Lattice-Based Methods Surpass Sum-of-Squares in Clustering

Ilias Zadik[a], Min Jae Song[b], Alexander S. Wein[c], and Joan Bruna[b,d,e]

[a]Department of Mathematics, Massachusetts Institute of Technology
[b]Courant Institute of Mathematical Sciences, New York University
[c]Simons Institute for the Theory of Computing, UC Berkeley
[d]Center for Data Science, New York University
[e]Center for Computational Mathematics, Flatiron Institute

January 10, 2022

Abstract

Clustering is a fundamental primitive in unsupervised learning which gives rise to a rich class of computationally-challenging inference tasks. In this work, we focus on the canonical task of clustering d-dimensional Gaussian mixtures with unknown (and possibly degenerate) covariance. Recent works (Ghosh et al. '20; Mao, Wein '21; Davis, Diaz, Wang '21) have established lower bounds against the class of low-degree polynomial methods and the sum-of-squares (SoS) hierarchy for recovering certain hidden structures planted in Gaussian clustering instances. Prior work on many similar inference tasks portends that such lower bounds strongly suggest the presence of an inherent statistical-to-computational gap for clustering, that is, a parameter regime where the clustering task is *statistically* possible but no *polynomial-time* algorithm succeeds.

One special case of the clustering task we consider is equivalent to the problem of finding a planted hypercube vector in an otherwise random subspace. We show that, perhaps surprisingly, this particular clustering model *does not exhibit* a statistical-to-computational gap, despite the aforementioned low-degree and SoS lower bounds. To achieve this, we give an algorithm based on Lenstra–Lenstra–Lovász lattice basis reduction which achieves the statistically-optimal sample complexity of $d + 1$ samples. This result extends the class of problems whose conjectured statistical-to-computational gaps can be "closed" by "brittle" polynomial-time algorithms, highlighting the crucial but subtle role of noise in the onset of statistical-to-computational gaps.

Contents

1 Introduction **3**
- 1.1 Main contributions . . . 4
- 1.2 Relation to prior work . . . 6
- 1.3 "Noiseless" problems and implications for SoS/low-degree lower bounds . . . 8

2 Preliminaries **10**

3 The LLL-based algorithm **12**
- 3.1 The algorithm and the main guarantee . . . 13

4 Implications of the success of LLL **15**
- 4.1 The planted sparse vector problem . . . 15
- 4.2 Continuous Learning with Errors . . . 15
- 4.3 Gaussian Clustering . . . 16

5 Information-theoretic lower bounds **17**
- 5.1 Optimality of $d+1$ samples for parameter recovery . . . 17
- 5.2 Information-theoretic lower bound for label recovery . . . 18

6 Proof of Algorithm 1 correctness **19**
- 6.1 Towards proving Theorem 3.3: auxiliary lemmas . . . 19
- 6.2 Proof of Theorem 3.3 . . . 23
- 6.3 Proof of Lemma 3.4 . . . 25

7 Proofs of information-theoretic lower bounds **27**
- 7.1 Proof of Theorem 5.1 . . . 28
- 7.2 Proof of Theorem 5.2 . . . 30
- 7.3 Proof of Lemma 7.3 . . . 35

1 Introduction

Many high-dimensional statistical inference problems exhibit a gap between what can be achieved by the optimal statistical procedure and what can be achieved by the best known *polynomial-time* algorithms. As a canonical example, finding a planted k-clique in a $G(n, 1/2)$ Erdős–Rényi random graph is statistically possible when k exceeds $2\log_2 n$ (via exhaustive search) but all known polynomial time algorithms require $k = \Omega(\sqrt{n})$, giving rise to a large conjectured "possible but hard" regime in between [Jer92, AKS98, BHK+19]. Such so-called *statistical-to-computational gaps* are prevalent in many other key learning problems including sparse PCA (principal component analysis) [BR13], community detection [DKMZ11], tensor PCA [MR14], and random constraint satisfaction problems [KMOW17], just to name a few. Unfortunately, since these are *average-case* problems where the input is drawn from a specific distribution, current techniques appear unable to prove computational hardness in these conjectured "hard" regimes based on standard complexity assumptions such as $\mathsf{P} \neq \mathsf{NP}$.

Still, a number of different methods have emerged for understanding these gaps and providing "rigorous evidence" for computational hardness of statistical problems. Many involve studying the power of restricted classes of algorithms that are tractable to analyze, including statistical query (SQ) algorithms [Kea98, FGR+17], the sum-of-squares (SoS) hierarchy [Par00, Las01], low-degree polynomial algorithms [HS17, HKP+17, Hop18], approximate message passing [DMM09], MCMC methods [Jer92], and various notions of "local" algorithms [GS17, GZ17, AGJ20]. As it turns out, the best known poly-time algorithms for a surprisingly wide range of statistical problems actually do belong to these restricted classes. As such, the above frameworks have been very successful at providing concrete explanations for statistical-to-computational gaps and allowing researchers to predict the location of the "hard" regime in new problems based on the location that such restricted classes of algorithms fail or succeed.

However, there are notorious exceptions where the above predictions turn out to be false. For example, the problem of learning parity (even in the absence of noise) is hard for SQ, SoS, and low-degree polynomials [Kea98, Gri01, Sch08], yet actually admits a simple poly-time solution via Gaussian elimination. Yet, to the best of our knowledge, *prior to the present work*, learning parities with no noise or other similar noiseless inference models based on linear equations, such as random 3-XOR-SAT, have been the only examples where some polynomial-time method (which appears to always be Gaussian elimination) works, while the SoS hierarchy and low-degree methods have been proven to fail.

In this work, we identify a new *class of problems* where the SoS hierarchy and low-degree lower bounds are provably bypassed by a polynomial-time algorithm. This class of problems is *not* based on linear equations, and the suggested optimal algorithm is *not* based on Gaussian elimination but on lattice basis reduction methods, which specifically seek to find a "short" vector in a lattice. Similar lattice-based method have over the recent years been able to "close" various statistical-to-computational gaps [ZG18, AHSS17, SZB21], yet this is the first example we are aware of that they are able to "close a gap" where the SoS hierarchy is known to fail to do so.

The problems we analyze can be motivated from several angles in theoretical computer science and machine learning, and can be thought of as important special cases of well-studied problems such as Planted Vector in a Subspace, Gaussian Clustering, and Non-Gaussian Component Analysis. While our result is more general, one specific problem that we solve is the following: for a hidden unit vector $u \in \mathbb{R}^d$, we observe n independent samples of the form

$$z_i \sim \mathcal{N}(x_i u, I_d - uu^\top), \quad i = 1, 2, \ldots, n, \tag{1}$$

where x_i are i.i.d. uniform ± 1, and the goal is to recover the hidden signs x_i and the hidden

direction u (up to a global sign flip). Prior to our work, the best known poly-time algorithm required $n \gg d^2$ samples[1] [MW21]. Furthermore, this was believed to be unimprovable due to lower bounds against SoS algorithms and low-degree polynomials [MS16, KB21, MRX20, GJJ+20, Kun20, MW21, DDW21]. Nevertheless, we give a poly-time algorithm under the much weaker assumption $n \geq d + 1$. In fact, this sample complexity is essentially optimal for the previous recovery problem, as shown by our information-theoretic lower bound (see Section 5). Our result makes use of the Lenstra–Lenstra–Lovász (LLL) algorithm for lattice basis reduction [LLL82], a powerful algorithmic paradigm that has seen recent, arguably surprising, success in solving to information-theoretic optimality a few different "noiseless" statistical inference problems, some even in regimes where it was conjectured that no polynomial-time method works: Discrete Regression [ZG18, GKZ21], Phase Retrieval [AHSS17, SZB21], Cosine Neuron Learning [SZB21], and Continuous Learning with Errors [BRST21, SZB21][2]. Yet, to the best of our knowledge, this work is the first to establish the success of an LLL-based method in a regime where low-degree and SoS lower bounds both suggest computational intractability. This raises the question of whether LLL can "close" any other conjectured statistical-to-computational gaps. We believe that understanding the power and limitations of the LLL approach is an important direction for future research towards understanding the computational complexity of inference.

We also point out one weakness of the LLL approach: our algorithm is brittle to the specifics of the model, and relies on the observations being "noiseless" in some sense. For instance, our algorithm only solves the model in (1) because the x_i values lie *exactly* in ± 1 and the covariance $\Sigma = I - uu^\top$ has quadratic form $u^\top \Sigma u$ *exactly* equal to zero (or, similarly to other LLL applications [ZG18], of exponentially small magnitude). If we were to perturb the model slightly, say by adding an inverse-polynomial amount of noise to the x_i's, our algorithm would break down because of the known non-robustness properties of the LLL algorithm. In fact, a noisy version (with inverse-polynomial noise) of one problem that we solve is the *homogeneous* Continuous Learning with Errors problem (hCLWE), which is provably hard based on the standard assumption [MR09, Conjecture 1.2] from lattice-based cryptography that certain worst-case lattice problems are hard against quantum algorithms [BRST21]. All existing algorithms for statistical problems based on LLL suffer from the same lack of robustness. In this sense, there is a strong analogy between LLL and the other known successful polynomial-time method for noiseless inference, namely the Gaussian elimination approach to learning parity: both exploit very precise algebraic structure in the problem and break down in the presence of even a small (inverse-polynomial) amount of noise.

As discussed above, our results "break" certain computational lower bounds based on SoS and low-degree polynomials. Still, we believe that these types of lower bounds are interesting and meaningful, but some care should be taken when interpreting them. It is in fact already well-established that such lower bounds can sometimes be beaten on "noiseless" problems (a key example being Gaussian elimination). However, there are some subtleties in how "noiseless" should be defined here, and whether fundamental problems with statistical-to-computational gaps such as planted clique—which has implications for many other inference problems via average-case reductions (e.g. [BR13, MW15, HWX15, BBH18])—should be considered "noiseless." We discuss these issues further in Section 1.3.

1.1 Main contributions

The main inference setting we consider in this work is as follows. Let $d \in \mathbb{N}$ be the growing ambient dimension the data lives in, n be the number of samples, and $a > 0$ be what we coin as the *spacing*

[1] Here are throughout, the notation $\gg$ hides logarithmic factors.
[2] in the exponentially-small noise regime

Problems	Low-degree LB	SoS LB	Previous Best	Our Results
Planted Vector (Rademacher)	$\tilde{\Omega}(d^2)$	$\tilde{\Omega}(d^{3/2})$	$\tilde{O}(d^2)$ [MW21, DDW21]	$d+1$
Gaussian Clustering (SNR $= \infty$)	$\tilde{\Omega}(d^2)$	$\tilde{\Omega}(d^{3/2})$	$\tilde{O}(d^2)$ [DDW21]	$d+1$
hCLWE (Noiseless)	-	-	$O(d^2)$ [BRST21]	$d+1$

Table 1: Sample complexity upper and lower bounds for *polynomial-time* exact recovery for Planted Hypercube Vector Recovery, Gaussian Clustering, and hCLWE. LB stands for "Lower Bound".

parameter. We consider arbitrary labels on the one-dimensional lattice: ax_i where $x_i \in \mathbb{Z}$ for $i = 1, \ldots, n$, under the weak constraints $|x_i| \le 2^d$, $i = 1, \ldots, n$ and $d^{-O(1)} \le |a| \le d^{O(1)}$. We also consider an arbitrary direction $u \in \mathcal{S}^{d-1}$ and an unknown covariance matrix Σ with $\Sigma u = 0$ and "reasonable" spectrum, in the sense that Σ does not have exponentially small or large eigenvalues in directions orthogonal to u (see Assumption 3.1). In particular, the choice of $\Sigma = I - uu^\top$ is permissible per our assumptions but much more generality is possible. Our goal is to learn *both* the labels x_i, $i = 1, \ldots, n$ and the hidden direction u (up to global sign flip applied to both x and u) from independent samples

$$z_i \sim \mathcal{N}(ax_i u, \Sigma), \quad i = 1, \ldots, n. \tag{2}$$

Exact recovery with polynomial-time algorithm. Our main algorithmic result is informally stated as follows.

Theorem 1.1 (Informal statement of Theorem 3.3)**.** *Under the above setting, if $n = d+1$ then there is an LLL-based algorithm (Algorithm 1) which terminates in polynomial time and outputs exactly, up to a global sign flip, both the correct labels x_i and the correct hidden direction u with probability* $1 - \exp(-\Omega(d))$.

Now, as explained in Section 1 our theorem has algorithmic implications for three previously studied problems: Planted Vector in a Subspace, Gaussian Clustering, and hCLWE, which is an instance of Non-Gaussian Component Analysis. In all three settings, the previous best algorithms required $\Omega(d^2)$ samples (formally this is true for the dense case of the planted vector in a subspace setting, but $\omega(d)$ samples are required for many sparse settings as well). As explained, in many of these cases lower bounds have been achieved for the classes of low-degree methods and the SoS hierarchy. In this work, we show that LLL can surpass these lower bounds and succeed with $n = d+1$ samples in all three problems. We provide more context in Section 1.2 and exact statements in Section 4. A high-level description of our contributions can be found in Table 1.

Information-theoretic lower bound for exact recovery of the hidden direction. One can naturally wonder whether for our setting there is something even better than LLL that can be achieved with bounded computational resources or even unbounded ones. We complement the previous result with an information-theoretic lower bound (Theorem 1.2) showing that *no* estimation procedure can succeed at exact recovery of the hidden direction u using at most $n = d - 1$ samples. This means LLL is information-theoretically optimal for recovering the hidden direction up to at most one additional sample.

Theorem 1.2 (Informal statement of Theorem 5.1)**.** *Under the setting of* (1)*, if* $n \le d-1$*, there is no estimation procedure that can guarantee with probability greater than* $1/2$ *exact recovery of the hidden direction* u *up to a global sign flip.*

In fact, our formal theorem (Theorem 5.1) shows that recovering the labels $\{x_i\}_{i=1}^n$ is strictly *easier* than recovering the hidden direction u in the sense that $d-1$ samples are insufficient for determining u even when we have exact knowledge of the labels $\{x_i\}_{i=1}^n$. In light of this fact, it is natural to ask about the sample complexity of recovering the labels. We provide an answer to this question by showing that no estimator can recover the labels $\{x_i\}_{i=1}^n$ (up to a global sign flip) with probability $1-o(1)$ when $n = \lfloor \rho d \rfloor$ for any constant $\rho \in (0,1)$ (Theorem 1.3). This implies that our sample complexity of $n = d+1$ is optimal up to a $1+o(1)$ factor for label recovery.

Theorem 1.3 (Informal statement of Theorem 5.2)**.** *Let* $\rho \in (0,1)$ *be a fixed constant. Under the setting of* (1)*, if* $n \le \lfloor \rho d \rfloor$*, there is no estimation procedure that can guarantee with probability* $1-o(1)$ *exact recovery of the labels* $\{x_i\}_{i=1}^n$ *up to a global sign flip.*

1.2 Relation to prior work

Non-gaussian component analysis. Non-gaussian component analysis (NGCA) is the problem of identifying a non-gaussian direction in random data. Concretely, this is a generalization of (1) where the x_i are drawn i.i.d. from some distribution μ on $\mathbb{R}$. When μ is the Rademacher ± 1 distribution, we recover the problem in (1) as a special case.

The NGCA problem was first introduced in [BKS+06], inspiring a long line of algorithmic results [KT06, SKBM08, DJSS10, DJNS13, Bea14, STS16, NOTV17, VX11, TV18, GS19]. This problem has also played a key role in many hardness results in the statistical query (SQ) model: starting from the work of [DKS17], various special cases of NGCA have provided SQ-hard instances for a variety of learning tasks [DKS17, DKS18, DKS19, DKP+21, BLPR19, GGK20, DKZ20, DKPZ21, GGJ+20, DKKZ20, DK20, NWR19]. More recently, a special case of NGCA has also been shown to be hard under the widely-believed assumption that certain worst-case lattice problems are hard [BRST21]. NGCA is a special case of the more general *spiked transport model* [NWR19].

Our main result (Theorem 1.1) solves NGCA with only $n = d+1$ samples in the case where μ is supported arbitrarily on an exponentially large subset of a 1-dimensional discrete lattice. While this case is essentially the noiseless, equispaced version of the "parallel Gaussian pancakes" problem, which was first introduced and shown to be SQ-hard by [DKS17], our result does not bypass known SQ lower bounds [DKS17, BRST21] as the hard construction involves Gaussian pancakes with non-negligible "thickness".

A *concurrent and independent work* by Diakonikolas and Kane [DK21] proposed a very similar LLL-based polynomial-time algorithm to ours for the case where μ is "nearly" supported on a finite subset of a finitely generated additive subgroup of $\mathbb{R}$, which includes (1) as a special case. Interestingly, while their algorithm provably works with $n = d+1$ samples in the noiseless case, it also tolerates a small exponential-in-d level of noise in the labels x_i at the expense of using $n = 2d$ samples. The exact noise tolerance of our proposed algorithm is left as an interesting open question.

Planted vector in a subspace. A line of prior work has studied the problem of finding a "structured" vector planted in an otherwise random d-dimensional subspace of $\mathbb{R}^n$, where $d < n$. A variety of algorithms have been proposed and analyzed in the case where the planted vector is *sparse* [DH14, BKS14, QSW16, HSSS16, QZL+20]. One canonical Gaussian generative model for this problem turns out to be equivalent to NGCA (with the same parameters d, n), where the entrywise distribution of the planted vector corresponds to the non-gaussian distribution μ in NGCA.

More specifically, the subspace in question is the *column span* of the matrix whose *rows* are the NGCA samples z_i; see e.g. Lemma 4.21 of [MW21] for the formal equivalence.

Motivated by connections to the *Sherrington–Kirkpatrick model* from spin glass theory, the setting of a planted *hypercube* (i.e. ± 1-valued) vector in a subspace has received recent attention; this is equivalent to the problem in (1). Specifically, sum-of-squares (SoS) lower bounds have been given for refuting the existence of a hypercube vector in (or close to) a purely random subspace. First, it was shown that when $n = O(d)$, SoS relaxations of degree 2 [MS16], 4 [KB21, MRX20], and 6 [Kun20] fail. A later improvement [GJJ+20] shows failure of degree-$n^{\Omega(1)}$ SoS when $n \ll d^{3/2}$ and conjectures that this condition can be improved to $n \ll d^2$ (see Conjectures 8.1 and 8.2 of [GJJ+20]).

The state-of-the-art algorithmic result for recovering a planted vector in a subspace is [MW21], which builds on [HSSS16] and in particular analyzes a spectral method proposed by [HSSS16]. For a planted ρ-sparse Rademacher vector (ρn entries are nonzero, and these nonzero entries are $\pm 1/\sqrt{\rho}$), this spectral method succeeds at recovering the vector provided $n \gg \rho^2 d^2$ [MW21]. On the other hand, if $n \ll \rho^2 d^2$ then all low-degree polynomial algorithms fail, implying in particular that *all* spectral methods (in a large class) fail [MW21]. These results cover the special case of a planted hypercube vector ($\rho = 1$), in which case there is a spectral method that succeeds when $n \gg d^2$, and failure of low-degree and spectral methods when $n \ll d^2$.

The above results suggest inherent computational hardness of planted sparse Rademacher vector when $n \ll \rho^2 d^2$. However, perhaps surprisingly, our main result (Theorem 1.1) implies that this problem can actually be solved via LLL in polynomial time whenever $n \geq d + 1$. Thus, LLL beats all low-degree algorithms whenever $\rho \gg 1/\sqrt{d}$. However, our algorithm requires the entries of the planted vector to *exactly* lie in $\{0, \pm 1/\sqrt{\rho}\}$, whereas the spectral method of [HSSS16, MW21] succeeds under more general conditions.

We remark that the planted hypercube vector problem is closely related to the *negatively-spiked Wishart model* with a hypercube spike, which can be thought of as a model for generating the *orthogonal complement* of the subspace. The work of [BKW20] gives low-degree lower bounds for this negative Wishart problem and conjectures hardness when $n = O(d)$ (Conjecture 3.1). However, our results do not refute this conjecture because the conjecture is for a slightly noisy version of the problem (since the SNR parameter β is taken to be strictly greater than -1).

Clustering. Our model (1) is an instance of a broader clustering problem (2) under Gaussian mixtures. In the binary case, it consists of n i.i.d. samples $\{(z_i, x_i)\}_{i=1\dots n} \in \mathbb{R}^d \times \{-1, +1\}$ of the Gaussian mixture $P(x_i = -1) = P(x_i = +1) = 1/2$, and $z_i \,|\, x_i \sim \mathcal{N}(x_i u, \Sigma)$, with unknown mean u and covariance Σ. The goal of clustering is to infer the mixture variables $\{x_i\}$ from the observations $\{z_i\}$. Clustering algorithms have been analysed extensively, both from the statistical and computational perspective.

The statistical performance is driven by the signal-to-noise ratio $\mathrm{SNR} = v^\top \Sigma^{-1} v$, in the sense that the error rate for recovering the mixture labels is $\exp(-\Omega(\mathrm{SNR}))$ [Fri89]. Exact recovery of the vector of n labels is thus possible only when $\mathrm{SNR} \gtrsim \log(n)$.

Recently, [DDW21] showed that the MLE estimator for the missing labels corresponds to a Max-Cut problem, which recovers the solution when $n = \tilde{\Omega}(d)$. Moreover, the authors argued that while SNR drives the inherent statistical difficulty of the estimation problem, a relaxed quantity $\mathrm{S} = \|v\|^2/\|\Sigma\|$ presumably controls the computational difficulty. In particular, the largest such gap is attained in the covariance choice of (1), for which $\mathrm{SNR} = \infty$ while $\mathrm{S} = 1$. In this regime, they identified a gap between the statistical and computational performance of multiple existing algorithms, raising the crucial question whether such guarantees can be obtained using polynomial-time algorithms. Several previous works [BV08, MV10, BDJ+20, CMZ19, FPB17] introduce algorithms

that either require larger sample complexity $n = \tilde{\Omega}(d^2)$, or have non-optimal error rates, for instance based on k-means relaxations [Roy17, MVW17, GV19, LLL+20]. Leveraging existing SoS lower bounds on the associated Max-Cut problem (see Section 1.3), [DDW21] suggest a statistical-to-computational gap for exact recovery in the binary Gaussian mixture. Our main result (Theorem 1.1) implies that this problem can be solved via the LLL basis reduction method in polynomial time whenever $n \geq d + 1$. Thus in the present work, we refute this conjecture for SNR $= \infty$ under a weak "niceness" assumption on the covariance matrix Σ.

LLL-based statistical recovery. Our algorithm is based on the breakthrough use of LLL for solving average-case subset sum problems in polynomial-time, specifically the works of [LO85] and [Fri86]. In these works, it is established that while the (promise) Subset-Sum problem is NP-hard, for some integer-valued distributions on the input weights it becomes polynomial-time solvable by applying the LLL basis reduction algorithm on a carefully designed lattice. Building on these ideas, [ZG18, GKZ21] proposed a new algorithm for noiseless discrete regression and discrete phase retrieval which provably solves these problems using only one sample, surpassing previous local-search lower bounds based on the so-called Overlap Gap Property [GZ17]. Again using the Subset-Sum ideas the LLL approach has also "closed" the gap for noiseless phase retrieval [AHSS17, SZB21] which was conjectured to be hard because of the failure of approximate message passing in this regime [MLKZ20]. Furthermore, for the problem of noiseless Continuous LWE (CLWE), the LLL algorithm has been shown to succeed with $n = \Omega(d^2)$ samples in [BRST21], and later in [SZB21] with the information-theoretically optimal $n = d + 1$ samples.

Our work adds a perhaps important new conceptual angle to the power of LLL for noiseless inference. A common feature of all the above inference models where LLL has been successfully applied is that they fall into the class of generalized linear models (GLMs). A GLM is generally defined as follows: for some hidden direction $w \in \mathcal{S}^{d-1}$ and "activation" function $\phi : \mathbb{R} \to \mathbb{R}$ one observes n i.i.d. samples of the form $y_i = \phi(\langle X_i, w\rangle) + \xi_i$, $i = 1, \ldots, n$ where $X_i \in \mathbb{R}^d$ and $\xi_i \in \mathbb{R}$ are i.i.d. random variables. Our work shows how to successfully apply LLL and achieve statistically optimal performance for the clustering setting (2), which importantly does not admit a GLM formulation. We consider this a potentially interesting conceptual contribution of the present work, since many "hard" inference settings, such as the planted clique model, also do not belong in the class of GLMs.

1.3 "Noiseless" problems and implications for SoS/low-degree lower bounds

SoS and low-degree lower bounds. The sum-of-squares (SoS) hierarchy [Par00, Las01, KMOW17, BHK+19] (see [BS16, RSS18, FKP19] for a survey) and low-degree polynomials [HS17, HKP+17, Hop18] (see [KWB19] for a survey) are two restricted classes of algorithms that are often studied in the context of statistical-to-computational gaps. These are not the only two such frameworks, but we will focus on these two because our result "breaks" lower bounds in these two frameworks. SoS is a powerful hierarchy of semidefinite programming relaxations. Low-degree polynomial algorithms are simply multivariate polynomials in the entries of the input, of degree logarithmic in the input dimension; notably, these can capture all *spectral methods* (subject to some technical conditions), i.e., methods based on the leading eigenvalue/eigenvector of some matrix constructed from the input (see Theorem 4.4 of [KWB19]). Both SoS and low-degree polynomials have been widely successful at obtaining the best known algorithms for a wide variety of high-dimensional "planted" problems, where the goal is to recover a planted signal buried in noisy data. While there is no formal connection between SoS and low-degree algorithms, they are believed to be roughly equivalent in power [HKP+17]. It is often informally conjectured that SoS and/or low-degree methods

are as powerful as the best poly-time algorithms for "natural" high-dimensional planted problems (nebulously defined). As a result, lower bounds against SoS and/or low-degree methods are often considered strong evidence for inherent computational hardness of statistical problems.

Issue of noise-robustness. In light of the above, it is tempting to conjecture optimality of SoS and/or low-degree methods among all poly-time methods for a wide variety of statistical problems. While this conjecture seems to hold up for a surprisingly long and growing list of problems, there are, of course, limits to the class of problems for which this holds. As discussed previously, a well-known counterexample is the problem of learning parity (or the closely-related XOR-SAT problem), where Gaussian elimination succeeds in a regime where SoS and low-degree algorithms provably fail. This counterexample is often tossed aside by the following argument: "Gaussian elimination is a brittle algebraic algorithm that breaks down if a small amount of noise is added to the labels, whereas SoS/low-degree methods are more robust to noise and are therefore capturing the limits of poly-time *robust* inference, which is a more natural notion anyway. If we restrict ourselves to problems that are sufficiently *noisy* then SoS/low-degree methods should be optimal." However, we note that in our setting, SoS/low-degree methods are strictly suboptimal for a problem that *does* have plenty of Gaussian noise; the issue is that the signal and noise have a particular joint structure that preserves certain exact algebraic relationships in the data. This raises an important question: what exactly makes a problem "noisy" or "noiseless", and under what kinds of noise should we believe that SoS/low-degree methods are unbeatable? In the following, we describe one possible answer.

The low-degree conjecture. The "low-degree conjecture" of Hopkins [Hop18, Conjecture 2.2.4] formalized one class of statistical problems for which low-degree polynomials are believed to be optimal among poly-time algorithms. These are certain *hypothesis testing* problems where the goal is to decide whether the input was drawn from a null (i.i.d. noise) distribution or a planted distribution (containing a planted signal). In our setting, one should imagine testing between n samples drawn from the model (1) and n samples drawn i.i.d. from $\mathcal{N}(0, I_d)$. Computational hardness of hypothesis testing generally implies hardness of the associated recovery/estimation/learning problem (which in our case is to recover x and u) as in Theorem 3.1 of [MW21]. The class of testing problems considered in Hopkins' conjecture has two main features: first, the problem should be highly symmetric, which is typical for high-dimensional statistical problems (although Hopkins' precise notion of symmetry does not quite hold for the problems we consider in this paper). Second, and most relevant to our discussion, the problem should be *noise-tolerant*. More precisely, Hopkins' conjecture states that if low-degree polynomials fail to distinguish a null distribution $\mathbb{Q}$ from a planted distribution $\mathbb{P}$, then no poly-time algorithm can distinguish $\mathbb{Q}$ from *a noisy version of* $\mathbb{P}$. For our setting, the appropriate "noise operator" to apply to $\mathbb{P}$ (which was refined in [HW20]) is to replace each sample z_i by

$$\sqrt{1-\delta^2}z_i + \delta z_i'$$

where $z_i' \sim \mathcal{N}(0, I_d)$ independently from z_i, for an arbitrarily small constant $\delta > 0$. This has the effect of replacing x_i with $\sqrt{1-\delta^2}x_i + \delta\tilde{z}_i$ where $\tilde{z}_i \sim \mathcal{N}(0,1)$. This noise is designed to "defeat" brittle algorithms such as Gaussian elimination, and indeed our LLL-based algorithm is also expected to be defeated by this type of noise.

To summarize, the problem we consider in this paper is *not* noise-tolerant in the sense of Hopkins' conjecture because the Gaussian noise depends on the signal (specifically, there is no noise in the direction of u) whereas Hopkins posits that the noise should be *oblivious* to the signal. Thus, in hindsight we should perhaps not be too surprised that LLL was able to beat SoS/low-degree for this problem. In other words, our result does not falsify the low-degree conjecture or the sentiment

behind it (low-degree algorithms are optimal for noisy problems), with the caveat that one must be careful about the precise meaning of "noisy." We feel that this lesson carries an often-overlooked conceptual message that may have consequences for other fundamental statistical problems such as planted clique, which we discuss next.

Planted clique. The *planted clique conjecture* posits that there is no polynomial-time algorithm for distinguishing between a random $G(n, 1/2)$ graph and a $G(n, 1/2)$ graph with a clique planted on k random vertices (by adding all edges between the clique vertices), when $k \ll \sqrt{n}$. The planted clique conjecture is central to the study of statistical-to-computational gaps because it has been used as a primitive to deduce computational hardness of many other problems via a web of average-case reductions (e.g. [BR13, MW15, HWX15, BBH18]). A refutation of the planted clique conjecture would be a major breakthrough that could cast doubts on whether other statistical-to-computational gaps are "real" or whether the gap can be closed by a better algorithm. As a result, it is important to ask ourselves why we believe the planted clique conjecture. Aside from the fact that it has resisted all algorithmic attempts so far, the primary concrete evidence for the conjecture comes in the form of lower bounds against SoS and low-degree polynomials [BHK$^+$19, Hop18]. However, it is perhaps unclear whether these should really be thought of as strong evidence for inherent hardness because (like the problem we study in the paper) planted clique is *not* noise-tolerant in the sense of Hopkins' conjecture (discussed above). Specifically, the natural noise operator would be to independently resample a small constant fraction of the edges, which would destroy the clique structure. In other words, the conjecture of Hopkins only implies that a noisy variant of planted clique (namely *planted dense subgraph*) is hard when $k \ll \sqrt{n}$.

While we do not have any concrete reason to believe that LLL could be used to solve planted clique, we emphasize that planted clique is in some sense a "noiseless" problem and so we do not seem to have a principled reason to conjecture its hardness based on SoS and low-degree lower bounds. On the other hand, we should perhaps be somewhat more confident in the "planted dense subgraph conjecture" because planted dense subgraph is a truly noisy problem in the sense of [Hop18, Conjecture 2.2.4].

2 Preliminaries

The key component of our algorithmic results is the LLL lattice basis reduction algorithm. The LLL algorithm receives as input d linearly independent vectors $v_1, \ldots, v_d \in \mathbb{Z}^d$ and outputs an integer linear combination of them with "small" ℓ_2 norm. Specifically, let us define the lattice generated by d *integer* vectors as simply the set of integer linear combination of these vectors.

Definition 2.1 (Lattice)**.** *Given linearly independent $v_1, \ldots, v_d \in \mathbb{Z}^d$, let*

$$L = L(v_1, \ldots, v_d) = \left\{ \sum_{i=1}^{d} \lambda_i v_i : \lambda_i \in \mathbb{Z}, i = 1, \ldots, d \right\} , \tag{3}$$

which we refer to as the lattice generated by integer-valued $v_1, \ldots, v_d$. We also refer to $(v_1, \ldots, v_d)$ as an (ordered) basis for the lattice L.

The LLL algorithm solves a search problem called the approximate Shortest Vector Problem (SVP) on a lattice L, given a basis of it.

Definition 2.2 (Shortest Vector Problem)**.** *An instance of the algorithmic α-approximate SVP for a lattice $L \subseteq \mathbb{Z}^d$ is as follows. Given a lattice basis $v_1, \ldots, v_d \in \mathbb{Z}^d$ for the lattice L, find a vector $\widehat{x} \in L$, such that*

$$\|\widehat{x}\|_2 \leq \alpha \cdot \mu(L) \ .$$

where $\mu(L) = \min_{x \in L, x \neq 0} \|x\|_2$.

The following theorem holds for the performance of the LLL algorithm, whose details can be found in [LLL82] or [Lov86].

Theorem 2.3 ([LLL82])**.** *There is an algorithm (namely the LLL lattice basis reduction algorithm), which receives as input a basis for a lattice L given by $v_1, \ldots, v_d \in \mathbb{Z}^d$ which*

(1) returns a vector $v \in L$ satisfying $\|v\|_2 \leq 2^{d/2}\mu(L)$,

(2) terminates in time polynomial in d and $\log\left(\max_{i=1}^{d} \|v_i\|_\infty\right)$.

In this work, we use the LLL algorithm for an integer relation detection application, a problem which we formally define below.

Definition 2.4 (Integer relation detection)**.** *An instance of the* integer relation detection problem *is as follows. Given a vector $b = (b_1, \ldots, b_k) \in \mathbb{R}^k$, find an $m \in \mathbb{Z}^k \setminus \{\mathbf{0}\}$, such that $\langle b, m \rangle := \sum_{j=1}^{k} b_j m_j = 0$. In this case, m is said to be an integer relation for the vector b.*

To define our class of problems, we make use of the following two standard objects.

Definition 2.5 (Bernoulli–Rademacher vector)**.** *We say that a random vector $v \in \mathbb{R}^n$ is a Bernoulli–Rademacher vector with parameter $\rho \in (0, 1]$ and write $v \sim \mathsf{BR}(n, \rho)$, if the entries of v are i.i.d. with*

$$v_i = \begin{cases} 0 & \text{with probability } 1 - \rho, \\ 1/\sqrt{n\rho} & \text{with probability } \rho/2, \\ -1/\sqrt{n\rho} & \text{with probability } \rho/2. \end{cases}$$

Definition 2.6 (Discrete Gaussian on $s\mathbb{Z}$)**.** *Let $r, s > 0$ be real numbers. We define the discrete Gaussian distribution with* width *r supported on the scaled integer lattice $s\mathbb{Z}$ to be the distribution whose probability mass function at each $x \in s\mathbb{Z}$ is proportional to $\exp\left(-\frac{x^2}{2r^2}\right)$.*

The following tail bound on the discrete Gaussian will be useful in Section 4, in which we reduce the hCLWE model (also known as "Gaussian pancakes") (Model 4.3) to our general model (Model 3.2) which is the central problem for our LLL-based algorithm.

Claim 2.7 (Adapted from [SZB21, Claim I.6])**.** *Let $\gamma \geq 1$ be a real number, and let ν be the discrete Gaussian of width 1 supported on $(1/\gamma)\mathbb{Z}$ such that the probability mass function at $x \in (1/\gamma)\mathbb{Z}$ is given by*

$$\nu(x) = \frac{1}{\mathcal{Z}} \exp(-x^2/2) \ ,$$

where $\mathcal{Z} = \sum_{x \in (1/\gamma)\mathbb{Z}} \exp(-x^2/2)$ is the normalization constant. Then, the following bound holds.

$$\mathcal{Z} \geq \gamma\sqrt{2\pi}\left(1 - 2(1 + 1/(4\pi\gamma)^2)\exp(-2\pi^2\gamma^2)\right) > 1 \ . \tag{4}$$

In particular, for $t \geq \gamma$,

$$\mathbb{P}_{x \sim \nu}[|x| \geq t] \leq 4\exp(-t^2/2) \ .$$

Proof. The first inequality in (4) follows from [SZB21, Claim 1.6]. The fact that $\mathcal{Z} > 1$ follows from our assumption that $\gamma \geq 1$. Since $\mathcal{Z} > 1$, we have

$$\mathbb{P}_{x\sim\nu}[|x| \geq t] = \frac{1}{\mathcal{Z}} \sum_{x\in(1/\gamma)\mathbb{Z},|x|\geq t} \exp(-x^2/2) \leq 2 \sum_{x\in(1/\gamma)\mathbb{Z},x\geq t} \exp(-x^2/2) \ .$$

Now notice that the terms of the series are decaying in a geometric fashion for $x \geq \gamma$ since

$$\exp\left(-\frac{(x+1/\gamma)^2}{2}\right) / \exp\left(-\frac{x^2}{2}\right) = \exp\left(-\frac{2x+1/\gamma}{2\gamma}\right) \leq \exp(-1) < 1/2 \ .$$

It follows that

$$\mathbb{P}_{x\sim\nu}[|x| \geq t] \leq 2 \sum_{x\in(1/\gamma)\mathbb{Z},x\geq t} \exp(-x^2/2) \leq 4\exp(-t^2/2) \ .$$

□

3 The LLL-based algorithm

We now present the main contribution of this work, which is an LLL-based polynomial-time algorithm that provably solves the general problem defined in Model 3.2 with access to only $n = d + 1$ samples.

We deal formally with samples coming from d-dimensional Gaussians, which have as their mean some unknown multiple of an unknown unit vector $u \in \mathcal{S}^{d-1}$, and also some unknown covariance Σ which nullifies u and satisfies the following weak "separability" condition.

Assumption 3.1 (Weak separability of the spectrum)**.** *Fix a unit vector $u \in \mathcal{S}^{d-1}$. We say that a positive semi-definite $\Sigma \in \mathbb{R}^{d\times d}$ is u-weakly separable if for some constant $C > 0$ it holds that*

(a) $\Sigma u = 0$ and,

(b) All other eigenvalues of Σ lie in the interval $[d^{-C}, d^C]$.

Notice that in particular the canonical case $\Sigma = I - uu^\top$ is u-weakly separable as all eigenvalues of Σ are equal to one, besides the zero eigenvalue which has multiplicity one and eigenvector u.

Under the weak separability assumption we establish the following generic result.

Model 3.2 (Our general model)**.** *Let $d, n \in \mathbb{N}$, known spacing level $a > 0$ satisfying $d^{-c} \leq a \leq d^c$ for some constant $c > 0$, and arbitrary $x_i \in \mathbb{Z} \cap [-2^d, 2^d]$, $i = 1, \ldots, n$. Consider also an arbitrary $u \in \mathcal{S}^{d-1}$ and an arbitrary unknown $\Sigma \in \mathbb{R}^{d\times d}$ which is u-weakly separable per Assumption 3.1. Conditional on u, Σ and $\{x_i\}_{i=1,\ldots,n}$, we then draw independent samples $z_1, \ldots, z_n \in \mathbb{R}^d$ where $z_i \sim \mathcal{N}((ax_i)u, \Sigma)$. The goal is to use z_i, $i = 1, \ldots, n$ to recover both $\{x_i\}_{i=1,\ldots,n}$ and u up to a global sign flip, with probability $1 - \exp(-\Omega(d))$ over the samples z_i, $i = 1, \ldots, n$.*

In the following section we discuss the guarantee of the our proposed Algorithm. After, we discuss how our results implies that learning in polynomial time with $d + 1$ samples is possible for: (1) the "planted sparse vector" problem (Model 4.1), (2) the homogeneous Continuous Learning with Errors (hCLWE) problem, also informally called the "Gaussian pancakes" problem (Model 4.3), and finally (3) Gaussian clustering (Model 4.6). As mentioned earlier in Section 1, our result bypasses known lower bounds for SoS and low-degree polynomials which suggested that the regime $n = \Theta(d)$ would not be achievable by polynomial-time algorithms.

In what follows and the rest of the paper, for some $N \in \mathbb{N}$ and $x \in \mathbb{R}$ we denote by $(x)_N := 2^{-N}\lfloor 2^N x \rfloor$ the truncation of x to its first N bits after zero.

Algorithm 1: LLL-based algorithm for recovering $u, (x_i)_{i=1,\dots,d+1}$

Input: $n = d+1$ samples $z_i \in \mathbb{R}^d$, $i = 1, \dots, d+1$, spacing $a > 0$.
Output: Estimated labels $\hat{x}_i \in \mathbb{Z}$, $i = 1, \dots, d+1$ and unit vector $\hat{u} \in S^{d-1}$.

Construct a $d \times d$ matrix Z with columns $z_2, \dots, z_{d+1}$, and let $N = \lceil d^4 (\log d)^2 \rceil$.
if $\det(Z) = 0$ **then**
  return $\hat{u} = 0$ and output FAIL.
Compute $\lambda_1 = 1$ and $\lambda_i = \lambda_i(z_1, \dots, z_{d+1})$ given by $(\lambda_2, \dots, \lambda_{d+1})^\top = -Z^{-1} z_1$.
Set $M = 2^{2d}$ and $\tilde{v} = \left((\lambda_2)_N, \dots, (\lambda_{d+1})_N, 2^{-N}\right) \in \mathbb{R}^{d+1}$.
Output $(t_1, t_2) \in \mathbb{Z}^{d+1} \times \mathbb{Z}$ from running the LLL basis reduction algorithm on the lattice generated by the columns of the following $(d+2) \times (d+2)$ integer-valued matrix B,

$$B = \left(\begin{array}{c|c} M 2^N (\lambda_1)_N & M 2^N \tilde{v} \\ \hline 0_{(d+1)\times 1} & I_{(d+1)\times(d+1)} \end{array} \right).$$

$\hat{u}_0 \leftarrow \text{SolveLinearEquation}(u', Z^\top u' = a t_1)$.
if $\hat{u}_0 = 0$ **then**
  return $\hat{u} = 0$ and output FAIL.
Set $\hat{x}_i = (t_1)_i / \|\hat{u}_0\|_2, i = 1, \dots, d+1$.
return $\hat{x}_i$, $i = 1, \dots, d+1$ and $\hat{u}_0 / \|\hat{u}_0\|_2$ and output SUCCESS.

3.1 The algorithm and the main guarantee

Our proposed algorithm for solving Model 3.2 is described in Algorithm 1. Specifically we assume the algorithm receives $n = d+1$ independent samples according to Model 3.2. As we see in the following theorem, the algorithm is able to recover exactly (up to a global sign flip) both the hidden direction u and the hidden labels x_i, $i = 1, \dots, n$ in polynomial time.

Theorem 3.3. *Algorithm 1, given as input independent samples $(z_i)_{i=1,\dots,d+1}$ from Model 3.2 with hidden direction u, covariance Σ, and true labels $\{x_i\}_{i=1,\dots,d+1}$ satisfies the following with probability $1 - \exp(-\Omega(d))$: there exists $\epsilon \in \{-1, 1\}$ such that the algorithm's outputs $\{\hat{x}_i\}_{i=1,\dots,d+1}$ and $\hat{u} \in S^{d-1}$ satisfy*

$$\hat{x}_i = \epsilon x_i \quad \text{for } i = 1, \dots, d+1$$
$$\text{and } \hat{u} = \epsilon u \,.$$

Moreover, Algorithm 1 terminates in $\mathsf{poly}(d)$ *steps.*

We now provide intuition behind the algorithm's success. Note that for the unknown u and x_i it holds that

$$\langle z_i, u \rangle = a x_i \quad \text{for all } i = 1, \dots, d+1; . \tag{5}$$

In the first step, the algorithm checks a certain general-position condition on the received samples, which naturally is satisfied almost surely for our random data. In the following crucial three steps, the algorithm attempts to recover only the hidden integer labels x_i without learning u. To do this, it exploits a certain random integer linear relation that the labels x_i's satisfy which *importantly does not involve any information about the unknown u*, besides its existence. The key observation leading to this relation is the following. Since we have $d+1$ vectors z_i in a d-dimensional space,

there exist scalars $\lambda_1, \ldots, \lambda_{d+1}$ (depending on the z_i's) such that $\sum_{i=1}^{d+1} \lambda_i z_i = 0$. These are exactly the λ_i's that the algorithm computes in the second step. Using them, observe that the following linear equation holds, due to (5),

$$\sum_{i=1}^{d+1} \lambda_i a x_i = \sum_{i=1}^{d+1} \lambda_i \langle z_i, u \rangle = \left\langle \sum_{i=1}^{d+1} \lambda_i z_i, u \right\rangle = \langle 0, u \rangle = 0 \,, \tag{6}$$

and therefore since $a > 0$ it gives the following integer linear equation

$$\sum_{i=1}^{d+1} \lambda_i x_i = 0. \tag{7}$$

Again note that the λ_i's can be computed from the given samples z_i, so this is an equation whose sole unknowns are the labels x_i. With this integer linear equation in mind, the algorithm in the following step employs the powerful LLL algorithm applied to an appropriate lattice. This application of the LLL is based on the breakthrough works of [Lag84, Fri86] for solving random subset-sum problems in polynomial-time, as well as its recent manifestations for solving various other noiseless inference settings such as binary regression [ZG18] and phase retrieval [AHSS17, SZB21]. To get some intuition for this connection, notice that in the case $x_i \in \{-1, 1\}$, (7) is really a (promise) subset-sum relation with weights λ_i and unknown subset $\{i : x_i = 1\}$ for which the corresponding λ_i's sum to $\frac{1}{2}\sum_{i=1}^{d+1} \lambda_i$. Now, after some careful technical work, including an appropriate truncation argument to work with integer-valued data, and various anti-concentration arguments such as the Carbery–Wright anticoncentration toolkit [CW01], one can show that the LLL step indeed recovers a constant multiple of the labels x_i, $i = 1, \ldots, d+1$ with probability $1 - \exp(-\Omega(d))$ (see also the next paragraph for more details on this). At this point, it is relatively straightforward to recover u using the linear equations (5).

Now we close by presenting the key technical lemma which ensures that LLL recovers the hidden labels x_i by finding a "short" vector in the lattice defined by the columns of the matrix B in Algorithm 1. Notice that if truncation at N bits was not present, that is we were "allowed" to construct the lattice basis with the non-integer numbers λ_i instead of $(\lambda_i)_N$, then a direct calculation based on (7) would give that the hidden labels are embedded in an element of the lattice simply because we would have

$$B(x_1, \ldots, x_{d+1}, 0)^\top = (0, x_1, \ldots, x_{d+1})^\top.$$

As this "hidden vector" in the lattice is M-independent (and M is taken to be very large) this naturally suggests that this vector may be "short" compared to the others in the lattice. The following lemma states that with probability $1 - \exp(-\Omega(d))$, this is indeed the case. The random lattice generated by the columns of B indeed does not contain any "spurious" short vectors other than the vector of the hidden labels and, naturally, its integer multiples. This implies that the LLL algorithm, despite its $2^{d/2}$ approximation ratio, will indeed return the integer relation that is "hidden in" the z_i's.

Lemma 3.4 (No spurious short vectors)**.** *Let $d \in \mathbb{N}$, $a \in [d^{-c}, d^c]$ for some constant $c > 0$ and $N = \lceil d^4 (\log d)^2 \rceil$. Let $u \in S^{d-1}$ be an arbitrary unit vector, $\Sigma \in \mathbb{R}^{d \times d}$ an arbitrary unknown u-separable matrix, and let $x_i \in \mathbb{Z} \cap [-2^d, 2^d]$ for $i = 1, \ldots, d+1$ be arbitrary but not all zero. Moreover, let $\{z_i\}_{i=1,\ldots,d+1}$ be independent samples from $\mathcal{N}((ax_i)u, \Sigma)$, and let B be the matrix constructed in Algorithm 1 using $\{z_i\}_{i=1,\ldots,d+1}$ as input and N-bit precision. Then, with probability $1 - \exp(-\Omega(d))$ over the samples, for any $t = (t_1, t_2) \in \mathbb{Z}^{d+1} \times \mathbb{Z}$ such that t_1 is not an integer multiple of $x = (x_1, \ldots, x_{d+1})$, the following holds:*

$$\|Bt\|_2 > 2^{2d} \,.$$

The proof of Lemma 3.4 is in Section 6.3.

4 Implications of the success of LLL

The algorithmic guarantee described in Theorem 3.3 has interesting implications for various recently studied problems.

4.1 The planted sparse vector problem

As we motivated in the introduction, a model of notable recent interest is the so-called "planted sparse vector" setting, where one plants a sparse vector in an unknown subspace.

Model 4.1 (Planted sparse vector)**.** *Let $d, n \in \mathbb{N}$ be such that $d \leq n$, and let the sparsity level $\rho \in (0, 1]$. First draw $v \sim \mathsf{BR}(n, \rho)$ and $R \in \mathbb{R}^{d\times d}$ be a uniformly at random chosen orthogonal matrix. Then, we sample i.i.d. $\tilde{z}_1, \ldots, \tilde{z}_{d-1} \sim \mathcal{N}(0, (1/n)I_n)$. Denote by $\tilde{Z} \in \mathbb{R}^{n\times d}$ the matrix whose columns we perceive as generating the "hidden subspace":*

$$\tilde{Z} = \left[v, \tilde{z}_1, \ldots, \tilde{z}_{d-1}\right] . \tag{8}$$

The statistician observes the rotated matrix $Z = \tilde{Z}R$. The goal is, given access to Z, to recover the ρ-sparse vector $v \in \mathbb{R}^n$.

As explained in Section 1, all low-degree polynomial methods, and therefore all spectral methods (in a large class) fail for this model when $n \ll \rho^2 d^2$ [MW21]. Furthermore, in the dense case where $\rho = 1$ and $n \ll d^{3/2}$, even degree-$n^{\Omega(1)}$ SoS algorithms are known to fail [GJJ+20].

Yet, using our Theorem 3.3 we can prove that for any $\rho \in (0, 1]$ the LLL algorithm solves Model 4.1 with $n = d + 1$ samples in polynomial-time. In particular, this improves the state-of-the-art for this task, and provably beats all low-degree algorithms, when $\rho = \omega(1/\sqrt{d})$. The corollary is based on the equivalence (up to rescaling) between Model 4.1 and some appropriate sub-case of Model 3.2 [DDW21].

Corollary 4.2. *Suppose $n, d \in \mathbb{N}$ with $d \leq n$ and $d \to +\infty$. Also let $\rho \in (0, 1]$ scale as $\rho = \omega(1/n)$. Given $n = d + 1$ samples z_i, $i = 1, \ldots, d + 1$ from Model 4.1, Algorithm 1 with input $\sqrt{d+1}z_i$, $i = 1, \ldots, d+1$ and spacing $a = 1/\sqrt{\rho}$, terminates in polynomial time and outputs v (up to a global sign flip) with probability $1 - \exp(-\Omega(\min\{d, \rho n\}))$.*

Proof. As established in [MW21, Lemma 4.21] for z_i a sample from Model 4.1 from a setting where the planted sparse vector is v, the distribution of $\sqrt{n}z_i = \sqrt{d+1}z_i$ is a sample from Model 3.2 with hidden direction v, spacing $a = 1/\sqrt{\rho}$ and covariance $\Sigma = I - vv^\top$. Note that the covariance Σ satisfies Assumption 3.1 for straightforward reasons and same for the condition on the spacing as $n = d + 1$ and $1 > \rho > 1/n$. On top of this, since $\rho n \to +\infty$ the output of $v \sim \mathsf{BR}(n, \rho)$ is not the zero vector with probability $1 - (1 - \rho)^n = 1 - \exp(-\Omega(\rho n))$. Hence, combining the above and Theorem 3.3 we immediately conclude the desired result. □

4.2 Continuous Learning with Errors

As we motivated in the introduction, a second model that is of interest, due to its connections to fundamental problems in lattice-based cryptography, is the homogeneous continuous learning with errors (hCLWE).

Model 4.3 (hCLWE). *Let $\gamma > 0$ be a real number and let ν be a discrete Gaussian of width 1 supported on $(1/\gamma)\mathbb{Z}$. Let $d, n \in \mathbb{N}$. First draw a random unit vector $u \sim \mathcal{S}^{d-1}$ and i.i.d. $x_1, \dots, x_n \sim \nu$. Conditional on u and $\{x_i\}_{i=1}^n$, draw independent samples $z_1, \dots, z_n \in \mathbb{R}^d$ where $z_i \sim \mathcal{N}(x_i u, I_d - uu^\top)$. The statistician observes the matrix $Z \in \mathbb{R}^{n \times d}$ with rows $z_1^\top, \dots, z_n^\top$. The goal is, given access to Z, to recover the unit vector $u \in \mathbb{R}^n$.*

It is known that for $\gamma \geq 2\sqrt{d}$, if we add any inverse-polynomial Gaussian noise in the hidden direction u of hCLWE, even detecting the existence of such discrete structure is hard, under standard worst-case hardness assumptions from lattice-based cryptography [BRST21]. Moreover, there are SQ lower bounds, which are unconditional, for this noisy version [DKS17, BRST21]. As a direct corollary of our Theorem 3.3 we show that in the noiseless case where no Gaussian noise is added, LLL can recover exactly the hidden direction in polynomial time with $d+1$ samples. We remark that while [BRST21] claim that their LLL-based algorithm for (inhomogeneous) CLWE could be generalized to the hCLWE setting, their algorithm uses $O(d^2)$ samples, which is suboptimal.

Corollary 4.4. *Suppose $d \in \mathbb{N}$ with $d \to +\infty$. Assume that $\gamma = \mathrm{poly}(d)$. Given $n = d+1$ samples z_i, $i = 1, \dots, d+1$ from Model 4.3, Algorithm 1 with input z_i, $i = 1, \dots, d+1$ and spacing $a = 1/\gamma$, terminates in polynomial time and outputs v (up to a global sign flip) with probability $1 - \exp(-\Omega(d))$.*

Proof. The proof follows from Claim 4.5. □

Claim 4.5 (hCLWE reduces to the general model). *Let $d \in \mathbb{N}$ and $\gamma = \mathsf{poly}(d)$. Then, with probability $1 - \exp(-\Omega(d))$, $n = \mathsf{poly}(d)$ random samples $\{z_i\}_{i=1}^n$ from Model 4.3 (Gaussian pancakes) with parameter γ is a valid instance of Model 3.2 (our general model) for $\Sigma = I - vv^\top$.*

Proof. Let ν denote the discrete Gaussian supported on $(1/\gamma)\mathbb{Z}$ in Model 4.3. By Claim 2.7, for any $t \geq \gamma$, it holds that

$$\mathbb{P}_{x \sim \nu}[|x| \geq t] \leq 4 \exp(-t^2/2) \, .$$

Let $t = \max\{\sqrt{d}, \gamma\}$. Then, $\mathbb{P}_{x \sim \nu}[|x| \geq t] = \exp(-\Omega(d))$. By a straightforward union bound, with probability $1 - \exp(-\Omega(d))$ the true "labels" $\{x_i\}_{i=1}^n$ of the samples $\{z_i\}_{i=1}^n$ satisfy $|x_i| \leq t = \mathsf{poly}(d)$ for all $i = 1, \dots, n$, and thus this is a valid instance of Model 3.2.

□

4.3 Gaussian Clustering

Finally, a third model which has been extensively studied in the theoretical machine learning community is the (Bayesian) Gaussian clustering model.

Model 4.6 (Gaussian clustering). *Let $d, n \in \mathbb{N}$. Fix some unknown positive semi-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. Now draw a random unit vector $u \sim \mathcal{S}^{d-1}$ and i.i.d. uniform Rademacher labels $x_1, \dots, x_n \sim \{-1, 1\}$. Conditional on u and $\{x_i\}_{i=1}^n$, draw independent samples $z_1, \dots, z_n \in \mathbb{R}^d$ where $z_i \sim \mathcal{N}(x_i u, \Sigma)$. The statistician observes the matrix $Z \in \mathbb{R}^{n \times d}$ with rows $z_1^\top, \dots, z_n^\top$. The goal is, given the observation matrix $Z \in \mathbb{R}^{n \times d}$, to recover (up to a global sign flip) the labels x_i, $i = 1, \dots, n$.*

As explained in the introduction, the recent work by [DDW21] shows that for Model 4.6, exact reconstruction of the labels is possible as long as $u^\top \Sigma^{-1/2} u > 2 \log n$ if Σ is invertible, and of course also in the regime where $\Sigma u = 0$. The authors of [DDW21] show that for any such Σ it is possible to achieve exact reconstruction with $\tilde{O}(d)$ samples by some computationally inefficient method, and

construct and analyze computationally efficient methods that work with $\tilde{O}(d^2)$ samples. On top of this, they conjecture that the regime $\Omega(d) = n = o(d^2)$ is computationally hard based on various forms of rigorous evidence such as failure of SoS and low-degree methods. Notably, the failure of the SoS hierarchy transfers even to the case $\Sigma u = 0$ in the regime $n \leq \tilde{\Omega}(d^{3/2})$. As a direct corollary of Theorem 3.3, we refute their conjecture for any covariance matrix Σ which nullifies u and satisfies a weak "niceness" assumption, namely Assumption 3.1. We show that under this assumption, exact reconstruction of the labels (and therefore of the clusters) is possible in polynomial time with only $n = d + 1$ samples.

Corollary 4.7. *Suppose $d \in \mathbb{N}$ with $d \to +\infty$. Given $n = d + 1$ samples z_i, $i = 1, \dots, d + 1$ from Model 4.6 with arbitrary covariance Σ which is u-weakly separable per Assumption 3.1. Then Algorithm 1 with input z_i, $i = 1, \dots, d + 1$ and spacing $a = 1$ terminates in polynomial-time and outputs exactly the labels x_i, $i = 1, \dots, n$ (up to a global sign flip) with probability $1 - \exp(-\Omega(d))$.*

The finite SNR regime. Our algorithmic guarantees for the linear sample complexity regime crucially depend on the "noiseless" aspect of the model, translated into our weak separability assumption (Assumption 3.1), which corresponds to SNR $= \infty$ in the notation of [DDW21]. An important question is to understand whether the linear sample complexity guarantees are possible in polynomial-time for finite (albeit growing with dimension) SNR. When SNR $= \exp(d^{\Theta(1)})$, the similar LLL-based procedure by [DK21] is proven to work and we do expect that our algorithm remains successful in this regime as well. On the other hand, from [DDW21] we know that SNR $> 2 \log d$ is necessary in order to ensure exact recovery. This leaves open a wide range of SNR. While we do not provide an answer to this question here, we want to note that the SNR $= \mathsf{poly}(d)$ regime would require some novel ideas, at least in terms of the analysis of the suggested algorithms. Both our algorithm and the algorithm by [DK21] are analyzed in a way that establishes success for both Gaussian Clustering (Rademacher labels) and hCLWE (Discrete Gaussians) simultaneously. Yet, the known computational hardness of hCLWE [BRST21] strongly suggests that hCLWE cannot be solved with inverse polynomial noise by any polynomial-time algorithm with polynomially many samples. Hence, both the analyses of the currently proposed algorithms for Gaussian Clustering with $\Theta(d)$ samples are expected to fail in this regime. We leave it as an interesting open question whether there exists a polynomial-time algorithm that succeeds in the Gaussian Clustering model for SNR $\ll \exp(d)$ and linear sample complexity.

5 Information-theoretic lower bounds

In this section we establish information-theoretic lower bounds associated with problem (1) both for *parameter recovery*, i.e., recovering the hidden direction u, as well as *label recovery*, i.e., recovering the hidden labels $\{x_i\}_{i=1}^n$. These lower bounds show that the sample complexity $n = d + 1$ for our LLL-based algorithm is optimal in the following sense: for exact recovery of the hidden direction u (up to a global sign flip), even $n = d - 1$ samples are not sufficient. For exact recovery of the labels (up to a global sign flip), $n = (1 - o(1))d$ samples are required.

5.1 Optimality of $d + 1$ samples for parameter recovery

In this section we establish that when $a = 1$, $n \leq d - 1$, and $x_i \in \{-1, 1\}$, for $i = 1, \dots, n$, one cannot information-theoretically exactly recover the hidden direction $u \in \mathcal{S}^{d-1}$ from n independent samples $z_i \sim \mathcal{N}(x_i u, I - uu^\top)$. This establishes the optimality of our LLL approach which achieves

a much more generic guarantee in Theorem 3.3 in recovering u exactly, up to one additional sample compared to the information-theoretic lower bound.

In fact, our lower bound is somewhat stronger. We assume that the statistician also has exact knowledge of the signs $x_i \in \{-1, 1\}$. We show that even in this setting exact recovery of the hidden direction u with probability greater than $1/2$ using at most $d-1$ samples is impossible. Notice that if x_i's are known to the statistician, there is no global sign ambiguity with respect to u.

Theorem 5.1 (Parameter recovery)**.** *Let arbitrary $x_i \in \{-1, 1\}$, $i = 1, \dots, d-1$ be fixed and known to the statistician. Moreover, let $u \in \mathcal{S}^{d-1}$ be a uniformly random unit vector. For each $i = 1, \dots, d-1$ let z_i be a sample generated independently from $\mathcal{N}(x_i u, I - uu^\top)$. Then it is information-theoretically impossible to construct an estimate $\hat{u} = \hat{u}(\{(x_i, z_i)\}_{i=1}^n) \in \mathcal{S}^{d-1}$ which satisfies $\hat{u} = u$ with probability larger than $1/2$.*

5.2 Information-theoretic lower bound for label recovery

In this section we establish that when $n = \lfloor \rho d \rfloor$ for any fixed $\rho \in (0, 1)$ and $x_i \in \{-1, 1\}$, $i = 1, \dots, n$, one cannot information-theoretically exactly recover the latent variables $\{x_i\}$ from independent samples $z_i \sim \mathcal{N}(x_i u, I - uu^\top)$, where $u \in \mathcal{S}^{d-1}$ is the unknown direction. Hence, combined with Theorem 3.3, we can conclude that our LLL approach achieves the optimal information-theoretic sample complexity for label recovery up to a $1 + o(1)$ factor.

Theorem 5.2 (Label recovery)**.** *Let $\rho \in (0, 1)$ be a fixed constant and let $n = \lfloor \rho d \rfloor$. Let $x \in \{-1, 1\}^n$ be drawn uniformly at random, and $u \in \mathcal{S}^{d-1}$ be a uniformly random unit vector. For each $i = 1, \dots, n$ let $z_i \in \mathbb{R}^d$ be a sample generated independently from $\mathcal{N}(x_i u, I - uu^\top)$. Then no estimator can exactly recover the labels $\{x_i\}_{i=1}^n$ up to a global sign flip from the observations $\{z_i\}_{i=1}^n$ with probability $1 - o(1)$.*

The main idea of the proof, which can be found in Section 7.2, is to first compute the conditional density (Lemma 7.3), which we denote by $f(Z|X = x)$, of the observations $Z \in \mathbb{R}^{n \times d}$ given labels $x \in \{-1, 1\}^n$. We achieve this by viewing Model 4.6 as the limit $\sigma \to 0$ of the Gaussian clustering model with covariance $\Sigma = \sigma^2 uu^\top + (I - uu^\top)$, and then identifying small perturbations in the labels x that result in small relative fluctuations in the conditional density $f(Z|X = x)$ that crucially remain uniformly bounded in d. In other words, we show that there exists a universal constant $\delta > 0$ such that for any $x \in \{-1, 1\}^n$ and "most" $Z \in \mathbb{R}^{n \times d}$ (under the measure induced by $f(Z|X = x)$), there exists $x' \in \{-1, 1\}^n$ such that $f(Z|X = x') \geq \delta f(Z|X = x)$. This spread in the conditional density, and consequently the posterior, directly impacts the accuracy of any estimator for the labels, including the MAP estimator. The scaling $n/d = \rho \in (0, 1)$ is used in our argument to obtain enough concentration of a certain quadratic form that in turn determines the conditional density (Remark 7.6 and Lemma 7.8). Our label recovery lower bound of Theorem 5.2 is thus slightly below the parameter recovery lower bound of Theorem 5.1, leaving open the regime $n = d - o(d)$. Closing this gap would require an alternative concentration argument, which we leave as an open question.

Finally, let us relate these lower bounds with those established in the literature. Statistical lower bounds for binary Gaussian mixtures have been extensively studied previously, e.g. in [LZ16, GV19, Roy17, Nda18, DDW21]. In our setting, although the mean and covariance are unknown, their particular relationship, i.e. $\Sigma = I - uu^\top$, is known to the statistician, resulting in a sharp transition at $n = (1 - o(1))d$ with explicit constant. We emphasize however that we focus on exact recovery rather than misclassification rate guarantees. Lastly, we operate in a regime where the covariance is singular, as opposed to [DDW21] where the covariance is assumed to be invertible, which requires extra technical care.

6 Proof of Algorithm 1 correctness

6.1 Towards proving Theorem 3.3: auxiliary lemmas

We present here three auxiliary lemmas for proving Theorem 3.3 and Lemma 3.4. The first lemma establishes that given a small (in ℓ_2-norm) "approximate" integer relation between real numbers, one can appropriately truncate each real number to a sufficiently large number of bits, so that the truncated numbers satisfy a small (in ℓ_2-norm) integer relation between them. This lemma, which is an immediate implication of [SZB21, Lemma D.6], is important for the appropriate application of the LLL algorithm, which needs to receive integer-valued input. Recall that for a real number x we denote by $(x)_N$ its truncation to its first N bits after zero, i.e. $(x)_N := 2^{-N}\lfloor 2^N x \rfloor$.

Lemma 6.1 ("Rounding" approximate integer relations [SZB21, Lemma D.6])**.** *Let $d \in \mathbb{N}$ be a number and let $n \in \mathbb{N}$ be such that $n \leq C_0 d$ for some constant $C_0 > 0$. Moreover, suppose for some constant $C_1 > 0$, a (real-valued) vector $s \in \mathbb{R}^n$ satisfies $\langle m, s\rangle = 0$ for some $m \in \mathbb{Z}^n$. Then for some sufficiently large constant $C > 0$, if $N = \lceil d^4(\log d)^2 \rceil$, there is an $m' \in \mathbb{Z}^{n+1}$ which is equal to m in the first n coordinates, satisfies $\|m'\|_2 \leq Cd^{\frac{1}{2}}\|m\|_2$, and is an integer relation for the numbers $(s_1)_N, \ldots, (s_n)_N, 2^{-N}$.*

We need the following anticoncentration result.

Lemma 6.2 (Anticoncentration of misaligned integer combinations)**.** *Assume that $d^c > a > 1/d^c$ for some $c > 0$ constant. Let $u \in S^{d-1}$ be an arbitrary unit vector and let $x_1, \ldots, x_{d+1} \in \mathbb{Z}$ be an arbitrary sequence of integers, which are not all equal to zero. Now for a sequence of integers $t = (t_1, \ldots, t_{d+1}) \in \mathbb{Z}^{d+1}$, we define the (multi-linear) polynomial $P_t(z_1, \ldots, z_{d+1})$ in $d(d+1)$ variables by*

$$P_t(z_1, \ldots, z_{d+1}) = \det(Z)t_1 + \sum_{i=2}^{d+1} \det(Z_{-i})t_i \,, \tag{9}$$

where each $z_1, \ldots, z_{d+1}$ is assumed to have a d-dimensional vector form, Z denotes the $d \times d$ matrix with $z_2, \ldots, z_{d+1}$ as its columns, and each Z_{-i} for $i = 2, \ldots, d+1$ denotes the $d \times d$ matrix formed by swapping out the $(i-1)$-th column of Z with $-z_1$.

Suppose z_i's are drawn independently from $\mathcal{N}((ax_i)u, \Sigma)$ for some $u \in \mathcal{S}^{d-1}$ and $\Sigma \in \mathbb{R}^{d\times d}$ which is u-weakly separable per Assumption 3.1 and eigenvalues $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_d$. Then, for any $t \in \mathbb{Z}^{d+1}$ it holds that

$$\mathbb{E}[P_t(z_1, \ldots, z_{d+1})] = 0 \tag{10}$$

and

$$\mathrm{Var}(P_t(z_1, \ldots, z_{d+1})) = (d-1)!a^{2d}\left(\prod_{i=2}^{d}\lambda_i\right)^2 \sum_{1\leq i<j\leq d+1}(t_i x_j - t_j x_i)^2 \,. \tag{11}$$

Furthermore, for some universal constant $B > 0$ the following holds. If $t \neq cx$ for any $c \in \mathbb{R}$, where we denote $x = (x_1, \ldots, x_{d+1})$, then for any $\epsilon > 0$,

$$\mathbb{P}(|P_t(z_1, \ldots, z_{d+1})| \leq \epsilon) \leq Bd^B\epsilon^{\frac{1}{d}}. \tag{12}$$

Proof. We first describe how (12) follows from (10) and (11). First, notice that under the assumption on the integer sequence $t_i, i = 1, \dots, d+1$ not being a multiple of the sequence of integers $x_i, i = 1, \dots, d+1$ it holds that for some $i, j = 1, \dots, d+1, i \neq j$ with $(t_i x_j - t_j x_i)^2 \geq 1$. In particular, using (11) we have

$$\mathrm{Var}(P_t(z_1, \dots, z_{d+1})) \geq (d-1)! a^{2d} \left(\prod_{i=2}^{d} \lambda_i \right)^2 .$$

But now notice that from Assumption 3.1 and $a > d^{-c}$, it holds for some constant $C' > 0$ that

$$a^{2d} \left(\prod_{i=2}^{d} \lambda_i \right)^2 \geq d^{-C'd}.$$

Hence, it holds that

$$\mathrm{Var}(P_t(z_1, \dots, z_{d+1})) \geq d^{-C'd}.$$

Now we employ [MNV16, Theorem 1.4] (originally proved in [CW01]) which implies that for some universal constant $B > 0$, since our polynomial is multilinear and has degree $d+1$, it holds for any $\epsilon > 0$ that

$$\mathbb{P}\left(|P_t(z_1, \dots, z_{d+1})| \leq \epsilon \sqrt{\mathrm{Var}(P_t(z_1, \dots, z_{d+1}))} \right) \leq B d \epsilon^{\frac{1}{d}}.$$

Using our lower bound on the variance we conclude the result.

Now we proceed with the mean and variance calculation. As this statement is about the first and second moment of P_t and the determinant operator is invariant up to basis transformations, we may assume without loss of generality that $u = e_1$, that is, u is equal to the first standard basis vector, and the remaining standard basis vectors are the remaining eigenvectors of Σ. Recall that z_i's are drawn in an independent fashion from $\mathcal{N}((a x_i) u, \Sigma)$. Hence for a sequence of i.i.d. $w_i \sim \mathcal{N}(0, I_{d-1}), i = 1, \dots, d+1$ we may assume from now on that,

$$z_i = \begin{bmatrix} a x_i \\ \Lambda w_i \end{bmatrix} \tag{13}$$

for $\Lambda := \mathrm{diag}(\lambda_2, \dots, \lambda_d)$.

Now let us define the $(d-1) \times (d-1)$ matrix W_{-j} for each $2 \leq j \leq d+1$ as the matrix formed using $w_2, \dots, w_{d+1}$ *except* w_j as its column vectors, and define functions $\psi_i : \mathbb{R}^{(d-1)\times(d-1)} \to \mathbb{R}$ for each $i = 2, \dots, d+1$ to be the determinant of W_{-j} with the column corresponding to w_i swapped by $-w_1$. For instance, if $2 \leq i \neq j \leq d+1$, then

$$\psi_i(W_{-j}) := \det(w_2, \dots, w_{i-1}, -w_1, w_{i+1}, \dots, w_{j-1}, w_{j+1}, \dots, w_{d+1}). \tag{14}$$

We abuse notation and also write

$$\psi_1(W_{-j}) := \det(w_2, \dots, w_{j-1}, w_{j+1}, \dots w_{d+1}) = \det(W_{-j}). \tag{15}$$

As the result is clearly a-homogeneous of degree $2d$ we assume in what follows that $a = 1$. Now by direct expansion along the first row of the corresponding matrices we have

$$\det(Z) = \sum_{j=2}^{d+1} (-1)^j x_j |\det(\Lambda)| \psi_1(W_{-j}) \, ,$$

and for each $i \geq 2$,

$$\det(Z_{-i}) := (-1)^{i+1} x_1 |\det(\Lambda)| \psi_1(W_{-i}) + \sum_{j=2, j\neq i}^{d+1} (-1)^j x_j |\det(\Lambda)| \psi_i(W_{-j}) \, .$$

Since $d > 1$ and w_i are i.i.d. $\mathcal{N}(0, I_d)$ we can immediately conclude that for all $i \geq 1, j \geq 2, i \neq j$,

$$\mathbb{E}[\psi_i(W_{-j})] = 0.$$

Hence,

$$\mathbb{E}[P_t(z_1, \ldots, z_{d+1})] = t_1 \mathbb{E}[\det(Z)] + \sum_{i=2}^{d+1} t_i \mathbb{E}[\det(Z_{-i})] = 0.$$

Now we calculate the second moment of the polynomial. In what follows, we slightly abuse notation and denote $Z_{-1} := Z$ for notational convenience. First, again by direct expansion of the determinant and the fact that w_i's for $i = 1, \ldots, d+1$ have i.i.d. standard Gaussian entries it holds by direct inspection that for all $i, j \in [d+1]$ with $i \neq j$,

$$\mathbb{E}[\psi_i(W_{-j})^2] = (d-1)! |\det(\Lambda)|^2 \, , \tag{16}$$

and unless $\{i, j\} = \{k, \ell\}$, it holds that

$$\mathbb{E}[\psi_i(W_{-j}) \psi_k(W_{-\ell})] = 0. \tag{17}$$

We now calculate for $2 \leq i \neq j \leq d+1$ the term $\mathbb{E}[\psi_i(W_{-j}) \psi_j(W_{-i})]$. We assume without loss of generality that $i < j$. Notice that for $\Pi_c \in \{0,1\}^{d-1 \times d-1}$, the permutation matrix corresponding to the cycle-permutation $c := (i-1, i, \ldots, j, j-1) \in \mathrm{Sym}([d-1])$, the matrix

$$(w_2, \ldots w_{i-1}, -w_1, w_{i+1}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_{d+1}) \, ,$$

equals

$$\Pi_c (w_2, \ldots w_{i-1}, w_{i+1}, \ldots, w_{j-1}, -w_1, w_{j+1}, \ldots, w_{d+1}) \, .$$

Hence,

$$\psi_i(W_{-j}) \psi_j(W_{-i}) = \det(\Pi_c) \psi_i^2(W_{-j}) = (-1)^{\mathrm{sgn}(c)} \psi_i^2(W_{-j}) = (-1)^{i-j+1} \psi_i^2(W_{-j}) \, .$$

In particular,

$$\mathbb{E}[\psi_i(W_{-j}) \psi_j(W_{-i})] = (-1)^{i-j+1} (d-1)! |\det(\Lambda)|^2 \, . \tag{18}$$

Now using (16), (17), we have for each $1 \leq i \leq d+1$,

$$\mathbb{E}[\det(Z_{-i})^2] = (d-1)! \sum_{j=1, j\neq i}^{d+1} x_j^2 |\det(\Lambda)|^2 \, , \tag{19}$$

and using (16), (17), and (18) we have for all $i \neq j$ that

$$\mathbb{E}[\det(Z_{-i}) \det(Z_{-j})] = -(d-1)! x_i x_j |\det(\Lambda)|^2 \, . \tag{20}$$

Hence, it holds that

$$\begin{aligned}
\mathbb{E}[P_t(z_1,\dots,z_{d+1})^2] &= |\det(\Lambda)|^2 \sum_{i,j=1}^{d+1} t_i t_j \mathbb{E}[\det(Z_{-i})\det(Z_{-j})] \\
&= |\det(\Lambda)|^2 \sum_{i=1}^{d+1} t_i^2 \mathbb{E}[\det(Z_{-i})^2] + \sum_{i,j=1, i\neq j}^{d+1} t_i t_j \mathbb{E}[\det(Z_{-i})\det(Z_{-j})] \\
&= (d-1)!|\det(\Lambda)|^2 \left(\sum_{i,j=1,i\neq j}^{d+1} t_i^2 x_j^2 - \sum_{i,j=1,i\neq j}^{d+1} t_i t_j x_i x_j \right) \\
&= (d-1)!|\det(\Lambda)|^2 \left(\sum_{1\le i<j\le d+1} (t_i x_j - t_j x_i)^2 \right) .
\end{aligned}$$

□

The following lemma establishes multiple structural properties of the $d+1$ samples.

Lemma 6.3. *Let $u \in S^{d-1}$ be an arbitrary unit vector and let $x_i \in \mathbb{Z} \cap [-2^d, 2^d]$ for $i = 1,\dots,d+1$ be arbitrary integers which are not all equal to zero. Let also spacing a with $d^{-c} < a < d^c$ for some $c > 0$ and Σ which is u-weakly separable per Assumption 3.1. We observe $d+1$ samples of the form z_i, where for each $i = 1,\dots,d+1$, z_i is an independent sample from $\mathcal{N}((ax_i)u, \Sigma)$. We denote by $Z \in \mathbb{R}^{d\times d}$ the (random) matrix with columns given by the d vectors $z_2,\dots,z_{d+1}$. The following properties hold.*

(1) The matrix Z is invertible almost surely.

(2) With probability $1-\exp(-\Omega(d))$ over the z_i's,

$$\|Z^{-1}z_1\|_\infty = O(2^{2d^2}).$$

(3) With probability $1-\exp(-\Omega(d))$ over the z_i's,

$$0 < |\det(Z)| = O(2^{d^2}).$$

Proof. For the fact that Z is invertible, consider its determinant, that is, the random variable $\det(Z)$. We claim that $\det(Z) \neq 0$ almost surely. Note that to prove this, by invariance of the determinant to the change of basis, we may assume without loss of generality that $u = e_1$, that is, u is the first standard basis vector, and the remaining standard basis vectors are the remaining eigenvectors of Σ. Under this assumption, for each $i = 1,\dots,d+1$, we can write using Assumption 3.1

$$z_i = \begin{bmatrix} ax_i \\ \Lambda w_i \end{bmatrix} ,$$

where $\Lambda = \mathrm{diag}(\lambda_2,\dots,\lambda_d)$ and w_i's are i.i.d. samples from $\mathcal{N}(0, I_{d-1})$. In other words, the first row of Z consists of $ax_2,\dots,ax_{d+1}$, and the rest are coordinates of Λw_i, where each w_i is a vector with i.i.d. standard Gaussian entries. Now the result follows from the fact that since not all x_i are equal to zero and also none of the λ_i's are zero from Assumption 3.1, the determinant $\det(Z)$ with fixed $x_2,\dots,x_{d+1}$ is a non-zero polynomial of the entries of $w_2,\dots,w_{d+1}$. As all entries of

w_i are distributed as i.i.d. standard Gaussians, the random polynomial $\det(Z)$ is almost surely non-zero [CT05].

For the second part, notice that by Cramer's rule for $i = 1, \ldots, d-1$, the i-th coordinate of $Z^{-1}z_1$ equals the quantity $\lambda_{i+1}(Z) := \det(z_2, \ldots, z_i, -z_1, z_{i+1}, \ldots, z_{d+1})/\det(Z)$ almost surely. Hence, again by the rotational invariance property of the determinant operator, we may assume that $u = e_1$ and the remaining standard basis vectors are the remaining eigenvectors of Σ. Let $q^{(i)} \in \mathbb{Z}^{d+1}$ be an integer-valued vector such that $q^{(i)}_j = 1$ if $i = j$ and $q^{(i)}_j = 0$ otherwise. Now using the notation of Lemma 6.2 we have that $P_{q^{(i)}}(z_1, \ldots, z_{d+1}) = \det(Z_{-i})$. By applying the anticoncentration result from Lemma 6.2 for the polynomial $P_{q^{(1)}}(z_1, \ldots, z_{d+1})$ and $\epsilon = 2^{-d^2}$ we conclude that

$$|\det(Z)| = |P_{q^{(1)}}(z_1, \ldots, z_{d+1})| \geq 2^{-d^2} \tag{21}$$

with probability $1 - \exp(-\Omega(d))$. Furthermore, for all $i = 1, \ldots, d+1$ it holds that

$$\mathbb{E}[P_{q^{(i)}}(z_1, \ldots, z_{d+1})^2] = \mathrm{Var}(P_{q^{(i)}}(z_1, \ldots, z_{d+1})) = a^{2d} d! \|x\|_2 \leq a^{2d} |\det(\Lambda)|^2 2^{10d \log d} \|x\|_2^2 ,$$

where $x := (x_1, \ldots, x_{d+1})^\top$ where $\Lambda = \mathrm{diag}(\lambda_2, \ldots, \lambda_d)$ and $\lambda_i, i > 1$ are the non-zero eigenvalues of Σ per Assumption 3.1. Hence, by Markov's inequality, the fact that $a < d^c$, the Assumption 3.1 and a union bound over i, we have for all $i = 1, \ldots, d+1$ that

$$|P_{q^{(i)}}(z_1, \ldots, z_{d+1})| \leq 2^{d^2/2} \|x\|_2^2 \tag{22}$$

with probability $1 - \exp(-\Omega(d))$.

Combining Eq.(21) and Eq.(22), we conclude that for all $i = 2, \ldots, d$,

$$|\lambda_i(Z)| = |P_{q^{(i)}}(z_1, \ldots, z_{d+1})/P_{q^{(1)}}(z_1, \ldots, z_{d+1})| \leq 2^{3d^2/2} \|x\|_2^2$$

with probability $1 - \exp(-\Omega(d))$. Since $\|x\|_2^2 = O(2^{2d})$ we have $\|Z^{-1}z_1\|_\infty \leq 2^{3d^2/2}\|x\|_2^2 \leq 2^{2d^2}$ with probability $1 - \exp(-\Omega(d))$. This concludes the proof of the second part.

Finally, Eq.(22) for $i = 1$ and the fact $\|x\|_2^2 = O(2^{2d})$ imply

$$|\det(Z)| = |P_{q^{(1)}}(z_1, \ldots, z_{d+1})| \leq 2^{d^2} \tag{23}$$

with probability $1 - \exp(-\Omega(d))$. This concludes the proof of the third part. □

6.2 Proof of Theorem 3.3

We now proceed with the proof of the Theorem 3.3 using the lemmas from the previous sections.

Theorem 3.3 (Restated). *Algorithm 1, given as input independent samples $(z_i)_{i=1,\ldots,d+1}$ from Model 3.2 with hidden direction u, covariance Σ, and true labels $\{x_i\}_{i=1,\ldots,d+1}$ satisfies the following with probability $1 - \exp(-\Omega(d))$: there exists $\epsilon \in \{-1, 1\}$ such that the algorithm's outputs $\{\hat{x}_i\}_{i=1,\ldots,d+1}$ and $\hat{u} \in S^{d-1}$ satisfy*

$$\hat{x}_i = \epsilon x_i \quad \text{for } i = 1, \ldots, d+1$$
$$\text{and } \hat{u} = \epsilon u .$$

Moreover, Algorithm 1 terminates in poly(d) *steps.*

Proof. We start with noticing that for an algorithm to recover u, x_i up a to global sign flip it suffices to recover the values of $\{x_i\}_{i=2,\dots,d+1}$ up to a global non-zero constant multiple. Indeed, since we already know the value of z_i's, if we learn the x_i's up to a constant, call it $C > 0$, then we can solve the linear system of d (independent) equations and with d unknowns given by $\langle z_i, v\rangle = Cax_i = C\langle z_i, u\rangle, i = 2, \dots, d+1$. Since by Lemma 6.3 the matrix Z, also formed in Algorithm 1, which is the $d \times d$ matrix with $z_2, \dots, z_{d+1}$ as its column vectors, is invertible almost surely, one can indeed solve this linear system to recover $v = Cu$, that is the same constant C times u. Since u is assumed to be unit norm one can then recover the quantity $|C| = \|v\|_2$, which is the absolute value of the unknown constant. Hence one can output for some $\epsilon = C/|C| \in \{-1, 1\}$ the estimated vector $Cu/|C| = \epsilon u$ and the estimated labels $Cx_i/|C| = \epsilon x_i, i = 1, \dots, d+1$ which are indeed the hidden direction u and the true labels $x_i, i = 1, \dots, d+1$ up to a global sign flip.

Now our proposed Algorithm 1 follows exactly this path: it first recovers a non-zero constant multiple of the x_i's (this is the values of the vector t_1 output by the LLL step) with probability $1 - \exp(-\Omega(d))$. Then it uses the simple procedure described above to output both the labels $x_i, i = 1, \dots, d+1$ and u up to a global constant multiple. This second part comprises exactly the last steps of the algorithm after the LLL step. The main procedure of our algorithm therefore is to use an appropriate application of LLL to learn the exact values of x_i up to a global sign flip. We now analyze the success of the LLL step to recover a global constant multiple of the x_i's with probability $1 - \exp(-\Omega(d))$.

Now the algorithm does not terminate in the second step exactly because of the almost sure invertibility of the matrix Z, per Lemma 6.3. Let us now analyze the (random) lattice $L = L(B)$ generated by the basis B, which is constructed in the next step of Algorithm 1.

First, observe that the real numbers $\{\lambda_i\}_{i=1,2,\dots,d+1}$ used in the top row of the lattice basis B, satisfy by definition

$$\sum_{i=1}^{d+1} \lambda_i z_i = 0 \, .$$

Hence, we conclude that since $\langle z_i, u\rangle = ax_i$ for the unknown direction $u \in S^{d-1}$ and spacing $a > 0$, it holds that

$$\sum_{i=1}^{d+1} \lambda_i a x_i = \sum_{i=1}^{d+1} \lambda_i \langle u, z_i\rangle = \langle u, \sum_{i=1}^{d+1} \lambda_i z_i\rangle = 0 \tag{24}$$

and therefore

$$\sum_{i=1}^{d+1} \lambda_i x_i = 0 \, . \tag{25}$$

We now show an upper bound on the shortest vector length of L, which we denote by $\mu(L)$. More precisely, we show that

$$\mu(L) = O(d2^d) \, .$$

To this end, define a real-valued vector $s \in \mathbb{R}^{d+1}$ with $s_i = \lambda_i$ for $i = 1, \dots, d+1$, and also an integer-valued vector $m \in \mathbb{Z}^{d+1}$ with $m_i = x_i$ for $i = 1, \dots, d+1$. Then, the integer relation (25) implies that $\langle s, m\rangle = 0$. Since $|x_i| \le 2^d$ for all $i = 1, \dots, d+1$ it also holds almost surely that $\|m\|_2 = \|x\|_2 \le \sqrt{d}2^d$. By Lemma 6.1, for the bit-precision N chosen by Algorithm 1, there exists

an integer $m'_{d+2} \in \mathbb{Z}$ such that $m' = (m, m'_{d+2}) \in \mathbb{Z}^{d+2}$ satisfies $\|m'\|_2 = O(d2^d)$ and is an integer relation for $(\lambda_1)_N, \ldots, (\lambda_{d+1})_N, 2^{-N}$.

Now define $b \in (2^{-N}\mathbb{Z})^{d+2}$ given by $b_i = (\lambda_i)_N$ for $i = 1, \ldots, d+1$, and $b_{d+2} = 2^{-N}$. Notice that $b_1 = (1)_N = 1$ and furthermore that the $\tilde{v}$ defined by the algorithm satisfies $\tilde{v} = (b_2, \ldots, b_{d+2})$. On top of this, we have that the m' defined in previous paragraph is an integer relation for b with $\|m'\|_2 = O(d2^d)$. Hence, $Bm' = (0, m')^\top$. It follows that $\mu(L) = O(d2^d)$ with probability $1 - \exp(-\Omega(d))$, since $\mu(L) \leq \|Bm'\|_2 = O(d2^d)$.

Recall from Theorem 2.3 that the LLL algorithm is guaranteed to return a lattice vector of ℓ_2-norm smaller than $2^{\frac{d+2}{2}}\mu(L)$. Now we employ Lemma 3.4 which combined with the fact that $2^{\frac{d+2}{2}}\mu(L) \leq 2^{2d}$ for sufficiently large d almost surely, allows us to conclude that the LLL algorithm returns a non-zero lattice vector $B(t_1, t_2)^\top$, where $t_1 \in \mathbb{Z}^{d+1}$ and $t_2 \in \mathbb{Z}$, such that t_1 is an integer multiple of $x = (x_1, \ldots, x_{d+1})$ with probability $1 - \exp(-\Omega(d))$. Hence, using t_1 the algorithm recovers a global non-zero constant multiple of the x_i's for $i = 1, \ldots, d+1$ with probability $1 - \exp(-\Omega(d))$.

For the termination time, it suffices to establish that the step using the LLL basis reduction algorithm can be performed in $\mathsf{poly}(d)$ time. To ensure $\mathsf{poly}(d)$ time for the LLL step, it suffices to show that the entries of the lattice basis B are not too large with probability $1 - \exp(-\Omega(d))$. More precisely, the running time of LLL depends on the logarithm of the largest entry in B by Theorem 2.3. Clearly, N and $\log M$ are polynomial in d. Finally, direct inspection and Lemma 6.3 implies that the quantity $\log \|\lambda\|_\infty$, where $\lambda = (\lambda_1, \ldots, \lambda_{d+1})^\top$ is as defined in Algorithm 1, is polynomially bounded with probability $1 - \exp(-\Omega(d))$. This establishes the $\mathsf{poly}(d)$ running time of the LLL step. □

6.3 Proof of Lemma 3.4

We focus this section on proving the key technical Lemma 3.4. As mentioned above, the proof of the lemma is quite involved, and, potentially interestingly, it requires the use of anticoncentration properties of the coefficients λ_i, which are rational functions of the coordinates of x_i, as discussed in Lemma 6.2.

Lemma 3.4 (Restated). *Let $d \in \mathbb{N}$, $a \in [d^{-c}, d^c]$ for some $c > 0$ and $N = \lceil d^4 (\log d)^2 \rceil$. Let $u \in S^{d-1}$ be an arbitrary unit vector, $\Sigma \in \mathbb{R}^{d \times d}$ an arbitrary u-weakly separable matrix and let $x_i \in \mathbb{Z} \cap [-2^d, 2^d]$ for $i = 1, \ldots, d+1$ be arbitrary but not all zero. Moreover, let $\{z_i\}_{i=1,\ldots,d+1}$ be independent samples from $N((ax_i)u, \Sigma)$, and let B be the matrix constructed in Algorithm 1 using $\{z_i\}_{i=1,\ldots,d+1}$ as input and N-bit precision. Then, with probability $1 - \exp(-\Omega(d))$ over the samples, for any $t = (t_1, t_2) \in \mathbb{Z}^{d+1} \times \mathbb{Z}$ such that t_1 is not an integer multiple of $x = (x_1, \ldots, x_{d+1})$, the following holds:*

$$\|Bt\|_2 > 2^{2d} .$$

Proof of Lemma 3.4. Let $t = (t_1, t_2) \in \mathbb{Z}^{d+1} \times \mathbb{Z}$ be arbitrary non-zero integer coefficients. Our proof consists of characterizing integer coefficients t for which the corresponding lattice vector Bt is "short", that is,

$$\|Bt\|_2 \leq 2^{2d}. \tag{26}$$

In what follows, by a *short lattice vector* we refer to the condition (26).

We first show that with probability $1 - \exp(-\Omega(d))$, lattice vectors can only be short for integer coefficients contained in some bounded rectangle $\mathcal{R} \subset \mathbb{Z}^{d+2}$, which we define below (see Eq.(28)).

Then, we apply our anticoncentration lemma (Lemma 6.2) and a union bound over a subset of $\mathcal{R}$ to conlcude that with probability $1-\exp(-\Omega(d))$, the only short lattice vectors are ones whose integer coefficients satisfy $t_1 = cx$ for some $c \in \mathbb{Z}$.

To this end, we first observe that entries of the first row of B are elements of $M\mathbb{Z}$, as by direct inspection $(Bt)_1 = M(\sum_{i=1}^{d+1} \left(2^N(\lambda_i)_N\right)(t_1)_i + t_2)$. It follows that if t is not an integer relation for the numbers $(\lambda_1)_N, \ldots, (\lambda_{d+1})_N, 2^{-N}$, then $\|Bt\|_2 \geq M = 2^{2d}$. Hence, it suffices to restrict our attention to t's which are integer relations, that is,

$$\sum_{i=1}^{d+1} (\lambda_i)_N (t_1)_i + t_2 2^{-N} = 0 \, .$$

Note that it cannot be the case that $t_1 = 0$ since this implies, by the integer relation above, $t_2 = 0$, and therefore the pair $t = (t_1, t_2)$ are zero, a contradiction. Hence, from now on we restrict ourselves only to the case where $t_1 \neq 0$.

Let us denote by t' the vector t without the first coordinate $(t_1)_1$, i.e., $t' = ((t_1)_2, \ldots, (t_1)_{d+1}, t_2)$. Our second observation is that $\|Bt\|_2 \geq \|t'\|_\infty$ because of the use of the submatrix I_{d+1} in the definition of B. This implies that any short lattice vector Bt must satisfy $\|t'\|_\infty \leq 2^{2d}$. Moreover, since t is an integer relation and $\lambda_1 = 1$, we have

$$|(t_1)_1| = \left| \sum_{i=2}^{d+1} (\lambda_i)_N (t_1)_i + t_2 2^{-N} \right| \leq \|t'\|_\infty \left(\|\lambda\|_1 + 2^{-N} \right) \, . \tag{27}$$

Now in the notation of Lemma 6.3 we have $\lambda = -Z^{-1} z_1$. Hence using Lemma 6.3 and the elementary fact that $\|\lambda\|_1 \leq (d+1)\|\lambda\|_\infty$, it holds with probability $1 - \exp(-\Omega(d))$ that $\|\lambda\|_1 = O(2^{2d^2})$. It follows that, for sufficiently large d, any short lattice vector Bt must satisfy $|(t_1)_1| \leq 2^{3d^2}$ with probability $1 - \exp(-\Omega(d))$. Hence, with probability $1 - \exp(-\Omega(d))$, every short vector Bt in the random lattice $L = L(B)$ has its integer coefficients t contained in $\mathcal{R}$, which is defined as

$$\mathcal{R} = \{(a, b) \in \mathbb{Z} \times \mathbb{Z}^{d+1} \, : \, |a| \leq 2^{3d^2}, \|b\|_\infty \leq 2^{2d}\} \, . \tag{28}$$

From $\mathcal{R}$, we also define $\mathcal{R}_1 \subset \mathbb{Z}^{d+1}$ such that $\mathcal{R}_1 = \{t_1 \in \mathbb{Z}^{d+1} \, : \, t = (t_1, t_2) \in \mathcal{R}\}$.

We now show using a union bound over $t \in \mathcal{R}$ that with probability $1 - \exp(-\Omega(d))$, the only short lattice vectors in L are ones whose integer coefficients $t = (t_1, t_2)$ satisfy $t_1 = cx$ for some $c \in \mathbb{Z}$. First, observe that since $|t_2| \leq 2^{2d}$, the following inequality holds if t is an integer relation:

$$\left| \sum_{i=1}^{d+1} (\lambda_i)_N (t_1)_i \right| \leq 2^{2d} 2^{-N} \, .$$

Consider $\mathcal{T}$ the set of all $t_1 \in \mathbb{Z}^{d+1} \setminus \bigcup_{c \in \mathbb{R}} \{c(x_1, \ldots, x_{d+1})^\top\}$. To prove our result it suffices to prove that

$$\mathbb{P}\left(\bigcup_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \left\{ \left| \sum_{i=1}^{d+1} (\lambda_i)_N (t_1)_i \right| \leq 2^{2d}/2^N \right\} \right) \leq \exp(-\Omega(d))$$

for which, since for any x it holds $|x - (x)_N| \leq 2^{-N}$ and $\|t\|_1 = |(t_1)_1| + \|t'\|_\infty \leq 2^{4d^2}$ for sufficiently large d, it suffices to prove that for large d,

$$\mathbb{P}\left(\bigcup_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \left\{ \left| \sum_{i=1}^{d+1} \lambda_i (t_1)_i \right| \leq 2^{5d^2}/2^N \right\} \right) \leq \exp(-\Omega(d)) \, .$$

Using the polynomial notation of Lemma 6.2 (specifically, Eq.(9)), as well as the fact that by Cramer's rule λ_i are rational functions of the coordinates of z_i satisfying $\lambda_i \det(z_2, \ldots, z_{d+1}) = \det(\ldots, z_{i-1}, -z_1, z_{i+1}, \ldots)$, it suffices to show

$$\mathbb{P}\left(\bigcup_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \{|P_{t_1}(z_1, \ldots, z_{d+1})| \le |\det(z_2, \ldots, z_{d+1})| 2^{5d^2}/2^N\} \right) \le \exp(-\Omega(d)) \, .$$

By Lemma 6.3, with probability $1 - \exp(-\Omega(d))$ there exists some constant $D > 0$ such that $\det(z_2, \ldots, z_{d+1}) \le D 2^{2d^2}$. Hence, it suffices to show

$$\mathbb{P}\left(\bigcup_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \{|P_{t_1}(z_1, \ldots, z_{d+1})| \le D 2^{7d^2}/2^N\} \right) \le \exp(-\Omega(d)) \, .$$

Now since $N = \omega(d^2 \log d)$, it suffices to show, for sufficiently large d,

$$\mathbb{P}\left(\bigcup_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \{|P_{t_1}(z_1, \ldots, z_{d+1})| \le 2^{-\frac{N}{2}}\} \right) \le \exp(-\Omega(d)) \, .$$

By a union bound, it suffices to show

$$\sum_{t_1 \in \mathcal{T} \cap \mathcal{R}_1} \mathbb{P}\left(|P_t(z_1, \ldots, z_{d+1})| \le 2^{-\frac{N}{2}} \right) \le 2^{-\Omega(d)}. \tag{29}$$

Now the number of integer points t_1 with ℓ_∞ norm at most 2^{3d^2} is at most $2^{3d^2(d+1)}$, since there are at most 2^{3d^2} choices per coordinate. Furthermore, using the anticoncentration inequality (12) of Lemma 6.2, we have for any $t_1 \in \mathcal{T}$ that for some universal constant $B > 0$,

$$\mathbb{P}\left(|P_{t_1}(z_1, \ldots, z_{d+1})| \le 2^{-\frac{N}{2}} \right) \le B d 2^{-\frac{N}{2d}} \, .$$

Using the above to upper bound the left hand side of (29), we see that the sum is at most

$$B d 2^{3d^2(d+1)} 2^{-\frac{N}{2d}} = \exp(O(d^3) - \Omega(N/d)) = \exp(-\Omega(d)) \, ,$$

where we used that $N/d = \Omega(d^3 \log d)$. This completes the proof. □

7 Proofs of information-theoretic lower bounds

We now provide the proofs of Theorem 5.1 and Theorem 5.2. As mentioned in Section 5, our lower bounds show that the sample complexity $n = d + 1$ for our LLL-based algorithm is indeed optimal. We first provide the proof of Theorem 5.1, which establishes that even when we have access to the true signs $x_i \in \{-1, 1\}$ for $i = 1, \ldots, n$, we cannot exactly recover the true hidden direction $u \in \mathcal{S}^{d-1}$ with probability larger than $1/2$ if $n \le d - 1$. Then, we prove Theorem 5.2 which shows that for exact recovery of the labels, $n = (1 - o(1))d$ samples are required.

7.1 Proof of Theorem 5.1

Theorem 5.1 (Restated). *Let arbitrary $x \in \{-1,1\}^n$ be fixed and known to the statistician. Moreover, let $u \in \mathcal{S}^{d-1}$ be a uniformly random unit vector. For each $i = 1,\dots,d-1$ let z_i be a sample generated independently from $\mathcal{N}(x_i u, I - uu^\top)$. Then it is information-theoretically impossible to construct an estimate $\hat{u} = \hat{u}(\{(x_i,z_i)\}_{i=1}^n) \in \mathcal{S}^{d-1}$ which satisfies $\hat{u} = u$ with probability larger than $1/2$.*

Proof. We establish the result by proving that the posterior is a uniform distribution on some finite subset of $\mathcal{S}^{d-1}$ of cardinality exactly equal to 2 almost surely. Notice that if we establish this, our impossibility is implied as follows: the optimal estimator in minimizing probability of error for exact recovery is the MAP estimator (see e.g., [SZB21, Lemma H.4]). The MAP estimator outputs the $v \in \mathcal{S}^{d-1}$ with the maximum posterior mass. Since the posterior is a uniform distribution between two points, the probability of the MAP estimator (and therefore any estimator) in recovering exactly u is at most $1/2$.

To this end, we calculate the posterior mass assigned to any arbitrary fixed vector $v \in \mathcal{S}^{d-1}$, given the samples $\{z_i\}_{i=1}^{d-1}$. Let us first complete v to an arbitrary ordered orthonormal basis of $\mathbb{R}^d$, say $v = q_1, q_2, \dots, q_d$. Then, since the labels x_i's are known, if u, the true hidden direction, were equal to v, the samples z_i for $i = 1,\dots,d-1$ would admit the basis representation

$$z_i = \sum_{j=1}^{d} c_{i,j} q_j = x_i v + \sum_{j=2}^{d} c_{i,j} q_j \ ,$$

where $c_{i,j}$ for $i \in [d-1]$, and $j \in [d]\backslash\{1\}$ are i.i.d. standard Gaussian random variables. Hence, given z_i the posterior mass assigned to v is zero if $\langle z_i, v\rangle \neq x_i$ and otherwise, it has mass proportional to

$$\frac{1}{(2\pi)^{(d-1)/2}} \exp\left(-\sum_{j=2}^{d} \langle q_j, z_i\rangle^2/2\right) = \frac{1}{(2\pi)^{(d-1)/2}} \exp(x_i^2 - \|z_i\|_2^2) = \frac{1}{(2\pi)^{(d-1)/2}} \exp(1 - \|z_i\|_2^2) \ .$$

Notice importantly that the computed quantity is constant with respect to the direction v due to the hard constraint $\langle z_i, v\rangle = x_i$ and the fact that $x_i^2 = 1$ for all $i \in [d-1]$. Hence, the posterior mass at $v \in \mathcal{S}^{d-1}$ given the single sample z_i is proportional to $\mathbb{1}[v \in \mathcal{S}^{d-1} \wedge \langle v, z_i\rangle = x_i]$. Since the z_i's are generated independently, the posterior measure given $\{z_i\}_{i=1}^{d-1}$ is proportional to $\mathbb{1}[v \in \mathcal{S}^{d-1} \wedge Zv = x]$, where Z is a matrix with $z_i^\top$'s as its rows. In what follows, let us call then

$$S := \{v \in \mathcal{S}^{d-1} \mid Zv = x\} \ .$$

To upper bound the success probability of any estimator by $1/2$, it suffices to show $|S| = 2$ almost surely. We first prove that $|S| \leq 2$ almost surely. Recall that $Z \in \mathbb{R}^{(d-1)\times d}$ is a matrix with rows $z_i^\top, i = 1,\dots,d-1$. Notice that to prove $|S| \leq 2$ almost surely, it suffices to show that the Kernel(Z) is a one-dimensional linear subspace (i.e., consists of points on a line passing through the origin) almost surely. Indeed, since the true direction $u \in \mathcal{S}^{d-1}$ satisfies $\langle u, z_i\rangle = x_i$ we have

$$S = \mathcal{S}^{d-1} \cap (u + \text{Kernel}(Z)) \ .$$

Any line can intersect the sphere in at most two points. Hence, if Kernel(Z) is one-dimensional almost surely, then $|S| \leq 2$ almost surely. We now show that the Kernel(Z) is one-dimensional almost surely. By invariance of the kernel of Z to the change of column basis, we may assume without loss of generality that $u = e_1$, that is, u is the first standard basis vector, and the remaining

orthonormal basis vectors of $\mathbb{R}^d$ are just the standard basis vectors $e_2, \ldots, e_d$. Under this assumption, we can write for each $i = 1, \ldots, d+1$,

$$z_i = \begin{bmatrix} ax_i \\ w_i \end{bmatrix} ,$$

where the w_i's are i.i.d. samples from $\mathcal{N}(0, I_{d-1})$.

In other words, the first column of $Z \in \mathbb{R}^{(d-1)\times d}$ consists of $ax_1, \ldots, ax_{d-1}$, and the remaining columns consist of coordinates of $w_i \in \mathbb{R}^{d-1}$. Now we proceed by establishing that the last $d-1$ columns of Z in this basis are linearly independent almost surely, which suffices to establish that the kernel is one-dimensional. Let $W \in \mathbb{R}^{(d-1)\times(d-1)}$ be the submatrix of Z consisting of the last $d-1$ columns of Z. Notice that the entries of W are i.i.d. Gaussian. Now using folklore results (e.g., [CT05]), we have that the determinant of an i.i.d. Gaussian matrix is non-zero almost surely. Hence, we conclude that indeed $\det(W) \neq 0$ almost surely. This implies that the column rank of $Z \in \mathbb{R}^{(d-1)\times d}$ is equal to $d-1$ almost surely, and thus Kernel(Z) is one-dimensional.

We now prove that $|S| \geq 2$ almost surely. First notice that by our assumption on the data generating process, the set S contains at least one unit vector, namely u, which is drawn uniformly at random from $\mathcal{S}^{d-1}$. Hence, $|S| \geq 1$. To show that $S \setminus \{u\}$ is non-empty, we claim there exists $y \in \mathbb{R}^d$ such that $y \in$ Kernel(Z) and $\langle y, u\rangle < 0$, almost surely. Suppose not. Then, by the hyperplane separation theorem $u \in \text{span}(z_1, \ldots, z_{d-1})$. That is, there exist scalars $\lambda_1, \ldots, \lambda_{d-1} \in \mathbb{R}$ such that

$$u = \sum_{i=1}^{d-1} \lambda_i z_i \ . \tag{30}$$

Let $u_2, \ldots, u_d$ be an orthonormal basis of the linear space which is perpendicular to u. Then, it holds

$$\sum_{i=1}^{d-1} \lambda_i \langle z_i, u_j\rangle = 0 \ \text{ for all } j = 2 \ldots, d \ . \tag{31}$$

By the sampling process of the z_i's, we have that $\langle z_i, u_j\rangle$ for all $i = 1, \ldots, d-1$ and $j = 2, \ldots, d$ are i.i.d. standard Gaussians $\mathcal{N}(0,1)$. Hence, by foklore results, the $(d-1)\times(d-1)$ matrix R consisting of the (Gaussian) entries $R_{i,j} = \langle z_i, u_{j+1}\rangle$ for $i \in [d-1], j \in [d-1]$ is invertible almost surely (this follows for example because the determinant is a non-zero polynomial of the entries of the matrix and by again appealing to [CT05]). But (31) implies that $R\lambda = 0$ where $\lambda = (\lambda_1, \ldots, \lambda_{d-1})^\top$. Hence, from the almost sure invertibility of R we conclude that necessarily $\lambda = R^{-1}\mathbf{0} = \mathbf{0}$ almost surely. But this condition readily contradicts (30) as we assumed that u is of unit norm. Hence, under almost sure properties of the samples z_i, we have established the existence of the desired $y \in \mathbb{R}^d$ almost surely. Now by employing the fact that $Zu = x$, we derive that for all $t \in \mathbb{R}$ and again for all $i = 1, \ldots, d-1$ it holds that

$$\langle z_i, u + ty\rangle = \langle z_i, u\rangle = x_i. \tag{32}$$

Furthermore, since $\langle u, y\rangle < 0$ we have

$$\inf_{t>0} \|u + ty\|_2 = \sqrt{1 - \langle u, y\rangle^2} < 1 < \sup_{t>0} \|u + ty\|_2 = +\infty. \tag{33}$$

Hence, by continuity of $\|u+ty\|_2$ as a function of $t \in \mathbb{R}$, there exists $t^* > 0$ such that $\|u+t^*y\|_2 = 1$. Combined with (32), this implies $u + t^*y \in S$. □

7.2 Proof of Theorem 5.2

Now we present the proof of Theorem 5.2, which establishes that one cannot information-theoretically recover the labels $\{x_i\}_{i=1}^n$ (up to a global sign flip) when $n = \lfloor \rho d \rfloor$ for any constant $\rho \in (0,1)$. This implies that our sample complexity of $n = d+1$ is optimal up to a $1+o(1)$ factor for label recovery. For the reader's convenience, we restate Theorem 5.2 below.

Theorem 5.2 (Restated). *Let $\rho \in (0,1)$ be a fixed constant and let $n = \lfloor \rho d \rfloor$. Let $x \in \{-1,1\}^n$ be drawn uniformly at random, and $u \in \mathcal{S}^{d-1}$ be a uniformly random unit vector. For each $i = 1, \dots, n$ let $z_i \in \mathbb{R}^d$ be a sample generated independently from $\mathcal{N}(x_i u, I - uu^\top)$. Then no estimator can exactly recover the labels $\{x_i\}_{i=1}^n$ up to a global sign flip from the observations $\{z_i\}_{i=1}^n$ with probability $1 - o(1)$.*

Before proving Theorem 5.2, we present auxiliary lemmas. We first show a key lemma which characterizes the conditional distribution of $Z = (z_1; \dots; z_n)^\top \in \mathbb{R}^{n \times d}$ given $X = x$, where $x \in \{-1,1\}^n$ is any Rademacher vector. We denote this conditional distribution by $f(Z|X = x)$.

Then, we define an "energy" function $F_H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, where $H = ZZ^\top$, and show that it satisfies useful concentration properties when Z is drawn from $f(Z|X = x)$ (see Definition 7.5, and Claims 7.8). The energy function F_H is useful because we can express the conditional density $f(Z' = Z|X = x)$ in terms of F_H (see Remark 7.6). Claim 7.9, which relates the value of $f(Z' = Z|X = x)$ to $f(Z' = Z|X = \tilde{x})$ for *some* $\tilde{x} \in \{-1,1\}^n$ such that $\|\tilde{x} - x\|_0 = 1$, will be crucial for the proof of Theorem 5.2.

Given the observed Z, we denote $Z \odot x := (x_1 z_1; \dots; x_n z_n)^\top \in \mathbb{R}^{n \times d}$, $\mathbf{1}^n = (1, \dots, 1) \in \mathbb{R}^n$, and $t_+ = \max(0, t)$. We have the following characterization of the conditional distribution of Z given the labels $X = x$.

Lemma 7.3 (Conditional density given labels). *The conditional distribution μ_x of Z given label assignment $x \in \{-1,1\}^n$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{nd}$, with a density $f(Z|X = x) := \frac{d\mu_x}{dZ}(Z)$ given by*

$$f(Z|X=x) = \begin{cases} \mathcal{Z}^{-1} \exp\left(-\frac{1}{2}\mathrm{Tr}(H)\right) \det(H)^{-\frac{1}{2}} (1 - x^\top H^{-1} x)_+^{\frac{d-n-2}{2}} & \text{if } \lambda_{min}(H) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{34}$$

where $H = ZZ^\top$ and $\mathcal{Z}$ is the normalization constant which does not depend on x.

The proof of Lemma 7.3 can be found in Section 7.3. This lemma establishes, via the Fisher-Neyman principle, that the Gram matrix H is a sufficient statistic for the label recovery. In particular, it reveals the invariance of the model with respect to orthogonal transformations of the d-dimensional input (since they do not modify its Gram matrix), as already observed by [DDW21, Section 2.2].

This Gram matrix will play a central role in the remainder of the proof. We now compute its conditional distribution:

Claim 7.4 (Distribution of H). *Let $x \in \{-1,1\}^n$ and let $Z \in \mathbb{R}^{n \times d}$ be drawn from the distribution $f(Z|X = x)$ (defined in Eq. (34)). Then, the matrix $H = ZZ^\top$ is distributed as* [3]

$$H \overset{\mathrm{d}}{=} xx^\top + Y \,, \tag{35}$$

where $Y = WW^\top$ for $W \in \mathbb{R}^{n \times (d-1)}$ with $W_{ij} \sim \mathcal{N}(0,1)$ for all $i \in [n], j \in [d-1]$.

[3] For random quantities X and Y, we write $X \overset{\mathrm{d}}{=} Y$ to denote that X and Y have the same distribution.

Proof. Let P_x^H denote the sought probability measure of H over the set of n-by-n positive semi-definite matrices. From the definition of the Gaussian clustering model, we have that

$$P_x^H = \int_{\mathcal{S}^{d-1}} P_{x,u}^H \, \nu(du) \,, \tag{36}$$

where $P_{x,u}^H$ is the distribution of $H = ZZ^T$, where $Z = (z_1; \ldots; z_n)^\top \in \mathbb{R}^{n\times d}$ and z_i's are independent Gaussian vectors of mean $x_i u$ and covariance $I - uu^\top$. Let $\mu_{x,u}$ denote the associated product measure on Z. Fix an arbitrary $u_0 \in \mathcal{S}^{d-1}$, and observe that Z has the same distribution as $Z_0 Q_u$, where Z_0 is drawn from μ_{x,u_0}, and Q_u is an orthogonal matrix which maps u_0 to u. It follows that $P_{x,u}^H$ does not depend on u, since for any orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ $(ZQ)(ZQ)^\top = ZZ^\top$.

Therefore, from Eq.(36) we obtain that $P_x^H = P_{x,e_1}^H$. By expressing Z in the canonical basis, we obtain $z_i \overset{\mathrm{d}}{=} (x_i, w_i)$, with $w_{i,j} \sim \mathcal{N}(0,1)$, and therefore $H_{i,j} = \langle z_i, z_j\rangle = x_i x_j + \langle w_i, w_j\rangle$ for $i \in [n]$ and $j \in [d-1]$. □

By the Sherman-Morrison formula [SM50], we have

$$H^{-1} = Y^{-1} - \frac{1}{1 + x^\top Y^{-1} x} (Y^{-1}x)(Y^{-1}x)^\top \,, \tag{37}$$

where Y follows the Wishart distribution $W_n(I_n, d-1)$. From H^{-1}, we define the following "energy" function F_H.

Definition 7.5 (Energy function)**.** *Given any positive definite $H \in \mathbb{R}^{n\times n}$, we define the energy function $F_H : \{-1,1\}^n \times \{-1,1\}^n \to \mathbb{R}$*

$$F_H(a,b) = a^\top H^{-1} b \,. \tag{38}$$

We abuse notation and write $F_H(a) = F_H(a,a)$.

Remark 7.6 (Conditional density using energy functions)**.** *Using the energy function F_H, we can equivalently write Eq.(34) as*

$$f(Z|X=x) = \mathcal{Z}^{-1} \exp\left(-\|Z\|_F^2\right) \det(ZZ^\top)^{-1/2} \cdot (1 - F_{ZZ^\top}(x))_+^{\frac{d-n-2}{2}} \,. \tag{39}$$

Claim 7.7. *For any $a, b \in \{-1,1\}^n$ and any positive definite $H \in \mathbb{R}^{n\times n}$ expressible as in Eq. (35), we have*

$$F_H(a,b) = a^\top Y^{-1} b - \gamma (a^\top Y^{-1} x)(b^\top Y^{-1} x) \tag{40}$$

where $\gamma = \frac{1}{1+x^\top Y^{-1}x} \in (0,1)$. In particular, if we take $a = b$, we have

$$F_H(a) = a^\top Y^{-1} a - \gamma (a^\top Y^{-1} x)^2$$

which immediately implies $0 \prec H^{-1} \preceq Y^{-1}$, in the positive semidefinite cone order.

Now, thanks to the proportional regime $n/d \to \rho \in (0,1)$ we have enough concentration to control the energy of planted labels, and, crucially, the size of the energy fluctuations for small label perturbations. This is formalised in Claims 7.8 and 7.9 respectively.

Claim 7.8 (Energy of planted labels)**.** *Let $d, n \in \mathbb{N}$ be such that $n = \lfloor \rho d \rfloor$ for some fixed constant $\rho \in (0,1)$. Moreover, let $F_H : \mathbb{R}^n \to \mathbb{R}$ be the energy function, where H is drawn according to Eq. (35) with planted labels x. Then, the following holds with probability $1 - \exp(-\Omega(n))$ over the randomness of Z.*

$$F_H(x) = \frac{x^\top Y^{-1} x}{1 + x^\top Y^{-1} x} \in [c, C] \ ,$$

where $0 < c < C < 1$ are fixed constants.

Proof. By straightforward calculation,

$$F_H(x) = x^\top Y^{-1} x - \frac{(x^\top Y^{-1} x)^2}{1 + x^\top Y^{-1} x} = \frac{x^\top Y^{-1} x}{1 + x^\top Y^{-1} x} \ .$$

Now note by [RV10, Theorem 3.3] that with probability at least $1 - \exp(-\Omega(n))$, the maximum eigenvalue of Y^{-1} is of order $O(1/n)$ with probability $1 - \exp(-\Omega(n))$. Hence, $x^\top Y^{-1} x \le \lambda_{\max}(Y^{-1}) \|x\|_2^2 = O(1)$. Moreover, by [RV10, Proposition 2.4], $\lambda_{\min}(Y^{-1}) = \Omega(1/n)$ with probability at least $1 - \exp(-\Omega(n))$. Therefore, $x^\top Y^{-1} x = \Theta(1)$ with probability $1 - \exp(-\Omega(n))$, and the conclusion follows. □

Claim 7.9 (Energy of 1-Hamming sign flip)**.** *Let $d, n \in \mathbb{N}$ be such that $n = \lfloor \rho d \rfloor$ for some fixed constant $\rho \in (0,1)$. Moreover, let $x \in \{-1,1\}^n$ and let $F_H : \{-1,1\}^n \to \mathbb{R}$ be the energy function, where H is drawn according to Eq. (35) with planted labels x. Then, there exists a constant $C > 0$ such that with probability $1 - \exp(-\Omega(n))$ over the randomness of H, there exists $\tilde{x} \in \{-1,1\}^n$ with $\|\tilde{x} - x\|_0 = 1$ such that*

$$F_H(\tilde{x}) - F_H(x) \le C/n \ . \tag{41}$$

Proof. By simple algebraic manipulation,

$$\begin{aligned}
(\tilde{x} - x)^\top H^{-1} (\tilde{x} - x) &= (\tilde{x} - x)^\top H^{-1} \tilde{x} - (\tilde{x} - x)^\top H^{-1} x \\
&= \tilde{x}^\top H^{-1} \tilde{x} - 2\tilde{x}^\top H^{-1} x + x^\top H^{-1} x \\
&= \tilde{x}^\top H^{-1} \tilde{x} - x^\top H^{-1} x - 2(\tilde{x} - x)^\top H^{-1} x \ .
\end{aligned}$$

Hence,

$$F_H(\tilde{x}) - F_H(x) = (\tilde{x} - x)^\top H^{-1} (\tilde{x} - x) + 2(\tilde{x} - x)^\top H^{-1} x \ .$$

Since $\tilde{x}$ is a neighbor of x with Hamming distance 1, we observe each $\tilde{x}$ corresponds to either e_i (or $-e_i$) in the sense that $\tilde{x} - x = 2x_i e_i$ for some $i \in [n]$. Using this expression, we have

$$F_H(\tilde{x}) - F_H(x) = 4 e_i^\top H^{-1} e_i - 4(x_i e_i)^\top H^{-1} \left(\sum_{j=1}^n x_j e_j \right) = 4F_H(e_i) - 4F_H(e_i, x) \ ,$$

where $F_H(e_i, x)$ is defined in (38).

We now observe that

$$F_H(x) = \left(\sum_{j=1}^n x_j e_j \right)^\top H^{-1} \left(\sum_{j=1}^n x_j e_j \right) = \sum_{j=1}^n F_H(x_j e_j, x) \ .$$

By an averaging argument, there exists at least one $i \in [n]$ such that $F_H(x_i e_i, x) \geq F_H(x)/n$. By Claim 7.8, with probability at least $1 - \exp(-\Omega(n))$, $F_H(x) \in [c_1, c_2]$ for some constants c_1, c_2 satisfying $0 < c_1 < c_2 < 1$. Hence, $F_H(x_i e_i, x) \geq c_1/n$ with probability $1 - \exp(-\Omega(n))$, which in particular implies $F(e_i, x) \geq 0$ with the same probability.

By the fact that $H^{-1} \preceq Y^{-1}$ and [RV10, Theorem 3.3], the maximum eigenvalue of H^{-1} is $O(1/n)$ with probability $1-\exp(-\Omega(n))$. Hence, $F_H(e_i) = O(1/n)$. It follows that $F_H(\tilde{x}) - F_H(x) \leq 4F_H(e_i) - 4F_H(x_i e_i, x) \leq C/n$ for some constant $C > 0$ with probability $1 - \exp(-\Omega(n))$. □

We are now ready to prove Theorem 5.2.

Proof of Theorem 5.2. Let μ be the marginal distribution over $\mathbb{R}^{n\times d} \times \{-1,1\}^n$, and let $f(Z' = Z|X = x)$ denote the conditional density of $Z \in \mathbb{R}^{n\times d}$ given label $x \in \{-1,1\}^n$ from Lemma 7.3. Let $\hat{x} : \mathbb{R}^{n\times d} \to \{-1,1\}^n$ be an arbitrary (deterministic) estimator and let $\mathrm{acc}(\hat{x}) \in [0,1]$ be its accuracy, defined as follows. For any $x \in \{-1,1\}^n$, let $S_{\hat{x},x} = \{Z \in \mathbb{R}^{n\times d} \mid \hat{x}(Z) = x \vee \hat{x}(Z) = -x\}$ denote the set of observations for which $\hat{x}$ is correct. Then, its accuracy (that is the probability of exactly recovering the correct sign pattern up to a global sign flip) is given by

$$\begin{aligned}
\mathrm{acc}(\hat{x}) &= \int \mathbb{1}\left[\hat{x}(Z) = x \vee \hat{x}(Z) = -x\right] d\mu(x, Z) \\
&= \sum_{x\in\{-1,1\}^n} \int_{S_{\hat{x},x}} f(Z|x)\mathbb{P}[X = x]dZ \\
&= \frac{1}{2^n} \sum_{x\in\{-1,1\}^n} \int_{S_{\hat{x},x}} f(Z|x)dZ \ . \qquad (42)
\end{aligned}$$

We show that there exists a constant $\delta > 0$ such that for any non-trivial estimator $\hat{x}$ achieving $\Omega(1)$ accuracy, we can construct a disjoint estimator $\hat{y}$, i.e., $\hat{x}(Z) \neq \pm\hat{y}(Z)$ for all $Z \in \mathbb{R}^{n\times d}$, such that $\mathrm{acc}(\hat{y}) \geq \delta \cdot \mathrm{acc}(\hat{x})$. Since $\hat{x}$ and $\hat{y}$ are disjoint, $\mathrm{acc}(\hat{x}) + \mathrm{acc}(\hat{y}) \leq 1$ (see (46) below). Hence, for any $\hat{x} : \mathbb{R}^{n\times d} \to \{-1,1\}^n$,

$$(1+\delta)\mathrm{acc}(\hat{x}) \leq 1 \Rightarrow \mathrm{acc}(\hat{x}) \leq \frac{1}{1+\delta} \ .$$

We note from Claim 7.9 that for each $x \in \{-1,1\}^n$, with probability $1 - \exp(-\Omega(n))$ over the distribution $f(Z|x)$, there exists a "1-Hamming" sign flip $q(x, Z) \in \{-1,1\}^n$ such that

$$F_H(x \odot q(x, Z)) - F_H(x) = O\left(\frac{1}{n}\right) \ .$$

By Lemma 7.3 and Remark 7.6, we claim that the $O(1/n)$ fluctuation upper bound in F_H implies that given $x \in \{-1,1\}^n$, with probability $1 - \exp(-\Omega(n))$ over $f(Z|x)$, there exists a 1-Hamming neighbor $\tilde{x}$ such that the likelihood ratio $f(Z' = Z|X = x)/f(Z' = Z|X = \tilde{x})$ is upper bounded by a constant $\delta > 0$ independent of the data dimension d.

To see this, let $x \in \{-1,1\}^n$ be any fixed label and let η be the random variable $\eta = 1 - F_H(x)$ induced by $f(Z|x)$. By Claim 7.8, with probability $1 - \exp(-\Omega(n))$, $\eta \in [c_1, c_2]$ for constants $0 < c_1 < c_2 < 1$. By Claim 7.9 with probability $1-\exp(-\Omega(n))$, there exists a 1-Hamming sign flip $q(x, Z)$ such that $F_H(x \odot q(x, Z)) - F_H(x) \leq c_3/n$ for some constant $c_3 > 0$. Hence, in combination

with Lemma 7.3, the following inequality holds with probability $1 - \exp(-\Omega(n))$ over $f(Z|x)$:

$$\begin{aligned}
\frac{f(Z' = Z|X = x)}{f(Z' = Z|X = x \odot q(x, Z))} &\leq \left(\frac{\eta}{\eta - c_3/n}\right)^{c_4 n} \\
&= \left(1 + \frac{c_3/n}{\eta - c_3/n}\right)^{c_4 n} \\
&= \left(1 + \frac{1}{(\eta/c_3)n - 1}\right)^{c_4 n} \\
&\leq \left(1 + \frac{1}{c_5 n}\right)^{c_4 n} \\
&\leq e^{c_4/c_5} \ , \qquad (43)
\end{aligned}$$

where we used the inequality $1 + t \leq e^t$ for all $t \in \mathbb{R}$.

Let us denote $\delta := e^{-c_4/c_5}$. Then, the following holds for any $x \in \{-1, 1\}^n$.

$$\int \mathbb{1}\left[\bigcup_{i=1}^{n} \{f(Z|x \odot e_i) \geq \delta f(Z|x)\}\right] f(Z|x)dZ \geq 1 - \exp(-\Omega(n)) \ . \qquad (44)$$

For each $x \in \{-1, 1\}^n$, let $A_x \subset \mathbb{R}^{n \times d}$ denote the support of the indicator in Eq. (44), so

$$\int_{A_x^c} f(Z|x)dZ \leq \exp(-\Omega(n)) \ . \qquad (45)$$

Moreover, the following holds for any estimator $\hat{x} : \mathbb{R}^{n \times d} \to \{-1, 1\}^n$:

$$\begin{aligned}
\mathrm{acc}(\hat{x}) &= \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \int_{S_{\hat{x},x}} f(Z|x)dZ \\
&= \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \int_{A_x \cap S_{\hat{x},x}} f(Z|x)dZ \\
&\quad + \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \int_{A_x^c \cap S_{\hat{x},x}} f(Z|x)dZ \\
&\leq \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \int_{A_x \cap S_{\hat{x},x}} f(Z|x)dZ + \exp(-\Omega(n)) \ .
\end{aligned}$$

As mentioned previously, given an estimator $\hat{x}$, we define a new estimator $\hat{y}$ that is disjoint from $\hat{x}$. In other words, $\hat{y}(Z) \neq \pm\hat{x}(Z)$ for all $Z \in \mathbb{R}^{n \times d}$. This implies $S_{\hat{x},x} \cap S_{\hat{y},x} = \emptyset$ and therefore

$$\begin{aligned}
\mathrm{acc}(\hat{x}) + \mathrm{acc}(\hat{y}) &= \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \left(\int_{S_{\hat{x},x}} f(Z|x)dZ + \int_{S_{\hat{y},x}} f(Z|x)dZ \right) \\
&\leq \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \int_{\mathbb{R}^{nd}} f(Z|x)dZ \\
&\leq 1 \ . \qquad (46)
\end{aligned}$$

Recall that we defined A_x as the set of Z's for which the following inequality between the conditional densities holds for some Hamming sign flip $q(x, Z) \in \{-1, 1\}^n$:

$$f(Z|x + q(x, Z)) \geq \delta f(Z|x) \ .$$

Recall that $S_{\hat{x},x} = \{Z \in \mathbb{R}^{n\times d} \mid \hat{x}(Z) = x \vee \hat{x}(Z) = -x\}$. We consider the following partition of $S_{\hat{x},x} \cap A_x$:

$$S_{\hat{x},x} \cap A_x = \bigcup_{i=1}^{n} T_i^x \ , \tag{47}$$

where $T_i^x = \{Z \in S_{\hat{x},x} \cap A_x \mid f(Z|x \odot e_i) \geq \delta f(Z|x)\}$. Note that we can make the T_i^x's disjoint by breaking ties between the e_i's arbitrarily.

Hence, Eq. (47) is indeed a partition. It follows that for each $x \in \{-1,1\}^n$,

$$\sum_{i=1}^{n} \int_{T_i^x} f(Z|x \odot e_i) dZ \geq \delta \int_{S_{\hat{x},x}\cap A_x} f(Z|x) dZ \ . \tag{48}$$

Now, summing over all $x \in \{-1,1\}^n$, we have

$$\begin{aligned}\frac{1}{2^n} \sum_{x\in\{-1,1\}^n} \sum_{i=1}^{n} \int_{T_i^x} f(Z|x \odot e_i) dZ &\geq \frac{\delta}{2^n} \sum_{x\in\{-1,1\}^n} \int_{S_{\hat{x},x}\cap A_x} f(Z|x) dZ \\ &\geq \frac{\delta}{2^n} \sum_{x\in\{-1,1\}^n} \int_{S_{\hat{x},x}} f(Z|x) dZ - \delta \exp(-\Omega(n)) \ .\end{aligned}$$

We define our new estimator $\hat{y}$ such that for any $x \in \{-1,1\}^n$ and $i \in [n]$,

$$\hat{y}(Z) = \begin{cases} \hat{x}(Z) \odot e_i & \text{for any } Z \in T_i^x \\ \hat{x}(Z) \odot e_1 & \text{for any } Z \in \mathbb{R}^{n\times d} \setminus \bigcup_{x,i} T_i^x \end{cases} . \tag{49}$$

Now that everything is in place, let α be the accuracy of a given estimator $\hat{x}$, i.e., $\alpha = \frac{1}{2^n}\sum_{x\in\{-1,1\}^n} \int_{S_x^*} f(Z|x) dZ$. Then, our proposed disjoint estimator $\hat{y}$ achieves accuracy at least $\delta(\alpha - \exp(-\Omega(n)))$. The two accuracies must add up to less than 1 since $\hat{x}$ and $\hat{y}$ are disjoint estimators. Therefore, if $\alpha = \Omega(1)$, then for sufficiently large n

$$\alpha + \delta(\alpha - \exp(-\Omega(n))) \leq 1 \Rightarrow \alpha \leq \frac{1 + \delta\exp(-\Omega(n))}{1+\delta} < 1 - \frac{\delta}{2(1+\delta)} = 1 - \Omega(1) \ .$$

Hence, we conclude that $\alpha \leq 1 - \Omega(1)$ necessarily, as claimed. □

7.3 Proof of Lemma 7.3

Lemma 7.3 (Restated). *The conditional distribution μ_x of Z given label assignment $x \in \{-1,1\}^n$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{nd}$, with a density $f(Z|X = x) := \frac{d\mu_x}{dZ}(Z)$ given by*

$$f(Z|X=x) = \begin{cases} \mathcal{Z}^{-1} \exp\left(-\frac{1}{2}\|Z\|_F^2\right) \det(H)^{-\frac{1}{2}} (1 - x^\top H^{-1} x)_+^{\frac{d-n-2}{2}} & \text{if } \lambda_{min}(H) > 0 \\ 0 & \text{otherwise} \end{cases} ,$$

where $H = ZZ^\top$ and $\mathcal{Z}$ is the normalization constant which does not depend on x.

Proof. Let us start by computing the joint probability distribution of x and Z by marginalizing over $u \in \mathcal{S}^{d-1}$. By definition, the probability distribution of $Z|x,u$, which we denote by $\mu_{x,u}$, is Gaussian with mean $(x_1 u, \ldots, x_n u) \in \mathbb{R}^{nd}$ and covariance $(I - uu^\top)^{\otimes n}$, and u is uniformly distributed on $\mathcal{S}^{d-1}$, independent of x. Hence, the corresponding (joint) measure of a rectangle $A \times B \times C \subseteq \mathbb{R}^{nd} \times \{-1,1\}^n \times \mathcal{S}^{d-1}$ is given by

$$\mathbb{P}\left\{Z \in A, x \in B, u \in C\right\} = 2^{-n} \sum_{x \in B} \int_C \mu_{x,u}(A) d\nu(u) \ , \tag{50}$$

where ν is the Haar measure on $\mathcal{S}^{d-1}$.

Now we fix an arbitrary label assignment $x' \in \{-1,1\}^n$, and consider the measure $\mu_{x'}$ on $\mathbb{R}^{nd}$ induced by marginalizing $u \in \mathcal{S}^{d-1}$. That is, for any (Lebesgue) measurable $A \subseteq \mathbb{R}^{nd}$,

$$\mu_{x'}(A) := \int_{\mathcal{S}^{d-1}} \mu_{x',u}(A) d\nu(u) \ . \tag{51}$$

Let $T_x : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ be defined as $T_x(Z) = Z \odot x$. We verify that $\mu_{x'} = T_{x'}^{\#} \mu_{\mathbf{1}^n}$, where $\mathbf{1}^n$ denotes the all-ones vector and $T^{\#}\mu$ denotes the pushforward measure of μ under the mapping T. Hence, it suffices to consider just $\mu_{\mathbf{1}^n}$. Observe that $\mu_{\mathbf{1}^n,u}$, the distribution of $Z|\mathbf{1}^n, u$, is Gaussian with mean $(u, u, \ldots, u)$ and same covariance $(I - uu^\top)^{\otimes n}$. Thus, $\mu_{\mathbf{1}^n,u}$ is a product measure of the form $\mu_{\mathbf{1}^n,u} = (\mu_u)^{\otimes n}$, where μ_u is the $(d-1)$-dimensional isotropic Gaussian measure supported on the hyperplane $S_u = \{z \in \mathbb{R}^d \mid \langle z, u \rangle = 1\}$.

Let $0 < \sigma < 1/\sqrt{2}$ and consider for each $u \in \mathcal{S}^{d-1}$ the mollified measure $\mu_{\mathbf{1}^n,u,\sigma} = (\mu_{u,\sigma})^{\otimes n}$, where $\mu_{u,\sigma}$ is now the Gaussian measure with mean u and covariance $\Sigma = \sigma^2 uu^T + (I_d - uu^\top)$. We verify immediately that $\Sigma^{-1} = I_d + (\sigma^{-2} - 1)uu^\top$. The density $f_{\mathbf{1}^n,u,\sigma}$ of $\mu_{\mathbf{1}^n,u,\sigma}$ with respect to the Lebesgue measure on $\mathbb{R}^{nd}$, which we denote by dZ, is thus

$$\begin{aligned} f_{\mathbf{1}^n,u,\sigma}(Z) := \frac{d\mu_{\mathbf{1}^n,u,\sigma}}{dZ} &= (2\pi)^{-nd/2} \det(\Sigma)^{-n/2} \exp\left(\sum_{i=1}^n -\frac{1}{2}(z_i - u)^\top (I + (\sigma^{-2} - 1)uu^\top)(z_i - u)\right) \\ &= (2\pi)^{-nd/2} \sigma^{-n} \exp\left(-\frac{1}{2}\sum_{i=1}^n \left(\|z_i\|_2^2 + (\sigma^{-2} - 1)\langle z_i z_i^\top, uu^\top \rangle - 2\sigma^{-2}\langle z_i, u \rangle + \sigma^{-2}\right)\right) \\ &= \mathcal{Z}^{-1} \sigma^{-n} e^{-n/(2\sigma^2)} \exp\left(-\frac{1}{2}\left(\|Z\|_F^2 + (\sigma^{-2} - 1)u^\top Z^\top Z u - 2\sigma^{-2}\langle Z^\top \mathbf{1}^n, u \rangle\right)\right) \ , \end{aligned} \tag{52}$$

where $\mathcal{Z} = (2\pi)^{nd/2}$ is the partition function, which does not depend on σ, u, or Z.

The resulting smoothed mixture $\mu_{\mathbf{1}^n,\sigma}$ is thus given by $\mu_{\mathbf{1}^n,\sigma}(A) = \int_{\mathcal{S}^{d-1}} \mu_{\mathbf{1}^n,u,\sigma}(A) \nu(du)$ for any measurable $A \subseteq \mathbb{R}^{nd}$. Since $\mu_{\mathbf{1}^n,u,\sigma}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{nd}$ for any $u \in \mathcal{S}^{d-1}$, the marginalized measure $\mu_{\mathbf{1}^n,\sigma}$ is also absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{nd}$ with density

$$\begin{aligned} f_{\mathbf{1}^n,\sigma}(Z) := \frac{d\mu_{\mathbf{1}^n,\sigma}}{dZ}(Z) &= \int_{\mathcal{S}^{d-1}} f_{\mathbf{1}^n,u,\sigma}(Z) d\nu(u) \\ &= \mathcal{Z}^{-1} \exp\left(-\frac{1}{2}\|Z\|_F^2\right) \sigma^{-n} \int_{\mathcal{S}^{d-1}} \exp\left(-\sigma^{-2} F(u)\right) d\nu(u) \ , \end{aligned} \tag{53}$$

where we define $F : \mathbb{R}^d \to \mathbb{R}$ to be

$$F(u) = \frac{1}{2}(1 - \sigma^2) u^\top Z^\top Z u - \langle Z^\top \mathbf{1}^n, u \rangle + \frac{n}{2} \ . \tag{54}$$

Since $Z^\top Z$ has rank $n < d$, we can marginalize over the remaining $d - n$ variables. Indeed, let $Z = V\Lambda U^\top$ be the singular value decomposition of Z, where $U \in \mathbb{R}^{d\times n}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$. Note that $Z^\top Z = U\Lambda^2 U^\top$. Now consider the projection $v = U^\top u \in \mathcal{B}_n$, where $\mathcal{B}_n$ is the n-dimensional ℓ_2 unit ball, and $r = U^\top Z^\top \mathbf{1}^n = \Lambda V^\top \mathbf{1}^n \in \mathbb{R}^n$. Then, $F(u) = E(v)$, where $E : \mathbb{R}^n \to \mathbb{R}$ is given by

$$E(v) = \frac{1}{2}(1-\sigma^2)v^\top \Lambda^2 v - \langle r, v\rangle + \frac{n}{2}\,. \tag{55}$$

Since u is uniformly distributed in $\mathcal{S}^{d-1}$, the joint distribution of $v = U^\top u$ is spherically symmetric, and the squared radius is distributed according to $\mathrm{Beta}(n/2, (d-n)/2)$ [DF87, Remark 2.10]. We refer to the proof of [NR03, Lemma 4] for a detailed derivation of this fact. Hence,

$$\begin{aligned} f_{\mathbf{1}^n,\sigma}(Z) &= \mathcal{Z}^{-1}\exp\left(-\frac{1}{2}\|Z\|_F^2\right)\sigma^{-n}\int_{\mathcal{B}_n}\exp(-\sigma^{-2}E(v))(1-\|v\|_2^2)^{\frac{d-n-2}{2}}dv \\ &= \mathcal{Z}^{-1}\exp\left(-\frac{1}{2}\|Z\|_F^2\right)\sigma^{-n}\int_{\mathcal{B}_n}\exp(-\sigma^{-2}E_1(v)+E_2(v))(1-\|v\|_2^2)^{\frac{d-n-2}{2}}dv\,, \end{aligned} \tag{56}$$

where we E_1 and E_2 comes from the following decomposition of E.

$$E_1(v) = \frac{1}{2}v^\top\Lambda^2 v - \langle r, v\rangle + \frac{n}{2},\ E_2(v) = \frac{1}{2}v^\top\Lambda^2 v\,. \tag{57}$$

We now apply Laplace's approximation method [Bre06, Chapter 5] to estimate the integral

$$\sigma^{-n}\int_{\mathcal{B}_n}\exp\left(-\sigma^{-2}E_1(v)+E_2(v)\right)\right)(1-\|v\|_2^2)^{\frac{d-n-2}{2}}dv\,. \tag{58}$$

Without loss of generality, we assume that Z is such that $\lambda_{\min}(ZZ^\top) = \lambda_n^2 > 0$ since the event $\lambda_{\min}(ZZ^\top) = 0$ has zero Lebesgue measure. Since E_1 is quadratic and $\Lambda^2 \succ 0$, E_1 is strongly convex. Let v^* be the (unique) global minimum of E_1, which is given by $v^* = \Lambda^{-2}r = \Lambda^{-1}V^\top\mathbf{1}^n$.

$$E_1(v^*) = -\frac{1}{2}r^\top\Lambda^{-2}r + \frac{n}{2} = -\frac{1}{2}(\mathbf{1}^n)^\top VV^\top(\mathbf{1}^n) + \frac{n}{2} = -\frac{1}{2}\|\mathbf{1}^n\|_2^2 + \frac{n}{2} = 0\,.$$

Furthermore,

$$E_2(v^*) = \frac{1}{2}r^\top\Lambda^{-2}r = \frac{1}{2}(\mathbf{1}^n)^\top VV^\top(\mathbf{1}^n) = \frac{1}{2}\|\mathbf{1}^n\|_2^2 = \frac{n}{2}\,.$$

We now distinguish two cases, depending whether v^* lies in the interior of $\mathcal{B}_n$ or not. If $v^* \in (\mathcal{B}_n)^\circ$, then by Laplace's approximation we have, for all Z such that $\lambda_{\min}(ZZ^\top) > 0$,

$$\sigma^{-n}\int_{\mathcal{B}_n}\exp\left(-\sigma^{-2}E_1(v)+E_2(v)\right)\right)(1-\|v\|_2^2)^{\frac{d-n-2}{2}}dv \tag{59}$$

$$\begin{aligned} &\overset{\sigma\to 0}{\to} (2\pi)^{n/2}\,\frac{\exp\left(-\sigma^{-2}E_1(v^*)+E_2(v^*)\right)\right)(1-\|v^*\|_2^2)^{\frac{d-n-2}{2}}}{|\nabla^2 E_1(v^*)|^{1/2}} \\ &= (2\pi)^{n/2}\,\frac{\exp\left(-\sigma^{-2}E_1(v^*)+E_2(v^*)\right)\right)(1-\|v^*\|_2^2)^{\frac{d-n-2}{2}}}{\sqrt{\det(ZZ^\top)}} \\ &= (2\pi)^{n/2}\,\frac{e^{n/2}(1-\|v^*\|_2^2)^{\frac{d-n-2}{2}}}{\sqrt{\det(ZZ^\top)}}\,. \end{aligned} \tag{60}$$

Let us now show that for any $Z \in \mathbb{R}^{n\times d}$ such that $\lambda_{\min}(ZZ^\top) > 0$ and $u^* \notin (\mathcal{B}_n)^\circ$, then

$$\sigma^{-n} \int_{\mathcal{B}_n} \exp(-\sigma^{-2} E_1(v)) \exp(E_2(v)) (1 - \|v\|_2^2)^{\frac{d-n-2}{2}} dv \to 0 \ , \ \text{as } \sigma \to 0 \ . \tag{61}$$

Indeed, notice that Eq.(58) can be equivalently written as

$$\sigma^{-n} \int_{\mathbb{R}^n} \exp\left(-\sigma^{-2} E_1(v) + E_2(v)\right) (1 - \|v\|_2^2)_+^{\frac{d-n-2}{2}} dv \ ,$$

where we denote $t_+ = \max(0, t)$. Observing that $(1 - \|v\|_2^2)_+^{\frac{d-n-2}{2}} = 0$ whenever $v^* \notin (\mathcal{B}_n)^\circ$, we conclude that the leading order term in the Laplace approximation vanishes, thereby proving (61).

We have just shown that, as $\sigma \to 0$, the density $f_{\mathbf{1}^n,\sigma}$ of $\mu_{\mathbf{1}^n,\sigma}$ admits a pointwise limit almost everywhere, i.e., $f_{\mathbf{1}^n,\sigma} \overset{\text{a.e.}}{\to} f_{\mathbf{1}^n}$ where the limit $f_{\mathbf{1}^n}$ is given explicitly by

$$f_{\mathbf{1}^n}(Z) = \mathcal{Z}^{-1} \exp\left(-\frac{1}{2}\|Z\|_F^2\right) \frac{\left(1 - (\mathbf{1}^n)^\top (ZZ^\top)^{-1} \mathbf{1}^n\right)_+^{\frac{d-n-2}{2}}}{\sqrt{\det(ZZ^\top)}} \ , \ \text{for } Z \ \text{ s.t. } \lambda_{\min}(ZZ^\top) > 0 \ . \tag{62}$$

Let us now show that the sequence $f_{\mathbf{1}^n,\sigma}$ is dominated by some function g in $L^1(\mathbb{R}^{nd})$, the set of all Lebesgue-integrable functions on $\mathbb{R}^{nd}$. To this end, observe first that

$$\sigma^{-n} \int_{\mathcal{B}_n} \exp\left(-\sigma^{-2} E_1(v) + E_2(v)\right)\right) (1 - \|v\|_2^2)^{\frac{d-n-2}{2}} dv \le \sigma^{-n} \int_{\mathbb{R}^n} \exp\left(-\sigma^{-2} E(v)\right)\right) dv \ . \tag{63}$$

Since $E(v)$ is quadratic in v with a symmetric positive-definite Hessian, the RHS is a Gaussian integral, which we can evaluate exactly. From the identity

$$\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} v^\top H v + \langle \beta, v\rangle + c\right) dv = (2\pi)^{n/2} (\det H)^{-1/2} \exp\left(c + \frac{1}{2}\beta^\top H^{-1} \beta\right) \ ,$$

and the definition of $E(v)$ in Eq.(55), we obtain

$$\begin{aligned} \sigma^{-n} \int_{\mathbb{R}^n} \exp\left(-\sigma^{-2} E(v)\right)\right) dv &= \sigma^{-n} \left(\frac{2\pi}{\sigma^{-2} - 1}\right)^{n/2} (\det \Lambda)^{-1} \exp\left(-\frac{1}{2\sigma^2}\left(n - \frac{1}{1-\sigma^2} r^\top \Lambda^{-2} r\right)\right) \\ &= \left(\frac{2\pi}{1-\sigma^2}\right)^{n/2} (\det \Lambda)^{-1} \exp\left(\frac{n}{2(1-\sigma^2)}\right) \ , \end{aligned} \tag{64}$$

by recalling that $r^\top \Lambda^{-2} r = n$. Thus, for any $\sigma < 1/\sqrt{2}$, from Eq.(63) and Eq.(64), we have

$$\sigma^{-n} \int_{\mathcal{B}_n} \exp\left(-\sigma^{-2} E_1(v) + E_2(v)\right)\right) (1 - \|v\|_2^2)^{\frac{d-n-2}{2}} dv \le (4\pi)^{n/2} e^n (\det \Lambda)^{-1} \ . \tag{65}$$

Hence, the following holds for any $\sigma < 1/\sqrt{2}$.

$$f_{\mathbf{1}^n,\sigma}(Z) \le C_n \det(ZZ^\top)^{-1/2} \exp\left(-\frac{1}{2}\|Z\|_F^2\right) := C_n g(Z) \ , \tag{66}$$

where C_n is a constant that depends only on n. We now show that $g \in L^1(\mathbb{R}^{nd})$.

Indeed, let $Z = LU^\top$ be the LQ decomposition of Z, where L is a lower triangular $n \times n$ matrix and $U \in \mathbb{R}^{d\times n}$ is orthogonal, belonging to the Stiefel manifold $\mathcal{V}(n, d)$. Furthermore, to ensure

uniqueness of the LQ decomposition, we assume that the diagonal entries of L are positive. Denote by T_n the space of such triangular matrices. By [Mui09, Theorem 2.1.13], the differential volume element dZ is expressed in the LQ decomposition as

$$dZ = \left| \prod_{i=1}^{n} L_{i,i}^{d-i} \right| dL dU \ . \tag{67}$$

As a result, we have

$$\begin{aligned} \int g(Z) dZ &\leq \int_{\mathbb{R}^{n\times d}} \exp\left(-\frac{1}{2}\|Z\|_F^2\right) \det(ZZ^\top)^{-1/2} dZ \\ &= \int_{T_n} \int_{\mathcal{V}(n,d)} \exp\left(-\frac{1}{2}\|L\|_F^2\right) \det(L)^{-1} \left| \prod_{i=1}^{n} L_{i,i}^{d-i} \right| dL dU \\ &= \int_{T_n} \int_{\mathcal{V}(n,d)} \exp\left(-\frac{1}{2}\|L\|_F^2\right) \det(L)^{d-n-1} \left| \prod_{i=1}^{n} L_{i,i}^{n-i} \right| dL dU \\ &\leq \mu(\mathcal{V}(n,d)) \int_{T_n} \exp\left(-\frac{1}{2}\|L\|_F^2\right) \|L\|_F^{nd} dL \\ &< \infty \ , \end{aligned} \tag{68}$$

where we used $\det(ZZ^\top)^{1/2} = |\prod_{i=1}^n L_{ii}|$, the fact that $\mathcal{V}(n,d)$ is a compact manifold with finite Haar measure, and that polynomial moments of the Gaussian distribution exist for any fixed order.

We therefore obtain from Eq.(68) that the sequence $f_{\mathbf{1}^n,\sigma}$ is dominated and converges pointwise a.e. to $f_{\mathbf{1}^n}$. As a result, by the dominated convergence theorem [EG18, Theorem 1.19], we obtain that $f_{\mathbf{1}^n}$ is integrable, with $\int f_{\mathbf{1}^n}(Z) dZ = 1$, and

$$\lim_{\sigma\to 0} \int |f_{\mathbf{1}^n,\sigma} - f_{\mathbf{1}^n}| \, dZ = 0 \ . \tag{69}$$

Let $\tilde{\mu}$ be the measure induced by the density $f_{\mathbf{1}^n}$. That is, $\tilde{\mu}(A) = \int_A f_{\mathbf{1}^n}(Z) dZ$, where $f_{\mathbf{1}^n}(Z) = 0$ for $Z \in \mathbb{R}^{nd}$ such that $\lambda_{\min}(Z^\top Z) = 0$. As a consequence of Eq.(69), for any measurable set $A \subseteq \mathbb{R}^{nd}$,

$$\begin{aligned} |\tilde{\mu}(A) - \mu_{\mathbf{1}^n,\sigma}(A)| &= \left| \int_A f_{\mathbf{1}^n}(Z) - f_{\mathbf{1}^n,\sigma}(Z) dZ \right| \\ &\leq \int_{\mathbb{R}^{nd}} |f_{\mathbf{1}^n}(Z) - f_{\mathbf{1}^n,\sigma}(Z)| dZ \ \to 0 \ , \quad \text{as } \sigma \to 0 \end{aligned} \tag{70}$$

which shows that $\mu_{\mathbf{1}^n,\sigma}$ weakly converges to $\tilde{\mu}$.

Finally, we argue that the measure $\mu_{\mathbf{1}^n}$ (defined in Eq.(51)) necessarily admits the desired density with respect to the Lebesgue measure on $\mathbb{R}^{nd}$. Since for each $u \in \mathcal{S}^{d-1}$, the distributions $\mu_{\mathbf{1}^n,u}$ and $\mu_{\mathbf{1}^n,u,\sigma}$ share the same mean, and have covariances $(\sigma^2 uu^T + (I - uu^\top))^{\otimes n}$ and $(I - uu^\top)^{\otimes n}$ that commute, it follows from [OP82, Theorem 4] that their 2-Wasserstein distance satisfies

$$W_2(\mu_{\mathbf{1}^n,u}, \mu_{\mathbf{1}^n,u,\sigma}) = \left\| \left((\sigma^2 uu^T + (I - uu^\top))^{\otimes n}\right)^{1/2} - \left((I - uu^\top)^{\otimes n}\right)^{1/2} \right\|_F^2 = O(\sigma^2) \ .$$

It follows that

$$W_2(\mu_{\mathbf{1}^n}, \mu_{\mathbf{1}^n,\sigma}) \leq \int_{\mathcal{S}^{d-1}} W_2(\mu_{\mathbf{1}^n,u}, \mu_{\mathbf{1}^n,u,\sigma}) d\nu(u) = O(\sigma^2) \ ,$$

and therefore that as $\sigma \to 0$, the smoothed measure $\mu_{\mathbf{1}^n,\sigma}$ converges weakly to $\mu_{\mathbf{1}^n}$, since W_2 metrizes weak convergence on Euclidean spaces [Vil09, Theorem 6.9]. Since $\mu_{\mathbf{1}^n,\sigma}$ weakly converges to both $\mu_{\mathbf{1}^n}$ and $\tilde{\mu}$, we conclude that $\mu_{\mathbf{1}^n} = \tilde{\mu}$, and the proof is complete. □

Acknowledgments

The authors would like to thank Jonathan Niles-Weed, Oded Regev, Kaizheng Wang, Tselil Schramm, Ilias Diakonikolas, Elchanan Mossel and Nike Sun for helpful discussions during different stages of this project.

IZ is supported by the Simons-NSF grant DMS-2031883 on the Theoretical Foundations of Deep Learning and the Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826. MS and JB are partially supported by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF-1845360, and DMS MoDL Scale 2134216. ASW is supported by a Simons-Berkeley Research Fellowship. Part of this work was done while IZ and ASW were visiting the Simons Institute for the Theory of Computing during Fall 2021.

References

[AGJ20] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *The Annals of Probability*, 48(4):2052–2087, 2020.

[AHSS17] Alexandr Andoni, Daniel Hsu, Kevin Shi, and Xiaorui Sun. Correspondence retrieval. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 105–126. PMLR, 2017.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

[BBH18] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166, 2018.

[BDJ+20] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. *arXiv preprint arXiv:2012.02119*, 2020.

[Bea14] Derek Merrill Bean. *Non-Gaussian component analysis*. University of California, Berkeley, 2014.

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

[BKS+06] Gilles Blanchard, Motoaki Kawanabe, Masashi Sugiyama, Vladimir Spokoiny, Klaus-Robert Müller, and Sam Roweis. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(2), 2006.

[BKS14] Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40, 2014.

[BKW20] Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained PCA problems. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, page 78. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[BLPR19] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840. PMLR, 2019.

[BR13] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *arXiv preprint arXiv:1304.0828*, 2013.

[Bre06] Karl W Breitung. *Asymptotic approximations for probability integrals*. Springer, 2006.

[BRST21] Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous LWE. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

[BS16] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*, 2016.

[BV08] S Charles Brubaker and Santosh S Vempala. Isotropic PCA and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

[CMZ19] T Tony Cai, Jing Ma, and Linjun Zhang. Chime: Clustering of high-dimensional gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

[CT05] Richard Caron and Tim Traynor. The zero set of a polynomial. *WSMR Report*, pages 05–02, 2005.

[CW01] A. Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8:233–248, 2001.

[DDW21] Damek Davis, Mateo Diaz, and Kaizheng Wang. Clustering a mixture of gaussians with unknown covariance. *arXiv preprint arXiv:2110.01602*, 2021.

[DF87] Persi Diaconis and David Freedman. A dozen de finetti-style results in search of a theory. In *Annales de l'IHP Probabilités et statistiques*, volume 23, pages 397–423, 1987.

[DH14] Laurent Demanet and Paul Hand. Scaling law for recovering the sparsest element in a subspace. *Information and Inference: A Journal of the IMA*, 3(4):295–309, 2014.

[DJNS13] Elmar Diederichs, Anatoli Juditsky, Arkadi Nemirovski, and Vladimir Spokoiny. Sparse non gaussian component analysis by semidefinite programming. *Machine learning*, 91(2):211–238, 2013.

[DJSS10] Elmar Diederichs, Anatoli Juditsky, Vladimir Spokoiny, and Christof Schutte. Sparse non-gaussian component analysis. *IEEE Transactions on Information Theory*, 56(6):3033–3047, 2010.

[DK20] Ilias Diakonikolas and Daniel M Kane. Hardness of learning halfspaces with massart noise. *arXiv preprint arXiv:2012.09720*, 2020.

[DK21] Ilias Diakonikolas and Daniel M. Kane. Non-gaussian component analysis via lattice basis reduction. *arXiv preprint arXiv:2004.08454*, 2021.

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

[DKMZ11] 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.

[DKP+21] Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, Thanasis Pittas, and Alistair Stewart. Statistical query lower bounds for list-decodable linear regression. *arXiv preprint arXiv:2106.09689*, 2021.

[DKPZ21] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals. *arXiv preprint arXiv:2102.04401*, 2021.

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[DKZ20] Ilias Diakonikolas, Daniel M Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *arXiv preprint arXiv:2006.16200*, 2020.

[DMM09] 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.

[EG18] Lawrence C Evans and Ronald F Garzepy. *Measure theory and fine properties of functions*. Routledge, 2018.

[FGR+17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):1–37, 2017.

[FKP19] Noah Fleming, Pravesh Kothari, and Toniann Pitassi. *Semialgebraic proofs and efficient algorithm design*. now the essence of knowledge, 2019.

[FPB17] Nicolas Flammarion, Balamurugan Palaniappan, and Francis Bach. Robust discriminative clustering with sparse regularizers. *The Journal of Machine Learning Research*, 18(1):2764–2813, 2017.

[Fri86] Alan M. Frieze. On the Lagarias-Odlyzko algorithm for the subset sum problem. *SIAM J. Comput.*, 15:536–539, 1986.

[Fri89] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[GGJ+20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

[GGK20] Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. *arXiv preprint arXiv:2006.15812*, 2020.

[GJJ+20] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for Sherrington-Kirkpatrick via planted affine planes. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–965. IEEE, 2020.

[GKZ21] David Gamarnik, Eren C. Kızıldağ, and Ilias Zadik. Inference in high-dimensional linear regression via lattice basis reduction and integer relation detection. *IEEE Transactions on Information Theory*, pages 1–1, 2021.

[Gri01] Dima Grigoriev. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theoretical Computer Science*, 259(1-2):613–622, 2001.

[GS17] David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *The Annals of Probability*, pages 2353–2376, 2017.

[GS19] Navin Goyal and Abhishek Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–851, 2019.

[GV19] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[GZ17] David Gamarnik and Ilias Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transition. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 948–953. PMLR, 07–10 Jul 2017.

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

[Hop18] Samuel Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

[HS17] Samuel B Hopkins and David Steurer. Bayesian estimation from few samples: community detection and related problems. *arXiv preprint arXiv:1710.00264*, 2017.

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

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

[HWX15] Bruce Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory*, pages 899–928, 2015.

[Jer92] Mark Jerrum. Large cliques elude the Metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[KB21] Dmitriy Kunisky and Afonso S Bandeira. A tight degree 4 sum-of-squares lower bound for the sherrington–kirkpatrick hamiltonian. *Mathematical Programming*, 190(1):721–759, 2021.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, November 1998.

[KMOW17] 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, 2017.

[KT06] Motoaki Kawanabe and Fabian J Theis. Estimating non-gaussian subspaces by characteristic functions. In *International Conference on Independent Component Analysis and Signal Separation*, pages 157–164. Springer, 2006.

[Kun20] Dmitriy Kunisky. Positivity-preserving extensions of sum-of-squares pseudomoments over the hypercube. *arXiv preprint arXiv:2009.07269*, 2020.

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

[Lag84] Jeffrey C Lagarias. Knapsack public key cryptosystems and diophantine approximation. In *Advances in cryptology*, pages 3–23. Springer, 1984.

[Las01] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[LLL82] Arjen Klaas Lenstra, Hendrik Willem Lenstra, and László Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534, 1982.

[LLL+20] Xiaodong Li, Yang Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179(1):295–341, 2020.

[LO85] J. C. Lagarias and A. M. Odlyzko. Solving low-density subset sum problems. *J. ACM*, 32(1):229–246, January 1985.

[Lov86] László Lovász. *An algorithmic theory of numbers, graphs and convexity.* SIAM, 1986.

[LZ16] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[MLKZ20] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions, 2020.

[MNV16] Raghu Meka, Oanh Nguyen, and Van Vu. Anti-concentration for polynomials of independent random variables. *Theory of Computing*, 12(11):1–17, 2016.

[MR09] Daniele Micciancio and Oded Regev. *Lattice-based Cryptography*, pages 147–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[MR14] Andrea Montanari and Emile Richard. A statistical model for tensor PCA. *arXiv preprint arXiv:1411.1076*, 2014.

[MRX20] Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–853, 2020.

[MS16] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827, 2016.

[Mui09] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, page 93–102, 2010.

[MVW17] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

[MW15] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

[MW21] Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

[Nda18] Mohamed Ndaoud. Sharp optimal recovery in the two-component gaussian mixture model. *arXiv preprint arXiv:1812.08078*, 2018.

[NOTV17] Klaus Nordhausen, Hannu Oja, David E Tyler, and Joni Virta. Asymptotic and bootstrap tests for the dimension of the non-gaussian subspace. *IEEE Signal Processing Letters*, 24(6):887–891, 2017.

[NR03] Assaf Naor and Dan Romik. Projecting the surface measure of the sphere of ℓ_p^n. In *Annales de l'IHP Probabilités et statistiques*, volume 39, pages 241–261, 2003.

[NWR19] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

[OP82] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[Par00] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. California Institute of Technology, 2000.

[QSW16] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *IEEE Transactions on Information Theory*, 62(10):5855–5880, 2016.

[QZL+20] Qing Qu, Zhihui Zhu, Xiao Li, Manolis C Tsakiris, John Wright, and René Vidal. Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. *arXiv preprint arXiv:2001.06970*, 2020.

[Roy17] Martin Royer. Adaptive clustering through semidefinite programming. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1793–1801, 2017.

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3389–3423. World Scientific, 2018.

[RV10] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*, pages 1576–1602. World Scientific, 2010.

[Sch08] 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.

[SKBM08] Masashi Sugiyama, Motoaki Kawanabe, Gilles Blanchard, and Klaus-Robert Muller. Approximating the best linear unbiased estimator of non-gaussian signals with gaussian noise. *IEICE transactions on information and systems*, 91(5):1577–1580, 2008.

[SM50] Jack Sherman and Winifred J. Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124 – 127, 1950.

[STS16] Hiroaki Sasaki, Voot Tangkaratt, and Masashi Sugiyama. Sufficient dimension reduction via direct estimation of the gradients of logarithmic conditional densities. In *Asian Conference on Machine Learning*, pages 33–48. PMLR, 2016.

[SZB21] Min Jae Song, Ilias Zadik, and Joan Bruna. On the cryptographic hardness of learning single periodic neurons. *arXiv preprint arXiv:2106.10744*, 2021.

[TV18] Yan Shuo Tan and Roman Vershynin. Polynomial time and sample complexity for non-gaussian component analysis: Spectral methods. In *Conference On Learning Theory*, pages 498–534. PMLR, 2018.

[Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[VX11] Santosh S Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order PCA. *arXiv preprint arXiv:1108.3329*, 2011.

[ZG18] Ilias Zadik and David Gamarnik. High dimensional linear regression using lattice basis reduction. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.