Computational Hardness of Certifying Bounds on Constrained PCA Problems

Afonso S. Bandeira
Dept. of Mathematics, ETH Zurich, Switzerland
https://people.math.ethz.ch/~abandeira/
bandeira@math.ethz.ch

Dmitriy Kunisky
Dept. of Mathematics, Courant Institute of Mathematical Sciences, New York University, USA
http://www.kunisky.com/
kunisky@cims.nyu.edu

Alexander S. Wein
Dept. of Mathematics, Courant Institute of Mathematical Sciences, New York University, USA
https://cims.nyu.edu/~aw128/
awein@cims.nyu.edu

Abstract

Given a random $n \times n$ symmetric matrix $\boldsymbol{W}$ drawn from the Gaussian orthogonal ensemble (GOE), we consider the problem of certifying an upper bound on the maximum value of the quadratic form $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$ over all vectors $\boldsymbol{x}$ in a constraint set $\mathcal{S} \subset \mathbb{R}^n$. For a certain class of normalized constraint sets we show that, conditional on a certain complexity-theoretic conjecture, no polynomial-time algorithm can certify a better upper bound than the largest eigenvalue of $\boldsymbol{W}$. A notable special case included in our results is the hypercube $\mathcal{S} = \{\pm 1/\sqrt{n}\}^n$, which corresponds to the problem of certifying bounds on the Hamiltonian of the Sherrington-Kirkpatrick spin glass model from statistical physics. Our results suggest a striking gap between optimization and certification for this problem.

Our proof proceeds in two steps. First, we give a reduction from the detection problem in the *negatively-spiked Wishart model* to the above certification problem. We then give evidence that this Wishart detection problem is computationally hard below the classical spectral threshold, by showing that no low-degree polynomial can (in expectation) distinguish the spiked and unspiked models. This method for predicting computational thresholds was proposed in a sequence of recent works on the sum-of-squares hierarchy, and is conjectured to be correct for a large class of problems. Our proof can be seen as constructing a distribution over symmetric matrices that appears computationally indistinguishable from the GOE, yet is supported on matrices whose maximum quadratic form over $\boldsymbol{x} \in \mathcal{S}$ is much larger than that of a GOE matrix.

2012 ACM Subject Classification Theory of computation → Computational complexity and cryptography

Keywords and phrases Certification, Sherrington-Kirkpatrick model, spiked Wishart model, low-degree likelihood ratio

Digital Object Identifier 10.4230/LIPIcs.ITCS.2020.78

Funding *Afonso S. Bandeira*: Part of this work was done while with the Courant Institute of Mathematical Sciences and the Center for Data Science at New York University and supported by NSF grants DMS-1712730, DMS-1719545, and by a grant from the Sloan Foundation.
Dmitriy Kunisky: Partially supported by NSF grants DMS-1712730 and DMS-1719545.
Alexander S. Wein: Partially supported by NSF grant DMS-1712730 and by the Simons Collaboration on Algorithms and Geometry.

Acknowledgements We thank Andrea Montanari and Samuel B. Hopkins for insightful discussions.

11th Innovations in Theoretical Computer Science Conference (ITCS 2020).
Editor: Thomas Vidick; Article No. 78; pp. 78:1–78:29

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

An important phenomenon in the study of the computational aspects of random problems is the appearance of *statistical-to-computational gaps*, wherein a problem may be solved by an inefficient algorithm – typically a brute-force search – but empirical evidence, heuristic formal calculations, and negative results for classes of powerful algorithms all suggest that the same problem cannot be solved by any algorithm running in polynomial time. Many examples of this phenomenon arise from Bayesian estimation tasks, in which the goal is to recover a planted *signal* from noisy observations. Bayesian problems exhibiting statistical-to-computational gaps in certain regimes include graph problems such as community detection [16], estimation for models of structured matrices and tensors [42, 30], statistical problems arising from imaging and microscopy tasks [53, 10], and many others. A different family of examples comes from random optimization problems that are *signal-free*, where there is no "planted" structure to recover; rather, the task is simply to optimize a random objective function as effectively as possible. Notable instances of problems of this kind that exhibit statistical-to-computational gaps include finding a large clique in a random graph [33], finding a submatrix of large entries of a random matrix [25], or finding an approximate solution to a random constraint satisfaction problem [1].

In this paper, we study a problem from the latter class, namely the problem of maximizing the quadratic form $\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}$ over a constraint set $\boldsymbol{x} \in \mathcal{S} \subset \mathbb{R}^n$, where $\boldsymbol{W}$ is a random matrix drawn from the Gaussian orthogonal ensemble,[1] $\boldsymbol{W} \sim \mathsf{GOE}(n)$. Unlike previous works that have studied whether an efficient algorithm can *optimize* and find $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{W})$ that achieves a large objective value, we study whether an efficient algorithm can *certify* an upper bound on the objective over all $\boldsymbol{x} \in \mathcal{S}$. In the notable case of the *Sherrington-Kirkpatrick (SK) Hamiltonian* [59, 49], where $\mathcal{S} = \{\pm 1/\sqrt{n}\}^n$, while there is an efficient algorithm believed to optimize arbitrarily close to the true maximum [46], we give evidence – based on the *low-degree likelihood ratio* recently studied in the sum-of-squares literature [9, 31, 29, 28], which we describe in Section 2.4 – that there is *no* efficient algorithm to certify an upper bound that improves on a simple spectral certificate. Thus, the certification task for this problem appears to exhibit a statistical-to-computational gap, while the optimization task does not.

1.1 Computational tasks in random optimization problems

To formalize the above discussion, consider a generic random optimization problem:

$$\begin{array}{ll} \text{maximize} & f_\omega(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{S} \\ \text{where} & \omega \sim \mathbb{P}, \end{array} \tag{1}$$

for $\mathbb{P}$ a probability distribution over some set of problem instances Ω. We will contrast two important computational tasks in this setting. The first, most obvious task is that of *optimization*, producing an algorithm that computes $\mathsf{alg}_{\mathsf{opt}} : \Omega \to \mathcal{S}$ such that $f_\omega(\mathsf{alg}_{\mathsf{opt}}(\omega))$ is as large as possible (say, in expectation, or with high probability as the size of the problem diverges).

Another task is that of *certification*, producing instead an algorithm that computes a scalar $\mathsf{alg}_{\mathsf{cert}} : \Omega \to \mathbb{R}$, such that *for all* $\omega \in \Omega$ and all $\boldsymbol{x} \in \mathcal{S}$ we have $f_\omega(\boldsymbol{x}) \le \mathsf{alg}_{\mathsf{cert}}(\omega)$. The main challenge specific to certification is that $\mathsf{alg}_{\mathsf{cert}}$ must produce a valid upper bound on

[1] Gaussian orthogonal ensemble (GOE): $\boldsymbol{W}$ is symmetric with $W_{ij} = W_{ji} \sim \mathcal{N}(0, 1/n)$ for $i \neq j$ and $W_{ii} \sim \mathcal{N}(0, 2/n)$, all independent.

f_ω for *every* possible instance $\omega \in \Omega$, no matter how unlikely ω is to occur under $\mathbb{P}$. Subject to this requirement, we seek to minimize $\mathsf{alg}_{\mathsf{cert}}(\omega)$ (again, in a suitable probabilistic sense when $\omega \sim \mathbb{P}$). *Convex relaxations* are a common approach to certification, where $\mathcal{S}$ is relaxed to a convex superset $\mathcal{S}' \supset \mathcal{S}$ over which it is possible to optimize exactly and efficiently using convex optimization. Often, such algorithms admit an alternative interpretation of *proving* a bound on $f_\omega(\boldsymbol{x})$ in some limited proof system (see, e.g., [27] for such discussion of sum-of-squares algorithms).

If $\boldsymbol{x}^\star = \boldsymbol{x}^\star(\omega)$ is the true maximizer of f_ω, then for any pair of optimization and certification algorithms as above, we have

$$f_\omega(\mathsf{alg}_{\mathsf{opt}}(\omega)) \leq f_\omega(\boldsymbol{x}^\star) \leq \mathsf{alg}_{\mathsf{cert}}(\omega). \tag{2}$$

Thus, in the case of a maximization problem, optimization algorithms approximate the true value $f_\omega(\boldsymbol{x}^\star)$ from below, while certification algorithms approximate it from above. We are then interested in how tight either inequality can be for random problems of growing dimension. Of course, we can achieve "perfect" optimization and certification (equality on either side of (2)) by exhaustive search over all $\boldsymbol{x} \in \mathcal{S}$, but we are interested in whether this is still possible when we restrict our attention to computationally-efficient algorithms.

To make these definitions concrete, we review an instance of each type of algorithm for the problem of optimizing the Sherrington-Kirkpatrick Hamiltonian.

▶ **Example 1.1.** The "SK problem" is the random optimization problem

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n \\ \text{where} & \boldsymbol{W} \sim \mathsf{GOE}(n). \end{array} \tag{3}$$

Here, two related *spectral algorithms* give simple examples of algorithms for both optimization and certification. For certification, writing $\lambda_{\max}$ for the largest eigenvalue of $\boldsymbol{W}$, we may use the bound

$$\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \leq \lambda_{\max} \cdot \|\boldsymbol{x}\|^2 = \lambda_{\max} \approx 2 \tag{4}$$

for all $\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n$, whereby $\lambda_{\max}$ is a certifiable upper bound on (3). From classical random matrix theory (see, e.g., [5]), it is known that $\lambda_{\max} \approx 2$ as $n \to \infty$.

For optimization, for $\boldsymbol{v}_{\max}$ the eigenvector of $\lambda_{\max}$, we may take $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{W}) := \mathrm{sgn}(\boldsymbol{v}_{\max})/\sqrt{n}$ where sgn denotes the $\{\pm 1\}$-valued sign function, applied entrywise. The vector $\boldsymbol{v}_{\max}$ is distributed as an uniform random unit vector in $\mathbb{R}^n$, so the quality of this solution may be computed as

$$\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} = \lambda_{\max} \cdot \langle \boldsymbol{x}, \boldsymbol{v}_{\max} \rangle^2 + O\left(\frac{1}{\sqrt{n}}\right) = \lambda_{\max} \cdot \frac{\|\boldsymbol{v}_{\max}\|_1^2}{n} + O\left(\frac{1}{\sqrt{n}}\right) \approx \frac{4}{\pi} \approx 1.2732 \tag{5}$$

with high probability as $n \to \infty$. (The error in the first equation is obtained as $\sum_i \lambda_i \langle \boldsymbol{v}_i, \boldsymbol{x} \rangle^2 \approx \frac{1}{n} \mathrm{Tr}(\boldsymbol{W})(1 - \langle \boldsymbol{v}_{\max}, \boldsymbol{x} \rangle^2)$, where the sum is over all eigenvectors $\boldsymbol{v}_i$ except $\boldsymbol{v}_{\max}$. This analysis appeared in [3], an early rigorous mathematical work on the SK model.)

On the other hand, deep results of statistical physics imply that the true optimal value approaches

$$\boldsymbol{x}^{\star\top} \boldsymbol{W} \boldsymbol{x}^\star \approx 2\mathsf{P}_* \approx 1.5264 \tag{6}$$

as $n \to \infty$, where the constant P_* is expressed via the celebrated Parisi formula for the free energy of the SK model [50, 49, 62]. The approximate value we give above was estimated with numerical experiments in previous works (see, e.g., [51, 15]).

Thus, neither for optimization nor for certification does the naive spectral algorithm achieve the optimal value, which suggests the question: can more sophisticated algorithms improve on the spectral algorithm for either task? For optimization, the recent result of [46] implies, assuming a widely-believed conjecture from statistical physics, that for any $\varepsilon > 0$ there exists a polynomial-time optimization algorithm achieving with high probability a value of $2\mathsf{P}_* - \varepsilon$ on the SK problem.[2] On the other hand, there are few results addressing certification in the SK problem. The only previous work we are aware of in this direction is [48], where a simple semidefinite programming relaxation (which coincides with degree-2 sum-of-squares) is shown to achieve the same value as the spectral certificate (4). More recently (after the initial appearance of this paper) the same was shown for degree-4 sum-of-squares [39, 45].

1.2 Our contributions

The main result of this paper, which we now state informally, gives formal evidence that for the SK certification problem, the simple spectral certificate (4) is optimal among efficient algorithms. See Corollary 3.8 for the precise statement.

▶ **Theorem 1.2** (Informal). *Conditional on the correctness of the low-degree likelihood ratio method (see Section 2.4), for any $\varepsilon > 0$, there is no polynomial-time algorithm that certifies the upper bound $2 - \varepsilon$ on the SK problem* (3) *with probability $1 - o(1)$ as $n \to \infty$.*

In fact, we expect that there is not even a subexponential-time algorithm; see Remark 3.6. Theorem 1.2 reveals a striking gap between optimization and certification: it is possible to efficiently give a tight lower bound on the maximum objective value by exhibiting a solution $\boldsymbol{x}$, but it seems impossible to efficiently give a tight upper bound. In other words, an algorithm can efficiently find a near-optimal solution, but cannot be sure that it has done so. The same result also holds for a wide variety of constraints other than $\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n$ (see Corollary 3.8). Due to the high-dimensional setting of the problem, we expect the value of an optimal certification algorithm to concentrate tightly; thus we expect Theorem 1.2 to still hold if $1 - o(1)$ is replaced by any positive constant.

Our result has important consequences for convex programming. A natural approach for optimizing the SK problem (3) would be to use a convex programming relaxation such as a semidefinite program based on the sum-of-squares hierarchy [60, 52, 41] (see [55] for a survey). Such a method would relax the constraints of the SK problem to weaker ones for which the associated optimization problem can be solved efficiently. One can either hope that the relaxation is *tight* and gives a valid solution $\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n$ (with high probability), or use a *rounding* procedure to extract a valid solution from the relaxation. The optimal value of any convex relaxation of (3) provides an upper bound on the optimal value of (3) and therefore gives a certification algorithm. Thus Theorem 1.2 implies that (conditional on the correctness of the low-degree likelihood ratio method) no polynomial-time convex relaxation of (3) can have value $\leq 2 - \varepsilon$ (resolving a question posed by [32]) and in particular cannot be tight.[3] As a result, we expect that natural relax-and-round approaches for optimization should fail to find a solution of value close to $2\mathsf{P}_*$. This would suggest a fundamental weakness of convex programs: even the most powerful convex programs (such as sum-of-squares relaxations)

[2] The work [46] builds on that of [2, 61], and these works taken together formalize the heuristic idea from statistical physics that optimization is tractable for certain optimization problems exhibiting *full replica symmetry breaking*.

[3] Our results suggest that $\Omega(n^{1-o(1)})$ rounds of sum-of-squares are required to certify a value $2 - \varepsilon$; see Remark 3.6.

seem to fail to optimize (3), even though other methods succeed (namely, the message-passing algorithm of [46]).[4] An explanation for this suboptimality is that convex relaxations are actually solving a fundamentally harder problem: certification.

1.3 Related work

A related example of an optimization–certification gap comes from random constraint satisfaction problems (CSPs).

▶ **Example 1.3.** In random MAX-3SAT, the decision variable is a boolean vector $\boldsymbol{x} \in \{0,1\}^n$, and the optimization task is to maximize the number of satisfied *clauses* $C_1, \ldots, C_m$, each of which is a boolean expression of the form $C_i = a_{i_1} \vee a_{i_2} \vee a_{i_3}$ where each a_{i_j} is chosen uniformly among the x_i and their boolean negations. Let $s_{\boldsymbol{C}}(\boldsymbol{x})$ denote the number of clauses satisfied by $\boldsymbol{x}$.

If $m/n \to \infty$ as $n \to \infty$, the optimal value $\max_{\boldsymbol{x}} s_{\boldsymbol{C}}(\boldsymbol{x})$ is $(7/8 + o(1))m$ with probability $1 - o(1)$ [11]. This is achieved by the trivial optimization algorithm that chooses a uniformly random assignment $\boldsymbol{x}$. On the other hand, sum-of-squares lower bounds suggest that it is hard to certify even $s_{\boldsymbol{C}}(\boldsymbol{x}) < m$ unless $m \gg n^{3/2}$ [26, 58, 38]. Along similar lines, a well-known conjecture of Feige asserts this cannot be certified (by *any* certification algorithm) in polynomial time for m/n an arbitrarily large constant [22].

As in the SK problem, there is an efficient algorithm for near-perfect optimization, while there does not seem to be such an algorithm for near-perfect certification. However, here the optimization algorithm is trivial (a random guess), so arguably a more natural optimization task would be to achieve the best possible advantage over random guessing, assessing the quality of a solution on a finer scale.

Prior work has used *sum-of-squares lower bounds* to argue for hardness of certification in problems such as random CSPs [26, 58, 38], planted clique [19, 44, 9], tensor injective norm [30, 29], graph coloring [8], community detection in hypergraphs [37], and others. These results prove that the sum-of-squares hierarchy (at some degree) fails to certify. If sum-of-squares fails at every constant degree (e.g., [9, 38, 29]), this suggests that all polynomial-time algorithms should also fail. In our case, it appears difficult to prove such sum-of-squares lower bounds for the SK problem, although recent work (appearing after the initial version of this paper) has shown lower bounds at degree 4 [39, 45]. We instead take a new approach based on a related heuristic for computational hardness, which we explain in Section 2.4. One advantage of this approach over sum-of-squares is that it is substantially simpler. Perhaps the prior work that is closest to our approach is [63], which also gives a reduction from a hypothesis testing problem to a certification problem.

Overview of techniques

The proof of Theorem 1.2 has two parts. First, we give a reduction from hypothesis testing in the *negatively-spiked Wishart model* [34, 6, 7, 54] to the SK certification problem. We then use a method introduced in the sum-of-squares literature based on the *low-degree likelihood ratio* [9, 31, 29, 28] to give formal evidence that detection in that negatively-spiked Wishart model is computationally hard.

[4] In contrast, simple rounded convex relaxations are believed to approximate many similar problems optimally in the worst-case (rather than average-case) setting [36].

In the spiked Wishart model, we observe either N i.i.d. samples $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N \sim \mathcal{N}(0, \boldsymbol{I}_n)$, or N i.i.d. samples $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N \sim \mathcal{N}(0, \boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top)$ where the "spike" $\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n$ is a uniformly random hypercube vector, and $\beta \in [-1, \infty)$. The goal is to distinguish between these two cases with probability $1 - o(1)$ as $n \to \infty$. In the *negatively-spiked* ($\beta < 0$) case with $\beta \approx -1$, this task amounts to deciding whether there is a hypercube vector $\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n$ that is nearly orthogonal to all of the samples $\boldsymbol{y}_i$. When $N = \Theta(n)$, a simple spectral method succeeds when $\beta^2 > n/N$ [6, 7], and we expect the problem to be computationally hard when $\beta^2 < n/N$.

Let us now intuitively explain the relation between the negatively-spiked Wishart model and the SK certification problem. Suppose we want to certify that

$$\mathsf{SK}(\boldsymbol{W}) := \max_{\boldsymbol{x} \in \{\pm 1/\sqrt{n}\}^n} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \leq 2 - \varepsilon$$

where $\boldsymbol{W} \sim \mathsf{GOE}(n)$, for some small constant $\varepsilon > 0$. Since the eigenvalues of $\boldsymbol{W}$ approximately follow the semicircle distribution on $[-2, 2]$ [64], we need to certify that the top δn-dimensional eigenspace of $\boldsymbol{W}$ does not (approximately) contain a hypercube vector, for a small $\delta > 0$ depending on ε. In particular, we need to distinguish a uniformly random δn-dimensional subspace (the distribution of the actual top eigenspace of $\boldsymbol{W} \sim \mathsf{GOE}(n)$) from a δn-dimensional subspace that contains a hypercube vector. Equivalently, taking orthogonal complements, we need to distinguish a uniformly random $(1-\delta)n$-dimensional subspace from a $(1-\delta)n$-dimensional subspace that is orthogonal to a hypercube vector. This is the problem of detection in the negatively-spiked Wishart model with $\beta \approx -1$ and $N = (1-\delta)n$, and these parameters lie in the "hard regime" $\beta^2 < n/N$.

Formally, we construct a distribution $\mathcal{D}(n)$ over $n \times n$ symmetric matrices with $\mathsf{SK}(\boldsymbol{W}) \geq 2 - \varepsilon/2$ when $\boldsymbol{W} \sim \mathcal{D}(n)$. This $\mathcal{D}(n)$ also has the property that, conditional on the hardness of the above detection problem, it is computationally hard to distinguish $\boldsymbol{W} \sim \mathcal{D}(n)$ from $\boldsymbol{W} \sim \mathsf{GOE}(n)$. The existence of such $\mathcal{D}(n)$ implies hardness of certification for the SK problem, because if an algorithm could certify that $\mathsf{SK}(\boldsymbol{W}) \leq 2 - \varepsilon$ when $\boldsymbol{W} \sim \mathsf{GOE}(n)$, then it could distinguish $\mathcal{D}(n)$ from $\mathsf{GOE}(n)$.

Borrowing terminology from [65, 66], we refer to this idea of "planting" a hidden solution (in our case, a hypercube vector $\boldsymbol{x}$) in such a way that it is difficult to detect, as *quiet planting*[5]. Our quiet planting scheme $\mathcal{D}(n)$ draws $\boldsymbol{W} \sim \mathsf{GOE}(n)$ and then rotates the top eigenspace of $\boldsymbol{W}$ to align with a random hypercube vector $\boldsymbol{x}$, while leaving the eigenvalues of $\boldsymbol{W}$ unchanged. (The more straightforward planting scheme, $\boldsymbol{W} + (2 - \varepsilon/2)\boldsymbol{x}\boldsymbol{x}^\top$ with $\boldsymbol{W} \sim \mathsf{GOE}(n)$, is not quiet because it changes the largest eigenvalue of $\boldsymbol{W}$ [23].) The question of how to design optimal quiet planting schemes in general remains an interesting open problem.

The final ingredient in our proof is to give formal evidence (in the form of the low-degree likelihood ratio) that detection in the spiked Wishart model is computationally hard below the spectral threshold. This consists of a calculation involving the projection of the likelihood ratio (between the "null" and "planted" distributions) onto the subspace of low-degree polynomials. This method suggests that the correct strategy for quiet planting is to match the low-degree moments of the distributions $\mathcal{D}(n)$ and $\mathsf{GOE}(n)$. We discuss the details of this method further in Section 2.4.

Our results on hardness in the spiked Wishart model may be of independent interest: our low-degree calculations suggest that, for a large class of spike priors, no polynomial-time algorithm can successfully distinguish the spiked and unspiked models below the classical

[5] However, our notion of quiet planting is not quite the same as that of [65, 66].

spectral threshold [6, 7], both in the negatively-spiked and positively-spiked regimes. (For positive spikes there was existing evidence for this based on failure of approximate message passing [24]; no such evidence was known for negative spikes.)

2 Background

2.1 Probability Theory

Our asymptotic notation (e.g., $O(\cdot), o(\cdot)$) always pertains to the limit $n \to \infty$. Parameters of the problem (e.g., $\beta, \gamma, \mathcal{X}, \mathcal{S}$) are held fixed as $n \to \infty$. Thus the constants hidden by $O(\cdot)$ and $o(\cdot)$ do not depend on n but may depend on the other parameters. When A_n is a sequence of events in probability spaces with measures $\mathbb{P}_n$, we say A_n holds *with high probability* if $\mathbb{P}_n[A_n] = 1 - o(1)$.

▶ **Definition 2.1.** *A real-valued random variable π with $\mathbb{E}[\pi] = 0$ is* subgaussian *if there exists $\sigma^2 \geq 0$ (the* variance proxy*) such that, for all $t \in \mathbb{R}$, $M(t) := \mathbb{E}[\exp(t\pi)] \leq \exp(\sigma^2 t^2/2)$.*

The name *subgaussian* refers to the fact that if $\pi \sim \mathcal{N}(0, \sigma^2)$, then $M(t) = \exp(\sigma^2 t^2/2)$. A random variable with law $\mathcal{N}(0, \sigma^2)$ is therefore subgaussian. Any bounded centered random variable is also subgaussian: if $\pi \in [a, b]$ almost surely, then π is subgaussian with $\sigma^2 = \frac{1}{4}(b-a)^2$ (see, e.g., [56]).

We next give some background facts from random matrix theory (see, e.g., [5]).

▶ **Definition 2.2.** *The* Gaussian orthogonal ensemble $\mathsf{GOE}(n)$ *is a probability distribution over symmetric matrices $\boldsymbol{W} \in \mathbb{R}^{n \times n}$, under which $W_{ii} \sim \mathcal{N}(0, 2/n)$ and $W_{ij} \sim \mathcal{N}(0, 1/n)$ when $i \neq j$, where the entries W_{ij} are independent for distinct pairs (i, j) with $i \leq j$.*

Our scaling of the entries of $\mathsf{GOE}(n)$ ensures a spectrum of constant width.

▶ **Proposition 2.3.** *Let $\boldsymbol{W}_n \sim \mathsf{GOE}(n)$. Then, almost surely, $\lambda_{\min}(\boldsymbol{W}_n) \to -2$ and $\lambda_{\max}(\boldsymbol{W}_n) \to 2$ as $n \to \infty$. In particular, for any $\varepsilon > 0$, $\|\boldsymbol{W}_n\| \leq 2 + \varepsilon$ with high probability. Also, the empirical distribution of eigenvalues of $\boldsymbol{W}_n$ converges weakly to a semicircle distribution supported on $[-2, 2]$.*

2.2 Constrained PCA

▶ **Definition 2.4.** *A* constraint set *is a sequence $\mathcal{S} = (\mathcal{S}_n)_{n \in \mathbb{N}}$ where $\mathcal{S}_n \subset \mathbb{R}^n$. The* constrained principal component analysis (PCA) problem with constraint set $\mathcal{S}$*, denoted* $\mathsf{PCA}(\mathcal{S})$*, is*

$$\begin{array}{ll} \textit{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \textit{subject to} & \boldsymbol{x} \in \mathcal{S}_n \\ \textit{where} & \boldsymbol{W} \sim \mathsf{GOE}(n). \end{array}$$

We will work only with constraint sets supported on vectors of approximately unit norm. General problems of this kind have been studied previously in, e.g., [20].

▶ **Example 2.5.** Problems that may be described in the constrained PCA framework include:

- the Sherrington-Kirkpatrick (SK) spin glass model: $\mathcal{S}_n = \{\pm 1/\sqrt{n}\}^n$ [59, 49],
- the Wigner sparse PCA null model: $\mathcal{S}_n = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \|\boldsymbol{x}\| = 1, \|\boldsymbol{x}\|_0 \leq \rho\}$ [17, 43],
- the spherical $2p$-spin spin glass model: $\mathcal{S}_{pn} = \{\boldsymbol{x}^{\otimes p} \ : \ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}$ [14, 13],
- the positive PCA null model: $\mathcal{S}_n = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ x_i \geq 0, \|\boldsymbol{x}\| = 1\}$ [47].

Our results apply to the first two examples: the SK model, and sparse PCA when $\rho = \Theta(n)$.

▶ **Definition 2.6.** *Let f be a (randomized) algorithm*[6] *that takes a square matrix $\boldsymbol{W}$ as input and outputs a number $f(\boldsymbol{W}) \in \mathbb{R}$. We say that f* certifies *a value B on $\mathsf{PCA}(\mathcal{S})$ if*
1. *for* any *symmetric matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$, $\max_{\boldsymbol{x} \in \mathcal{S}_n} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \le f(\boldsymbol{W})$, and*
2. *if $\boldsymbol{W}_n \sim \mathsf{GOE}(n)$ then $f(\boldsymbol{W}_n) \le B + o(1)$ with high probability.*

2.3 Spiked Wishart Models

▶ **Definition 2.7.** *A* normalized spike prior *is a sequence $\mathcal{X} = (\mathcal{X}_n)_{n \in \mathbb{N}}$ where $\mathcal{X}_n$ is a probability distribution over $\mathbb{R}^n$, such that if $\boldsymbol{x} \sim \mathcal{X}_n$ then $\|\boldsymbol{x}\| \to 1$ in probability as $n \to \infty$.*

▶ **Definition 2.8** (Spiked Wishart model)**.** *Let $\mathcal{X}$ be a normalized spike prior, let $\gamma > 0$, and let $\beta \in [-1, \infty)$. Let $N = \lceil n/\gamma \rceil$. We define two probability distributions over $(\mathbb{R}^n)^N$:*
1. *Under $\mathbb{Q}$, the* null *model, draw $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ independently for $i \in [N]$.*
2. *Under $\mathbb{P}$, the* planted *model, draw $\boldsymbol{x} \sim \mathcal{X}_n$. If $\beta\|\boldsymbol{x}\|^2 \ge -1$, then draw $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top)$ independently for $i \in [N]$. Otherwise, draw $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ independently for $i \in [N]$.*

Taken together, $\mathbb{P}$ and $\mathbb{Q}$ form the spiked Wishart model $(\mathbb{P}, \mathbb{Q}) =: \mathsf{Wishart}(n, \gamma, \beta, \mathcal{X})$*. For fixed γ and β we denote the sequence $(\mathsf{Wishart}(n, \gamma, \beta, \mathcal{X}))_{n \in \mathbb{N}}$ by $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$.*

Several remarks on this definition are in order. First, we make the explicit choice $N = \lceil n/\gamma \rceil$ for concreteness, but our results apply to any choice of $N = N(n)$ for which $n/N \to \gamma$ as $n \to \infty$. Second, often the Wishart model is described in terms of the distribution of the sample covariance matrix $\frac{1}{N}\sum_{i=1}^N \boldsymbol{y}_i \boldsymbol{y}_i^\top$. We instead work directly with the samples $\boldsymbol{y}_i$ so as not to restrict ourselves to algorithms that only use the sample covariance matrix. (This modification only makes our results on computational hardness of detection more general.) Finally, the definition of $\mathbb{P}$ has two cases to ensure that the covariance matrix $\boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top$ is positive semidefinite. We will work in the setting $\beta > -1$ where the first case ($\beta\|\boldsymbol{x}\|^2 \ge -1$) occurs with high probability.

We consider the algorithmic task of distinguishing between $\mathbb{P}$ and $\mathbb{Q}$ in the following sense.

▶ **Definition 2.9.** *For sequences of distributions $\mathbb{P} = (\mathbb{P}_n)_{n \in \mathbb{N}}$ and $\mathbb{Q} = (\mathbb{Q}_n)_{n \in \mathbb{N}}$ over measurable spaces $(\Omega_n, \mathcal{F}_n)_{n \in \mathbb{N}}$, an algorithm $f_n : \Omega_n \to \{0, 1\}$ achieves* strong detection *between $\mathbb{P}$ and $\mathbb{Q}$ if*

$$\mathbb{Q}_n[f_n(\boldsymbol{y}) = 0] = 1 - o(1) \qquad \text{and} \qquad \mathbb{P}_n[f_n(\boldsymbol{y}) = 1] = 1 - o(1).$$

The celebrated BBP transition [6] implies a spectral algorithm for strong detection in the spiked Wishart model whenever $\beta^2 > \gamma$.

▶ **Theorem 2.10** ([6, 7])**.** *Let $\mathcal{X}$ be any normalized spike prior. If $\beta^2 > \gamma$ then there exists a polynomial-time algorithm for strong detection in $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$.*

The algorithm thresholds the largest eigenvalue (if $\beta > 0$) or smallest eigenvalue (if $\beta < 0$) of the sample covariance matrix $\frac{1}{N}\sum_{i=1}^N \boldsymbol{y}_i \boldsymbol{y}_i^\top$. This eigenvalue converges almost surely to a limiting value which is different under $\mathbb{P}$ and $\mathbb{Q}$.

[6] We allow f to be randomized; i.e., it may use randomness in its computations, but the output B must be an upper bound almost surely. We do not expect certification algorithms to require randomness, but it may be convenient, e.g., to randomly initialize an iterative optimization procedure.

We will give evidence that (see Corollary 3.5) if $\mathcal{X}$ has i.i.d. subgaussian entries with suitable scaling, then no polynomial-time algorithm achieves strong detection below the BBP threshold ($\beta^2 < \gamma$). Exponential-time strong detection is possible below the BBP threshold for some priors, such as i.i.d. Rademacher when $\beta < -0.84$ [54]. Very sparse priors with $\boldsymbol{x}$ supported on $O(\sqrt{n})$ entries give rise to the *sparse PCA* regime, where polynomial-time strong detection is possible below the BBP threshold [35, 4, 18]; our results will not apply in this setting (although see [29, 21] for some related work that addresses this regime).

2.4 The Low-Degree Likelihood Ratio

Inspired by the sum-of-squares hierarchy (e.g., [60, 52, 41]) and in particular the pseudo-calibration approach [9], recent works [31, 29, 28] have proposed a strikingly simple method for predicting computational hardness of Bayesian inference problems. This method recovers widely-conjectured computational thresholds for high-dimensional inference problems such as planted clique [9, 28], community detection [31, 28], sparse PCA [21], tensor PCA [29, 28, 40], and the spiked Wigner matrix model [40]. We now give an overview of this method (see also [40] for a survey).

Consider the problem of distinguishing two simple hypotheses $\mathbb{P}_n$ and $\mathbb{Q}_n$ which are probability distributions on some domain $\Omega_n = \mathbb{R}^{d(n)}$ (where typically the dimension $d(n)$ grows with n). One example is the spiked Wishart model $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$ for some fixed choice of the parameters $\beta, \gamma, \mathcal{X}$. The idea is to take low-degree polynomials as a proxy for polynomial-time algorithms and consider whether there are such polynomials that can distinguish $\mathbb{P}_n$ from $\mathbb{Q}_n$.

▶ **Definition 2.11.** *Let $D : \mathbb{N} \to \mathbb{N}$. We say that distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ is $D(n)$-low-degree easy if there exists a sequence of nonzero polynomials $f_n \in \mathbb{R}[y_1, \ldots, y_{d(n)}]$ with $\deg f_n \leq D(n)$ such that*

$$\lim_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{y}\sim\mathbb{P}_n} f_n(\boldsymbol{y})}{\sqrt{\mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_n} f_n(\boldsymbol{y})^2}} = +\infty, \tag{7}$$

and $D(n)$-low-degree hard otherwise.

We view $\mathbb{Q}_n$ as the "null" distribution, which is often i.i.d. Gaussian (as in the Wishart model) or i.i.d. Rademacher (± 1-valued). $\mathbb{Q}_n$ induces an inner product on L^2 functions $f : \Omega_n \to \mathbb{R}$ given by $\langle f, g\rangle_{L^2(\mathbb{Q}_n)} = \mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_n}[f(\boldsymbol{y})g(\boldsymbol{y})]$, and a norm $\|f\|^2_{L^2(\mathbb{Q}_n)} = \langle f, f\rangle_{L^2(\mathbb{Q}_n)}$. For $D \in \mathbb{N}$, let $\mathbb{R}[\boldsymbol{y}]_{\leq D}$ denote the polynomials $\Omega_n \to \mathbb{R}$ of degree at most D. For $f : \Omega_n \to \mathbb{R}$, let $f^{\leq D}$ denote the orthogonal projection (with respect to $\langle\cdot,\cdot\rangle_{L^2(\mathbb{Q}_n)}$) of f onto $\mathbb{R}[\boldsymbol{y}]_{\leq D}$. The following relates low-degree hardness to the *low-degree likelihood ratio*.

▶ **Theorem 2.12** ([31])**.** *Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be probability distributions on Ω_n for each $n \in \mathbb{N}$. Suppose $\mathbb{P}_n$ is absolutely continuous with respect to $\mathbb{Q}_n$, so that the likelihood ratio $L_n = \frac{d\mathbb{P}_n}{d\mathbb{Q}_n}$ is defined. Then*

$$\max_{f\in\mathbb{R}[\boldsymbol{y}]_{\leq D}\setminus\{0\}} \frac{\mathbb{E}_{\boldsymbol{y}\sim\mathbb{P}_n} f(\boldsymbol{y})}{\sqrt{\mathbb{E}_{\boldsymbol{y}\sim\mathbb{Q}_n} f(\boldsymbol{y})^2}} = \|L_n^{\leq D}\|_{L^2(\mathbb{Q}_n)}. \tag{8}$$

Proof. The objective can be rewritten as $\langle f, L_n\rangle_{L^2(\mathbb{Q}_n)}/\|f\|_{L^2(\mathbb{Q}_n)}$, so by basic Hilbert space theory, the maximum is attained by taking $f = L_n^{\leq D}$. ◀

▶ **Corollary 2.13.** *In the setting of Theorem 2.12,*

1. *if $\|L_n^{\leq D(n)}\|_{L^2(\mathbb{Q}_n)} = O(1)$, then distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ is $D(n)$-low-degree hard;*
2. *if $\|L_n^{\leq D(n)}\|_{L^2(\mathbb{Q}_n)} = \omega(1)$, then distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ is $D(n)$-low-degree easy.*

We take $O(\log n)$-degree polynomials $\Omega_n \to \mathbb{R}$ as a proxy for functions computable in polynomial-time. One justification for this is that many polynomial-time algorithms compute the leading eigenvalue of a matrix $\boldsymbol{M}$ whose entries are constant-degree polynomials in the data; in fact, there is formal evidence that such *low-degree spectral methods* are as powerful as the sum-of-squares hierarchy [29]. Typically, $O(\log n)$ rounds of power iteration are sufficient to compute the leading eigenvalue accurately, which amounts to evaluating the $O(\log n)$-degree polynomial $\mathrm{Tr}(\boldsymbol{M}^{2q})$ for some $q = O(\log n)$. This argument can be made formal: if $\|L_n^{\leq D}\|_{L^2(\mathbb{Q}_n)} = O(1)$ then all low-degree spectral methods must fail in a certain sense [40]. This motivates the following informal conjecture, which is based on [31, 29, 28], particularly Conjecture 2.2.4 of [28].

▶ **Conjecture 2.14** (Informal). *For "nice" distributions $\mathbb{P}_n$ and $\mathbb{Q}_n$, if distinguishing $\mathbb{P}_n$ and $\mathbb{Q}_n$ is $\log^{1+\Omega(1)}(n)$-low-degree hard, then there is no randomized polynomial-time algorithm for strong detection between $\mathbb{P}$ and $\mathbb{Q}$.*

This conjecture is useful because the norm of the low-degree likelihood ratio, $\|L_n^{\leq D}\|_{L^2(\mathbb{Q}_n)}$, can be computed (or at least bounded) for various distributions such as the stochastic block model [31] and the spiked tensor model [29, 28]. More generally, Hypothesis 2.1.5 of [28] conjectures that degree-D polynomials are a proxy for time-$n^{\widetilde{\Theta}(D)}$ algorithms.

▶ Remark 2.15. We do not expect the converse of Conjecture 2.14 to hold. If detection is $O(\log n)$-low-degree easy then we expect an $n^{O(\log n)}$-time algorithm but not necessarily a polynomial-time algorithm, because not every $O(\log n)$-degree polynomial can be evaluated in polynomial time.

Conjecture 2.14 is informal in that we do not specify what is meant by "nice" distributions. See Conjecture 2.2.4 of [28] for a precise variant of Conjecture 2.14; however, this variant uses the more refined notion of *coordinate degree* and so does not apply to the calculations we will perform. Roughly speaking, "nice" distributions $\mathbb{P}$ and $\mathbb{Q}$ should satisfy the following:

1. $\mathbb{Q}$ should be a product distribution, e.g., i.i.d. Gaussian or i.i.d. Rademacher;
2. $\mathbb{P}$ should be sufficiently symmetric with respect to permutations of its coordinates; and
3. we should be able to add a small amount of noise to $\mathbb{P}$, ruling out distributions with brittle algebraic structure (such as random satisfiable instances of XOR-SAT, which can be identified using Gaussian elimination [12]).

We refer the reader to [31, 28, 40] for further details and evidence in favor of Conjecture 2.14.

3 Main Results

3.1 Spiked Wishart Models

We now study the low-degree hardness of the spiked Wishart model. The following technical definitions will be important to specify the priors to which our results apply.

▶ **Definition 3.1.** *Let $\beta \in (-1, \infty)$ and let $\mathcal{X}$ be a normalized spike prior. We say that $\mathcal{X}$ is* β-good *if when $\boldsymbol{x} \sim \mathcal{X}_n$ then $\beta\|\boldsymbol{x}\|^2 > -1$ almost surely.*

We consider spike priors having i.i.d. entries, and will sometimes need to slightly modify the spike prior to ensure that it is β-good and has bounded norm.

▶ **Definition 3.2.** *Let π be a probability distribution over $\mathbb{R}$ such that $\mathbb{E}[\pi] = 0$ and $\mathbb{E}[\pi^2] = 1$. Let $\mathsf{iid}(\pi/\sqrt{n})$ denote the normalized spike prior $\mathcal{X} = (\mathcal{X}_n)$ that draws each entry of $\boldsymbol{x}$ independently from $\frac{1}{\sqrt{n}}\pi$. (We do not allow π to depend on n.)*

▶ **Definition 3.3.** *For a normalized spike prior $\mathcal{X}$, let the β-truncation* $\mathrm{trunc}_\beta(\mathcal{X})$ *of $\mathcal{X}$ denote the following normalized spike prior. To sample $\boldsymbol{x}$ from* $(\mathrm{trunc}_\beta(\mathcal{X}))_n$*, first sample $\boldsymbol{x}' \sim \mathcal{X}_n$. Then, let $\boldsymbol{x} = \boldsymbol{x}'$ if $\beta\|\boldsymbol{x}'\|^2 > -1$ and $\|\boldsymbol{x}'\|^2 \le 2$, and let $\boldsymbol{x} = \boldsymbol{0}$ otherwise.*

If $\beta > -1$ then since $\mathcal{X}$ is normalized ($\|\boldsymbol{x}'\| \to 1$ in probability), the first case of Definition 3.3 occurs with high probability. The upper bound $\|\boldsymbol{x}'\| \le 2$ is for technical convenience, and the constant 2 is not essential. Note also that the β-truncation of an i.i.d. prior is no longer i.i.d.

▶ **Theorem 3.4.** *Fix constants $\gamma > 0$ and $\beta > -1$.*

1. *Suppose $\beta^2 < \gamma$. Let $\mathcal{X} = \mathrm{trunc}_\beta(\mathsf{iid}(\pi/\sqrt{n}))$ where π is subgaussian with $\mathbb{E}[\pi] = 0$ and $\mathbb{E}[\pi^2] = 1$. For any $D = o(n/\log n)$, distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ is $D(n)$-low-degree hard.*
2. *Suppose $\beta^2 > \gamma$. Let $\mathcal{X} = \mathsf{iid}(\pi/\sqrt{n})$ be β-good with π symmetric about zero, $\mathbb{E}[\pi] = 0$, and $\mathbb{E}[\pi^2] = 1$. For any $D = \omega(1)$, distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ is $D(n)$-low-degree easy.*

We prove Theorem 3.4 in Section 5. Part 1 of Theorem 3.4, combined with Conjecture 2.14, suggests that for i.i.d. subgaussian priors, strong detection is hard below the BBP threshold.

▶ **Corollary 3.5.** *Suppose Conjecture 2.14 holds for the spiked Wishart model. Fix constants $\gamma > 0$ and $\beta > -1$. Let π be subgaussian with $\mathbb{E}[\pi] = 0$ and $\mathbb{E}[\pi^2] = 1$. Let $\mathcal{X}$ be either $\mathsf{iid}(\pi/\sqrt{n})$ or $\mathrm{trunc}_\beta(\mathsf{iid}(\pi/\sqrt{n}))$. If $\beta^2 < \gamma$, then there is no randomized polynomial-time algorithm for strong detection in $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$.*

Proof. The case $\mathcal{X} = \mathrm{trunc}_\beta(\mathsf{iid}(\pi/\sqrt{n}))$ follows immediately from Part 1 of Theorem 3.4. If strong detection is impossible for $\mathcal{X} = \mathrm{trunc}_\beta(\mathsf{iid}(\pi/\sqrt{n}))$, then strong detection is also impossible for $\mathcal{X} = \mathsf{iid}(\pi/\sqrt{n})$, as these two spike priors differ with probability $o(1)$ (under the natural coupling). ◀

▶ Remark 3.6. We make some technical remarks regarding Theorem 3.4 and Corollary 3.5.

1. Even if Conjecture 2.14 does not hold, note that Theorem 3.4 still implies unconditional lower bounds against low-degree polynomials in the sense of Definition 2.11.
2. Part 2 of Theorem 3.4 serves only to check that we do not predict computational hardness when $\beta^2 > \gamma$ (as polynomial-time strong detection is possible in this regime; see Theorem 2.10). The assumption that π is symmetric about zero should not be essential.
3. In Part 1 of Theorem 3.4 and in Corollary 3.5, the requirement that $\mathcal{X}$ be a β-truncated i.i.d. prior can be relaxed. We only require that $\mathcal{X}$ is the β-truncation of a normalized prior admitting a *local Chernoff bound* (see Definition 5.11).
4. Part 1 of Theorem 3.4 holds for any $D = o(n/\log n)$, much larger than the $D = \log^{1+\Omega(1)}(n)$ required by Conjecture 2.14. Since Hypothesis 2.1.5 of [28] conjectures that degree-D polynomials are a proxy for $n^{\widetilde{\Theta}(D)}$-time algorithms [28], this suggests that the conclusion of Corollary 3.5 also holds for $2^{n^{1-\delta}}$-time algorithms, for any $\delta > 0$. In other words, strong detection requires nearly-exponential time.

3.2 Constrained PCA

The following result gives a reduction from strong detection in the spiked Wishart model to certification in the constrained PCA problem.

▶ **Theorem 3.7.** *Let $\mathcal{S}$ be a constraint set and let $\mathcal{X}$ be a normalized spike prior such that if $\boldsymbol{x} \sim \mathcal{X}_n$ then $\boldsymbol{x} \in \mathcal{S}_n$ with high probability. Suppose there exists $\varepsilon > 0$ and a randomized polynomial-time algorithm that certifies the value $2 - \varepsilon$ on $\mathsf{PCA}(\mathcal{S})$. Then there exist $\gamma > 1$ and $\beta \in (-1, 0)$ such that there is a randomized polynomial-time algorithm for strong detection in $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$.*

We give the proof in Section 4. Note for the parameters γ, β above, $\beta^2 < \gamma$ (the "hard regime").

▶ **Corollary 3.8.** *Suppose Conjecture 2.14 holds for the spiked Wishart model. Let π be subgaussian with $\mathbb{E}[\pi] = 0$ and $\mathbb{E}[\pi^2] = 1$. Let $\mathcal{S}$ be a constraint set such that, if $\boldsymbol{x} \sim \mathsf{iid}(\pi/\sqrt{n})$, then $\boldsymbol{x} \in \mathcal{S}_n$ with high probability. Then, for any $\varepsilon > 0$, there is no randomized polynomial-time algorithm to certify the value $2 - \varepsilon$ on $\mathsf{PCA}(\mathcal{S})$.*

Proof. The result is immediate from Theorem 3.4 and Theorem 3.7. ◀

In particular, we obtain the hardness of improving on the spectral certificate in the SK model.

▶ **Corollary 3.9.** *If Conjecture 2.14 holds for the spiked Wishart model, then for any $\varepsilon > 0$, there is no randomized polynomial-time algorithm to certify the value $2 - \varepsilon$ on the SK problem* (3)*.*

Proof. Apply Corollary 3.8 with π having the Rademacher distribution and $\mathcal{S}_n = \{\pm 1/\sqrt{n}\}^n$. ◀

4 Proof of Reduction from Spiked Wishart to Constrained PCA

Our proof will rely on the following crucial invariance property of $\mathsf{GOE}(n)$.

▶ **Proposition 4.1.** *For any orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(n)$, if $\boldsymbol{W} \sim \mathsf{GOE}(n)$, then the law of $\boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}^\top$ is also $\mathsf{GOE}(n)$.*

Proof of Theorem 3.7. Let $\mathcal{S}$ be a constraint set and let $\mathcal{X}$ be a normalized spike prior such that, if $\boldsymbol{x} \sim \mathcal{X}_n$, then $\boldsymbol{x} \in \mathcal{S}_n$ with high probability. Suppose that for some $\varepsilon > 0$ there is a randomized polynomial-time algorithm f that certifies the value $2 - \varepsilon$ on $\mathsf{PCA}(\mathcal{S})$. We will show that this implies that there is a polynomial-time algorithm for strong detection in $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$ with certain parameters $\gamma > 1$ and $\beta \in (-1, 0)$ (depending on ε). Note that these parameters lie in the "hard" regime $\beta^2 < \gamma$.

Our algorithm for detection in the Wishart model is as follows. Fix $\gamma > 1$, to be chosen later. Since $n/N \to \gamma$ we have $n > N$ (for sufficiently large n). Given samples $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top)$, let $V = \mathrm{span}\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\} \subseteq \mathbb{R}^n$ and let $V^\perp$ be its orthogonal complement. We sample $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ having the distribution $\mathsf{GOE}(n)$ conditioned on the event that the span of the top $n - N$ eigenvectors of $\boldsymbol{W}$ is $V^\perp$. Concretely, we can obtain a sample in the following way. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ be a uniformly random orthonormal basis for V and let $\boldsymbol{v}_{N+1}, \dots, \boldsymbol{v}_n$ be a uniformly random orthonormal basis for $V^\perp$. Sample $\boldsymbol{W}' \sim \mathsf{GOE}(n)$ and let $\lambda_1 < \cdots < \lambda_n$ be the eigenvalues of $\boldsymbol{W}'$. Then, let $\boldsymbol{W} := \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$. Finally, run the certification algorithm f for $\mathsf{PCA}(\mathcal{S})$ on $\boldsymbol{W}$. The detection algorithm $\widetilde{f} : (\mathbb{R}^n)^N \to \{0, 1\}$ then thresholds $f(\boldsymbol{W})$:

$$\widetilde{f}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) = \begin{cases} 0 \text{ (report } \mathbb{Q}_n) & \text{if } f(\boldsymbol{W}) \le 2 - \varepsilon/2, \\ 1 \text{ (report } \mathbb{P}_n) & \text{if } f(\boldsymbol{W}) > 2 - \varepsilon/2. \end{cases} \tag{9}$$

We now prove that $\widetilde{f}$ indeed achieves strong detection in $\mathsf{Wishart}(\gamma, \beta, \mathcal{X})$. First, if the samples $\boldsymbol{y}_i$ are drawn from the null model $\mathbb{Q}_n$, then V is a uniformly random N-dimensional subspace of $\mathbb{R}^n$, so by Proposition 4.1 the law of $\boldsymbol{W}$ constructed above is $\mathsf{GOE}(n)$. Thus $f(\boldsymbol{W}) \le 2 - \varepsilon/2$ with high probability by assumption, and therefore $\widetilde{f}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) = 0$ with high probability, i.e., our algorithm correctly reports that the samples were drawn from the null model.

Next, suppose the samples $\boldsymbol{y}_i$ are drawn from the planted model $\mathbb{P}_n$ with planted spike $\boldsymbol{x} \sim \mathcal{X}_n$. We will choose $\gamma > 1$ and $\beta \in (-1, 0)$ so that $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \geq 2 - \varepsilon/3$ with high probability. Since $\boldsymbol{x} \in \mathcal{S}_n$ with high probability, this would imply $f(\boldsymbol{W}) \geq 2 - \varepsilon/3$, so we will have $\widetilde{f}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) = 1$ with high probability, i.e., our algorithm will correctly report that the samples were drawn from the planted model.

It remains to show that $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \geq 2 - \varepsilon/3$. Let $\lambda_1 < \cdots < \lambda_n$ be the eigenvalues of $\boldsymbol{W}$ and let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ be the corresponding (unit-norm) eigenvectors. By Proposition 2.3, with high probability, for all $i \in [n]$, $\lambda_i \in [-2 - o(1), 2 + o(1)]$. Furthermore, by the semicircle law [64], with high probability, $\lambda_{N+1} \geq 2 - g(\gamma)$ where $g(\gamma) > 0$ is a function satisfying $g(\gamma) \to 0$ as $\gamma \to 1^+$ (recalling that $n/N \to \gamma$). Letting $\|\boldsymbol{x}\|_V$ denote the norm of the orthogonal projection of $\boldsymbol{x}$ onto V, we have, with high probability,

$$\begin{aligned}
\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} &= \boldsymbol{x}^\top \left(\sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \right) \boldsymbol{x} \\
&= \sum_{i=1}^n \lambda_i \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^2 \\
&\geq \lambda_1 \|\boldsymbol{x}\|_V^2 + \lambda_{N+1} \|\boldsymbol{x}\|_{V^\perp}^2 \\
&\geq (-2 - o(1)) \|\boldsymbol{x}\|_V^2 + (2 - g(\gamma))(\|\boldsymbol{x}\|^2 - \|\boldsymbol{x}\|_V^2) \\
&= (2 - g(\gamma)) \|\boldsymbol{x}\|^2 + (-4 + g(\gamma) - o(1)) \|\boldsymbol{x}\|_V^2 \\
&\geq 2 - g(\gamma) - 4\|\boldsymbol{x}\|_V^2 - o(1). \qquad (10)
\end{aligned}$$

Thus we need to upper bound $\|\boldsymbol{x}\|_V^2$. Let $\boldsymbol{P}_V$ denote the orthogonal projection matrix onto V. Since V is the span of $\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$, we have $\boldsymbol{P}_V \preceq \frac{1}{\mu} \boldsymbol{Y}$ where

$$\boldsymbol{Y} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{y}_i \boldsymbol{y}_i^\top$$

and μ is the smallest nonzero eigenvalue of $\boldsymbol{Y}$. (Here $\preceq$ denotes Loewner order.) Since $\boldsymbol{Y}$ is a spiked Wishart matrix, it follows from Theorem 1.2 of [7] that its smallest nonzero eigenvalue converges almost surely to $(1 - \sqrt{\gamma})^2$ as $n \to \infty$. Thus we have $\mu = (1 - \sqrt{\gamma})^2 + o(1)$. Therefore,

$$\|\boldsymbol{x}\|_V^2 = \|\boldsymbol{P}_V \boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{P}_V \boldsymbol{x} \leq \frac{1}{\mu} \boldsymbol{x}^\top \boldsymbol{Y} \boldsymbol{x} = \frac{1}{\mu N} \sum_{i=1}^N \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle^2.$$

We have $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top)$ and so $\langle \boldsymbol{x}, \boldsymbol{y}_i \rangle \sim \mathcal{N}(0, \boldsymbol{x}^\top (\boldsymbol{I}_n + \beta \boldsymbol{x}\boldsymbol{x}^\top) \boldsymbol{x}) = \mathcal{N}(0, \|\boldsymbol{x}\|^2 + \beta \|\boldsymbol{x}\|^4)$. Therefore, letting $a_N^2 = \sum_{i=1}^N g_i^2$ for g_i i.i.d. standard gaussian random variables, so that a_N^2 has the χ^2 distribution with N degrees of freedom, we have conditional on $\boldsymbol{x}$ the distributional equality

$$\boldsymbol{x}^\top \boldsymbol{Y} \boldsymbol{x} \stackrel{(d)}{=} (\|\boldsymbol{x}\|^2 + \beta \|\boldsymbol{x}\|^4) \frac{a_N^2}{N}.$$

Standard concentration inequalities imply $a_N^2 / N \in [1 - o(1), 1 + o(1)]$ with high probability, and therefore $\boldsymbol{x}^\top \boldsymbol{Y} \boldsymbol{x} = 1 + \beta + o(1)$ with high probability. Thus, with high probability, we find

$$\|\boldsymbol{x}\|_V^2 = \frac{1 + \beta}{(1 - \sqrt{\gamma})^2} + o(1). \qquad (11)$$

Finally, we choose $\gamma > 1$ close enough to 1 so that $g(\gamma) \le \varepsilon/8$. By (11), we can also choose $\beta \in (-1, 0)$ close enough to -1 so that $\|\boldsymbol{x}\|_V^2 \le \varepsilon/32$ with high probability. Combining these, from (10) it follows that $\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x} \ge 2 - \varepsilon/4 - o(1) \ge 2 - \varepsilon/3$ with high probability, completing the proof. ◀

▶ Remark 4.2. We remark that we have ignored issues of numerical precision by assuming a model of computation where eigendecomposition computations can be done exactly in polynomial time. However, we believe all the operations we have used are stable, so that our reduction should also hold for weaker models of computation. (In particular, if we want to compute polynomially-many bits of precision of the $\mathsf{PCA}(\mathcal{S})$ instance, this should require only polynomially-many bits of precision in the eigendecomposition computation.)

5 Proofs for Spiked Wishart Models

5.1 Preliminaries

Spiked Wishart model statistics

The following formulae pertaining to the spiked Wishart model are derived in [54]. (Recall that in the spiked Wishart model, the parameter N is determined by n and γ as $N = \lceil n/\gamma \rceil$.)

▶ **Proposition 5.1.** *Suppose $\gamma > 0$, $\beta \in [-1, \infty)$, and $\mathcal{X}$ is a β-good normalized spike prior. Then, the likelihood ratio of the null and planted probability distributions of Definition 2.8 is*

$$L_{n,\gamma,\beta,\mathcal{X}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) := \frac{d\mathbb{P}_n}{d\mathbb{Q}_n}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{X}_n} \left[\left(1 + \beta \|\boldsymbol{x}\|^2\right)^{-N/2} \prod_{i=1}^{N} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta\|\boldsymbol{x}\|^2} \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle^2 \right) \right]. \quad (12)$$

If furthermore $\|\boldsymbol{x}\|^2 < 1/|\beta|$ almost surely when $\boldsymbol{x} \sim \mathcal{X}_n$, then the second moment of the likelihood ratio is given by

$$\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathbb{Q}_n} \left(L_{n,\gamma,\beta,\mathcal{X}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N)\right)^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{X}_n} \left[(1 - \beta^2 \langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle^2)^{-N/2} \right] \quad (13)$$

where $\boldsymbol{x}^1, \boldsymbol{x}^2$ are drawn independently from $\mathcal{X}_n$.

Hermite polynomials

We recall the classical one-dimensional Hermite polynomials.

▶ **Definition 5.2.** *The polynomials $h_k \in \mathbb{R}[x]$ for $k \ge 0$ are defined by the recursion*

$$h_0(x) = 1,$$
$$h_{k+1}(x) = x h_k(x) - h_k'(x),$$

and we define normalized versions

$$\widehat{h}_k(x) = \frac{1}{\sqrt{k!}} h_k(x).$$

▶ **Proposition 5.3.** *The $\widehat{h}_k$ are an orthonormal polynomial system for the standard Gaussian measure:*

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0,1)} \left[\widehat{h}_k(g) \widehat{h}_\ell(g) \right] = \delta_{k\ell}.$$

Similarly, we define the product Hermite polynomials. It is helpful to first define some notations for vectors of indices, which will also be used in the later derivations.

▶ **Definition 5.4.** *Let* $\mathbb{N} = \{n \in \mathbb{Z} : n \geq 0\}$. *For* $\boldsymbol{\alpha} \in \mathbb{N}^n$ *and* $\boldsymbol{x} \in \mathbb{R}^n$, *let*

$$|\boldsymbol{\alpha}| := \sum_{i=1}^{n} \alpha_i,$$
$$\boldsymbol{\alpha}! := \prod_{i=1}^{n} \alpha_i!,$$
$$\boldsymbol{x}^{\boldsymbol{\alpha}} := \prod_{i=1}^{n} x_i^{\alpha_i}.$$

▶ **Definition 5.5.** *For* $\boldsymbol{\alpha} \in \mathbb{N}^n$ *and* $\boldsymbol{x} \in \mathbb{R}^n$,

$$H_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \prod_{i=1}^{n} h_{\alpha_i}(x_i),$$
$$\widehat{H}_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \prod_{i=1}^{n} \widehat{h}_{\alpha_i}(x_i) = \frac{1}{\sqrt{\boldsymbol{\alpha}!}} H_{\boldsymbol{\alpha}}(\boldsymbol{x}).$$

▶ **Proposition 5.6.** *The* $\widehat{H}_{\boldsymbol{\alpha}}$ *are an orthonormal polynomial system for the product measure of* n *standard Gaussian measures:*

$$\mathop{\mathbb{E}}_{\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)} \left[\widehat{H}_{\boldsymbol{\alpha}}(\boldsymbol{g}) \widehat{H}_{\boldsymbol{\beta}}(\boldsymbol{g}) \right] = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}.$$

Combinatorics

We will also need the (ordinary) generating function of the central binomial coefficients.

▶ **Proposition 5.7.** *For any* $x \in \mathbb{R}$ *with* $|x| < \frac{1}{4}$,

$$(1 - 4x)^{-1/2} = \sum_{k \geq 0} \binom{2k}{k} x^k.$$

5.2 Norm of the Low-Degree Projection

In this section, we describe the formulas for the norm of the low-degree likelihood ratio in the spiked Wishart model, $\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}$. The full calculations are given in Appendix A.

The following result, the main technical one of this portion of the argument, computes the norm of the projection of the likelihood ratio onto a single Hermite polynomial.

▶ **Lemma 5.8.** *Let* $\boldsymbol{\alpha} \in (\mathbb{N}^n)^N$, *and let* $\boldsymbol{\alpha}_i \in \mathbb{N}^n$ *denote the* i*th component. Let* $|\boldsymbol{\alpha}| = \sum_{i=1}^{N} |\boldsymbol{\alpha}_i|$. *Suppose* $\gamma > 0$, $\beta \in [-1, \infty)$, *and* $\mathcal{X}$ *is a* β*-good normalized spike prior. Then,*

$$\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}} \rangle_{L^2(\mathbb{Q}_n)}^2 = \begin{cases} \beta^{|\boldsymbol{\alpha}|} \cdot \prod_{i=1}^{N} \frac{(|\boldsymbol{\alpha}_i|-1)!!^2}{\boldsymbol{\alpha}_i!} \cdot \left(\mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}_n} \boldsymbol{x}^{\sum_{i=1}^{N} \boldsymbol{\alpha}_i} \right)^2 & \text{if } |\boldsymbol{\alpha}_i| \text{ even for all } i \in [N], \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where when $\beta = 0$ *and* $\boldsymbol{\alpha} = \boldsymbol{0}$ *we interpret* $0^0 = 1$.

Note in particular that the quantity in question does not depend on the sign of β; thus the calculation of the norm of the low-degree projection of the likelihood ratio will not distinguish between the positively and negatively spiked Wishart models. Interestingly, in our proof, which involves generalized Hermite polynomials that form families of orthogonal polynomials with respect to Gaussian measures of different variances, this corresponds to the fact that an "umbral" analogue of the Hermite polynomials corresponding to a fictitious Gaussian measure with *negative* variance satisfies many of the same identities as the ordinary Hermite polynomials.

Combining these quantities, we may give a simple description of the norm of the low-degree projection of the likelihood ratio.

▶ **Lemma 5.9.** *Suppose $\gamma > 0$, $\beta \in [-1, \infty)$, and $\mathcal{X}$ is a β-good normalized spike prior. Define*

$$\varphi_N(x) := (1-4x)^{-N/2}, \tag{15}$$

$$\varphi_{N,k}(x) := \sum_{d=0}^{k} x^d \sum_{\substack{d_1,\dots,d_N \\ \sum d_i = d}} \prod_{i=1}^{N} \binom{2d_i}{d_i}, \tag{16}$$

so that $\varphi_{N,k}(x)$ is the Taylor series of φ_N around $x = 0$ truncated to degree k (as may be justified by Proposition 5.7). Then,

$$\|L^{\le D}_{n,\gamma,\beta,\mathcal{X}}\|^2_{L^2(\mathbb{Q}_n)} = \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2 \sim \mathcal{X}_n} \left[\varphi_{N,\lfloor D/2 \rfloor} \left(\frac{\beta^2 \langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle^2}{4} \right) \right] \tag{17}$$

where $\boldsymbol{x}^1, \boldsymbol{x}^2$ are drawn independently from $\mathcal{X}$.

▶ Remark 5.10. The squared norm of the low-degree likelihood ratio (17) is closely related via Taylor expansion to the squared norm (or second moment) of the full likelihood ratio (13), which is recovered by taking $D \to \infty$ while n and N remain fixed.

5.3 Asymptotics as $n \to \infty$

In this section, we use the formula from Lemma 5.9 to prove Part 1 of Theorem 3.4 (the case $\beta^2 < \gamma$). The proof of Part 2 ($\beta^2 > \gamma$) is deferred to Appendix B.4.

The following concentration result is the key property that we require from the spike prior $\mathcal{X}$.

▶ **Definition 5.11.** *A normalized spike prior $\mathcal{X}$ admits a* local Chernoff bound *if for every $\eta > 0$ there exist $\delta > 0$ and $C > 0$ such that, for all n,*

$$\Pr\left\{ |\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle| \ge t \right\} \le C \exp\left(-\frac{1}{2}(1-\eta) n t^2 \right) \quad \text{for all } t \in [0, \delta] \tag{18}$$

where $\boldsymbol{x}^1, \boldsymbol{x}^2$ are drawn independently from $\mathcal{X}_n$.

▶ **Proposition 5.12.** *If π is subgaussian with $\mathbb{E}[\pi] = 0$ and $\mathbb{E}[\pi^2] = 1$ then $\mathsf{iid}(\pi/\sqrt{n})$ and $\mathrm{trunc}_\beta(\mathsf{iid}(\pi/\sqrt{n}))$ (for any $\beta > -1$) each admit a local Chernoff bound.*

We defer the proof to Appendix B.1.

Proof of Theorem 3.4 (Part 1). Let $\beta^2 < \gamma$. We decompose the norm of the low-degree likelihood ratio into two parts, which we will bound separately:

$$\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right] = R_1 + R_2$$

where

$$R_1 := \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right],$$

$$R_2 := \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|>\varepsilon}\,\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right].$$

Here $\varepsilon > 0$ is a small constant to be chosen later. We call R_1 the *small deviations* and call R_2 the *large deviations*.

The following two lemmas bound these two terms, respectively. First, we bound the large deviations.

▶ **Lemma 5.13** (Large Deviations). *Let $\beta^2 < \gamma$. Suppose $\mathcal{X}$ is a β-good normalized spike prior that admits a local Chernoff bound. Suppose that for any n, $\boldsymbol{x} \sim \mathcal{X}_n$ satisfies $\|\boldsymbol{x}\|^2 \leq 2$ almost surely. If $D = o(n/\log n)$ and $\varepsilon > 0$ is any constant, then $R_2 = o(1)$.*

We give a proof summary, with the full proof deferred to Appendix B.2. Since $\|\boldsymbol{x}^1\|^2 \leq 2$ and $\|\boldsymbol{x}^2\|^2 \leq 2$,

$$R_2 \leq \Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle| > \varepsilon\right\}\varphi_{N,\lfloor D/2\rfloor}(\beta^2).$$

By the local Chernoff bound, $\Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle| > \varepsilon\right\}$ decays exponentially in n. To complete the proof, we use elementary combinatorial bounds to control the polynomial expression (16) for $\varphi_{N,\lfloor D/2\rfloor}(\beta^2)$. Its growth is roughly of order $O(n^D)$, which is counteracted by the exponential decay of $\Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle| > \varepsilon\right\}$ so long as $D = o(n/\log n)$.

Next, we bound the small deviations. For this part of the argument, it is irrelevant that the likelihood ratio is truncated to its low-degree component, and we essentially reuse an existing argument for the full likelihood ratio from [54].

▶ **Lemma 5.14** (Small Deviations). *Let $\beta^2 < \gamma$. Suppose $\mathcal{X}$ is a β-good normalized spike prior that admits a local Chernoff bound. Let $D = D(n)$ be any function of n. If $\varepsilon > 0$ is a sufficiently small constant then $R_1 = O(1)$.*

We again give a proof summary, with the full proof deferred to Appendix B.3. As mentioned above, unlike in the proof of Lemma 5.13, here we simply bound $\varphi_{N,\lfloor D/2\rfloor} \leq \varphi_N$ in the expression for R_1. To bound the resulting expression, we borrow an argument from [54]. This step crucially uses the local Chernoff bound, and amounts to showing that the exponential decay from the Chernoff bound sufficiently counteracts the exponential growth of the likelihood ratio term $\varphi_N(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2)$ when $\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle$ is small.

Combining Proposition 5.12 with Lemmas 5.13 and 5.14 completes the proof of Part 1 of Theorem 3.4. (The proof of Part 2 is deferred to Appendix B.4.) ◀

References

1 Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 793–802. IEEE, 2008.

2 Louigi Addario-Berry and Pascal Maillard. The algorithmic hardness threshold for continuous random energy models. *arXiv preprint*, 2018. arXiv:1810.05129.

3 Michael Aizenman, Joel L Lebowitz, and David Ruelle. Some rigorous results on the Sherrington-Kirkpatrick spin glass model. *Communications in mathematical physics*, 112(1):3–20, 1987.

4 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.

5 Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. An introduction to random matrices, 2010.

6 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.

7 Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

8 Jess Banks, Robert Kleinberg, and Cristopher Moore. The Lovász Theta Function for Random Regular Graphs and Community Detection in the Hard Regime. *arXiv preprint*, 2017. arXiv:1705.01194.

9 Boaz Barak, Samuel B Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 428–437. IEEE, 2016.

10 Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer. Heterogeneous multireference alignment: A single pass approach. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pages 1–6. IEEE, 2018.

11 Amin Coja-Oghlan, Andreas Goerdt, and André Lanka. Strong refutation heuristics for random k-SAT. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 310–321. Springer, 2004.

12 Nadia Creignou and Hervé Daude. Satisfiability threshold for random XOR-CNF formulas. *Discrete Applied Mathematics*, 96:41–53, 1999.

13 A Crisanti, H Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.

14 A Crisanti and HJ Sommers. The spherical p-spin interaction spin glass model: the statics. *Phys. B*, 87:341, 1992.

15 Andrea Crisanti and Tommaso Rizzo. Analysis of the infinity-replica symmetry breaking solution of the Sherrington-Kirkpatrick model. *Physical Review E*, 65(4):046137, 2002.

16 Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

17 Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. *arXiv preprint*, 2014. arXiv:1402.2238.

18 Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

19 Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Conference on Learning Theory*, pages 523–562, 2015.

20 Yash Deshpande, Andrea Montanari, and Emile Richard. Cone-constrained principal component analysis. In *Advances in Neural Information Processing Systems*, pages 2717–2725, 2014.

21 Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-Time Algorithms for Sparse PCA. *arXiv preprint*, 2019. arXiv:1907.11635.

22 Uriel Feige. Relations between average case complexity and approximation complexity. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 534–543. ACM, 2002.

23 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.

24 Alyson K Fletcher and Sundeep Rangan. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562, 2018.

25 David Gamarnik and Quan Li. Finding a large submatrix of a Gaussian random matrix. *The Annals of Statistics*, 46(6A):2511–2561, 2018.

26 Dima Grigoriev. Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. *Theoretical Computer Science*, 259(1-2):613–622, 2001.

27 Dima Grigoriev and Nicolai Vorobjov. Complexity of Null-and Positivstellensatz proofs. *Annals of Pure and Applied Logic*, 113(1-3):153–160, 2001.

28 Samuel Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

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

30 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.

31 Samuel B Hopkins and David Steurer. Bayesian estimation from few samples: community detection and related problems. *arXiv preprint*, 2017. `arXiv:1710.00264`.

32 Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. *arXiv preprint*, 2018. `arXiv:1808.07226`.

33 Mark Jerrum. Large cliques elude the Metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

34 Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

35 Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 7, 2004.

36 Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

37 Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Community detection in hypergraphs, spiked tensor models, and Sum-of-Squares. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 124–128. IEEE, 2017.

38 Pravesh K Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer. Sum of squares lower bounds for refuting any CSP. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 132–145. ACM, 2017.

39 Dmitriy Kunisky and Afonso S Bandeira. A Tight Degree 4 Sum-of-Squares Lower Bound for the Sherrington-Kirkpatrick Hamiltonian. *arXiv preprint*, 2019. `arXiv:1907.11686`.

40 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*, 2019. `arXiv:1907.11636`.

41 Jean B Lasserre. An explicit exact SDP relaxation for nonlinear 0-1 programs. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 293–303. Springer, 2001.

42 Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. *arXiv preprint*, 2015. `arXiv:1507.03857`.

43 Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse PCA. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1635–1639. IEEE, 2015.

44 Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 87–96. ACM, 2015.

45 Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting Sum-of-Squares Lower Bounds: Degree-2 to Degree-4, 2019. `arXiv:1911.01411`.

46 Andrea Montanari. Optimization of the Sherrington-Kirkpatrick Hamiltonian. *arXiv preprint*, 2018. `arXiv:1812.10897`.

47 Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3):1458–1484, 2016.

48 Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. *arXiv preprint*, 2015. `arXiv:1504.05910`.

49 Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

50 Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

51 Giorgio Parisi. A sequence of approximated solutions to the SK model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

52 Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

53 Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-Passing Algorithms for Synchronization Problems over Compact Groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

54 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.

55 Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint*, 2018. `arXiv:1807.11419`.

56 Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *Lecture notes*, 2018.

57 Steven Roman. *The umbral calculus*. Springer, 2005.

58 Grant Schoenebeck. Linear level Lasserre lower bounds for certain k-CSPs. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 593–602. IEEE, 2008.

59 David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

60 Naum Z Shor. Class of global minimum bounds of polynomial functions. *Cybernetics*, 23(6):731–734, 1987.

61 Eliran Subag. Following the ground-states of full-RSB spherical spin glasses. *arXiv preprint*, 2018. `arXiv:1812.04588`.

62 Michel Talagrand. The Parisi formula. *Annals of mathematics*, pages 221–263, 2006.

63 Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of RIP certification. In *Advances in Neural Information Processing Systems*, pages 3819–3827, 2016.

64 Eugene P Wigner. Characteristic vectors of bordered matrices with infinite dimensions I. In *The Collected Works of Eugene Paul Wigner*, pages 524–540. Springer, 1993.

65 Lenka Zdeborova and Florent Krzakala. Hiding quiet solutions in random constraint satisfaction problems. *Physical Review Letters*, 102(LA-UR-08-08090; LA-UR-08-8090), 2008.

66 Lenka Zdeborová and Florent Krzakala. Quiet planting in the locked constraint satisfaction problems. *SIAM Journal on Discrete Mathematics*, 25(2):750–770, 2011.

A Proofs for Computing the Low-Degree Likelihood Ratio

A.1 Generalized and Umbral Hermite Polynomials

We introduce some calculations with a useful generalization of the Hermite polynomials. While the usual Hermite polynomials are a family of orthogonal polynomials for the Gaussian measure with variance 1, and a straightforward generalization yields orthogonal polynomials for the Gaussian measure with any positive variance, we will use the surprising further generalization to fictitious Gaussian measures with *negative* variance, as described by the so-called *umbral calculus*. We follow the presentation of [57] (specifically, Section 2.1 of Chapter 4).

▶ **Definition A.1.** *For any $v \in \mathbb{R}$, the* Hermite polynomials with variance v *are defined by the recursion*

$$h_0(x; v) = 1, \tag{19}$$
$$h_{k+1}(x; v) = x h_k(x; v) - v \partial_x [h_k](x; v). \tag{20}$$

The next facts are useful for translating between different versions of the basic recursion and other properties of the Hermite polynomials.

▶ **Proposition A.2** (Differentiation Identity)**.** *For any $v, x \in \mathbb{R}$,*

$$\partial_x [h_k](x; v) = k h_{k-1}(x; v). \tag{21}$$

▶ **Proposition A.3** (Alternate Recursion)**.** *For any $v \in \mathbb{R}$, the Hermite polynomials are equivalently defined by the recursion*

$$h_0(x; v) = 1, \tag{22}$$
$$h_{k+1}(x; v) = x h_k(x; v) - v k h_{k-1}(x; v). \tag{23}$$

The following is yet another common way of defining the Hermite polynomials, in terms of the derivatives of the corresponding Gaussian density (or, in the negative variance case, a suitable generalization thereof).

▶ **Proposition A.4** (Rodrigues Formula)**.** *Let $v \in \mathbb{R}$ with $v \neq 0$. Then,*

$$\frac{d^k}{dx^k}\left[\exp\left(-\frac{1}{2v}x^2\right)\right] = (-v)^{-k} h_k(x; v) \exp\left(-\frac{1}{2v}x^2\right). \tag{24}$$

The next fact shows how the generalized Hermite polynomials transform under scaling.

▶ **Proposition A.5** (Scaling Identity)**.** *Let $v, w, x \in \mathbb{R}$, then*

$$h_k(wx; v) = w^k h_k\left(x; \frac{v}{w^2}\right). \tag{25}$$

Finally, the following is a generalized version of Gaussian integration by parts. We only provide the version of this identity for the standard Hermite polynomials, which is the only one we will use, but analogous statements hold for the generalized and umbral Hermite polynomials.

▶ **Proposition A.6** (Integration by Parts)**.** *Let $f \in \mathcal{C}^k(\mathbb{R})$ have $|f^{(i)}(x)| \leq e^{Cx}$ for all $i \in \{0, 1, \ldots, k\}$ and some $C > 0$. Then,*

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0,1)} [h_k(g; 1) f(g)] = \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0,1)} \left[f^{(k)}(g)\right]. \tag{26}$$

While the above results are standard, we now give two results we will use in our calculation that do not seem to appear explicitly in the previous literature, although they are straightforward to obtain from the preceding facts. First, we will use the following slightly more general version of the Rodgrigues formula (Proposition A.4) in our calculations.

▶ **Proposition A.7** (Multidimensional Rodrigues Formula). *Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{\alpha} \in \mathbb{N}^n$, and $v \in \mathbb{R}$ with $v \neq 0$. Then,*

$$\partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}} \left[\exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \right] = (-v)^{-|\boldsymbol{\alpha}|} \boldsymbol{x}^{\boldsymbol{\alpha}} h_{|\boldsymbol{\alpha}|}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) \exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right). \tag{27}$$

Proof. We proceed by induction on $|\boldsymbol{\alpha}|$. Clearly the result holds for $\boldsymbol{\alpha} = \mathbf{0}$. Suppose the result holds for all $|\boldsymbol{\alpha}'| \leq k$, and $|\boldsymbol{\alpha}| = k + 1 > 0$. Let $\boldsymbol{\alpha}'$ having $|\boldsymbol{\alpha}'| = k$ differ from $\boldsymbol{\alpha}$ only in coordinate i, so that $\alpha'_i = \alpha_i - 1$ and $\alpha'_j = \alpha_j$ for all $j \neq i$. Then,

$$\begin{aligned}
&\partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}} \left[\exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \right] \\
&\quad = \partial_{y_i} \left[\partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}'} \left[\exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \right] \right] \\
&\quad = (-v)^{-k} \boldsymbol{x}^{\boldsymbol{\alpha}'} \partial_{y_i} \left[h_k(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) \exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \right] \\
&\quad = (-v)^{-k} \boldsymbol{x}^{\boldsymbol{\alpha}'} \left(x_i \partial_x[h_k](\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) - v^{-1} \langle \boldsymbol{x}, \boldsymbol{y} \rangle x_i h_k(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) \right) \exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \\
&\quad = (-v)^{-(k+1)} \boldsymbol{x}^{\boldsymbol{\alpha}} \left(\langle \boldsymbol{x}, \boldsymbol{y} \rangle h_k(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) - v \partial_x[h_k](\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) \right) \exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \\
&\quad = (-v)^{-(k+1)} \boldsymbol{x}^{\boldsymbol{\alpha}} h_{k+1}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; v) \exp\left(-\frac{1}{2v} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right),
\end{aligned}$$

completing the proof. ◀

Second, we will need the following calculation evaluating the expectation of any Hermite polynomial under any centered Gaussian measure.

▶ **Proposition A.8** (Expectation Under Mismatched Variance). *Let $v \in \mathbb{R}$ and $k \geq 0$. Then,*

$$\mathbb{E}_{g \sim \mathcal{N}(0, \sigma^2)} \left[h_k(g; v) \right] = \begin{cases} 0 & \text{if } k \text{ odd,} \\ (k-1)!!(\sigma^2 - v)^{k/2} & \text{if } k \text{ even.} \end{cases}$$

Proof. The result for odd k holds since $h_k(\cdot\,; v)$ is an odd function for any $v \in \mathbb{R}$ in this case. For even k, we argue by induction on k. The result clearly holds for $k = 0$. If the result holds for a given k, then we may compute

$$\begin{aligned}
\mathbb{E}_{g \sim \mathcal{N}(0,\sigma^2)}[h_{k+2}(g; v)] &= \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_{k+2}(\sigma g)] \\
&= \sigma \mathbb{E}_{g \sim \mathcal{N}(0,1)}[g h_{k+1}(\sigma g; v)] - v(k+1) \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_k(\sigma g)] \\
&= \sigma^2 \mathbb{E}_{g \sim \mathcal{N}(0,1)}[\partial_x[h_{k+1}](\sigma g; v)] - v(k+1) \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_k(\sigma g)] \\
&= (k+1)\sigma^2 \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_k(\sigma g; v)] - v(k+1) \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_k(\sigma g)] \\
&= (k+1)(\sigma^2 - v) \mathbb{E}_{g \sim \mathcal{N}(0,1)}[h_k(\sigma g)] \\
&= (k+1)(\sigma^2 - v) \mathbb{E}_{g \sim \mathcal{N}(0,\sigma^2)}[h_k(g)]
\end{aligned}$$

completing the proof. ◀

We note two interesting features of this result. First, it generalizes two simple cases, on the one hand $v = \sigma^2$ where the expectation is zero unless $k = 0$, as may be seen from the orthogonality relations, and on the other $v = 0$ where it recovers the moments of a Gaussian measure. Second, the quantities appearing on the right-hand side formally resemble the moments of a Gaussian measure of suitable variance, but the formula in fact still holds for $\sigma^2 < v$, in which case case these quantities may be viewed as the moments of a fictitious Gaussian measure of negative variance (the same as inspired the umbral Hermite polynomials).

A.2 Individual Hermite Components of the Likelihood Ratio

Proof of Lemma 5.8. By Proposition A.6, we find

$$\begin{aligned}
&\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}} \rangle^2 \\
&= \frac{1}{\boldsymbol{\alpha}!} \left(\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathbb{Q}_n} \partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}} L_{n,\gamma,\beta,\mathcal{X}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) \right)^2 \\
&= \frac{1}{\boldsymbol{\alpha}!} \left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{X}_n, \boldsymbol{y} \sim \mathbb{Q}_n} \left(1 + \beta \|\boldsymbol{x}\|^2\right)^{-N/2} \partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}} \prod_{i=1}^N \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|\boldsymbol{x}\|^2} \langle \boldsymbol{x}, \boldsymbol{y}_i \rangle^2 \right) \right)^2 \\
&= \frac{1}{\prod_{i=1}^N \boldsymbol{\alpha}_i!} \left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{X}_n} \left(1 + \beta \|\boldsymbol{x}\|^2\right)^{-N/2} \prod_{i=1}^N \mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)} \left[\partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}_i} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|\boldsymbol{x}\|^2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \right] \right)^2
\end{aligned} \tag{28}$$

where the $\boldsymbol{\alpha}_i \in \mathbb{N}^n$ are the components of $\boldsymbol{\alpha}$ corresponding to $\boldsymbol{y}_i$, for each $i \in [N]$. When $\beta = 0$, our result follows from the above, giving $\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}} \rangle^2 = \delta_{0,|\boldsymbol{\alpha}|}$. (Indeed, in this case the null and planted models are identical, so $L_{n,\gamma,\beta,\mathcal{X}} = 1$ is a constant, which is compatible with the above.) Let us suppose $\beta \neq 0$ below.

In this case, using Proposition A.7, we have

$$\begin{aligned}
&\partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}_i} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|\boldsymbol{x}\|^2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right) \\
&\quad = \left(\frac{1 + \beta \|\boldsymbol{x}\|^2}{\beta} \right)^{-|\boldsymbol{\alpha}_i|} \boldsymbol{x}^{\boldsymbol{\alpha}_i} h_{|\boldsymbol{\alpha}_i|}\left(\langle \boldsymbol{x}, \boldsymbol{y} \rangle; -\frac{1 + \beta \|\boldsymbol{x}\|^2}{\beta} \right) \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|\boldsymbol{x}\|^2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \right).
\end{aligned} \tag{29}$$

(Note that the sign of the spike, or equivalently the sign of β, is the opposite of the sign of the variance of the Hermite polynomials that appear; thus, it is the negatively spiked case that corresponds to the more natural positive variance Hermite polynomials.) Since when $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ then $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \sim \mathcal{N}(0, \|\boldsymbol{x}\|^2)$, we find

$$\begin{aligned}
\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}} \rangle^2 = &\frac{\beta^{2|\boldsymbol{\alpha}|}}{\prod_{i=1}^N \boldsymbol{\alpha}_i!} \left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{X}_n} \frac{\boldsymbol{x}^{\sum_{i=1}^N \boldsymbol{\alpha}_i}}{(1 + \beta \|\boldsymbol{x}\|^2)^{|\boldsymbol{\alpha}| + N/2}} \right. \\
&\left. \prod_{i=1}^N \mathop{\mathbb{E}}_{g \sim \mathcal{N}(\boldsymbol{0}, \|\boldsymbol{x}\|^2)} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1 + \beta \|\boldsymbol{x}\|^2}{\beta} \right) \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|\boldsymbol{x}\|^2} g^2 \right) \right)^2.
\end{aligned} \tag{30}$$

We next focus on the innermost expectation. We may rewrite:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{g\sim\mathcal{N}(\mathbf{0},\|\boldsymbol{x}\|^2)} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1+\beta\|\boldsymbol{x}\|^2}{\beta}\right)\exp\left(\frac{1}{2}\frac{\beta}{1+\beta\|\boldsymbol{x}\|^2}g^2\right)\\
&= \frac{1}{\sqrt{2\pi\|\boldsymbol{x}\|^2}}\int_{-\infty}^{\infty} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1+\beta\|\boldsymbol{x}\|^2}{\beta}\right)\exp\left(-\frac{1}{2}\left(\frac{1}{\|\boldsymbol{x}\|^2}-\frac{\beta}{1+\beta\|\boldsymbol{x}\|^2}\right)g^2\right)dg\\
&= \frac{(1+\beta\|\boldsymbol{x}\|^2)^{1/2}}{\sqrt{2\pi\|\boldsymbol{x}\|^2(1+\beta\|\boldsymbol{x}\|^2)}}\int_{-\infty}^{\infty} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1+\beta\|\boldsymbol{x}\|^2}{\beta}\right)\exp\left(-\frac{1}{2\|\boldsymbol{x}\|^2(1+\beta\|\boldsymbol{x}\|^2)}g^2\right)dg\\
&= (1+\beta\|\boldsymbol{x}\|^2)^{1/2}\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\|\boldsymbol{x}\|^2(1+\beta\|\boldsymbol{x}\|^2))} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1+\beta\|\boldsymbol{x}\|^2}{\beta}\right). \qquad (31)
\end{aligned}$$

By Proposition A.8, this quantity will be zero unless $|\boldsymbol{\alpha}_i|$ is even, and thus $\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}}\rangle^2$ will be zero unless $|\boldsymbol{\alpha}_i|$ is even for all i. In this case, by Proposition A.8,

$$\mathop{\mathbb{E}}_{g\sim\mathcal{N}(\mathbf{0},\|\boldsymbol{x}\|^2)} h_{|\boldsymbol{\alpha}_i|}\left(g; -\frac{1+\beta\|\boldsymbol{x}\|^2}{\beta}\right)\exp\left(\frac{1}{2}\frac{\beta}{1+\beta\|\boldsymbol{x}\|^2}g^2\right) = (|\boldsymbol{\alpha}_i|-1)!!\frac{(1+\beta\|\boldsymbol{x}\|^2)^{|\boldsymbol{\alpha}_i|+1/2}}{\beta^{|\boldsymbol{\alpha}_i|/2}}. \qquad (32)$$

Substituting into (30), we find many cancellations after which we are left with

$$\langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}}\rangle^2 = \frac{\prod_{i=1}^N (|\boldsymbol{\alpha}_i|-1)!!^2}{\prod_{i=1}^N \boldsymbol{\alpha}_i!}\beta^{|\boldsymbol{\alpha}|}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{X}_n} \boldsymbol{x}^{\sum_{i=1}^N \boldsymbol{\alpha}_i}\right)^2, \qquad (33)$$

the final result. ◀

A.3 Norm of the Low-Degree Likelihood Ratio

Proof of Lemma 5.9. Recall that

$$\|L^{\le D}_{n,\gamma,\beta,\mathcal{X}}\|^2_{L^2(\mathbb{Q}_n)} = \sum_{\substack{\boldsymbol{\alpha}\in(\mathbb{N}^n)^N\\ |\boldsymbol{\alpha}|\le D}} \langle L_{n,\gamma,\beta,\mathcal{X}}, \widehat{H}_{\boldsymbol{\alpha}}\rangle^2. \qquad (34)$$

We substitute in the result of Lemma 5.8, which, after introducing independent replicas $\boldsymbol{x}^1, \boldsymbol{x}^2 \sim \mathcal{X}_n$, may be rewritten as

$$\begin{aligned}
\|L^{\le D}_{n,\gamma,\beta,\mathcal{X}}\|^2_{L^2(\mathbb{Q}_n)} &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{\substack{\boldsymbol{\alpha}_i\in\mathbb{N}^n, i\in[N]\\ |\boldsymbol{\alpha}_i| \text{even}\\ \sum_{i=1}^N|\boldsymbol{\alpha}_i|\le D}} \prod_{i=1}^N \frac{(|\boldsymbol{\alpha}_i|-1)!!^2}{\boldsymbol{\alpha}_i!}\beta^{|\boldsymbol{\alpha}_i|}(\boldsymbol{x}^1)^{\boldsymbol{\alpha}_i}(\boldsymbol{x}^2)^{\boldsymbol{\alpha}_i}\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{d=0}^D \beta^d \sum_{\substack{d_1,\dots,d_N \text{ even}\\ \sum d_i = d}} \left(\prod_{i=1}^N \frac{(d_i-1)!!^2}{d_i!}\right) \sum_{\substack{\boldsymbol{\alpha}_i\in\mathbb{N}^n, i\in[N]\\ |\boldsymbol{\alpha}_i|=d_i}} \prod_{i=1}^N \binom{d_i}{\boldsymbol{\alpha}_i} \prod_{j=1}^n (x_j^1 x_j^2)^{\boldsymbol{\alpha}_i(j)}.
\end{aligned}$$

By the multinomial theorem,

$$\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{d_i} = \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}^n\\ |\boldsymbol{\alpha}|=d_i}} \binom{d_i}{\boldsymbol{\alpha}} \prod_{j=1}^n (x_j^1 x_j^2)^{\boldsymbol{\alpha}(j)},$$

and therefore

$$\begin{aligned}\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{\sum_{i=1}^N d_i} &= \prod_{i=1}^N \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{d_i} \\ &= \prod_{i=1}^N \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}^n \\ |\boldsymbol{\alpha}|=d_i}} \binom{d_i}{\boldsymbol{\alpha}} \prod_{j=1}^n (x_j^1 x_j^2)^{\boldsymbol{\alpha}(j)} \\ &= \sum_{\substack{\boldsymbol{\alpha}_i\in\mathbb{N}^n, i\in[N] \\ |\boldsymbol{\alpha}_i|=d_i}} \prod_{i=1}^N \binom{d_i}{\boldsymbol{\alpha}_i} \prod_{j=1}^n (x_j^1 x_j^2)^{\boldsymbol{\alpha}_i(j)}.\end{aligned}$$

In our case, this shows

$$\begin{aligned}\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{\substack{0\leq d\leq D \\ d \text{ even}}} \beta^d \sum_{\substack{d_1,\dots,d_N \text{ even} \\ \sum d_i = d}} \left(\prod_{i=1}^N \frac{(d_i-1)!!^2}{d_i!}\right)\left(\prod_{i=1}^N \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{d_i}\right) \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{\substack{0\leq d\leq D \\ d \text{ even}}} \beta^d \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^d \sum_{\substack{d_1,\dots,d_N \text{ even} \\ \sum d_i = d}} \prod_{i=1}^N \frac{d_i!}{d_i!!^2} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{\substack{0\leq d\leq D \\ d \text{ even}}} 2^{-d}\beta^d \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^d \sum_{\substack{d_1,\dots,d_N \text{ even} \\ \sum d_i = d}} \prod_{i=1}^N \binom{d_i}{d_i/2}\end{aligned}$$

where we have used the identities $n! = n!! \cdot (n-1)!!$ and $(2n)!! = 2^n \cdot n!$. We now pass to a notation making the restriction to even degrees clearer:

$$\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \sum_{0\leq d\leq \lfloor D/2\rfloor} \left(\frac{\beta^2\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2}{4}\right)^d \sum_{\substack{d_1,\dots,d_N \\ \sum d_i = d}} \prod_{i=1}^N \binom{2d_i}{d_i}.$$

The remaining function may be understood in terms of the generating function of the central binomial coefficients: using Proposition 5.7, we have that for any $x \in (-\frac{1}{4}, \frac{1}{4})$,

$$\varphi_N(x) := (1-4x)^{-N/2} = \left(\sum_{d\geq 0} \binom{2d}{d} x^d\right)^N = \sum_{d\geq 0} x^d \sum_{\substack{d_1,\dots,d_N \\ \sum d_i = d}} \prod_{i=1}^N \binom{2d_i}{d_i}.$$

Writing $\varphi_{N,k}(x)$ for the truncation of this Taylor series to degree k, we see that

$$\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n} \left[\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2}{4}\right)\right],$$

the final result. ◀

B Proofs for Bounding the Low-Degree Likelihood Ratio

B.1 Local Chernoff Bound

Proof of Proposition 5.12. It is sufficient to show that $\mathsf{iid}(\pi/\sqrt{n})$ admits a local Chernoff bound. Since π is subgaussian, $\pi^2 - \mathbb{E}[\pi^2]$ is subexponential (see, e.g., [56]), i.e., the moment-generating function $M(t) = \mathbb{E}[\exp(t(\pi^2 - \mathbb{E}[\pi^2]))]$ satisfies $M(t) \leq \exp(\frac{t^2}{2s^2})$ for all $|t| \leq s$ for

a suitable choice of a constant $s > 0$. In particular, $\mathbb{E}[\exp(t\pi^2)] < \infty$ for all $|t| \le s$ for this choice of $s > 0$.

Let $\Pi = \pi\pi'$, the product of two independent copies of π. Let σ^2 be the variance proxy of π (see Definition 2.1). The moment-generating function of Π is

$$M(t) = \mathbb{E}[\exp(t\Pi)] = \mathbb{E}_\pi \mathbb{E}_{\pi'}[\exp(t\pi\pi')] \le \mathbb{E}_\pi \left[\exp\left(\sigma^2 t^2 \pi^2/2\right)\right] < \infty$$

provided $\frac{1}{2}\sigma^2 t^2 < s$, i.e., $|t| < \sqrt{2s/\sigma^2}$. Thus $M(t)$ exists in an open interval containing $t = 0$, which implies $M'(0) = \mathbb{E}[\Pi] = 0$ and $M''(0) = \mathbb{E}[\Pi^2] = 1$ (this is the defining property of the moment-generating function: its derivatives at $t = 0$ are the moments of Π).

Let $\eta > 0$ and $f(t) = \exp\left(\frac{t^2}{2(1-\eta)}\right)$. Since $M(0) = 1, M'(0) = 0, M''(0) = 1$ and $f(0) = 1, f'(0) = 0, f''(0) = \frac{1}{1-\eta} > 1$, there exists $\delta > 0$ such that for all $t \in [-\delta, \delta]$, $M(t)$ exists and $M(t) \le f(t)$.

We now apply the standard Chernoff bound argument to $\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle = \frac{1}{n}\sum_{i=1}^n \Pi_i$, where $\Pi_1, \ldots, \Pi_n$ are i.i.d. copies of Π. For any $\lambda > 0$,

$$\begin{aligned}
\Pr\left\{\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle \ge t\right\} &= \Pr\left\{\exp(\lambda\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle) \ge \exp(\lambda t)\right\} \\
&\le \exp(-\lambda t)\mathbb{E}[\exp(\lambda\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle)] && \text{(by Markov's inequality)} \\
&= \exp(-\lambda t)\mathbb{E}[\exp(\lambda n^{-1}\sum_{i=1}^n \Pi_i)] \\
&= \exp(-\lambda t)[M(\lambda/n)]^n \\
&\le \exp(-\lambda t)[f(\lambda/n)]^n && \text{(provided } \lambda/n \le \delta\text{)} \\
&\le \exp(-\lambda t)\exp\left(\frac{\lambda^2}{2(1-\eta)n}\right).
\end{aligned}$$

Taking $\lambda = (1-\eta)nt$,

$$\Pr\left\{\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle \ge t\right\} \le \exp\left(-(1-\eta)nt^2 + \frac{1}{2}(1-\eta)nt^2\right) = \exp\left(-\frac{1}{2}(1-\eta)nt^2\right)$$

as desired. This holds provided $\lambda/n \le \delta$, i.e., $t \le \delta/(1-\eta)$. The same argument (with $-\Pi$ instead of Π) holds for the other tail bound $\Pr\left\{\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle \le -t\right\}$. ◀

B.2 Bounding the Large Deviations

Proof of Lemma 5.13. Recall that

$$\varphi_{N,\lfloor D/2 \rfloor}(x) = \sum_{d=0}^{\lfloor D/2 \rfloor} x^d \sum_{\substack{d_1,\ldots,d_N \\ \sum d_i = d}} \prod_{i=1}^N \binom{2d_i}{d_i}.$$

Note that the first sum above has $\lfloor D/2 \rfloor + 1$ terms and the second sum has at most $N^d \le N^{\lfloor D/2 \rfloor} \le N^{D/2}$ terms. It is combinatorially clear that $\binom{2d_i}{d_i}\binom{2d_j}{d_j} \le \binom{2(d_i+d_j)}{d_i+d_j}$, and therefore

$$\prod_{i=1}^N \binom{2d_i}{d_i} \le \binom{2\sum_{i=1}^N d_i}{\sum_{i=1}^N d_i} = \binom{2d}{d} \le (2d)^d \le D^{D/2}.$$

Since $\|\boldsymbol{x}^1\|^2 \le 2$ and $\|\boldsymbol{x}^2\|^2 \le 2$ we have $\frac{1}{4}\beta^2\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle^2 \le \beta^2$. Since $d \le D/2$ we have $(\beta^2)^d \le (1+\beta^2)^{D/2}$, and therefore

$$\varphi_{N,\lfloor D/2 \rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle^2\right) \le (D/2+1)(1+\beta^2)^{D/2} N^{D/2} D^{D/2} \le (1+\beta^2)^{D/2} D N^{D/2} D^{D/2}.$$

Combining these bounds,

$$\begin{aligned} R_2 &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|>\varepsilon}\,\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right] \\ &\leq \Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|>\varepsilon\right\}(1+\beta^2)^{D/2}DN^{D/2}D^{D/2}. \end{aligned}$$

Since R_2 increases as ε decreases, we can assume without loss of generality that ε is small enough that we may apply the local Chernoff bound (18):

$$\begin{aligned} &\leq \exp\left(-\frac{1}{3}n\varepsilon^2\right)(1+\beta^2)^{D/2}DN^{D/2}D^{D/2} \\ &= \exp\left(-\frac{1}{3}n\varepsilon^2+\frac{D}{2}\log(1+\beta^2)+\log D+\frac{D}{2}\log N+\frac{D}{2}\log D\right) \\ &= o(1) \end{aligned}$$

provided $D=o(n/\log n)$, completing the proof. ◀

B.3 Bounding the Small Deviations

Proof of Lemma 5.14. We use the argument from Appendix K of [54]. Since the Taylor series for $\varphi_N(x)$ has nonnegative coefficients, we have $\varphi_{N,\lfloor D/2\rfloor}(x)\leq\varphi_N(x)$ for all $x\in[0,1/4)$. Taking $\varepsilon<1/|\beta|$, we have

$$\begin{aligned} R_1 &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\varphi_{N,\lfloor D/2\rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right] \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\varphi_{N}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\left(1-\beta^2\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)^{-N/2}\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\exp\left(-\frac{N}{2}\log\left(1-\beta^2\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right)\right]. \end{aligned}$$

By the convexity of $t\mapsto-\log(1-\beta^2 t)$, we have $-\log(1-\beta^2 t)\leq-\frac{t}{\varepsilon^2}(1-\beta^2\varepsilon^2)$ for all $t\in[0,\varepsilon^2]$. Letting $c:=-\frac{N}{2\varepsilon^2}\log(1-\beta^2\varepsilon^2)>0$, we proceed bounding

$$\begin{aligned} &\leq \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2\sim\mathcal{X}_n}\left[\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\exp\left(c\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)\right] \\ &= \int_0^\infty \Pr\left\{\mathbb{1}_{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon}\,\exp\left(c\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)>u\right\}du \\ &= \int_0^\infty \Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon \quad\text{and}\quad \exp\left(c\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)>u\right\}du \\ &\leq 1+\int_1^\infty \Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon \quad\text{and}\quad \exp\left(c\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2\right)>u\right\}du. \end{aligned}$$

Applying the change of variables $u=\exp(ct)$,

$$\begin{aligned} &= 1+\int_0^\infty \Pr\left\{|\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle|\leq\varepsilon \quad\text{and}\quad \langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2>t\right\}c\exp(ct)\,dt \\ &\leq 1+\int_0^{\varepsilon^2} \Pr\left\{\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle^2>t\right\}c\exp(ct)\,dt. \end{aligned}$$

Provided ε is sufficiently small, we can apply the local Chernoff bound (18):

$$\leq 1 + Cc \int_0^{\varepsilon^2} \exp\left(-\frac{1}{2}(1-\eta)nt + ct\right) dt.$$

Let $\widehat{\gamma} := n/N$, so that $\widehat{\gamma} \to \gamma$ as $n \to \infty$. Letting $c =: \widehat{c}n$ where $\widehat{c} = -\log(1-\beta^2\varepsilon^2)/(2\varepsilon^2\widehat{\gamma})$,

$$\leq 1 + C \cdot \widehat{c}n \int_0^{\varepsilon^2} \exp\left[\left(-\frac{1}{2}(1-\eta) + \widehat{c}\right) nt\right] dt.$$

We have $\lim_{\varepsilon \to 0^+} \widehat{c} = \frac{\beta^2}{2\widehat{\gamma}}$. Since $\beta^2 < \gamma$, we have that for sufficiently large n, $\frac{\beta^2}{2\widehat{\gamma}} < \frac{1}{2}$. Thus we can choose ε and η small enough so that for sufficiently large n, $-\frac{1}{2}(1-\eta) + \widehat{c} \leq -\alpha$ for some $\alpha > 0$. Now

$$\begin{aligned}
&\leq 1 + C \cdot \widehat{c}n \int_0^{\infty} \exp\left(-\alpha nt\right) dt \\
&= 1 + \frac{C \cdot \widehat{c}}{\alpha} \\
&= O(1),
\end{aligned}$$

completing the proof. ◀

B.4 Above the BBP Threshold

Proof of Theorem 3.4 (Part 2). Let $\beta^2 > \gamma$. Recall

$$\begin{aligned}
\|L^{\leq D}_{n,\gamma,\beta,\mathcal{X}}\|^2_{L^2(\mathbb{Q}_n)} &= \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2 \sim \mathcal{X}_n} \left[\varphi_{N,\lfloor D/2 \rfloor}\left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2\right)\right] \\
&= \sum_{d=0}^{\lfloor D/2 \rfloor} \mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2 \sim \mathcal{X}_n} \left(\frac{\beta^2}{4}\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^2\right)^d \sum_{\substack{d_1,\ldots,d_N \\ \sum d_i = d}} \prod_{i=1}^{N} \binom{2d_i}{d_i}. \qquad (35)
\end{aligned}$$

Since each term in the outer summation of (35) is nonnegative, it sufficies to fix a single $d \leq D/2$ and show that the corresponding term is $\omega(1)$. We can write $\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle = \frac{1}{n}\sum_{i=1}^n \Pi_i$ where $\Pi_1, \ldots, \Pi_n$ are i.i.d. with distribution of the product $\Pi = \pi\pi'$ of two independent copies of π. This means

$$\mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2 \sim \mathcal{X}_n} \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{2d} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \Pi_i\right)^{2d} = n^{-2d} \sum_{i_1,\ldots,i_{2d} \in [n]} \mathbb{E}[\Pi_{i_1}\Pi_{i_2}\cdots\Pi_{i_{d2}}].$$

Since π is symmetric about zero, Π is also symmetric about zero, so all moments $\mathbb{E}[\Pi^k]$ are nonnegative. This means each term in the remaining sum is nonnegative, so we can obtain a lower bound by only considering terms where each index occurring among the $i_1, \ldots, i_{2d}$ occurs exactly twice:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}^1,\boldsymbol{x}^2 \sim \mathcal{X}_n} \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^{2d} \geq n^{-2d}\binom{n}{d}\frac{(2d)!}{2^d}\left(\mathbb{E}[\Pi^2]\right)^d = n^{-2d}\binom{n}{d}\frac{(2d)!}{2^d}.$$

Next, we bound the inner summation of (35) below by taking only the terms with $d_i \in \{0, 1\}$ for all $i \in [N]$:

$$\sum_{\substack{d_1,\ldots,d_N \\ \sum d_i = d}} \prod_{i=1}^{N} \binom{2d_i}{d_i} \geq \binom{N}{d} 2^d.$$

Combining these bounds, we find that for any fixed $0 \leq d \leq \lfloor D/2 \rfloor$,

$$\begin{aligned} \|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 &\geq \frac{\beta^{2d}}{4^d} n^{-2d} \binom{n}{d} \frac{(2d)!}{2^d} \binom{N}{d} 2^d \\ &= \left(\frac{\beta^2}{4n^2}\right)^d \frac{(2d)!\, n!\, N!}{(d!)^2\, (n-d)!\, (N-d)!} \\ &\geq \left(\frac{\beta^2 (n-d)(N-d)}{4n^2}\right)^d \binom{2d}{d}. \end{aligned}$$

Using the standard bound $\binom{2d}{d} \geq 4^d/(2\sqrt{d})$,

$$\|L_{n,\gamma,\beta,\mathcal{X}}^{\leq D}\|_{L^2(\mathbb{Q}_n)}^2 \geq \frac{1}{2\sqrt{d}} \left(\frac{\beta^2 (n-d)(N-d)}{n^2}\right)^d.$$

This final expression will be $\omega(1)$ provided that $1 \ll d \ll n$, since $n/N \to \gamma$ and $\beta^2 > \gamma$. ◀