Proceedings of Machine Learning Research vol 134:1–27, 2021 34th Annual Conference on Learning Theory

Hypothesis testing with low-degree polynomials in the Morris class of exponential families

Dmitriy Kunisky KUNISKY@CIMS.NYU.EDU
Department of Mathematics, Courant Institute of Mathematical Sciences, New York University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Analysis of low-degree polynomial algorithms is a powerful, newly-popular method for predicting computational thresholds in hypothesis testing problems. One limitation of current techniques for this analysis is their restriction to Bernoulli and Gaussian distributions. We expand this range of possibilities by performing the low-degree analysis of hypothesis testing for the Morris class of natural exponential families with quadratic variance function, giving a unified treatment of Gaussian, Poisson, gamma (including exponential and chi-squared), binomial (including Bernoulli), negative binomial (including geometric), and generalized hyperbolic secant distributions. We then give several algorithmic applications.

1. In models where a random signal is observed through coordinatewise-independent noise applied in an exponential family, the success or failure of low-degree polynomials is governed by the *z-score overlap*, the inner product of z-score vectors with respect to the null distribution of two independent copies of the signal.

2. In the same models, testing with low-degree polynomials exhibits *channel monotonicity*: the above distributions admit a total ordering by computational cost of hypothesis testing, according to a scalar parameter describing how the variance depends on the mean in an exponential family.

3. In a spiked matrix model with a particular non-Gaussian noise distribution, the low-degree prediction is incorrect unless polynomials with arbitrarily large degree in individual matrix entries are permitted. This shows that polynomials summing over self-avoiding walks and variants thereof, as proposed recently by Ding, Hopkins, and Steurer (2020) for spiked matrix models with heavy-tailed noise, are strictly suboptimal for this model. Thus low-degree polynomials appear to offer a tradeoff between robustness and strong performance fine-tuned to specific models. Inspired by this, we suggest that a class of problems requiring *exploration before inference*, where an algorithm must first examine the input and then use some intermediate computation to choose a suitable inference subroutine, appears especially difficult for low-degree polynomials.

Keywords: low-degree polynomials, hypothesis testing, exponential families, computational complexity, computational phase transitions

1. Introduction

A powerful framework has emerged recently for making predictions of and producing evidence for computational thresholds in high-dimensional average-case algorithmic problems, which analyzes the performance of algorithms that compute low-degree polynomials. Originating in results on sum-of-squares optimization (Barak et al., 2019; Hopkins et al., 2017; Hopkins and Steurer, 2017; Hopkins, 2018), this idea has since been fruitfully applied to a wide range of problems (Bandeira et al., 2020b; Ding et al., 2019; Brennan and Bresler, 2020; Gamarnik et al., 2020; Schramm and Wein, 2020; Bandeira et al., 2020a; Wein, 2020). Perhaps the main attraction of this so-called *low-degree method* is the simplicity of low-degree polynomial algorithms, which are both quite natural

and often easier to study than other putatively-optimal families of algorithms like convex relaxations and message-passing techniques.

Before applying the low-degree method in a given situation, we must decide how exactly to make sense of "a low-degree polynomial succeeding" at a given computational task. Besides just producing a sensible definition for a given task—different definitions have been used in the literature for hypothesis testing (Hopkins, 2018; Kunisky et al., 2019), statistical estimation (Hopkins and Steurer, 2017; Ding et al., 2020a; Schramm and Wein, 2020), and optimization (Gamarnik et al., 2020; Wein, 2020)—practitioners also sometimes make choices in this step to keep the low-degree method analytically tractable. On the other hand, the problems studied to date with the low-degree method involve only a few different families of probability distributions over their inputs, and often the low-degree analysis takes advantage of an intimate coupling between simplifications made in the success criteria and convenient distribution-specific identities. Our goal will be to investigate the most common measures of efficacy for low-degree polynomial algorithms for a wider range of probability distributions, probing whether some of the success of the low-degree method might be due to this kind of fortuitous coincidence.

We focus on hypothesis testing: given two sequences of probability distributions $\mathbb{P}_n$ and $\mathbb{Q}_n$, we consider whether a low-degree polynomial can correctly distinguish $\boldsymbol{y} \sim \mathbb{P}_n$ from $\boldsymbol{y} \sim \mathbb{Q}_n$ with high probability as $n \to \infty$. The most direct way to formalize this question, by analogy with the classical Neyman-Pearson lemma (Neyman and Pearson, 1933), is to ask, for some degree bound $D(n)$ corresponding to a computational budget: *do there exist polynomials* $f_n \in \mathbb{R}[\boldsymbol{y}]$ *with* $\mathsf{deg}(f(n)) \leq D(n)$ *and thresholds* $\xi_n \in \mathbb{R}$ *such that*

$$\lim_{n\to\infty} \mathbb{P}_n[f_n(\boldsymbol{y}) > \xi_n] = \lim_{n\to\infty} \mathbb{Q}_n[f_n(\boldsymbol{y}) < \xi_n] = 1 \text{ ?} \tag{1}$$

Unfortunately, it appears difficult to prove lower bounds against this class of algorithms. Instead, recent research has focused on the following "averaged" version of the above: *do there exist polynomials* $f_n \in \mathbb{R}[\boldsymbol{y}] \cap L^2(\mathbb{Q}_n)$ *with* $\mathsf{deg}(f(n)) \leq D(n)$ *such that*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathbb{P}_n} f_n(\boldsymbol{y}) = +\infty\,, \text{ while } \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathbb{Q}_n} f_n(\boldsymbol{y})^2 = 1 \text{ ?} \tag{2}$$

As we will see in Section 1.2, here the optimal f_n may be computed explicitly using orthogonal polynomials, at least for simple models of hypothesis testing. The resulting predictions have been remarkably consistent with other, more technically-challenging methods for models including the planted clique problem (Barak et al., 2019), the stochastic block model (Hopkins and Steurer, 2017), and spiked matrix and tensor models (Hopkins, 2018; Kunisky et al., 2019; Bandeira et al., 2020b; Ding et al., 2019).

Yet, our current understanding of this heuristic and the circumstances under which it is accurate remains incomplete in the regard mentioned above: these examples all involve observations $\boldsymbol{y}$ having only Gaussian or Bernoulli distributions. Moreover, the low-degree analysis often hinges on special algebraic and analytic properties of these distributions and their orthogonal polynomials. In this paper, we will describe the results of applying the low-degree method to hypothesis testing in many further exponential families, making new predictions and suggesting some challenging examples that we hope will stimulate further research in this direction.

1.1. Natural exponential families with quadratic variance function

We first introduce the distributions we will study, which form *exponential families*. Throughout this section we follow the presentation of the seminal papers of Morris (1982, 1983), which first recognized the many shared statistical properties of these families. We start by recalling the basic notions of exponential families.

Definition 1 *Let ρ_0 be a probability measure over $\mathbb{R}$ which is not a single atom. Let $\psi(\theta) := \log \mathbb{E}_{x\sim\rho_0}[\exp(\theta x)]$ and $\Theta := \{\theta \in \mathbb{R} : \psi(\theta) < \infty\}$. Then, the* natural exponential family (NEF) generated by ρ_0 *is the family of probability measures ρ_θ, for $\theta \in \Theta$, given by*

$$d\rho_\theta(x) := \exp(\theta x - \psi(\theta))d\rho_0(x). \tag{3}$$

Sometimes, the "natural parameter" θ is the mean of ρ_θ or a translation thereof; however, as the next example shows, the mapping $\theta \mapsto \mathbb{E}_{x\sim\rho_\theta}[x]$ need not be particularly simple in general.

Example 1 *Taking $d\rho_0(x) = e^{-x}\mathbb{1}\{x \geq 0\}dx$, we have $\Theta = (-\infty, 1)$, and this generates the NEF of exponential distributions, $d\rho_\theta(x) = (1-\theta)e^{-(1-\theta)x}\mathbb{1}\{x \geq 0\}dx$. The mean of ρ_θ is $\mathbb{E}_{x\sim\rho_\theta}[x] = \frac{1}{1-\theta}$.*

Nonetheless, it is always possible to reparametrize any NEF in terms of the mean in the following way. The cumulant generating functions of the ρ_θ are merely translations of ψ, $\psi_\theta(\eta) := \log \mathbb{E}_{x\sim\rho_\theta}[\exp(\eta x)] = \psi(\theta+\eta) - \psi(\theta)$. Therefore, the means and variances of ρ_θ are

$$\mu_\theta := \mathbb{E}_{x\sim\rho_\theta}[x] = \psi'_\theta(0) = \psi'(\theta), \tag{4}$$

$$\sigma^2_\theta := \mathbb{E}_{x\sim\rho_\theta}[x^2] - (\mathbb{E}_{x\sim\rho_\theta}[x])^2 = \psi''_\theta(0) = \psi''(\theta). \tag{5}$$

Since ρ_0 is not an atom, neither is any ρ_θ, and thus $\psi''(\theta) = \sigma^2_\theta > 0$ for all $\theta \in \mathbb{R}$. Therefore, ψ' is strictly increasing, and thus one-to-one. Letting $\Omega \subseteq \mathbb{R}$ equal the image of $\mathbb{R}$ under ψ' (some open interval, possibly infinite on either side, of $\mathbb{R}$), we see that ρ_θ admits an alternative mean parametrization, as follows.

Definition 2 *If ρ_0 generates the NEF ρ_θ, then we let $\widetilde{\rho}_\mu = \rho_{(\psi')^{-1}(\mu)}$ over $\mu \in \Omega$. The* mean-parametrized NEF generated by ρ_0 *is the family of probability measures $\widetilde{\rho}_\mu$, for $\mu \in \Omega$.*

By the same token, within an NEF, the variance is a function of the mean. In the above setting, we denote this function as follows.

Definition 3 *For $\mu \in \Omega$, define the* variance function $V(\mu) := \sigma^2_{(\psi')^{-1}(\mu)} = \psi''((\psi')^{-1}(\mu))$.

The function $V(\mu)$ is simple for many NEFs that are theoretically important, and its simplicity appears to be a better measure of the "canonicity" of an NEF than, e.g., the simplicity of the probability density or mass function. Specifically, the most important NEFs have $V(\mu)$ a low-degree polynomial: $V(\mu)$ is constant only for the Gaussian NEF with some fixed variance, and linear only for the Poisson NEF and affine transformations thereof.

The situation becomes more interesting for $V(\mu)$ quadratic, which NEFs Morris gave the following name.

Definition 4 *If $V(\mu) = v_0 + v_1\mu + v_2\mu^2$ for some $v_i \in \mathbb{R}$, then we say that ρ_0 generates a* natural exponential family with quadratic variance function (NEF-QVF).

Name	$d\rho_0(x)$	**Support**	$V(\mu)$
Gaussian (variance $\sigma^2 > 0$)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}x^2)dx$	$\mathbb{R}$	σ^2
Poisson	$\frac{1}{e}\frac{1}{x!}$	$\mathbb{Z}_{\geq 0}$	μ
Gamma (shape $\alpha > 0$)	$\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}dx$	$(0, +\infty)$	$\frac{1}{\alpha}\mu^2$
Binomial (m trials)	$\frac{1}{2^m}\binom{m}{x}$	$\{0, \dots, m\}$	$-\frac{1}{m}\mu^2 + \mu$
Negative Binomial (m successes)	$\frac{1}{2^{m+x}}\binom{x+m-1}{x}$	$\mathbb{Z}_{\geq 0}$	$\frac{1}{m}\mu^2 + \mu$
Hyperbolic Secant (shape $r > 0$)	(Morris (1982), Section 5)	$\mathbb{R}$	$\frac{1}{r}\mu^2 + r$

Table 1: **The six basic NEF-QVFs.** We describe the six natural exponential families with quadratic variance function from which, according to the results of Morris (1982), any such family can be generated by an affine transformation. The sixth "generalized hyperbolic secant" family is more complicated to describe, but one representative distribution generating the $r = 1$ family has density $\frac{1}{2}\operatorname{sech}(\frac{\pi}{2}x)dx$, and may be thought of as a smoothed Laplace distribution.

NEF-QVFs are also sometimes called the *Morris class* of exponential families. One of the main results of Morris (1982) is a complete classification of the NEF-QVFs, as follows.

Proposition 5 *Any NEF-QVF can be obtained by an affine transformation ($X \mapsto aX + b$ applied to the underlying random variables) of one of the six families listed in Table 1. Conversely, any affine transformation of an NEF-QVF yields another NEF-QVF.*

We will study *hypothesis testing* in high-dimensional products of NEF-QVFs. That is, we seek to distinguish with high probability two sequences of distributions: the *null* distributions $\mathbb{Q}_n$, and the *planted* or *alternative* distributions $\mathbb{P}_n$. We consider two possible relationships between the two sequences: (1) *kin spiking*, where $\mathbb{P}_n$ belongs to the same NEF as $\mathbb{Q}_n$ but has a different mean, and (2) *additive spiking*, where $\mathbb{P}_n$ is a translation of $\mathbb{Q}_n$, possibly not belonging to the same NEF. Kin spiking will be mathematically more elegant, but additive spiking will allow us to treat a spiked matrix model similar to those studied in earlier works.

Definition 6 (Kin-spiked NEF-QVF model) *Let $\widetilde{\rho}_\mu$ be a mean-parametrized NEF-QVF over $\mu \in \Omega \subseteq \mathbb{R}$.[1] Let $N = N(n) \in \mathbb{N}$ and $\mu_{n,i} \in \Omega$ for each $n \in \mathbb{N}$ and $i \in [N(n)]$. Let $\mathcal{P}_n$ be a probability measure over $\Omega^{N(n)}$. Then, define sequences of probability measures $\mathbb{P}_n, \mathbb{Q}_n$ as follows:*

- *Under $\mathbb{Q}_n$, draw $y_i \sim \widetilde{\rho}_{\mu_{n,i}}$ independently for $i \in [N(n)]$.*
- *Under $\mathbb{P}_n$, draw $\boldsymbol{x} \sim \mathcal{P}_n$, and then draw $y_i \sim \widetilde{\rho}_{x_i}$ independently for $i \in [N(n)]$.*

Definition 7 (Additively-spiked NEF-QVF model) *In the same setting as Definition 6 but with $\mathcal{P}_n$ now a probability measure over $\mathbb{R}^{N(n)}$, define:*

- *Under $\mathbb{Q}_n$, draw $y_i \sim \widetilde{\rho}_{\mu_{n,i}}$ independently for $i \in [N(n)]$ (the same as above).*

1. It will not matter for our purposes what the base measure ρ_0 is.

- *Under $\mathbb{P}_n$, first draw $\boldsymbol{x} \sim \mathcal{P}_n$ and $z_i \sim \widetilde{\rho}_{\mu_{n,i}}$ independently for $i \in [N(n)]$, and observe $y_i = x_i + z_i$.*

In either case, we will focus on the problem of *strong detection*: given a particular time budget $T(n)$, does there exist a sequence of tests $f_n : \mathbb{R}^{N(n)} \to \{\mathtt{p}, \mathtt{q}\}$ computable in time $T(n)$ and having

$$\lim_{n\to\infty} \mathbb{Q}_n[f_n(\boldsymbol{y}) = \mathtt{q}] = \lim_{n\to\infty} \mathbb{P}_n[f_n(\boldsymbol{y}) = \mathtt{p}] = 1\,? \tag{6}$$

To the best of our knowledge, with the exception of negatively-spiked Gaussian Wishart models (Bandeira et al., 2020b,a), all previous applications of the low-degree method to hypothesis testing in the literature may be expressed as kin-spiked models in the Gaussian or Bernoulli NEF-QVFs.[2]

1.2. The low-degree likelihood ratio method

We now describe the calculation, based on the question (2) mentioned earlier, that we will take as a heuristic proxy for the difficulty of hypothesis testing. This will suggest that the problem of efficient strong detection described above may be addressed with computations involving the likelihood ratio,

$$L_n(\boldsymbol{y}) := \frac{d\mathbb{P}_n}{d\mathbb{Q}_n}(\boldsymbol{y}). \tag{7}$$

These techniques work in the Hilbert space $L^2(\mathbb{Q}_n)$, having inner product $\langle f, g\rangle := \mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_n}[f(\boldsymbol{y})g(\boldsymbol{y})]$ and associated norm $\|f\|^2 := \langle f, f\rangle$.

The following is the basic statement relating (2) to the likelihood ratio, which follows from a linear-algebraic calculation.

Proposition 8 (Hopkins and Steurer (2017); Hopkins et al. (2017); Hopkins (2018)) *Denote by $L_n^{\leq D}$ the orthogonal projection of L_n to the subspace of $L^2(\mathbb{Q}_n)$ consisting of polynomials having degree at most D. Then,*

$$\left\{ \begin{array}{ll} \textit{maximize} & \mathbb{E}_{\boldsymbol{y}\sim\mathbb{P}_n} f_n(\boldsymbol{y}) \\ \textit{subject to} & f_n \in \mathbb{R}[\boldsymbol{y}]_{\leq D} \cap L^2(\mathbb{Q}_n) \\ & \mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_n} f_n(\boldsymbol{y})^2 = 1 \end{array} \right\} = \|L_n^{\leq D}\|, \tag{8}$$

with optimizer $f_n^\star = L_n^{\leq D}/\|L_n^{\leq D}\|$.

We remark that if we took $D = +\infty$ here and optimized over all of $L^2(\mathbb{Q}_n)$, then the optimizer, again by a straightforward calculation, would simply be the likelihood ratio $f_n^\star = L_n/\|L_n\|$, and the optimal value its norm $\|L_n\|$. The likelihood ratio is closely related to optimal hypothesis testing via the Neyman-Pearson lemma (Neyman and Pearson, 1933), and the norm $\|L_n\|$ also plays a role through Le Cam's notion of *contiguity* and the associated *second moment method* (Le Cam and Yang, 2012). In particular, if $\|L_n\|$ is bounded as $n \to \infty$, then no test can achieve strong detection, a fact that has been used to great effect for several high-dimensional problems in recent literature (Montanari et al., 2015; Banks et al., 2018; Perry et al., 2018).

The "truncated" variant $L_n^{\leq D}$ is called the *low-degree likelihood ratio*, and emerges according to the above result as the main object conjecturally controlling the computational cost of hypothesis testing. The basic conjecture concerning the norm of the low-degree likelihood ratio is as follows.

2. We remark that kin and additive spiking are equivalent in the Gaussian NEF-QVF.

Conjecture 9 (Informal; Conjecture 2.2.4 of Hopkins (2018)) *For "sufficiently nice" sequences of probability measures $\mathbb{P}_n$ and $\mathbb{Q}_n$, if there exists $\epsilon > 0$, $D = D(n) \geq (\log n)^{1+\epsilon}$, and a constant K such that $\|L_n^{\leq D}\| \leq K$ for all n, then there is no sequence of f_n that are computable in polynomial time in n and that achieve the strong detection conditions (6).*

Different scalings of $D(n)$ are also conjectured to capture hardness of testing with various other computational time budgets ranging from polynomial to exponential (see, e.g., Section 3 of Kunisky et al. (2019) and Ding et al. (2019, 2020b)); roughly speaking, if $\|L_n^{\leq D(n)}\|$ is bounded, then we expect strong detection in time $T(n) = \exp(D(n))$ to be impossible.

These preliminaries established, we may outline our contributions at a high level: we will compute and bound $\|L_n^{\leq D}\|$ for kin-spiked and additively-spiked NEF-QVF models, and compare these results with Conjecture 9 and variants thereof to produce new predictions and challenges for the low-degree method.

2. Main results

2.1. Low-degree analysis and z-score overlap

Our first result is a general bound on $\|L_n^{\leq D}\|$ in kin-spiked NEF-QVF models. Before presenting the statement, we highlight a previously-known special case that the general result will resemble: for the NEF-QVF of Gaussian distributions with unit variance, Kunisky et al. (2019) showed the following formula:

$$\|L_n^{\leq D}\|^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2} \exp^{\leq D}(\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle), \tag{9}$$

where $\boldsymbol{x}^i$ are drawn independently from $\mathcal{P}_n$ and $\exp^{\leq D}(t) = \sum_{k=0}^{D} t^k/k!$ is the order-D Taylor expansion of the exponential function. This formula is both elegant in principle, showing that the behavior of $\|L_n^{\leq D}\|$ may be reduced to the behavior of the scalar "overlap" random variables $\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle$, and leads to simplified proofs of low-degree analyses in practice. We show that a similar result holds in any NEF-QVF, so long as we (1) replace the equality with a suitable inequality, (2) replace $\boldsymbol{x}^i$ with suitably centered and normalized z-scores, and (3) replace the exponential function with a suitable relative, which will depend on the NEF-QVF's value of v_2, the coefficient of μ^2 in the variance function.

We first briefly describe the set of all possible values of v_2.

Proposition 10 *Let $\mathcal{V} := [0, +\infty) \cup \{-\frac{1}{m} : m \in \mathbb{Z}_{\geq 1}\} \subset \mathbb{R}$. Then, for any NEF-QVF, $v_2 \in \mathcal{V}$. Conversely, for any $v \in \mathcal{V}$, there exists an NEF-QVF with $v_2 = v$. The only NEF-QVFs with $v_2 < 0$ are the binomial families (including Bernoulli), and the only NEF-QVFs with $v_2 = 0$ are the Gaussian and Poisson families.*

Definition 11 *For $t \in \mathbb{R}$ and $v \in \mathcal{V}$, define*

$$f(t; v) := \begin{cases} e^t & \text{if } v = 0, \\ (1 - vt)^{-1/v} & \text{if } v \neq 0 \text{ and } t < 1/|v|, \\ +\infty & \text{if } v > 0 \text{ and } t \geq 1/|v|. \end{cases} \tag{10}$$

Moreover, for $D \in \mathbb{N}$, let $f^{\leq D}(t; v)$ denote the order-D Taylor expansion of $f(t; v)$ about $t = 0$ for fixed v, and let $f^{\leq +\infty}(t; v) := f(t; v)$.

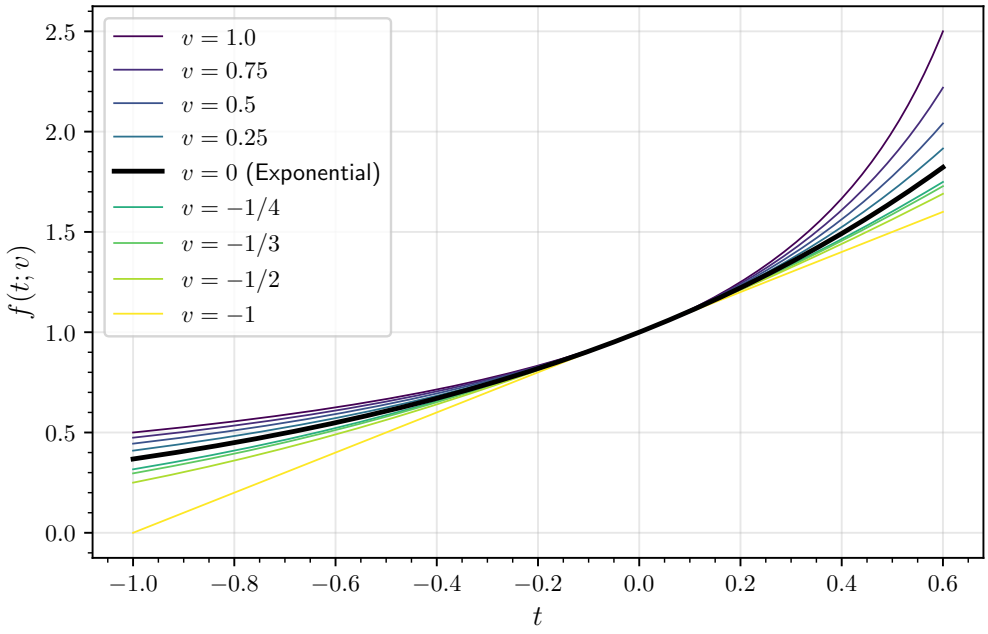

Figure 1: We plot $f(t; v)$ near $t = 0$ for various values of v, emphasizing the monotonicity in v and the appearance of the exponential function for $v = 0$.

See Figure 1 for an illustration of these functions "sandwiching" the exponential.

Theorem 12 *Let $\widetilde{\rho}_\mu$ be a mean-parametrized NEF-QVF over $\mu \in \Omega \subseteq \mathbb{R}$, with variance function $V(\mu) = v_0 + v_1\mu + v_2\mu^2$. For $\mu, x \in \Omega$, define the z-*score *as*

$$z_\mu(x) := \frac{x - \mu}{\sqrt{V(\mu)}}. \tag{11}$$

*Let $\mu_{n,i} \in \Omega$ and $\mathcal{P}_n$ be as in Definition 6 of the kin-spiked NEF-QVF model. Define the z-*score overlap*,*

$$r_n := \sum_{i=1}^{N(n)} z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2), \tag{12}$$

where $\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{P}_n$ independently. Let $L_n^{\leq D}$ denote the low-degree likelihood ratio.

- *If $v_2 \geq 0$, then for any $n \in \mathbb{N}$ and $D \in \mathbb{N} \cup \{+\infty\}$,*

$$\|L_n^{\leq D}\|^2 \leq \mathbb{E}\left[f^{\leq D}(r_n; v_2)\right], \tag{13}$$

  and equality holds if $v_2 = 0$ (i.e., in the Gaussian and Poisson NEFs).

- *If $v_2 < 0$, then for any $n \in \mathbb{N}$ and $D \in \mathbb{N} \cup \{+\infty\}$,*

$$\mathbb{E}\left[f^{\leq D}(r_n; v_2)\right] \leq \|L_n^{\leq D}\|^2 \leq \mathbb{E}\left[f^{\leq D}(r_n; 0)\right]. \tag{14}$$

We also give more cumbersome exact formulae in Section A, but emphasize these bounds here for their similarity to the simpler Gaussian case. In particular, since $f(t; v) \approx \exp(t)$ near $t = 0$ for any v, to a first approximation it appears reasonable to estimate $\|L_n^{\leq D}\|^2 \approx \mathbb{E}[\exp^{\leq D}(r_n)]$. We suggest this as a powerful heuristic, far simpler than the full low-degree likelihood ratio analysis, for making quick predictions of computational thresholds. As an example, we use this to informally derive the Kesten-Stigum threshold in the symmetric stochastic block model with two communities in Section B.1, in merely a few lines of calculation requiring no graph-theoretic intuition.

2.2. Channel monotonicity

The monotonicity of the functions $f(t; v)$ in v evident in Figure 1 suggests that we might expect $\|L_n^{\leq D}\|$ to be monotone across different kin-spiked NEF-QVF models with the same mean distribution $\mathcal{P}_n$. While this does not follow directly from the above result, a slightly more careful argument shows that it is indeed the case.

Theorem 13 *Suppose $L_n^{(i)}$ for $i \in \{1, 2\}$ are the likelihood ratios for the hypothesis testing problems in two kin-spiked NEF-QVF models, with mean domains $\Omega^{(i)}$ and variance functions $V^{(i)}(\mu) = v_0^{(i)} + v_1^{(i)}\mu + v_2^{(i)}\mu^2$. Suppose that the null means $\mu_{n,j}$ and the distribution $\mathcal{P}_n$ are the same in both problems (in particular, $\Omega^{(1)} \cap \Omega^{(2)}$ must contain the support of $\mathcal{P}_n$). If $v_2^{(1)} \leq v_2^{(2)}$, then, for any $D \in \mathbb{N} \cup \{+\infty\}$, $\|(L_n^{(1)})^{\leq D}\|^2 \leq \|(L_n^{(2)})^{\leq D}\|^2$.*

Informally, this says that if $v_2^{(1)} \leq v_2^{(2)}$, then "Problem 1 is at least as hard as Problem 2," for any given computational budget. For example, for a fixed collection of null means $\mu_{n,i}$ and a fixed spike mean distribution $\mathcal{P}_n$, we would predict the following relationships among output "channels" or observation distributions, with "$\geq$" denoting greater computational difficulty:

$$\text{Bernoulli} \geq \text{Binomial} \geq \text{Gaussian} = \text{Poisson} \geq \text{Exponential}. \tag{15}$$

This suggests two intriguing open problems further probing the low degree method. First, are these predictions in fact accurate, i.e., can they be corroborated with any other form of evidence of computational hardness? (One intriguing possibility is average-case reductions in the style of Berthet and Rigollet (2013); Brennan and Bresler (2020) between different NEF-QVFs.) And second, if these predictions are accurate, then does *strict* inequality hold in computational cost between any of these versions of a given problem, or does *channel universality* hold (we borrow the term from Lesieur et al. (2015) but use it in a slightly different sense), where in fact computational complexity of testing does not depend on the NEF-QVF through which the data are observed?

2.3. Hyperbolic secant spiked matrix model

We now turn our attention to additively-spiked NEF-QVF models, and isolate one particular model where we can both perform explicit calculations and draw a comparison to prior works. This is a *spiked matrix model*—asking us to determine whether an unstructured random matrix has been deformed by a rank-one perturbation—with noise distributed according to ρ^{sech} the probability measure on $\mathbb{R}$ which has the following density $w(x)$ with respect to Lebesgue measure:

$$w(x) := \frac{1}{2\cosh(\pi x/2)} = \frac{1}{2}\operatorname{sech}(\pi x/2). \tag{16}$$

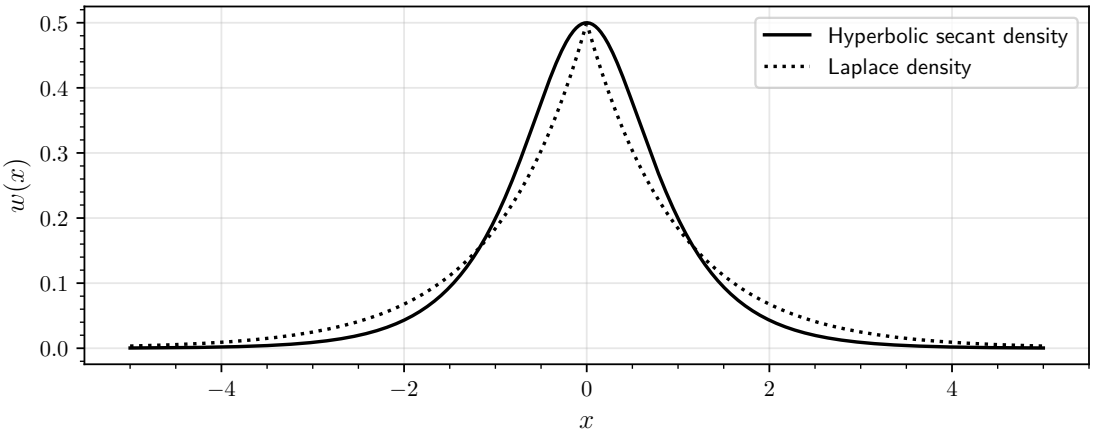

Figure 2: We plot the density $w(x)$ of the hyperbolic secant distribution used in Theorem 17, showing that it is effectively a smoothed version of the density $\frac{1}{2}\exp(-|x|)$ of the better-known Laplace distribution.

This density belongs to the rather obscure class of "generalized hyperbolic secant" NEFs mentioned in Table 1. It may be viewed as a smoothing of the Laplace distribution; see Figure 2.[3]

We next specify the spiked matrix model we will study.

Definition 14 (Rademacher-spiked Wigner matrix models) *Given a probability measure ρ over $\mathbb{R}$, we write $\mathsf{Wig}(\rho, \lambda) = ((\mathbb{Q}_n, \mathbb{P}_n))_{n=1}^{\infty}$ for the sequence of pairs of probability measures $\mathbb{Q}_n$ and $\mathbb{P}_n$ over $\mathbb{R}^{\binom{[n]}{2}}$ defined as follows:*

- *Under $\boldsymbol{Y} \sim \mathbb{Q}_n$, we draw $Y_{\{i,j\}} \sim \rho$ independently for all $i < j$.*
- *Under $\boldsymbol{Y} \sim \mathbb{P}_n$, we first draw $\boldsymbol{x} \sim \mathsf{Unif}(\{\pm 1\}^n)$ and $Z_{\{i,j\}} \sim \rho$ independently for all $i < j$, and then set $Y_{\{i,j\}} = \frac{\lambda}{\sqrt{n}} x_i x_j + Z_{\{i,j\}}$.*

Thus $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$ is an additively-spiked NEF-QVF model, as described in Definition 7. We may also view $\boldsymbol{Y} = \frac{\lambda}{\sqrt{n}} \boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{Z}$ as symmetric matrices, where we omit the diagonal from the observations (this is a technical convenience that is straightforward but tedious to eliminate).

The following result characterizes two testing algorithms related to computing the largest eigenvalue. For $\boldsymbol{Y}$ itself, for sufficiently large λ, the largest eigenvalue undergoes a "pushout" effect under $\mathbb{P}_n$ and becomes larger than the typical largest eigenvalue under $\mathbb{Q}_n$, as characterized by a variant of the Baik–Ben Arous–Péché (BBP) transition of random matrix theory (Baik et al., 2005; Capitaine et al., 2009). It turns out, however, that it is suboptimal to merely compute and threshold the largest eigenvalue of $\boldsymbol{Y}$; instead, the optimal algorithm is to first apply an entrywise transformation and only then compute and threshold the largest eigenvalue.

Proposition 15 (Better-than-BBP testing (Capitaine et al., 2009; Perry et al., 2018)) *Define the constant $\lambda_* := 2\sqrt{2}/\pi \approx 0.9$.*

3. This density has some other remarkable mathematical properties: (1) like the Gaussian density, up to dilation $w(x)$ is its own Fourier transform, and (2) $w(x)$ is the Poisson kernel over the strip $\{z : \mathsf{Im}(z) \in [-1, 1]\} \subset \mathbb{C}$.

- *If $\lambda > 1$, then strong detection in $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$ can be achieved in polynomial time by the* PCA test*,*

$$f^{\mathsf{PCA}}(\boldsymbol{Y}) := \begin{cases} p & \text{if } \frac{1}{\sqrt{n}}\lambda_{\mathsf{max}}(\boldsymbol{Y}) \geq \frac{1}{2}(2 + \lambda + \lambda^{-1}), \\ q & \text{otherwise.} \end{cases} \tag{17}$$

- *If $\lambda < 1$, then f^{PCA} does not achieve strong detection in $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$.*

- *If $\lambda > \lambda_*$, then strong detection in $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$ can be achieved in polynomial time by the* pre-transformed PCA test*:*

$$f^{\mathsf{tPCA}}(\boldsymbol{Y}) := \begin{cases} p & \text{if } \frac{1}{\sqrt{n}}\lambda_{\mathsf{max}}\left(\frac{\pi}{2}\tanh\left(\frac{\pi}{2}\boldsymbol{Y}\right)\right) \geq \frac{1}{2}(2\lambda_* + \lambda_*^2 \cdot \lambda + \lambda^{-1}), \\ q & \text{otherwise.} \end{cases} \tag{18}$$

  Here, $\tanh(\cdot)$ is applied entrywise to the matrix argument.

- *If $\lambda < \lambda_*$, then there exists no test (efficiently computable or not) that can achieve strong detection in $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$.*

The threshold λ_* is related to the *Fisher information* in the family of translates of ρ^{sech} as $\lambda_* = (\int_{-\infty}^{\infty} w'(x)^2/w(x)dx)^{-1/2}$, and the optimal entrywise transformation is the logarithmic derivative $\frac{\pi}{2}\tanh(\frac{\pi}{2}x) = -w'(x)/w(x)$; the results of Perry et al. (2018) show that both relationships hold for optimal tests in non-Gaussian spiked matrix models in great generality.

While low-degree polynomials can approximate the test f^{PCA} via the power method, the transcendental entrywise $\tanh(\cdot)$ transformation used by f^{tPCA} seems rather ill-suited to the low-degree analysis. We show below that, indeed, if we attempt to carry out the low-degree method for this problem while bounding the *entrywise degree* of the polynomials involved—the greatest power with which any given entry of $\boldsymbol{Y}$ can appear—then we obtain an incorrect threshold. Loosely speaking, this suggests that some analytic computation like the transcendental $\tanh(\cdot)$ operation is in fact *necessary* to obtain an optimal test.

Definition 16 (Entrywise degree) *For a polynomial $p \in \mathbb{R}[y_1, \ldots, y_N]$, write $\mathsf{deg}_i(p)$ for the greatest power with which y_i occurs in a monomial having non-zero coefficient in p.*

Theorem 17 *Suppose $D \in \mathbb{N}$ and $0 < \lambda < \lambda_* + \frac{1}{20D}$. Then,*

$$\limsup_{n\to\infty} \left\{ \begin{array}{ll} \textit{maximize} & \mathbb{E}_{\boldsymbol{Y}\sim\mathbb{P}_n} f_n(\boldsymbol{Y}) \\ \textit{subject to} & f_n \in \mathbb{R}[\boldsymbol{Y}], \\ & \mathsf{deg}_{\{i,j\}}(f_n) \leq D \textit{ for all } \{i,j\} \in \binom{[n]}{2}, \\ & \mathbb{E}_{\boldsymbol{Y}\sim\mathbb{Q}_n} f_n(\boldsymbol{Y})^2 = 1 \end{array} \right\} < +\infty. \tag{19}$$

That is, when we restrict our attention to polynomials of entrywise degree at most D a constant not growing with n, the apparent computational threshold suggested by the corresponding low-degree calculation shifts by $\Omega(1/D)$ from the true value.

This limitation applies, for example, to the approach of the recent work of Ding et al. (2020a). The authors propose to build tests and estimators for spiked matrix models that remain effective under heavy-tailed noise distributions by using polynomials that sum over monomials indexed by

self-avoiding walks on the matrix $\boldsymbol{Y}$. In particular, they show that, for $\lambda > 1$—the optimal threshold for Gaussian noise—such polynomials can successfully achieve strong detection in $\mathsf{Wig}(\rho, \lambda)$ for a wide variety of measures ρ, ranging from Gaussian ρ to very heavy-tailed ρ for which f^{PCA} fails severely. However, our result implies that, since these polynomials have entrywise degree 1 (that is, they are multilinear), such polynomials (and many generalizations thereof to higher but bounded entrywise degree) *cannot* achieve strong detection for all $\lambda > \lambda_*$, and thus are suboptimal for this model.

2.4. Exploration before inference: a challenge for low-degree polynomials

The discussion above suggests that, for algorithms computing low-degree polynomials, there is a tension between robustness to heavy-tailed noise distributions and optimality for specific rapidly-decaying (and, in the case above, non-Gaussian) noise distributions. We propose the following hypothesis testing problem to capture a simple case of this challenge.

Definition 18 (Mixed spiked matrix model) *Fix $\alpha > 1$ and $\lambda > 0$. Let ρ^{heavy} be the probability measure over $\mathbb{R}$ whose density is proportional to $(1 + x^2)^{-\alpha/2}$. Denote $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda) =: ((\mathbb{Q}_n^{(1)}, \mathbb{P}_n^{(1)}))_{n=1}^{\infty}$ and $\mathsf{Wig}(\rho^{\mathsf{heavy}}, \lambda) =: ((\mathbb{Q}_n^{(2)}, \mathbb{P}_n^{(2)}))_{n=1}^{\infty}$. Let $\mathsf{Mix}(\alpha, \lambda) = ((\mathbb{Q}_n, \mathbb{P}_n))_{n=1}^{\infty}$ for $\mathbb{Q}_n$ and $\mathbb{P}_n$ defined as follows:*

- *Under $\mathbb{Q}_n$, draw $i \sim \mathsf{Unif}(\{1, 2\})$, and observe $\boldsymbol{Y} \sim \mathbb{Q}_n^{(i)}$.*
- *Under $\mathbb{P}_n$, observe $\boldsymbol{Y} \sim \mathbb{P}_n^{(1)}$.*

In words, in the null distribution we flip a coin to choose between a heavy-tailed and a rapidly-decaying but non-Gaussian entrywise distribution for the noise matrix, while in the planted distribution we always observe the latter kind of noise.

The motivation for this definition should be clear at a technical level in the context of the low-degree likelihood ratio analysis: the possibility of receiving heavy-tailed inputs in effect restricts the available polynomials to low entrywise degree, which in turn precludes optimal testing in $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$. Still, we suggest that this example is not as artificial as it might appear; on the contrary, it captures an important reality of statistical practice. When faced with a dataset, the statistician knows to first examine the data, and in particular assess what distributional assumptions could be justified, before applying a particular algorithm. We might, for instance, think of the case $i = 2$ in $\mathbb{Q}_n$ as observing the results of a severely miscalibrated experiment, in which case we should not use an inference procedure fine-tuned to our distributional expectations. As this initial examination is often said to fall under the rubric of "exploratory data analysis," we call this algorithmic strategy *exploration before inference*.

It appears difficult for low-degree polynomials to match the performance of algorithms performing exploration before inference. In particular, we have the following result.

Theorem 19 *For all $\alpha > 1$ and $\lambda > \lambda_*$, there is a polynomial-time algorithm achieving strong detection in $\mathsf{Mix}(\alpha, \lambda)$. However, for all $0 < \lambda < \lambda_* + \frac{1}{10\alpha}$, we have*

$$\limsup_{n \to \infty} \left\{ \begin{array}{ll} \textit{maximize} & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{P}_n} f_n(\boldsymbol{Y}) \\ \textit{subject to} & f_n \in \mathbb{R}[\boldsymbol{Y}] \cap L^2(\mathbb{Q}_n) \\ & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n} f_n(\boldsymbol{Y})^2 = 1 \end{array} \right\} < +\infty. \tag{20}$$

We remark that the optimization above is over *all* polynomials, with no degree constraint; the only constraint is that the polynomials belong to $L^2(\mathbb{Q}_n)$. While we have specifically engineered $\mathsf{Mix}(\alpha, \lambda)$ to have fewer polynomials in $L^2(\mathbb{Q}_n)$, it seems reasonable to conjecture that even algorithms that threshold polynomials in the sense of (1) would fail to achieve strong detection in this model. On the other hand, the algorithm achieving the first statement is a simple application of exploration before inference: first, it examines the entrywise maximum of $\boldsymbol{Y}$ to determine with high probability whether $i = 2$ was chosen under $\mathbb{Q}_n$. If so, it returns q; if not, it applies the test f^{tPCA} from Proposition 15.

One possible resolution of this difficulty is the notion of *coordinate degree* suggested in the formalization of the low-degree method in the thesis of Hopkins (2018). Functions of low coordinate degree are those spanned by functions depending only on a small number of variables; this gives a generalization of low-degree polynomials that is more natural in that it does not depend upon the specific functional basis of low-degree monomials. At the very least, $\mathsf{Mix}(\cdot, \cdot)$ gives a fairly natural model where optimizing over functions of low coordinate degree rather than low-degree polynomials is necessary for the low-degree method to make correct predictions.

3. Open problems

In conclusion, we propose two directions for future research. The first is to determine whether the overlap form of the norm of the full or low-degree likelihood ratio holds in greater generality than the NEF-QVF distributions considered here or the further example of the Gaussian Wishart distribution considered in Bandeira et al. (2020b,a). These may be viewed as "integrable" statistical models, where powerful systems of combinatorial identities for orthogonal polynomials allow us to derive overlap formulae. There are some further examples where this may be possible; one tantalizing possibility is to analyze exponential families with variance functions given by polynomials of small degree greater than two. Such families have been studied in the statistics literature; see, e.g., Letac and Mora (1990); Hassairi and Zarai (2004) for cubic variance functions. Better yet would be a general overlap form not depending on the detailed structure of orthogonal polynomials; however, it remains unclear how to complete such a derivation without algebraic tools.

Another problem is to determine the extent to which low-degree polynomials can capture exploration before inference. One specific intriguing question is whether replacing degree with coordinate degree in the setting of Section 2.4 allows low-degree functions to solve the detection problem posed there, that is, whether functions of low coordinate degree can in fact achieve strong detection in $\mathsf{Mix}(\alpha, \lambda)$ for all $\alpha > 1$ and $\lambda > \lambda_*$. Of course, the same question applies to other caricatures of situations requiring exploration before inference. Generally, it seems difficult to encode the "branching" or "if statement" step of algorithms including exploration into a function of low coordinate degree, and thus plausible that the low-degree method, even augmented with the notion of coordinate degree, might make an incorrect prediction for $\mathsf{Mix}(\alpha, \lambda)$ and similar models.

We note also that exploration before inference need not always coincide with *robustness*, where we want an algorithm to perform well under poorly-behaved random or adversarial corruptions of the inputs. Exploration before inference could also arise in, e.g., a model choosing one of several different rapidly-decaying noise distributions, where an algorithm might estimate the noise distribution before performing inference. One such example was studied by Montanari et al. (2018) who give an algorithm with optimal performance on spiked matrix models that is fully agnostic to the noise distribution (over a broad class). Their approach proceeds by performing kernel density esti-

mation of the noise distribution from the data, followed by applying a pre-transformation tailored to this noise distribution as proposed by Perry et al. (2018), and then using PCA on the transformed matrix. This is a prototypical and more realistic example of exploration before inference, and our results suggest that such an algorithm again likely cannot be captured in the low-degree polynomial class; as before, we see little reason that changing degree to coordinate degree would help. The twin problems of formulating precisely what it means for an algorithm to be adaptive in this way and showing that low-degree algorithms cannot achieve such adaptivity would establish an important limitation on what kinds of problems are well-suited to the low-degree paradigm.

Acknowledgments

I thank Alex Wein for comments on an early version of the manuscript, and Sam Hopkins, Jingqiu Ding, and Afonso Bandeira for helpful discussions.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Afonso S Bandeira, Jess Banks, Dmitriy Kunisky, Cristopher Moore, and Alexander S Wein. Spectral planting and the hardness of refuting cuts, colorability, and communities in random graphs. *arXiv preprint arXiv:2008.12237*, 2020a.

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)*, volume 151, pages 78:1–78:29, 2020b.

Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018.

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.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *26th Annual Conference on Learning Theory (COLT 2013)*, pages 1046–1066, 2013.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. *arXiv preprint arXiv:2005.08099*, 2020.

Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1):1–47, 2009.

Jingqiu Ding, Samuel B Hopkins, and David Steurer. Estimating rank-one spikes from heavy-tailed noise via self-avoiding walks. *arXiv preprint arXiv:2008.13735*, 2020a.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-time algorithms for sparse PCA. *arXiv preprint arXiv:1907.11635*, 2019.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. The average-case time complexity of certifying the restricted isometry property. *arXiv preprint arXiv:2005.11270*, 2020b.

David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Low-degree hardness of random optimization problems. *arXiv preprint arXiv:2004.12063*, 2020.

Abdelhamid Hassairi and Mohammed Zarai. Characterization of the cubic exponential families by orthogonality of polynomials. *Annals of probability*, 32(3B):2463–2476, 2004.

Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

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

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 *58th Annual Symposium on Foundations of Computer Science (FOCS 2017)*, pages 720–731. IEEE, 2017.

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.

HO Lancaster. Joint probability distributions in the Meixner classes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(3):434–443, 1975.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2012.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 680–687. IEEE, 2015.

Gérard Letac and Marianne Mora. Natural real exponential families with cubic variance functions. *The Annals of Statistics*, 18(1):1–37, 1990.

Joseph Meixner. Orthogonale polynomsysteme mit einer besonderen gestalt der erzeugenden funktion. *Journal of the London Mathematical Society*, 1(1):6–13, 1934.

Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015.

Andrea Montanari, Feng Ruan, and Jun Yan. Adapting to unknown noise distribution in matrix denoising. *arXiv preprint arXiv:1810.02954*, 2018.

Cristopher Moore. The computer science and physics of community detection: landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

Carl N Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

Carl N Morris. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, 11(2):515–529, 1983.

Jerzy Neyman and Egon Sharpe Pearson. IX. 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(694-706):289–337, 1933.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA I: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

Tselil Schramm and Alexander S Wein. Computational barriers to estimation from low-degree polynomials. *arXiv preprint arXiv:2008.02269*, 2020.

Alexander S Wein. Optimal low-degree hardness of maximum independent set. *arXiv preprint arXiv:2010.06563*, 2020.

Appendix A. Low-degree likelihood ratio analysis in NEF-QVFs

In this section, we give an explicit formulae for $\|L_n^{\leq D}\|^2$ in kin- and additively-spiked NEF-QVF models, and prove the bounds in Theorem 12. Our strategy will be to decompose L_n according to orthogonal polynomials in $L^2(\mathbb{Q}_n)$, and sum the masses of the components of low-degree polynomials.

Therefore, in Section A.1, we first review the general description of orthogonal polynomials in NEF-QVFs, and prove some minor further results that will be useful. Then, in Section A.2 we give formulae for the coefficients of each orthogonal polynomial component of the likelihood ratio in both the kin- and additively-spiked models. Finally, in Sections A.3 and A.4, we prove results concerning the norms of the full and low-degree likelihood ratios in kin-spiked models (including Theorem 12), where further simplifications are possible.

A.1. Orthogonal polynomials in NEF-QVFs

Our main tool will be that, in NEF-QVFs, there is a remarkable connection between the likelihood ratio and the orthogonal polynomials of ρ_θ. The likelihood ratio in any NEF is simple:

$$L(y;\theta) := \frac{d\rho_\theta}{d\rho_0}(y) = \mathsf{exp}(y\theta - \psi(\theta)), \tag{21}$$

where $\psi(\theta) = \mathbb{E}_{x\sim\rho_0}[\exp(\theta x)]$. We may also reparametrize in terms of the mean:

$$\widetilde{L}(y;\mu) := L(y;(\psi')^{-1}(\mu)) = \exp(y(\psi')^{-1}(\mu) - \psi((\psi')^{-1}(\mu))). \tag{22}$$

As the following result of Morris shows, in an NEF-QVF, $\widetilde{L}(y;\mu)$ is a kind of generating function of the orthogonal polynomials of $\widetilde{\rho}_\mu$.

Definition 20 *For $v \in \mathbb{R}$, define the sequences of constants*

$$\widehat{a}_k(v) := \prod_{j=0}^{k-1}(1+vj), \tag{23}$$

$$a_k(v) := k! \cdot \widehat{a}_k(v). \tag{24}$$

Proposition 21 (NEF-QVF Rodrigues Formula; Theorem 4 of Morris (1982)) *Let $\mu_0 = \psi'(0) = \mathbb{E}_{x\sim\rho_0}[x]$. Define the polynomials*

$$p_k(y;\mu_0) := \frac{V(\mu_0)^k}{\widetilde{L}(y,\mu_0)} \cdot \frac{d^k\widetilde{L}}{d\mu^k}(y,\mu_0). \tag{25}$$

Then, $p_k(y;\mu_0)$ is a degree k monic polynomial in y, and this family satisfies the orthogonality relation

$$\mathop{\mathbb{E}}_{y\sim\widetilde{\rho}_{\mu_0}} p_k(y;\mu_0)p_\ell(y;\mu_0) = \delta_{k\ell} \cdot a_k(v_2)V(\mu_0)^k. \tag{26}$$

In particular, defining the normalized polynomials

$$\widehat{p}_k(y;\mu_0) := \frac{1}{V(\mu_0)^{k/2}\sqrt{a_k(v_2)}}p_k(y;\mu_0), \tag{27}$$

the $\widehat{p}_k(y;\mu_0)$ are orthonormal polynomials for $\widetilde{\rho}_{\mu_0}$.

The main property of these polynomials that will be useful for us is the following identity, also obtained by Morris, giving the expectation of a given orthogonal polynomial under the kin spiking operation, i.e., under a different distribution from the same NEF-QVF.

Proposition 22 (Corollary 1 of Morris (1982)) *For all $k \in \mathbb{N}$ and $x, \mu \in \Omega$,*

$$\mathop{\mathbb{E}}_{y\sim\widetilde{\rho}_x} p_k(y;\mu) = \widehat{a}_k(v_2)(x-\mu)^k. \tag{28}$$

We may obtain a straightforward further corollary by including the normalization, which allows us to incorporate the variance factor into a z-score, as follows.

Corollary 23 (Kin-spiked expectation) *For all $k \in \mathbb{N}$ and $x, \mu \in \Omega$,*

$$\mathop{\mathbb{E}}_{y\sim\widetilde{\rho}_x} \widehat{p}_k(y;\mu) = \sqrt{\frac{\widehat{a}_k(v_2)}{k!}}z_\mu(x)^k. \tag{29}$$

We will also be interested in the analogous result for additive spiking. This result is less elegant, and is expressed in terms of another polynomial sequence. First, we write the precise generating function relation between the likelihood ratio and the orthogonal polynomials.

Proposition 24 (Generating function) *Let $\mu \in \Omega$ and write $\psi(\eta) = \mathbb{E}_{x\sim\widetilde{\rho}_\mu}[\exp(\eta x)]$. Then,*

$$\sum_{k\geq 0} \frac{z_\mu(t)^k}{k!} p(y;\mu) = \exp\left(y(\psi')^{-1}(t) - \psi((\psi')^{-1}(t))\right). \tag{30}$$

Note that here we are "rebasing" the NEF-QVF to have $\widetilde{\rho}_\mu$ as the base measure by our definition of $\psi(\cdot)$. One may view this result as generalizing to NEF-QVFs the generating function $\exp(ty - \frac{1}{2}t^2)$ for Hermite polynomials. The key property of such generating functions is that y appears *linearly* in the exponential. (Indeed, as early as 1934, Meixner had essentially discovered the NEF-QVFs, albeit only recognizing their significance in terms of this distinctive property of their orthogonal polynomials Meixner (1934); Lancaster (1975).)

This linearity allows us to prove an addition formula, expanding the translation operator in orthogonal polynomials.

Definition 25 (Translation polynomials) *Let $\tau_k(y;\mu) \in \mathbb{R}[y]$ be defined by the generating function*

$$\sum_{k\geq 0} \frac{z_\mu(t)^k}{k!} \tau_k(y;\mu) := \exp\left(y(\psi')^{-1}(t)\right). \tag{31}$$

Also, define the normalized versions

$$\widehat{\tau}(y;\mu) := \frac{1}{V(\mu)^{k/2}\sqrt{a_k(v_2)}} \tau_k(x;\mu). \tag{32}$$

Proposition 26 (Addition formula) *For all $x, y \in \mathbb{R}$ and $\mu \in \Omega$,*

$$p_k(x+y;\mu) = \sum_{\ell=0}^{k} \binom{k}{\ell} \tau_{k-\ell}(x;\mu) p_\ell(y;\mu). \tag{33}$$

Proof This follows from expanding the generating function (30) at $x + y$ as a product of two exponential generating functions. ■

Finally, we obtain the additively-spiked version of Corollary 23 by taking expectations and using the orthogonality of the p_k.

Proposition 27 (Additively-spiked expectation) *For all $k \in \mathbb{N}$, $\mu \in \Omega$, and $x \in \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{y\sim\widetilde{\rho}_\mu} \widehat{p}_k(x+y;\mu) = \widehat{\tau}_k(x;\mu). \tag{34}$$

Proof This follows from taking expectations on either side of (33), observing that the only non-zero term is for $\ell = 0$ by the orthogonality of the p_ℓ, and noting that $p_0(y;\mu) = 1$. ■

A.2. Components of the likelihood ratio

Returning to the multivariate setting of our results, let $\mathbb{Q}_n$ and $\mathbb{P}_n$ be as in Definition 6 of the kin-spiked NEF-QVF model. Then, the likelihood ratio is

$$L_n(\boldsymbol{y}) := \frac{d\mathbb{P}_n}{d\mathbb{Q}_n}(\boldsymbol{y}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N \frac{d\widetilde{\rho}_{x_i}}{d\widetilde{\rho}_{\mu_{n,i}}}(y_i)\right]. \tag{35}$$

An orthonormal system of polynomials for $\mathbb{Q}_n$ is given by the product basis formed from the $\widehat{p}_k(y;\mu_{n,i})$ that we defined in Proposition 21:

$$\widehat{P}_{\boldsymbol{k}}(\boldsymbol{y};\boldsymbol{\mu}_n) := \prod_{i=1}^N \widehat{p}_{k_i}(y_i;\mu_{n,i}) \tag{36}$$

for $\boldsymbol{k}\in\mathbb{N}^N$, where $\boldsymbol{\mu}_n := (\mu_{n,1},\dots,\mu_{n,N(n)})$.

We show that the projection of L_n onto any component $\widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)$ admits the following convenient expression in terms of the z-score.

Lemma 28 (Components under kin spiking) *In the kin-spiked NEF-QVF model, for all $\boldsymbol{k}\in\mathbb{N}^N$,*

$$\langle L_n, \widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)\rangle = \sqrt{\frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N z_{\mu_{n,i}}(x_i)^{k_i}\right]. \tag{37}$$

Proof Performing a change of measure using the likelihood ratio and factorizing the inner product using independence of coordinates under $\mathbb{Q}_n$, we find

$$\begin{aligned}
\langle L_n, \widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)\rangle &= \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathbb{Q}_n}\left[L_n(\boldsymbol{y})\widehat{P}_{\boldsymbol{k}}(\boldsymbol{y};\boldsymbol{\mu}_n)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathbb{P}_n}\left[\widehat{P}_{\boldsymbol{k}}(\boldsymbol{y};\boldsymbol{\mu}_n)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N \mathop{\mathbb{E}}_{y_i\sim\widetilde{\rho}_{x_i}}[\widehat{p}_{k_i}(y_i;\mu_{n,i})]\right]
\end{aligned}$$

and using Corollary 23,

$$= \sqrt{\frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N z_{\mu_{n,i}}(x_i)^{k_i}\right], \tag{38}$$

completing the proof. ■

Following the same argument for the additively-spiked model and using Proposition 27 instead of Corollary 23 gives the following similar result.

Lemma 29 (Components under additive spiking) *In the additively-spiked NEF-QVF model, for all $\boldsymbol{k}\in\mathbb{N}^N$,*

$$\langle L_n, \widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)\rangle = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N \widehat{\tau}_{k_i}(x_i;\mu_{n,i})\right]. \tag{39}$$

A.3. Full likelihood ratio norm

First, we give an exact formula for the norm of the untruncated likelihood ratio in a kin-spiked NEF-QVF model.

Theorem 30 *In the kin-spiked NEF-QVF model, for all $n \in \mathbb{N}$,*

$$\|L_n\|^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{P}_n} \left[\prod_{i=1}^{N} f(z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2); v_2) \right]. \tag{40}$$

The key technical step is to recognize that the function $f(\cdot; v)$ from Definition 11 is in fact the exponential generating function of the $\widehat{a}_k(v)$, as follows.

Proposition 31 *For all $t \in \mathbb{R}$ and $v \in \mathcal{V}$,*

$$f(t; v) = \sum_{k=0}^{\infty} \frac{\widehat{a}_k(v)}{k!} t^k. \tag{41}$$

Proof Differentiating the power series termwise and using the formula from Definition 20 gives the differential equation

$$\frac{\partial}{\partial t} f(t; v) = f(t; v) + vt \frac{\partial}{\partial t} f(t; v), \tag{42}$$

and the result follows upon solving the equation. ■

Proof (of Theorem 30) We have by Lemma 28

$$\begin{aligned}
\|L_n\|^2 &= \sum_{\boldsymbol{k} \in \mathbb{N}^N} \langle L_n, \widehat{P}_{\boldsymbol{k}}(\cdot; \boldsymbol{\mu}_n) \rangle^2 \\
&= \sum_{\boldsymbol{k} \in \mathbb{N}^N} \frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!} \left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{P}_n} \left[\prod_{i=1}^{N} z_{\mu_{n,i}}(x_i)^{k_i} \right] \right)^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{P}_n} \left[\sum_{\boldsymbol{k} \in \mathbb{N}^N} \prod_{i=1}^{N} \left\{ \frac{\widehat{a}_{k_i}(v_2)}{k_i!} (z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2))^{k_i} \right\} \right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{P}_n} \left[\prod_{i=1}^{N} \left\{ \sum_{k=0}^{\infty} \frac{\widehat{a}_k(v_2)}{k!} (z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2))^{k} \right\} \right],
\end{aligned} \tag{43}$$

and the result follows from Proposition 31. ■

A.4. Low-degree likelihood ratio norm: Proof of Theorem 12

For this result, we use two more ancillary facts about the constants $\widehat{a}_k(\cdot)$.

Proposition 32 (Monotonicity) *For $k \in \mathbb{N}$, $\widehat{a}_k(v)$ is non-negative and monotonically non-decreasing in v over $v \in \mathcal{V}$.*

Proof Recall from (23) that, by definition,

$$\widehat{a}_k(v) = \prod_{j=0}^{k-1}(1+vj). \tag{44}$$

Thus clearly $\widehat{a}_k(v)$ is monotonically non-decreasing over $v \geq 0$, since each factor is monotonically non-decreasing.

If $v \in \mathcal{V}$ with $v < 0$, then $v = -\frac{1}{m}$ for some $m \in \mathbb{Z}_{\geq 1}$. Thus for $k \geq m+1$, $\widehat{a}_k(v) = 0$. So, in this case we may rewrite

$$\widehat{a}_k(v) = \mathbb{1}\{k \leq m\} \prod_{j=0}^{\min\{k-1, m-1\}} (1+vj). \tag{45}$$

Now, each factor belongs to $[0, 1)$, and again each factor is monotonically non-decreasing with v, so the result follows. ■

Proposition 33 (Multiplicativity relations) *For all* $\boldsymbol{k} \in \mathbb{N}^N$,

$$\begin{aligned} &\textstyle\prod_{i=1}^N \widehat{a}_{k_i}(v) \leq \widehat{a}_{\sum_{i=1}^N k_i}(v) \quad \textit{if } v > 0, \\ &\textstyle\prod_{i=1}^N \widehat{a}_{k_i}(v) = \widehat{a}_{\sum_{i=1}^N k_i}(v) \quad \textit{if } v = 0, \\ &\textstyle\prod_{i=1}^N \widehat{a}_{k_i}(v) \geq \widehat{a}_{\sum_{i=1}^N k_i}(v) \quad \textit{if } v < 0. \end{aligned} \tag{46}$$

Proof When $v = 0$, then $\widehat{a}_k(v) = 1$ for all k, so the result follows immediately. When $v > 0$, we have

$$\begin{aligned} \prod_{i=1}^N \widehat{a}_{k_i}(v) &= \prod_{i=1}^N \prod_{j=0}^{k_i - 1} (1+vj) \\ &\leq \prod_{i=1}^N \prod_{j=\sum_{a=1}^{i-1} k_a}^{\sum_{a=1}^{i} k_a} (1+vj) \\ &= \prod_{j=1}^{\sum_{i=1}^N k_i} (1+vj) \\ &= \widehat{a}_{\sum_{i=1}^N k_i}(v). \end{aligned} \tag{47}$$

When $v < 0$, a symmetric argument together with the observations from Proposition 32 gives the result. ■

Proof (of Theorem 12) Suppose first that $v_2 \geq 0$. We have by Lemma 28

$$\begin{aligned}
\|L_n^{\leq D}\|^2 &= \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \langle L_n, \widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)\rangle^2 \\
&= \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!} \left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N z_{\mu_{n,i}}(x_i)^{k_i}\right]\right)^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{P}_n}\left[\sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!} \prod_{i=1}^N (z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2))^{k_i}\right],
\end{aligned}$$

and using Proposition 33,

$$\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{P}_n}\left[\sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \frac{\widehat{a}_{|\boldsymbol{k}|}(v_2)}{\prod_{i=1}^N k_i!} \prod_{i=1}^N (z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2))^{k_i}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{P}_n}\left[\sum_{d=0}^D \frac{\widehat{a}_d(v_2)}{d!} \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|=d}} \binom{d}{k_1\cdots k_N} \prod_{i=1}^N (z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2))^{k_i}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{P}_n}\left[\sum_{d=0}^D \frac{\widehat{a}_d(v_2)}{d!} \left(\sum_{i=1}^N z_{\mu_{n,i}}(x_i^1) z_{\mu_{n,i}}(x_i^2)\right)^d\right], \qquad (48)
\end{aligned}$$

giving the upper bound from (13) for $v_2 > 0$. When $v_2 = 0$, then equality holds above, so we obtain equality in (13). Also, when $v_2 < 0$, then the above argument holds with the inequality reversed, giving the lower bound of (14).

Finally, for the upper bound of (14), note that when $v_2 < 0$, we may bound $\|L_n^{\leq D}\|^2$ using Proposition 32 and the result for $v_2 = 0$ by

$$\begin{aligned}
\|L_n^{\leq D}\|^2 &= \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \frac{\prod_{i=1}^N \widehat{a}_{k_i}(v_2)}{\prod_{i=1}^N k_i!} \left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N z_{\mu_{n,i}}(x_i)^{k_i}\right]\right)^2 \\
&\leq \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|\leq D}} \frac{\prod_{i=1}^N \widehat{a}_{k_i}(0)}{\prod_{i=1}^N k_i!} \left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{i=1}^N z_{\mu_{n,i}}(x_i)^{k_i}\right]\right)^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} f^{\leq D}(r_n;0), \qquad (49)
\end{aligned}$$

giving the result. ■

Appendix B. Applications

B.1. Example: Kesten-Stigum on the back of an envelope

Let us show how to use Theorem 12 to predict a computational threshold in the symmetric stochastic block model with two communities (see, e.g., Abbe (2017); Moore (2017) for surveys of this model). In this model, $N(n) = \binom{n}{2}$, and indexing in N is identified with pairs $\{i, j\} \in \binom{[n]}{2}$, which represent edges in a graph. There are two external parameters, $a, b > 0$. The model belongs to the Bernoulli NEF-QVF, with $v_2 = -1$, and the means in the null and planted models are as follows:

- Under $\mathbb{Q}_n$, $\mu_{n,\{i,j\}} = \frac{a+b}{2n}$, with variances $V(\mu_{n,\{i,j\}}) = \mu_{n,\{i,j\}}(1 - \mu_{n,\{i,j\}}) \approx \mu_{n,\{i,j\}}$.
- Under $\mathbb{P}_n$, we generate $\boldsymbol{\sigma} \in \{\pm 1\}^N$ either i.i.d. uniformly or conditioned to have equal numbers of plus and minus coordinates (this does not make a significant difference), and set $x_{\{i,j\}} = \frac{a}{n}$ if $\sigma_i = \sigma_j$ and $x_{\{i,j\}} = \frac{b}{n}$ if $\sigma_i \neq \sigma_j$. We may summarize this as

$$x_{\{i,j\}} = \frac{a+b}{2n} + \frac{a-b}{2n}\sigma_i\sigma_j. \tag{50}$$

We first compute the z-scores:

$$z_{\mu_{n,\{i,j\}}}(x_{\{i,j\}}) = \frac{\left(\frac{a+b}{2n} + \frac{a-b}{2n}\sigma_i\sigma_j\right) - \frac{a+b}{2n}}{\sqrt{\frac{a+b}{2n}}} = \frac{1}{\sqrt{2n}} \cdot \frac{a-b}{\sqrt{a+b}} \cdot \sigma_i\sigma_j. \tag{51}$$

As a shorthand, let us write $\boldsymbol{z}(\boldsymbol{x})$ for the coordinatewise application of this function. We then compute the inner product of the z-scores of two independent copies of $\boldsymbol{x}$, which may be expressed in terms of two copies of $\boldsymbol{\sigma}$:

$$r_n := \langle \boldsymbol{z}(\boldsymbol{x}^1), \boldsymbol{z}(\boldsymbol{x}^2) \rangle = \frac{1}{2n}\frac{(a-b)^2}{a+b} \sum_{1 \le i < j \le n} \sigma_i^1\sigma_i^2\sigma_j^1\sigma_j^2 = \frac{(a-b)^2}{4(a+b)} \cdot \frac{\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle^2 - n}{n}. \tag{52}$$

By our heuristic based on Theorem 12, we then expect

$$\|L_n^{\le D}\|^2 \approx \mathop{\mathbb{E}}_{\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2} \left[\mathsf{exp}^{\le D} \left(\frac{(a-b)^2}{4(a+b)} \cdot \left(\frac{\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle^2}{n} - 1 \right) \right) \right]. \tag{53}$$

(Alternatively, if we work in the Poisson NEF instead of the Bernoulli NEF, then this will be an exact equality by Theorem 12.)

We assume heuristically that $\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle / \sqrt{n}$ is distributed approximately as $\mathcal{N}(0, 1)$, and that $D(n)$ grows slowly enough that only after this convergence does the convergence $\mathsf{exp}^{\le D}(\cdot) \to \mathsf{exp}(\cdot)$ occur. Thus, we expect

$$\limsup_{n \to \infty} \|L_n^{\le D(n)}\|^2 \lesssim \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0,1)} \left[\mathsf{exp} \left(\frac{(a-b)^2}{4(a+b)} (g^2 - 1) \right) \right], \tag{54}$$

where the right-hand side evaluates the moment-generating function of a χ^2 random variable, which is finite if and only if $\frac{(a-b)^2}{4(a+b)} < \frac{1}{2}$, or if and only if $(a-b)^2 < 2(a+b)$, which is the Kesten-Stigum threshold.

B.2. Channel monotonicity: Proof of Theorem 13

This result is a simple consequence of the arguments we have made already to prove Theorem 12.

Proof (of Theorem 13) We have by Lemma 28

$$\begin{aligned}\|(L_n^{(i)})^{\leq D}\|^2 &= \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N\\ |\boldsymbol{k}|\leq D}} \langle L_n^{(i)}, \widehat{P}_{\boldsymbol{k}}(\cdot;\boldsymbol{\mu}_n)\rangle^2 \\ &= \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N\\ |\boldsymbol{k}|\leq D}} \frac{\prod_{j=1}^N \widehat{a}_{k_j}(v_2^{(i)})}{\prod_{j=1}^N k_j!} \left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_n}\left[\prod_{j=1}^N z_{\mu_{n,j}}(x_j)^{k_j}\right]\right)^2. \end{aligned} \tag{55}$$

In each term on the right-hand side, the only factor that depends on i is $\prod_{j=1}^N \widehat{a}_{k_j}(v_2^{(i)})$, so the result follows from the monotonicity described by Proposition 32. (Indeed, this shows slightly more, that the monotonicity holds even for the norm of the projection of $L_n^{(i)}$ onto the orthogonal polynomial of any given index $\boldsymbol{k}$.) ■

B.3. Hyperbolic secant spiked matrix model: Proof of Theorem 17

To prove this result, we will analyze the translation polynomials τ_k, from Definition 25, for the NEF generated by ρ^{sech}. First, note that the mean and variance of ρ^{sech} are $\mu = 0$ and $V(0) = 1$, and more generally the variance function in the generated NEF is $V(\mu) = \mu^2 + 1$ (per Table 1), where in particular the quadratic coefficient is $v_2 = 1$. Thus the associated normalizing constants are $a_k(v_2) = (k!)^2$ and $\widehat{a}_k(v_2) = k!$.

Recall that the translation polynomials admit a generating function expressed in terms of the cumulant generating function of ρ^{sech}. We therefore compute

$$\psi(\theta) := \mathop{\mathbb{E}}_{y\sim\rho^{\mathsf{sech}}} \exp(\theta y) = \frac{1}{2}\int_{-\infty}^{\infty} \operatorname{sech}\left(\frac{\pi y}{2}\right)\exp(\theta y)\,dy = -\log(\cos\theta), \tag{56}$$

$$\psi'(\theta) = \tan(\theta), \tag{57}$$

whereby the translation polynomials for $\mu = 0$ (the mean of ρ^{sech}) have the generating function

$$\sum_{k\geq 0}\frac{t^k}{k!}\tau_k(y;0) = \sum_{k\geq 0} t^k \widehat{\tau}_k(y;0) = \exp\left(y\tan^{-1}(t)\right). \tag{58}$$

Before proceeding, we also establish some preliminary bounds on the coefficients and values of these polynomials. We denote by $[x^\ell](p(x))$ the coefficient of x^ℓ in a polynomial or formal power series $p(x)$.

Proposition 34 *For all $k \geq 1$ and $\ell \geq 0$,*

$$|[x^\ell](\widehat{\tau}_k(x))| \leq \mathbb{1}\{k \equiv \ell \pmod 2, \ell > 0\}\frac{(2\log(ek))^{\ell-1}}{k\,\ell!}. \tag{59}$$

Proof Expanding the generating function, we have

$$[x^\ell](\widehat{\tau}_k(x)) = [t^k x^\ell](\exp(x \tanh^{-1}(t))) = \frac{1}{\ell!}[t^k]((\tanh^{-1}(t))^\ell). \tag{60}$$

If $k \geq 1$ and $\ell = 0$, then this is zero. Since the coefficients in the Taylor series of $\tanh^{-1}(t)$ are $[t^k](\tanh^{-1}(t)) = \mathbb{1}\{k \equiv 1 \pmod 2\}(-1)^{(k-1)/2}/k$, we may bound

$$|[x^\ell](\widehat{\tau}_k(x))| \leq \mathbb{1}\{k \equiv \ell \pmod 2, \ell > 0\}\frac{1}{\ell!}\underbrace{\sum_{\substack{a_1,\ldots,a_\ell \geq 1 \\ a_1+\cdots+a_\ell = k}} \frac{1}{\prod_{i=1}^\ell a_i}}_{c(k,\ell)}. \tag{61}$$

We now show that $c(k,\ell) \leq (2\log(ek))^{\ell-1}/k$ by induction on ℓ. Since $c(k,1) = 1/k$, the base case holds. We note the bound on harmonic numbers

$$\sum_{a=1}^k \frac{1}{a} \leq \log(ek) \text{ for all } k \geq 1. \tag{62}$$

Supposing the result holds for $c(k,\ell-1)$, we expand $c(k,\ell)$ according to the value that a_ℓ takes:

$$\begin{aligned}
c(k,\ell) &\leq \sum_{a=1}^{k-1} \frac{1}{a} c(k-a, \ell-1) \\
&\leq (2\log(ek))^{\ell-2} \sum_{a=1}^{k-1} \frac{1}{a} \cdot \frac{1}{k-a} && \text{(inductive hypothesis)} \\
&\leq \frac{(2\log(ek))^{\ell-2}}{k} \sum_{a=1}^{k-1} \left(\frac{1}{a} + \frac{1}{k-a}\right) \\
&\leq \frac{(2\log(ek))^{\ell-2}}{k} \cdot 2\log(ek), && \text{(by (62))}
\end{aligned}$$

completing the argument. ■

This yields the following pointwise bound. As we will ultimately be evaluating this on quantities of order $O(n^{-1/2})$, what is most important to us is the precision for very small arguments.

Corollary 35 *For all $k \geq 1$ and $x > 0$,*

$$|\widehat{\tau}_k(x)| \leq \begin{cases} x \cdot \frac{1}{k} \cdot (ek)^{2x} & \text{if } k \text{ odd}, \\ x^2 \cdot \frac{2\log(ek)}{k} \cdot (ek)^{2x} & \text{if } k \text{ even}. \end{cases} \tag{63}$$

Proof Write $\ell_0 = 1$ if k is odd and $\ell_0 = 2$ if k is even. We bound by Proposition 34,

$$\begin{aligned}
|\widehat{\tau}_k(x)| &\leq \frac{1}{k} \sum_{\ell=\ell_0}^k \frac{(2\log(ek))^{\ell-1}}{\ell!} x^\ell \\
&\leq \frac{x^{\ell_0}(2\log(ek))^{\ell_0-1}}{k} \sum_{\ell=\ell_0}^k \frac{(2\log(ek)x)^{\ell-\ell_0}}{(\ell-\ell_0)!} \\
&\leq \frac{x^{\ell_0}(2\log(ek))^{\ell_0-1}}{k} \exp((2\log(ek)x), && (64)
\end{aligned}$$

and the result follows upon rearranging. ■

Proof (of Theorem 17) First, applying Lemma 29 to the hyperbolic secant spiked matrix model, the coefficients of the likelihood ratio are given by, for any $\boldsymbol{k} \in \mathbb{N}^{\binom{[n]}{2}}$,

$$\begin{aligned}\langle L_n(\boldsymbol{Y}), \widehat{P}_{\boldsymbol{k}} \rangle &= \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \mathcal{P}_n} \left[\prod_{1 \le i < j \le n} \widehat{\tau}_{k_{\{i,j\}}}(X_{\{i,j\}}) \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathsf{Unif}(\{\pm 1\}^n)} \left[\prod_{1 \le i < j \le n} \widehat{\tau}_{k_{\{i,j\}}} \left(\frac{\lambda}{\sqrt{n}} x_i x_j \right) \right]. \end{aligned} \tag{65}$$

Our specific choice of $\boldsymbol{x} \in \{\pm 1\}^n$ allows an interesting further simplification: thanks to this choice, we can decouple the dependence of the components of L_n on λ from the dependence on $\boldsymbol{x}$. Note that, by the generating function identity (58), for all $k \ge 0$ we have that $\tau_k(x)$ contains only monomials of the same parity as k. Therefore, for all $\boldsymbol{k} \in \mathbb{N}^{\binom{[n]}{2}}$, we have

$$\langle L_n, \widehat{P}_{\boldsymbol{k}} \rangle = \prod_{i<j} \widehat{\tau}_{k_{ij}} \left(\frac{\lambda}{\sqrt{n}} \right) \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\prod_{i<j} (x_i x_j)^{k_{ij}} \right]. \tag{66}$$

Here and in the remainder of the proof, we write $k_{ij} = k_{\{i,j\}}$ and $i < j$ for $1 \le i < j \le n$ to lighten the notation. Let us also write $|\boldsymbol{k}|_\infty := \max_{i<j} k_{ij}$.

We note that, since the second factor above is either 0 or 1, we may further bound

$$|\langle L_n, \widehat{P}_{\boldsymbol{k}} \rangle| \le \left| \prod_{i<j} \widehat{\tau}_{k_{ij}} \left(\frac{\lambda}{\sqrt{n}} \right) \right| \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\prod_{i<j} (x_i x_j)^{k_{ij}} \right]$$

When $|\boldsymbol{k}|_\infty \le D$, then, by Corollary 35, we may continue

$$\le \prod_{\substack{i<j \\ k_{ij}>0}} \frac{(eD)^{\frac{\lambda}{\sqrt{n}}}}{k_{ij}} (2 \log(e k_{ij}))^{\mathbb{1}\{k_{ij} \text{ even}\}} \left(\frac{\lambda}{\sqrt{n}} \right)^{1+\mathbb{1}\{k_{ij} \text{ even}\}} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\prod_{i<j} (x_i x_j)^{k_{ij}} \right]. \tag{67}$$

Squaring and rewriting this as an expectation over two independent $\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathsf{Unif}(\{\pm 1\}^n)$, we find

$$|\langle L_n, \widehat{P}_{\boldsymbol{k}} \rangle|^2 \le \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2} \prod_{\substack{i<j \\ k_{ij}>0}} \frac{(eD)^{2\frac{\lambda}{\sqrt{n}}}}{k_{ij}^2} (2 \log(e k_{ij}))^{2\,\mathbb{1}\{k_{ij} \text{ even}\}} \left(\frac{\lambda}{\sqrt{n}} \right)^{2(1+\mathbb{1}\{k_{ij} \text{ even}\})} (x_i^1 x_i^2 x_j^1 x_j^2)^{k_{ij}}. \tag{68}$$

Summing over $|\boldsymbol{k}|_\infty \le D$, we then find

$$\begin{aligned}&\sum_{\substack{\boldsymbol{k} \in \mathbb{N}^N \\ |\boldsymbol{k}|_\infty \le D}} |\langle L_n, \widehat{P}_{\boldsymbol{k}} \rangle|^2 \\ &\le \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2} \prod_{i<j} \left(1 + (eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^2}{n} \sum_{\substack{k=1 \\ k \text{ odd}}}^{D} \frac{1}{k^2} (x_i^1 x_i^2 x_j^1 x_j^2)^k + (eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^4}{n^2} \sum_{\substack{k=1 \\ k \text{ even}}}^{D} \frac{4 \log(ek)^2}{k^2} (x_i^1 x_i^2 x_j^1 x_j^2)^k \right)\end{aligned}$$

and, using that the x_i^a are Rademacher-valued,

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} \prod_{1\le i<j\le n} \left(1 + (eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^2}{n} x_i^1 x_i^2 x_j^1 x_j^2 \sum_{\substack{k=1 \\ k \text{ odd}}}^{D} \frac{1}{k^2} + (eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^4}{n^2} \sum_{\substack{k=1 \\ k \text{ even}}}^{D} \frac{4\log(ek)^2}{k^2}\right)$$

Here, using that $\sum_{\ell\ge 0} \frac{1}{(2\ell+1)^2} = \frac{\pi^2}{8} = \lambda_*^{-2}$ and $\sum_{\ell \ge D/2} \frac{1}{(2\ell+1)^2} \ge \int_{D/2}^{\infty} \frac{dx}{(2x+1)^2} = \frac{1}{2D+2} \ge \frac{1}{3D}$, we may write

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} \prod_{1\le i<j\le n} \left(1 + (eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^2}{n} x_i^1 x_i^2 x_j^1 x_j^2 \left(\frac{1}{\lambda_*^2} - \frac{1}{3D}\right) + O\left(\frac{1}{n^2}\right)\right)$$

$$\le \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} \exp\left((eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^2}{n} \left(\frac{1}{\lambda_*^2} - \frac{1}{3D}\right) \sum_{1\le i<j\le n} x_i^1 x_i^2 x_j^1 x_j^2 + O(1)\right)$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} \exp\left((eD)^{2\frac{\lambda}{\sqrt{n}}} \frac{\lambda^2}{2} \left(\frac{1}{\lambda_*^2} - \frac{1}{3D}\right) \frac{\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2}{n} + O(1)\right),$$

where we absorb the diagonal terms from $\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2/n$ into the $O(1)$ term. Finally, by our assumption we have $\lambda < \lambda_* + \frac{1}{20D}$. Therefore, $\lambda^2 < \lambda_*^2 + \frac{41}{400D}$. So, $\lambda^2(\lambda_*^{-2} - \frac{1}{3D}) < 1 - \frac{1}{6D} + \frac{41}{400D} < 1 - \frac{1}{20D}$, and thus, for sufficiently large n, we will have

$$\le \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2} \exp\left(\frac{1}{2}\left(1 - \frac{1}{20D}\right) \frac{\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2}{n} + O(1)\right). \tag{69}$$

This is precisely the quantity arising in the analysis of computationally-unbounded strong detection for the Gaussian Rademacher-spiked matrix model in Perry et al. (2018) (invoking Le Cam's second moment method as mentioned above), where it is shown that this quantity is bounded as $n \to \infty$, since the factor multiplying $\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2/n$ is strictly smaller than $\frac{1}{2}$. (This makes formal the heuristic argument that $\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle/\sqrt{n}$ converges to a standard Gaussian random variable, whereby the above is asymptotically an evaluation of the moment generating function of a χ^2 random variable, which we also alluded to in Section B.1.) Thus we find

$$\limsup_{n\to\infty} \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^N \\ |\boldsymbol{k}|_\infty \le D}} |\langle L_n, \widehat{P}_{\boldsymbol{k}}\rangle|^2 < +\infty, \tag{70}$$

as claimed. ∎

Remark 36 (A general "Rademacher trick") *Step 1 in the proof above, where we take advantage of the Rademacher prior to decouple the dependence of the likelihood ratio's components on the signal-to-noise ratio λ from that on the actual spike vector $\boldsymbol{x}$, should apply in much greater generality. Indeed, we expect a similar property to hold in any additive model where (1) the spike distribution $\boldsymbol{x} \sim \mathcal{P}_n$ has the property that $|x_i| = \lambda(n)$ for some constant $\lambda(n)$ for all $i \in [N(n)]$, and (2) the noise distribution is symmetric, whereby the polynomials playing the role of $\widehat{\tau}_k$ will be even polynomials for even k and odd polynomials for odd k. Thus a similar analysis is likely possible in a wide range of models with "flat" signals $\boldsymbol{x}$, reducing the low-degree analysis to analytic questions about $\widehat{\tau}_k$.*

Remark 37 *The argument of Perry et al. (2018) derives the critical value λ_* for $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda)$ in terms of the Fisher information in the family of translates of the distribution ρ^{sech}, while our calculation, if we consider $D = D(n)$ growing slowly, obtains the same value using orthogonal polynomials. It appears that the connection between these derivations lies in the summation identity $\sum_{\ell \geq 0} \frac{1}{(2\ell+1)^2} = \frac{\pi^2}{8}$. We suspect that there are similar identities associated to these two approaches to calculating the critical signal-to-noise ratio in $\mathsf{Wig}(\rho, \lambda)$ for other well-behaved measures ρ. It would be interesting to understand what class of summation identities arises in this way, and whether equating these two derivations can give novel proofs of such identities.*

B.4. Mixed spiked matrix model: Proof of Theorem 19

This result follows almost immediately from Theorem 17 upon expanding the definitions.

Proof (of Theorem 19) Recall that ρ^{heavy} has density proportional to $(1+x^2)^{-\alpha/2}$, and we write $\mathsf{Wig}(\rho^{\mathsf{sech}}, \lambda_* \cdot \lambda) =: ((\mathbb{Q}_n^{(1)}, \mathbb{P}_n^{(1)}))_{n=1}^{\infty}$ and $\mathsf{Wig}(\rho^{\mathsf{heavy}}, \lambda) =: ((\mathbb{Q}_n^{(2)}, \mathbb{P}_n^{(2)}))_{n=1}^{\infty}$. Note that, since $\mathbb{E}_{x \sim \rho^{\mathsf{heavy}}} |x|^\beta = +\infty$ for all $\beta \geq \alpha - 1$, all polynomials in $L^2(\mathbb{Q}_n^{(2)})$ have entrywise degree at most $\alpha/2$ in every coordinate.

Since $\mathbb{Q}_n$ is a mixture of $\mathbb{Q}_n^{(1)}$ and $\mathbb{Q}_n^{(2)}$ with weight $\frac{1}{2}$ for each, we have $L^2(\mathbb{Q}_n) = L^2(\mathbb{Q}_n^{(1)}) \cap L^2(\mathbb{Q}_n^{(2)})$ and $\mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n} f_n(\boldsymbol{Y})^2 = \frac{1}{2}\mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n^{(1)}} f_n(\boldsymbol{Y})^2 + \frac{1}{2}\mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n^{(2)}} f_n(\boldsymbol{Y})^2$. Therefore,

$$
\begin{aligned}
&\left\{ \begin{array}{ll} \text{maximize} & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{P}_n} f_n(\boldsymbol{Y}) \\ \text{subject to} & f_n \in \mathbb{R}[\boldsymbol{Y}] \cap L^2(\mathbb{Q}_n) \\ & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n} f_n(\boldsymbol{Y})^2 = 1 \end{array} \right\} \\
&\qquad \leq \left\{ \begin{array}{ll} \text{maximize} & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{P}_n^{(1)}} f_n(\boldsymbol{Y}) \\ \text{subject to} & f_n \in \mathbb{R}[\boldsymbol{Y}] \setminus \{0\} \\ & \mathsf{deg}_{\{i,j\}}(f_n) \leq \alpha/2 \text{ for all } \{i,j\} \in \binom{[n]}{2} \\ & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n^{(1)}} f_n(\boldsymbol{Y})^2 \leq 2 \end{array} \right\} \\
&\qquad \leq \left\{ \begin{array}{ll} \text{maximize} & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{P}_n^{(1)}} f_n(\boldsymbol{Y}) \\ \text{subject to} & f_n \in \mathbb{R}[\boldsymbol{Y}] \\ & \mathsf{deg}_{\{i,j\}}(f_n) \leq \alpha/2 \text{ for all } \{i,j\} \in \binom{[n]}{2} \\ & \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_n^{(1)}} f_n(\boldsymbol{Y})^2 = 1 \end{array} \right\} \cdot \sqrt{2}, \qquad (71)
\end{aligned}
$$

where the final expression is precisely that controlled by Theorem 17, and the result follows upon taking $D = \alpha/2$ in that result. ∎