Counterexamples to the Low-Degree Conjecture

Justin Holmgren
NTT Research, Palo Alto, CA, USA
justin.holmgren@ntt-research.com

Alexander S. Wein
Courant Institute of Mathematical Sciences, New York University, NY, USA
awein@cims.nyu.edu

─── **Abstract** ───────────

A conjecture of Hopkins (2018) posits that for certain high-dimensional hypothesis testing problems, no polynomial-time algorithm can outperform so-called "simple statistics", which are low-degree polynomials in the data. This conjecture formalizes the beliefs surrounding a line of recent work that seeks to understand statistical-versus-computational tradeoffs via the *low-degree likelihood ratio*. In this work, we refute the conjecture of Hopkins. However, our counterexample crucially exploits the specifics of the noise operator used in the conjecture, and we point out a simple way to modify the conjecture to rule out our counterexample. We also give an example illustrating that (even after the above modification), the symmetry assumption in the conjecture is necessary. These results do not undermine the low-degree framework for computational lower bounds, but rather aim to better understand what class of problems it is applicable to.

1 Introduction

A primary goal of computer science is to understand which problems can be solved by efficient algorithms. Given the formidable difficulty of proving unconditional computational hardness, state-of-the-art results typically rely on unproven conjectures. While many such results rely only upon the widely-believed conjecture $\mathsf{P} \neq \mathsf{NP}$, other results have only been proven under stronger assumptions such as the unique games conjecture [19, 20], the exponential time hypothesis [16], the learning with errors assumption [25], or the planted clique hypothesis [17, 4].

It has also been fruitful to conjecture that a specific *algorithm* (or limited class of algorithms) is optimal for a suitable class of problems. This viewpoint has been particularly prominent in the study of average-case noisy statistical inference problems, where it appears that optimal performance over a large class of problems can be achieved by methods such as the sum-of-squares hierarchy (see [24]), statistical query algorithms [18, 5], the approximate message passing framework [9, 22], and low-degree polynomials [15, 14, 13]. It is helpful to have such a conjectured-optimal meta-algorithm because this often admits a *systematic* analysis of hardness. However, the exact class of problems for which we believe these methods

12th Innovations in Theoretical Computer Science Conference (ITCS 2021).
Editor: James R. Lee; Article No. 75; pp. 75:1–75:9

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

are optimal has typically not been precisely formulated. In this work, we explore this issue for the class of low-degree polynomial algorithms, which admits a systematic analysis via the low-degree likelihood ratio.

The *low-degree likelihood ratio* [15, 14, 13] has recently emerged as a framework for studying computational hardness in high-dimensional statistical inference problems. It has been shown that for many "natural statistical problems," all known polynomial-time algorithms only succeed in the parameter regime where certain "simple" (low-degree) statistics succeed. The power of low-degree statistics can often be understood via a relatively simple explicit calculation, yielding a tractable way to precisely predict the statistical-versus-computational tradeoffs in a given problem. These "predictions" can rigorously imply lower bounds against a broad class of spectral methods [21, Theorem 4.4] and are intimately connected to the *sum-of-squares hierarchy* (see [14, 13, 24]). Recent work has (either explicitly or implicitly) carried out this type of low-degree analysis for a variety of statistical tasks [3, 15, 14, 13, 2, 1, 21, 8, 6, 23, 7]. For more on these methods, we refer the reader to the PhD thesis of Hopkins [13] or the survey article [21].

Underlying the above ideas is the belief that for certain "natural" problems, low-degree statistics are as powerful as all polynomial-time algorithms – we refer broadly to this belief as the "low-degree conjecture". However, formalizing the notion of "natural" problems is not a straightforward task. Perhaps the easiest way to illustrate the meaning of "natural" is by example: prototypical examples (studied in the previously mentioned works) include planted clique, sparse PCA, random constraint satisfaction problems, community detection in the stochastic block model, spiked matrix models, tensor PCA, and various problems of a similar flavor. All of these can be stated as simple hypothesis testing problems between a "null" distribution (consisting of random noise) and a "planted" distribution (which contains a "signal" hidden in noise). They are all high-dimensional problems but with sufficient symmetry that they can be specified by a small number of parameters (such as a "signal-to-noise ratio"). For all of the above problems, the best known polynomial-time algorithms succeed precisely in the parameter regime where simple statistics succeed, i.e., where there exists a $O(\log n)$-degree polynomial of the data whose value behaves noticeably different under the null and planted distributions (in a precise sense). Thus, barring the discovery of a drastically new algorithmic approach, the low-degree conjecture seems to hold for all the above problems. In fact, a more general version of the conjecture seems to hold for runtimes that are not necessarily polynomial: degree-D statistics are as powerful as all $n^{\tilde{\Theta}(D)}$-time algorithms, where $\tilde{\Theta}$ hides factors of $\log n$ [13, Hypothesis 2.1.5] (see also [21, 8]).

A precise version of the low-degree conjecture was formulated in the PhD thesis of Hopkins [13]. This includes precise conditions on the null distribution ν and planted distribution μ which capture most of the problems mentioned above. The key conditions are that there should be sufficient symmetry, and that μ should be injected with at least a small amount of noise. Most of the problems above satisfy this symmetry condition (a notable exception being the spiked Wishart model[1], which satisfies a mild generalization of it), but it remained unclear whether this assumption was needed in the conjecture. On the other hand, the noise assumption is certainly necessary, as illustrated by the example of solving a system of linear equations over a finite field: if the equations have an exact solution then it can be obtained via Gaussian elimination even though low-degree statistics suggest that the

---

[1] Here we mean the formulation of the spiked Wishart model used in [1], where we directly observe Gaussian samples instead of only their covariance matrix.

problem should be hard; however, if a small amount of noise is added (so that only a $1 - \varepsilon$ fraction of the equations can be satisfied) then Gaussian elimination is no longer helpful, and the low-degree conjecture seems to hold.

In this work we investigate more precisely what kinds of noise and symmetry conditions are needed in the conjecture of Hopkins [13]. Our first result (Theorem 4) actually refutes the conjecture in the case where the underlying random variables are real-valued. Our counterexample exploits the specifics of the noise operator used in the conjecture, along with the fact that a single real number can be used to encode a large (but polynomially bounded) amount of data. In other words, we show that a *stronger* noise assumption than the one in [13] is needed; Remark 6 explains a modification of the conjecture that we do not know how to refute. Our second result (Theorem 7) shows that the symmetry assumption in [13] cannot be dropped, i.e., we give a counterexample for a weaker conjecture that does not require symmetry. Both of our counterexamples are based on efficiently decodable error-correcting codes.

Notation

Asymptotic notation such as $o(1)$ and $\Omega(1)$ pertains to the limit $n \to \infty$. We say that an event occurs with *high probability* if it occurs with probability $1 - o(1)$, and we use the abbreviation w.h.p. ("with high probability"). We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. The *Hamming distance* between vectors $x, y \in F^n$ (for some field F) is $\Delta(x, y) = |\{i \in [n] : x_i \neq y_i\}|$ and the *Hamming weight* of x is $\Delta(x, 0)$.

2 The Low-Degree Conjecture

We now state the formal variant of the low-degree conjecture proposed in the PhD thesis of Hopkins [13, Conjecture 2.2.4]. The terminology used in the statement will be explained below.

▶ **Conjecture 1.** *Let $\mathcal{X}$ be a finite set or $\mathbb{R}$, and let $k \geq 1$ be a fixed integer. Let $N = \binom{n}{k}$. Let ν be a product distribution on $\mathcal{X}^N$. Let μ be another distribution on $\mathcal{X}^N$. Suppose that μ is S_n-invariant and $(\log n)^{1+\Omega(1)}$-wise almost independent with respect to ν. Then no polynomial-time computable test distinguishes $T_\delta \mu$ and ν with probability $1 - o(1)$, for any $\delta > 0$. Formally, for all $\delta > 0$ and every polynomial-time computable $t : \mathcal{X}^N \to \{0, 1\}$ there exists $\delta' > 0$ such that for every large enough n,*

$$\frac{1}{2} \mathop{\mathbb{P}}_{x \sim \nu} (t(x) = 0) + \frac{1}{2} \mathop{\mathbb{P}}_{x \sim T_\delta \mu} (t(x) = 1) \leq 1 - \delta'.$$

We now explain some of the terminology used in the conjecture, referring the reader to [13] for the full details. We will be concerned with the case $k = 1$, in which case S_n-invariance of μ means that for any $x \in \mathcal{X}^n$ and any $\pi \in S_n$ (the symmetric group) we have $\mathbb{P}_\mu(x) = \mathbb{P}_\mu(\pi \cdot x)$ where π acts by permuting coordinates. The notion of *D-wise almost independence* captures how well degree-D polynomials can distinguish μ and ν. For our purposes, we do not need the full definition of D-wise almost independence (see [13]), but only the fact that it is implied by *exact D-wise independence*, defined as follows.

▶ **Definition 2.** *A distribution μ on $\mathcal{X}^N$ is D-wise independent with respect to ν if for any $S \subseteq [N]$ with $|S| \leq D$ we have equality of the marginal distributions $\mu|_S = \nu|_S$.*

Finally, the *noise operator T_δ* is defined as follows.

▶ **Definition 3.** *Let ν be a product distribution on $\mathcal{X}^N$ and let μ be another distribution on $\mathcal{X}^N$. For $\delta \in [0, 1]$, let $T_\delta \mu$ be the distribution on $\mathcal{X}^N$ generated as follows. To sample $z \sim T_\delta \mu$, first sample $x \sim \mu$ and $y \sim \nu$ independently, and then, independently for each i, let*

$$z_i = \begin{cases} x_i & \text{with probability } 1 - \delta, \\ y_i & \text{with probability } \delta. \end{cases}$$

(Note that T_δ depends on ν but the notation suppresses this dependence; ν will always be clear from context.)

3   Main Results

We first give a counterexample that refutes Conjecture 1 in the case $\mathcal{X} = \mathbb{R}$.

▶ **Theorem 4.** *The following holds for infinitely many n. Let $\mathcal{X} = \mathbb{R}$, and $\nu = \text{Unif}([0, 1]^n)$. There exists a distribution μ on $\mathcal{X}^n$ such that μ is S_n-invariant (with $k = 1$) and $\Omega(n)$-wise independent with respect to ν, and for some constant $\delta > 0$ there exists a polynomial-time computable test distinguishing $T_\delta \mu$ and ν with probability $1 - o(1)$.*

The proof is given in Section 5.1.

▶ Remark 5. We assume a standard model of finite-precision arithmetic over $\mathbb{R}$, i.e., the algorithm t can access polynomially-many bits in the binary expansion of its input.

Note that we refute an even weaker statement than Conjecture 1 because our counterexample has *exact* $\Omega(n)$-wise independence instead of only $(\log n)^{1+\Omega(1)}$-wise *almost* independence.

▶ Remark 6. Essentially, our counterexample exploits the fact that a single real number can be used to encode a large block of data, and that the noise operator T_δ will leave many of these blocks untouched (effectively allowing us to use a super-constant alphabet size). We therefore propose modifying Conjecture 1 in the case $\mathcal{X} = \mathbb{R}$ by using a different noise operator that applies a small amount of noise to *every* coordinate instead of resampling a small number of coordinates. If ν is i.i.d. $\mathcal{N}(0, 1)$ then the standard Ornstein-Uhlenbeck noise operator is a natural choice (and in fact, this is mentioned by [13]). Formally, this is the noise operator T_δ that samples $T_\delta \mu$ as follows: draw $x \sim \mu$ and $y \sim \nu$ and output $\sqrt{1 - \delta}x + \sqrt{\delta}y$.

Our second result illustrates that in the case where $\mathcal{X}$ is a finite set, the S_n-invariance assumption cannot be dropped from Conjecture 1. (In stating the original conjecture, Hopkins [13] remarked that he was not aware of any counterexample when the S_n-invariance assumption is dropped).

▶ **Theorem 7.** *The following holds for infinitely many n. Let $\mathcal{X} = \{0, 1\}$ and $\nu = \text{Unif}(\{0, 1\}^n)$. There exists a distribution μ on $\mathcal{X}^n$ such that μ is $\Omega(n)$-wise independent with respect to ν, and for some constant $\delta > 0$ there exists a polynomial-time computable test distinguishing $T_\delta \mu$ and ν with probability $1 - o(1)$.*

The proof is given in Section 5.2.

Both of our counterexamples are in fact still valid in the presence of a stronger noise operator T_δ that *adversarially* changes any δ-fraction of the coordinates.

The rest of the paper is organized as follows. Our counterexamples are based on error-correcting codes, so in Section 4 we review the basic notions from coding theory that we will need. In Section 5 we construct our counterexamples and prove our main results.

4   Coding Theory Preliminaries

Let $F = \mathbb{F}_q$ be a finite field. A *linear code C* (over F) is a linear subspace of F^n. Here n is called the *(block) length*, and the elements of C are called *codewords*. The *distance* of C is the minimum Hamming distance between two codewords, or equivalently, the minimum Hamming weight of a nonzero codeword.

▶ **Definition 8.** *Let C be a linear code. The* dual distance *of C is the minimum Hamming weight of a vector in F^n that is orthogonal to all codewords. Equivalently, the dual distance is the distance of the* dual code $C^\perp = \{x \in F^n : \langle x, c \rangle = 0 \;\; \forall c \in C\}$.

The following standard fact will be essential to our arguments.

▶ **Proposition 9.** *If C is a linear code with dual distance d, then the uniform distribution over codewords is $(d-1)$-wise independent with respect to the uniform distribution on F^n.*

Proof. This is a standard fact in coding theory, but we give a proof here for completeness. Fix $S \subseteq [n]$ with $|S| \leq d - 1$. For some k, we can write $C = \{x^\top G : x \in F^k\}$ for some $k \times n$ *generator matrix G* whose rows form a basis for C. Let G_S be the $k \times |S|$ matrix obtained from G by keeping only the columns in S. It is sufficient to show that if x is drawn uniformly from F^k then $x^\top G_S$ is uniform over $F^{|S|}$. The columns of G_S must be linearly independent, because otherwise there is a vector $y \in F^n$ of Hamming weight $\leq d - 1$ such that $Gy = 0$, implying $y \in C^\perp$, which contradicts the dual distance. Thus there exists a set $T \subseteq [k]$ of $|S|$ linearly independent rows of G_S (which form a basis for $F^{|S|}$). For any fixed choice of $x_{[k] \setminus T}$ (i.e., the coordinates of x outside T), if the coordinates x_T are chosen uniformly at random then $x^\top G_S$ is uniform over $F^{|S|}$. This completes the proof.   ◀

▶ **Definition 10.** *A code C admits efficient decoding from r errors and s erasures if there is a deterministic polynomial-time algorithm $\mathcal{D} : (F \cup \{\perp\})^n \to F^n \cup \{\mathsf{fail}\}$ with the following properties.*
- *For any codeword $c \in C$, let $c' \in F^n$ be any vector obtained from c by changing the values of at most r coordinates (to arbitrary elements of F), and replacing at most s coordinates with the erasure symbol $\perp$. Then $\mathcal{D}(c') = c$.*
- *For any arbitrary $c' \in F^n$ (not obtained from some codeword as above), $\mathcal{D}(c')$ can output any codeword or* fail *(but must never output a vector that is not a codeword[2]).*

Note that the decoding algorithm knows where the erasures have occurred but does not know where the errors have occurred.

Our first counterexample (Theorem 4) is based on the classical *Reed-Solomon codes* $\mathsf{RS}_q(n,k)$, which consist of univariate polynomials of degree at most k evaluated at n canonical elements of the field $\mathbb{F}_q$, and are known to have the following properties.

▶ **Proposition 11** (Reed-Solomon Codes). *For any integers $0 \leq k < n$ and for any prime power $q \geq n$, there is a length-n linear code C over $\mathbb{F}_q$ with the following properties:*
- *the dual distance is $k + 2$,*
- *C admits efficient decoding from r errors and s erasures whenever $2r + s < n - k$.*

Proof. See e.g., [11] for the construction and basic facts regarding Reed-Solomon codes $\mathsf{RS}_q(n,k)$. The distance of $\mathsf{RS}_q(n,k)$ is $n - k$. It is well known that the dual code of $\mathsf{RS}_q(n,k)$ is $\mathsf{RS}_q(n, n - k - 2)$, which proves the claim about dual distance. Efficient decoding is discussed e.g., in Section 3.2 of [11].   ◀

---

[2] The codes we deal with in this paper can be efficiently constructed, and so it is easy to test whether a given vector is a codeword. Thus, this assumption is without loss of generality.

Our second counterexample (Theorem 7) is based on the following construction of efficiently correctable binary codes with large dual distance. A proof (by Guruswami) can be found in [26, Theorem 4]. Similar results were proved earlier [27, 10, 12].

▶ **Proposition 12.** *There exists a universal constant $\zeta \geq 1/30$ such that for every integer $i \geq 1$ there is a linear code C over $\mathbb{F}_2 = \{0, 1\}$ of block length $n = 42 \cdot 8^{i+1}$, with the following properties:*

- *the dual distance is at least ζn, and*
- *C admits efficient decoding from $\zeta n/2$ errors (with no erasures).*

5   Proofs

Before proving the main results, we state some prerequisite notation and lemmas.

▶ **Definition 13.** *Let F be a finite field. For $S \subseteq [n]$ (representing erased positions), the S-restricted Hamming distance $\Delta_S(x, y)$ is the number of coordinates in $[n] \setminus S$ where x and y differ:*

$$\Delta_S(x, y) = |\{i \in [n] \setminus S \ : \ x_i \neq y_i\}|.$$

We allow x, y to belong to either F^n or $F^{[n]\setminus S}$, or even to $(F \cup \{\bot\})^n$ so long as the "erasures" $\bot$ occur only in S.

The following lemma shows that a random string is sufficiently far from any codeword.

▶ **Lemma 14.** *Let C be a length-n linear code over a finite field F. Suppose C admits efficient decoding from $2r$ errors and s erasures, for some r, s satisfying $r \leq (n-s)/(8e)$. For any fixed choice of at most s erasure positions $S \subseteq [n]$, if x is a uniformly random element of $F^{[n]\setminus S}$ then*

$$\mathbb{P}_x(\exists c \in C \ : \ \Delta_S(c, x) \leq r) \leq (r+1)2^{-r}.$$

Proof. Let $B_r(c) = \{x \in F^{[n]\setminus S} \ : \ \Delta_S(c, x) \leq r\} \subseteq F^{[n]\setminus S}$ denote the Hamming ball (with erasures S) of radius r and center c, and let $|B_r|$ denote its cardinality (which does not depend on c). We have the following basic bounds on $|B_r|$:

$$\binom{n - |S|}{r}(|F| - 1)^r \leq |B_r| \leq (r+1)\binom{n - |S|}{r}(|F| - 1)^r.$$

Since decoding from $2r$ errors and s erasures is possible, the Hamming balls $\{B_{2r}(c)\}_{c \in C}$ are disjoint, and so

$$\mathbb{P}_x(\exists c \in C \ : \ \Delta_S(c, x) \leq r) \leq |B_r|/|B_{2r}|$$

$$\leq (r+1)\binom{n - |S|}{r}\binom{n - |S|}{2r}^{-1}(|F| - 1)^{-r}$$

$$\leq (r+1)\binom{n - |S|}{r}\binom{n - |S|}{2r}^{-1}.$$

Using the standard bounds $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$, this becomes

$$\leq (r+1)\left(\frac{4er}{n - |S|}\right)^r$$

$$\leq (r+1)\left(\frac{4er}{n - s}\right)^r$$

which is at most $(r+1)2^{-r}$ provided $r \leq (n-s)/(8e)$.   ◀

▶ **Lemma 15.** *Let $j_1, \ldots, j_n$ be uniformly and independently chosen from $[n]$. For any constant $\alpha < 1/e$, the number of indices $i \in [n]$ that occur exactly once among $j_1, \ldots, j_n$ is at least αn with high probability.*

Proof. Let $X_i \in \{0, 1\}$ be the indicator that i occurs exactly once among $j_1, \ldots, j_n$, and let $X = \sum_{i=1}^n X_i$. We have (as $n \to \infty$)

$$\mathbb{E}[X_i] = n(1/n)(1 - 1/n)^{n-1} \to e^{-1},$$

$\mathbb{E}[X_i^2] = \mathbb{E}[X_i]$, and for $i \neq i'$,

$$\mathbb{E}[X_i X_{i'}] = n(n-1)(1/n)^2(1 - 2/n)^{n-2} \to e^{-2}.$$

This means $\mathbb{E}[X] = (1 + o(1))n/e$ and

$$
\begin{aligned}
\mathrm{Var}(X) &= \sum_i (\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2) + \sum_{i \neq i'} (\mathbb{E}[X_i X_{i'}] - \mathbb{E}[X_i]\mathbb{E}[X_{i'}]) \\
&\leq n(1 + o(1))(e^{-1} - e^{-2}) + n(n-1) \cdot o(1) \\
&= o(n^2).
\end{aligned}
$$

The result now follows by Chebyshev's inequality.   ◀

5.1 Proof of Theorem 4

The idea of the proof is as follows. First imagine the setting where μ is not required to be S_n-invariant. By using each real number to encode an element of $F = \mathbb{F}_q$, we can take ν to be the uniform distribution on F^n and take μ to be a random Reed-Solomon codeword in F^n. Under $T_\delta \mu$, the noise operator will corrupt a few symbols ("errors"), but the decoding algorithm can correct these and thus distinguish $T_\delta \mu$ from ν.

In order to have S_n-invariance, we need to modify the construction. Instead of observing an ordered list $y = (y_1, \ldots, y_n)$ of symbols, we will observe pairs of the form (i, y_i) (with each pair encoded by a single real number) where i is a random index. If the same i value appears in two different pairs, this gives conflicting information; we deal with this by simply throwing it out and treating y_i as an "erasure" that the code needs to correct. If some i value does not appear in any pairs, we also treat this as an erasure. The full details are given below.

Proof of Theorem 4. Let C be the length-n code from Proposition 11 with $k = \lceil \alpha n \rceil$ for some constant $\alpha \in (0, 1)$ to be chosen later. Fix a scheme by which a real number encodes a tuple (j, y) with $j \in [n]$ and $y \in F = \mathbb{F}_q$, in such a way that a uniformly random real number in $[0, 1]$ encodes a uniformly random tuple (j, y). More concretely, we can take $n = q = 2^m$ for some integer $m \geq 1$, in which case (j, y) can be directly encoded using the first $2m$ bits of the binary expansion of a real number. Under $x \sim \nu$, each coordinate x_i encodes an independent uniformly random tuple (j_i, y_i). Under μ, let each coordinate x_i be a uniformly random encoding of (j_i, y_i), drawn as follows. Let $\tilde{c}$ be a uniformly random codeword from C. Draw $j_1, \ldots, j_n \in [n]$ independently and uniformly. For each i, if j_i is a unique index (in the sense that $j_i \neq j_{i'}$ for all $i' \neq i$) then set $y_i = \tilde{c}_{j_i}$; otherwise choose y_i uniformly from F.

Note that μ is S_n-invariant. Since the dual distance of C is $k + 2 = \Omega(n)$, it follows (using Proposition 9) that μ is $\Omega(n)$-wise independent with respect to ν. By choosing $\delta > 0$ and $\alpha > 0$ sufficiently small, we can ensure that $16\delta n + (2n/3 + 4\delta n) < n - k$ and so C admits efficient decoding from $8\delta n$ errors and $2n/3 + 4\delta n$ erasures (see Proposition 11). The algorithm to distinguish $T_\delta \mu$ and ν is as follows. Given a list of (j_i, y_i) pairs, produce $c' \in F^n$

by setting $c'_{j_i} = y_i$ wherever j_i is a unique index (in the above sense), and setting all other positions of c' to $\perp$ (an "erasure"). Let $S \subseteq [n]$ be the indices i for which $c'_i = \perp$. Run the decoding algorithm on c'; if it succeeds and outputs a codeword c such that $\Delta_S(c, c') \leq 4\delta n$ then output "$T_\delta \mu$", and otherwise output "ν".

We can prove correctness as follows. If the true distribution is ν then Lemma 15 guarantees $|S| \leq 2n/3$ (w.h.p.). Since the non-erased values in c' are uniformly random, Lemma 14 ensures there is no codeword c with $\Delta_S(c, c') \leq 4\delta n$ (w.h.p.), provided we choose $\delta \leq 1/(96e)$, and so our algorithm outputs "ν" (w.h.p).

Now suppose the true distribution is $T_\delta \mu$. In addition to the $\leq 2n/3$ erasures caused by non-unique j_i's sampled from μ, each coordinate resampled by T_δ can create up to 2 additional erasures and can also create up to 2 errors (i.e., coordinates i for which $c'_i \neq \perp$ but $c'_i \neq \tilde{c}_i$). Since at most $2\delta n$ coordinates get resampled (w.h.p.), this means we have a total of (up to) $4\delta n$ errors and $2n/3 + 4\delta n$ erasures. This means decoding succeeds and outputs $\tilde{c}$ (i.e., the true codeword used to sample μ), and furthermore, $\Delta_S(\tilde{c}, c') \leq 4\delta n$.   ◄

5.2   Proof of Theorem 7

The idea of the proof is similar to the previous proof, and somewhat simpler (since we do not need S_n-invariance). We take ν to be the uniform distribution on binary strings and take μ to be the uniform distribution on codewords, using the binary code from Proposition 12. The decoding algorithm is able to correct the errors caused by T_δ.

Proof of Theorem 7. Let C be the code from Proposition 12. Each codeword $c \in C$ is an element of $\mathbb{F}_2^n$, which can be identified with $\{0, 1\}^n = \mathcal{X}^n$. Let μ be the uniform distribution over codewords. Since the dual distance of C is at least ζn, we have from Proposition 9 that μ is $\Omega(n)$-wise independent with respect to the uniform distribution ν. We also know that C admits efficient decoding from $\zeta n/2$ errors.

Let $\delta = \min\{1/(16e), \zeta/8\}$. The algorithm to distinguish $T_\delta \mu$ and ν, given a sample c', is to run the decoding algorithm on c'; if decoding succeeds and outputs a codeword c such that $\Delta(c, c') \leq 2\delta n$ then output "$T_\delta \mu$", and otherwise output "ν".

We can prove correctness as follows. If c' is drawn from $T_\delta \mu$ then c' is separated from some codeword c by at most $2\delta n \leq \zeta n/4$ errors (w.h.p.), and so decoding will find c. If instead c' is drawn from ν then (since $\delta \leq 1/(16e)$) by Lemma 14, there is no codeword within Hamming distance $2\delta n$ of c' (w.h.p.).   ◄

──── **References** ────

1   Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained PCA problems. *arXiv preprint*, 2019. `arXiv:1902.07324`.

2   Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (Nearly) efficient algorithms for the graph matching problem on correlated random graphs. In *Advances in Neural Information Processing Systems*, pages 9186–9194, 2019.

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

4   Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *arXiv preprint*, 2013. `arXiv:1304.0828`.

5   Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using fourier analysis. In Frank Thomson Leighton and Michael T. Goodrich, editors, *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 253–262. ACM, 1994. `doi:10.1145/195058.195147`.

6 Matthew Brennan and Guy Bresler. Average-case lower bounds for learning sparse mixtures, robust estimation and semirandom adversaries. *arXiv preprint*, 2019. `arXiv:1908.06130`.

7 Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. *arXiv preprint*, 2019. `arXiv:1912.11071`.

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

9 David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

10 G-L Feng and Thammavarapu RN Rao. Decoding algebraic-geometric codes up to the designed minimum distance. *IEEE Transactions on Information Theory*, 39(1):37–45, 1993.

11 Venkatesan Guruswami and Madhu Sudan. Improved decoding of reed-solomon and algebraic-geometric codes. In *Proceedings 39th Annual Symposium on Foundations of Computer Science*, pages 28–37. IEEE, 1998.

12 Venkatesan Guruswami and Madhu Sudan. On representations of algebraic-geometry codes. *IEEE Transactions on Information Theory*, 47(4):1610–1613, 2001.

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

14 Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

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

16 Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-SAT. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001. `doi:10.1006/jcss.2000.1727`.

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

18 Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In S. Rao Kosaraju, David S. Johnson, and Alok Aggarwal, editors, *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 392–401. ACM, 1993. `doi:10.1145/167088.167200`.

19 Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 767–775, 2002.

20 Subhash Khot. On the unique games conjecture. In *FOCS*, volume 5, page 3, 2005.

21 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`.

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

23 Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: Degree-2 to degree-4. *arXiv preprint*, 2019. `arXiv:1911.01411`.

24 Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint*, 6, 2018. `arXiv:1807.11419`.

25 Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56(6):34:1–34:40, 2009. `doi:10.1145/1568318.1568324`.

26 Amir Shpilka. Constructions of low-degree and error-correcting ε-biased generators. *computational complexity*, 18(4):495, 2009.

27 Alexei N Skorobogatov and Serge G Vladut. On the decoding of algebraic-geometric codes. *IEEE Transactions on Information Theory*, 36(5):1051–1060, 1990.