

High-Accuracy List-Decodable Mean Estimation

Ziyun Chen*

Spencer Compton†

Daniel Kane‡

Jerry Li§

Abstract

In list-decodable learning, we are given a set of data points such that an α -fraction of these points come from a “nice” distribution D , for some small $\alpha \ll 1$, and the goal is to output a short list of candidate solutions, such that at least one element of this list recovers some non-trivial information about D . By now, there is a large body of work on this topic; however, while many algorithms can achieve optimal list size in terms of α , all known algorithms must incur error which decays, in some cases quite poorly, with $1/\alpha$. In this paper, we ask if this is inherent: is it possible to trade off list size with accuracy in list-decodable learning? More formally, given $\varepsilon > 0$, can we output a slightly larger list in terms of α and ε , but so that one element of this list has error at most ε with the ground truth? We call this problem *high-accuracy list-decodable learning*.

Our main result is that non-trivial high-accuracy guarantees, both information-theoretically and algorithmically, are possible for the canonical setting of list-decodable mean estimation of identity-covariance Gaussians. Specifically, we demonstrate that there exists a list of candidate means of size at most $L = \exp\left(O\left(\frac{\log^2 1/\alpha}{\varepsilon^2}\right)\right)$ so that one of the elements of this list has ℓ_2 distance at most ε to the true mean. We also design an algorithm that outputs such a list with runtime and sample complexity $n = d^{O(\log L)} + \exp(\tilde{O}(\log L))$. In particular, our results demonstrate that in the natural regime where α and ε are both small constants, it is possible to achieve error ≤ 0.01 in fully-polynomial time, where all prior work suffered error which was much larger than 1. We do so by demonstrating a completely novel proof of identifiability, as well as a new algorithmic way of leveraging this proof without the sum-of-squares hierarchy, which may be of independent technical interest.

1 Introduction

Learning in the presence of outliers is a central challenge in statistics and machine learning. One of the most challenging, but important, formulations of this problem is the setting of *list-decodable learning* [BBV08, CSV17]. Here, the learner is given a dataset of n points where all but an α -fraction of the data is corrupted, for $\alpha \ll 1$, so that the vast majority of the dataset is complete noise. In this setting, it is easy to see that it is generally impossible to recover the ground truth answer. However, the influential work of [CSV17] demonstrated that it is possible to output a short list of candidate solutions, one of which must achieve non-trivial closeness to the ground truth.

*University of Washington, ziyuncc@cs.washington.edu. Supported by a Simons investigator award 928589, and an NSF grant CCF-2203541.

†Stanford University, compton@stanford.edu. Supported by the NDSEG Fellowship Program, Tselil Schramm’s NSF CAREER Grant no. 2143246, and Gregory Valiant’s Simons Foundation Investigator Award and NSF award AF-2341890.

‡University of California, San Diego, dakane@ucsd.edu. Supported by NSF Medium Award CCF-2107547.

§University of Washington, jerryzli@cs.washington.edu.

While these results are very interesting, they run into a fundamental bottleneck: namely, that the quantitative guarantees these estimates can attain are quite bad. Consider the canonical setting of list-decodable mean estimation: given a dataset of points where an α -fraction of the samples are drawn from a well-behaved distribution (say, an isotropic Gaussian) and the remaining $(1 - \alpha)$ -fraction are arbitrary outliers, all known efficient algorithms incur estimation error at least on the order of $\sqrt{\log(1/\alpha)}$. In particular, the statistical distance of the best estimate in the list and the ground truth Gaussian approaches 1 as α goes to zero. Moreover, it is known that such error is unavoidable, assuming the algorithm outputs a list of size at most $\text{poly}(1/\alpha)$ [DK23].

However, these lower bounds leave open an intriguing possibility: namely, that we could potentially achieve significantly higher accuracy—indeed, even arbitrarily good error—if we are willing to tolerate a slightly larger list size. For instance, suppose that α is some small constant, say $\alpha = 0.25$. Then, current algorithmic results would only seek to output a constant-sized list, and the error of the best estimator in this list would have error $\Omega(1)$. But here, it is very reasonable to ask if, for $\varepsilon > 0$ (say another small constant, e.g. $\varepsilon = 0.01$), it is possible to output a list of slightly larger size that depends mildly on $1/\varepsilon$, so that the error of the best guess in the list is ε . We call this problem *high-accuracy list-decodable mean estimation*. Despite the wealth of work on list-decodable learning, there are no non-trivial guarantees, even information-theoretically, for this natural question. Even basic questions in this vein are unresolved: for instance, prior to this work, it was not even known whether or not there exists a list of dimension-independent size which achieves such a guarantee. In particular, all known identifiability proofs (both efficient and inefficient) for list-decodable learning fundamentally cannot achieve high-accuracy recovery guarantees. Motivated by this discussion, we ask:

What are the statistical and computational limits of high-accuracy list-decodable learning?

Not only is this question natural in its own right, it also has immediate applications to the question of *semi-verified learning*, first proposed by [CSV17]. Here, we are given a large dataset of noisy data, such that an $(1 - \alpha)$ -fraction of it is arbitrary noise, and in addition, we are given a much smaller subset of k trusted data points, which are guaranteed to be from the true distribution, and the goal is to synthesize the combined information to obtain better guarantees than are achievable with just the noisy or the trusted data points alone. It is well-known that any algorithm for list-decodable learning implies non-trivial guarantees for this semi-verified setting, as we can use the list-decodable learning to output a small list of hypotheses, and then perform hypothesis selection on this list using our trusted data points. In particular, doing so allows us to obtain non-trivial guarantees for semi-verified learning when the number of trusted points is much smaller than the dimension. However, because previous algorithms for list-decodable learning could only obtain a low-accuracy list, the resulting semi-verified learner also suffered bad error as a result. In contrast, if one could obtain a high-accuracy list-decodable learner, where the size of the list is sufficiently small, then one could hope to obtain significantly better error rates. Indeed, the difficulty of optimally combining information with different levels of error (even in the special case of semi-verified learning) was observed in [CLC25], where obtaining optimal statistical rates for semi-verified learning was posed as an interesting open question.

From a technical perspective, this question is also very interesting. To somewhat oversimplify the state of affairs, all previous techniques for list-decodable learning sought to recover a subset of points that shares $\alpha^2 n$ points with the true set of good points, usually with some additional regularity conditions on the recovered subset. They then argued that this overlap, plus the regularity

conditions, ensured that the statistics of the recovered set of points cannot deviate too far from the ground truth statistics. However, it is not hard to see that, except in very special cases, such an argument fundamentally cannot obtain high-accuracy guarantees, because it cannot distinguish between the ground truth distribution, and the ground truth distribution conditioned on an event of probability α , and the statistics of these two events can differ wildly. This is to all to say that conceptually novel ideas are necessary to obtain any high-accuracy guarantees in the list-decodable learning setting.

1.1 Our Results

In this work, we obtain the first non-trivial guarantees for high-accuracy list-decodable learning. Specifically, we consider the canonical setting of list-decodable mean estimation for isotropic Gaussians in d -dimensions. Formally, we consider the following, standard noise model for list-decodable estimation:

Definition 1.1. *We say a set of points S is an α -pure set of points with respect to a distribution D if it contains a subset $S_{\text{good}} \subset S$ so that $|S_{\text{good}}| = \alpha|S|$, and S_{good} is a collection of independent samples from D .*

Note that we make no assumptions about the other points in S ; for instance, they could even be chosen adversarially depending on the points in S_{good} . We also note that this is also equivalent to saying that S is a $(1 - \alpha)$ -additively corrupted dataset in the terminology of [DKS18], however, we find that for this setting it is slightly more intuitive to measure the fraction of inliers rather than the fraction of outliers.

Definition 1.2 (List-decodable Gaussian mean estimation). *Let $\alpha, \varepsilon > 0$, and let $\mu \in \mathbb{R}^d$. Given a dataset α -pure dataset S with respect to $N(\mu, I)$ of size n , output a list of L candidate means $\mu_1, \dots, \mu_L$ so that with high probability,*

$$\min_i \|\mu_i - \mu\|_2 \leq \varepsilon.$$

We refer to this quantity ε as the error of the list-decodable learning algorithm.

Our first result is a tight characterization of the information-theoretic limits of this problem. Specifically, we show:

Theorem 1.3 (informal, see Theorem 4.1 and Lemma 4.8). *In the setting of Definition 1.2, there is an (inefficient) estimator which, for n sufficiently large, outputs a list of size*

$$L = \exp\left(O\left(\frac{\log^2(1/\alpha)}{\varepsilon^2}\right)\right)$$

candidate means, which achieves error ε with high probability. Moreover, any algorithm which achieves error ε with constant probability must output a list of size $\exp\left(\Omega\left(\frac{\log^2(1/\alpha)}{\varepsilon^2}\right)\right)$.

We pause to make several remarks about this result. First, note that the list size is completely independent of the dimension. Second, we observe that in the aforementioned setting where α, ε are both small constants, this is the first result that demonstrates that error which is significantly smaller than a large constant larger than 1 is possible for list-decodable mean estimation. The key

technical idea is a new identifiability proof which directly argues that the set of possible candidate means must have bounded size, using ideas from Gaussian process theory and isoperimetry. See [Sections 3.1](#) and [4](#) for a more in-depth discussion of these ideas.

The estimator which achieves the upper bound in [Theorem 1.3](#) is inefficient, and moreover, works in the asymptotic setting, i.e. when n could be very large. Our second result is a new, efficient estimator with non-asymptotic guarantees, which achieves the same error:

Theorem 1.4 (informal, see [Theorem 5.1](#)). *In the setting of [Definition 1.2](#), there is an algorithm which outputs a list of size at most*

$$L = \exp \left(O \left(\frac{\log^2(1/\alpha)}{\varepsilon^2} \right) \right)$$

candidate means, which achieves error ε with probability at least 0.99 when the sample size $n \geq d^{O(\log L)} + \exp(L)$. The time complexity is $d^{O(\log L)} + \exp \exp(\tilde{O}(\log L))$.

In particular, in the regime where α, ε are small constants, our runtime and sample complexity are fully polynomial, demonstrating that it is possible to achieve small constant error efficiently in the list-learning setting.

By combining this with standard hypothesis selection routines, an immediate implication of this is a new algorithm that achieves high-accuracy guarantees for semi-verified learning, with very few trusted points:

Corollary 1.5. *Let $\alpha, \varepsilon > 0$, and let $\mu \in \mathbb{R}^d$. Suppose we are given an α -pure dataset S_{noisy} for $N(\mu, I)$ of size n_1 , as well as a dataset of S_{trusted} of n_2 points drawn independently from $N(\mu, I)$. Suppose that*

$$n_1 \geq d^{O(\log(L))} + \exp(L), n_2 \geq \Omega \left(\frac{\log L}{\varepsilon^2} \right),$$

Then, there is an algorithm which runs in time $d^{O(\log L)} + \exp \exp(\tilde{O}(\log L))$, and which outputs $\hat{\mu}$ so that $\|\hat{\mu} - \mu\|_2 \leq \varepsilon$ with probability at least 0.99.

We outline a short proof in [Section A.1](#).

From a technical perspective, an interesting aspect of our algorithm is that it once again departs from the “standard” paradigm for algorithm design in the list-learning setting. To once again oversimplify, typically, after one has established a complex identifiability proof such as [Theorem 1.3](#), the “standard” approach is to use techniques from SDP hierarchies such as the sum-of-squares hierarchy to convert the proof into an efficient algorithm, see e.g. [\[KSS18, KKK19, RY20, DKK⁺22a, BK21, IK22\]](#). However, our identifiability proof uses somewhat sophisticated machinery, including the previously mentioned Gaussian isoperimetry, and does not seem to easily lift into SoS.

Instead, we propose a two-step algorithm. First, we efficiently identify a low-dimensional subspace which must essentially contain all possible candidate means, by using a new filtering algorithm based on high-degree Hermite polynomials, which may be of independent technical interest. Then, within this subspace, we are able to prune the list of possible candidate means down to the correct size. Crucially, this second step appeals to our identifiability theorem ([Theorem 1.3](#)) in a black-box fashion, to demonstrate that the number of candidates our pruning procedure can return is small. See [Sections 3.2](#) and [5](#) for a more detailed description of our algorithm.

1.2 Related work

List-decodable learning was first proposed in work of [BBV08, CSV17]. By now there is a rich literature on the topic, including efficient algorithms for list-decodable mean estimation [KSS18, DKS18, RY20, DKK⁺21, DKK⁺22b], sparse mean estimation [ZS22, DKK⁺22a], covariance estimation [IK22], linear regression [KKK19, RY20, BK21, DJKS23], and even more general settings [CP23, KSTV25], as well as computational lower bounds [DKP⁺21]. However, none of these works achieve high-accuracy guarantees similar to the ones we are interested in. List-decodable learning is also closely related to the larger literature on robust statistics [Hub92, Tuk60, Ans60, Tuk75], and in particular, the recent wave of interest in algorithmic robust statistics, beginning with work of [DKK⁺19, LRV16]; we defer the interested reader to [DK23] for a more detailed description of this literature.

As described above, list-decodable learning is also closely related to the semi-verified learning problem [CSV17, MV18, ZS23, CLC25], however, none of these results bear any technical relevance to our setting.

2 Preliminaries

Throughout this work, for conciseness of exposition, $c > 0$ will denote any sufficiently small constant, and $C > 0$ will denote any sufficiently large constant. In different lines, the value of c, C may change. For a matrix M , we will let $\|M\|_2$ denote its spectral norm, and we will let $\|M\|_F$ denote its Frobenius norm. Unless the base is otherwise specified, $\log$ refers to the natural log.

We will also require the following classic result:

Theorem 2.1 (Sudakov’s Minoration Inequality [Sud69]). *For a mean zero Gaussian process, for any $\varepsilon \geq 0$ we have*

$$\mathbb{E} \left[\sup_{i \in [l]} X_t \right] \geq c\varepsilon \sqrt{\log(\mathcal{N}(T, d, \varepsilon))}$$

2.1 Low-accuracy list-decodable mean estimation

As a preprocessing routine, we will first use the previous work that efficiently obtains low-accuracy list learning guarantees but with a small list size. Specifically, we will need:

Theorem 2.2 ([DKK⁺22a]). *Let $\alpha > 0$. There is an algorithm which, given an α -pure dataset S with respect to $N(\mu, I)$ of size $n = d^{O(\log(1/\alpha))}$, outputs a list of $L = O(1/\alpha)$ candidate means $\mu_i, \dots, \mu_L$ so that with probability at least 0.99, there exists $i \in [L]$ so that $\|\mu_i - \mu\|_2 \leq O(\sqrt{\log 1/\alpha})$.*

3 Technical overview

The main contributions of our paper are an information-theoretic proof of the near-optimal list size, and an accompanying efficient algorithm. In this section, we highlight the key ideas for both.

3.1 Information-theoretic bound

In this section, we will discuss the existence of a small list for high-accuracy list-decodable learning in an infinite-sample setting. These techniques will later yield desired guarantees for the typical finite-sample list-decodable setting.

Consider the infinite-sample regime for list-decodable learning: the inliers will be distributed according to the distribution $N(\mu, I)$, and the only power of the adversary is where to add the rest of the mass. Hence, in the infinite-sample limit, we expect to see a distribution D where $D(x) \geq \alpha[N(\mu, I)](x) \forall x \in \mathbb{R}^d$. When we observe such a distribution, there may be many values of $\mu \in \mathbb{R}^d$ for which a distribution D satisfies this condition. We similarly define a notion for whether a potential mean $\mu \in \mathbb{R}^d$ is consistent with plausibly being an α -fraction of D :

Definition 3.1 (α -consistent). $\mu \in \mathbb{R}^d$ is α -consistent if $D(x) \geq \alpha \cdot [N(\mu, I)](x) \forall x \in \mathbb{R}^d$.

We desire to output a small list where every α -consistent μ is close to an item in our list. Our procedure for inefficiently constructing a list for D is simple: while there exists an α -consistent μ that is ε -separated from all previous items in our list, then add any such μ . In this sense, our task boils down to the core question: what is the maximum-size list $\mu_1, \dots, \mu_l$ of α -consistent and ε -separated μ_i ?

Warm-up: sketch for $\varepsilon = 10\sqrt{\log(1/\alpha)}$. When ε is large enough, the work of Diakonikolas, Kane, and Stewart [DKS18] yields a simple yet illuminating answer to our core question. Here we present a proof implied by theirs. For sake of contradiction, consider a list $\mu_1, \dots, \mu_l$ of α -consistent and $10\sqrt{\log(1/\alpha)}$ -separated μ_i , where $l = \frac{4}{\alpha}$. Let us define the regions $R_1, \dots, R_l$ that correspond to regions where the density of $N(\mu_i, I)$ is larger than all other $N(\mu_j, I)$. Without proof, we give intuition that because the μ_i are all well-separated, at least half of the mass of each $N(\mu_i, I)$ must be in the region R_i where its density is the largest:

$$\Pr_{X \sim N(\mu_i, I)}[X \in R_i] \geq 1 - \sum_{j \neq i} \Pr_{X \sim N(\mu_i, I)}[[N(\mu_i, I)](x) \leq [N(\mu_j, I)](x)] \geq \frac{1}{2}$$

Yet, by definition of α -consistency, the density of D must always be at least α -fraction the density of any $N(\mu_i, I)$. We are then able to conclude a contradiction that the total mass of D must exceed 1, since too much mass must be in each region R_i :

$$1 \geq \sum_{i=1}^l \Pr_{X \sim D}[X \in R_i] \geq \sum_{i=1}^l \alpha \Pr_{X \sim N(\mu_i, I)}[X \in R_i] \geq \frac{\alpha \cdot l}{2} = 2$$

Accordingly, this proves that any list $\mu_1, \dots, \mu_l$ of α -consistent and $10\sqrt{\log(1/\alpha)}$ -separated μ_i must have size at most $O(1/\alpha)$.

In the actual proof of [DKS18], the analysis mostly goes as we have just described. Their conceptual emphasis is less explicitly about geometric regions R_i , and more about how each μ_i has an associated set, and these sets have limited overlap. In our upcoming proof, we will crucially use how our R_i are the *Voronoi cells*, meaning the regions where μ_i has the largest density.

Intuition for our result. Recall how R_i correspond to the regions where $N(\mu_i, I)$ has the largest density, meaning $R_i = \{\|x - \mu_i\|^2 < \|x - \mu_j\|^2, \forall j \neq i\}$. Let $q_i = \Pr_{x \sim N(\mu_i, I)}[X \in R_i]$. In this language, our analysis will follow a similar program to the warm-up but with sharper guarantees: we will closely analyze the q_i , and then similarly conclude the sum of q_i must exceed $1/\alpha$ if there are too many α -consistent μ_i that are ε -separated.

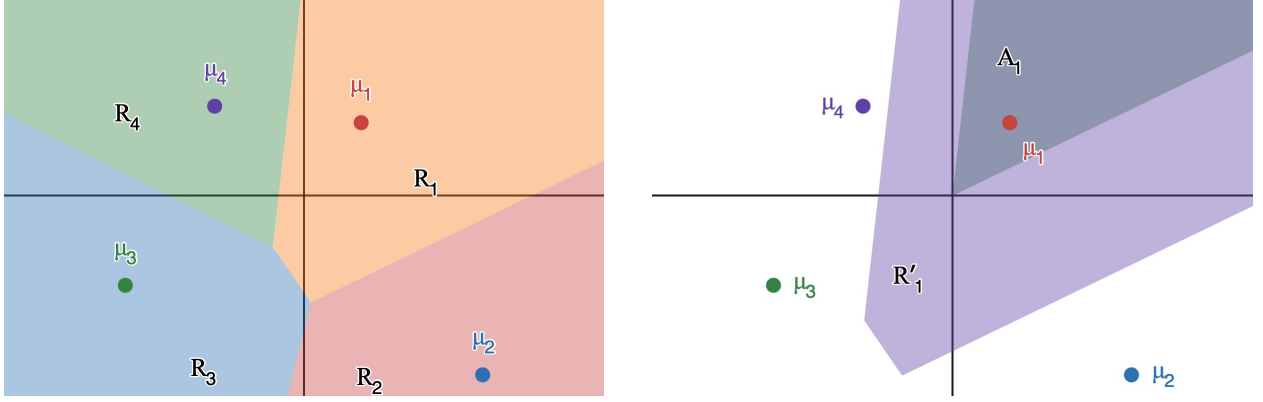


Figure 1: The left figure illustrates the regions R_i where their density is the maximum for the points $\mu_1, \mu_2, \mu_3, \mu_4$. The right figure focuses on A_1 and R'_1 , illustrating how R'_1 contains a fattening of A_1 . Since $q_i \geq \Pr_{X \sim N(0, I)}[X \in R'_1]$, this perspective will enable lower bounds for q_i in terms of A_i .

Our first observation is that we find it helpful to localize our analysis to small balls. We use the initial warm-up result to reduce our task to $O(1/\alpha)$ subproblems, where after re-centering we may assume μ is in a bounded ball such that $\|\mu\|_2 \leq O(\sqrt{\log(1/\alpha)})$.

We now introduce the following auxiliary probability sequence $(p_i)_{i \in [l]}$, defined as $p_i = \Pr_{x \sim N(0, I)}[x \in A_i]$, where $A_i = \{x : \langle x, \mu_i \rangle > \langle x, \mu_j \rangle, \forall j \neq i\}$. This region A_i will be a surprisingly insightful reference for analyzing the mass in regions R_i .

Let R'_i correspond to the region R_i translated by $-\mu_i$. By definition,

$$q_i = \Pr_{X \sim N(\mu_i, I)}[X \in R_i] = \Pr_{X \sim N(0, I)}[X \in R'_i]$$

This indicates how relationships between A_i and R'_i may inform the value of q_i . We initially observe that $q_i \geq p_i$ via the following simple analysis that $R'_i \supseteq A_i$:

$$R'_i = \{x : \|x\|^2 < \|x + (\mu_i - \mu_j)\|^2, \forall j \neq i\} \supseteq \{x : \langle x, \mu_i \rangle > \langle x, \mu_j \rangle, \forall j \neq i\} = A_i$$

More consequentially, after a short calculation, we observe R'_i actually contains an $\varepsilon/2$ -fattening of A_i (e.g. see Fig. 1). Combined with the Gaussian isoperimetry inequality, this will give us a favorable lower bound for q_i in terms of p_i . For example, in the regime where $\varepsilon \in (0, 1]$, we use

$$q_i \geq p_i \cdot \exp(c\varepsilon \sqrt{\log(1/p_i)}). \quad (1)$$

Let us describe intuition for how this bound leads us towards a proof. For simplicity, consider a list $\mu_1, \dots, \mu_l$ where the auxiliary probabilities p_i happen to be uniform: $p_1, \dots, p_l = \frac{1}{l}$. This is not generally true, but is an insightful example. In this case, using Eq. (1), we could conclude that for any list of size $l > \exp(C \log^2(1/\alpha)/\varepsilon^2)$, the sum of q_i must exceed $1/\alpha$ and cause a contradiction.

Of course, the auxiliary probabilities p_i need not be uniform. Moreover, the lower bound of Eq. (1) is not fruitful in many cases, such as an example where $(p_1, \dots, p_l) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots)$. A key observation, however, is that the values of p cannot be arbitrary. Indeed, $p_1, \dots, p_l$ correspond to the distribution of the arg max for a Gaussian process $\sup_{i \in [l]} \langle x, \mu_i \rangle$. With a careful technical argument, we show that the distribution p must be well-spread enough for Eq. (1) to yield a favorable bound, or otherwise the distribution p would cause a violation of Sudakov's minoration inequality (and

hence is not consistent with the distribution of the optimizer of a Gaussian process). Eventually, this yields our desired information-theoretic guarantee.

3.2 Efficient list-decodable learning

Our plan will be to design an algorithm that efficiently leverages our information-theoretic proof. In that proof, our procedure was to simply choose ε -separated, α -consistent μ_i repeatedly until there were no more such μ_i . For this section, it is both impossible to know whether a μ_i is α -consistent from only finite samples, and also non-obvious how to efficiently find μ_i with our desired properties.

As a remedy for this, we will choose μ_i for which there exists a collection of αn samples, where these samples centered around μ_i have empirical moments that are very close to the standard Gaussian $N(0, I)$. Relaxing this slightly, we may look for a $\mu_i \in \mathbb{R}^d$ and sample weights $w_1, \dots, w_n \in [0, 1]$ with $\sum_i w_i = \alpha n$, where μ_i is ε -separated from the previously chosen list entries, and

$$\frac{1}{\alpha n} \sum_i w_i \cdot \langle X_i - \mu_i, v \rangle^j \approx \mathbb{E}_{X \sim N(0, I)} [x^j]$$

for all $\|v\|_2 = 1$ and $j \in \{1, \dots, k\}$ for some bounded k .

First, we will employ prior work of [DKK⁺22a] to efficiently reduce our task to $O(1/\alpha)$ subproblems, where each has the restriction $\|\mu_i\|_2 \leq O(\sqrt{\log 1/\alpha})$.

Finding a low-dimensional subspace. Next, we will find a low-dimensional subspace that is close to the true μ with high probability. Crucially, the dimension of this subspace will not depend on d . We will do so via a two-step process. The first step will be an iterative filtering algorithm similar to those in the robust statistics literature [DKK⁺19] on the empirical high-order Hermite polynomial tensors of the dataset. More specifically, let $H_t(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d^t}$ denote the t -th order (probabilist's) Hermite polynomial tensor (see Definition 5.6), for t appropriately chosen but independent of d . By standard results in high-dimensional probability, we know that if we take enough samples, the empirical statistics over the good samples of the Hermite polynomials concentrate very tightly around the population statistics, and the population statistics of the Hermite polynomials satisfies

$$\left\| \mathbb{E}_{X \sim N(\mu, I)} [H_t(X) \otimes H_t(X)] \right\|_2 \leq t! \cdot C^t \exp\left(\sqrt{t} \|\mu\|_2\right),$$

which is in particular, much smaller than $\exp(L)$ for the list size L we are targeting. We show that these facts imply that we can use iterative filtering to remove samples from the dataset so that (1) we remove almost no good points, and (2) the ℓ_2 -norm of the empirical Hermite polynomial (when treated as a d^t -length vector) over the remaining data points is bounded. See Lemma 5.12 for the full analysis. Once we have this, the ℓ_2 bound on the degree t empirical Hermite polynomial implies that if we flatten it into a $(d \times d^{t-1})$ matrix, this matrix must have bounded Frobenius norm, and so in particular, it has a small number of large singular values. As a final step, we demonstrate that if we take the union of the span of the left singular vectors with large singular values of these flattened matrices, this subspace must approximately contain μ . Intuitively, this is because our condition implies that in all directions orthogonal to this subspace, the low-degree moments of dataset match that of a standard normal Gaussian. However, if there was a candidate mean that was not contained in this subspace, then this would necessarily induce a large moment, and thus we can conclude that no such candidate mean can exist.

Searching for moment-matching candidates. With a low-dimensional subspace in hand, we may now afford running times with exponential dependence in the dimension of the subspace. Thus, we may exhaustively search over a net for points that approximately match moments as desired. Checking whether a fixed center has sample weights $w_1, \dots, w_n$ that match moments is an efficiently-solvable convex program. This subroutine produces an $\varepsilon/2$ -separated list of points $\mu_1, \dots, \mu_l$ satisfying our moment-matching condition, and at least one of these points will be ε -close to the true μ with high probability.

Bounding list size via fooling. The main remaining question is whether this list $\mu_1, \dots, \mu_l$ must be small. We will show any collection of $\varepsilon/2$ -separated μ_i that satisfy this moment condition must have bounded size. Recall $R_i \triangleq \{\|x - \mu_i\| < \|x - \mu_j\|\}$: the region where $N(\mu_i, I)$ has the largest density. In our information-theoretic proof, we showed that in the case where the ε -separated μ_i were all α -consistent, then for large enough l there is a contradiction

$$1 \geq \sum_{i=1}^l \Pr_{X \sim D}[X \in R_i] \geq \sum_{i=1}^l \alpha \cdot \Pr_{X \sim N(\mu_i, I)}[X \in R_i] > 1.$$

In the original proof, most of the difficulty was in the last step of the inequality. When we adapt this argument for moment-matching μ_i , the main difficulty is the second inequality. For any moment-matching μ_i , let $w^{(i)}$ be the corresponding vector of sample weights in $[0, 1]$, and let D_i be the normalized empirical distribution over these samples: $D_i = \frac{1}{\alpha n} \sum_{j=1}^n w_j^{(i)} \cdot X_j$. Our same information-theoretic proof will show that our list of moment-matching μ_i cannot be too large, if we prove

$$\sum_{i=1}^l \alpha \cdot \Pr_{X \sim D_i}[X \in R_i] \approx \sum_{i=1}^l \alpha \cdot \Pr_{X \sim N(\mu_i, I)}[X \in R_i].$$

Meaning, we must prove the normalized empirical distribution from $w^{(i)}$ has approximately as many points in R_i as the distribution $N(\mu_i, I)$. In general, just because the lower-order moments of D_i match the moments of $N(\mu_i, I)$, does not mean the proportion of their samples within some region will be approximately the same. However, observe that our particular region R_i is exactly the intersection of $l - 1$ halfspaces. There is a rich body of work in pseudorandomness (e.g. [Baz09, Bra08, DGJ⁺10, DKN10, KM13, KKM13]) and more recently, testable learning (e.g. [RV23, GKK23, DKK⁺23, KSV24, STW24, DKLZ24]) studying how matching moments implies fooling concept classes like halfspaces; hence, distributions that match moments will have roughly the same proportion of samples within an intersection of halfspaces. The works of [DKN10, GKK23] guide us towards the technical fooling result we desire. However, we cannot immediately use either of these results without modification, since our application only matches moments *approximately* (the statement of [DKN10] uses exact matching), and we desire tighter guarantees for super-constant numbers of halfspaces than given by [GKK23].

All together, since we efficiently find a list $\mu_1, \dots, \mu_l$ where each μ_i has a corresponding sample weights $w^{(i)}$ that match moments with $N(\mu_i, I)$, and since moment-matching fools the intersection of halfspaces, then our information-theoretic proof bounds the size of our list.

4 Information-theoretic bound

In this section, we will show the existence of a small list for high-accuracy list-decodable learning. We will discuss an infinite-sample setting, but this will later give us results that yield the desired

guarantee for the typical list-decodable setting.

Recall the notion of α -consistency (Definition 3.1) describing whether a potential mean $\mu \in \mathbb{R}^d$ is consistent with plausibly being an α -fraction of D . We now state our main information-theoretic result for list-decodable mean estimation:

Theorem 4.1. *Consider any distribution D over $\mathbb{R}^d$, inlier parameter $\alpha \in (0, 1/2]$, and accuracy parameter $\varepsilon \in (0, 10\sqrt{\log(1/\alpha)}]$. There exists a list L such that for any α -consistent $\mu \in \mathbb{R}^d$, there is a $\mu' \in L$ such that $\|\mu - \mu'\|_2 \leq \varepsilon$. Further, L has size at most $\exp(C \log^2(1/\alpha)/\varepsilon^2)$ for sufficiently large $C > 0$.*

This is an information-theoretic version of our eventual goal; we will not worry about leveraging only a finite number of samples or having desirable runtime.

Diakonikolas, Kane, and Stewart [DKS18] showed that you may learn a list of size $O(1/\alpha)$ with accuracy $\|\mu - \mu'\|_2 \leq O(\sqrt{\log(1/\alpha)})$. Our improvement in this work is improving the accuracy beyond this barrier. The information-theoretic version of the classical result is quite simple, and we will present a sketch of a proof implied by theirs as a warm-up.

4.1 Warm-up: sketch for learning up to accuracy $O(\sqrt{\log(1/\alpha)})$

Lemma 4.2. *Consider any distribution D over $\mathbb{R}^d$ and an inlier parameter $\alpha \in (0, 1/2]$. There exists a list L such that for any α -consistent $\mu \in \mathbb{R}^d$, there is a $\mu' \in L$ such that $\|\mu - \mu'\|_2 \leq 10\sqrt{\log(1/\alpha)}$. Further, L has size at most $\frac{4}{\alpha}$.*

Proof. Suppose there existed a list of means $L = [\mu_1, \dots, \mu_l]$, where all μ_i are α -consistent, and all pairs are $10\sqrt{\log(1/\alpha)}$ -separated (meaning, $\|\mu_i - \mu_j\|_2 \geq 10\sqrt{\log(1/\alpha)}$ for $i \neq j$). We will show by contradiction that it is impossible for such a list to have size $\lceil 4/\alpha \rceil$.

Let us define regions $R_1, \dots, R_l$ that correspond to regions where the density of $N(\mu_i, I)$ is larger than all other $N(\mu_j, I)$ for $i \neq j$.

First, we observe how most of the density of $N(\mu_i, I)$ is contained within R_i :

$$\begin{aligned} \Pr_{X \sim N(\mu_i, I)}[X \in R_i] &\geq 1 - \sum_{j \neq i} \Pr_{X \sim N(\mu_i, I)}[[N(\mu_i, I)](x) \leq [N(\mu_j, I)](x)] \\ &= 1 - \sum_{j \neq i} \Pr_{Z \sim N(0, 1)}[z \geq \|\mu_i - \mu_j\|] \geq 1 - \left\lceil \frac{4}{\alpha} \right\rceil \cdot \Pr_{Z \sim N(0, 1)}\left[z \geq 10\sqrt{\log(1/\alpha)}\right] \\ &\geq 1 - \left\lceil \frac{4}{\alpha} \right\rceil \cdot e^{-50 \log(1/\alpha)} \geq \frac{1}{2} \end{aligned}$$

After this observation, we may conclude by definition of α -consistency,

$$1 \geq \sum_{i=1}^l \Pr_{X \sim D}[X \in R_i] \geq \sum_{i=1}^l \alpha \cdot \Pr_{X \sim N(\mu_i, I)}[X \in R_i] \geq l \cdot \alpha \cdot \frac{1}{2} = 2,$$

which is a contradiction. Let us define $q_i \triangleq \Pr_{X \sim N(\mu_i, I)}[X \in R_i]$. In summary, we have shown that a list of too many well-separated means that are α -consistent would cause the sum of q_i to be large enough to cause a contradiction.

This proof sketch is enough to easily show the information-theoretic existence of a small list. Let us construct the list iteratively, starting with an empty list $L = []$. At each step, if there is an α -consistent μ that it is at least $10\sqrt{\log(1/\alpha)}$ away from everything in L , then arbitrarily add any such μ . Otherwise, our list is finished. By our above proof, this will terminate in less than $\frac{4}{\alpha}$ steps, and hence yield a list of size $\leq \frac{4}{\alpha}$. $\square$

4.2 Our result: learning up to accuracy ε

We will now detail how to get a list with better accuracy ε . First, we may use the previous warm-up as a subroutine, to give us a preliminary list L_0 where $|L_0| \leq \frac{4}{\alpha}$, and any α -consistent μ will be within $10\sqrt{\log(1/\alpha)}$ distance of some element in L_0 .

This will let us consider separately learning within balls of radius $10\sqrt{\log(1/\alpha)}$. Focusing within these balls may not immediately seem helpful, but will prove useful later.

Recall $R_i = \{\|x - \mu_i\|^2 < \|x - \mu_j\|^2, \forall j \neq i\}$ and $q_i = \Pr_{x \sim N(\mu_i, I)}[X \in R_i]$. Our analysis will follow a similar program to the warm-up but with sharper guarantees: we will closely analyze the q_i , and then similarly conclude the sum of q_i must exceed $1/\alpha$ if there are too many α -consistent μ that are β -separated. (Using $\beta = \varepsilon$ will clearly give some helpful guarantee, but later we will get sharper results by using this for general β .) First, we lower bound the sum of q_i :

Theorem 4.3. *Assume $\mu_1, \dots, \mu_l$ are β -separated and $\|\mu_i\| \leq r$, where $l \geq 2^{Cr^4/\beta^6}$ for sufficiently large $C > 0$, $0 < \alpha \leq \frac{1}{2}$, and $0 < \beta \leq C\sqrt{\log(1/\alpha)}$. Then,*

$$\text{poly}(1/\alpha) \cdot \sum_{i=1}^L q_i \geq 2^{\frac{c\beta^3}{r^2} \sqrt{\log l}}. \quad (2)$$

Deferring the proof of [Theorem 4.3](#) to [Section 4.2.1](#), let us first see how this implies [Theorem 4.1](#).

Concluding Theorem 4.1. First, use [Lemma 4.2](#) to obtain a list L_0 of size at most $\frac{4}{\alpha}$ such that any α -consistent μ is within $r_0 \triangleq 10\sqrt{\log(1/\alpha)}$ of some item in L_0 .

For each $\mu' \in L_0$, we will create a sub-list L' . Without loss of generality, translate the distribution so μ' is the origin. We now aim for a list with error at most ε . Our goal will be complete if for any α -consistent μ where $\|\mu - \mu'\| \leq r_0$, there is a $\mu^* \in L'$ where $\|\mu - \mu^*\| \leq \varepsilon$.

A naive application of [Theorem 4.3](#) would give some upper bound on the size of each list $\mu_1, \dots, \mu_l$ by setting $r = 10\sqrt{\log(1/\alpha)}$ and $\beta = \varepsilon$. This would yield a list upper bound of size $\exp(C \log^4(1/\alpha)/\varepsilon^6)$. However, we may apply [Theorem 4.3](#) more carefully and do better.

Consider the union of all sub-lists to be $\mu_1, \dots, \mu_l$, where all μ_i are α -consistent and ε -separated. Let $|M_i|$ denote the size of the smallest $\frac{r_0}{2^i}$ -cover of the list $\mu_1, \dots, \mu_l$. By construction, $|M_0| \leq 4/\alpha$.

Claim 4.4. *For any positive integer i , it holds that $|M_i| \leq |M_{i-1}| \cdot \exp(C \log^2(1/\alpha)/r_i^2)$.*

Proof. For constructing the $r_i \triangleq \frac{r_0}{2^i}$ -cover, we may separately consider covering the list entries within the r_{i-1} -radius balls centered at the cover centers M_{i-1} . The covering number within each r_{i-1} -ball is at most the size of the largest set of r_i -separated μ_j in this ball. Suppose there were a collection of l' items μ_j in our list that are r_i -separated in this r_{i-1} -radius ball. Then, using [Theorem 4.3](#) with

$r = r_{i-1}$ and $\beta = r_i$, we conclude:

$$\begin{aligned}
\sum_{i=1}^{l'} \alpha \Pr_{X \sim D}[X \in R_i] \leq 1 &\iff \sum_{i=1}^{l'} q_i \leq 1/\alpha \\
&\implies 2^{\frac{cr_i^3}{2^{r_{i-1}}}} \sqrt{\log l'} \cdot \text{poly}(\alpha) \leq 1 \\
&\iff 2^{r_i \sqrt{\log l'}} \cdot \text{poly}(\alpha) \leq 1 \\
&\implies l' \leq \exp(C \log^2(1/\alpha)/r_i^2) \quad \square
\end{aligned}$$

Let $i^* \triangleq \lceil \log_2(\frac{r_0}{\varepsilon/4}) \rceil$. Observe how it must hold that $l = |M_{i^*}|$, since the list $\mu_1, \dots, \mu_l$ is ε -separated and $r_{i^*} < \varepsilon/2$. Using [Claim 4.4](#), we may conclude our proof by computing

$$\begin{aligned}
l = |M_{i^*}| &= |M_0| \cdot \prod_{i=1}^{i^*} \frac{|M_i|}{|M_{i-1}|} \\
&\leq \frac{4}{\alpha} \cdot \prod_{i=1}^{i^*} \exp(C \log^2(1/\alpha)/r_i^2) \\
&\leq \exp(C \log^2(1/\alpha)/\varepsilon^2).
\end{aligned}$$

This completes the proof of our information-theoretic result.

Remark for finite samples. Let us roughly discuss how this proof technique would yield results for the typical list-decodable setting. (We only discuss roughly since the later result [Theorem 5.1](#) will directly prove this.) Observe how the above proof mostly leverages that if some μ' is α -consistent, then $\Pr_{X \sim D}[X \in R_i] \geq \alpha \Pr_{X \sim N(\mu', I)}[X \in R_i]$. Imagine if we defined a similar notion of being *finitely*-(α, η, l)-consistent: where a μ' is finitely-(α, η, l)-consistent on a normalized empirical distribution D' if $\Pr_{X \sim D'}[X \in H] \geq \alpha \Pr_{X \sim N(\mu_i, I)}[X \in H] - \eta$, for all H that are the intersection of at most l halfspaces. Since the intersection of l halfspaces is a concept class with bounded VC dimension, we may show that the true μ will be finitely-(α, η, l)-consistent for a sufficiently large n with high probability. Moreover, this condition is enough for our proof to work: we could follow a similar proof structure as before, repeatedly choosing any μ' that is ε -separated from the existing list and is finitely-(α, η, l)-consistent, until there is no such μ' .

Employing this plan efficiently poses numerous challenges that we approach in [Section 5](#).

4.2.1 Proof of [Theorem 4.3](#)

Proof. We now prove the main theorem for our identifiability result. We consider the following auxiliary probability sequence $(p_i)_{i \in [l]}$, defined as $p_i = \Pr_{x \sim N(0, I)}[x \in A_i]$, where $A_i = \{x : \langle x, \mu_i \rangle > \langle x, \mu_j \rangle, \forall j \neq i\}$. Let R'_i correspond to the region R_i translated by $-\mu_i$. By definition,

$$q_i = \Pr_{X \sim N(\mu_i, I)}[X \in R_i] = \Pr_{X \sim N(0, I)}[X \in R'_i] = \Pr_{X \sim N(0, I)}[\|x\|^2 < \|x - \mu_j + \mu_i\|, \forall j \neq i].$$

We initially observe that $q_i \geq p_i$ via the following simple analysis that $R'_i \supseteq A_i$:

$$R'_i = \{x : \|x\|^2 < \|x + (\mu_i - \mu_j)\|^2, \forall j \neq i\} \supseteq \{x : \langle x, \mu_i \rangle > \langle x, \mu_j \rangle, \forall j \neq i\} = A_i$$

More consequentially, we may show R'_i actually contains an *fattening* of A_i ; combined with the Gaussian isoperimetry inequality, this will give us a favorable lower bound for q_i in terms of p_i :

Lemma 4.5. $q_i \geq \alpha^{c'} \cdot p_i \cdot \exp(c\beta\sqrt{\log(1/p_i)})$.

Proof. We show that R'_i contains a fattening of A_i :

Claim 4.6. $(A_i)_{\beta/2} \triangleq \{x : \text{dist}(x, A_i) \leq \beta\} \subseteq \{x : \|x\|^2 < \|x - \mu_j + \mu_i\|^2, \forall j \neq i\} = R'_i$

Proof. For $y \in (A_i)_{\beta/2}$, write it as $x + h$, where $x \in A_i$ and $\|h\| \leq \beta/2$. Then for any $j \neq i$

$$\langle y, \mu_j - \mu_i \rangle = \langle x, \mu_j - \mu_i \rangle + \langle h, \mu_j - \mu_i \rangle < \langle h, \mu_j - \mu_i \rangle \leq \|h\| \|\mu_i - \mu_j\| \leq \frac{1}{2} \|\mu_j - \mu_i\|^2,$$

which implies $\|y\|^2 < \|y - \mu_j + \mu_i\|^2$ as we desired. $\square$

Since $\Pr_{x \in N(0, I)}[x \in A_i] = p_i$, we will use the Gaussian isoperimetry inequality to get a lower bound for the measure of the fattening $\Pr_{x \in N(0, I)}[x \in (A_i)_{\beta/2}]$. We use the following implication of the Gaussian isoperimetry inequality and defer the proof to [Section A.2](#):

Lemma 4.7 (Gaussian isoperimetry). *For a measurable set A in $\mathbb{R}^n$ under standard gaussian measure μ , let $A_\beta = \{x \in \mathbb{R}^n : \text{dist}(x, A) \leq \beta\}$. If $\mu(A) \leq 1/2$, then*

$$\mu(A_\beta) \geq \mu(A) \cdot \exp\left(-c'(1 + \beta^2) + c\beta\sqrt{\log(1/\mu(A))}\right).$$

Using [Claim 4.6](#), [Lemma 4.7](#), $\beta \leq O(\sqrt{\log(1/\alpha)})$, and that $p_i \leq \frac{1}{2}$ by symmetry, we conclude

$$q_i \geq \Pr_{x \in N(0, I)}[x \in (A_i)_{\beta/2}] \geq \alpha^{c'} \cdot p_i \cdot \exp\left(c\beta\sqrt{\log(1/p_i)}\right),$$

which completes the proof of our lemma. $\square$

We now hope to use [Lemma 4.5](#) to prove [Theorem 4.3](#). We remark that if $p_1, \dots, p_l$ could be any arbitrary non-negative values that sum to 1, then our approach would not yield [Theorem 4.3](#). For example, if $\beta = 1$ and our vector p was the infinite-length vector $\frac{1}{2}, \frac{1}{4}, \dots$, then the quantity $\sum_i p_i \cdot \exp(c\beta\sqrt{\log(1/p_i)})$ from [Lemma 4.5](#) would converge to a constant, yet our desired bound for [Theorem 4.3](#) will go to infinity as $l \rightarrow \infty$.

Luckily for our approach, the values of p_i are not arbitrary. Each p_i corresponds to probability that μ_i is the maximizer for a Gaussian process $\max_i \langle x, \mu_i \rangle$ for $X \sim N(0, I)$. We will prove this distribution of maximizers is constrained in a way that enforces the distribution p must take values such that the quantity $\sum_i p_i \cdot \exp(c\beta\sqrt{\log(1/p_i)})$ grows according to our desired theorem.

Warm-up intuition: infinite dyadic $p = [1/2, 1/4, \dots]$ is impossible. As a warm-up, let us quickly show why it is impossible for the vector p to take value $[1/2, 1/4, \dots]$. By Sudakov's minoration inequality ([Theorem 2.1](#)), we know the expected maximum $\mathbb{E}[\sup_{i \in [l]} \langle x, \mu_i \rangle] \geq c\beta\sqrt{\log(l)}$ is infinite in this example where l is infinite. However, if p is dyadic and corresponds to the distribution of maximizers, then we may conclude this expected maximum is finite:

$$\begin{aligned} \mathbb{E}_{X \sim N(0, I)}[\sup_{i \in [l]} \langle x, \mu_i \rangle] &= \sum_{i=1}^{\infty} \mathbb{E}_{X \sim N(0, I)}[\mathbb{1}_{\mu_i \text{ maximizes}} \cdot \langle x, \mu_i \rangle] \\ &\leq \sum_{i=1}^{\infty} \sqrt{\mathbb{E}_{X \sim N(0, I)}[\mathbb{1}_{\mu_i \text{ maximizes}}^2] \cdot \mathbb{E}_{X \sim N(0, I)}[\langle x, \mu_i \rangle^2]} \quad (\text{via Cauchy-Schwarz}) \\ &= \sum_{i=1}^{\infty} \sqrt{2^{-i} \cdot \|\mu_i\|^2} \leq Cr \end{aligned}$$

This contradicts the lower bound from Sudakov's minoration inequality, and shows that p could not possibly take the value of this infinite dyadic distribution. In general, this gives some sense that the distribution of p must be well-spread in a way that hopefully gives us a favorable lower bound for $\sum_i p_i \cdot \exp(c\beta\sqrt{\log(1/p_i)})$. In the remaining proof, we will judiciously apply Holder's inequality in a way that attains our desired result.

General proof. By [Lemma 4.5](#), we know

$$\sum_{i \in [l]} q_i \geq \alpha^{c'} \cdot \sum_{i \in [l]} p_i \cdot \exp(\beta\sqrt{\log(1/p_i)}).$$

We use Holder's inequality on the expectation of the Gaussian process, using positive constants s, t that satisfy $\frac{1}{s} + \frac{1}{t} = 1$ (we choose s, t later),

$$\sum_{i \in [l]} \mathbb{E}[\mathbb{1}_{x \in A_i} \cdot \langle x, \mu_t \rangle] \leq \sum_{i \in [l]} (\mathbb{E}[\mathbb{1}_{x \in A_i}])^{1/t} \cdot \mathbb{E}[|\langle x, \mu_t \rangle|^s]^{1/s} = \sum_{i \in [l]} (p_i)^{1-1/s} \cdot O(r\sqrt{s}).$$

By Sudakov's minoration inequality ([Theorem 2.1](#)) we know the left-hand side is at least $c\beta\sqrt{\log l}$. We will choose $s = \frac{c^2\beta^2 \log l}{9r^2}$, which satisfies $s > 1$ by our condition on l . Using this choice of s :

$$c\beta\sqrt{\log l} \leq \sum_{i \in [l]} (p_i)^{1-1/s} \cdot O(r\sqrt{s}) \implies \sum_{i \in [l]} (p_i)^{1-1/s} \geq 3.$$

We observe this sum is still large even when restricting to $S = \{i \in [l] : p_i \geq 1/l^2\}$:

$$\begin{aligned} \sum_{i \in S} (p_i)^{1-1/s} &= \sum_{i \in [l]} (p_i)^{1-1/s} - \sum_{i \notin S} (p_i)^{1-1/s} \geq 3 - \sum_{i \notin S} (p_i)^{1-1/s} \\ &\geq 3 - \sum_{i \notin S} (p_i)^{1/2} \geq 3 - \sum_{i \notin S} 1/l \geq 2 \quad (\text{using } s \geq 2 \text{ from our condition on } l) \end{aligned}$$

Further, for $i \in S$, we have $(p_i)^{1-1/s} \leq p_i \cdot \exp\left(\frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}}\right)$, this is because

$$\begin{aligned} (p_i)^{1-1/s} &\leq p_i \cdot \exp\left(\frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}}\right) \\ \iff (1/p_i)^{1/s} &\leq \exp\left(\frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}}\right) \\ \iff \frac{9r^2}{c^2\beta^2 \log l} \log(1/p_i) &\leq \frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}} \\ \iff \sqrt{\log(1/p_i)} &\leq 2\sqrt{\log l} \\ \iff p_i &\geq 1/l^4. \end{aligned}$$

Thus,

$$\sum_{i \in [l]} p_i \cdot \exp\left(\frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}}\right) \geq \sum_{i \in S} p_i \cdot \exp\left(\frac{18r^2\sqrt{\log(1/p_i)}}{c^2\beta^2\sqrt{\log l}}\right) \geq \sum_{i \in S} p_i^{1-1/s} \geq 2. \quad (3)$$

With this, we may conclude the proof of our theorem:

$$\begin{aligned}
\alpha^{-c'} \cdot \sum_{i \in [l]} q_i &\geq \sum_{i \in [l]} p_i \cdot \exp(\beta \sqrt{\log(1/p_i)}) \\
&\geq \left(\sum_{i \in [l]} p_i \cdot \exp\left(\frac{18r^2 \sqrt{\log(1/p_i)}}{c^2 \beta^2 \sqrt{\log l}}\right) \right)^{\frac{c^2 \beta^3}{18r^2} \sqrt{\log l}} \quad (\text{Jensen's inequality; } l \text{ chosen s.t. } \frac{c^2 \beta^3}{18r^2} \sqrt{\log l} \geq 1) \\
&\geq 2^{\frac{c^2 \beta^3}{18r^2} \sqrt{\log l}}. \quad \square \quad (\text{via Eq. (3)})
\end{aligned}$$

4.3 List size lower bound

Theorem 1.5 of [DKS18] also shows how it is information-theoretically impossible to get better than $\Omega(\sqrt{\log 1/\alpha})$ accuracy with a list of only size $\text{poly}(1/\alpha)$. For a general lower bound in terms of the list size required for ε accuracy, the same construction of [DKS18] will yield a lower bound qualitatively similar to our upper bound in the previous section. Interestingly, we note that for list-decodable learning up to accuracy ε , list size of $2^{\Omega(\log^2(1/\alpha)/\varepsilon^2)}$ was necessary. Observe how when $\varepsilon = \Theta(\sqrt{\log(1/\alpha)})$, this size is $\text{poly}(1/\alpha)$, as expected. The proof for this lower bound follows the proof strategy of Proposition 5.11 in [DK23]; we present this for completeness in Section A.3.

Lemma 4.8. *Let $0 < \varepsilon < \frac{\sqrt{\log(1/2\alpha)}}{2}$ and $0 < \alpha < 1/3$. There exists a distribution D such that for any list containing an ε -close mean for all possible α -consistent means, the list must have size at least $\exp(c \log^2(1/\alpha)/\varepsilon^2)$.*

5 Efficient list-decodable learning

In this section, we will design an algorithm that efficiently constructs a list of size comparable to our information-theoretic upper bound of Theorem 4.1, given access to a finite number of samples. We gave an overview of the proof in Section 3.2, and formally state our result here:

Theorem 5.1. *In the setting of Definition 1.2, there is an algorithm which outputs a list of at most*

$$L = \exp\left(O\left(\frac{\log^2(1/\alpha)}{\varepsilon^2}\right)\right)$$

candidate means, which achieves error ε with probability at least 0.99 when $n \geq d^{C \log^2(1/\alpha)/\varepsilon^2} + \exp(L)$. The time complexity is

$$d^{C \log^2(1/\alpha)/\varepsilon^2} + \exp\left(\exp\left(\tilde{O}\left(\frac{\log^2(1/\alpha)}{\varepsilon^2}\right)\right)\right).$$

In Section 5.1, we will find a low-dimensional subspace that is close to μ with high probability. In Section 5.2, we focus on this low-dimensional subspace and exhaustively search over a net for points that approximately match moments as desired. In Section 5.3, we will detail the required technical result for how approximately matching moments fools halfspaces. In Section 5.4, we finally combine these ingredients to conclude our algorithm.

5.1 Subspace learning

In this section, we give the first key algorithmic subroutine we will need for our efficient algorithm, namely, a method which allows us to recover a low-dimensional subspace which approximately contains all of the candidate means. More formally, we show:

Theorem 5.2. *Let $\alpha, \varepsilon > 0$. Let S be an α -pure dataset with respect to $N(\mu, I)$ of size*

$$n = \left(\frac{d \log(1/\alpha)}{\varepsilon} \right)^{\Omega(\log^2(1/\alpha)/\varepsilon^2)},$$

and assume that $\|\mu\|_2 \leq C\sqrt{\log(1/\alpha)}$. There is an algorithm which, given S , outputs a subspace $V \subset \mathbb{R}^d$ of dimensionality $k^{O(k)}$, where $k = 100 \log^2(1/\alpha)/\varepsilon^2$, so that with probability at least 0.99, we have that $\|\Pi_{V^\perp} \mu\|_2 \leq \varepsilon$. Moreover, this algorithm runs in time $\text{poly}(n)$.

We begin with preliminaries, then give an overview, and finally detail our analysis.

5.1.1 Tensor preliminaries

We will often have to work with high order tensors, so here we will record some useful notation for them.

Definition 5.3 (Tensor indexing). *Let t be a positive integer. For any partition of $[t]$ into sets $S_1, \dots, S_k$, and any tensors $T_1, \dots, T_k$ of orders $|S_1|, \dots, |S_k|$ respectively, we let*

$$T_1^{S_1} \otimes \dots \otimes T_k^{S_k}$$

to denote the order t tensor obtained by taking the tensor product of T_1 in the indices S_1 with T_2 in the indices S_2 , etc.

Definition 5.4 (Tensor symmetrization). *Let A, B be tensors of rank t_1 and t_2 respectively over $\mathbb{R}^d$. We let the symmetrization of $A \otimes B$, denoted $\text{Sym}(A \otimes B)$ be given by*

$$\text{Sym}(A \otimes B) = \frac{1}{\binom{t_1+t_2}{t_1}} \sum_{\substack{[t]=S_1 \sqcup S_2 \\ |S_1|=t_1, |S_2|=t_2}} A^{S_1} \otimes B^{S_2},$$

that is, we sum over all possible ways of combining A and B into a rank $t_1 + t_2$ tensor.

We note that we will be sometimes a bit loose, and associate a rank t tensor over $\mathbb{R}^d$ with a vector in $\mathbb{R}^{d^t}$ in the natural way. We will also need the following way of associating tensors with a canonical matrix:

Definition 5.5. *Let $t \geq 1$. For any order t tensor T , we let $\text{flat}(T)$ be the $\mathbb{R}^d \times \mathbb{R}^{d^{t-1}}$ flattening of T on its first mode. That is, $\text{flat}(u \otimes H) = uH^\top$ for any $H \in \mathbb{R}^{d^{t-1}}$, and we can linearly extend the definition of flat onto all rank- t tensors.*

Hermite polynomial tensors Our algorithm will heavily leverage the structure of the Hermite polynomial tensors, the natural multivariate analog of the univariate Hermite polynomials.

Definition 5.6 (The (probabilist's) Hermite polynomial tensor). *For all non-negative integers k , let $h_k : \mathbb{R} \rightarrow \mathbb{R}$ denote the degree- k (probabilist's) Hermite polynomial, defined recursively by $h_0 = 1$, $h_1(x) = x$, and*

$$h_{k+1}(x) = xh_k(x) - kh_{k-1}(x) .$$

Similarly, The degree- k Hermite polynomial tensor $H_k : \mathbb{R}^d \rightarrow (\mathbb{R}^d)^{\otimes k}$ is similarly defined recursively by $H_0 = 1$, $H_1(x) = x$, and

$$H_{k+1}(x) = \text{Sym}(x \otimes H_k(x)) - k\text{Sym}(I \otimes H_{k-1}(x)) . \quad (4)$$

The Hermite polynomials have several important properties, some of which we record below:

Claim 5.7 (Properties of Hermite polynomials). *Fix any integer $k \geq 0$. Then, we have the following:*

- *For any $x, y \in \mathbb{R}$, we have that*

$$h_k(x + y) = \sum_{j=0}^k \binom{k}{j} y^{k-j} h_j(x) .$$

- *For any $v \in \mathbb{R}^d$, we have that $\langle H_k(x), v^{\otimes k} \rangle = h_k(\langle v, x \rangle)$.*
- *Let $\mu \in \mathbb{R}$. Then $\mathbb{E}_{X \sim N(\mu, 1)}[h_k(X)] = \mu^k$. Similarly, for $\mu \in \mathbb{R}^d$, we have that*

$$\mathbb{E}_{X \sim N(\mu, I)}[H_k(X)] = \mu^{\otimes k} .$$

- *For all k, ℓ , we have that*

$$\mathbb{E}_{X \sim N(0, 1)}[h_k(x)h_\ell(x)] = k! \cdot \delta_{k=\ell} .$$

The following is a useful identity for the second moment of Hermite polynomial tensor, which is a special case of the formula derived in [Kan21] (see also Claim 9.6 of [LL22] for this specialization):

Lemma 5.8. *For all $k \geq 0$, and for all $\mu \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{X \sim N(\mu, I)}[H_k(X) \otimes H_k(X)] = \sum_{S_1, S_2 \subset [k], |S_1|=|S_2|} \sum_{\substack{\text{Matchings } P \\ \text{of } S_1, S_2}} \bigotimes_{\{a,b\} \in P} I^{(a,k+b)} \bigotimes_{c \notin S_1} \mu^{(c)} \bigotimes_{c \notin S_2} \mu^{(k+c)} . \quad (5)$$

An important implication is the following pair of second moment bounds for Hermite polynomials under Gaussians with non-zero mean:

Lemma 5.9. *For any $\mu \in \mathbb{R}$ and integer $k \geq \mu^2$, we have*

$$\mathbb{E}_{X \sim N(0, 1)} h_k^2(X + \mu) = k! \cdot \exp(\Theta(\sqrt{k}|\mu|))$$

Proof. By explicit calculation, we have:

$$\begin{aligned}
\frac{1}{k!} \cdot \mathbb{E}_{X \sim N(0,1)} h_k^2(X + \mu) &= \sum_{j=0}^k \left(\binom{k}{j} \mu^{k-j} \right)^2 \cdot \frac{j!}{k!} \\
&= \sum_{j=0}^k \frac{k!}{((k-j)!)^2 \cdot j!} \mu^{2(k-j)} \\
&= \sum_{j=0}^k \frac{k!}{(j!)^2 \cdot (k-j)!} \mu^{2j} \\
&= \sum_{j=0}^k \binom{k}{j} \frac{1}{j!} \cdot \mu^{2j}.
\end{aligned}$$

To show the upper bound, since $\binom{k}{j} \leq \frac{k^j}{j!}$, we have that

$$\frac{1}{k!} \cdot \mathbb{E}_{X \sim N(0,1)} h_k^2(X + \mu) \leq \sum_{j=0}^k \frac{(k\mu^2)^j}{(j!)^2} \leq I_0(2\sqrt{k}|\mu|).$$

where $I_0(x) = \sum_{j=0}^{\infty} \frac{(\frac{1}{4}x^2)^j}{(j!)^2}$ is the modified Bessel function of the first kind. Then, since $I_0(x) = (1 + o(1)) \frac{e^x}{\sqrt{2\pi x}}$ (see e.g. 10.30 of [Olv10]), this implies the upper bound.

To obtain the lower bound, we let $j = \lceil \frac{\sqrt{k}|\mu|}{2} \rceil (\leq k/2)$. Then, we have that

$$\binom{k}{j} \frac{1}{j!} \mu^{2j} \geq \frac{(k-j)^j \cdot \mu^{2j}}{j^{2j}} \geq \left(\frac{k\mu^2}{2j^2} \right)^j \geq 2^j = \exp(\Omega(\sqrt{k}|\mu|)),$$

which gives us the lower bound. □

Lemma 5.10.

$$\left\| \mathbb{E}_{X \sim N(\mu, I)} [H_k(X) \otimes H_k(X)] \right\|_2 = k! \exp(O(\sqrt{k} \|\mu\|)).$$

Proof. By Lemma 5.8,

$$\mathbb{E}_{X \sim N(\mu, I)} [H_k(X) \otimes H_k(X)] = \sum_{S_1, S_2 \subset [k], |S_1|=|S_2|} \sum_{\text{Matchings } P \text{ of } S_1, S_2} \bigotimes_{\{a,b\} \in P} I^{(a,k+b)} \bigotimes_{c \notin S_1} \mu^{(c)} \bigotimes_{c \notin S_2} \mu^{(k+c)}.$$

Taking the operator norm on both sides, we get

$$\begin{aligned}
\left\| \mathbb{E}_{X \sim N(\mu, I)} [H_k(X) \otimes H_k(X)] \right\| &\leq \sum_{j=0}^k \sum_{S_1, S_2 \subset [k], |S_1|=|S_2|=j} \sum_{\text{Matchings } P \text{ of } S_1, S_2} \|\mu\|^{2(k-j)} \\
&= \sum_{j=0}^k \binom{k}{j}^2 j! \cdot \|\mu\|^{2(k-j)}
\end{aligned}$$

which is same to the term in [Lemma 5.9](#) up to factor $k!$, so we get this is upper bounded by $k! \cdot \exp(\Theta(\sqrt{k} \|\mu\|))$ as desired. $\square$

5.1.2 Overview

Overview of Algorithm Let $k = 100 \log^2(1/\alpha)/\varepsilon^2$, $\gamma = e^{-100k}$ and $\lambda = e^{-10k}$. Recall that H_t denotes the t -th Hermite polynomial tensor. For any set $T \subseteq S$, we define

$$M_t(T) := \frac{1}{|T|} \sum_{i \in T} H_t(X_i) .$$

Our algorithm proceeds in two steps. First, we perform an iterative filtering algorithm to obtain a subset T of S with the guarantee that (i) we have removed essentially no good points from T , and (ii) $\|M_t(T)\|_2 \leq \frac{1}{\gamma} (C \sqrt{\log(1/\alpha)})^t$, for all $t \in [2k]$.

Once we have done this, we show that a simple spectral method applied on a flattening of the moment tensors allows us to extract the desired subspace. Formally, observe that every M_t can be written naturally as $(d \times d^{t-1})$ -sized matrix by flattening the first tensor mode, and denote this matrix $\text{flat}(M_t(T))$. Our algorithm will simply take the top left singular vectors of $\text{flat}(M_t(T))$ for all $t = 1, \dots, 2k$, and take the span of their union. We provide formal pseudocode for this algorithm in [Algorithm 1](#). In the remaining proof, may use M_t as shorthand for $M_t(T)$.

Algorithm 1 Subspace learning algorithm

Let $T = [n]$.

while there exists a $t \in \{1, 2, \dots, 2k\}$ such that $\|M_t(T)\|_2 > \gamma^{-3} t! (C \sqrt{\log(1/\alpha)})^t$ **do**

For every $i \in T$, let

$$\tau_i = \max \left(0, \langle H_t(X_i), M_t(T) \rangle + \gamma^{-2} t! \left(C \sqrt{\log 1/\alpha} \right)^t \cdot \|M_t\|_2 \right) .$$

Sample $i \propto \tau_i$, and remove i from T .

end while

for $t = 1, 2, \dots, 2k$ **do**

Let Q_t denote the set of the s_t left singular vectors of $\text{flat}(M_t(T))$, where

$$s_t = \gamma^{-6} (t!)^2 (C \sqrt{\log(1/\alpha)})^{2t} / \lambda^2 .$$

end for

Return

$$V = \text{span} \left(\bigcup_{i=1}^{2k} Q_t \right) .$$

Regularity condition Before we analyze the algorithm, it will be convenient to establish the following regularity condition on the good set of points. Recall that by assumption, we know that there exists a set $S_{\text{good}} \subset S$ so that $|S_{\text{good}}| = \alpha n$, and S_{good} is a set of independent draws from $N(\mu, I)$. We will require the following lemma, which follows by [Lemma A.1](#), [Lemma 5.10](#) (for [Eq. \(6\)](#)) and [Lemma A.2](#) (for [Eqs. \(7\)](#) and [\(8\)](#)).

Lemma 5.11. *Let $X_1, \dots, X_m \sim N(\mu, I)$, and let k be a positive integer. Then, for*

$$n \geq (dk\sqrt{\log(1/\alpha)})^{\Omega(k)},$$

with probability 0.99 the following holds: for all $t \leq 2k$, and for all $u \in \mathbb{R}^d$, we have that

$$\frac{1}{m} \sum_{i=1}^m \langle u, H_k(X_i) \rangle^2 \leq t! \cdot C^t \cdot \exp\left(\sqrt{t} \|\mu\|_2\right) \cdot \|u\|_2^2, \quad (6)$$

for some universal constant C , and for all unit vector w , we have that

$$\frac{1}{m} \sum_{i=1}^m h_k^4(\langle X_i, w \rangle) \leq \mathbb{E}_{X \sim N(\mu, I)} [h_k^4(\langle X, w \rangle)] + O(1) \quad (7)$$

$$\frac{1}{m} \sum_{i=1}^m h_k^2(\langle X_i, w \rangle) \geq \mathbb{E}_{X \sim N(\mu, I)} [h_k^2(\langle X, w \rangle)] - O(1) \quad (8)$$

For the rest of the proof, condition on the event that Lemma 5.11 holds for S_{good} . By our condition on n , this happens with probability at least 0.999.

5.1.3 Analysis of the filter

First, observe that the filtering routine clearly runs in time which is polynomial in n , since each loop runs in time $\text{poly}(n, d^k) = \text{poly}(n)$, and we will run for at most n loops, since we remove one data point each time we run the filter.

Our main result for the filtering subroutine is the following:

Lemma 5.12. *For the output set T , we have $\|M_t(T)\|_2 \leq \frac{1}{\gamma} (C\sqrt{\log(1/\alpha)})^t$ and with probability $1 - \delta$, we have $|S_{\text{good}} \cap T| \geq (1 - \delta)|S_{\text{good}}|$, where $\delta = \frac{\gamma^{1/2}}{\alpha} = e^{-\Omega(k)}$.*

Proof. By the termination condition of the filter, we have that $\|M_t(T)\|_2 \leq \frac{1}{\gamma} (C\sqrt{\log(1/\alpha)})^t$.

Thus, it remains to prove the second statement. We first observe that for any iteration for t , we have

$$\begin{aligned} \frac{1}{|T|} \sum_{i \in T} \tau_i &\geq \frac{1}{|T|} \sum_{i \in T} \langle H_t(X_i), M_t(T) \rangle \\ &= \|M_t(T)\|_2^2. \end{aligned}$$

On other other hand, we claim that

$$\frac{1}{|T|} \sum_{i \in S_{\text{good}}} \tau_i \leq \gamma \|M_t(T)\|_2^2, \quad (9)$$

where for simplicity, we say that $\tau_i = 0$ for any i that has already been removed from the dataset. Given this, we have that

$$\frac{\sum_{i \in S_{\text{good}}} \tau_i}{\sum_{i \in T} \tau_i} \leq \gamma,$$

which implies that if we define the potential function

$$\Phi := |T| + \frac{1}{\gamma}(|S_{\text{good}}| - |S_{\text{good}} \cap T|),$$

then this potential function is a super-martingale. Therefore, with probability $1 - \gamma^{1/2}$, at termination, $\Phi \leq \gamma^{-1/2}n$, which implies that $|S_{\text{good}} \cap T| \geq (1 - \frac{\gamma^{1/2}}{\alpha})|S_{\text{good}}|$, as claimed. Thus it suffices to prove (9).

First, by Jensen's inequality, we observe that

$$\begin{aligned} \left| \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} \langle M_t, H_t(X_i) \rangle \right| &\leq \left(\frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} \langle M_t, H_t(X_i) \rangle^2 \right)^{1/2} \\ &\leq \left(t! C^t \exp(\sqrt{t} \|\mu\|_2) \right)^{1/2} \|M_t\|_2 \\ &\ll \frac{\gamma}{10} \|M_t\|_2^2. \end{aligned}$$

Let A_t denote the set of points in S_{good} satisfying $\langle M_t, H_t(X_i) \rangle \leq -\gamma^{-2}t! \left(C\sqrt{\log 1/\alpha} \right)^t \|M_t\|_2$. By Chebyshev's inequality with Lemma 5.11,

$$\begin{aligned} \frac{|A_t|}{|S_{\text{good}}|} &\leq \frac{t! C^t \exp(\sqrt{t} \|\mu\|_2) \cdot \|M_t\|_2^2}{\left(\gamma^{-2}t! (C\sqrt{\log 1/\alpha})^t \right)^2 \|M_t\|_2^2} \\ &\ll \gamma^3, \end{aligned}$$

by our choice of γ . In particular, this implies that

$$\begin{aligned} \left| \frac{1}{|S_{\text{good}}|} \sum_{i \in A_t} \langle M_t, H_t(X_i) \rangle \right| &\leq \gamma^{1.5} \cdot \left(t! C^t \exp(\sqrt{t} \|\mu\|_2) \right)^{1/2} \|M_t\|_2 \\ &\ll \frac{\gamma}{10} \|M_t\|_2^2. \end{aligned}$$

Thus, overall, we have that

$$\begin{aligned} \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} \tau_i &= \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}} \setminus A_t} \langle M_t, H_t(X_i) \rangle + \gamma^{-2}t! (C\sqrt{\log 1/\alpha})^t \|M_t\|_2 \\ &\leq \frac{\gamma}{5} \|M_t\|_2^2 + \gamma^{-2}t! (C\sqrt{\log 1/\alpha})^t \|M_t\|_2 \\ &\ll \gamma \|M_t\|_2^2, \end{aligned}$$

as claimed. This completes the proof. $\square$

5.1.4 Analysis of the spectral truncation

First, notice that since $\|M_t(T)\|_2^2 \leq \gamma^{-6}t!(C\sqrt{\log(1/\alpha)})^{2t}$, this implies that the $(s_t + 1)$ -th singular value of $\text{flat}(M_t(T))$ is at most λ . By the definition of Q_t , for any $1 \leq t \leq 2k$, we have for any $u \in \mathbb{R}^d : u \perp V$ and $H \in \mathbb{R}^{d^{t-1}}$,

$$\langle M_t(T), u \otimes H \rangle \leq \lambda \|u\|_2 \|H\|_2.$$

We next require the following lemma:

Lemma 5.13. *For any unit vector $w \in \mathbb{R}^d$ that satisfies $w \perp V$, we have*

$$\frac{1}{|T|} \sum_{i \in T} h_k^2(\langle X_i, w \rangle) \leq 2 \mathbb{E}_{x \sim N(0,1)} [h_k^2(x)] = 2 \cdot k!.$$

Proof. We decompose $h_k^2(x)$ by Hermite polynomials: $h_k^2(x) = \sum_{t=0}^{2k} c_t h_t(x)$, then

$$c_0 = \mathbb{E}_{x \sim N(0,1)} [h_k^2(x)] = k!.$$

Hence

$$\begin{aligned} \frac{1}{|T|} \sum_{i \in T} h_k^2(\langle X_i, w \rangle) &= \frac{1}{|T|} \sum_{i \in T} \left(c_0 + \sum_{t=1}^{2k} c_t h_t(\langle X_i, w \rangle) \right) \\ &= k! + \sum_{t=1}^{2k} \frac{c_t}{|T|} \sum_{i \in T} h_t(\langle X_i, w \rangle) \\ &= k! + \sum_{t=1}^{2k} \frac{c_t}{|T|} \sum_{i \in T} \langle H_t(X_i), w^{\otimes t} \rangle && \text{(By Claim 5.7)} \\ &= k! + \sum_{t=1}^{2k} c_t \langle M_t(T), w^{\otimes t} \rangle \\ &\leq k! + \lambda \sum_{t=1}^{2k} c_t. && \text{(by } w \perp V) \end{aligned}$$

We have following bound on $\sum_{t=1}^{2k} c_t$:

$$\begin{aligned} \left(\sum_{t=1}^{2k} c_t \right)^2 &\leq \sum_{t=1}^{2k} \frac{1}{t!} \cdot \sum_{t=1}^{2k} (t!) c_t^2 && \text{(Cauchy-Schwarz)} \\ &\leq e \mathbb{E}_{x \in N(0,1)} \left[\left(\sum_{t=1}^{2k} c_t h_t(x) \right)^2 \right] \\ &\leq 3 \mathbb{E}_{x \in N(0,1)} [h_k^4(x)] \\ &\leq 3 \cdot 3^{2k} \left(\mathbb{E}_{x \in N(0,1)} [h_k^2(x)] \right)^2 = 3^{2k+1} (k!)^2, && \text{(By hypercontractivity)} \end{aligned}$$

and so the lemma follows by our choice of $\lambda = e^{-10k}$. $\square$

Now we are ready to prove the main theorem of this section:

Proof of Theorem 5.2. Suppose the distance of μ to V is larger than ε . Let w be the unit vector parallel to $(\Pi_{V^\perp})\mu$.

On the one hand, by Lemma 5.13, we know that

$$\frac{1}{|T|} \sum_{i \in T} h_k^2(\langle X_i, w \rangle) \leq 2 \cdot k!.$$

On the other hand, we also have that:

$$\frac{1}{|T|} \sum_{i \in T} h_k^2(\langle X_i, w \rangle) \geq \frac{\alpha}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}} \cap T} h_k^2(\langle X_i, w \rangle)$$

Since

$$\begin{aligned} \mathbb{E}_{i \sim S_{\text{good}}} [\mathbf{1}_{i \notin S_{\text{good}} \cap T} \cdot h_k^2(\langle X_i, w \rangle)] &\leq \left(\mathbb{E}_{i \sim S_{\text{good}}} [\mathbf{1}_{i \notin S_{\text{good}} \cap T}] \right)^{\frac{1}{2}} \cdot \left(\mathbb{E}_{i \sim S_{\text{good}}} [h_k^4(\langle X_i, w \rangle)] \right)^{\frac{1}{2}} \\ &\quad \text{(Cauchy-Schwarz)} \\ &\leq \delta^{1/2} \cdot \left(\mathbb{E}_{X \sim N(\mu, I)} [h_k^4(\langle X, w \rangle)] + O(1) \right)^{\frac{1}{2}} \quad \text{(By Eq. (7))} \\ &\leq \delta^{1/2} \cdot \left(\mathbb{E}_{x \sim N(0, 1)} h_k^4(x + y) \right)^{\frac{1}{2}} + O(1) \\ &\leq \delta^{1/2} \cdot 3^k \mathbb{E}_{x \sim N(0, 1)} h_k^2(x + y) + O(1) \quad \text{(By hypercontractivity)} \\ &\leq O(1), \end{aligned}$$

and so we obtain that

$$\begin{aligned} \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}} \cap T} h_k^2(\langle X_i, w \rangle) &\geq \mathbb{E}_{i \sim S_{\text{good}}} [h_k^2(\langle X_i, w \rangle)] - O(1) \\ &\geq \mathbb{E}_{X \sim N(\mu, I)} [h_k^2(\langle X, w \rangle)] - O(1) \quad \text{(By Eq. (8))} \\ &= \mathbb{E}_{x \sim N(0, 1)} h_k^2(x + \langle \mu, w \rangle) - O(1) \\ &\geq \exp(\Omega(\sqrt{k}\varepsilon)) \cdot k! - O(1) \geq (3/\alpha) \cdot (k!), \quad \text{(By Lemma 5.9)} \end{aligned}$$

which implies that

$$\frac{1}{|T|} \sum_{i \in T} h_k^2(\langle X_i, w \rangle) \geq 3 \cdot k!,$$

which contradicts to the upper bound $2 \cdot (k!)$. This completes the proof. $\square$

5.2 Exhaustive search for moment-matching candidates

Suppose we receive an α -pure set of n samples where the inliers are distributed according to $N(\mu, I)$, and we know a bound $\|\mu\| \leq r$. In this section, we will focus on how to construct a set $\mu_1, \dots, \mu_l$ such that some μ_i is ε -close to μ , all μ_i have a subset of αn points where the centered moments are close to $N(0, I)$, and all pairs μ_i, μ_j are $\varepsilon/2$ -separated.

Our ingredients from other sections will later imply that such a list has bounded size. Further, when we invoke the following theorem, we will be invoking it for a learned subspace of small dimension, so the dependence on d is not as concerning as it may appear at first glance.

Theorem 5.14. *Suppose there are n samples $X_1, \dots, X_n \in \mathbb{R}^d$, αn of which are sampled according to $X_i \sim N(\mu, I)$. Consider parameters $0 < \alpha, \delta, \Delta < 1$, $\|\mu\| \leq r$, $0 < \varepsilon \leq r$, and a positive integer*

k^* . There exists an algorithm that outputs a list $\mu_1, \dots, \mu_l$ such that the μ_i are $\varepsilon/2$ -separated, each μ_i has a corresponding vector $w_1^{(i)}, \dots, w_n^{(i)} \in [0, 1]$ such that $\sum_{j=1}^n w_j^{(i)} = \alpha n$, and

$$\left\| \frac{1}{\alpha n} \sum_{i=1}^n w_j^{(i)} \cdot (X_j - \mu_i)^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \leq \Delta \quad \forall k \in \{1, \dots, k^*\}.$$

Moreover, with probability at least $1 - \delta$, there exists a $\mu^* \in \mu_1, \dots, \mu_l$ such that $\|\mu - \mu^*\|_2 \leq \varepsilon$. The algorithm has running time $O(\frac{r}{\varepsilon \Delta})^d \cdot (2dk^*)^{Cdk^*} \cdot (\frac{1}{\alpha \delta \Delta})^C$ and uses $n = \lceil \frac{4(k^*)^{1.5}(8dk^*)^{k^*}}{\alpha \delta \Delta} \rceil$ samples.

Proof. Overview. Our plan is to consider an ε' -cover $\mathcal{N}$ of the ball $\mathcal{B}$ in which μ is restricted, where $\varepsilon' \ll \varepsilon/2$. For each center in the cover, we will check whether there exists a subset of samples where the moments around the center look sufficiently close to Gaussian (in terms of the Frobenius norm; this may be computed efficiently, we discuss later). For all possible centers that pass this test, we will iterate in an arbitrary order, and greedily add a center to our set if it is not $\varepsilon/2$ -close to any center already in our set. Clearly, our final set will consist of centers where their moments satisfy our condition, and the centers are $\varepsilon/2$ -separated. What remains is to show that one of these centers will be ε -close to the true μ . Let μ^* be an arbitrary point in $\mathcal{N}$ that is ε' -close to μ . We will show that with probability at least $1 - \delta$, μ^* will satisfy the moment conditions. Hence, either μ^* will be in our set, or μ^* is $\varepsilon/2$ -close to an item in our set. Since μ^* itself is ε' -close to μ (where $\varepsilon' \leq \varepsilon/2$), this immediately implies μ is ε -close to an item in our set.

ε' -cover construction. Let us define a simple ε' -cover $\mathcal{N}$ for $\mathcal{B}$. Since each coordinate is within a bounded range of width $2r$, for each coordinate let us consider possible values $\{-r, \dots, -2\varepsilon'/\sqrt{d}, -\varepsilon'/\sqrt{d}, 0, \varepsilon'/\sqrt{d}, 2\varepsilon'/\sqrt{d}, \dots, r\}$; there are at most $1 + 2\lceil r\sqrt{d}/\varepsilon' \rceil \leq 5r\sqrt{d}/\varepsilon'$ options. We will choose our net $\mathcal{N}$ to be the set of all $\leq (5r\sqrt{d}/\varepsilon')^d$ options that are within the ball. This is a valid ε' net, because for any $p \in \mathcal{B}$, consider the point p' where each coordinate is rounded to the nearest value in the net that is closer to 0: we observe $p' \in \mathcal{B}$ and $\|p - p'\|_2 \leq \varepsilon'$. Enumerating over this net takes $O(r\sqrt{d}/\varepsilon')^d$ time.

Moment-checking procedure. For any fixed $\mu' \in \mathcal{N}$, we hope to find a collection of values $w_1, \dots, w_n \in [0, 1]$ such that $\sum_i w_i = \alpha n$, and the following condition holds for the moment tensors:

$$\left\| \frac{1}{\alpha n} \sum_{i=1}^n w_i \cdot (X_i - \mu')^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \leq \Delta \quad \forall k \in \{1, \dots, k^*\}$$

This condition would be implied by

$$\sum_{k=1}^{k^*} \left\| \frac{1}{\alpha n} \sum_{i=1}^n w_i \cdot (X_i - \mu')^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F^2 \leq \Delta^2. \quad (10)$$

This condition is quite convenient to check: the constraints on w are linear constraints, and the objective is simply a least-squares regression task, so we may run in time $\text{poly}(n, d^{k^*})$ for a fixed μ' .

Showing μ^* passes the moment-checking procedure. Recall μ^* is an arbitrary point in our set $\mathcal{N}$ that is ε' -close to μ . We desire to show that with probability at least $1 - \delta$, μ^* will pass the moment-checking procedure. Our plan is to show that choosing the w_i corresponding to the αn inliers sampled from $N(\mu, I)$ will pass the test. Our moment-checking condition will certainly pass if for all $k \in \{1, \dots, k^*\}$, it holds:

$$\left\| \frac{1}{\alpha n} \sum_{i \in \alpha n \text{ inliers}} (X_i - \mu^*)^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \leq \Delta/\sqrt{k^*} \quad (11)$$

where $X_i \sim N(\mu, I)$. This is equivalent to the following condition for $Y_1, \dots, Y_{\alpha n} \sim N(\mu - \mu^*, I)$:

$$\left\| \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} Y_i^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \leq \Delta / \sqrt{k^*} \quad (12)$$

The following technical lemma allows us to show this condition (Eq. (12)) occurs with high probability when $\|\mu - \mu^*\|$ is small enough (proof deferred to Section A.5).

Lemma 5.15. *For parameters $0 < \delta, \gamma < 1$, suppose it holds that $m \geq \frac{4(8dk)^k}{\delta\gamma}$ and $\|\mu\|_\infty \leq \frac{\gamma}{2(4dk)^{k/2}}$. Then,*

$$\Pr_{X_1, \dots, X_m \sim N(\mu, I)} \left[\left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \geq \gamma \right] \leq \delta.$$

This immediately lets us conclude that Eq. (12) holds with probability at least $1 - \delta$:

Corollary 5.16. *The moment-checking condition will hold for μ^* with probability at least $1 - \delta$ if $n \geq \frac{4(k^*)^{1.5}(8dk^*)^{k^*}}{\alpha\delta\Delta}$ and $\varepsilon' \leq \frac{\Delta}{2\sqrt{k^*}(4dk^*)^{k^*/2}}$.*

Proof. We invoke Lemma 5.15 for each $k \in \{1, \dots, k^*\}$ with $\gamma = \Delta/\sqrt{k^*}$, $\delta' = \delta/k^*$, $\|\mu\|_\infty \leq \varepsilon'$, and $m = \alpha n$. The first condition holds since

$$m \geq \frac{4(8dk)^k}{\delta'\gamma} \iff n \geq \frac{4(k^*)^{1.5}(8dk^*)^{k^*}}{\alpha\delta\Delta}.$$

The second condition holds since

$$\|\mu\|_\infty \leq \frac{\gamma}{2(4dk)^{k/2}} \iff \varepsilon' \leq \frac{\Delta}{2\sqrt{k^*}(4dk^*)^{k^*/2}}.$$

□

Concluding. Hence, our algorithm succeeds with probability at least $1 - \delta$ with our choice of parameters $\varepsilon' = \min(\varepsilon/2, \frac{\Delta}{2\sqrt{k^*}(4dk^*)^{k^*/2}})$ and $n = \lceil \frac{4(k^*)^{1.5}(8dk^*)^{k^*}}{\alpha\delta\Delta} \rceil$. Our running time is $O(r\sqrt{d}/\varepsilon')^d \cdot \text{poly}(n, d^{k^*}) = O(\frac{r}{\varepsilon\Delta})^d \cdot (2dk^*)^{Cdk^*} \cdot (\frac{1}{\alpha\delta\Delta})^C$. This concludes the proof of our theorem. □

5.3 Sharper moment-matching fooling

The guarantees from Theorem 5.6 of [GKK23] are lossy in the setting where moments are subgaussian and the number of halfspaces are super-constant, as is the case in our application. We provide a sharper result (proof deferred to Section A.6):

Lemma 5.17. *Suppose two distributions G, T over $\mathbb{R}^d$ match moments up to order k (which is even), meaning*

$$\left| \mathbb{E}_G[\langle G, v \rangle^i] - \mathbb{E}_T[\langle T, v \rangle^i] \right| \leq \Delta$$

for all $\|v\|_2 = 1$ and $i \in \{1, \dots, k\}$. Moreover, suppose G has subgaussian moments where

$$\mathbb{E}_G[\langle G, v \rangle^i] \leq (C_1 \sqrt{i})^i$$

and its linear projections are anticoncentrated such that

$$\mathbb{E}_G[\langle G, v \rangle \in [L, R]] \leq C_2 \cdot (R - L).$$

for constants $C_1, C_2 > 0$. Consider any collection of halfspaces $H_1, \dots, H_m$ where $H_i \triangleq \mathbb{1}[\langle x, a_i \rangle \geq b_i]$. There exist constants $C_3, C_4 \geq 1$ where if $k \geq C_3 \cdot \max(m^6/\varepsilon^2, \log^8(1/\varepsilon)/\varepsilon^2)$, $\Delta \leq \left(\frac{1}{C_4 k}\right)^k$, and $0 < \varepsilon \leq 1/2$, then for the intersection $H \triangleq \bigcap_{i=1}^m H_i$ it holds

$$|\Pr_G[X \in H] - \Pr_T[X \in H]| \leq C\varepsilon.$$

5.4 Combining ingredients: proof of Theorem 5.1

Let us recall our general plan. First, we will use prior work of [DKK⁺22a] to reduce the task to $O(1/\alpha)$ subproblems where $\|\mu\|_2 \leq O(\sqrt{\log(1/\alpha)})$. Next, for each subproblem, we will learn the low-dimensional subspace via Section 5.1. Afterwards, we will exhaustively search over a net within this subspace via Section 5.2, where this subroutine returns μ_i satisfying the moment matching condition, and one of them is close to μ , yet we have not bounded the size of this output list. Finally, we will combine our moment-matching fooling result in Section 5.3 with the ideas in our information-theoretic proof of Theorem 4.1 to prove the list must have bounded size. We now combine all the ingredients.

Throughout this algorithm, we will need three groups of samples of size n_1, n_2, n_3 such that $n = n_1 + n_2 + n_3$. We will randomly distribute our n samples into these three groups, and since n is sufficiently large, the number of good inliers in each group will be at least $\alpha n_i/2$ with high probability. For the remaining description of our algorithm subroutines, we will leverage $\alpha/2$ as our inlier parameter.

Using prior work of Diakonikolas, Kane, Karmalkar, Pensia, and Pittas [DKK⁺22a]. For any constant $\delta > 0$, Theorem 1.2 of [DKK⁺22a] outputs a list of size $O(1/\alpha)$, that with probability at least $1 - \delta$ contains a point within distance $C\sqrt{\log(1/\alpha)}$ of the true μ . Moreover, this algorithm runs in time $(d \log(1/\alpha))^{C \log(1/\alpha)}$ and uses $n_1 = (d \log(1/\alpha))^{C \log(1/\alpha)}$ samples. From this output list, we will focus on $O(1/\alpha)$ subproblems, where after translation we may assume that $\|\mu\|_2 \leq C\sqrt{\log(1/\alpha)}$. We now focus on each of the subproblems separately.

Learning a low-dimensional subspace. We invoke Theorem 5.2 with $\alpha' = \alpha/2$ and $\varepsilon' = \varepsilon/2$. This gives a subspace with dimension at most $\exp(C \log^2(1/\alpha)/\varepsilon^2 \cdot (\log(\log(1/\alpha)/\varepsilon)))$, where μ is within $\varepsilon/2$ of the subspace with constant probability. We may run this constantly many times on different samples to boost the failure probability to any small constant, while keeping the subspace dimension at most $\exp(C \log^2(1/\alpha)/\varepsilon^2 \cdot (\log(\log(1/\alpha)/\varepsilon)))$. Learning the subspace has sample/time complexity $(d \log(1/\alpha)/\varepsilon)^{C \log^2(1/\alpha)/\varepsilon^2}$.

Exhaustively searching within the low-dimensional subspace. We now invoke Theorem 5.14 to search the low-dimensional subspace. We may use $d' = (C \log^2(1/\alpha)/\varepsilon^2)^{C \log^2(1/\alpha)/\varepsilon^2}$, $\varepsilon = \varepsilon'/2$, $r \leq C\sqrt{\log(1/\alpha)}$, and δ to be a sufficiently small constant. We defer the choice of Δ and k^* until later. All n_3 samples will be input to the algorithm after projecting to the subspace. Observe how with probability at least $1 - \delta$, the list contains a μ^* in the subspace where $\|\mu^* - \Pi_V \mu\|_2 \leq \varepsilon/2$, and hence $\|\mu - \mu^*\| \leq \varepsilon$. Moreover, all μ_i are $\varepsilon/4$ -separated. The main remaining question is to bound the size of this list.

Bounding the list size. Let $\mu_1, \dots, \mu_l$ denote the output list from some call to the exhaustive search subroutine. We will show how the list size must be bounded by using similar ideas to

our information-theoretic proof [Theorem 4.1](#), in addition to our moment-matching fooling technical result [Lemma 5.17](#). Let $r_0 = C\sqrt{\log(1/\alpha)}$ be the bound on the error from the algorithm of [\[DKK⁺22a\]](#). For non-negative integer i , let $|M_i|$ denote the minimum size $r_i \triangleq \frac{r_0}{2^i}$ -cover of $\mu_1, \dots, \mu_l$. If we choose Δ and k^* appropriately, we may bound the ratio of consecutive $|M_i|/|M_{i-1}|$. Eventually, for some i^* where $r_{i^*} < \varepsilon/8$, we may conclude $|M_{i^*}| = l$ because all μ_i are $\varepsilon/4$ -separated. Since the ratio bound will let us bound $|M_{i^*}|$, this will in turn bound l . We now prove the ratio bound:

Claim 5.18. *For any positive integer i where $r_i \geq \frac{\varepsilon}{100}$, it holds that $\frac{|M_i|}{|M_{i-1}|} \leq \exp(C \log^2(1/\alpha)/r_i^2)$.*

Proof. This proof mirrors the proof of [Claim 4.4](#), except we must adjust that our μ_i only satisfy moment-matching conditions, instead of the α -consistent property.

We will construct a cover M_i by considering covering each r_{i-1} -radius ball of M_{i-1} separately. The size of the cover for each ball is bounded by the maximum-possible size subset $\mu_1, \dots, \mu_{l'}$ such that all μ_i are r_i -separated. We will prove that such a set cannot exist for l' too large. After translating all such μ_i accordingly, recall all $\|\mu_i\|_2 \leq r_{i-1}$. Let us again define corresponding regions R_i where the density of μ_i is maximal: $R_i \triangleq \{ \|x - \mu_i\| < \|x - \mu_j\| \}$. Let D denote the normalized empirical distribution over the n_3 samples. Similarly, let D_i denote the normalized empirical distribution corresponding to each $w^{(i)}$. Since the total mass of D is 1, we know:

$$\sum_{i=1}^{l'} \Pr_{X \sim D}[X \in R_i] \leq 1 \implies \sum_{i=1}^{l'} \alpha' \Pr_{X \sim D_i}[X \in R_i] \leq 1 \implies \sum_{i=1}^{l'} \frac{\alpha}{2} \Pr_{X \sim D_i}[X \in R_i] \leq 1 \quad (13)$$

We now want to show that $\Pr_{X \sim D_i}[X \in R_i] \approx \Pr_{X \sim N(\mu_i, I)}[X \in R_i]$. Note how we may translate the distribution D_i by $-\mu_i$ (and similarly define R'_i as the region R_i with the same translation), so we instead study

$$|\Pr_{X \sim D_i}[X \in R_i] - \Pr_{X \sim N(\mu_i, I)}[X \in R_i]| = |\Pr_{X \sim (D_i - \mu_i)}[X \in R'_i] - \Pr_{N(0, I)}[R'_i]|$$

This is a quantity that we may bound via our moment-matching fooling technical result of [Lemma 5.17](#). Note how we are aiming to show contradiction for a sufficiently large l' ; we will choose l' so that is always at most $\exp(C \log^2(1/\alpha)/r_i^2)$. We now want to use [Lemma 5.17](#) such that each D_i will fool an intersection of at most $l' \leq \exp(C \log^2(1/\alpha)/r_i^2)$ halfspaces with additive error at most $\frac{1}{l'}$. In the language of [Lemma 5.17](#), $G = N(0, I)$, and H is the distribution D_i translated by $-\mu_i$. The subgaussian moments assumption and anticoncentration assumption follows immediately since $G = N(0, I)$. The lemma implies fooling halfspaces up to error $\frac{1}{l'}$ if we choose $k^* = \exp(C \log^2(1/\alpha)/\varepsilon^2) \geq \exp(C \log^2(1/\alpha)/r_i^2)$ and $\Delta = \exp(C \log^2(1/\alpha)/\varepsilon^2) - \exp(C \log^2(1/\alpha)/\varepsilon^2)$. Using this, we continue:

$$\begin{aligned} \text{Eq. (13)} &\iff \sum_{i=1}^{l'} \frac{\alpha}{2} (\Pr_{X \sim N(\mu_i, I)}[X \in R_i] + (\Pr_{X \sim D_i}[X \in R_i] - \Pr_{X \sim N(\mu_i, I)}[X \in R_i])) \leq 1 \\ &\implies \sum_{i=1}^{l'} \frac{\alpha}{2} \Pr_{X \sim N(\mu_i, I)}[X \in R_i] \leq 1 + \frac{\alpha}{2} \\ &\implies \sum_{i=1}^{l'} \Pr_{X \sim N(\mu_i, I)}[X \in R_i] \leq \frac{4}{\alpha} \end{aligned}$$

$$\begin{aligned}
&\iff \sum_{i=1}^{l'} q_i \leq \frac{4}{\alpha} && \text{(recalling the definition of } q_i\text{)} \\
&\implies \text{poly}(\alpha) \cdot 2^{\frac{cr_i^3}{r_{i-1}^2} \sqrt{\log l'}} \leq \frac{4}{\alpha} && \text{(invoking Theorem 4.3 with } r = r_{i-1} \text{ and } \beta = r_i\text{)} \\
&\iff \text{poly}(\alpha) \cdot 2^{cr_i \sqrt{\log l'}} \leq \frac{4}{\alpha} \\
&\implies l' \leq \exp(C \log^2(1/\alpha)/r_i^2).
\end{aligned}$$

This completes the proof of our claim, since the maximal size r_i -separated set inside each r_{i-1} -radius ball must have size at most $\exp(C \log^2(1/\alpha)/r_i^2)$. $\square$

With this ratio bound in hand, the observation that $l = |M_{i^*}|$, and setting $i^* = \lceil \log_2(r_0/(\varepsilon/16)) \rceil$, we conclude

$$\begin{aligned}
t = |M_{i^*}| &\leq |M_0| \cdot \prod_{i=1}^{i^*} \frac{|M_i|}{|M_{i-1}|} = \prod_{i=1}^{i^*} \frac{|M_i|}{|M_{i-1}|} \\
&\leq \prod_{i=1}^{i^*} \exp(C \log^2(1/\alpha)/r_i^2) \leq \exp(C \log^2(1/\alpha)/r_{i^*}^2) \\
&\leq \exp(C \log^2(1/\alpha)/\varepsilon^2).
\end{aligned}$$

This bounds the size of each of the list returned in each of the C/α subproblems, and in total we attain a list size bound of $(C/\alpha) \cdot \exp(C \log^2(1/\alpha)/\varepsilon^2) \leq \exp(C \log^2(1/\alpha)/\varepsilon^2)$.

Concluding runtime and sample usage. Finally, we revisit runtime and sample complexity now that all parameters have been chosen (namely, Δ and k^*). The algorithm of [DKK⁺22a] still has sample/time complexity of $(d \log(1/\alpha))^{C \log(1/\alpha)}$. Learning the subspace still has sample/time complexity $(d \log(1/\alpha)/\varepsilon)^{C \log^2(1/\alpha)/\varepsilon^2}$. The exhaustive search has sample complexity $\exp(\exp(C \log^2(1/\alpha)/\varepsilon^2))$ and time complexity $\exp(\exp(C \log^2(1/\alpha)/\varepsilon^2 \cdot (\log(\log(1/\alpha)/\varepsilon))))$. In total, the sample complexity is $d^{C \log^2(1/\alpha)/\varepsilon^2} + \exp(\exp(C \log^2(1/\alpha)/\varepsilon^2))$, and the time complexity is $d^{C \log^2(1/\alpha)/\varepsilon^2} + \exp(\exp(C \log^2(1/\alpha)/\varepsilon^2 \cdot (\log(\log(1/\alpha)/\varepsilon))))$.

References

- [Ans60] Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.
- [Ban38] Stefan Banach. Über homogene polynome in (l^2) . *Studia Mathematica*, 7(1):36–44, 1938.
- [Baz09] Louay MJ Bazzi. Polylogarithmic independence can fool dnf formulas. *SIAM Journal on Computing*, 38(6):2220–2272, 2009.
- [BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680, 2008.
- [BK21] Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1279–1297. SIAM, 2021.

- [Bra08] Mark Braverman. Polylogarithmic independence fools ac 0 circuits. *Journal of the ACM (JACM)*, 57(5):1–10, 2008.
- [BS71] Jacek Bochnak and Józef Siciak. Polynomials and multilinear mappings in topological vector-spaces. *Studia Mathematica*, 39(1):59–76, 1971.
- [CLC25] Syomantak Chaudhuri, Jerry Li, and Thomas A Courtade. Robust estimation under heterogeneous corruption rates. *arXiv preprint arXiv:2508.15051*, 2025.
- [CP23] Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1713–1726, 2023.
- [CR19] Daniel Carando and Jorge Tomás Rodríguez. Symmetric multilinear forms on hilbert spaces: Where do they attain their norm? *Linear Algebra and its Applications*, 563:178–192, 2019.
- [CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, pages 47–60, 2017.
- [DGJ⁺10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.
- [Din12] Seán Dineen. *Complex analysis on infinite dimensional spaces*. Springer Science & Business Media, 2012.
- [DJKS23] Abhimanyu Das, Ayush Jain, Weihao Kong, and Rajat Sen. Efficient list-decodable regression using batches. In *International Conference on Machine Learning*, pages 7025–7065. PMLR, 2023.
- [DK23] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.
- [DKK⁺19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.
- [DKK⁺21] Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. List-decodable mean estimation in nearly-pca time. *Advances in Neural Information Processing Systems*, 34:10195–10208, 2021.
- [DKK⁺22a] Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pitas. List-decodable sparse mean estimation via difference-of-pairs filtering. *Advances in Neural Information Processing Systems*, 35:13947–13960, 2022.
- [DKK⁺22b] Ilias Diakonikolas, Daniel M Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. Clustering mixture models in almost-linear time via list-decodable mean estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1262–1275, 2022.

- [DKK⁺23] Ilias Diakonikolas, Daniel Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 36:39470–39490, 2023.
- [DKLZ24] Ilias Diakonikolas, Daniel Kane, Sihan Liu, and Nikos Zarifis. Testable learning of general halfspaces with adversarial label noise. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1308–1335. PMLR, 2024.
- [DKN10] Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2010.
- [DKP⁺21] Ilias Diakonikolas, Daniel Kane, Ankit Pensia, Thanasis Pittas, and Alistair Stewart. Statistical query lower bounds for list-decodable linear regression. *Advances in Neural Information Processing Systems*, 34:3191–3204, 2021.
- [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.
- [GKK23] Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1657–1670, 2023.
- [Hub92] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.
- [IK22] Misha Ivkov and Pravesh K Kothari. List-decodable covariance estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1276–1283, 2022.
- [Kan21] Daniel M Kane. Robust learning of mixtures of gaussians. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1258. SIAM, 2021.
- [KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. *Advances in neural information processing systems*, 32, 2019.
- [KKM13] Daniel Kane, Adam Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *Conference on Learning Theory*, pages 522–545. PMLR, 2013.
- [KM13] Adam Klivans and Raghu Meka. Moment-matching polynomials. *arXiv preprint arXiv:1301.0820*, 2013.
- [KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

- [KSTV25] Adam R Klivans, Konstantinos Stavropoulos, Kevin Tian, and Arsen Vasilyan. The power of iterative filtering for supervised learning with (heavy) contamination. *arXiv preprint arXiv:2505.20177*, 2025.
- [KSV24] Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2887–2943. PMLR, 2024.
- [LL22] Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261, 2022.
- [LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.
- [MV18] Michela Meister and Gregory Valiant. A data prism: Semi-verified learning in the small- α regime. In *Conference On Learning Theory*, pages 1530–1546. PMLR, 2018.
- [Olv10] Frank WJ Olver. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.
- [PST07] A Pappas, Y Sarantopoulos, and A Tonge. Norm attaining polynomials. *Bulletin of the London Mathematical Society*, 39(2):255–264, 2007.
- [RV23] Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1643–1656, 2023.
- [RY20] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.
- [STW24] Lucas Slot, Stefan Tiegel, and Manuel Wiedmer. Testably learning polynomial threshold functions. *Advances in Neural Information Processing Systems*, 37:3781–3831, 2024.
- [Sud69] Vladimir N Sudakov. Gaussian measures, cauchy measures and ε -entropy. In *Soviet Math. Dokl*, volume 10, pages 310–313, 1969.
- [Tuk60] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.
- [Tuk75] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.
- [Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

- [ZS22] Shiwei Zeng and Jie Shen. List-decodable sparse mean estimation. *Advances in Neural Information Processing Systems*, 35:24031–24045, 2022.
- [ZS23] Shiwei Zeng and Jie Shen. Semi-verified pac learning from the crowd. In *International Conference on Artificial Intelligence and Statistics*, pages 505–522. PMLR, 2023.

A Deferred proofs

A.1 Proof of [Corollary 1.5](#)

First, we run our list-decodable mean estimation algorithm from [Theorem 5.1](#) and obtain a list $\mu_1, \dots, \mu_l$ where $l \leq L$. Let μ^* be any arbitrary item in the list where $\|\mu - \mu^*\|_2 \leq \varepsilon$. Given this list, the rest of the proof follows from standard ideas.

We now design a subroutine where, given two items μ_i, μ_j such that $\|\mu_i - \mu_j\|_2 \geq 4\varepsilon$, we will output a “winner” between μ_i and μ_j . Our desired guarantee is that for all such pairwise tests, if one of μ_i, μ_j is μ^* , then μ^* will always win.

Our test is simple. For each pairwise comparison, we use our n_2 samples $X_1, \dots, X_{n_2}$, and transform them into $Y_k = \langle X_k - \frac{\mu_i + \mu_j}{2}, \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|_2} \rangle$. Then, our test will output μ_i if the average $\frac{1}{n_2} \sum_{k=1}^{n_2} Y_k > 0$, and μ_j otherwise.

We will show that μ^* wins all its tests with high probability. Without loss of generality, suppose μ^* is μ_i , we prove that $\frac{1}{n_2} \sum_{k=1}^{n_2} Y_k > 0$ with probability at least $\delta/|L|$. Observe how the Y_k are distributed according to $N(\langle \mu - \frac{\mu^* + \mu_j}{2}, \frac{\mu^* - \mu_j}{\|\mu^* - \mu_j\|_2} \rangle, 1)$. The mean of this univariate Gaussian distribution is at least

$$\begin{aligned} \langle \mu - \frac{\mu^* + \mu_j}{2}, \frac{\mu^* - \mu_j}{\|\mu^* - \mu_j\|_2} \rangle &= \langle \mu^* + (\mu - \mu^*) - \frac{\mu^* + \mu_j}{2}, \frac{\mu^* - \mu_j}{\|\mu^* - \mu_j\|_2} \rangle \\ &\geq \langle \frac{\mu^* - \mu_j}{2}, \frac{\mu^* - \mu_j}{\|\mu^* - \mu_j\|_2} \rangle - \varepsilon \\ &= \frac{\|\mu^* - \mu_j\|_2}{2} - \varepsilon \geq \varepsilon \end{aligned}$$

With probability at least $1 - \delta/|L|$, the empirical mean will concentrate within $\varepsilon/2$ of its true mean from $n_2 \geq \frac{C \log(|L|/\delta)}{\varepsilon^2}$ samples. By union bound, μ^* will pass all its tests with probability at least $1 - \delta$.

Finally, we output any μ_i that passes all its tests. In the event where μ^* was undefeated, this means μ_i is within 4ε of μ^* , and hence within 8ε of μ . If we rescale ε by a factor of 8, then we obtain the desired guarantee.

A.2 Proof of [Lemma 4.7](#)

Proof. Denote $\mu(A)$ by $p(\leq 1/2)$. We define $x(\geq 0)$ be the solution of $\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = p$.

Gaussian isoperimetry inequality tells us,

$$\mu(A_\epsilon) \geq \int_{x-\epsilon}^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-(t-\epsilon)^2/2} dt$$

We know that when $p = 1/3$, $x \approx 0.43$, so when $p \geq 1/3$, we have $x \leq 0.44$, and when $p \leq 1/3$, we have $x \geq 0.42$ and by Mills ratio, we know $p \geq \frac{x}{1+x^2} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \geq \frac{e^{-x^2/2}}{10x}$, so $x \geq \Omega(\sqrt{\log(1/p)})$. Therefore in either case we have $x(p) = \Omega(\sqrt{\log(1/p)}) - O(1)$, so

$$\frac{\mu(A_\epsilon)}{\mu(A)} \geq \min_{t \geq x} \frac{e^{-(t-\epsilon)^2/2}}{e^{-t^2/2}} = \frac{e^{-(x-\epsilon)^2/2}}{e^{-x^2/2}} = e^{-\epsilon^2/2} \cdot e^{x\epsilon} = \exp(-\epsilon^2/2 - O(\epsilon) + \Omega(\epsilon\sqrt{\log(1/p)})),$$

which right hand side is $\Omega(1) \exp(\Omega(\epsilon\sqrt{\log(1/p)}))$ when $\epsilon = O(1)$ and is $\exp(-O(\epsilon^2) + \Omega(\epsilon\sqrt{\log(1/p)}))$ when $\epsilon = \Omega(1)$, so the lemma follows. $\square$

A.3 Proof of Lemma 4.8

Proof. Let D_0 be the distribution $\frac{1}{2}N(0, I)$, and let e_i be the d -dimensional vector that has a 1 in the i -th coordinate and 0 for all other coordinates. Additionally, for $i \in \{1, \dots, d\}$, let D_i be the function $D_i(x) = \max(0, \alpha[N(2\epsilon e_i, I)](x) - D_0(x))$.

Suppose the total mass of each D_i is at most $\frac{1}{2d}$. In this case, we could define our distribution D such that its density is at least $D(x) \geq \sum_{i=0}^d D_i(x)$ everywhere, and any valid list must have size at least d since all $2\epsilon e_i$ are α -consistent and $> 2\epsilon$ -separated. Our plan is to analyze the total mass for any D_i , and then choose d accordingly.

First, let us bound the region where D_i is nonzero. This occurs when the ratio $\frac{D_0(x)}{\alpha[N(2\epsilon e_i, I)](x)} < 1$. We analyze this ratio:

$$\begin{aligned} \frac{D_0(x)}{\alpha[N(2\epsilon e_i, I)](x)} &= \frac{1}{2\alpha} e^{(-\|x\|_2^2 + \|x - 2\epsilon e_i\|_2^2)/2} \\ &= \frac{1}{2\alpha} e^{-\langle x, 2\epsilon e_i \rangle + \|2\epsilon e_i\|_2^2/2} \\ &= \frac{1}{2\alpha} e^{-2\epsilon x_i + \epsilon^2/2} \end{aligned}$$

Hence, we may observe conditions under which D_i is nonzero:

$$D_i(x) > 0 \iff x_i > \frac{\epsilon^2/2 + \log(1/2\alpha)}{\epsilon}$$

From here, we may bound the total mass of D_i :

$$\begin{aligned} \int_{\mathbb{R}^d} D_i(x) &\leq \Pr_{X \sim N(2\epsilon e_i, I)} \left[x_i > \frac{\epsilon^2/2 + \log(1/2\alpha)}{\epsilon} \right] \\ &= \Pr_{X \sim N(0,1)} \left[x > \frac{\epsilon^2/2 + \log(1/2\alpha)}{\epsilon} - 2\epsilon \right] \\ &\leq \Pr_{X \sim N(0,1)} \left[x > \frac{\log(1/2\alpha)}{2\epsilon} \right] && (\text{using } \epsilon < \frac{\sqrt{\log(1/2\alpha)}}{2}) \\ &\leq \frac{2\epsilon}{\log(1/2\alpha)} \cdot e^{-\frac{\log^2(1/2\alpha)}{8\epsilon^2}} \leq e^{-\frac{\log^2(1/2\alpha)}{8\epsilon^2}} \end{aligned}$$

Hence, we may choose any d (and thus force any list size) where

$$d \leq \max(1, \lfloor e^{\frac{\log^2(1/2\alpha)}{8\epsilon^2}}/2 \rfloor).$$

This means we can force a list of size $\exp(c \log^2(1/2\alpha)/\epsilon^2)$. $\square$

A.4 Omitted proofs from Section 5.1

Lemma A.1.

$$\Pr_{X_1, \dots, X_m \sim N(\mu, I)} \left[\left\| \frac{1}{m} \sum_{i=1}^m H_k(X_i) \otimes H_k(X_i) - \mathbb{E}_{X \sim N(\mu, I)} [H_k(X) \otimes H_k(X)] \right\|_F \geq a \right] \leq \frac{(dk \max(1, \|\mu\|_\infty))^{O(k)}}{a^2 m},$$

Proof of Lemma A.1. By

$$H_k(X) = \sum_{\substack{P \text{ Partition of set } S \\ \text{subset with size 1 and 2}}} \bigotimes_{\{a,b\} \in P} (-I)^{(a,b)} \bigotimes_{\{c\} \in P} X^{(c)}. \quad (14)$$

we know $H_k(X) \otimes H_k(X)$ can be written as the sum of terms in the form of

$$\bigotimes_{\{a,b\} \in P} (-I)^{(a,b)} \bigotimes_{\{c\} \in P} X^{(c)} \bigotimes_{\{a,b\} \in P'} (-I)^{(a+k, b+k)} \bigotimes_{\{c\} \in P} X^{(c+k)}.$$

For each such term, using concentrated [Claim A.4](#) and let $\theta = \max(1, \|\mu\|_\infty)$, we know the difference between empirical average and the expectation is at most $a/(k!)^2$ with probability at least $1 - \frac{(k!)^4 (8dk\theta^2)^k}{a^2 m} = 1 - \frac{(dk\theta)^{O(k)}}{a^2 m}$.

Since the number of summands is bounded by $(k!)^2$, by union bound we complete the proof. $\square$

Lemma A.2. For $\mu \in \mathbb{R}$, and integer $k \geq 0$, we have

$$\Pr_{X_1, \dots, X_m \sim N(\mu, 1)} \left[\left\| \frac{1}{m} \sum_{i=1}^m f(X_i) - \mathbb{E}_{X \sim N(\mu, 1)} [f(X)] \right\|_2 \geq a \right] \leq \frac{\exp(O(k + \sqrt{k}|\mu|))}{a^2 m},$$

for $f = h_k^2$ and h_k^4 .

Proof. By Chebyshev's inequality, it suffices to show $\mathbb{E}_{x \sim N(\mu, 1)} [f^2(x)] \leq \exp(O(k + \sqrt{k}|\mu|))$. Indeed by hypercontractivity, we have

$$\left(\mathbb{E}_{x \sim N(\mu, 1)} [h_k^8(x)] \right)^{1/8} \leq e^{O(k)} \left(\mathbb{E}_{x \sim N(\mu, 1)} [h_k^4(x)] \right)^{1/4} \leq e^{O(k)} \left(\mathbb{E}_{x \sim N(\mu, 1)} [h_k^2(x)] \right)^{1/2}$$

which is at most $\exp(O(\sqrt{k}|\mu|))$ due to [Lemma 5.9](#), so the lemma follows. $\square$

A.5 Proof of [Lemma 5.15](#)

Proof. We may bound the desired quantity by decomposing into two components:

$$\left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] \right\|_F \leq \left\| \mathbb{E}_{Z \sim N(0, I)} [Z^{\otimes k}] - \mathbb{E}_{X \sim N(\mu, I)} [X^{\otimes k}] \right\|_F + \left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{X \sim N(\mu, I)} [X^{\otimes k}] \right\|_F \quad (15)$$

We now bound the first term:

Claim A.3. $\| \mathbb{E}_{z \sim N(0, I)} [(\mu + z)^{\otimes k}] - \mathbb{E}_{z \sim N(0, I)} [z^{\otimes k}] \|_F^2 \leq (4dk)^{k/2} \|\mu\|_\infty$

Proof.

$$\begin{aligned} & \left\| \mathbb{E}_{z \sim N(0, I)} [(\mu + z)^{\otimes k}] - \mathbb{E}_{z \sim N(0, I)} [z^{\otimes k}] \right\|_F^2 \\ &= \sum_{i_1, \dots, i_k \in [d]} \left(\mathbb{E}_{z \sim N(0, I)} [\Pi_{j=1}^k (\mu_{i_j} + z_{i_j})] - \mathbb{E}_{z \sim N(0, I)} [\Pi_{j=1}^k z_{i_j}] \right)^2 \end{aligned}$$

Consider each monomial separately. The monomials of only z terms will cancel out. Otherwise, a monomial will have at least one μ term. Since the expected value of a monomial of z with at most $k-1$ terms is at most $(k-2)!!$, then:

$$\begin{aligned} & \leq d^k \cdot \left((2^k - 1) \cdot \|\mu\|_\infty \cdot (k-2)!! \right)^2 \\ & \leq (4dk)^k \|\mu\|_\infty^2 \\ & \Rightarrow \left\| \mathbb{E}_{z \sim N(0, I)} [(\mu + z)^{\otimes k}] - \mathbb{E}_{z \sim N(0, I)} [z^{\otimes k}] \right\|_F \leq (4dk)^{k/2} \|\mu\|_\infty \quad \square \end{aligned}$$

We may also bound the second term of Eq. (15):

Claim A.4. $\Pr_{X_1, \dots, X_m \sim N(\mu, I)} \left[\left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{X \sim N(\mu, I)} [X^{\otimes k}] \right\|_F \geq a \right] \leq \frac{(8dk \cdot \max(1, \|\mu\|_\infty^2))^k}{a^2 m}$

Proof. Using Chebyshev's inequality we observe

$$\Pr_{X_1, \dots, X_m \sim N(\mu, I)} \left[\left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{X \sim N(\mu, I)} [X^{\otimes k}] \right\|_F \geq a \right] \leq \frac{\text{Var}_{X \sim N(\mu, I)} (\|X^{\otimes k}\|_F)}{a^2}.$$

Hence, we will bound

$$\begin{aligned} \text{Var}_{X \sim N(\mu, I)} (\|X^{\otimes k}\|_F) &\leq \mathbb{E}_{X_1, \dots, X_m \sim N(\mu, I)} \left[\sum_{i_1, \dots, i_k \in [d]} \left(\frac{1}{m} \sum_{j=1}^m \left(\Pi_{\ell=1}^k (X_j)_{i_\ell} - \mathbb{E}_{X \sim N(\mu, I)} [\Pi_{\ell=1}^k X_{i_\ell}] \right) \right)^2 \right] \\ &= \frac{1}{m} \sum_{i_1, \dots, i_k \in [d]} \mathbb{E}_{X \sim N(\mu, I)} \left[\left(\Pi_{\ell=1}^k X_{i_\ell} - \mathbb{E}_{Y \sim N(\mu, I)} [\Pi_{\ell=1}^k Y_{i_\ell}] \right)^2 \right] \\ &\leq \frac{1}{m} \sum_{i_1, \dots, i_k \in [d]} \mathbb{E}_{X \sim N(\mu, I)} \left[\left(\Pi_{\ell=1}^k X_{i_\ell} \right)^2 \right] \\ &= \frac{1}{m} \sum_{i_1, \dots, i_k \in [d]} \mathbb{E}_{Z \sim N(0, I)} \left[\left(\Pi_{\ell=1}^k (Z_{i_\ell} + \mu_{i_\ell}) \right)^2 \right] \\ &\leq \frac{1}{m} \sum_{i_1, \dots, i_k \in [d]} 2^{2k} \cdot (2k-1)!! \cdot \max(1, \|\mu\|_\infty^{2k}) \\ &\leq \frac{(8dk \cdot \max(1, \|\mu\|_\infty^2))^k}{m}. \end{aligned}$$

All together, this implies our claim that

$$\Pr_{X_1, \dots, X_m \sim N(\mu, I)} \left[\left\| \frac{1}{m} \sum_{i=1}^m X_i^{\otimes k} - \mathbb{E}_{X \sim N(\mu, I)} [X^{\otimes k}] \right\|_F \geq a \right] \leq \frac{(8dk \cdot \max(1, \|\mu\|_\infty^2))^k}{a^2 m}. \quad \square$$

Concluding the proof of our lemma, we will set our parameters so that both terms of Eq. (15) are at most $\gamma/2$ with probability at least $1 - \delta$.

Bounding the first term with Claim A.3, it is sufficient to set $\|\mu\|_\infty$ such that:

$$(4dk)^{k/2} \|\mu\|_\infty \leq \gamma/2 \iff \|\mu\|_\infty \leq \frac{\gamma}{2(4dk)^{k/2}}$$

Bounding the second term with Claim A.4 and $\|\mu\|_\infty \leq 1$, it is sufficient to set m such that:

$$\frac{(8dk)^k}{(\gamma/2)^2 m} \leq \delta \iff m \geq \frac{4(8dk)^k}{\delta \gamma}.$$

This completes the proof of our lemma. $\square$

A.6 Proof of Lemma 5.17

Proof. The proof will mostly follow the structure of [DKN10], while making adjustments that allow for moments matching *approximately* instead of exactly in their work.

Consider the transformation $F : \mathbb{R}^d \rightarrow \mathbb{R}^m$ where $F(x)_i = \langle x, a_i \rangle$. Then the region of intersection for halfspaces is $\mathbb{1}[F(x)_i \geq b_i \ \forall i \in [m]]$. Without loss of generality, all $\|a_i\|_2 = 1$.

Let $I_R : \mathbb{R}^m \rightarrow \{0, 1\}$ be the indicator of this region, and $\tilde{I}_R^\eta : \mathbb{R}^m \rightarrow \mathbb{R}$ is the *FT-mollification* of I_R (you may think of it as a smoothed version of I_R with smoothing parameter η ; see Sections 3 and 4 of [DKN10] for an overview of FT-mollification, we will only use a few properties and will cite them clearly later). In this language, our desired guarantee is simply a bound on $|\mathbb{E}_G[I_R(F(X))] - \mathbb{E}_T[I_R(F(X))]|$. Set some parameters $\rho = \varepsilon/m$ and $\eta = m/\rho = m^2/\varepsilon$. The proof of this will entail three steps to show

$$\mathbb{E}_G[I_R(F(X))] \approx_{(a)} \mathbb{E}_G[\tilde{I}_R^\eta(F(X))] \approx_{(b)} \mathbb{E}_T[\tilde{I}_R^\eta(F(X))] \approx_{(c)} \mathbb{E}_T[I_R(F(X))].$$

These are shown in parts (a), (b), and (c), respectively.

Part (a). We will bound $\mathbb{E}_G[I_R(F(X))] \approx_{(a)} \mathbb{E}_G[\tilde{I}_R^\eta(F(X))]$. This part follows exactly as done in [DKN10]. We will provide it (almost copied exactly) for completeness. Let $d_2(x, \partial R)$ denote the L_2 distance from some point $x \in \mathbb{R}^m$ to the boundary of R . Then,

$$\begin{aligned} & |\mathbb{E}_G[I_R(F(X))] - \mathbb{E}_G[\tilde{I}_R^\eta(F(X))]| \\ & \leq \mathbb{E}_G[|I_R(F(X)) - \tilde{I}_R^\eta(F(X))|] \end{aligned} \tag{16}$$

Using Theorem 4.10 of [DKN10], which states $|I_R(x) - \tilde{I}_R^\eta(x)| \leq \min \left\{ 1, O \left(\left(\frac{m}{\eta d_2(x, \partial R)} \right)^2 \right) \right\}$:

$$\begin{aligned} & \leq \Pr_G[d_2(F(X), \partial R) \leq 2\rho] + O \left(\sum_{s=1}^{\infty} \left(\frac{m^2}{\eta^2 2^{2s} \rho^2} \right) \cdot \Pr_G[d_2(F(X), \partial R) \leq 2^{s+1}\rho] \right) \\ & \leq \sum_{i=1}^m \Pr_G[|F(X)_i - b_i| \leq 2\rho] + O \left(\sum_{s=1}^{\infty} \sum_{i=1}^m 2^{-2s} \cdot \Pr_G[|F(X)_i - b_i| \leq 2^{s+1}\rho] \right) \\ & \leq 2m\rho C_2 + O \left(\sum_{s=1}^{\infty} 2^{-2s} \cdot (2^{s+1}\rho m C_2) \right) \quad (\text{by anti-concentration}) \\ & = O(\varepsilon C_2) = O(\varepsilon) \end{aligned} \tag{17}$$

Part (b). Let us introduce some notation for Taylor expansions around 0. Let $p_{\leq r}$ denote the degree- r Taylor expansion of $\tilde{I}_R^\eta$, let $p_{=r}$ denote the terms of $p_{\leq r}$ with degree exactly r , and let $p_{>r}(x) \triangleq \tilde{I}_R^\eta(x) - p_{\leq r}(x)$. For any polynomial p and multi-index α , let $p[\alpha]$ denote the coefficient for this multi-index. We follow the plan of [DKN10] and decompose into two terms

$$|\mathbb{E}_G[\tilde{I}_R^\eta] - \mathbb{E}_T[\tilde{I}_R^\eta]| \leq |\mathbb{E}_G[p_{>k-1}] - \mathbb{E}_T[p_{>k-1}]| + |\mathbb{E}_G[p_{\leq k-1}] - \mathbb{E}_T[p_{\leq k-1}]|.$$

We start by bounding the first term:

$$\begin{aligned} & |\mathbb{E}_G[p_{>k-1}] - \mathbb{E}_T[p_{>k-1}]| \\ & \leq \mathbb{E}_G[|p_{>k-1}|] + \mathbb{E}_T[|p_{>k-1}|] \end{aligned}$$

Using Theorem 4.8 of [DKN10] to bound $\|\partial^\beta \tilde{I}_R^\eta\|_\infty \leq (2\eta)^{|\beta|}$ for all $\beta \in \mathbb{N}^m$, and Taylor's theorem $|p_{>k-1}(F(x))| \leq \sup_{|\beta|=k} \|\partial^\beta \tilde{I}_R^\eta\|_\infty \cdot \frac{\|F(X)\|_1^k}{k!}$:

$$\begin{aligned} & \leq \mathbb{E}_G \left[\frac{(2\eta)^k}{k!} \|F(X)\|_1^k \right] + \mathbb{E}_T \left[\frac{(2\eta)^k}{k!} \|F(X)\|_1^k \right] \\ & \leq \frac{(2\eta m)^k}{k!} \left(\mathbb{E}_G[\max_{i \in [m]} (F(X)_i)^k] + \mathbb{E}_T[\max_{i \in [m]} (F(X)_i)^k] \right) \\ & \leq \frac{m(2\eta m)^k}{k!} \left(\max_{i \in [m]} \mathbb{E}_G[(F(X)_i)^k] + \max_{i \in [m]} \mathbb{E}_T[(F(X)_i)^k] \right) \\ & \leq \frac{m(2\eta m)^k}{k!} \left(2 \sup_{\|v\|_2=1} \mathbb{E}_G[\langle X, v \rangle^k] + \sup_{\|v\|_2=1} \left| \mathbb{E}_G[\langle X, v \rangle^k] - \mathbb{E}_T[\langle X, v \rangle^k] \right| \right) \\ & \leq \frac{m(2\eta m)^k}{k!} \left(2(C_1 \sqrt{k})^k + \Delta \right) \\ & \leq m \cdot \left(\frac{2em^3}{k\varepsilon} \right)^k \cdot \left(2(C_1 \sqrt{k})^k + \Delta \right) \end{aligned}$$

This quantity is at most ε if we choose $k \geq Cm^6/\varepsilon^2$ for sufficiently large $C > 0$, and $\Delta \leq 1$, which are both satisfied by our choice in the theorem statement.

For the remaining summand, bounding the difference in the expectation of $p_{\leq k-1}$ will leverage the lower-order moment bounds as you might expect. The following lemma will be helpful for bounding monomial moments in terms of moments for projections:

Lemma A.5 (Optimizer of symmetric k -linear form. Implied by Equation (2) in [CR19]; references proofs in [Ban38, BS71, PST07, Din12]). *Consider any two random variables X, Y over $\mathbb{R}^d$. Then,*

$$\sup_{\|u_1\|_2, \dots, \|u_k\|_2=1} \mathbb{E}_X[\Pi_{j=1}^k \langle X, u_j \rangle] - \mathbb{E}_Y[\Pi_{j=1}^k \langle Y, u_j \rangle] = \sup_{\|v\|_2=1} \mathbb{E}_X[\langle X, v \rangle^k] - \mathbb{E}_Y[\langle Y, v \rangle^k].$$

First, we may use this to get a bound in the moment projections for $F(X)$ under G, H with only a blowup depending on k, m (not d):

Claim A.6. For any unit $v \in \mathbb{R}^j$ ($\|v\|_2 = 1$) and $k' \in \{0, \dots, k\}$,

$$\left| \mathbb{E}_G[\langle F(X), v \rangle^{k'}] - \mathbb{E}_T[\langle F(X), v \rangle^{k'}] \right| \leq m^{k/2} \Delta$$

Proof.

$$\begin{aligned} & \left| \mathbb{E}_G[\langle F(X), v \rangle^{k'}] - \mathbb{E}_T[\langle F(X), v \rangle^{k'}] \right| \\ &= \left| \mathbb{E}_G \left[\left(\sum_{i=1}^m \langle X, a_i \rangle \cdot v_i \right)^{k'} \right] - \mathbb{E}_T \left[\left(\sum_{i=1}^m \langle X, a_i \rangle \cdot v_i \right)^{k'} \right] \right| \\ &= \left| \mathbb{E}_G \left[\sum_{i_1, \dots, i_{k'} \in [m]} \Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \cdot v_{i_j} \right] - \mathbb{E}_T \left[\sum_{i_1, \dots, i_{k'} \in [m]} \Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \cdot v_{i_j} \right] \right| \\ &\leq \sum_{i_1, \dots, i_{k'} \in [m]} \left| \mathbb{E}_G \left[\Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \cdot v_{i_j} \right] - \mathbb{E}_T \left[\Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \cdot v_{i_j} \right] \right| \\ &= \sum_{i_1, \dots, i_{k'} \in [m]} \left(\Pi_{j=1}^{k'} |v_{i_j}| \right) \cdot \left| \mathbb{E}_G \left[\Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \right] - \mathbb{E}_T \left[\Pi_{j=1}^{k'} \langle X, a_{i_j} \rangle \right] \right| \\ &\leq \|v\|_1^{k'} \sup_{\|u_1\|_2, \dots, \|u_{k'}\|_2 = 1} \left| \mathbb{E}_G[\Pi_{j=1}^{k'} \langle X, u_j \rangle] - \mathbb{E}_T[\Pi_{j=1}^{k'} \langle X, u_j \rangle] \right| \\ &\leq m^{k'/2} \Delta \quad \square \end{aligned} \tag{using Lemma A.5}$$

Corollary A.7. For any $k' \in \{0, \dots, k\}$

$$\sup_{i_1, \dots, i_{k'} \in [m]} \left| \mathbb{E}_G[\Pi_{j=1}^{k'} f(X)_{i_j}] - \mathbb{E}_T[\Pi_{j=1}^{k'} f(X)_{i_j}] \right| \leq m^{k'/2} \Delta$$

Proof. This follows immediately from Lemma A.5 and Claim A.6. $\square$

With these moment bounds in hand, we now bound

$$\left| \mathbb{E}_G[p_{\leq k-1}(F(x))] - \mathbb{E}_T[p_{\leq k-1}(F(x))] \right| \tag{18}$$

$$\leq k \max_{i \in [k]} \left| \mathbb{E}_G[p_{=i}(F(X))] - \mathbb{E}_T[p_{=i}(F(X))] \right| \tag{19}$$

$$\leq k \max_{i \in [k]} \sum_{\alpha=(j_1, \dots, j_i), j_\ell \in [m]} |p_{=i}[\alpha]| \cdot \left| \mathbb{E}_G[F(X)_\alpha] - \mathbb{E}_T[F(X)_\alpha] \right| \tag{20}$$

$$\leq k \max_{i \in [k]} m^i \cdot \left(\max_{\alpha=(j_1, \dots, j_i), j_\ell \in [m]} |p_{=i}[\alpha]| \right) \cdot \max_{\alpha=(j_1, \dots, j_i), j_\ell \in [m]} \left| \mathbb{E}_G[F(X)_\alpha] - \mathbb{E}_T[F(X)_\alpha] \right| \tag{21}$$

Using Theorem 4.8 of [DKN10] to bound $\|\partial^\beta \tilde{I}_R^\eta\|_\infty \leq (2\eta)^{|\beta|}$ for all $\beta \in \mathbb{N}^m$:

$$\begin{aligned}
&\leq k \max_{i \in [k]} m^i \cdot \left(\frac{(2\eta)^i}{i!} \right) \cdot \max_{\alpha=(j_1, \dots, j_i), j_\ell \in [m]} \left| \mathbb{E}_G[F(X)_\alpha] - \mathbb{E}_T[F(X)_\alpha] \right| \quad (22) \\
&\leq k \max_{i \in [k]} m^i \cdot \left(\frac{(2\eta)^i}{i!} \right) \cdot \left(m^{i/2} \Delta \right) \quad (\text{using Corollary A.7}) \\
&\leq k \cdot (2m^{3.5}/\varepsilon)^k \Delta \leq (4k)^k \Delta \quad (\text{using } k \geq m^6/\varepsilon^2)
\end{aligned}$$

which is at most ε if we choose $\Delta \leq (8k)^{-k}$ as is chosen in the theorem statement.

Part (c). This will follow very similarly to part (a). We may identically argue using the steps from Eq. (16) to Eq. (17). What remains is to show H is sufficiently anti-concentrated in its one-dimensional projections. Note how this is implied if we could show the special case where H fools any pair of $m = 2$ halfspaces. One convenient way to prove this, is noting how the result of [GKK23] was only too lossy for super-constant m , so we may use the results after certifying how the conditions hold for moment-matching $F(X)$ under G and $F(X)$ under H . We rephrase Theorem 5.6 of [GKK23] (their result holds for arbitrary functions of the indicators of halfspaces, but for simplicity we just state their result for intersections of halfspaces).

Theorem A.8 (Theorem 5.6 of [GKK23]). *Suppose D and D' are distributions over $\mathbb{R}^d$ such that D satisfies*

1. *α -strictly subexponential tails: For all $\|u\| = 1$, $\Pr_D[|\langle x, u \rangle| > t] \leq \exp(-C_1 t^{1+\alpha})$ for some positive constant C_1 .*
2. *Anticoncentration: For all $\|u\| = 1$ and continuous intervals $T \subset \mathbb{R}$, we have $\Pr_D[\langle x, u \rangle \in T] \leq C_2 |T|$ for some positive constant C_2 .*

Moreover, for some $k \in \mathbb{N}$, every j -moment tuple is close (for $j \in \{0, \dots, k\}$):

$$\left| \mathbb{E}_D[\Pi_{\ell=1}^j X_{i_\ell}] - \mathbb{E}_{D'}[\Pi_{\ell=1}^j X_{i_\ell}] \right| \leq \frac{\sqrt{m}}{2k} \frac{j!}{d^j} \left(\frac{1}{C_5 k^{\alpha/(1+\alpha)}} \right)^{j+1}$$

for some positive constant C_5 .

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be the indicator of the intersection of m halfspaces. Then, for some constant $C > 0$,

$$\left| \mathbb{E}_D[f] - \mathbb{E}_{D'}[f] \right| \leq k^{-\alpha/(1+\alpha)} \sqrt{m} (C \log(\sqrt{m} k^{\alpha/(1+\alpha)}))^{2m}.$$

We will invoke this result simply with $d = 1$ and $m = 2$, using the following immediate corollary:

Corollary A.9. *Suppose D and D' are distributions over $\mathbb{R}$ where D has subexponential tails (with $\alpha = 1$) and anticoncentration as stated in Theorem A.8. There exists some constant $C^* > 0$ where if for some $k \in \mathbb{N}$, every j -moment tuple is close (for $j \in \{0, \dots, k\}$):*

$$\left| \mathbb{E}_D[\Pi_{\ell=1}^j X_{i_\ell}] - \mathbb{E}_{D'}[\Pi_{\ell=1}^j X_{i_\ell}] \right| \leq \frac{j!}{k} \left(\frac{1}{C^* \sqrt{k}} \right)^{j+1}$$

Then, for any $0 < \varepsilon \leq \frac{1}{2}$, as long as $k \geq C^ \cdot \frac{1}{\varepsilon^2} \cdot \log^8(1/\varepsilon)$, it holds that*

$$|\Pr_D[x \in [L, R]] - \Pr_{D'}[x \in [L, R]]| \leq \varepsilon.$$

Proof. This is just invoking [Theorem A.8](#) with $d = 1$ and $m = 2$. The moment condition of [Theorem A.8](#) follows trivially from the assumption of this corollary. We now simply compute

$$\begin{aligned}
& |\Pr_D[x \in [L, R]] - \Pr_{D'}[x \in [L, R]]| \leq k^{-\alpha/(1+\alpha)} \sqrt{m} (C_1 \log(\sqrt{m} k^{\alpha/(1+\alpha)}))^{2m} \\
&= \frac{\sqrt{2}}{\sqrt{k}} (C_1 \log(\sqrt{2}) + \frac{1}{2} \log(k))^4 \\
&\leq C \cdot \frac{\log(k)^4}{\sqrt{k}} \\
&\leq C \cdot (\log(C^*)^4 + \log(1/\varepsilon)^4) \cdot \left(\frac{\varepsilon}{C^* \log^4(1/\varepsilon)} \right) \quad (\text{by monotonicity and } k \geq \frac{C^* \log^8(1/\varepsilon)}{\varepsilon^2}) \\
&\leq \varepsilon \quad \square \quad (\text{for sufficiently large } C^*)
\end{aligned}$$

We may invoke [Corollary A.9](#) on $F(X)_i$ under G and T , with $\varepsilon' = \varepsilon/m$. The anticoncentration property holds under G by the assumption of our theorem, and the subexponential tails property holds under G since it is equivalent (up to the choice of constant) to our theorem's subgaussian moment assumption (e.g. see Proposition 2.5.2 of [Ver18]). Using [Corollary A.7](#), the moment tuple difference condition holds as long as

$$\begin{aligned}
m^{k/2} \Delta &\leq \frac{1}{k} \left(\frac{1}{C^* \sqrt{k}} \right)^{k+1} \\
&\iff \Delta \leq \frac{1}{k \cdot m^{k/2}} \left(\frac{1}{C^* \sqrt{k}} \right)^{k+1} \\
&\iff \Delta \leq \left(\frac{1}{C \cdot C^* \cdot k} \right)^k \quad (\text{for sufficiently large } C > 0, \text{ using } k \geq m, C^* \geq 1)
\end{aligned}$$

which follows from our choice of Δ . Hence, as long as $k \geq C^* \cdot \frac{1}{\varepsilon'^2} \cdot \log^8(1/\varepsilon') = C^* \cdot \frac{m^2}{\varepsilon^2} \cdot \log^8(m/\varepsilon)$, we may conclude, using the steps from [Eq. \(16\)](#) to [Eq. \(17\)](#) in part(a):

$$\begin{aligned}
& |\mathbb{E}_T[I_R(F(X))] - \mathbb{E}_T[\tilde{I}_R^\eta(F(X))]| \\
&\leq \sum_{i=1}^m \Pr_T[|F(X)_i - b_i| \leq 2\rho] + O\left(\sum_{s=1}^{\infty} \sum_{i=1}^m 2^{-2s} \cdot \Pr_H[|F(X)_i - b_i| \leq 2^{s+1}\rho]\right)
\end{aligned}$$

By [Corollary A.9](#):

$$\begin{aligned}
&\leq O(m\varepsilon') + \sum_{i=1}^m \Pr_T[|F(X)_i - b_i| \leq 2\rho] + O\left(\sum_{s=1}^{\infty} \sum_{i=1}^m 2^{-2s} \cdot \Pr_H[|F(X)_i - b_i| \leq 2^{s+1}\rho]\right) \\
&\leq O(m\varepsilon') + O(\rho m C_2) = O(\varepsilon C_2) = O(\varepsilon) \quad \square
\end{aligned}$$