^2)$ time for each $u \in U_t$. Similarly, the computation of $\boldsymbol{\alpha}$ can be done in $\mathcal{O}(pn^p)$ time and we can perform a linear scan in $\mathcal{O}(n)$ time to obtain the largest k entries. So, an arbitrary round runs in $\mathcal{O}(pn^{p+t})$ time. We perform the entire process r times. □

Algorithm 3 recovers the exact k-sparse support of some signal $x_{(\pi(i))}$ in each round. That is, $\mathcal{I}_{x_{(\pi(i))}} = \operatorname{supp}\left(x_{(\pi(i))}\right)$ for $i \in [r]$. It succeeds when these two claims hold for any round $i \in [r]$:

(I) The t-sparse maximizer $\boldsymbol{v}_*$ shares $\geq (1-\epsilon) \cdot t$ support coordinates with some signal $x_{(\pi(i))}$.

(II) The k largest entries of $\boldsymbol{\alpha}$ belong to the support $\operatorname{supp}\left(x_{(\pi(i))}\right)$ of $x_{(\pi(i))}$.

Lemma 23 and Lemma 24 address these claims respectively. See Appendix F.2 for their proofs.

Lemma 23. *Consider Model 4 and an arbitrary round $i \in [r]$. Suppose*

$$\lambda_r \geq \frac{32\kappa}{(A\epsilon)^p} \sqrt{t \left(\frac{k}{t}\right)^p \ln(n)}, \quad \lambda_r \geq \kappa \cdot \lambda_1, \quad \textit{and} \quad \kappa \geq 5A^{2p} \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}.$$

Then,

$$\mathbb{P}\left[\left|\operatorname{supp}\left(\boldsymbol{v}_*\right) \cap \operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon) \cdot t\right] \geq 1 - 4\exp\left(-\lambda_r^2 \frac{\epsilon^2}{128A^4}\left(\frac{t}{k}\right)^p\right)$$

Lemma 24. *Consider Model 4 and an arbitrary round $i \in [r]$. Suppose*

$$\lambda_r \geq \frac{32\kappa}{(A\epsilon)^p}\sqrt{t\left(\frac{k}{t}\right)^p \ln(n)}, \quad \lambda_r \geq \kappa \cdot \lambda_1, \quad \textit{and} \quad \kappa \geq 5A^{2p}\left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}.$$

Further suppose that $\left|\operatorname{supp}\left(\boldsymbol{v}_*\right) \cap \operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon) \cdot t$. Then, the largest (in magnitude) coordinates of $\boldsymbol{\alpha}$ are $\operatorname{supp}\left(x_{(\pi(i))}\right)$ with probability at least $1 - 2n\exp\left(-\lambda_r^2 \frac{A^{2p}\epsilon^{2p-2}}{16\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$.

We now prove Theorem 21 using Lemma 22, Lemma 23, and Lemma 24.

Proof of Theorem 21. Lemma 22 gives the running time. The correctness of Algorithm 3 hinges on Lemma 23 and Lemma 24 always succeeding. In each round, we need Lemma 23 to succeed once and Lemma 24 to succeed at most n times. There are a total of at most $r(1+n)$ events, each failing with probability at most $4n\exp\left(-\lambda_r^2 \frac{A^{2p-4}\epsilon^{2p-2}}{128\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$. By union bound, the probability of *any* event fails is at most $r(1+n) \cdot 4n\exp\left(-\lambda_r^2 \frac{A^{2p-4}\epsilon^{2p-2}}{128\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$. By the disjoint signal support assumption, we have $r \leq n$. So, when $\lambda_r \gtrsim \frac{\kappa}{(A\epsilon)^p}\sqrt{t\left(\frac{k}{t}\right)^p \ln\left(\frac{n}{\delta}\right)}$, Algorithm 3 succeeds with probability at least $1-\delta$. □

B.2 General tensors for single spike

We now briefly describe how to extend the model of Model 4 to the case where the single tensor signal could be made up of $1 \leq \ell \leq p$ distinct k-sparse vectors[30]: instead of the signal being $x^{\otimes p}$, it is $x_{(1)} \otimes \ldots \otimes x_{(p)}$ involving ℓ distinct vectors. The discussions in this section can be further generalized to the case of multiple approximately flat spikes using the techniques from Section 2.2 and Appendix B.1.

Given ℓ, we can modify Algorithm 2 to search over $U_t^{\otimes \ell}$ and compute $\boldsymbol{v}_*$ that maximizes

$$\langle \boldsymbol{Y}^{(1)}, u_{(1)} \otimes \ldots \otimes u_{(p)} \rangle$$

where there are $\binom{p-1}{\ell-1}$ possible ways[31] to form the signal using ℓ distinct t-sparse vectors. By Lemma 20, $\langle \boldsymbol{W}^{(1)}, u_{(1)} \otimes \ldots \otimes u_{(p)} \rangle \lesssim \sqrt{\ell t \log(n)}$ whenever one (or more) of the t-sparse vectors used to form $u_{(1)} \otimes \ldots \otimes u_{(p)}$ is *not* part of the actual signal. Suppose the maximizer $\boldsymbol{v}_*$ involves ℓ distinct vectors $\boldsymbol{v}_{*,(1)}, \ldots, \boldsymbol{v}_{*,(\ell)}$. For notational convenience, let us write $\boldsymbol{v}_* \setminus \boldsymbol{v}_{*,(q)}$ to mean the tensor of order $p-1$ derived by removing one copy of $\boldsymbol{v}_{*,(q)}$ from $\boldsymbol{v}_*$. Then, for each distinct vector $\boldsymbol{v}_{*,(q)}$ in the maximizer, define $\boldsymbol{\alpha}_{(q)} \in \mathbb{R}$ where $\alpha_{(q),i} = \langle \boldsymbol{Y}^{(2)}, (\boldsymbol{v}_* \setminus \boldsymbol{v}_{*,(q)}) \otimes e_i \rangle$ and output the k largest entries of $\boldsymbol{\alpha}_{(q)}$ as the support of $\boldsymbol{v}_{*,(q)}$.

This modified algorithm will run in time $\mathcal{O}\left(\ell (pe)^\ell n^{p+\ell t}\right)$.[32] Adapting our analysis for the single spike accordingly (by using Lemma 20 with ℓ distinct vectors) will show that we can recover the signal supports of each $u_{(q)}$ whenever $\lambda \gtrsim \sqrt{\ell t (k/t)^p \log n}$.[33] Notice that this is an improvement over the algorithm of [LZ20] for $\ell \in o(p)$ or $t \geq 2$ when $p \in \omega(1)$.

[30] This model has been studied by [LZ20]. To be precise, they actually allow different *known* sparsity levels for each $x_{(q)}$ vector. Here, we assume that all of them are k-sparse. It is conceptually straightforward (but complicated and obfuscates the key idea) to extend the current discussion to allow different sparsity values.

[31] There are $\binom{p-1}{\ell-1}$ ways to obtain integer solutions to $x_1 + \ldots + x_\ell = p$ assuming $x_1 \geq 1, \ldots, x_\ell \geq 1$.

[32] The ℓ factor is due to using ℓ copies of $\boldsymbol{\alpha}$. The increase from p to p^ℓ is due to trying $\binom{p-1}{\ell-1}$ combinations. The increase of n^t to $n^\ell e^\ell$ is due to searching over $U_t^{\otimes \ell}$.

[33] The extra $\sqrt{\ell}$ factor follows from Lemma 20 to accomodate ℓ distinct vectors in the maximization.

C Computational Low-Degree Bounds

In this section, we formalize our results on the computational low-degree bounds for sparse tensor PCA. We will first show a low-degree lower bound on the distinguishing problem using k-sparse scaled Rademacher unit vectors and then will give a low-degree distinguishing algorithm showing that the lower-bound is tight in certain parameter regimes.

Following the discussion in Appendix A.4, we design the following distinguishing problem.

Problem 25 (Hypothesis testing for single-spiked k-sparse scaled Rademacher vectors)**.** *Given an observation tensor $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$, decide whether:*

$$\textit{Null distribution } H_0 : \boldsymbol{Y} = \boldsymbol{W}$$

$$\textit{Planted distribution } H_1 : \boldsymbol{Y} = \boldsymbol{W} + \lambda \boldsymbol{x}^{\otimes p}$$

where $\boldsymbol{W}$ is a noise tensor with i.i.d. $N(0,1)$ entries and $\boldsymbol{x}$ is a k-sparse scaled Rademacher unit vector whose entries are independently drawn as follows:

$$\boldsymbol{x}_i = \begin{cases} 1/\sqrt{k} & \textit{with probability } k/(2n), \\ -1/\sqrt{k} & \textit{with probability } k/(2n), \\ 0 & \textit{with probability } 1 - k/n\,. \end{cases}$$

Formally speaking, the vector $\boldsymbol{x}$ in Problem 25 is not necessarily a unit vector, as compared to the planted signal in the single-spike case of Model 4. However, since $\boldsymbol{x}$ is $k(1+o(1))$-sparse with high probability, a lower bound given by Problem 25 implies a distinguishing lower bound for single-spike sparse tensor model with a planted $k(1+o(1))$-sparse vector and signal strength $\frac{\lambda}{1+o(1)}$. We study Problem 25 through the lens of low-degree polynomials. Since $\boldsymbol{W}$ is Gaussian noise, we use the set of normalized probabilists' Hermite polynomials $\{h_\alpha\}_\alpha$ as our orthogonal basis. Our strategy is similar to prior works such as [HKP+17, HS17, DKWB19]: By examining the low-degree analogue of the χ^2-divergence between probability measures, we will show that low-degree polynomial estimators cannot distinguish H_0 and H_1.

We now state the two main theorems that we will prove in the following subsections. For a cleaner exposition, we defer some proofs to Appendix F.3.

Theorem 26 (Single-spike low-degree distinguishability lower bound)**.** *Let $p \geq 2$, $1 \leq D \leq 2n/p$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. If $0 \leq \epsilon \leq 1/2$ and*

$$\lambda \leq \sqrt{\frac{\epsilon D}{e 4^p}} \min\left\{ \left(\frac{n}{pD}\right)^{p/4}, \left(\frac{k}{pD}\left(1 + \left|\ln\left(\frac{npD}{ek^2}\right)\right|\right)\right)^{p/2} \right\},$$

then

$$\chi^2(H_1 \parallel H_0) = \sup_{|\alpha| \leq D} \frac{\left(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}) - \mathbb{E}_{H_0} h_\alpha(\boldsymbol{Y})\right)^2}{Var_{H_0} h_\alpha(\boldsymbol{Y})} = \sum_{|\alpha| \leq D} \left(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y})\right)^2 \leq 2\epsilon.$$

Theorem 27 (Single-spike low-degree distinguishability lower bound)**.** *Let $p \geq 2$, $1 \leq D \leq 2n/p$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. If either of the following holds:*

1. *If D is even and*

$$\lambda \geq \epsilon^{\frac{1}{2D}} e^{\frac{p}{2}} \sqrt{D} \left(\frac{n}{pD}\right)^{\frac{p}{4}}$$

2. *If $p \leq n$, $D \leq \frac{\ln^2(n/p)}{4e^2}$ is even, $\sqrt{np} \cdot \left(\frac{e\sqrt{D}}{\ln(n/k)}\right) \leq k \leq \sqrt{np}$, and*

$$\lambda \geq \epsilon^{\frac{1}{2D}} \sqrt{D} \left(\frac{k}{pD} \ln\left(\frac{n}{k}\right)\right)^{\frac{p}{2}}$$

then

$$\chi^2(H_1 \parallel H_0) = \sup_{|\alpha| \leq D} \frac{\left(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}) - \mathbb{E}_{H_0} h_\alpha(\boldsymbol{Y})\right)^2}{Var_{H_0} h_\alpha(\boldsymbol{Y})} = \sum_{|\alpha| \leq D} \left(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y})\right)^2 \geq \epsilon.$$

C.1 Low-degree lower bound

To prove our computational lower bound, we first compute $(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}))^2$ explicitly in Lemma 28 using properties of the normalized probabilists' Hermite polynomials for a given degree parameter D. Then, we upper bound $\sum_{|\alpha| \leq D} \left(\mathbb{E}_{H_1}[h_\alpha(\boldsymbol{Y})]\right)^2$ using Lemma 29 and Lemma 30. Solving for the condition on λ such that $\sum_\alpha \left(\mathbb{E}_{H_1}[h_\alpha(\boldsymbol{Y})]\right)^2 \ll \epsilon$ yields our computational lower bound Theorem 26.

Lemma 28. *Let $p \geq 2$, $d \geq 1$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. An entry of $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ can be indexed by either an integer from $[n^p]$ or a p-tuple. Define $\phi : [n^p] \to [n]^p$, α, $c(\alpha)$, $s(\alpha)$, and $\mathbb{1}_{even(c(\alpha))}$ as follows:*

- *$\phi(i)$ maps to a p-tuple indicating the p (possibly repeated) entries of $\boldsymbol{x}$ that are used.*
- *$\alpha = (\alpha_1, \ldots, \alpha_{n^p})$ is an n^p-tuple that corresponds to a Hermite polynomial of degree $|\alpha| = \sum_{i=1}^{n^p} \alpha_i$. For each i, α_i is the number of times entry $\boldsymbol{Y}_{\phi(i)}$ was chosen, where each $\boldsymbol{Y}_{\phi(i)}$ references p coordinates of $\boldsymbol{x}$.*
- *$c(\alpha) = (c_1, \ldots, c_n)$, where c_j is the number of times $\boldsymbol{x}_j$ is used in α.*
- *$s(\alpha)$ is the number of distinct non-zero $\boldsymbol{x}_j$'s in $c(\alpha)$.*
- *$\mathbb{1}_{even(c(\alpha))}$ be the indicator whether* all *c_j's are even.*

Under these definitions, we have the following:

$$\left(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y})\right)^2 = \lambda^{2d} k^{-pd} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right).$$

Lemma 29. *Let $p \geq 2$, $1 \leq D \leq 2n/p$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. Then,*

$$\sum_{|\alpha| \leq D} \left(\mathbb{E}_{H_1}[h_\alpha(\boldsymbol{Y})]\right)^2 \leq \sum_{d=1}^{D} \frac{\lambda^{2d}}{d!} \sum_{s=1}^{pd/2} \left(\frac{ek^2}{sn}\right)^s \left(\frac{s}{k}\right)^{pd}.$$

Lemma 30. *For $p \geq 2$, $d \geq 1$, $1 \leq k \leq n$ and $1 \leq s \leq pd/2$, we have*

$$\left(\frac{ek^2}{sn}\right)^s \left(\frac{s}{k}\right)^{pd} \leq \left[\frac{2pd}{\min\left\{\sqrt{npd},\ k\left(1 + \left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd}.$$

We are now ready to prove Theorem 26.

Proof of Theorem 26. Lemma 29 and Lemma 30 together tell us that

$$
\begin{aligned}
\sum_{|\alpha|\le D} \left(\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})]\right)^2 &\le \sum_{d=1}^{D} \frac{\lambda^{2d}}{d!}\frac{pd}{2}\left[\frac{2pd}{\min\left\{\sqrt{npd},\ k\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd}\\
&= \sum_{d=1}^{D} \lambda^{2d}\left[\left(\frac{1}{2d!}\right)^{\frac{1}{pd}}(pd)^{\frac{1}{pd}}\frac{2pd}{\min\left\{\sqrt{npd},\ k\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd}\\
&\le \sum_{d=1}^{D} \lambda^{2d}\left[\left(\frac{e}{d}\right)^{\frac{1}{p}}\frac{4pd}{\min\left\{\sqrt{npd},\ k\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd} && (\star)\\
&= \sum_{d=1}^{D} \lambda^{2d}\left(\frac{e}{d}\right)^{d}\frac{4^{pd}}{\min\left\{\left(\frac{n}{pd}\right)^{pd/2},\ \left(\frac{k}{pd}\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right)^{pd}\right\}}\\
&\le \sum_{d=1}^{D} \epsilon^{d}\left(\frac{D}{d}\right)^{d}\frac{\min\left\{\left(\frac{n}{pD}\right)^{pd/2},\ \left(\frac{k}{pD}\left(1+\left|\ln\left(\frac{npD}{ek^2}\right)\right|\right)\right)^{pd}\right\}}{\min\left\{\left(\frac{n}{pd}\right)^{pd/2},\ \left(\frac{k}{pd}\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right)^{pd}\right\}} && (*)\\
&\le \sum_{d=1}^{D} \epsilon^{d} && (\dagger)
\end{aligned}
$$

where $(\star)$ is because $\frac{1}{2d!} \le \frac{1}{d!} \le \left(\frac{e}{d}\right)^d$ and $(pd)^{\frac{1}{pd}} \le 2$, $(*)$ is the theorem assumption on λ, and $(\dagger)$ is because $p \ge 2$. The statement follows since $\sum_{d=1}^{D} \epsilon^d \le \frac{\epsilon}{1-\epsilon} \le 2\epsilon$ for $0 \le \epsilon \le 1/2$. □

C.2 Low-degree distinguishing algorithm

The starting point of our distinguishing algorithm is the explicit expression from Lemma 28 and Claim 34. Assuming D is even[34], we show that degree D Hermite polynomials is "sufficiently large" by considering a subset of terms in the explicit summation.

Proof of Theorem 27. Under the assumption of $D \le \frac{2n}{p}$ and D is even, Lemma 28 and Claim 34 together tell us that

$$
\begin{aligned}
\sum_{|\alpha|\le D} \left(\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})]\right)^2 &= \sum_{d=1}^{D} \lambda^{2d}k^{-pd}\frac{1}{d!}\sum_{s=1}^{\lfloor pd/2\rfloor}\binom{n}{s}\left(\frac{k}{n}\right)^{2s}\sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}}\binom{pd}{2\beta_1,\ldots,2\beta_s}\\
&\ge \frac{\lambda^{2D}}{D!}k^{-pD}\sum_{s=1}^{pD/2}\binom{n}{s}\left(\frac{k}{n}\right)^{2s}\sum_{\substack{\beta_1+\ldots+\beta_s=pD/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}}\binom{pD}{2\beta_1,\ldots,2\beta_s}\\
&\ge \frac{\lambda^{2D}}{D!}k^{-pD}\binom{n}{pD/2}\left(\frac{k}{n}\right)^{pD}\binom{pD}{2,\ldots,2} && (\dagger)\\
&\ge \frac{\lambda^{2D}}{D!}k^{-pD}\left(\frac{2n}{pD}\right)^{pD/2}\left(\frac{k}{n}\right)^{pD}\left(\frac{pD}{e}\right)^{pD}2^{-pD/2} && (\ddagger)\\
&\ge \left(\frac{\lambda^2}{D}\left(\frac{kpD}{ek\sqrt{npD}}\right)^{p}\right)^{D} && (\star)
\end{aligned}
$$

where $(\dagger)$ is by only using $s = pD/2$ and $\beta_1 = \ldots = \beta_{pD/2} = 1$, $(\ddagger)$ is due to $\binom{n}{k} \ge (n/k)^k$ and $n! \ge (n/e)^n$, and $(\star)$ is because $D! \le D^D$.

[34] For $D \ge 2$, we consider Hermite polynomials of degree either D or $D-1$ (whichever is even).

When $\lambda \geq \epsilon^{\frac{1}{2D}} e^{\frac{p}{2}} \sqrt{D} \left(\frac{n}{pD} \right)^{\frac{p}{4}}$, we see that

$$\chi^2(H_1 \parallel H_0) = \sum_{|\alpha| \leq D} (\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})])^2 \geq \left(\frac{\lambda^2}{D} \left(\frac{kpD}{ek\sqrt{npD}} \right)^p \right)^D = \left(\frac{\lambda^2}{De^p} \left(\frac{pD}{n} \right)^{\frac{p}{2}} \right)^D \geq \epsilon$$

We now assume that $p \leq n$, $D \leq \frac{\ln^2(n/p)}{4e^2}$ and $\sqrt{np} \cdot \left(\frac{e\sqrt{D}}{\ln(n/k)} \right) \leq k \leq \sqrt{np}$. Then,

$$\left(\frac{\lambda^2}{D} \left(\frac{kpD}{ek\sqrt{npD}} \right)^p \right)^D = \left(\frac{\lambda^2}{D} \left(\frac{pD}{k\ln(\frac{n}{k})} \right)^p \left(\frac{k\ln(\frac{n}{k})}{\sqrt{e^2 npD}} \right)^p \right)^D \geq \left(\frac{\lambda^2}{D} \left(\frac{pD}{k\ln(\frac{n}{k})} \right)^p \right)^D$$

where the last inequality is because $\sqrt{np} \cdot \left(\frac{e\sqrt{D}}{\ln(n/k)} \right) \leq k$. The constraints $p \leq n$ and $\sqrt{D}e \leq \frac{1}{2}\ln(n/p) = \ln(n/\sqrt{np}) \leq \ln(n/k)$ ensure that there exists valid values of k.

So when $\lambda \geq \epsilon^{\frac{1}{2D}} \sqrt{D} \left(\frac{k}{pD} \ln\left(\frac{n}{k}\right) \right)^{\frac{p}{2}}$, we see that

$$\chi^2(H_1 \parallel H_0) = \sum_{|\alpha| \leq D} (\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})])^2 \geq \left(\frac{\lambda^2}{D} \left(\frac{pD}{k\ln(\frac{n}{k})} \right)^p \right)^D \geq \epsilon$$

□

D Information-theoretic Lower Bound

In this section, we will use standard techniques[35] in information theory to lower bound the minimax risk for approximate signal recovery in the single-spike sparse tensor PCA.

Remark We notice that equivalent results appeared in [PWB16, NZ20]. Nevertheless we include it for completeness.

Consider the following setting: Given a single tensor observation $\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p}$ generated from an underlying signal $x \in U_t$ (i.e. the parameter of the observation is $\theta(\boldsymbol{Y}) = x$), an estimator $\widehat{\theta}(\boldsymbol{Y})$ outputs some unit vector in $\widehat{x} \in U_k$. For two vectors x and x', we use the pseudometric[36] $\rho(x, x') = \min\{\|x - x'\|_2, \|x + x'\|_2\}$ and the loss function $\Phi(t) = t^2/2$. Thus, $\Phi(\rho(x, x')) = 1 - |\langle x, x' \rangle|$ with the corresponding minimax risk being

$$\inf_{\widehat{\theta}} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[\Phi\left(\rho\left(\widehat{\theta}(\boldsymbol{Y}), \theta(\boldsymbol{Y}) \right) \right) \right] = \inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle| \right]$$

Let $\mathcal{X} \subseteq U_k$ be an ϵ-packing of U_k of size $|\mathcal{X}| = m \geq P(U_k, \rho, \epsilon)$. Then, one can show that the minimax risk can be lower bounded as follows:

$$\inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle| \right] \geq \frac{\epsilon^2}{4} \cdot \left(1 - \frac{\max_{u,v \in \mathcal{X}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v} \right) + 1}{\log m} \right) \tag{7}$$

where $D_{KL}(\cdot \| \cdot)$ is the KL-divergence function and $\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u}$ is the probability distribution of observing $\boldsymbol{Y}$ from signal u with additive standard Gaussian noise tensor $\boldsymbol{W}$. The following information-theoretic lower bound is shown by lower bounding Eq. (7).

Theorem 31 (Single-spike info-theoretic lower bound)**.** *Given $\boldsymbol{Y} = \boldsymbol{W} + \lambda x^{\otimes p} \in \otimes^p \mathbb{R}^n$ where $\boldsymbol{W}$ is a noise tensor with i.i.d. $N(0, 1)$ entries and the planted signal $x \in U_k$ has signal strength λ. Let $\widehat{x} \in U_k$ be the recovered signal by* any *estimator. If $n \geq 2k$ and $\lambda \leq \sqrt{\frac{k}{12} \log\left(\frac{n-k}{k}\right) - \frac{1}{2}}$, then*

$$\inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}} \left[1 - |\langle \widehat{x}, x \rangle| \right] \geq 0.05.$$

[35] See Appendix A.3 for a brief introduction.

[36] Instead of the "standard" $\|x - x'\|_2$ loss, we want a loss function that captures the "symmetry" that $\langle x, x' \rangle^p = \langle x, -x' \rangle^p$ for even tensor powers p. Clearly, $\rho(x, y) = \rho(y, x)$ and one can check that $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$. Observe that ρ is a pseudometric (and not a metric) because $\rho(x, y) = 0$ holds for $x = -y$.

Remark With $n \geq 2k$, we see that $\log(\frac{n-k}{k}) \geq \log(\frac{n}{2k})$. Then, for if $n \geq 2^{\frac{1}{1-c}} k$ for any constant $c \in (0,1)$, we see that $\log(\frac{n}{2k}) \geq c \log(\frac{n}{k})$. In particular, when $n \geq 4k$, we have $\log(\frac{n}{2k}) \geq \frac{1}{2}\log(\frac{n}{k})$ and can write $\lambda \lesssim \sqrt{k \log(n/k)}$.

To prove the result, we lower bound m and upper bound the KL-divergence. Since $\|x - x'\|_2 \geq \rho(x, x')$, we see that $N(U_k, \|\cdot\|_2, \epsilon) \leq P(U_k, \|\cdot\|_2, \epsilon) \leq P(U_k, \rho, \epsilon) \leq m$. To lower bound m, we lower bound $N(U_k, \|\cdot\|_2, \epsilon)$ via Lemma 32. Then, we upper bound the KL-divergence in Lemma 33 by the triangle inequality and the KL-divergence of Gaussian vectors. We defer the proofs of Lemma 32 and Lemma 33 to Appendix F.4.

Lemma 32. *Let U_k be the set of k-sparse flat unit vectors and $N(U_k, \|\cdot\|_2, \epsilon)$ be the ϵ-covering number of U_k with respect to Euclidean distance. For $\epsilon \in (0,1]$ and $n \geq 2k$,*

$$N(U_k, \|\cdot\|_2, \epsilon) \geq \left(\frac{n-k}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)}.$$

Lemma 33. *Denote $\mathcal{S}_k^{n-1}$ as the set of k-sparse unit vectors. Then,*

$$\max_{u,v \in \mathcal{S}_k^{n-1}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) \leq 2\lambda^2$$

where $D_{KL}(\cdot\|\cdot)$ is the KL-divergence function and $\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u}$ is the probability distribution of observing $\boldsymbol{Y}$ from signal u with additive standard Gaussian noise tensor $\boldsymbol{W}$.

We are now ready to prove Theorem 31 by using the above two lemmata.

Proof of Theorem 31. The theorem follows by computing a lower bound for Eq. (7) with $\mathcal{X}$ as an ϵ-packing of U_k of size $|\mathcal{X}| = m \geq P(U_k, \rho, \epsilon)$.

Lemma 32 tells us that $N(U_k, \|\cdot\|_2, \epsilon) \geq \left(\frac{n-k}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)}$. Since $\|x - x'\|_2 \geq \rho(x, x')$, we see that $N(U_k, \|\cdot\|_2, \epsilon) \leq P(U_k, \|\cdot\|_2, \epsilon) \leq P(U_k, \rho, \epsilon) \leq m$. Thus,

$$k\left(1 - \frac{\epsilon^2}{2}\right) \log\left(\frac{n-k}{k}\right) \leq \log m$$

Meanwhile, Lemma 33 tells us that $\max_{u,v \in \mathcal{S}_k^{n-1}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) \leq 2\lambda^2$. Since $U_k \subseteq \mathcal{S}_k^{n-1}$, this implies that

$$\max_{u,v \in \mathcal{X}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) \leq 2\lambda^2$$

Let $\tau > 0$ be a lower bound constant (which we fix later). Putting the above bounds together, we have

$$\begin{aligned}
\inf_{\widehat{x} \in U_k} \sup_{x \in U_k} \mathbb{E}_{\boldsymbol{Y}}\left[1 - |\langle \widehat{x}, x\rangle|\right] &\geq \frac{\epsilon^2}{4} \cdot \left(1 - \frac{\max_{u,v \in \mathcal{X}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) + 1}{\log m}\right) && \text{Eq. (7)} \\
&\geq \frac{\epsilon^2}{4} \cdot \left(1 - \frac{2\lambda^2 + 1}{k\left(1 - \frac{\epsilon^2}{2}\right) \log\left(\frac{n-k}{k}\right)}\right) \\
&\geq \tau
\end{aligned}$$

Rearranging, we get $\lambda \leq \sqrt{\frac{(1-(4\tau)/\epsilon^2)(1-\epsilon^2/2)}{2} k \log\left(\frac{n-k}{k}\right) - \frac{1}{2}}$. The claims follows[37] by setting $\tau = 0.05$ and $\epsilon = 1/2$. □

In the setting context of our interest, the works of [NWZ20, PWB+20] both papers outline a phase transition at $\lambda = \Theta(\sqrt{k \log(n/k)})$. Specifically, they prove that weak recovery is possible when $\lambda \gtrsim \sqrt{k \log(n/k)}$ and information theoretically impossible when $\lambda \lesssim \sqrt{k \log(n/k)}$. Our information theoretic bounds presented here are equivalent to these results, up to constant factors.

[37] Observe that $(1 - (4 * 0.05)/(0.5^2)) * (1 - (0.5^2)/2)/2 = 0.0875 > 1/12$.

E Related Model: Planted sparse densest sub-hypergraph

The planted k-densest sub-hypergraph model ([CPMB19, BCPS20, CPSB20]) is closely related to our sparse spiked tensor model Model 4. While not directly reducible from/to one another, techniques developed in one model can inform the other.

The planted k-densest sub-hypergraph model is a weighted complete hypergraph where a subset of k planted vertices, denoted by $S \subseteq [n]$, is *drawn uniformly at random* and each hyperedge involves p vertices, for $2 \leq p \leq k \leq n$. Except for the $\binom{k}{p}$ *one-sided biased* hyperedges (belonging to the planted subgraph induced by the k vertices in S) whose weights follow the Gaussian distribution $N(\beta, \sigma^2)$, the weight of all remaining $\binom{n}{p} - \binom{k}{p}$ hyperedges follow a Gaussian distribution $N(0, \sigma^2)$. In other words, the hyperedge defined by $\{i_1, \ldots, i_p\}$, for $i_1, \ldots, i_p \in [n]$, has weight

$$\boldsymbol{Y}_{i_1,\ldots,i_p} = \begin{cases} \beta + \boldsymbol{W}_{i_1,\ldots,i_p} & \text{if } i_1, \ldots, i_p \in S \\ \boldsymbol{W}_{i_1,\ldots,i_p} & \text{otherwise} \end{cases}$$

where each $\boldsymbol{W}_{i_1,\ldots,i_p} \sim N(0, \sigma^2)$ is independent and planted entries have a $\beta > 0$ bias.

As one can see, the planted k-densest sub-hypergraph model (PDSM) is very similar to the single-spike ($r = 1$) sparse spiked tensor model (SSTM) that we study[38]. However, there are two key model differences that one needs to be aware of. Firstly, there are n^p observations in SSTM instead of $\binom{n}{p}$ in PDSM as the former is not constrained to hyperedges (e.g. $\boldsymbol{Y}_{1,\ldots,1}$ is a valid data observation in SSTM but not in PDSM). Secondly, our signal bias is not one-sided and is scaled by a factor of $k^{-p/2}$: For a planted coordinate $(i_1, \ldots, i_p)$, SSTM observes $\boldsymbol{Y}_{i_1,\ldots,i_p} = \boldsymbol{W}_{i_1,\ldots,i_p} \pm \lambda k^{-p/2}$ instead of $\boldsymbol{Y}_{i_1,\ldots,i_p} = \boldsymbol{W}_{i_1,\ldots,i_p} + \beta$ in PDSM, where the sign of bias in SSTM depends on the polarities of signal entries $x_{i_1}, \ldots, x_{i_p}$.

While the signal scaling discrepancies can be handled by replacing $\beta\sqrt{\binom{k}{p}}$ terms in PDSM bounds with λ in SSTM[39], the one-sidedness of the signal bias has implications on the computational hardness of the two models. In a recent work, [CPSB20] showed that an Approximate Message Passing (AMP) algorithm succeeds in signal recovery for PDSM when $\lambda \gtrsim \frac{k}{p\sqrt{n}} \left(\frac{n}{k}\right)^{p/2}$. In constrast, Theorem 2 (for constant p) tells us that our polynomial time algorithm for SSTM succeeds when $k \leq \sqrt{np}$ and $\lambda \gtrsim \sqrt{k^p \log\left(\frac{np}{k}\right)}$. Meanwhile, Theorem 3 implies that it is impossible to recover the signal using low-degree polynomials whenever the signal-to-noise ratio satisfies $\lambda \lesssim \widetilde{O}\left(\min\left\{\left(\frac{n}{p}\right)^{p/4}, \left(\frac{k}{p}\left(1 + \left|\ln\left(\frac{np}{k^2}\right)\right|\right)\right)^{p/2}\right\}\right)$. Indeed, the one-sidedness of the signal bias in PDSM appears to make the problem *computationally easier* than SSTM in some regimes[40].

Nevertheless, we believe that techniques used in either model are generally applicable in the other and we expect a variant of our limited brute force algorithm to work in PDSM. From a statistical viewpoint, [CPSB20] proved information-theoretic lower bounds for recovery in PDSM of $\lambda \lesssim \sqrt{k \log n}$ while we have $\lambda \lesssim \sqrt{k \log((n-k)/k)}$ for approximate signal recovery[41] in SSTM, which matches the PDSM bounds when $k \in o(n)$. These results in both models rely on standard techniques such as Fano's inequality.

[38] We believe that handling a more general σ^2 is a non-issue when comparing these models because the σ^2 factor could be propagated throughout our analysis by appropriately adjusting the sub-Gaussian concentration arguments.

[39] This discrepancy arises due to having $\binom{k}{p}$ planted hyperedges in PDSM, as opposed to k^p signal entries in SSTM, and the signal strength scaling of $k^{-p/2}$ in SSTM.

[40] e.g. Large k regimes such as $k = n^{0.9}$. For large k, a heuristic adaptation of our low-degree analysis to PDSM shows that the relationship between parameters λ, n, k and p in a low-degree bound is roughly of the form $\lambda \gtrsim \left(\frac{n}{k}\right)^{p/2}$ instead of $\lambda \gtrsim n^{p/4}$. This roughly matches the AMP bounds shown by [CPSB20] and further provides credence to our claim that techniques from one model can applied to the other.

[41] i.e. It is enough to find a strongly correlated estimate $\widehat{x}$ of the signal x where $\widehat{x}$ could be "wrong on a few coordinates". The bounds for exact and approximate recovery in [CPSB20] differ by constant factors.

F Deferred proofs and details

This section provides the formal proofs that were deferred in favor for readability. For convenience, we will restate the statements before proving them.

F.1 Sparse norm bounds

Lemma 20 (Tensor sparse bound)**.** *Let $\boldsymbol{T} \in \otimes^p \mathbb{R}^n$ be an order $p \geq 2$ tensor with i.i.d. standard Gaussian entries and $x_{(1)}, \ldots, x_{(p)} \in \mathcal{S}_s^{n-1}$ be r distinct (at most) s-sparse unit vectors. Then, for $1 \leq s \leq n$, $1 \leq r \leq p$, and $\gamma \in (0,1)$,*

$$\mathbb{P}\left[\left|\boldsymbol{T}(x_{(1)}, \ldots, x_{(p)})\right| \geq \sqrt{8 \cdot \left(4rs \ln\left(\frac{np}{s}\right) + \ln\left(\frac{1}{\gamma}\right)\right)}\right] \leq 2\gamma$$

Proof of Lemma 20. We will focus on proving the following statement:

$$\mathbb{P}\left[\max_{\substack{x_{(1)}, \ldots, x_{(p)} \in \mathcal{S}_s^{n-1}, \\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) \geq \sqrt{8 \cdot \left(4rs \ln\left(\frac{np}{s}\right) + \ln\left(\frac{1}{\gamma}\right)\right)}\right] \leq \gamma \tag{8}$$

By a similar argument, one can obtain

$$\mathbb{P}\left[\min_{\substack{x_{(1)}, \ldots, x_{(p)} \in \mathcal{S}_s^{n-1}, \\ r \text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)}, \ldots, x_{(p)}\right) \leq -\sqrt{8 \cdot \left(4rs \ln\left(\frac{np}{s}\right) + \ln\left(\frac{1}{\gamma}\right)\right)}\right] \leq \gamma \tag{9}$$

The lemma follows from a union bound of Eq. (8) and Eq. (9).

It now remains to prove Eq. (8). Let λ, t, ϵ be proof parameters which we fix later. Define $\mathcal{N}(\mathcal{S}_s^{n-1})$ as an ϵ-cover of $\mathcal{S}_s^{n-1}$ of smallest cardinality $N(\mathcal{S}_s^{n-1}, \epsilon)$. It is known[42] that $\left(\frac{1}{\epsilon}\right)^n \leq N(\mathcal{S}^{n-1}, \epsilon) \leq \left(\frac{2}{\epsilon} + 1\right)^n \leq \left(\frac{3}{\epsilon}\right)^n$. Treating each unit sphere defined on s coordinates independently and then taking union bound gives us $N(\mathcal{S}_s^{n-1}, \epsilon) \leq \binom{n}{s} \cdot N(\mathcal{S}^{n-1}, \epsilon) \leq \left(\frac{en}{s}\right)^s \left(\frac{3}{\epsilon}\right)^s$. So, $\left|\mathcal{N}(\mathcal{S}_s^{n-1})\right| \leq \left(\frac{3en}{\epsilon s}\right)^s$. For any r distinct vectors $x_{(1)}, \ldots, x_{(p)} \in \mathcal{N}(\mathcal{S}_s^{n-1})$,

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) \geq t\right] \\
&= \mathbb{P}\left[\sum_{i_1, \ldots, i_p = 1}^{n} \boldsymbol{T}_{i_1, \ldots, i_p} x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p} \geq t\right] \\
&= \mathbb{P}\left[\exp\left(\lambda \sum_{i_1, \ldots, i_p = 1}^{n} \boldsymbol{T}_{i_1, \ldots, i_p} x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p}\right) \geq e^{\lambda t}\right] \\
&\leq e^{-\lambda t} \cdot \mathbb{E}\left[\exp\left(\lambda \sum_{i_1, \ldots, i_p = 1}^{n} \boldsymbol{T}_{i_1, \ldots, i_p} x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p}\right)\right] && \text{Markov's inequality} \\
&= e^{-\lambda t} \cdot \Pi_{i_1, \ldots, i_p = 1}^{n} \mathbb{E}\left[\exp\left(\boldsymbol{T}_{i_1, \ldots, i_p} \lambda x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p}\right)\right] \\
&= e^{-\lambda t} \cdot \Pi_{i_1, \ldots, i_p = 1}^{n} \exp\left(\frac{\left(\lambda x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p}\right)^2}{2}\right) && \boldsymbol{T}_{i_1, \ldots, i_p} \sim N(0,1) \\
&= \exp\left(-\lambda t + \sum_{i_1, \ldots, i_p = 1}^{n} \frac{\left(\lambda x_{(1), i_1} x_{(2), i_2} \ldots x_{(p), i_p}\right)^2}{2}\right) \\
&= \exp\left(-\lambda t + \frac{\lambda^2}{2}\right) && \sum_{i_1 = 1}^{n} x_{(1), i_1}^2 = \ldots = \sum_{i_p = 1}^{n} x_{(p), i_p}^2 = 1 \\
&\leq \exp\left(-\frac{t^2}{2}\right) && \text{Maximized when } \lambda = t
\end{aligned}$$

[42] e.g. See [Ver18, Corollary 4.2.13].

By union bound over all $\left(N(\mathcal{S}_s^{n-1}, \epsilon)\right)^r \leq \left(\frac{3en}{\epsilon s}\right)^{rs}$ r distinct points in $\otimes^p \mathcal{N}(\mathcal{S}_s^{n-1})$,

$$\begin{aligned}
\mathbb{P}\left[\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{N}(\mathcal{S}_s^{n-1}),\\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) \geq t\right] &\leq \sum_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{N}(\mathcal{S}_s^{n-1}),\\ r \text{ distinct vectors}}} \mathbb{P}\left[\boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) \geq t\right] \\
&\leq \left(\frac{3en}{\epsilon s}\right)^{rs} \exp\left(-\frac{t^2}{2}\right) \\
&= \exp\left(rs \ln\left(\frac{3en}{\epsilon s}\right) - \frac{t^2}{2}\right)
\end{aligned}$$

As $\otimes^p \mathcal{S}_s^{n-1}$ is compact, there are r distinct vectors $x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(p)} \in \mathcal{S}_s^{n-1}$ such that

$$\left(x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(p)}\right) = \operatorname{argmax}_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)}, \ldots, x_{(p)}\right)$$

By definition of ϵ-cover, there are vectors $x_{(1)}, x_{(2)}, \ldots, x_{(p)} \in \mathcal{N}(\mathcal{S}_s^{n-1})$ such that $x^*_{(1)} = x_{(1)} + \delta_{(1)}, \ldots, x^*_{(p)} = x_{(p)} + \delta_{(p)}$, where $\left\|\delta_{(z)}\right\|_2 \leq \epsilon$ for $z \in \{1, \ldots, p\}$. Let $z \in \{1, \ldots, p\}$. Since $x^*_{(z)}$ and $x_{(z)}$ are s-sparse, $\delta_{(z)}$ is at most $(2s)$-sparse. We can express $\delta_{(z)} = \delta^{(1)}_{(z)} + \delta^{(2)}_{(z)}$ as a sum of two s-sparse vectors where $\frac{\delta^{(1)}_{(z)}}{\left\|\delta^{(1)}_{(z)}\right\|_2}, \frac{\delta^{(2)}_{(z)}}{\left\|\delta^{(2)}_{(z)}\right\|_2} \in \mathcal{S}_s^{n-1}$, $\left\|\delta^{(1)}_{(z)}\right\|_2 \leq \left\|\delta_{(z)}\right\|_2 \leq \epsilon$, and $\left\|\delta^{(2)}_{(z)}\right\|_2 \leq \left\|\delta_{(z)}\right\|_2 \leq \epsilon$. We can relate $\boldsymbol{T}(x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(p)})$ to $\boldsymbol{T}(x_{(1)}, x_{(2)}, \ldots, x_{(p)})$ by expanding the definition:

$$\begin{aligned}
&\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, x_{(2)}, \ldots, x_{(p)}) \\
&= \sum_{i_1,\ldots,i_p=1}^{n} \boldsymbol{T}_{i_1,\ldots,i_p} x^*_{(1),i_1} x^*_{(2),i_2} \cdots x^*_{(p),i_p} \\
&= \sum_{i_1,\ldots,i_p=1}^{n} \boldsymbol{T}_{i_1,\ldots,i_p} \left(x_{(1)} + \delta^{(1)}_{(1)} + \delta^{(2)}_{(1)}\right)_{i_1} \left(x_{(2)} + \delta^{(1)}_{(2)} + \delta^{(2)}_{(2)}\right)_{i_2} \cdots \left(x_{(p)} + \delta^{(1)}_{(p)} + \delta^{(2)}_{(p)}\right)_{i_p} \\
&\leq \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) + \left(\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)})\right) \cdot \left(\epsilon \cdot 2\binom{p}{1} + \epsilon^2 \cdot 2^2\binom{p}{2} + \ldots + \epsilon^p 2^p \binom{p}{p}\right) \\
&\leq \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) + \left(\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)})\right) \cdot \left(\frac{2p\epsilon}{1!} + \frac{(2p\epsilon)^2}{2!} + \ldots + \frac{(2p\epsilon)^p}{p!}\right) \\
&\leq \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)}) + \left(\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}(x_{(1)}, \ldots, x_{(p)})\right) \cdot \left(e^{2p\epsilon} - 1\right)
\end{aligned}$$

The first inequality is by counting how the δ's group together, factoring out their norms so that they belong to $\mathcal{S}_s^{n-1}$ (so $\max_{(\ldots)} \boldsymbol{T}\left(x_{(1)}, \ldots, x_{(p)}\right)$ applies), then using $\left\|\delta^{(1)}_{(z)}\right\|_2, \left\|\delta^{(2)}_{(z)}\right\|_2 \leq \epsilon$. The second inequality is due to $\binom{n}{k} \leq \frac{n^k}{k!}$. The third is due to the definition of $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Note that $1 > e^{2p\epsilon} - 1$ if and only if $\epsilon < \frac{\ln 2}{2p}$. Set $\epsilon = \frac{\ln 2}{4p}$, then $\frac{1}{2 - e^{2p\epsilon}} < 2$. Rearranging, we get

$$\begin{aligned}
\max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{S}_s^{n-1},\\ r \text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)}, x_{(2)}, \ldots, x_{(p)}\right) &\leq \frac{\boldsymbol{T}\left(x_{(1)}, x_{(2)}, \ldots, x_{(p)}\right)}{2 - e^{2p\epsilon}} \\
&\leq 2 \max_{\substack{x_{(1)},\ldots,x_{(p)} \in \mathcal{N}(\mathcal{S}_s^{n-1}),\\ r \text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)}, x_{(2)}, \ldots, x_{(p)}\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{x_{(1)},\ldots,x_{(p)}\in\mathcal{S}_s^{n-1},\\ r\text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)},x_{(2)},\ldots,x_{(p)}\right)\geq t\right]\\
&\leq \mathbb{P}\left[\max_{\substack{x_{(1)},\ldots,x_{(p)}\in\mathcal{N}(\mathcal{S}_s^{n-1}),\\ r\text{ distinct vectors}}} \boldsymbol{T}\left(x_{(1)},x_{(2)},\ldots,x_{(p)}\right)\geq \frac{t}{2}\right]\\
&\leq \exp\left(rs\ln\left(\frac{3en}{\epsilon s}\right)-\frac{(t/2)^2}{2}\right) && \text{From above}\\
&\leq \exp\left(4rs\ln\left(\frac{np}{s}\right)-\frac{t^2}{8}\right) && \text{Since } \epsilon=\frac{\ln 2}{4p} \text{ and } \ln\left(\frac{12e}{\ln 2}\right)<4
\end{aligned}$$

Setting $t^2 = 8 \cdot \left(4rs\ln\left(\frac{np}{s}\right)+\ln\left(\frac{1}{\gamma}\right)\right)$ yields Eq. (8). □

F.2 Proofs for recovery algorithms

Lemma 23. *Consider Model 4 and an arbitrary round $i \in [r]$. Suppose*

$$\lambda_r \geq \frac{32\kappa}{(A\epsilon)^p}\sqrt{t\left(\frac{k}{t}\right)^p \ln(n)}, \quad \lambda_r \geq \kappa\cdot\lambda_1, \quad \textit{and} \quad \kappa \geq 5A^{2p}\left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}.$$

Then,

$$\mathbb{P}\left[\left|\operatorname{supp}\left(\boldsymbol{v}_*\right)\cap\operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon)\cdot t\right] \geq 1-4\exp\left(-\lambda_r^2\frac{\epsilon^2}{128A^4}\left(\frac{t}{k}\right)^p\right)$$

Proof of Lemma 23. Without loss of generality, suppose that $x_{(1)},\ldots,x_{(s)}$ are the remaining s (where $1\leq s\leq r$) unrecovered signals with signal strengths $\lambda_1,\ldots,\lambda_s$ such that $\lambda_1\geq\ldots\geq\lambda_s\geq\lambda_r$. Let $u_*\in U_t$ lie completely in some signal $\hat{x}$ with signal strength $\hat{\lambda}$ such that

$$\hat{\lambda}\langle\hat{x},u_*\rangle^p \geq \max_{q\in[s]}\max_{u\in U_t}\lambda_q\langle x_{(q)},u\rangle^p \quad \text{and} \quad \langle\boldsymbol{Y}^{(1)},u_*^{\otimes p}\rangle \geq \frac{\hat{\lambda}}{\sqrt{2}}\langle\hat{x},u_*\rangle^p+\langle\boldsymbol{W}^{(1)},u_*^{\otimes p}\rangle.$$

By optimality, $\langle\boldsymbol{Y}^{(1)},\boldsymbol{v}_*^{\otimes p}\rangle\geq\langle\boldsymbol{Y}^{(1)},u_*^{\otimes p}\rangle$. So, the claim holds if we can show that $\langle\boldsymbol{Y}^{(1)},u_*^{\otimes p}\rangle > \langle\boldsymbol{Y}^{(1)},u^{\otimes p}\rangle$ for *any* $u\in U_t$ such that

$$\left|\operatorname{supp}(u)\cap\operatorname{supp}\left(x_{(1)}\right)\right| < (1-\epsilon)\cdot t,\ldots,\left|\operatorname{supp}(u)\cap\operatorname{supp}\left(x_{(s)}\right)\right| < (1-\epsilon)\cdot t\,. \tag{10}$$

For any $u \in U_t$ that satisfies Eq. (10), we see that

$$\begin{aligned}
&\langle \boldsymbol{Y}^{(1)}, u^{\otimes p}\rangle \\
&= \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \sum_{q=1}^{s} \frac{\lambda_q}{\sqrt{2}} \langle u, x_{(q)}\rangle^p \\
&\leq \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \frac{\hat{\lambda}}{\sqrt{2}} \left(\langle \hat{x}, u_*\rangle - \frac{\epsilon}{A}\sqrt{\frac{t}{k}}\right)^p + \frac{\lambda_1 \epsilon^p A^p}{\sqrt{2}} \left(\frac{t}{k}\right)^{\frac{p}{2}} && \text{Eq. (10)} \\
&\leq \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \frac{\hat{\lambda}}{\sqrt{2}} \left(\left(\langle \hat{x}, u_*\rangle - \frac{\epsilon}{A}\sqrt{\frac{t}{k}}\right)^p + \frac{\epsilon^p A^p}{\kappa} \left(\frac{t}{k}\right)^{\frac{p}{2}}\right) && \hat{\lambda} \geq \lambda_r \geq \kappa \lambda_1 \\
&\leq \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \frac{\hat{\lambda}}{\sqrt{2}} \langle \hat{x}, u_*\rangle^p \left(\left(1 - \frac{\epsilon}{A^2}\right)^p + \frac{\epsilon^p A^p}{\kappa}\right) && \sqrt{\frac{t}{k}} \leq \langle \hat{x}, u_*\rangle \leq A\sqrt{\frac{t}{k}} \\
&\leq \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \frac{\hat{\lambda}}{\sqrt{2}} \langle \hat{x}, u_*\rangle^p \left(\left(1 - \frac{\epsilon}{A^2}\right)^p + \frac{\epsilon(1-\epsilon)^{p-1}}{A^p}\right) && \kappa \geq A^{2p} \left(\frac{\epsilon}{1-\epsilon}\right)^{p-1} \\
&\leq \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle + \frac{\hat{\lambda}}{\sqrt{2}} \langle \hat{x}, u_*\rangle^p \left(1 - \frac{\epsilon}{A^2}\right)^{p-1} && A \geq 1, \epsilon \leq \frac{1}{2}
\end{aligned}$$

Let us set parameters (r, s, γ) as $\left(1, t, \exp\left(-\frac{\lambda_r^2 \epsilon^2}{128 A^4}\left(\frac{t}{k}\right)^p\right)\right)$ in Lemma 20. Since

$$\frac{\lambda_r^2 \epsilon^2}{128 A^4} \left(\frac{t}{k}\right)^p \geq 8t \ln(n) \geq 4t \ln\left(\frac{np}{t}\right),$$

we see that

$$\frac{\lambda_r \epsilon}{2\sqrt{2}A} \left(\frac{t}{k}\right)^{\frac{p}{2}} \geq \sqrt{8\left(4t \ln\left(\frac{np}{t}\right) + \ln\left(\frac{1}{\gamma}\right)\right)}.$$

Thus, Lemma 20 gives us that, for any $u \in U_t$,

$$\mathbb{P}\left[\max_{u \in U_t} \left|\langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle\right| \geq \frac{\lambda_r \epsilon}{2\sqrt{2}A^2} \left(\frac{t}{k}\right)^{\frac{p}{2}}\right] \leq 2\exp\left(-\frac{\lambda_r^2 \epsilon^2}{128 A^4} \left(\frac{t}{k}\right)^p\right).$$

Then, with probability at least $1 - 4\exp\exp\left(-\frac{\lambda_r^2 \epsilon^2}{128 A^4}\left(\frac{t}{k}\right)^p\right)$,

$$\langle \boldsymbol{W}^{(1)}, u_*^{\otimes p}\rangle - \langle \boldsymbol{W}^{(1)}, u^{\otimes p}\rangle < \frac{\hat{\lambda}}{\sqrt{2}} \langle \hat{x}, u_*\rangle^p \frac{\epsilon}{A^2} \leq \frac{\hat{\lambda}}{\sqrt{2}} \langle \hat{x}, u_*\rangle^p \left(1 - \left(1 - \frac{\epsilon}{A^2}\right)^{p-1}\right).$$

and so $\langle \boldsymbol{Y}^{(1)}, \boldsymbol{v}_*^{\otimes p}\rangle > \langle \boldsymbol{Y}^{(1)}, u_*^{\otimes p}\rangle > \langle \boldsymbol{Y}^{(1)}, u^{\otimes p}\rangle$. for any $u \in U_t$ that satisfies Eq. (10). □

Lemma 24. *Consider Model 4 and an arbitrary round $i \in [r]$. Suppose*

$$\lambda_r \geq \frac{32\kappa}{(A\epsilon)^p} \sqrt{t\left(\frac{k}{t}\right)^p \ln(n)}, \quad \lambda_r \geq \kappa \cdot \lambda_1, \quad \text{and} \quad \kappa \geq 5A^{2p}\left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}.$$

Further suppose that $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon) \cdot t$. Then, the largest (in magnitude) coordinates of $\boldsymbol{\alpha}$ are $\operatorname{supp}\left(x_{(\pi(i))}\right)$ with probability at least $1 - 2n\exp\left(-\lambda_r^2 \frac{A^{2p}\epsilon^{2p-2}}{16\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$.

Proof of Lemma 24. Recall that

$$\boldsymbol{\alpha}_\ell = \sum_{q \in [r]} \frac{\lambda_q}{\sqrt{2}} x_{(q),\ell} \langle x_{(q)}, \boldsymbol{v}_*\rangle^{p-1} + \langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell\rangle.$$

Since $\boldsymbol{W}^{(2)}$ is independent from $\boldsymbol{W}^{(1)}$, we can apply standard Gaussian bounds. That is,

$$\mathbb{P}\left[\left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle\right| \geq \lambda_r \frac{A^p \epsilon^{p-1}}{2\kappa\sqrt{2t}} \left(\frac{t}{k}\right)^{\frac{p}{2}}\right] \leq 2\exp\left(-\lambda_r^2 \frac{A^{2p}\epsilon^{2p-2}}{16\kappa^2 t}\left(\frac{t}{k}\right)^p\right).$$

Now, conditioned on

$$\left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell \rangle\right| < \lambda_r \frac{A^p \epsilon^{p-1}}{2\kappa\sqrt{2k}} \left(\frac{t}{k}\right)^{\frac{p-1}{2}} < \lambda_r \frac{(1-\epsilon)^{p-1}}{2A^p\sqrt{2k}} \left(\frac{t}{k}\right)^{\frac{p-1}{2}},$$

we consider cases of $\ell \in \operatorname{supp}\left(x_{(\pi(i))}\right)$ and $\ell \notin \operatorname{supp}\left(x_{(\pi(i))}\right)$ separately. To be precise, we will show the following two results:

1. $\mathbb{P}\left[|\boldsymbol{\alpha}_\ell| < \lambda_r \frac{(1-\epsilon)^{p-1}}{2A^p\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}} \;\middle|\; \ell \in \operatorname{supp}\left(x_{(\pi(i))}\right)\right] \leq 2\exp\left(-\lambda_r^2 \frac{A^{2p}\epsilon^{2p-2}}{16\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$
2. $\mathbb{P}\left[|\boldsymbol{\alpha}_\ell| > \lambda_r \frac{2A^p\epsilon^{p-1}}{\kappa\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}} \;\middle|\; \ell \notin \operatorname{supp}\left(x_{(\pi(i))}\right)\right] \leq 2\exp\left(-\lambda_r^2 \frac{A^{2p}\epsilon^{2p-2}}{16\kappa^2 t}\left(\frac{t}{k}\right)^p\right)$

As $\kappa > 4A^{2p}\left(\frac{\epsilon}{1-\epsilon}\right)^{p-1}$, there will be a value gap in $|\boldsymbol{\alpha}_\ell|$ for $\ell \in \operatorname{supp}\left(x_{(\pi(i))}\right)$ versus $\ell \notin \operatorname{supp}\left(x_{(\pi(i))}\right)$. The result follows by taking a union bound over all n coordinates.

Case 1 $\left(\ell \in \operatorname{supp}\left(x_{(\pi(i))}\right)\right)$**:** Since $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon)\cdot t$,

$$\left|\frac{\lambda_{\pi(i)}}{\sqrt{2}} \cdot x_{(\pi(i)),\ell} \cdot \langle \boldsymbol{v}_*, x_{(\pi(i))}\rangle^{p-1}\right| \geq \frac{\lambda_r}{A\sqrt{2k}} \cdot \left|\langle \boldsymbol{v}_*, x_{(\pi(i))}\rangle^{p-1}\right| \geq \lambda_r \frac{(1-\epsilon)^{p-1}}{A^p\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}}.$$

By reverse triangle inequality, we have

$$|\boldsymbol{\alpha}_\ell| = \lambda_r \frac{(1-\epsilon)^{p-1}}{A^p\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}} - \left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell\rangle\right| > \lambda_r \frac{(1-\epsilon)^{p-1}}{2A^p\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}}.$$

Case 2 $\left(\ell \notin \operatorname{supp}\left(x_{(\pi(i))}\right)\right)$**:** Since signals have disjoint support and $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(\pi(i))}\right)\right| \geq (1-\epsilon)\cdot t$, we have $\left|\operatorname{supp}(\boldsymbol{v}_*) \cap \operatorname{supp}\left(x_{(j)}\right)\right| < \epsilon \cdot t$.
By triangle inequality, we have

$$\begin{aligned}
|\boldsymbol{\alpha}_\ell| &\leq \lambda_1 \frac{A^p\epsilon^{p-1}}{\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}} + \left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell\rangle\right| \\
&\leq \lambda_r \frac{A^p\epsilon^{p-1}}{\kappa\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}} + \left|\langle \boldsymbol{W}^{(2)}, \boldsymbol{v}_*^{\otimes p-1} \otimes e_\ell\rangle\right| \\
&\leq \lambda_r \frac{2A^p\epsilon^{p-1}}{\kappa\sqrt{2k}}\left(\frac{t}{k}\right)^{\frac{p-1}{2}}.
\end{aligned}$$

□

F.3 Proofs for computational bounds

Claim 34 relates the counting of $\boldsymbol{Y}$ entries with coordinates of the signal $\boldsymbol{x}$. In the claim, $s \in [n]$ is the number of entries of $\boldsymbol{x}$ that is considered in the summation. We only need to consider s up to $\lfloor pd/2 \rfloor$ because the expectation is 0 if some coordinate of $\boldsymbol{x}$ is used an odd number of times. Each α can be viewed as d consecutive chunks of p entries, and each $(\beta_1, \ldots, \beta_s)$ counts the number of times $\boldsymbol{x}_j$ occurs in α.

Claim 34. *For a fixed degree $d \leq 2n/p$,*

$$\begin{aligned}
&\sum_{|\alpha|=d} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right) \\
=&\sum_{s=1}^{\lfloor pd/2 \rfloor} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}} \binom{pd}{2\beta_1,\ldots,2\beta_s} \frac{1}{\binom{d}{\alpha_1,\ldots,\alpha_{n^p}}} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right) \\
=&\frac{1}{d!} \sum_{s=1}^{\lfloor pd/2 \rfloor} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}} \binom{pd}{2\beta_1,\ldots,2\beta_s}
\end{aligned}$$

Proof. The second equality is by definition of multinomial coefficients. To prove the first equality, we consider two equivalent ways of viewing the summation.

From the viewpoint of choosing entries of $\boldsymbol{Y}$, one chooses d (possibly repeated) entries of $\boldsymbol{Y}$ and computes $\mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right)$ directly on the corresponding α.

From the viewpoint of choosing entries from $\boldsymbol{x}$, first observe that each α considered actually involves pd (possibly repeated) entries of $[n]$ and can be mapped to a multi-set of pd numbers[43], where multiple α's could map to the same multi-set of pd numbers[44]. Thus, one can first pick a multi-set and then go over the different α's corresponding to all possible permutations[45]. Under constraint of $\mathbb{1}_{even(c(\alpha))}$, a multi-set is *valid* (contributes a non-zero term to the summation) only when the multiplicity of each number is even. So, one can view the summation as a process of first choosing s distinct coordinates from $[n]$ such that each coordinate is used a non-zero even number of times when forming a multi-set of pd numbers. Naturally, we have $1 \leq s \leq \lfloor pd/2 \rfloor \leq n$ and $s(\alpha) = s$. For a fixed choice of s coordinates, $\sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}} \binom{pd}{2\beta_1,\ldots,2\beta_s}$ sums over all valid multi-sets involving s entries of $[n]$. However, since every permutation of a fixed multi-set corresponds to a possibly repeated α's, we divide by $\binom{d}{\alpha_1,\ldots,\alpha_{n^p}}$[46]. Finally, each such α is then scaled by $\left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right)$. □

Example illustrating Claim 34

We illustrate the counting process with an example where $p = 2$, $n = 2$, and $d = 3$. Denote $\alpha, \beta, \gamma \in [n]^2$ as three distinct coordinates of $\boldsymbol{Y}$. By picking entries $\{\boldsymbol{Y}_\alpha, \boldsymbol{Y}_\beta, \boldsymbol{Y}_\gamma\}$, the corresponding Hermite polynomial $h_1(\boldsymbol{Y}_\alpha)h_1(\boldsymbol{Y}_\beta)h_1(\boldsymbol{Y}_\gamma) = \boldsymbol{Y}_\alpha \boldsymbol{Y}_\beta \boldsymbol{Y}_\gamma$ is multi-linear. With repeated entries such as $\{\boldsymbol{Y}_\alpha, \boldsymbol{Y}_\alpha, \boldsymbol{Y}_\beta\}$ and $\{\boldsymbol{Y}_\alpha, \boldsymbol{Y}_\alpha, \boldsymbol{Y}_\alpha\}$, the corresponding Hermite polynomials are $h_2(\boldsymbol{Y}_\alpha)h_1(\boldsymbol{Y}_\beta)$ and $h_3(\boldsymbol{Y}_\alpha)$ respectively.

By the constraint of $\mathbb{1}_{even(c(\alpha))}$, it suffices to only consider choices such that there are an even number of 1's and 2's. Ignoring permutations, there are 10 such selections. Including permutations, there are $\binom{3}{3} \cdot 2 + \binom{3}{2,1} \cdot 6 + \binom{3}{1,1,1} \cdot 2 = 32$ such selections. Note that only the last 2 are multi-linear.

1 distinct: $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{11}, \boldsymbol{Y}_{11}\}$, $\{\boldsymbol{Y}_{22}, \boldsymbol{Y}_{22}, \boldsymbol{Y}_{22}\}$

2 distinct: $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{11}, \boldsymbol{Y}_{22}\}$, $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \boldsymbol{Y}_{12}\}$, $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{21}, \boldsymbol{Y}_{21}\}$, $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{22}, \boldsymbol{Y}_{22}\}$, $\{\boldsymbol{Y}_{12}, \boldsymbol{Y}_{12}, \boldsymbol{Y}_{22}\}$, $\{\boldsymbol{Y}_{21}, \boldsymbol{Y}_{21}, \boldsymbol{Y}_{22}\}$

3 distinct: $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \boldsymbol{Y}_{21}\}$, $\{\boldsymbol{Y}_{12}, \boldsymbol{Y}_{21}, \boldsymbol{Y}_{22}\}$

[43] E.g. We can identify the polynomial $\boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{21}\boldsymbol{Y}_{11}$ with the multi-set of its indices $\{1, 1, 1, 1, 1, 1, 2, 2\}$.

[44] E.g. $\boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{21}\boldsymbol{Y}_{11}$, $\boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{12}\boldsymbol{Y}_{11}$, and $\boldsymbol{Y}_{11}\boldsymbol{Y}_{11}\boldsymbol{Y}_{11}\boldsymbol{Y}_{22}$ all map to $\{1, 1, 1, 1, 1, 1, 2, 2\}$.

[45] E.g. $(1, 1, 1, 2, 2, 1, 1, 1) \equiv \boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{21}\boldsymbol{Y}_{11}$ and $(1, 1, 1, 2, 1, 2, 1, 1) \equiv \boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{12}\boldsymbol{Y}_{11}$ are counted differently.

[46] E.g. Suppose $\beta_1 = 3$, $\beta_2 = 1$ and $pd = 8$ in the combinatorial summation. $\binom{pd}{2\beta_1,\ldots,2\beta_s}$ will include permutations such as $(1, 1, 1, 2, 1, 2, 1, 1)$ and $(1, 1, 1, 1, 1, 2, 1, 2)$. However, both of $(1, 1, 1, 2, 1, 2, 1, 1)$ and $(1, 1, 1, 1, 1, 2, 1, 2)$ actually refer to the same α term since $\boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{12}\boldsymbol{Y}_{11} = \boldsymbol{Y}_{11}\boldsymbol{Y}_{11}\boldsymbol{Y}_{12}\boldsymbol{Y}_{12}$.

We first compute the summation on the left hand side of Claim 34. An α with 1 distinct entry such as $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{11}, \boldsymbol{Y}_{11}\}$ contributes $\left(\frac{k}{n}\right)^2 \frac{1}{3!}$ to the summation. With 2 distinct entries, such as $\{\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \boldsymbol{Y}_{12}\}$, we get $\left(\frac{k}{n}\right)^4 \frac{1}{1!2!}$. Finally, each multi-linear polynomial contributes $\left(\frac{k}{n}\right)^4 \frac{1}{1!1!1!}$. So,

$$\sum_{|\alpha|=d} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right) = \frac{2}{3!}\left(\frac{k}{n}\right)^2 + \frac{6}{1!2!}\left(\frac{k}{n}\right)^4 + \frac{2}{1!1!1!}\left(\frac{k}{n}\right)^4 = \frac{1}{3}\left(\frac{k}{n}\right)^2 + 5\left(\frac{k}{n}\right)^4$$

On the right hand side of Claim 34, we count by viewing the selection of 3 entries of Y as filling up $pd = 6$ slots with values from $\{1, 2\}$:

$$\begin{aligned}
&\frac{1}{d!} \sum_{s=1}^{\lfloor pd/2 \rfloor} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s = pd/2 \\ \beta_1 \neq 0, \ldots, \beta_s \neq 0}} \binom{pd}{2\beta_1, \ldots, 2\beta_s} \\
=&\frac{1}{3!} \sum_{s=1}^{2} \binom{2}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s = 3 \\ \beta_1 \neq 0, \ldots, \beta_s \neq 0}} \binom{6}{2\beta_1, \ldots, 2\beta_s} \\
=&\frac{1}{6}\binom{2}{1}\left(\frac{k}{n}\right)^2 \binom{6}{6} + \frac{1}{6}\binom{2}{2}\left(\frac{k}{n}\right)^4 \left[\binom{6}{2,4} + \binom{6}{4,2}\right] \\
=&\frac{1}{3}\left(\frac{k}{n}\right)^2 + 5\left(\frac{k}{n}\right)^4
\end{aligned}$$

Claim 35. *For $x > 0$ and $0 < a < 1$, we have $xa^x \leq \min\left\{x, \frac{1}{e \ln(1/a)}\right\}$.*

Proof. When $0 < a < 1$, we have $xa^x \leq x$ trivially. For $x > 0$ and $0 < a < 1$, we see that $(1/a)^x > 0$. So,

$$\left(\frac{1}{a}\right)^x \geq e \ln\left(\frac{1}{a}\right)^x = ex \ln\left(\frac{1}{a}\right) \iff xa^x \leq \frac{1}{e \ln\left(\frac{1}{a}\right)}$$

Thus, $xa^x \leq \min\left\{x, \frac{1}{e \ln(1/a)}\right\}$. □

Lemma 28. *Let $p \geq 2$, $d \geq 1$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. An entry of $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ can be indexed by either an integer from $[n^p]$ or a p-tuple. Define $\phi : [n^p] \to [n]^p$, α, $c(\alpha)$, $s(\alpha)$, and $\mathbb{1}_{even(c(\alpha))}$ as follows:*

- *$\phi(i)$ maps to a p-tuple indicating the p (possibly repeated) entries of $\boldsymbol{x}$ that are used.*
- *$\alpha = (\alpha_1, \ldots, \alpha_{n^p})$ is an n^p-tuple that corresponds to a Hermite polynomial of degree $|\alpha| = \sum_{i=1}^{n^p} \alpha_i$. For each i, α_i is the number of times entry $\boldsymbol{Y}_{\phi(i)}$ was chosen, where each $\boldsymbol{Y}_{\phi(i)}$ references p coordinates of $\boldsymbol{x}$.*
- *$c(\alpha) = (c_1, \ldots, c_n)$, where c_j is the number of times $\boldsymbol{x}_j$ is used in α.*
- *$s(\alpha)$ is the number of distinct non-zero $\boldsymbol{x}_j$'s in $c(\alpha)$.*
- *$\mathbb{1}_{even(c(\alpha))}$ be the indicator whether* all *c_j's are even.*

Under these definitions, we have the following:

$$(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}))^2 = \lambda^{2d} k^{-pd} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right).$$

Proof of Lemma 28. For fixed multi-index $\alpha = (\alpha_1, \dots, \alpha_{n^p})$ such that $|\alpha| = d$, we now compute $\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y})$.

$$\begin{aligned}
&\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}) && \\
=&\mathbb{E}_{H_1} \Pi_{i=1}^{n^p} h_{\alpha_i}(\boldsymbol{Y}_{\phi(i)}) && \text{Product of Hermite polys} \\
=&\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{W}_{\phi(i)} \sim N(0,1)} \Pi_{i=1}^{n^p} h_{\alpha_i}(\boldsymbol{Y}_{\phi(i)}) && \text{Definition of } H_1 \\
=&\mathbb{E}_{\boldsymbol{x}} \Pi_{i=1}^{n^p} \mathbb{E}_{\boldsymbol{W}_{\phi(i)} \sim N(0,1)} h_{\alpha_i}(\boldsymbol{Y}_{\phi(i)}) && \text{Independence of } \boldsymbol{W} \text{ entries} \\
=&\mathbb{E}_{\boldsymbol{x}} \Pi_{i=1}^{n^p} \mathbb{E}_{\boldsymbol{W}_{\phi(i)} \sim N(0,1)} h_{\alpha_i}(\boldsymbol{W}_{\phi(i)} + \lambda \Pi_{j \in \phi(i)} x_j) && \text{Definition of } \boldsymbol{Y}_{\alpha_i} \\
=&\mathbb{E}_{\boldsymbol{x}} \Pi_{i=1}^{n^p} \mathbb{E}_{\boldsymbol{z} \sim N(\lambda \Pi_{j \in \phi(i)} x_j, 1)} h_{\alpha_i}(z) && \text{Translation property of Hermite} \\
=&\mathbb{E}_{\boldsymbol{x}} \Pi_{i=1}^{n^p} \sqrt{\frac{1}{(\alpha_i)!}} (\lambda \Pi_{j \in \phi(i)} x_j)^{\alpha_i} && \text{Expectation of deg } \alpha_i \text{ Hermite on } \boldsymbol{z} \sim (\mu, 1) \\
=&\left(\Pi_{i=1}^{n^p} \sqrt{\frac{1}{(\alpha_i)!}} \right) \lambda^d \mathbb{E}_{\boldsymbol{x}} \Pi_{i=1}^{n^p} \Pi_{j \in \phi(i)} x_j^{\alpha_i} && \sum_i^{n^p} \alpha_i = |\alpha| = d \\
=&\left(\Pi_{i=1}^{n^p} \sqrt{\frac{1}{(\alpha_i)!}} \right) \lambda^d \mathbb{E}_{\boldsymbol{x}} \Pi_{j=1}^{n} x_j^{c_j} && \text{Definition of } c(\alpha) = (c_1, \dots, c_n) \\
=&\left(\Pi_{i=1}^{n^p} \sqrt{\frac{1}{(\alpha_i)!}} \right) \lambda^d \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n} \right)^{s(\alpha)} k^{-\frac{pd}{2}} &&
\end{aligned}$$

The last equality is because $\mathbb{E}_{\boldsymbol{x}} \Pi_{j=1}^{n} x_j^{c_j} = 0$ if there is an odd c_j. So, for $|\alpha| = d$,

$$(\mathbb{E}_{H_1} h_\alpha(\boldsymbol{Y}))^2 = \lambda^{2d} k^{-pd} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n} \right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!} \right)$$

□

Lemma 29. *Let $p \geq 2$, $1 \leq D \leq 2n/p$, $\boldsymbol{Y} \in \otimes^p \mathbb{R}^n$ be an observation tensor, $\boldsymbol{x}$ be a k-sparse scaled Rademacher vector, and $\{h_\alpha\}_\alpha$ be the set of normalized probabilists' Hermite polynomials. Then,*

$$\sum_{|\alpha| \leq D} (\mathbb{E}_{H_1}[h_\alpha(\boldsymbol{Y})])^2 \leq \sum_{d=1}^{D} \frac{\lambda^{2d}}{d!} \sum_{s=1}^{pd/2} \left(\frac{ek^2}{sn} \right)^s \left(\frac{s}{k} \right)^{pd}.$$

Proof of Lemma 29. To upper bound $\sum_{|\alpha| \leq D} (\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})])^2$, we use an equality that relates the counting of $\boldsymbol{Y}$ entries with coordinates of the signal $\boldsymbol{x}$. For a fixed d, it can be shown (see Claim 34) that

$$\sum_{|\alpha|=d} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n} \right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!} \right) = \frac{1}{d!} \sum_{s=1}^{\lfloor pd/2 \rfloor} \binom{n}{s} \left(\frac{k}{n} \right)^{2s} \sum_{\substack{\beta_1 + \ldots + \beta_s = pd/2 \\ \beta_1 \neq 0, \ldots, \beta_s \neq 0}} \binom{pd}{2\beta_1, \dots, 2\beta_s}$$

This allows us to perform combinatoric arguments on the coordinates of the signal $\boldsymbol{x}$ instead of over the tensor coordinates of $\boldsymbol{Y}$.

$$
\begin{aligned}
&\sum_{|\alpha|\leq D} \left(\mathbb{E}_{H_1}[f_\alpha(\boldsymbol{Y})]\right)^2 \\
&= \sum_{d=1}^{D} \sum_{|\alpha|=d} \lambda^{2d} k^{-pd} \mathbb{1}_{even(c(\alpha))} \left(\frac{k}{n}\right)^{2s(\alpha)} \left(\Pi_{i=1}^{n^p} \frac{1}{(\alpha_i)!}\right) && \text{From above} \\
&= \sum_{d=1}^{D} \lambda^{2d} k^{-pd} \frac{1}{d!} \sum_{s=1}^{\lfloor pd/2 \rfloor} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}} \binom{pd}{2\beta_1,\ldots,2\beta_s} && \text{Claim 34} \\
&\leq \sum_{d=1}^{D} \frac{\lambda^{2d} k^{-pd}}{d!} \sum_{s=1}^{pd/2} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\substack{\beta_1+\ldots+\beta_s=pd/2\\ \beta_1\neq 0,\ldots,\beta_s\neq 0}} \binom{pd}{2\beta_1,\ldots,2\beta_s} && \text{Drop floor} \\
&\leq \sum_{d=1}^{D} \frac{\lambda^{2d} k^{-pd}}{d!} \sum_{s=1}^{pd/2} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\beta_1+\ldots+\beta_s=pd/2} \binom{pd}{2\beta_1,\ldots,2\beta_s} && \text{Drop } \beta_i \neq 0 \\
&\leq \sum_{d=1}^{D} \frac{\lambda^{2d} k^{-pd}}{d!} \sum_{s=1}^{pd/2} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} \sum_{\gamma_1+\ldots+\gamma_s=pd} \binom{pd}{\gamma_1,\ldots,\gamma_s} && \text{Drop ``evenness constraint''} \\
&= \sum_{d=1}^{D} \frac{\lambda^{2d} k^{-pd}}{d!} \sum_{s=1}^{pd/2} \binom{n}{s} \left(\frac{k}{n}\right)^{2s} s^{pd} && \text{Multinomial theorem} \\
&\leq \sum_{d=1}^{D} \frac{\lambda^{2d} k^{-pd}}{d!} \sum_{s=1}^{pd/2} \left(\frac{ek^2}{sn}\right)^{s} s^{pd} && \binom{n}{s} \leq \left(\frac{en}{s}\right)^s \\
&= \sum_{d=1}^{D} \frac{\lambda^{2d}}{d!} \sum_{s=1}^{pd/2} \left(\frac{ek^2}{sn}\right)^{s} \left(\frac{s}{k}\right)^{pd}
\end{aligned}
$$

□

Lemma 30. *For $p \geq 2$, $d \geq 1$, $1 \leq k \leq n$ and $1 \leq s \leq pd/2$, we have*

$$
\left(\frac{ek^2}{sn}\right)^{s} \left(\frac{s}{k}\right)^{pd} \leq \left[\frac{2pd}{\min\left\{\sqrt{npd},\ k\left(1+\left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd}.
$$

Proof of Lemma 30. We will first push all terms into $[\cdots]^{pd}$ and then upper bound the terms inside[47]. We start by recalling three useful inequalities:

- For $x > 0$, we have $x^{\frac{1}{x}} \leq 2$.
- For $x > 0$ and $0 < a < 1$, we have $xa^x \leq \min\left\{x, \frac{1}{e\ln(1/a)}\right\}$.
- For $x \geq \frac{1}{e}$, we have $\min\left\{\frac{1}{2}, \frac{1}{e\ln x}\right\} \leq \frac{1}{1+|\ln(x)|}$.

Using the first inequality, we get

$$
\left(\frac{ek^2}{sn}\right)^{s} \left(\frac{s}{k}\right)^{pd} = \left[\left(\frac{ek^2}{sn}\right)^{\frac{s}{pd}} \frac{s}{k}\right]^{pd} = \left[\left(\frac{ek^2}{npd}\right)^{\frac{s}{pd}} \left(\frac{pd}{s}\right)^{\frac{s}{pd}} \frac{s}{k}\right]^{pd} \leq \left[2\left(\frac{ek^2}{npd}\right)^{\frac{s}{pd}} \frac{s}{k}\right]^{pd}
$$

[47] This works because the terms inside are greater than 0 and $pd \geq 1$.

When $ek^2 \geq npd$, we use $s \leq pd/2$ to get

$$\left(\frac{ek^2}{npd}\right)^{\frac{s}{pd}} \frac{s}{k} \leq \left(\frac{ek^2}{npd}\right)^{\frac{pd/2}{pd}} \frac{pd/2}{k} = \sqrt{\frac{epd}{4n}} \leq \sqrt{\frac{pd}{n}}$$

When $ek^2 < npd$, we use the second and third inequalities[48] to get

$$\left(\frac{ek^2}{npd}\right)^{\frac{s}{pd}} \frac{s}{k} = \frac{s}{pd} \left(\frac{ek^2}{npd}\right)^{\frac{s}{pd}} \frac{pd}{k} \leq \min\left\{\frac{1}{2}, \frac{1}{e \ln\left(\frac{npd}{ek^2}\right)}\right\} \cdot \frac{pd}{k} \leq \frac{pd}{k\left(1 + \left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)}$$

Putting together, we see that

$$\left(\frac{ek^2}{sn}\right)^s \left(\frac{s}{k}\right)^{pd} \leq \left[2 \max\left\{\sqrt{\frac{pd}{n}}, \frac{pd}{k\left(1 + \left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)}\right\}\right]^{pd} = \left[\frac{2pd}{\min\left\{\sqrt{npd},\ k\left(1 + \left|\ln\left(\frac{npd}{ek^2}\right)\right|\right)\right\}}\right]^{pd}$$

□

F.4 Proofs for information-theoretic lower bound

Lemma 32. *Let U_k be the set of k-sparse flat unit vectors and $N(U_k, \|\cdot\|_2, \epsilon)$ be the ϵ-covering number of U_k with respect to Euclidean distance. For $\epsilon \in (0, 1]$ and $n \geq 2k$,*

$$N(U_k, \|\cdot\|_2, \epsilon) \geq \left(\frac{n-k}{k}\right)^{k\left(1 - \frac{\epsilon^2}{2}\right)}.$$

Proof of Lemma 32. For $x, x' \in U_k$, let us denote $\alpha = |\{i \in [n] : i \in \mathcal{I}_x \cap \mathcal{I}_{x'} \text{ and } x_i = x'_i\}|$ be the intersecting indices with agreeing signs, $\beta = |\{i \in [n] : i \in \mathcal{I}_x \cap \mathcal{I}_{x'} \text{ and } x_i = -x'_i\}|$ be the intersecting indices with disagreeing signs, and $\gamma = |\{i \in [n] : i \notin \mathcal{I}_x \cap \mathcal{I}_{x'}\}|$ be the non-intersecting indices. By definition, $\alpha \geq 0$, $\beta \geq 0$, $\gamma \geq 0$, $\alpha + \beta = |\mathcal{I}_x \cap \mathcal{I}_{x'}|$, $\alpha + \beta + \gamma = 2k - |\mathcal{I}_x \cap \mathcal{I}_{x'}| = |\mathcal{I}_x| + |\mathcal{I}_{x'}| - |\mathcal{I}_x \cap \mathcal{I}_{x'}|$, and $\gamma = 2(k - |\mathcal{I}_x \cap \mathcal{I}_{x'}|)$. Then, for $x, x' \in U_k$,

$$\|x - x'\|_2 = \sqrt{\beta\left(\frac{2}{\sqrt{k}}\right)^2 + \gamma\left(\frac{1}{\sqrt{k}}\right)^2} = \sqrt{\frac{4\beta + \gamma}{k}} \geq \sqrt{\frac{\gamma}{k}} = \sqrt{2 - \frac{2|\mathcal{I}_x \cap \mathcal{I}_{x'}|}{k}}$$

So, $\|x - x'\|_2 \leq \epsilon$ implies that $|\mathcal{I}_x \cap \mathcal{I}_{x'}| \geq k(1 - \frac{\epsilon^2}{2})$. This means that for any *fixed* $x \in U_k$, there are at most[49] $\sum_{i=0}^{\lfloor \epsilon^2 k/2 \rfloor} \binom{k}{i}\binom{n-k}{i}$ vectors in U_k (including x itself) that are of distance at most ϵ from x. By definition of covering number, we know that

$$N(U_k, \|\cdot\|_2, \epsilon) \cdot \sum_{i=0}^{\lfloor \epsilon^2 k/2 \rfloor} \binom{k}{i}\binom{n-k}{i} \geq |U_k| = 2^k \binom{n}{k}$$

Thus, to argue that $N(U_k, \|\cdot\|_2, \epsilon) \geq \left(\frac{n-k}{k}\right)^{k\left(1 - \frac{\epsilon^2}{2}\right)}$, it suffices to show

$$2^k \binom{n}{k} \geq \left(\frac{n-k}{k}\right)^{k\left(1 - \frac{\epsilon^2}{2}\right)} \cdot \sum_{i=0}^{\lfloor \epsilon^2 k/2 \rfloor} \binom{k}{i}\binom{n-k}{i}$$

[48] Observe that $0 < \frac{s}{pd} \leq \frac{1}{2}$, $0 < \frac{ek^2}{npd} < 1$, and $\frac{1}{e} \leq 1 < \frac{npd}{ek^2}$.

[49] First pick i out of k coordinates of x to be different, then pick the i different coordinates amongst the $n - k$ coordinates outside of $\mathcal{I}_x$. The summation is from 0 to $\lfloor \epsilon^2 k/2 \rfloor$ because we need to have $|\mathcal{I}_x \cap \mathcal{I}_{x'}| \geq k(1 - \frac{\epsilon^2}{2})$.

Observe that since $n \geq 2k$, the term $\binom{n-k}{i}$ increases with i:

$$\begin{aligned}
\sum_{i=0}^{\lfloor \epsilon^2 k/2 \rfloor} \binom{k}{i}\binom{n-k}{i} &\leq \binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor} \sum_{i=0}^{\lfloor \epsilon^2 k/2 \rfloor} \binom{k}{i} && (\star) \\
&\leq \binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor} \sum_{i=0}^{k} \binom{k}{i} && \lfloor \epsilon^2 k/2 \rfloor \leq k \\
&= \binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor} 2^k && \text{Binomial theorem}
\end{aligned}$$

where $(\star)$ is because $n \geq 2k$ and $\epsilon \in (0,1]$ implies that $n-k \geq \epsilon^2 k$ so $\binom{n-k}{i} \leq \binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor}$ for $0 \leq i \leq \lfloor \epsilon^2 k/2 \rfloor$. Thus, it suffices to show

$$\binom{n}{k} \geq \left(\frac{n-k}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)} \cdot \binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor}$$

We will now show that $\frac{\binom{n}{k}}{\binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor}} \geq \left(\frac{n-k}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)}$:

$$\begin{aligned}
\frac{\binom{n}{k}}{\binom{n-k}{\lfloor \epsilon^2 k/2 \rfloor}} &= \frac{n!}{k!(n-k)!} \frac{(\lfloor \epsilon^2 k/2 \rfloor)!(n-k-\lfloor \epsilon^2 k/2 \rfloor)!}{(n-k)!} \\
&= \frac{n!}{(n-k)!} \frac{(\lfloor \epsilon^2 k/2 \rfloor)!}{k!} \frac{(n-k-\lfloor \epsilon^2 k/2 \rfloor)!}{(n-k)!} \\
&= [(n) \cdot \ldots \cdot (n-k+1)] \cdot \left[\frac{1}{(k) \cdot \ldots \cdot (k - \lfloor \epsilon^2 k/2 \rfloor + 1)}\right] \cdot \\
&\quad \left[\frac{1}{(n-k) \cdot \ldots \cdot (n-k-\lfloor \epsilon^2 k/2 \rfloor + 1)}\right] \\
&\geq (n-k)^k \left(\frac{1}{k}\right)^{\lfloor \epsilon^2 k/2 \rfloor} \left(\frac{1}{n-k}\right)^{\lfloor \epsilon^2 k/2 \rfloor} \\
&\geq (n-k)^k \left(\frac{1}{k}\right)^{\frac{\epsilon^2 k}{2}} \left(\frac{1}{n-k}\right)^{\frac{\epsilon^2 k}{2}} \\
&\geq (n-k)^k \left(\frac{1}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)} \left(\frac{1}{n-k}\right)^{\frac{\epsilon^2 k}{2}} \\
&= \left(\frac{n-k}{k}\right)^{k\left(1-\frac{\epsilon^2}{2}\right)}
\end{aligned}$$

where the last inequality is because $\epsilon \leq 1$ implies that $1 - \frac{\epsilon^2}{2} \geq \frac{\epsilon^2}{2}$. □

Lemma 33. *Denote $\mathcal{S}_k^{n-1}$ as the set of k-sparse unit vectors. Then,*

$$\max_{u,v \in \mathcal{S}_k^{n-1}} D_{KL}\left(\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|v}\right) \leq 2\lambda^2$$

where $D_{KL}(\cdot \| \cdot)$ is the KL-divergence function and $\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u}$ is the probability distribution of observing $\boldsymbol{Y}$ from signal u with additive standard Gaussian noise tensor $\boldsymbol{W}$.

Proof of Lemma 33. Define $vec(T)$ as vectorization of a tensor from $\otimes^p \mathbb{R}^n$ to $\mathbb{R}^{n^p}$. Then, for $u \in \mathcal{S}_k^{n-1}$, we see that $\|vec(\lambda u^{\otimes p})\|_2^2 = \lambda^2$ and the distribution $\mathbb{P}_{\boldsymbol{Y} \sim \mathcal{Y}|u}$ follows the distribution of a Gaussian vector $\boldsymbol{g} \sim N(vec(\lambda u^{\otimes p}), I_{n^p})$.

For two Gaussian vectors $\boldsymbol{g} \sim N(\mu_0, I_{n^p})$ and $\boldsymbol{h} \sim N(\mu_1, I_{n^p})$, we know that $D_{KL}(\boldsymbol{g} \| \boldsymbol{h}) = \frac{1}{2}(\mu_1 - \mu_0)^\top (\mu_1 - \mu_0) = \frac{1}{2}\|\mu_1 - \mu_0\|_2^2 \leq \frac{1}{2}(\|\mu_1\|_2 + \|\mu_0\|_2)^2$ by triangle inequality[50].

[50] We get an equality if $\mu_1 = -\mu_0$.

Thus, $D_{KL}\left(\mathbb{P}_{\boldsymbol{Y}\sim\mathcal{Y}|u} \,\middle\|\, \mathbb{P}_{\boldsymbol{Y}\sim\mathcal{Y}|v}\right) \leq \frac{1}{2}(\lambda^2+\lambda^2)^2 = 2\lambda^2$. □