Efficient Certificates of Anti-Concentration Beyond Gaussians

Ainesh Bakshi* ainesh@mit.edu MIT

Pravesh K. Kothari kothari@cs.princeton.edu IAS & Princeton University

Goutham Rajendran gouthamr@cmu.edu CMU

Madhur Tulsiani† madhurt@ttic.edu TTIC

Aravindan Vijayaraghavan‡ aravindv@northwestern.edu Northwestern

Abstract

A set of high dimensional points $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ in isotropic position is said to be δ-anti concentrated if for every direction v, the fraction of points in X satisfying $|\langle x_i, v\rangle| \leqslant \delta$ is at most $O(\delta)$. Motivated by applications to list-decodable learning and clustering, three recent works [KKK19; RY20a; BK21] considered the problem of constructing efficient certificates of anti-concentration in the average case, when the set of points X corresponds to samples from a Gaussian distribution. Their certificates played a crucial role in several subsequent works in algorithmic robust statistics on list-decodable learning and settling the robust learnability of arbitrary Gaussian mixtures. Unlike related efficient certificates of concentration properties that are known for wide class of distributions [KSS18], the aforementioned approach has been limited only to *rotationally invariant* distributions (and their affine transformations) with the only prominent example being Gaussian distributions.

This work presents a new (and arguably the most natural) formulation for anti- concentration. Using this formulation, we give quasi-polynomial time verifiable sum-of-squares certificates of anti-concentration that hold for a wide class of *non-Gaussian* distributions including anti-concentrated bounded product distributions and uniform distributions over L_p balls (and their affine transformations). Consequently, our method upgrades and extends results in algorithmic robust statistics e.g., list-decodable learning and clustering, to such distributions.

As in the case of previous works, our certificates are also obtained via relaxations in the sum-of-squares hierarchy. However, the nature of our argument differs significantly from prior works that formulate anti-concentration as the non-negativity of an explicit polynomial. Our argument constructs a canonical integer program for anti-concentration and analysis a SoS relaxation of it, independent of the intended application. The explicit polynomials appearing in prior works can be seen as specific dual certificates to this program.

From a technical standpoint, unlike existing works that explicitly construct sum-of-squares certificates, our argument relies on duality and analyzes a pseudo-expectation on large subsets of the input points that take a small value in some direction. Our analysis uses the method of polynomial *reweightings* to reduce the problem to analyzing only *analytically dense* or *sparse* directions.

*Supported by the NSF TRIPODS program (award DMS2022448) and Ankur Moitra's ONR grant.

†Supported by the NSF grants CCF-1816372 and CCF-2326685.

‡Supported by NSF grants CCF-1652491, CCF-2154100 and ECCS-2216970.

Contents

1 Introduction

A high-dimensional random variable $X \in \mathbb{R}^d$ is said to be (ε, δ) anti-concentrated, if along any direction $v \in S^{d-1}$ and $a \in \mathbb{R}$, we have $\Pr[|\langle X, v\rangle - a| \leqslant \delta] \leqslant \varepsilon$. For example, a spherical Gaussian can easily be shown to be (δ, δ) anti-concentrated for any $\delta > 0$ (for sufficiently large d). While concentration properties of distributions can be studied via tools such as moment bounds, the study of anti-concentration often requires more delicate structural information about the underlying distribution [Kri16].

In this work, motivated by various statistical applications, we consider the problem of *certifying* anti-concentration properties for a set of points. In analogy with the above definition of random variables, a set of high dimensional points $X = \{x_1, x_2, \ldots, x_n\} \subseteq R^d$ in isotropic position is said to be (ε, δ) anti-concentrated if for every unit vector v and a shift $a \in \mathbb{R}$, the fraction of points in X satisfying $|\langle x_i, v\rangle| \leqslant \delta$ is at most ε i.e., $\Pr_{i\in[n]}[|\langle x_i, v\rangle - a| \leqslant \delta] \leqslant \varepsilon$.

The problem of deciding if a given set X is (ε, δ) anti-concentrated is NP-hard in the worst case.[1] Moreover, for sufficiently small values of the anti-concentration scale $\delta \approx 1/\sqrt{d}$, efficient algorithms for certifying δ anti-concentration imply algorithms for the (homogeneous version of) the Continuous Learning With Errors (hCLWE) problem considered by Bruna, Regev, Song, and Tang [BRST21].

In the average case, when the point set X corresponds to independent samples of certain distributions, certificates of anti-concentration have served as important primitives in several algorithms for problems in list-decodable learning and algorithmic robust statistics, such as list-decodable linear regression [KKK19], subspace recovery [RY20b; BK21], and covariance estimation [IK22], and learning mixtures of Gaussians [BDJKKV22; LM21]. Moreover, certifying anti-concentration is also shown to be *necessary* for list-decodable regression [KKK19].

The certificates for anti-concentration in prior works were obtained by formulating it as the non-negativity of an explicit polynomial and exhibit the non-negativity by expressing it as a sum-of-squares polynomial. However, discovering such sum-of-squares certificates relied crucially on the properties of the underlying distributions, such as sub-Gaussian tails and spherical symmetry, with the only prominent examples being Gaussian distributions and the uniform distribution over the Euclidean ball.

Moreover, while the algorithms for list-decoding and clustering were often based on anti-concentration properties, the certificates of anti-concentration played a secondary role, and varied from one application to the next, as per convenience. Motivated by the above, we consider the following natural question:

Does there exist an application-agnostic certificate of anti-concentration that applies to a broad class of distributions?

In this work, we obtain such certificates of anti-concentration via sum-of-squares relaxations of a natural integer program. Our certificates apply to a broad class of distributions that goes beyond Gaussians, or rotationally invariant distributions (or their affine transformations). We also unify several prior applications of anti-concentration and resolve other open problems raised in prior works.

[1] In the special case when $\delta = 0$, this corresponds to deciding whether there exists an ε fraction of points that lie on a $d-1$ dimensional subspace. This is known to be NP-hard, see e.g. [HM13; Kha95].

1.1 Our Results

We begin by defining the class of distributions that would be the focus of this paper (refer to Definition 2.23 for a quantitative definition).

Definition 1.1 (Reasonably anti-concentrated distributions)**.** A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is *reasonably anti-concentrated* if it satisfies the following: (a) $\mathcal{D}$ is anti-concentrated (b) $\mathcal{D}$ has *strictly sub-exponential* tails, (c) $\mathcal{D}$ satisfies *hyper-contractivity* of linear forms and (d) $\mathcal{D}$ is almost k-wise independent.

Apart from Gaussian-like distributions, we show that a broad family of distributions are *reasonably anti-concentrated*, including arbitrary affine transformations of anti-concentrated bounded product distributions e.g., the solid hypercube (see Lemma 2.24) and uniform distributions over ℓ_p balls for any $p > 1$ (see Theorem 9.1).

Now, given a point set $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ of samples from a distribution with mean μ and covariance Σ, consider the following natural optimization problem to certify anti-concentration:

$$\max_{w \in \mathbb{R}^n, v \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i \in [n]} w_i$$
$$\text{s.t.} \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ \forall i \in [n] & w_i \langle x_i - \mu, v \rangle^2 \leqslant w_i \cdot \delta^2 \cdot (v^\top \Sigma v) \end{Bmatrix} \tag{1}$$

Here, the w_i's are indicators, and the constraints count how many samples are smaller than δ^2 fraction of the variance along v. This program is *the* natural integer program formulation of anti-concentration. Intuitively, this program finds a direction v such that projecting the samples along v witnesses the largest fraction of *concentrated* samples.

Our main result shows that we can efficiently *certify* an upper bound on the objective value of this program as long as X is drawn i.i.d. from any affine transformation of a *reasonably anti-concentrated* distribution.

Theorem 1.2 (Efficient Anti-Concentration certificate, Informal version of Theorem 3.1)**.** *Given a point set $X = \{ x_i \}_{i \in [n]}$ in $\mathbb{R}^d$ such that X is drawn iid from an unknown affine transformation of a reasonably anti-concentrated distribution $\mathcal{D}$ and $0 < \delta < 1$, there exists an algorithm that with high probability over the samples, certifies $\max_{w,v} \sum_{i \in [n]} w_i / n \leqslant \delta$ and runs in $n^{\mathcal{O}_\delta(\log(d))}$ time.* [2]

Remark 1.3 (Comparison to prior work)**.** Prior works [KKK19; RY20a; RY20b; BK21; BK20] have exhibited similar certificates of anti-concentration, in the context of list-decodable regression, subspace recovery and clustering, but there are two main differences compared to our work.

1. While their certificates also also obtained via SoS relaxations, they have worked with non-standard formulations which are usually adapted to the problem at hand and consider explicit polynomial approximations to the box indicator function. Furthermore, the analyses of such certificates in the corresponding algorithmic applications have been ad-hoc. In contrast, our certificate considers the canonical integer program formulation Eq. (1) that essentially mimics the definition of anti-concentration.
2. The certificates in prior works only work for spherically symmetric distributions such as Gaussians or uniform distribution on the sphere. This essentially reduces certifying anti-concentration to a univariate problem. Our certificate on the other hand goes beyond

[2] Throughout, we use $\mathcal{O}_\delta$ to suppress the explicit dependence on δ

spherical symmetry and allows us to handle the more general notion of *reasonably anti-concentrated* distributions instead (Definition 1.1), albeit at the cost of quasi-polynomial running time.

Finally, we note that we indirectly infer the existence of a sum-of-squares proof by showing that any low-degree pseudo-distribution consistent with the program in Eq. (1) must in fact satisfy that the pseudo-expected objective value is at most δ. It then follows from SDP duality that a low-degree sum-of-squares certificate of this inequality must exist. Our inability to obtain a direct sum-of-squares proof perhaps illustrates why finding such certificates has remained elusive in the past.

1.2 Applications

We show that we can use our certificate of anti-concentration to obtain efficient algorithms for various statistics problems, including the first algorithms for clustering mixtures of *reasonably anti-concentrated* distributions with arbitrary covariances, and list-decodable regression for such distributions. We note that all prior work on these problems was limited to affine transformations of rotationally invariant distributions, specifically Gaussians.

Warmup: Spectrally separated mixture with two clusters. As a warmup, we use our new anti-concentration certificate to give a quasi-polynomial time algorithm for clustering any mixture of two reasonably anti-concentrated distributions that are spectrally separated i.e., there is a direction v along which the variances of the two components differ significantly.

Theorem 1.4 (Spectrally Separated Mixture, Informal version of Theorem 6.1)**.** *Consider a set of n i.i.d. samples drawn from $\mathcal{M} = \frac{1}{2}\mathcal{D}(0,\Sigma_1) + \frac{1}{2}\mathcal{D}(0,\Sigma_2)$ such that there exists a direction v satisfying $v^\top\Sigma_1 v < \delta^3 v^\top\Sigma_2 v$, for some $\delta < 1$. Then, there exists an algorithm that runs in $n^{\mathcal{O}_\delta(\log^2(d))}$ time and with probability at least 99/100 outputs a direction z such that $z^\top\Sigma_1 z < \mathcal{O}(\delta) z^\top\Sigma_2 z$, whenever $n = d^{\Omega_\delta(\log(d))}$.*

The above algorithm when given samples from such a spectrally separated mixture finds a direction z where the variances of the two components differ by a factor of $\Omega(1/\delta)$. One can then project along this direction to recovers the two clusters up to $O(\delta)$ error.

This setting of spectrally separated mixtures captures as a special case the well-studied problem of *subspace clustering* [Vid03; BK21], where the components are supported on two distinct subspaces. Hence there is a direction v along which the first component has non-zero variance, and the second component has variance 0 i.e., the spectral separation is arbitrarily large. The above algorithm provides quasi-polynomial time guarantees when the points in each cluster are drawn from *reasonably anti-concentrated* distributions supported on the corresponding subspace. To the best of our knowledge, previous guarantees only hold in the special case when the points in the subspace is drawn from a Gaussian (or a rotationally invariant distribution) restricted to the subspace, or assume other structural assumptions on the subspaces [BK21; CGKMS24]. Moreover, Theorem 6.1 guarantees approximate recovery of the clusters even when the spectral separation is only a (sufficiently large) constant factor.

Clustering mixtures of *reasonably anti-concentrated* distributions. Next, we consider the more general setting when the points are drawn from a mixture of k *reasonably anti-concentrated* distributions. There is rich body of work on clustering mixture models where efficient polynomial time algorithms are known when the means of the clusters are separated sufficiently

see e.g., [VW04; AM05; KK10; HL18; KS17b; DKS17a]. In the specific setting when the component distributions are Gaussians, recent works provide polynomial time guarantees under significantly less stringent notion of separation called the "clusterable setting", which includes spectral separation, or Frobenius separation, or mean separation; see Definition 7.1 for a formal definition.

To the best of our knowledge, there are no known algorithms in the clusterable setting for more general distributions beyond Gaussians. We make progress on this question by giving an algorithm that , with an additional technical condition of certified hypercontractivity of quadratic forms (Definition 2.19)), recovers the clusters up to $O(\delta)$ error when the underlying mixture is clusterable.

Theorem 1.5 (Clustering Mixtures, Informal version of Theorem 7.2)**.** *Given $0 < \varepsilon < 1$ and n samples from a mixture of $k = O(1)$ reasonably anti-concentrated distributions that also have certifiably hypercontractive quadratic forms (Definition 2.19) such that the resulting mixture is clusterable (Definition 7.1), there exists an algorithm that runs in $n^{\mathcal{O}_{k,\varepsilon}(\log^2(d))}$ time and outputs an ε-accurate clustering whenever $n = d^{\Omega_{k,\varepsilon}(\log(d))}$.*

Remark 1.6 (Outlier-robust algorithms)**.** We remark that using standard machinery in robust statistics, the sum-of-squares relaxation we analyze can be made to work in the strong contamination model. Here, an η-fraction of input samples are adversarially corrupted, and the rest are drawn from a mixture of reasonably anti-concentrated distributions. We refer the reader to [FKP+19; BDHKKK20] for details.

List-decodable linear regression. Finally, we consider the setting of list-decodable linear regression. Here an α-fraction of the samples are *inliers* and are drawn i.i.d. from some *reasonably anti-concentrated* distribution $\mathcal{D}$. The remaining $(1-\alpha)$ fraction of the samples are arbitrary, and potentially adversarially generated. The underlying statistical model is simply $y_i = \langle \Theta, x_i \rangle$, where $x_i \sim \mathcal{D}$, and Θ is some unknown hyperplane. For the formal definition, see Model 8.1. The following theorem gives a quasi-polynomial time guarantee for this problem when the inliers are drawn from any resonably anti-concentration distribution.

Theorem 1.7 (List-Decodable Regression, Informal version of Theorem 8.2)**.** *For any $0 < \alpha, \varepsilon < 1$, there exists an algorithm that takes n samples from the above model where an α-fraction are from a reasonably anti-concentrated distribution, runs in $n^{\mathcal{O}_{\varepsilon,\alpha}(\log(d))}$ time and outputs a list of $\mathcal{O}(1/\alpha)$ vectors such that with probability at least $99/100$, the list contains at least one vector $\hat{\Theta}$ such that $\|\hat{\Theta} - \Theta\|_2^2 \leqslant \varepsilon$, when $n = d^{\Omega_{\varepsilon,\alpha}(\log(d))}$.*

1.3 Technical Overview

In this section, we provide an outline of the techniques we introduce to obtain our certificate of anti-concentration.

Inefficient certificates of anti-concentration. As a starting point, we note that the uniform distribution over any convex body is anti-concentrated, and therefore the optimization problem in Eq. (1) achievies a value of at most δ, for any $\delta < 1$. The proof of anti-concentration simply follows from observing that the density of any univariate projection is upper bounded by $\exp(-t/c)$, for a fixed constant c. One naïve certificate that demonstrates $\sum_i w_i$ is bounded in all directions is one where we simply create an ε-net and check each univariate projection, i.e. for each vector in the net, we project the distribution along this vector and check that it is

anti-concentrated. However, it is easy to see that the size of such a certificate is at least $\exp(d)$ and our goal is to avoid this exponential dependence on the dimension.

Existing approaches to certifying anti-concentration. As alluded to earlier, the notion of certifiable anti-concentration was introduced in two concurrent works, one by Raghavendra and Yau [RY20a] and one by Karmalkar, Klivans and Kothari [KKK19], and they formulate anti-concentration as a polynomial inequality. First, let $f(x)$ be the univariate box-indicator function:

$$f(x) = \begin{cases} 1 & \text{if } -\delta \leqslant x \leqslant \delta \\ 0 & \text{otherwise} \end{cases}$$

Then, there exists a polynomial $p(x)$ that approximates f everywhere in the interval $[-1, 1]$, and the degree of this polynomial is roughly $\mathcal{O}(1/\delta^2)$ (in fact, there is a sum-of-squares polynomial satisfying these criteria). The formulation of anti-concentration considered in prior works is as follows:

$$\forall v \in \mathbb{R}^d, \quad \frac{1}{n} \sum_{i \in [n]} p(\langle x_i, v \rangle^2) \leqslant \delta \|v\|_2^2. \tag{2}$$

Since p is large in the interval $[-\delta, \delta]$, such a certificate implies that the number of samples for which $|\langle x_i, v \rangle| \leqslant \delta$ is bounded. To avoid brute force search over an ε-net, [RY20a; KKK19; BK20] show that there is a sum-of-squares proof of the inequality in Eq. (2) in the indeterminate v (see Section 2.1 for background on sum-of-squares proofs) and this in turn implies an efficient certificate.

However, all known proofs of such a certificate assume that the input distribution is rotationally invariant. Therefore, $p(\langle x_i, v \rangle^2) = p(g\|v\|^2)$, where g is a standard Gaussian, implying p is a univariate function, and any univariate inequality admits a sum-of-squares proof (see Fact 2.10). It is easy to see that all prior works thus heavily rely on the input distribution being rotationally invariant.

Anti-concentration along analytically dense directions. It is well known that the marginal distribution obtained by projecting the uniform distribution over a convex body along a random direction is well-approximated by a Gaussian. Restricting our attention to well-behaved convex bodies, we develop an analytic formulation of this statement: For any direction v such that $\|v\|_4^4 \leqslant \lambda \|v\|_2^4$, we show that there is a sum-of-squares proof in v that $\frac{1}{n} \sum_{i \in [n]} p^2(\langle x_i, v \rangle^2) \leqslant \delta$, where p^2 is the sum-of-squares polynomial that approximates the box-indicator function. In the language of sum-of-squares proofs,

$$\left\{ \|v\|_4^4 \leqslant \exp(-1/\delta^2) \|v\|_2^4 \right\} \vdash^{v} \left\{ \frac{1}{n} \sum_{i \in [n]} p^2(\langle x_i, v \rangle^2) \leqslant \delta \right\}, \tag{3}$$

i.e. under the axiom that the ℓ_4^4 norm of v is bounded, there is an efficient certificate that p^2 is small on the input samples in all directions (see Theorem 4.1). The captures the intuition that v cannot be a sparse direction or have large magnitude coordinates. Our proof proceeds by explicitly bounding the higher moments of $\langle x, v \rangle^2$ and showing that under the axiom $\{ \|v\|_4^4 \leqslant \exp(-1/\delta^2) \|v\|_2^4 \}$, the moments are $\exp(-1/\delta^2)$-close to that of a Gaussian, additively (see Lemmas 4.2 and 4.5)[3].

[3]Our proof is inspired by Lemma 2.6 in Kindler, Noar and Schechtman [KNS10]

Additionally, we show that p^2 certifies anti-concentration around any bounded point ϕ, as opposed to just the origin, i.e.

$$\left\{ \|v\|_4^4 \leqslant \exp(-1/\delta^2)\|v\|_2^4 \right\} \vdash^{v} \left\{ \frac{1}{n}\sum_{i\in[n]} p^2(\langle x_i, v\rangle^2 - \phi) \leqslant \delta \right\},$$

only at the cost of p^2 having degree $\mathcal{O}(1/\delta^2 + \phi)$. This is a qualitative improvement over prior works, and is necessary to the remainder of our argument.

Anti-concentration on the remaining directions. Directions that are not analytically dense include, in particular, sparse directions, and marginals along sparse directions do not behave like the Gaussian distribution. Nevertheless, it is tempting to try to show that the remaining set of directions is small, and perhaps admits a small ε-net. Certainly, the set of sparse directions satisfies this property. However, consider the direction $v = [1/2, 1/2, c/\sqrt{n}, \ldots, c/\sqrt{n}]$, for a fixed constant c. This direction is neither sparse nor analytically dense, and we cannot discard the large or the small coordinates, since they contribute an equal amount of ℓ_2-mass. Surprisingly, we do not know how to obtain an explicit sum-of-squares proof of anti-concentration for such directions. Instead, we show that we can still indirectly certify anti-concentration.

Inspired by this example, let us assume we have the ability to condition indeterminates, i.e. coordinates of v that are large. Given this ability, it is not too hard to show that if we condition on all the coordinates that are at least λ, and denote this set by u_S, the resulting residual vector, $v - u_S$ would have ℓ_∞ norm bounded by λ, and in turn $\|v - u_S\|_4^4 \leqslant \lambda$. Further, u_S will have at most $1/\lambda$ non-zero entries. We could then appeal to univariate anti-concentration on $\langle x, u_S\rangle$ and analytically dense certificates on $\langle x, v - u_S\rangle$. While we do not posess the ability to condition large coordinates of v in the proof system, we show how to execute such a proof strategy by looking at dual objects to sum-of-squares proofs, i.e. *pseudo-distributions*.

Re-weighting pseudo-distributions. Consider a pseudo-distribution μ over a vector valued indeterminate v such that $\|v\|^2 \leqslant 1$. Since v is a unit vector, it does not take values in a discrete set and we cannot condition μ on a specific value (unlike the boolean setting, e.g. [BRS11]). Instead, we show that we can re-weight the pseudo-distribution by a low degree polynomial, such as v_i^{2t}, where v_i is the i-th coordinate of v and t is a large integer, and this morally has the same effect as that of conditioning (we refer the reader to Section 2.2 for background on re-weighting pseudo-distributions).

The main theorem we obtain is that after an iterative re-weighting process (Algorithm 5.2), there exists a new pseudo-distribution μ' and a sparse vector u_S (this is a vector without any indeterminates) such that either $\widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_S\|_2^{2t}\right] \leqslant \eta$, i.e. all the low-degree moments of the residual vector are small, or $\widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_S\|_{2t}^{2t}\right] \leqslant \eta$, i.e. the residual vector is analytically dense. Here η can be made arbitrarily small at the cost of running time, however, t must be $\Omega(\log(d))$, which results in the quasi-polynomial running time in all our results.

The iterative re-weighting proceeds as follows: given the pseudo-distribution μ we find a coordinate v_i such that $\widetilde{\mathbb{E}}_{\mu}\left[v_i^{2t}\right] \geqslant \lambda^{2t}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2t}\right]$. If no such coordinate exists, it's easy to show that $\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_{2t}^{2t}\right]$ is small, since we essentially have an ℓ_∞ bound on v. Otherwise, we re-weight μ by v_i^{2t}, and set the new vector $v = v - \widetilde{\mathbb{E}}_{\mu}[v_i]$. We iterate until there is either no such coordinate or the norm of the residual vector is small.

The analysis has two key components. The first is a scalar re-weighting statement that was obtained by Barak, Kothari and Steurer [BKS17]: Given a bounded indeterminate z, there exists a re-weighting of μ, denoted by μ', such that $\widetilde{\mathbb{E}}_{\mu'}\left[(z-\widetilde{\mathbb{E}}_{\mu}[z])^{2t}\right] \leqslant \varepsilon^{2t}\widetilde{\mathbb{E}}_{\mu}[z]^{2t}$. Using this statement, we can conclude that in a single iteration, the re-weighting process essentially fixes the i-th coordinate. However, since the pseudo-distribution gets re-weighted over and over again, we might undo this conditioning. We instead design a carefully chosen potential that monotonically decreases, showing that the re-weighting process must terminate after a few iterations (see the proof of Theorem 5.1 for details).

Finally, we note that $t = \Omega(\log(d))$ is necessary for our re-weighting strategy. Consider the case where μ is simply the uniform distribution over the standard basis vectors (a special case of a pseudo-distribution). Each element in the support has only one non-zero coordinate, but $\mathbb{E}_{\mu}\left[v_i^{2t}\right]$ remains a negligible fraction of the ℓ_2 norm until $t = \Omega(\log(d))$.

Pseudo-expectations vs. axioms. Given our re-weighting scheme, it is not too hard to show that for the re-weighted pseudo-distribution μ', $\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_4^4\right] \leqslant \eta$. We would now like to invoke our certificates of anti-concentration along the direction $v - u_{\mathcal{S}}$. However, we only obtain the axiom $\{\, \|v-u_{\mathcal{S}}\|_4^4 \leqslant \eta \,\}$ if the pseudo-expectation remains small for all low-degree sum-of-squares polynomials multiplied by $\|v-u_{\mathcal{S}}\|_4^4$. Instead, we show that our certificates used the axiom by only multiplying it by $\|v-u_{\mathcal{S}}\|_2^{2\ell}$, for different integral ℓ, and we can indeed prove that $\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_2^{2\ell}\cdot\|v-u_{\mathcal{S}}\|_4^4\right] \leqslant \eta$ (see Lemma 3.5). This allows us to invoke our certificates for directions that satisfy $\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_4^4\right] \leqslant \eta$.

Objective is bounded under a re-weighting. With the aforementioned components in place, we are ready to analyze the objective value of the anti-concentration program. Let μ be any pseudo-distribution that is consistent with the constraints in Eq. (1). For ease of exposition, we assume that $w_i\langle x_i, v\rangle^2 = 0$, and this constraint can be easily relaxed. Then, there exists a re-weighting of μ and a sparse vector $u_{\mathcal{S}}$ such that either $\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_4^4\right]$ is small or $\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_2^{2t}\right]$ is small. Consider the first case. Since $p^2(0) = 1$, we can bound the objective as follows:

$$\begin{aligned}\frac{1}{n}\widetilde{\mathbb{E}}_{\mu'}\left[\sum_{i\in[n]} w_i\right] &= \frac{1}{n}\widetilde{\mathbb{E}}_{\mu'}\left[\sum_{i\in[n]} w_i p^2(0)\right] \\ &= \frac{1}{n}\widetilde{\mathbb{E}}_{\mu'}\left[\sum_{i\in[n]} w_i p^2(w_i\langle x_i, v\rangle)\right] \\ &\leqslant \frac{1}{n}\widetilde{\mathbb{E}}_{\mu'}\left[\sum_{i\in[n]} p^2(\langle x_i, u_{\mathcal{S}}\rangle + \langle x_i, v-u_{\mathcal{S}}\rangle)\right]\end{aligned} \tag{4}$$

where the second equality follows from p^2 being a sum-of-squares polynomial and our constraint. We can then drop the indicators since they are at most 1. Now, we treat the vector $\langle x_i, u_{\mathcal{S}}\rangle$ as a shift (since this is a fixed vector) and invoke our analytic certificate on the direction $v - u_{\mathcal{S}}$. For most samples $\langle x_i, u_{\mathcal{S}}\rangle$ is bounded, and therefore, the objective value is small (see Lemma 3.4 for details).

Now, consider the second case. We again proceed as Eq. (4) and split into the $p^2(\langle x_i, u_{\mathcal{S}}\rangle + \langle x, v-u_{\mathcal{S}}\rangle)$. We then just expand out the polynomial by writing it in the monomial basis, and show that we can treat all the terms involving $\langle x, v-u_{\mathcal{S}}\rangle$ as additive error. Since all low-degree moments of this vector are bounded, the additive error remains small (see Lemma 3.2). We now

invoke anti-concentration properties on $p^2(\langle x, u_S \rangle)$, but since u_S is sparse, and fixed, we no longer need sum-of-squares certificates, and can simply reduce to the univariate case. Taking a step back, we managed to show that any pseudo-distribution μ can be re-weighted to one for which $\frac{1}{n}\sum_{i\in[n]} w_i$ is small.

Inferring the existence of SoS proof. Semi-definite programming duality implies that if an inequality is true for all pseudo-distributions of degree-k, then there must be a degree-k sum-of-squares proof of the inequality (see Fact 2.9). Thus far, we have managed to show that any pseudo-distribution μ can be re-weighted to obtain the desired inequality, i.e. $\frac{1}{n}\widetilde{\mathbb{E}}_{\mu'}\left[\sum_{i\in[n]} w_i\right] \leqslant \delta$. In general, the existence of a re-weighting need not imply a sum-of-squares proof. However, we use the specific structure of our re-weighting to show that such a proof must indeed exist.

We proceed via contradiction. Let μ be a pseudo-distribution under which $\frac{1}{n}\widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right] > 10\delta$. Then, we re-weight μ by $(\sum_{i\in[n]} w_i)^{2t}$ for a large t, to essentially fix the objective value, and let the resulting pseudo-distribution be μ_1. The scalar fixing lemma described above ensures that the objective value remains large even if we apply low-degree polynomials to the objective. Now, we invoke our re-weighting result on μ_1, we can show that there is a re-weighting of μ_1, say μ_2, such that $\frac{1}{n}\widetilde{\mathbb{E}}_{\mu_2}\left[\sum_{i\in[n]} w_i\right] < \delta/100$. However, since this re-weighting is a low-degree polynomial, it could not have altered the objective value under μ_1 by too much, therefore yielding a contradiction (we refer the reader to Theorem 3.1 for a complete proof). Formally, we can infer a certificate of anti-concentration of the following form:

$$\delta - \frac{1}{n}\sum_{i\in[n]} w_i = \mathrm{sos}(v,w) + \sum_{i\in[n]} q_i(v,w)\Big(w_i(\delta\|v\|^2 - \langle x_i, v\rangle^2)\Big) + \sum_{i\in[n]} r_i(w_i^2 - w_i), \tag{5}$$

where $\mathrm{sos}(v,w), q_i(v,w)$ are sum-of-squares polynomials and r_i are arbitrary low-degree polynomials.

Clustering via certificates. We note that the certificates we obtain do not go via the certificates obtained in prior works that were discussed in (2) and therefore, does not imply the algorithmic applications of anti-concentration in a black-box manner. We describe how to use our certificates to design algorithms. For the purposes of the overview, we consider a simple example: let $X = \{x_i\}_{i\in[n}$ be n samples from a mixture of two *reasonably anti-concentrated* distributions, denoted by $\mathcal{M} = \frac{1}{2}\mathcal{D}(0,\Sigma_1) + \frac{1}{2}\mathcal{D}(0,\Sigma_2)$. The two distributions are concentric (both means are zero) and the mixture is isotropized, so the mixture covariance is I. Further, there is some unknown direction v such that $v^\top\Sigma_1 v \leqslant \delta v^\top\Sigma_2 v$, for some sufficiently small δ. Our goal is to simply recover some direction that continues to separate Σ_1 and Σ_2. Note, we can use such a direction to cluster by simply projecting all the samples along this direction and thresholding them.

To find such a direction we consider the following program:

$$\mathcal{C}_\delta = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i\in[n]} w_i = n/2 \\ \forall i \in [n] & w_i\langle x_i, v\rangle^2 \leqslant c\delta w_i\|v\|^2 \end{Bmatrix} \tag{6}$$

This program is feasible, since we can set the w_i's to indicate the samples from $\mathcal{D}(0,\Sigma_1)$. We then compute a pseudo-distribution that is consistent with $\mathcal{C}_\delta$ and our goal is to obtain a separating

direction. A natural candidate is to compute the covariance matrix of the indeterminates, i.e. $\widetilde{\mathbb{E}}_\mu \left[vv^\top\right]$ and then sample $g \sim \mathcal{N}(0, \widetilde{\mathbb{E}}_\mu \left[vv^\top\right])$ (often referred to as Gaussian rounding). However, there can be several directions where the variance of $\mathcal{D}(0, \Sigma_2)$ is significantly smaller than $\mathcal{D}(0, \Sigma_1)$ and μ could be the uniform distribution over all such directions. In fact, it is not to hard to construct an instance where $\widetilde{\mathbb{E}}_\mu \left[vv^\top\right] = I$ and therefore, our program would not succeed in finding a separating direction.

Instead, we pick one of the samples uniformly at random and re-weight the pseudo-distribution by $\langle x_i, v\rangle^{2t}$. With probability $1/2$, we sample x_i from $\mathcal{D}(0, \Sigma_2)$ and the re-weighting essentially conditions the pseudo-distribution to be supported on directions where Σ_2 has larger variance, and the aforementioned Gaussian rounding scheme succeeds.

A bit more formally, using the certificate from Eq. (5) we prove the following key sum-of-squares inequality (see Lemma 6.4 for a formal statement and proof):

$$\mathcal{C}_\delta \vdash \left\{ \left\langle \Sigma_1, vv^\top \right\rangle^2 \leqslant \mathcal{O}(1) \cdot \left(\delta^2 \|v\|^4 - \sum_{x_i \in \mathcal{D}(0,\Sigma_2)} \underbrace{q_i(w,v) w_i \left(\delta v^\top \Sigma_2 v - \frac{\delta \|v\|_2^2}{100} \right)}_{(7).(1)} \right) \right\}, \quad (7)$$

where the term (7).(1) need not axiomatically be non-negative. Since we compute a pseudo-distribution consistent with $\mathcal{C}_\delta$, we can conclude

$$\left\langle \Sigma_1, \widetilde{\mathbb{E}}_\mu \left[vv^\top\right] \right\rangle^2 \leqslant \widetilde{\mathbb{E}}_\mu \left[\left\langle \Sigma_1, vv^\top \right\rangle^2 \right] \leqslant \delta^2 \|v\|_2^4 - \sum_i \widetilde{\mathbb{E}}_\mu \left[q_i(w,v) w_i \left(\delta v^\top \Sigma_2 v - \frac{\delta \|v\|_2^2}{100} \right) \right].$$

Observe, as long as $\widetilde{\mathbb{E}}_\mu \left[q_i(w,v) v^\top \Sigma_2 v\right] \geqslant 0.1$, we are in good shape, since the additive term disappears, and we have $\left\langle \Sigma_1, \widetilde{\mathbb{E}}_\mu \left[vv^\top\right] \right\rangle \leqslant \delta \|v\|_2^2$, an obtaining a separating direction from here is straight-forward.

As a thought experiment, consider the case where we know the scalar quantity $v^\top \Sigma_2 v$. Then, re-weighting μ by $(v^\top \Sigma_2 v)^{2t}$ outputs a pseudo-distribution μ' such that the expected value of $v^\top \Sigma_2 v$ remains large, even when multiplied by low-degree sum-of-squares polynomials, in particular, the q_i's and therefore $\widetilde{\mathbb{E}}_\mu \left[q_i(w,v) v^\top \Sigma_2 v\right] \geqslant 0.1$ (see Lemma 6.5). Therefore, we can now simply perform Gaussian rounding on $\widetilde{\mathbb{E}}_{\mu'} \left[vv^\top\right]$. Of course, re-weighting by $v^\top \Sigma_2 v$ is as hard as learning Σ_2, but if we pick a uniformly random sample, and condition on x_i being from $\mathcal{D}(0, \Sigma_2)$, in expectation, we are re-weighting $v^\top \Sigma_2 v$ and the above argument continues to work (see Lemma 6.6 for a full proof).

1.4 Related work

Anti-concentration is a vast field and has been subject to intense study in probability theory (e.g. [Erd45; CTV06; Lov10; BK21]) and statistics (e.g. [LS01; CCK15]). Similar to concentration inequalities [Tro15; MP16; RT23], anti-concentration inequalities have found numerous applications in computer science, e.g. [Sim15; KKK19; RY20a; AA11; CR11; Vid12; She12; Nie22]. Among such applications, the works most closely related to ours are the ones exploiting anticoncentration certificates to design approximation algorithms [KKK19; RY20a; BK20; RY20b; DHKK20; BK21; IK22]. Notably, many of these certificates are based on the Sum of Squares (SoS) hierarchy. In particular, these SoS certificates of anti-concentration have been modified in various ways to design statistical algorithms. However, they suffer from two drawbacks – Most of the programs used have been application-specific, requiring ad-hoc analyses designed for the

problem at hand; and for technical reasons, they work only for spherical distributions, e.g. Gaussians. In this work, we present a unified and the most canonical way to certify anti-concentration phenomena for a wide class of distributions that go beyond spherical distributions.

The sum of squares hierarchy (SoS) [Sho87; Nes00; Par00; Gri01] uses insights from proof complexity to design algorithms, leading to the powerful framework known as "proofs to algorithms" [FKP19]. This has been highly effective in theoretical computer science, capturing state-of-the-art approximation algorithms for many problems, e.g. [ARV04; GW95; HSS15; Rag08]. Due to its effectiveness in capturing many algorithmic reasoning techniques [FKP19], lower bounds against SoS is also an active area of study [BHKKMP16; KMOW17; GJJPR20; PR20; Raj22; Kun20; PR22; Jon22; JPRTX22; JPRX23]. In recent years, many works have exploited SoS hierarchy for applications in algorithmic robust statistics [DK23], leading to to breakthrough algorithms for long-standing open problems [BDJKKV22; LM21; Hop20; KKM18; FKP19; KS17b; BK20; BK21; SS17]. Highlights include robustly learning mixtures of high dimensional Gaussians (a long line of works culminating in [BDJKKV22; LM21]), efficient algorithms for the fundamental problems of regression [KKM18], moment estimation [KS17b], clustering [BK20] and subspace recovery [BK21] in the presence of outliers. In recent years, many more applications are being discovered, e.g. in differential privacy [HKM22; GH22; KMV22; HKMN23] and quantum information [BBHKSZ12; BKS17; BLMT23].

Finally, we discuss works related to our applications. The literature on clustering and regression is incredibly vast. These fundamental problems have many connections to other standard statistical tasks such as mixture modeling, density estimation, latent variable modeling, etc. For the most closely related works on clustering and regression, see [BK20; LL22; DKKLT22; KKK19; DKS17b; KS17a; HL18; CLS20; RY20a] and citations therein.

Organization of the paper. In Section 2, we present the preliminaries of the sum-of-squares framework and define the distributional properties we use. In Section 3, we present our main anti-concentration program and state our main theorem, with the proof details filled out in Sections 3, 4 and 5. We then show the application of our program to spectral clustering in Section 6, k-clustering in Section 7 and list decodable regression in Section 8.

2 Preliminaries

2.1 The Sum-of-Squares framework

We now provide an overview of the Sum-of-Squares proof system. We closely follow the exposition as it appears in the lecture notes of Barak [Bar].

Pseudo-Distributions. A discrete probability distribution over $\mathbb{R}^m$ is defined by its probability mass function, $D\colon \mathbb{R}^m \to \mathbb{R}$, which must satisfy $\sum_{x\in \mathrm{supp}(D)} D(x) = 1$ and $D \geqslant 0$. We extend this definition by relaxing the non-negativity constraint to merely requiring that D passes certain low-degree non-negativity tests. We call the resulting object a pseudo-distribution.

Definition 2.1 (Pseudo-distribution)**.** A *degree-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^m \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x)p(x)^2 \geqslant 0$ for every polynomial p of degree at most $\ell/2$, where the summation is over all x in the support of D.

Next, we define the related notion of pseudo-expectation.

Definition 2.2 (Pseudo-expectation)**.** The *pseudo-expectation* of a function f on $\mathbb{R}^m$ with respect to a pseudo-distribution μ, denoted by $\widetilde{\mathbb{E}}_{\mu(x)}[f(x)]$, is defined as

$$\widetilde{\mathbb{E}}_{\mu(x)}[f(x)] = \sum_x \mu(x) f(x).$$

We use the notation $\widetilde{\mathbb{E}}_{\mu(x)}\left[(1, x_1, x_2, \ldots, x_m)^{\otimes \ell}\right]$ to denote the degree-ℓ moment tensor of the pseudo-distribution μ. In particular, each entry in the moment tensor corresponds to the pseudo-expectation of a monomial of degree at most ℓ in x.

Definition 2.3 (Constrained pseudo-distributions)**.** Let μ be a degree-ℓ pseudo-distribution over $\mathbb{R}^d$. Let $\mathcal{A} = \{ p_1 \geqslant 0, p_2 \geqslant 0, \ldots p_m \geqslant 0 \}$ be a system of m polynomial inequality constraints of degree at most r. We define μ to satisfy $\mathcal{A}$ at degree $\ell \geqslant 1$ if for every subset $\mathcal{S} \subset [m]$, and every sum-of-squares polynomial q such that $\deg(q) + \sum_{i \in \mathcal{S}} \max(p_i, r) \leqslant \ell$, $\widetilde{\mathbb{E}}_{\mu}\left[q \prod_{i \in \mathcal{S}} p_i\right] \geqslant 0$. Further, we define μ to *approximately satisfy* the system of constraints $\mathcal{A}$ if the above inequalities are satisfied up to additive error $2^{-n^\ell} \|q\| \prod_{i \in \mathcal{S}} \|p_i\|$, where $\|\cdot\|$ denotes the euclidean norm of the coefficients of the polynomial, represented in the monomial basis.

Crucially, there's an efficient separation oracle for moment tensors of constrained pseudo-distributions.

Fact 2.4 ([Sho87; Nes00; Par00; Gri01])**.** *For any $m, \ell \in \mathbb{N}$, the following set has a $m^{O(\ell)}$-time weak separation oracle (in the sense of [GLS81]):*

$$\left\{ \widetilde{\mathbb{E}}_{\mu(x)}\left[(1, x_1, x_2, \ldots, x_m)^{\otimes \ell}\right] \,\middle|\, \mu \text{ is a degree-}\ell \text{ pseudo-distribution over } \mathbb{R}^m \right\}$$

This fact, together with the equivalence of weak separation and optimization [GLS81] forms the basis of the sum-of-squares algorithm, as it allows us to efficiently approximately optimize over pseudo-distributions.

Given a system of polynomial constraints, denoted by $\mathcal{A}$, we say that it is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. Then, the following fact follows from Fact 2.4 and [GLS81]:

Fact 2.5 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(m+t)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system $\mathcal{A}$ of t polynomial constraints in m variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately (see Definition 2.3)* [4].

We now define basic facts for pseudo-distributions, which extend facts that hold for standard probability distributions, which can be found in several prior works listed above.

Fact 2.6 (Cauchy-Schwarz for Pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Then, for any degree d pseudo-distribution μ, $\widetilde{\mathbb{E}}_{\mu}[fg] \leqslant \sqrt{\widetilde{\mathbb{E}}_{\mu}[f^2]} \cdot \sqrt{\widetilde{\mathbb{E}}_{\mu}[g^2]}$.*

Fact 2.7 (Hölder's Inequality for Pseudo-Distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Fix $t \in \mathbb{N}$. Then, for any degree dt pseudo-distribution μ,*

$$\widetilde{\mathbb{E}}_{\mu}\left[f^{t-1} g\right] \leqslant \left(\widetilde{\mathbb{E}}_{\mu}\left[f^t\right]\right)^{\frac{t-1}{t}} \cdot \left(\widetilde{\mathbb{E}}_{\mu}\left[g^t\right]\right)^{\frac{1}{t}}.$$

In particular, for all even integers k, $\widetilde{\mathbb{E}}_{\mu}[f]^k \leqslant \widetilde{\mathbb{E}}_{\mu}\left[f^k\right]$.

[4] Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(m+t)^{O(1)}$.

Sum-of-squares proofs. Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in the indeterminate x. Given constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$, a *sum-of-squares proof* that implies the identity $\{g \geqslant 0\}$ consists of polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S^2 \cdot \prod_{i \in S} f_i$$

We say that this proof has *degree-ℓ* if for every set $S \subseteq [m]$, the polynomial $p_S^2 \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \left|\frac{}{\ell}\right. \{g \geqslant 0\} \tag{8}$$

For all polynomials $f, g \colon \mathbb{R}^n \to \mathbb{R}$ and for all functions $F \colon \mathbb{R}^n \to \mathbb{R}^m$, $G \colon \mathbb{R}^n \to \mathbb{R}^k$, $H \colon \mathbb{R}^p \to \mathbb{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have the following inference rules.

The first one derives new inequalities by addition/multiplication:

$$\frac{\mathcal{A} \left|\frac{}{\ell}\right. \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \left|\frac{}{\ell}\right. \{f + g \geqslant 0\}}, \frac{\mathcal{A} \left|\frac{}{\ell}\right. \{f \geqslant 0\}, \mathcal{A} \left|\frac{}{\ell'}\right. \{g \geqslant 0\}}{\mathcal{A} \left|\frac{}{\ell+\ell'}\right. \{f \cdot g \geqslant 0\}} \qquad \text{(Addition/Multiplication Rule)}$$

The next one derives new inequalities by transitivity:

$$\frac{\mathcal{A} \left|\frac{}{\ell}\right. \mathcal{B}, \mathcal{B} \left|\frac{}{\ell'}\right. C}{\mathcal{A} \left|\frac{}{\ell \cdot \ell'}\right. C} \qquad \text{(Transitivity Rule)}$$

Finally, the last rule derives new inequalities via substitution:

$$\frac{\{F \geqslant 0\} \left|\frac{}{\ell}\right. \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \left|\frac{}{\ell \cdot \deg(H)}\right. \{G(H) \geqslant 0\}} \qquad \text{(Substitution Rule)}$$

Sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 2.8 (Soundness)**.** *If $\mu \left|\models_{\ell}\right. \mathcal{A}$ for a level-ℓ pseudo-distribution μ and there exists a sum-of-squares proof $\mathcal{A} \left|\frac{}{\ell'}\right. \mathcal{B}$, then $\mu \left|\models_{\ell \cdot \ell' + \ell'}\right. \mathcal{B}$.*

If the pseudo-distribution D satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \left|\frac{}{r'}\right. B$ (number of bits required to write down the proof). In our applications, the bit complexity of all sum of squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 2.9 (Completeness)**.** *Suppose $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^n x_i^2 \leqslant B\}$ for some finite B. Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \left|\models_{r}\right. \mathcal{A}$ also satisfies $D \left|\models_{r'}\right. \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \left|\frac{}{d}\right. \{g \geqslant -\varepsilon\}$.*

Next, we recall the fact that any univariate polynomial inequality admits a sum-of-squares proof over the reals.

Fact 2.10 (Univariate Polynomial Inequalities admit SoS Proofs, [Lau09])**.** *Given a univariate degree-t polynomial p such that for all $x \in \mathbb{R}$, $p(x) \geqslant 0$, we have $\left|\frac{x}{d}\right. \{ p(x) \geqslant 0 \}$. Further, if for all $a \leqslant x \leqslant b$, we have $p(x) \geqslant 0$, then $\{ x \geqslant a, x \leqslant b \} \left|\frac{x}{d}\right. \{ p(x) \geqslant 0 \}$.*

Fact 2.11. *(Non-negative Quadratic Polynomials Inequalities admit SoS Proofs) Given a multivariate polynomial $p(x_1, x_2, \ldots, x_m) \in \mathbb{R}[x_1, \ldots, x_m]$ such that p has degree 2 and $p(x_1, x_2, \ldots, x_m) \geqslant 0$ for all $x_1, x_2, \ldots x_m \in \mathbb{R}$. Then, $\left|\frac{x}{2}\right. \{ p(x) \geqslant 0 \}$.*

Proof. Note that there is a unique $(m+1) \times (m+1)$ symmetric matrix M such that if we let $v(x) = (1, x_1, \ldots, x_m)^\dagger$ then

$$p(x_1, \ldots, x_m) = v(x)^\dagger M v(x) .$$

Now we claim that M must be PSD. Indeed for any vector $v = (v_1, \ldots, v_{m+1}) \in \mathbb{R}^{m+1}$, if $v_1 \neq 0$, then we can consider plugging in $x_1 = v_2/v_1, \ldots, x_m = v_{m+1}/v_1$ and since $p(x_1, \ldots, x_m) \geqslant 0$ we get that $v^\dagger M v \geqslant 0$. If $v_1 = 0$, then we can consider plugging in $x_1 = cv_2, \ldots, x_m = cv_{m+1}$ and take the limit as $c \to \infty$ and again we deduce that $v^\dagger M v \geqslant 0$. Thus, M must be PSD. We can now write $M = \sum_{i=1}^{m+1} u_i u_i^\dagger$ for some vectors $u_i \in \mathbb{R}^{m+1}$. Thus, we can write

$$p(x_1, \ldots, x_m) = v(x)^\dagger M v(x) = \sum_{i=1}^{m+1} \langle u_i, v(x) \rangle^2$$

which is a degree-2 SoS polynomial and we are done. □

Basic Sum-of-Squares Proofs. Next, we use the following basic facts regarding sum-of-squares proofs. For further details, we refer the reader to a recent monograph [FKP+19].

Fact 2.12 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix and v be a vector in $\mathbb{R}^d$. Then,*

$$\left|\frac{v}{2}\right. \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\} .$$

Fact 2.13 (SoS Almost Triangle Inequality)**.** *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then,*

$$\left|\frac{f_1, f_2, \ldots, f_r}{2t}\right. \left\{ \left(\sum_{i \leqslant r} f_i \right)^{2t} \leqslant r^{2t-1} \left(\sum_{i=1}^{r} f_i^{2t} \right) \right\} .$$

Fact 2.14 (SoS Hölder's Inequality)**.** *Let $w_1, \ldots w_n$ be indeterminates and let $f_1, \ldots f_n$ be polynomials of degree m in vector valued variable x. Let k be a power of 2. Then,*

$$\{ w_i^2 = w_i, \forall i \in [n] \} \left|\frac{x, w}{2km}\right. \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} w_i f_i \right)^k \leqslant \left(\frac{1}{n} \sum_{i=1}^{n} w_i \right)^{k-1} \left(\frac{1}{n} \sum_{i=1}^{n} f_i^k \right) \right\} .$$

Fact 2.15 (Almost square-root)**.** *Given a scalar indeterminate v, $\{ v^2 \leqslant 1 \} \left|\frac{v}{}\right. \{ -1 \leqslant v \leqslant 1 \}$.*

Proof. We know that $\left\{ (1-v)^2 = 1 + v^2 - 2v \geqslant 0 \right\}$ and $\left\{ (1+v)^2 = 1 + v^2 + 2v \geqslant 0 \right\}$. Further, by assumption, $\{ 1 - v^2 \geqslant 0 \}$ and by the addition rule we have $\{ 2 + 2v \geqslant 0 \}$ and $\{ 2 - 2v \geqslant 0 \}$. Rearranging yields the claim. □

2.2 Re-weighting Pseudo-distributions

Let μ be a pseudo-distribution of degree $\geqslant t$ on d-dimensional vector valued indeterminate v. Let q be a sum-of-squares polynomial in v of degree $t' < t$. Then, μ' defined by $\mu'(v) = \mu(v) \cdot q(v)$ is called the *reweighting* of μ by the polynomial q. It is easy to observe that μ' is a pseudo-distribution of degree at least $t - t'$. Further, if μ satisfying a polynomial equality constraint $\{r = 0\}$ and degree of r is at most $t - t'$, then, so does μ'. See [BKS17] for background on reweighting pseudo-distributions.

We also require the following straightforward fact that ensures subsequent re-weightings of pseudo-distributions do not decrease the pseudo-expectation of the indeterminate:

Fact 2.16 (Scalar Re-weightings are Monotone)**.** *Let μ be a pseudo-distribution over an indeterminate z. Let μ' be a re-weighting of μ obtained by using the polynomial $z^{2\ell}$, for some $\ell \in \mathbb{N}$. Then, $\widetilde{\mathbb{E}}_{\mu'}\left[z^2\right] \geqslant \widetilde{\mathbb{E}}_{\mu}\left[z^2\right]$, given μ is has degree at least $3\ell + 3$.*

Proof. Recall, by the definition of re-weighting a pseudo-distribution, we have

$$\widetilde{\mathbb{E}}_{\mu'}\left[z^2\right] = \frac{\widetilde{\mathbb{E}}_{\mu}\left[z^2 \cdot z^{2\ell}\right]}{\widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell}\right]}. \tag{9}$$

By Hölder's inequality for pseudo-distributions (Fact 2.7 applied with $f = z^2$ and $k = \ell$), we have $\widetilde{\mathbb{E}}_{\mu}\left[z^2\right] \leqslant \left(\widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell}\right]\right)^{1/\ell}$ and thus

$$\widetilde{\mathbb{E}}_{\mu}\left[z^2\right]\widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell}\right] \leqslant \widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell}\right]^{\frac{\ell+1}{\ell}} \leqslant \left(\widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell+2}\right]^{\frac{2\ell}{2\ell+2}} \cdot \widetilde{\mathbb{E}}_{\mu}\left[1\right]^{\frac{2}{2\ell+2}}\right)^{\frac{\ell+1}{\ell}} = \widetilde{\mathbb{E}}_{\mu}\left[z^{2\ell+2}\right], \tag{10}$$

where the second inequality follows from applying Hölder's inequality again. Therefore, we can rearrange Equation (10) and conclude $\widetilde{\mathbb{E}}_{\mu'}\left[z^2\right] \geqslant \widetilde{\mathbb{E}}_{\mu}\left[z^2\right]$. □

Fact 2.17 (Bounding moments of subtracting a coordinate)**.** *Let v be a vector-valued indeterminate over $\mathbb{R}^d$ and $y \in \mathbb{N}$. Then, for any $i \in [d]$,*

$$\left|\frac{v}{2y}\right. \left\{ \left(\|v\|_2^2 - v_i^2\right)^y \leqslant \|v\|_2^{2y} - v_i^{2y} \right\}$$

Proof. We use the fact that $\left|\frac{a,b}{2t}\right. \left\{ a^{2t} + b^{2t} \leqslant (a+b)^{2t} \right\}$ as follows:

$$\left|\frac{v}{2y}\right. \left\{ \left(\|v\|_2^2 - v_i^2\right)^{2y} + v_i^{2y} \leqslant \left(\|v\|_2^2 - v_i^2 + v_i^2\right)^{2y} = \|v\|_2^{2y} \right\} \tag{11}$$

Rearranging yields the claim. □

2.3 Distributions

Definition 2.18 (Certifiable-Hypercontractivity of linear forms)**.** An distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to be h-certifiably c_h-hypercontractive if there's a degree-h sum-of-squares proof of the following unconstrained polynomial inequality in d-dimensional vector-valued indeterminate v:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \langle x, v\rangle^{2h} \leqslant (c_h)^{2h} \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \langle x, v\rangle^2\right)^h,$$

for h being an even integer. A set of points $X \subseteq \mathbb{R}^d$ is said to be (c_h, h)-certifiably hypercontractive if the uniform distribution on X is h-certifiably c_h-hypercontractive.

Definition 2.19 (Certifiable Hypercontractivity of quadratic forms)**.** An isotropic distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to be h-certifiably c_h-hypercontractive if there's a degree h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^{2h} \leqslant (c_h)^{2h}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^2\right)^h,$$

for h being an even integer. A set of points $X \subseteq \mathbb{R}^d$ is said to be (c_h, h)-certifiably hypercontractive if the uniform distribution on X is h-certifiably c_h-hypercontractive.

Definition 2.20 (Almost k-wise independent)**.** For an integer $k \geqslant 1$, define a distribution $\mathcal{D}$ to be almost k-wise independent if for all subsets $S \subseteq [n]$ of size at most k, we have $\mathbb{E}_{x\sim\mathcal{D}}[\prod_{i\in S} x_i^2] = (1 \pm \gamma_S)\prod_{i\in S}\mathbb{E}_{x\sim\mathcal{D}}[x_i^2]$ where $\gamma_S = o_d(1)$.

Definition 2.21 (Strictly sub-exponential distribution)**.** A distribution $\mathcal{D}$ in isotropic position is ε-strictly sub-exponential if for all $v \in \mathbb{R}^d$,

$$\Pr[|\langle x, v\rangle| \geqslant t\sqrt{v^\top \Sigma v}] \leqslant \exp\left(-t^{1+\varepsilon}/c\right),$$

for some constant c.

Fact 2.22 (Sub-Exponential Distributions are bounded)**.** *Let $x \sim \mathcal{D}$ be a sample from a d-dimensional sub-exponential distribution. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, for all $u \in \mathbb{R}^d$, $\langle x, u\rangle^2 \leqslant \log(1/\delta)\|u\|_2^2$.*

The main definition we consider in this paper is as follows:

Definition 2.23 ((δ, k, ε)-Reasonably anti-concentrated distributions)**.** A distribution $\mathcal{D}$ is (δ, k, ε)-reasonably anti-concentrated if it satisfies the following:

1. $\mathcal{D}$ is δ-anti-concentrated
2. $\mathcal{D}$ is ε-strictly sub-exponential.
3. $\mathcal{D}$ has $\exp\left((1/\delta^2)^{\exp(1/\delta^4)}\right)$-certifiably $O(1)$-hypercontractive linear forms.
4. $\mathcal{D}$ is almost k-wise independence.

The following lemma shows that bounded product distributions, a natural class of non-spherical distributions are reasonably anti-concentrated.

Lemma 2.24 (Product Distributions)**.** *A δ-anti-concentrated product distribution $\mathcal{D}$ with strictly sub-exponential coordinate marginals is reasonably anti-concentrated.*

Proof. The coordinates are k-wise independent by definition. To show strict sub-exponentiality, we use an analogue of the results from [Ver20, Lemma 3.4.2] which states that sub-gaussian norm of the vector is bounded by the maximum sub-gaussian norm of each coordinate (upto a constant). This argument directly extends to sub-exponential norm (for instance, see [Ver20, Exercise 2.7.3]). Certifiable hypercontractivity of linear forms of product distributions follows from [KS18]. □

2.4 Polynomial approximation to the box indicator

Our arguments hinge on polynomial approximations to the box indicator function $\mathbb{1}(x \in [\zeta - \delta, \zeta + \delta])$. Standard approximation theory provides us versatile tools to construct such polynomials and indeed, this has been used extensively in related prior works.

In this work, we work with the following explicit polynomial:

Definition 2.25 (Box Indicator Polynomial)**.** For $\eta, \zeta, \delta > 0$ and $L \geqslant 1$, choose an even d and let p be a degree $O(d)$ polynomial such that

$$p(x) = \eta T_d\left(\frac{(x-\zeta)^2}{L^2} - 1 - \delta^2\right),$$

where T_d is the d-th Chebyshev polynomial of the first kind. We assume that the constants are chosen sufficiently small for the following facts to hold.

Fact 2.26. *If $1 \leqslant x \leqslant 1.4$, and we write $x = 1 + \varepsilon$, then we have $T_d(x) = e^{\Theta(d\sqrt{\varepsilon})}$*

Proof. Note that for $y = O(1)$, $\cosh(y) = 1 + O(y^2)$. Therefore, setting $\cosh(y) = x$ gives $\operatorname{arcCosh} x = \Theta(\sqrt{\varepsilon})$ which implies

$$T_d(x) = \cosh(d \operatorname{arcCosh} x) = \frac{e^{d \operatorname{arcCosh} x} + e^{-d \operatorname{arcCosh} x}}{2} = e^{\Theta(d\sqrt{\varepsilon})}$$

□

Lemma 2.27. *When $|x - \zeta| \leqslant \Delta L < \delta L$, we have that $p(x) \geqslant \Omega(\eta e^{\Theta(d\sqrt{\delta^2 + \Delta^2})})$.*

Proof. When $x \in (\zeta - \Delta L, \zeta + \Delta L)$, we have $p(x) = \eta T_d(y)$ where y lies in the range $[-1 - \delta^2, -1 - \delta^2 + \Delta^2]$ but T_d monotonically decreases in this range, therefore $p(x) \geqslant \eta T_d(-1 - \delta^2 + \Delta^2) \geqslant \eta e^{\Theta(d\sqrt{\delta^2+\Delta^2})}$. □

Lemma 2.28. *For constants η, ζ, δ, L, the coefficients of $p(x)$ are bounded by $2^{O(d)}$.*

Proof. We have

$$T_d(y) = \sum_{k=0}^{d/2} \binom{d}{2k} (y^2 - 1)^k y^{n-2k}$$

Note that we have each binomial coefficient above being at most 2^d. Finally, we plug in $y = \frac{(x-\zeta)^2}{L^2} - 1 - \delta^2$ and expand further using the binomial theorem. Using the same bounds and putting them together implies the result. □

Fact 2.29. *If $x \leqslant \zeta - \sqrt{2 + \delta^2}L$ or $x \geqslant \zeta + \sqrt{2 + \delta^2}L$, we have $p(x) \leqslant \eta \left(2\frac{x-\zeta}{L}\right)^{2d}$*

Proof. We have $p(x) = \eta T_d(y)$ where $y = (x - \zeta)^2/L^2 - 1 - \delta^2$. For the given range of x, we have $y \geqslant 1$ and therefore,

$$\begin{aligned} T_d(y) &= \frac{1}{2}((y - \sqrt{y^2 - 1})^d + (y + \sqrt{y^2 - 1})^d) \\ &\leqslant (y + \sqrt{y^2 - 1})^d \\ &\leqslant (2y)^d \end{aligned}$$

$$\leqslant \left(2\frac{x-\zeta}{L}\right)^{2d}$$

which implies the bound. □

Lemma 2.30 (Gaussian anti-concentration with shifts)**.** *Given $\delta, \zeta > 0$, let $L = 1/\delta + \zeta, \eta = \delta/L$, and let p be as above. Then,*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}\left[p^2(x-\zeta)\right] \leqslant \mathcal{O}(\delta),$$

when $d = \Omega\Big(\log\Big(\frac{1}{\eta\delta} + \frac{\zeta}{\delta}\Big) \cdot \Big(\frac{1}{1/\delta^2} + \zeta^2\Big)\Big)$.

Proof. The proof idea is as follows. We split $\mathbb{R}$ into the following intervals: $\mathcal{I}_0 = (-\infty, \zeta - \sqrt{2+\delta^2}L), \mathcal{I}_1 = [\zeta - \sqrt{2+\delta^2}L, \zeta - \delta L], \mathcal{I}_2 = (\zeta - \delta L, \zeta + \delta L], \mathcal{I}_3 = (\zeta + \delta L, \zeta + \sqrt{2+\delta^2}L]$ and $\mathcal{I}_4 = (\zeta + \sqrt{2+\delta^2}L, \infty)$. For intervals $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, we use the bound $\mu(x) \leqslant 1$, where $\mu(x)$ is the pdf at x. For $\mathcal{I}_0$ and $\mathcal{I}_4$, we use the tail estimate.

Firstly, if $\delta^2 \leqslant 0.1$,

$$p(\zeta) = \eta T_d\left(-1-\delta^2\right) \leqslant \eta e^{\Theta(d\delta)}$$

For the interval $\mathcal{I}_2$, $p(x) = \eta T_d(y)$ where y lies in the range $[-1-\delta^2, -1]$ but T_d monotonically decreases in this range, therefore $p(x)$ lies in the interval $[\eta, \eta e^{\Theta(d\delta)}]$, implying

$$\int_{x\in\mathcal{I}_2} p^2(x)\mu(x)dx \leqslant \int_{\mathcal{I}_2} 1dx \leqslant 2\delta L.$$

Next, observe that for $x \in \mathcal{I}_1 \cup \mathcal{I}_3$, we have $p(x) = \eta T_d(y)$ where $y \in [-1,1]$ but for such a y, we have $T_d(y) \in [-1,1]$ (simply because $T_d(\cos\theta) = \cos(d\theta)$), which means $p(x) \in [-\eta, \eta]$. Therefore,

$$\int_{x\in\mathcal{I}_1} p^2(x)\mu(x)dx + \int_{x\in\mathcal{I}_3} p^2(x)\mu(x)dx \leqslant \int_{x\in\mathcal{I}_1\cup\mathcal{I}_3} \eta^2\mu(x)dx \leqslant 4\eta^2(L+\zeta).$$

Finally, for $\mathcal{I}_0 \cup \mathcal{I}_4$, it follows from [KKK19, Fact A.3], and that in this interval, $T_d\left(\left(\frac{x-\zeta}{L}\right)^2 - 1 - \delta^2\right) \leqslant \left(2\frac{x-\zeta}{L}\right)^{2d}$,

$$\begin{aligned}
\int_{x\in\mathcal{I}_0\cup\mathcal{I}_4} p^2(x)\mu(x)dx &\leqslant \int_{x\in\mathcal{I}_0\cup\mathcal{I}_4} \frac{\eta^2 2^{4d}}{L^{4d}}(x+\zeta)^{4d}\exp(-|x|^\alpha)dx \\
&\leqslant \int_{x\in\mathcal{I}_0\cup\mathcal{I}_4} \frac{\eta^2 2^{8d}}{L^{4d}}\cdot\left(\zeta^{4d} + x^{4d}\right)\exp(-|x|^\alpha)dx \\
&\leqslant \frac{\eta^2 2^{8d}\zeta^{4d}}{L^{4d}}\int_{x\in\mathcal{I}_0}\exp(-|x|^a)dx + \frac{\eta^2 2^{2d}}{L^{4d}}\exp(-L^\alpha)\left(L^{8d} + (32d/\alpha)^{4d/\alpha}\right)
\end{aligned}$$

Then, combining all bounds above, and setting $L = 1/\delta + \zeta$, and $\eta = \delta/L$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[p^2(x)\right] &\leqslant p(0) + 2\int_{x\in\mathcal{I}_1} p^2(x)\mu(x)dx + 2\int_{x\in\mathcal{I}_2} p^2(x)\mu(x)dx + \int_{x\in\mathcal{I}_3} p^2(x)\mu(x)dx \\
&\leqslant \eta e^{\Theta(d\delta)} + 2\delta L + \eta^2(L+\zeta) + \frac{\eta^2 2^{2d}}{L^{4d}}\exp(-L^\alpha)\left(L^{8d} + (32d/\alpha)^{4d/\alpha}\right) + \frac{\eta^2 2^{2d}\zeta^{4d}}{L^{4d}} \\
&\leqslant O(\delta),
\end{aligned}$$

when $d = \Omega(\log(1/\eta)(L^2)) = \Omega\Big(\log\Big(\frac{1}{\eta\delta} + \frac{\zeta}{\delta}\Big) \cdot \Big(\frac{1}{1/\delta^2} + \zeta^2\Big)\Big)$. □

We also have the following fact, which can be shown using the same argument as [KKK19].

Fact 2.31 (Anti-Concentration for Sub-exponential distributions). *Given samples $\{ x_i \}_{i\in[n]}$ from a δ anti-concentrated strictly sub-exponential distribution $\mathcal{D}$, it follows that*

$$\frac{1}{n}\sum_{i\in[n]} p^2\Big(\langle x_i, v\rangle^2\Big) \leqslant \delta.$$

3 Anti-concentration program

In this section, we introduce a polynomial system that certifies an upper bound on anti-concentration for a set of n samples drawn from a *reasonably anti-concentrated* distribution (Definition 2.23). Formally, given n i.i.d. samples $\{ x_i \}_{i\in[n]}$ drawn from a *reasonably anti-concentrated* distribution $\mathcal{D}$ with mean 0 and covariance I, and $\delta > 0$, consider the following program:

$$\mathcal{A}_\delta = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ \forall i \in [n] & w_i\langle x_i, v\rangle^2 \leqslant w_i\delta^2 \\ & \|v\|_2^2 = 1 \end{Bmatrix} \tag{12}$$

Intuitively, w_i are indicators indicating the concentrated points. When $w_i = 1$, we must have $\langle x_i, v\rangle^2 \leqslant \delta^2$ which ensures that x_i is concentrated in the direction v, where we search over unit vectors v. The main theorem we establish is that we can certify an upper bound of $O(\delta)$ on $\sum w_i$ over all feasible regions of the program above. Formally,

Theorem 3.1 (Certifying Anti-Concentration). *Given $\delta > 0$, and $n \geqslant n_0$ samples $\{ x_1, x_2, \ldots, x_n \} \subseteq \mathbb{R}^d$ for some $n_0 \geqslant d^t$, sampled from a $(\delta, \exp(1/\delta^2), O(1))$ reasonably anti-concentrated distribution $\mathcal{D}$ (Definition 2.23), let $t = \mathcal{O}\left(\frac{\log(d)}{\exp(\delta^4)^{\exp(1/\delta^2)}}\right)$. Then, there is a degree-t certificate of anti-concentration, i.e.*

$$\mathcal{A}_\delta \vdash_t^{w,v} \left\{ \frac{1}{n}\sum_{i\in[n]} w_i \leqslant \delta \right\},$$

At a high level, we do not provide a direct sum-of-squares proof of this inequality. Instead, we reason about degree-t pseudo-distributions to infer the existence of a sum-of-squares proof (see Fact 2.9). Given a degree-t pseudo-distribution μ that is consistent with the constraint system $\mathcal{A}_\delta$ in Equation (12), we perform several re-weightings that correspond to conditions on entries in v such that either the resulting vector is analytically dense (i.e. the q-norm is bounded by a small constant times the 2-norm) or $1/\exp(1/\delta)$ moments of ℓ_2 norm are small. Then, we use the analytic properties of the re-weighted pseudo-distribution to certify an upper bound on the objective in (12).

More concretely, we abstract out three lemmas that help us complete the proof of Theorem 3.1. We defer the proofs of these lemmas to the subsequent subsections.

We begin by showing that if the re-weighting from Algorithm 5.2 outputs a pseudo-distribution μ' and a constant u_S such that the pseudo-moments of $\|v - u_S\|_2$ are small, then we can simply decompose the polynomial approximation to the box indicator function and appeal to anti-concentration along the direction u_S. Formally, we have the following lemma:

Lemma 3.2 (Decomposing the box-indicator for *almost-sparse* directions)**.** *Given $1 > \gamma > 0$, let $t \in \mathbb{N}$, and let p be the degree-$2t$ polynomial from Definition 2.25. Let $\{ x_i \}_{i\in[n]}$ be n i.i.d. samples from a reasonably anti-concentrated distribution $\mathcal{D}$ (see Definition 2.23). Let μ' be a degree-ϕ pseudo-distribution such that $\phi \geqslant 4t$ and let u_S be a fixed constant vector supported on $|S| \ll d$ coordinates. If for all $\ell \in [t]$, $\widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_S\|_2^{2\ell}\right] \leqslant \gamma^{2\ell}$, then*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\frac{1}{n}\sum_{i\in[n]} p^2\Big(\langle x_i, v\rangle^2\Big)\right] \leqslant \frac{1}{n}\sum_{i\in[n]} p^2\Big(\langle x_i, u_S\rangle^2\Big) + O\Big(2^{O(t)} c_{2t}\gamma\Big),$$

where c_{2t} is the hypercontractivity constant of $\mathcal{D}$.

Next, we show that if instead, the re-weighting from Algorithm 5.2 outputs a pseudo-distribution such that the pseudo-expectation of $\|v - u_S\|_{2z}^{2z}$ is small, we can invoke our sum-of-squares certificate of anti-concentration with shifts from Section 4 to certify a bound on the box-indicator polynomial.

Lemma 3.3 (Anti-Concentration with Shifts)**.** *Given $\zeta > 0$, a pseudo-distribution μ of degree $t = \Omega\left(\frac{\log(d)}{\exp(\zeta^4)^{\exp(1/\zeta^4)}}\right)$, and a vector $u_S \in \mathbb{R}^d$ supported on $|S|$ coordinates for some $S \subseteq [d]$, let $\{ x_i \}_{i\in[n]}$ be a set of n points such that for all $i \in [n]$, $\langle x_i, u_S\rangle^2 \leqslant \mathcal{O}(\log(1/\zeta))$. Let q be the polynomial from Theorem 4.1 with degree $\mathcal{O}(\exp(1/\zeta^2))$. If*

$$\widetilde{\mathbb{E}}_{\mu} \left[\|v - u_S\|_{2z}^{2z}\right] \leqslant \zeta^{2z}\widetilde{\mathbb{E}}_{\mu} \left[\|v - u_S\|_2^{2z}\right],$$

for $z = \Omega\big(\log(d)\exp(1/\zeta^2)\big)$, then we have

$$\widetilde{\mathbb{E}}_{\mu} \left[\frac{1}{n}\sum_{i\in[n]} q^2(\langle x_i, v\rangle)\right] \leqslant \zeta \cdot \widetilde{\mathbb{E}}_{\mu} \left[\|v - u_S\|^2\right] + \frac{1}{\exp(1/\zeta^2)}$$

Combining Lemmas 3.2 and 3.3, we can now obtain the following re-weighting statement:

Lemma 3.4 (Re-weighting certifies Anti-Concentration)**.** *Given $0 < \delta < 1$ and a degree t pseudo-distribution μ over a vector valued indeterminate $v \in \mathbb{R}^d$, there exists a re-weighting of μ, denoted by μ' such that*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\sum_{i\in[n]} w_i\right] < \delta n,$$

if $t = \Omega\left(\log(d)\left(\exp\big(1/\delta^4\big)^{\exp\left(1/\delta^4\right)}\right)\right)$.

We defer the proofs of Lemmas 3.2 to 3.4 to subsequent sections. We are now ready to complete the proof of our main Theorem 3.1 by inferring the existence of a sum-of-squares proof from the existence of the re-weighting in Lemma 3.4. We note that such an implication is not true in general, and we heavily exploit the structure of our re-weighting strategy in the following proof.

Proof of Theorem 3.1. Assume for the sake of contradiction that there exists some degree $t = \Omega\left(\log(d) \cdot \left(\exp\big(1/\delta^4\big)^{\exp\left(1/\delta^4\right)}\right)\right)$ pseudo-distribution μ such that $\widetilde{\mathbb{E}}_{\mu} \left[\sum_{i\in[n]} w_i\right] > \delta n$. We now re-weight μ by $\left(\frac{\sum_{i\in[n]} w_i}{n}\right)^q$ for some $q \geqslant \Omega(t)$ to obtain μ', which essentially fixes the value

of $\sum_{i\in[n]} w_i$. It follows from Lemma 5.3 that

$$\widetilde{\mathbb{E}}_{\mu'}\left[\left(\frac{\sum_{i\in[n]} w_i}{n} - \widetilde{\mathbb{E}}_{\mu}\left[\frac{\sum_{i\in[n]} w_i}{n}\right]\right)^{2q}\right] \leqslant \varepsilon^{2q}\left(\widetilde{\mathbb{E}}_{\mu}\left[\frac{\sum_{i\in[n]} w_i}{n}\right]\right)^{2q}. \tag{13}$$

for ε as per that lemma. Next, it follows from Lemma 3.4 that there exists a re-weighting μ'' of μ' such that $\widetilde{\mathbb{E}}_{\mu''}\left[\sum_{i\in[n]} w_i\right] \leqslant \delta n/10$. Now, as we will see in Section 5, μ' was constructed by re-weighting μ iteratively by large powers of co-ordinates of v, and this process repeats $\gamma = \mathcal{O}(\exp(1/\delta^4)^{\exp 1/\delta^2})$ times. In particular, let $\mu \to \mu' \to \mu_1 \to \mu_2 \to \ldots \to \mu_\gamma = \mu''$ be the intermediate re-weightings, where μ_ℓ is obtained by picking a coordinate $i_\ell \in [d]$ such that $\widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[v_{i_\ell}^{2t'}\right] \geqslant \zeta^{t'}\widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[\|v\|^{2t'}\right]$ where $\zeta = 1/\exp(1/\delta^2)$. Let $\mathcal{S} \subset [d]$ be the subset of indices that are selected in this process and let $z = \prod_{i\in\mathcal{S}} v_i^{2t'}$. Therefore,

$$\widetilde{\mathbb{E}}_{\mu'}[z] = \widetilde{\mathbb{E}}_{\mu'}\left[\prod_{i\in\mathcal{S}} v_i^{2t'}\right] = \widetilde{\mathbb{E}}_{\mu_{\gamma-1}}\left[v_{i_\gamma}^{2t'}\right] \geqslant \zeta^{2t'}. \tag{14}$$

Next, since $\{\, \|v\|^2 = 1 \,\}$ is a constraint in our constraint system, we have $\widetilde{\mathbb{E}}_{\mu'}\left[z^2\right] \leqslant 1$.

Combining these two observations, we note that

$$\begin{aligned}
\left(\widetilde{\mathbb{E}}_{\mu''}\left[\sum_{i\in[n]} w_i - \widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right]\right]\right)^{4q} &\leqslant \left(\widetilde{\mathbb{E}}_{\mu''}\left[\left(\sum_{i\in[n]} w_i - \widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right]\right)^{2q}\right]\right)^2 \\
&= \frac{\left(\widetilde{\mathbb{E}}_{\mu'}\left[\left(\sum_{i\in[n]} w_i - \widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right]\right)^{2q} z\right]\right)^2}{\left(\widetilde{\mathbb{E}}_{\mu'}[z]\right)^2} \\
&\leqslant \frac{n^{4q}\left(\widetilde{\mathbb{E}}_{\mu'}\left[\left(\sum_{i\in[n]} w_i/n - \widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i/n\right]\right)^{4q}\right]\right)\left(\widetilde{\mathbb{E}}_{\mu'}\left[z^2\right]\right)}{\left(\widetilde{\mathbb{E}}_{\mu'}[z]\right)^2} \\
&\leqslant \varepsilon^{2q}\cdot n^{4q}/\zeta^{4t'}
\end{aligned} \tag{15}$$

Taking the $4q$-th root and setting $\varepsilon = \zeta/(10\delta)$, we have

$$\widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right] \leqslant \widetilde{\mathbb{E}}_{\mu''}\left[\sum_{i\in[n]} w_i\right] + \delta n/10, \tag{16}$$

which is a contradiction. This allows us to conclude that for any degree-t pseudo-distribution $\widetilde{\mathbb{E}}_{\mu}\left[\sum_{i\in[n]} w_i\right] \leqslant \delta n$, and therefore

$$\mathcal{A}_\delta \left|\frac{v,w}{t}\right. \left\{\, \delta n - \sum_{i\in[n]} w_i \geqslant 0 \,\right\},$$

which concludes the proof. □

3.1 Handling *almost-sparse* directions

In this subsection, we prove Lemma 3.2, restated for convenience.

Lemma 3.2 (Decomposing the box-indicator for *almost-sparse* directions)**.** *Given $1 > \gamma > 0$, let $t \in \mathbb{N}$, and let p be the degree-$2t$ polynomial from Definition 2.25. Let $\{ x_i \}_{i\in[n]}$ be n i.i.d. samples from a reasonably anti-concentrated distribution $\mathcal{D}$ (see Definition 2.23). Let μ' be a degree-ϕ pseudo-distribution such that $\phi \geqslant 4t$ and let $u_{\mathcal{S}}$ be a fixed constant vector supported on $|\mathcal{S}| \ll d$ coordinates. If for all $\ell \in [t]$, $\widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_{\mathcal{S}}\|_2^{2\ell}\right] \leqslant \gamma^{2\ell}$, then*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\frac{1}{n}\sum_{i\in[n]} p^2\Big(\langle x_i, v\rangle^2\Big)\right] \leqslant \frac{1}{n}\sum_{i\in[n]} p^2\Big(\langle x_i, u_{\mathcal{S}}\rangle^2\Big) + O\Big(2^{O(t)} c_{2t}\gamma\Big),$$

where c_{2t} is the hypercontractivity constant of $\mathcal{D}$.

Proof. We begin by recalling that by definition p^2 is an even polynomial, and therefore we have

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]} p^2(\langle x_i, v\rangle) &= \frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]} c_j \langle x_i, v\rangle^{2j} \\ &= \sum_{j\in[t]} c_j \left(\frac{1}{n}\sum_{i\in[n]} \langle x_i, v\rangle^{2j}\right)\end{aligned} \tag{17}$$

where for all j, $|c_j| \leqslant 2^{O(t)}$ (Lemma 2.28). For each sample x_i, consider the decomposition $\langle x_i, v\rangle = \langle x_i, v - u_{\mathcal{S}}\rangle + \langle x_i, u_{\mathcal{S}}\rangle$. Consider the case where $c_j \geqslant 0$. Now, using the binomial theorem, for $j \in [t]$, we have

$$\langle x_i, v\rangle^{2j} = \langle x_i, u_{\mathcal{S}}\rangle^{2j} + \sum_{\ell=1}^{2j} \binom{2j}{\ell} \langle x_i, u_{\mathcal{S}}\rangle^{\ell} \langle x_i, v - u_{\mathcal{S}}\rangle^{2j-\ell}$$

Then, summing over all $i \in [n]$ and taking pseudo-expectation, we have,

$$\begin{aligned}&\widetilde{\mathbb{E}}_{\mu'} \left[\frac{1}{n}\sum_{i\in[n]} \langle x_i, v\rangle^{2j}\right] \\ &= \frac{1}{n}\sum_{i\in[n]} \langle x_i, u_{\mathcal{S}}\rangle^{2j} + \underbrace{\sum_{\ell=1}^{2j} \binom{2j}{\ell} \widetilde{\mathbb{E}}_{\mu'} \left[\frac{1}{n}\sum_{i\in[n]} \langle x_i, u_{\mathcal{S}}\rangle^{\ell} \langle x_i, v - u_{\mathcal{S}}\rangle^{2j-\ell}\right]}_{(18).1}\end{aligned} \tag{18}$$

Observe, for a fixed ℓ, the term in (18).1 is a scalar quantity (since we take pseudo-expectation over μ') and therefore, it suffices to bound the absolute value of this scalar. Recall, by pseudo-Jensen's, for any indeterminate x, $\left(\widetilde{\mathbb{E}}_{\mu}[x]\right)^2 \leqslant \widetilde{\mathbb{E}}_{\mu}[x^2]$. Therefore, for a fixed ℓ, we have

$$
\begin{aligned}
&\left(\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{\ell}\langle x_i, v-u_{\mathcal{S}}\rangle^{2j-\ell}\right]\right)^2 \\
&\quad\leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\left(\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{\ell}\langle x_i, v-u_{\mathcal{S}}\rangle^{2j-\ell}\right)^2\right] \\
&\quad\leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{2\ell}\cdot\frac{1}{n}\sum_{i\in[n]}\langle x_i, v-u_{\mathcal{S}}\rangle^{4j-2\ell}\right] \\
&\quad= \underbrace{\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{2\ell}}_{(19).(1)}\underbrace{\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, v-u_{\mathcal{S}}\rangle^{4j-2\ell}\right]}_{(19).(2)}
\end{aligned}
\tag{19}
$$

To bound term (19).(1), we use true hypercontractivity of the samples:

$$
\begin{aligned}
\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{2\ell} &\leqslant K_{2\ell}\left(\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{2}\right)^{\ell} \\
&= K_{2\ell}\left(u_{\mathcal{S}}^{\top}\left(\frac{1}{n}\sum_{i\in[n]}x_i x_i^{\top}\right)u_{\mathcal{S}}\right)^{\ell} \\
&\leqslant 2K_{2\ell}\|u_{\mathcal{S}}\|_2^{2\ell} \leqslant 2K_{2\ell},
\end{aligned}
\tag{20}
$$

where $K_{2\ell}$ is the hypercontractivity constant.

To bound term (19).(2), we appeal to certifiable hypercontractivity of the uniform distribution over the samples:

$$
\begin{aligned}
&\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, v-u_{\mathcal{S}}\rangle^{4j-2\ell}\right] \\
&\leqslant K_{2\ell}\widetilde{\mathbb{E}}_{\mu'}\left[\left(\frac{1}{n}\sum_{i\in[n]}\langle x_i, v-u_{\mathcal{S}}\rangle^{2}\right)^{2j-\ell}\right] \\
&= K_{2\ell}\widetilde{\mathbb{E}}_{\mu'}\left[\left((v-u_{\mathcal{S}})^{\top}\left(\frac{1}{n}\sum_{i\in[n]}x_i x_i^{\top}\right)(v-u_{\mathcal{S}})\right)^{2j-\ell}\right] \\
&\leqslant 2K_{2\ell}\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_2^{4j-2\ell}\right] \leqslant 2K_{2\ell}\gamma^{4j-2\ell}.
\end{aligned}
\tag{21}
$$

Combining equations (20) and (21), we have

$$
|\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{\ell}\langle x_i, v-u_{\mathcal{S}}\rangle^{2j-\ell}\right]| \leqslant 2K_{2\ell}\gamma^{4j-2\ell}
$$

Now we can bound the term (18).1 as follows:

$$
\begin{aligned}
\sum_{\ell\in[2j]}\binom{2j}{\ell}|\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{\ell}\langle x_i, v-u_{\mathcal{S}}\rangle^{2j-\ell}\right]| &\leqslant \sum_{\ell\in[2j]}\binom{2j}{\ell}4K_{2\ell}\gamma^{2j-\ell} \\
&\leqslant 4K_{2j}2^{2j}\gamma
\end{aligned}
\tag{22}
$$

Therefore, for any fixed j, we have

$$\widetilde{\mathbb{E}}_{\mu'}\left[\frac{c_{2j}}{n}\sum_{i\in[n]}\langle x_i, v\rangle^{2j}\right] = \frac{c_{2j}}{n}\sum_{i\in[n]}\langle x_i, u_{\mathcal{S}}\rangle^{2j} + c_{2j}\eta_j,$$

where $|\eta_j| \leqslant 4K_{2j}2^{2j}\gamma$. Summing over all $j \in [t]$,

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]}p_\gamma^2(\langle x_i, v\rangle)\right] &\leqslant \frac{1}{n}\sum_{i\in[n]}p_\gamma^2(\langle x_i, u_{\mathcal{S}}\rangle) + \sum_{j\in[t]}c_{2j}\eta_j \\ &\leqslant \frac{1}{n}\sum_{i\in[n]}p_\gamma^2(\langle x_i, u_{\mathcal{S}}\rangle) + 2^{O(t)}K_t\gamma\end{aligned} \tag{23}$$

□

3.2 Handling analytically dense directions

In this subsection, we prove Lemma 3.3, restated below for convenience.

Lemma 3.3 (Anti-Concentration with Shifts). *Given $\zeta > 0$, a pseudo-distribution μ of degree $t = \Omega\left(\frac{\log(d)}{\exp(\zeta^4)^{\exp(1/\zeta^4)}}\right)$, and a vector $u_{\mathcal{S}} \in \mathbb{R}^d$ supported on $|\mathcal{S}|$ coordinates for some $\mathcal{S} \subseteq [d]$, let $\{x_i\}_{i\in[n]}$ be a set of n points such that for all $i \in [n]$, $\langle x_i, u_{\mathcal{S}}\rangle^2 \leqslant \mathcal{O}(\log(1/\zeta))$. Let q be the polynomial from Theorem 4.1 with degree $\mathcal{O}(\exp(1/\zeta^2))$. If*

$$\widetilde{\mathbb{E}}_\mu\left[\|v - u_{\mathcal{S}}\|_{2z}^{2z}\right] \leqslant \zeta^{2z}\widetilde{\mathbb{E}}_\mu\left[\|v - u_{\mathcal{S}}\|_2^{2z}\right],$$

for $z = \Omega\big(\log(d)\exp(1/\zeta^2)\big)$, then we have

$$\widetilde{\mathbb{E}}_\mu\left[\frac{1}{n}\sum_{i\in[n]}q^2(\langle x_i, v\rangle)\right] \leqslant \zeta\cdot\widetilde{\mathbb{E}}_\mu\left[\|v - u_{\mathcal{S}}\|^2\right] + \frac{1}{\exp(1/\zeta^2)}$$

We begin by showing that if the pseudo-expectation of the $2z$-norm is bounded, then, the pseudo-expectation of the 4 norm is bounded, even with additional vectors of the ℓ_2^2 norm.

Lemma 3.5 ($(2z, 2)$-Hypercontractivity implies $(4, 2)$-Hypercontractivity). *Given a pseudo-distribution μ over a vector valued indeterminate v such that $\|v\|^2 \leqslant 1$, and an even integer $z \geqslant 2$ such that $\widetilde{\mathbb{E}}_\mu\left[\|v\|_{2z}^{2z}\right] \leqslant \lambda^{2z}\cdot\widetilde{\mathbb{E}}_\mu\left[\|v\|_2^{2z}\right]$, we have that for any integer $k \geqslant 0$,*

$$\widetilde{\mathbb{E}}_\mu\left[\|v\|_2^{2k}\cdot\|v\|_4^4\right] \leqslant \lambda^2\Big(\widetilde{\mathbb{E}}_\mu\left[\|v\|_2^{2z}\right].\Big)^{1/z}$$

Proof. We first use induction on z' to prove $\left\{\|v\|_4^{4z'} \leqslant \|v\|_{2z'+2}^{2z'+2}\cdot\|v\|_2^{2(z'-1)}\right\}$ for integers $z' \geqslant 1$. If $z' = 1$, then this is trivial and when $z' = 2$, we have

$$\left|\overset{v}{-}\right\}\left\{\|v\|_4^8 = \left(\sum_{i\in[d]}v_i^4\right)^2 = \left(\sum_{i\in[d]}v_i^3\cdot v_i\right)^2 \leqslant \left(\sum_{i\in[d]}v_i^6\right)\cdot\left(\sum_{i\in[d]}v_i^2\right) = \|v\|_6^6\cdot\|v\|_2^2\right\},$$

where the inequality follows from Cauchy-Schwarz, and this completes the base case. For the induction step, assume $\left\{ \|v\|_4^{4z'} \leqslant \|v\|_{2z'+2}^{2z'+2} \cdot \|v\|_2^{2(z'-1)} \right\}$. Then, using the inductive hypothesis

$$\vdash^{v} \left\{ \|v\|_4^{4z'+4} = \|v\|_4^{4z'} \cdot \|v\|_4^4 \leqslant \|v\|_{2z'+2}^{2z'+2} \cdot \|v\|_2^{2(z'-1)} \cdot \|v\|_4^4 \right\}$$

We wish to prove this is at most $\|v\|_{2z'+4}^{2z'+4} \cdot \|v\|_2^{2z'}$ in SoS. For this, it suffices to prove

$$\begin{aligned}
\|v\|_{2z'+2}^{2z'+2} \cdot \|v\|_2^{2(z'-1)} \cdot \|v\|_4^4 &\leqslant \|v\|_{2z'+4}^{2z'+4} \cdot \|v\|_2^{2z'} \\
\|v\|_{2z'+2}^{2z'+2} \cdot \|v\|_4^4 &\leqslant \|v\|_{2z'+4}^{2z'+4} \cdot \|v\|_2^2 \\
\sum_{i \neq j} v_i^{2z'+2} \cdot v_j^4 &\leqslant \sum_{i \neq j} v_i^{2z'+4} \cdot v_j^2 \\
0 &\leqslant \sum_{i \neq j} (v_i^{2z'+4} \cdot v_j^2 - v_i^{2z'+2} \cdot v_j^4) \\
0 &\leqslant \sum_{i<j} (v_i^{2z'+4} \cdot v_j^2 + v_j^{2z'+4} \cdot v_i^2 - v_i^{2z'+2} \cdot v_j^4 - v_j^{2z'+2} \cdot v_i^4)
\end{aligned}$$

In the first step, we dropped the power of $\|v\|_2$, and in the second step, we just expanded the norms. The remaining steps are just rearranging the given expression. Therefore, it suffices to prove that for each $i < j$, $(v_i^{2z'+4} \cdot v_j^2 + v_j^{2z'+4} \cdot v_i^2 - v_i^{2z'+2} \cdot v_j^4 - v_j^{2z'+2} \cdot v_i^4)$ is SoS. But this is true because

$$\begin{aligned}
&v_i^{2z'+4} \cdot v_j^2 + v_j^{2z'+4} \cdot v_i^2 - v_i^{2z'+2} \cdot v_j^4 - v_j^{2z'+2} \cdot v_i^4 \\
&\quad = (v_i^2 - v_j^2)^2 (v_i^{2z'} v_j^2 + v_i^{2z'-2} v_j^4 + \ldots + v_i^2 v_j^{2z'})
\end{aligned}$$

which is clearly a sum of squares. We verify the last equality in the following display.

$$\begin{aligned}
&(v_i^2 - v_j^2)^2 (v_i^{2z'} v_j^2 + v_i^{2z'-2} v_j^4 + \ldots + v_i^2 v_j^{2z'}) \\
&\quad = (v_i^2 - v_j^2)(v_i^{2z'+2} v_j^2 + \ldots + v_i^4 v_j^{2z'} - v_i^{2z'} v_j^4 - \ldots - v_i^2 v_j^{2z'+2}) \\
&\quad = (v_i^2 - v_j^2)(v_i^{2z'+2} v_j^2 - v_i^2 v_j^{2z'+2}) \\
&\quad = v_i^{2z'+4} \cdot v_j^2 + v_j^{2z'+4} \cdot v_i^2 - v_i^{2z'+2} \cdot v_j^4 - v_j^{2z'+2} \cdot v_i^4
\end{aligned}$$

as desired.

Now, we complete the proof. Using Fact 2.7, we have

$$\begin{aligned}
\left(\widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \cdot \|v\|_4^4 \right] \right)^z &\leqslant \left(\widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \cdot \|v\|_4^4 \right] \right)^{z-1} \\
&\leqslant \widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k(z-1)} \cdot \|v\|_4^{4z-4} \right] \\
&\leqslant \widetilde{\mathbb{E}}_\mu \left[\|v\|_{2z}^{2z} \cdot \|v\|_2^{2(z-2)+2k(z-1)} \right] \\
&\leqslant \widetilde{\mathbb{E}}_\mu \left[\|v\|_{2z}^{2z} \right] \\
&\leqslant \lambda^{2z} \widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2z} \right]
\end{aligned} \tag{24}$$

Taking zth roots completes the proof. □

We can now complete the proof of our lemma, building upon Theorem 4.1, which certifies anticoncentration of dense directions with shifts.

Proof of Lemma 3.3. Since for all $i \in [n]$, we know that $\langle x_i, u_{\mathcal{S}} \rangle \leqslant O(\log(1/\zeta))$, it suffices to consider the box indicator polynomial q^2 corresponding to Theorem 4.1, with $\Delta^2 = \Theta(\log(1/\zeta))$. Recall, we know that

$$\widetilde{\mathbb{E}}_{\mu} \left[\|v - u_{\mathcal{S}}\|_{2z}^{2z} \right] \leqslant \eta^{2z} \widetilde{\mathbb{E}}_{\mu} \left[\|v - u_{\mathcal{S}}\|_{2}^{2z} \right],$$

where $\eta = 1/\exp\left(1/\zeta^4\right)$. Let $k = \exp(1/\zeta^4)$. It follows from Lemma 3.5 that for any $\ell \in [k]$,

$$\widetilde{\mathbb{E}}_{\mu} \left[\|v - u_{\mathcal{S}}\|_2^{2\ell} \cdot \|v - u_{\mathcal{S}}\|_4^4 \right] \leqslant \eta. \tag{25}$$

However, observe such a statement is not sufficient to invoke the axiom $\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \}$ required to apply Theorem 4.1. Instead, writing out the guarantee of Theorem 4.1 (from Eq. (33) with the formal substitution $v = v - u_{\mathcal{S}}$) for some $\lambda < 1/\exp\left(1/\zeta^4\right)$, we have

$$\zeta \|v - u_{\mathcal{S}}\|^2 - \frac{1}{n} \sum_{i \in [n]} q^2(\langle x_i, v - u_{\mathcal{S}} \rangle + \langle x_i, u_{\mathcal{S}} \rangle) = \mathrm{sos}(v - u_{\mathcal{S}}) + z(v - u_{\mathcal{S}}) \Big(\lambda \|v - u_{\mathcal{S}}\|_2^4 - \|v - u_{\mathcal{S}}\|_4^4 \Big) \tag{26}$$

Recall, $z(v - u_{\mathcal{S}}) = \sum_{a \in [\exp(1/\zeta^2)]} c_a \|v - u_{\mathcal{S}}\|_2^{2a}$ (which we can see by tracking our proof of Theorem 4.1), where $|c_a| \leqslant \exp\left(1/\zeta^2\right)$, and thus using Eq. (25), we have

$$\begin{aligned} \widetilde{\mathbb{E}}_{\mu'} \left[z(v - u_{\mathcal{S}}) \|v - u_{\mathcal{S}}\|_4^4 \right] &= \widetilde{\mathbb{E}}_{\mu'} \left[\left(\sum_{a \in [\exp(1/\zeta^2)]} c_a \|v - u_{\mathcal{S}}\|_2^{2a} \right) \|v - u_{\mathcal{S}}\|_4^4 \right] \\ &\leqslant \eta \cdot \exp\left(2/\zeta^2\right) \leqslant 1/\exp\left(1/\zeta^3\right), \end{aligned} \tag{27}$$

Similarly, since λ is small and $\|v - u_{\mathcal{S}}\|^2 \leqslant 1$, we can bound

$$\left| \widetilde{\mathbb{E}}_{\mu} \left[z(v - u_{\mathcal{S}}) \lambda \|v - u_{\mathcal{S}}\|_2^4 \right] \right| \leqslant 1/\exp\left(1/\zeta^3\right).$$

Taking pseudo-expectations on both sides in Eq. (26) and plugging in the above bounds we have

$$\widetilde{\mathbb{E}}_{\mu} \left[\frac{1}{n} \sum_{i \in [n]} q^2(\langle x_i, v \rangle) \right] \leqslant \zeta \cdot \widetilde{\mathbb{E}}_{\mu} \left[\|v - u_{\mathcal{S}}\|^2 \right] + \frac{1}{\exp(1/\zeta^2)},$$

as desired. □

3.3 Re-weightings to Refutation

In this subsection, we prove that given any pseudo-distribution μ we can find a re-weighting μ' such that $\widetilde{\mathbb{E}}_{\mu'} \left[\sum_{i \in [n]} w_i \right] \leqslant \delta n$, formally,

Lemma 5.1 (Re-weighting certifies Anti-Concentration). *Given $0 < \delta < 1$ and a degree t pseudo-distribution μ over a vector valued indeterminate $v \in \mathbb{R}^d$, there exists a re-weighting of μ, denoted by μ' such that*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\sum_{i \in [n]} w_i \right] < \delta n,$$

if $t = \Omega\left(\log(d) \left(\exp\left(1/\delta^4\right)^{\exp\left(1/\delta^4\right)} \right) \right)$.

We proceed by case analysis, where we case on whether the re-weighting outputs a sparse vector $u_{\mathcal{S}}$ and pseudo-distribution μ' for which pseudo-moment of $\|v - u_{\mathcal{S}}\|_2^2$ are small, or $v - u_{\mathcal{S}}$ is analytically dense, in the sense that $\widetilde{\mathbb{E}}_{\mu}\left[\|v - u_{\mathcal{S}}\|_{2z}^{2z}\right]$ is small. In the first case, we invoke Lemma 3.2 to appeal to anti-concentration of the sparse coordinates, and in the second case we invoke the analytic sum-of-squares certificates from Lemma 3.3. We abstract out reweighting analyses to later sections and utilize Theorem 5.1 in our proof.

Proof of Lemma 3.4. We assume the hypercontractivity constant $K_{2\ell} \leqslant c^{2\ell}$ for a fixed constant c, where $\ell \leqslant \exp(1/\delta^2)$. Let μ be any degree-$\Omega\left(\frac{\log(d)}{\exp(1/\zeta^4)^{\exp(1/\zeta^4)}}\right)$ pseudo-distribution that is consistent with $\mathcal{A}_\delta$. Then, we can appeal to Theorem 5.1 with parameters $\delta = \delta/2, k = 1/\exp(1/\delta^2), z = \log(d)/\exp(1/\delta^2), \eta = 1/\exp(1/\delta^2)$ to obtain a re-weighted pseudo-distribution μ' and a vector $u_{\mathcal{S}}$ supported on at most $|\mathcal{S}| = \mathcal{O}(1/(\delta^{4k+2}\eta^2))$ coordinates. We first consider the case where the resulting pseudo-distribution μ' satisfies that for all $y \in [k]$,

$$\widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_{\mathcal{S}}\|_2^{2y}\right] \leqslant \eta.$$

Next, observe that $p^2(z^2) \geqslant 1/2$ for all $z \in [-\delta, \delta]$ is a univariate inequality in a bounded interval. Further, it follows from the constraints $\mathcal{A}_\delta$ that $w_i^2\langle x_i, v\rangle^2 \leqslant w_i\delta^2 \leqslant \delta^2$ and therefore, invoking the univariate inequality with $z = w_i^2\langle x_i, v\rangle^2$, and using Fact 2.10 we have

$$\{\,\mathcal{A}_\delta\,\} \vdash^{w_i^2\langle x_i, v\rangle^2} \left\{\frac{1}{2} \leqslant p^2\left(w_i\langle x_i, v\rangle^2\right)\right\}, \tag{28}$$

and therefore,

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]} w_i\right] &\leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\frac{2}{n}\sum_{i\in[n]} w_i \cdot p^2\left(w_i\langle x_i, v\rangle^2\right)\right] \\ &\leqslant \underbrace{\frac{1}{n}\sum_{i\in[n]} p^2\left(\langle x_i, u_{\mathcal{S}}\rangle^2\right)}_{(29).(1)} + \delta,\end{aligned} \tag{29}$$

where the last inequality follows from Lemma 3.2 and our choice of $\eta \leqslant \delta/\left(2^{O(\ell)}K_{2\ell}\right)$ that makes the additive error small. Finally, we note that (29).(1) is bounded by δ since Fact 2.31 is a univariate inequality in $\langle x, u_{\mathcal{S}}\rangle^2$ and thus admits a sum-of-squares proof.

Next, consider the alternative, where μ' satisfies

$$\widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_{\mathcal{S}}\|_{2z}^{2z}\right] \leqslant (2\delta)^{2z} \cdot \widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_{\mathcal{S}}\|_2^{2z}\right] \text{ and } \widetilde{\mathbb{E}}_{\mu'}\left[\|v - u_{\mathcal{S}}\|\right] \geqslant \eta.$$

It follows from Fact 2.22 that with for all but a δ-fraction of the samples x_i, $\langle x_i, u_{\mathcal{S}}\rangle^2 \leqslant \log(1/\delta)$. Conditioned on this event, we can partition the set of points x_i into $\mathcal{P} \subset [n]$ such that

$$\mathcal{P} = \left\{\, i \in [n] \mid \langle x_i, u_{\mathcal{S}}\rangle^2 \leqslant \log(1/\delta)\,\right\}.$$

Then,

$$\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in[n]} w_i\right] \leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in\mathcal{P}} w_i\right] + \frac{1}{n}\left|[n] \setminus \mathcal{P}\right| \leqslant \underbrace{\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in\mathcal{P}} w_i\right]}_{(30).(1)} + \delta, \tag{30}$$

and thus it suffices to bound term (30).(1). Using Eq. (28) again, we have

$$\widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in\mathcal{P}} w_i\right] \leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\frac{1}{n}\sum_{i\in\mathcal{P}} w_i q^2\Big(\langle x_i, v\rangle^2\Big)\right] \leqslant \mathcal{O}(\delta), \tag{31}$$

which follows from Lemma 3.3, since each point in $\mathcal{P}$ has a bounded shift. □

Finally, in Appendix A, we will show how to bound the bit complexity of our SoS proofs. The idea is to instantiate the identity for various carefully chosen w, v and then extract bounds on the coefficients.

3.4 Certifying Anti-Concentration around a shift

We observe that we can execute the same proof as above to certify anti-concentration w.r.t. the following set of constraints:

$$\mathcal{A}_{\delta,\phi} = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ \forall i \in [n] & w_i\Big(\langle x_i, v\rangle^2 - \phi\Big) \leqslant w_i\delta^2 \\ & \|v\|_2^2 = 1 \end{Bmatrix} \tag{32}$$

Corollary 3.6 (Certifying Anti-Concentration around a shift)**.** *Given $\delta > 0$, $\phi \geqslant 1$, and $n \geqslant n_0$ samples $\{ x_1, x_2, \ldots, x_n \} \subseteq \mathbb{R}^d$ for some $n_0 \geqslant d^t$, sampled from a $(\delta, \exp(1/\delta^2), O(1))$-reasonably anti-concentrated distribution $\mathcal{D}$ (Definition 2.23), let $t = \mathcal{O}\left(\frac{\log(d)\cdot\phi}{\exp(\delta^4)^{\exp(1/\delta^2)}}\right)$. Then, there is a degree-t certificate of anti-concentration, i.e.*

$$\mathcal{A}_\delta \vdash_t^{w,v} \left\{ \frac{1}{n}\sum_{i\in[n]} w_i \leqslant \delta \right\},$$

Proof. The proof is exactly the same as that of Theorem 3.1, with the modification that we use the shifted version of p^2, with the shift being ϕ. It follows from Theorem 4.1 that $\mathbb{E}_{i\in[n]}\left[p^2\Big(\langle x_i, v\rangle^2 - \phi\Big)\right] \leqslant \delta$ as long as p^2 has degree $\mathcal{O}(1/\delta^2 + \phi)$. □

4 Certifiable Anti-Concentration along analytically dense directions

We show that for reasonably anticoncentrated distributions, we can obtain a sum-of-squares certificate of anti-concentration along *analytically-dense* directions, not necessarily centered around the origin:

Theorem 4.1 (Anti-concentration certificates with bounded shifts)**.** *Let v be a vector such that $0 < \|v\|_2^2 \leqslant 1$. Let $\{ x_i \}_{i\in[n]}$ be n iid samples from a distribution $\mathcal{D}$ which is reasonably anticoncentrated. Let p^2 be a degree $t = O\Big(\log\Big(\frac{1}{\delta} + \frac{\zeta}{\delta}\Big)\cdot\Big(\frac{1}{1/\delta^2} + \Delta^2\Big)\Big)$ polynomial such that for all $\zeta^2 \leqslant \Delta^2$ and $\lambda \leqslant \frac{\delta}{(2t)^{4t}\Delta^t}$,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \vdash_t^v \left\{ \mathbb{E}\left[p^2(\langle x_i, v\rangle - \zeta)\right] \leqslant O(\delta)\|v\|_2^2 \right\},$$

Further, let $\mathcal{D}'$ be a distribution such that $\|y\|_2 \leqslant \Delta$ almost surely for all $y \sim \mathcal{D}'$ and let $\{ y_i \}_{i\in[n]}$ be n iid samples from $\mathcal{D}'$. Then,

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash \left\{ \frac{1}{n} \sum_{i\in[n]} p^2(\langle x_i, v\rangle - y_i) \leqslant O(\delta)\|v\|_2^2 \right\}.$$

The rest of the section is devoted to proving this theorem and relies on the following strategy.

- In Lemmas 4.2-4.3, we first obtain a SoS certificate of anti-concentration along a dense direction v that satisfies $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$. The analysis proceeds by utilizing an explicit polynomial p^2 that approximates the box-indicator, and is inspired by [KNS10].
- In Theorem 4.6, we show that given a sufficiently ℓ_4-dense direction, we can obtain a certificate of anti-concentration around a point that need not be the origin.

We can re-write the final conclusion above as the following polynomial identity:

$$\delta\|v\|^2 - \frac{1}{n}\sum_{i\in[n]} p\Big(\langle x_i, v\rangle^2 - y_i\Big) = \mathrm{sos}(v) + z(v)\Big(\lambda\|v\|_2^4 - \|v\|_4^4\Big) \tag{33}$$

where $\mathrm{sos}(v), z(v)$ are sum-of-squares polynomials, p^2 has degree $t = O\Big(\log\Big(\frac{1}{\delta} + \frac{\zeta}{\delta}\Big) \cdot \Big(\frac{1}{1/\delta^2} + \Delta^2\Big)\Big)$ and $\lambda \leqslant \frac{\delta}{(2t)^{4t}\Delta^t}$. Further, $z(v)$ is explicit in our proof of Theorem 4.1.

4.1 Anti-concentration certificates along dense directions

The goal of this section is to obtain a SoS certificate of anti-concentration along the direction v that satisfies $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$. Our approach is to get a lower and upper bound on all 'monomials', i.e. $\mathbb{E}\left[\langle x, v\rangle^{2k}\right]$ and then compute an explicit polynomial p^2 that approximates the box-indicator.

Lemma 4.2 (Lower Bound on $2k$-th Moment). *Let $d/2 \geqslant k \geqslant 2$ and let $x \sim \mathcal{D}$, where $\mathcal{D}$ is a reasonably anticoncentrated distribution. Further, for any set $S \subset [d]$ such that $|S| = k$, denote $\mathbb{E}\left[\prod_{i\in S} x_i^2\right] = \Phi_k$. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$ for some $\lambda \in (0, 2/k^2)$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \vdash_{4k}^{v} \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \geqslant \left(\frac{(2k)!}{2^k k!}\right) \cdot \Phi_k\left(1 - \frac{(k-1)\lambda}{2}\right)\|v\|_2^{2k} \right\}.$$

Next, we consider the upper bound certificate. We show that such a certificate holds whenever the distribution we consider is reasonably anticoncentrated.

Lemma 4.3 (Upper Bound on $2k$-th Moment). *Let $k \geqslant 2$ and let $x \sim \mathcal{D}$ where $\mathcal{D}$ is a reasonably anticoncentrated distribution. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$ for some $\lambda \in (0, 1/(k^k c_k^k))$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \vdash_{4k}^{v} \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \leqslant \left(\frac{(2k)!}{2^k k!}\right)\left(1 + \left(\lambda c_k^{k-2} k^k\right)\right)\|v\|_2^{2k} \right\}.$$

We prove these lemmas in the subsequent sections.

4.1.1 Warmup: The case of independent coordinates

In this section, as a warmup, we consider the case when the x_is are independent with mean 0 and variance 1 and prove the bounds from the previous section. To this end, we obtain lower and upper bounds on the $2k$th moments in Lemma 4.4 and Lemma 4.5 respectively.

We first obtain a lower bound on $\mathbb{E}\left[\langle x, v\rangle^{2k}\right]$ that is captured by sos. The proof is inspired by Lemma 2.6 in Kindler, Naor and Schechtman [KNS10].

Lemma 4.4 (Lower Bound on $2k$-th moment)**.** *Let $d/2 \geqslant k \geqslant 2$ and let $\{x_i\}_{i\in[d]}$ be d iid random variables with mean 0 and variance 1. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$ for some $\lambda \in (0, 2/k^2)$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \geqslant \left(\frac{(2k)!}{2^k k!}\right)\left(1 - \frac{k(k-1)\lambda}{2}\right)\|v\|_2^{2k} \right\}$$

Proof. For ease of notation in this proof, denote by $w_i := x_i v_i$ for $i \leqslant d$. Using the substitution rule, we have

$$\left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] = \sum_{\substack{\ell_1,\ldots\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{(2k)!}{\prod_{j\in[d]}(2\ell_j)!}\,\mathbb{E}\left[\prod_{i\in[d]} w_j^{2\ell_j}\right] \geqslant \left(\frac{(2k)!}{2^k}\right)\left(\sum_{\substack{S\subset[n]\\ |S|=k}}\prod_{j\in S}\mathbb{E}\left[w_j^2\right]\right)\right\}, \quad (34)$$

where the inequality follows from only summing over terms for which $\ell_j \leqslant 1$ for all $j \in [d]$. On the other hand, using the substitution rule and $\mathbb{E}[w_j^2] = \mathbb{E}[x_j^2 v_j^2] = v_j^2$, we have

$$\left|\frac{v}{4k}\right. \left\{ \|v\|_2^{2k} = \left(\sum_{j\in[d]}\mathbb{E}\left[w_j^2\right]\right)^k = \sum_{\substack{\ell_1,\ldots,\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{k!}{\prod_{j\in[d]}\ell_j!}\prod_{j\in[d]}(\mathbb{E}\left[w_j^2\right])^{\ell_j}\right\}. \quad (35)$$

We split the term above into terms where all the ℓ_j's are at most 1 and the terms where at least one ℓ_j is at least 2, to get

$$\begin{aligned}\|v\|_2^{2k} &\leqslant k!\sum_{\substack{S\subset[n]\\ |S|=k}}\prod_{j\in S}\mathbb{E}\left[w_j^2\right] + \frac{k}{2}\sum_{j\in[d]}\mathbb{E}\left[w_j^2\right]\sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-1\\ r_j\geqslant 1}}\frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]}(\mathbb{E}\left[w_i^2\right])^{r_i}\\ &= k!\left\{\sum_{\substack{S\subset[n]\\ |S|=k}}\prod_{j\in S}\mathbb{E}\left[w_j^2\right] + \frac{k}{2}\sum_{j\in[d]}(\mathbb{E}\left[w_j^2\right])^2\sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}}\frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]}(\mathbb{E}\left[w_i^2\right])^{r_i}\right\}.\end{aligned} \quad (36)$$

Let us focus on upper bounding the second term above. Observe, the inner summation is independent of j and thus

$$\begin{aligned}&\left|\frac{v}{4k}\right. \left\{ \sum_{j\in[d]}(\mathbb{E}\left[w_j^2\right])^2\sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}}\frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]}(\mathbb{E}\left[w_i^2\right])^{r_i}\right.\\ &\qquad\left. = \left(\sum_{j\in[d]}(\mathbb{E}\left[w_j^2\right])^2\right)\cdot(k-1)\sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}}\frac{(k-2)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]}(\mathbb{E}\left[w_i^2\right])^{r_i}\right\}.\end{aligned} \quad (37)$$

We now bound each of these terms above as follows: using the bound on the ℓ_4^4 norm, along with $\mathbb{E}\left[x_j^2\right] = 1$,

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \left|\frac{v}{4}\right. \left\{ \sum_{j\in[d]}(\mathbb{E}\left[w_j^2\right])^2 = \sum_{j\in[d]} v_j^4 \leqslant \lambda\|v\|_2^4 \right\}, \quad (38)$$

$$\left|\frac{v}{4k}\right. \left\{ \sum_{\substack{r_1,\dots r_d \in \mathbb{Z}_{\geqslant 0} \\ r_1+\dots+r_d=k-2}} \frac{(k-2)!}{\prod_{i\in[d]} r_i!} \prod_{i\in[d]} \mathbb{E}\left[w_i^2\right]^{r_i} = \|v\|_2^{2k-4} \right\}. \tag{39}$$

Combining the above, we conclude

$$\left|\frac{v}{4k}\right. \left\{ \|v\|_2^{2k} \leqslant k! \sum_{\substack{S\subset[n] \\ |S|=k}} \prod_{j\in S} \mathbb{E}\left[w_j^2\right] + \frac{k(k-1)}{2}\left(\lambda\|v\|_2^{2k}\right) \right\} \tag{40}$$

Together with (34), we obtain the result. □

A similar proof technique can be used for the upper bound.

Lemma 4.5 (Upper Bound on $2k$-th Moment)**.** *Let $k \geqslant 2$ and let $\{x_i\}_{i\in[d]}$ be d iid random variables drawn from a distribution $\mathcal{D}$ with mean 0 and variance 1 that is (c_k, k)-certifiably hypercontractive. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$ for some $\lambda \in (0, 1/(k^k c_k^k))$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \leqslant \left(\frac{(2k)!}{2^k k!} + \lambda c_k^k k^k\right)\|v\|_2^{2k} \right\}.$$

Proof. The proof is similar to the proof of Lemma 4.4. We have

$$\begin{aligned}\left|\frac{v}{4k}\right. \Bigg\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] &= \sum_{\substack{\ell_1,\dots\ell_d \in \mathbb{Z}_{\geqslant 0} \\ \ell_1+\dots+\ell_d=k}} \frac{(2k)!}{\prod_{j\in[d]}(2\ell_j)!} \prod_{i\in[d]} \mathbb{E}\left[x_j^{2\ell_j} v_j^{2\ell_j}\right] \\ &\leqslant \left(\frac{(2k)!}{2^k}\right) \sum_{\substack{S\subset[d] \\ |S|=k}} \prod_{j\in S} \mathbb{E}\left[x_j^2 v_j^2\right] + \sum_{j\in[d]} v_j^4 \sum_{\substack{r_1,\dots r_d \in \mathbb{Z}_{\geqslant 0} \\ r_1+\dots+r_d=k-2}} \frac{(2k)!}{\prod_{i\in[d]}(2r_i)!} \mathbb{E}\left[x_j^4 \prod_{i\in[d]} x_i^{2r_i} v_i^{2r_i}\right] \Bigg\},\end{aligned} \tag{41}$$

To bound the first term, we use $\mathbb{E}\left[x_j^2 v_j^2\right] = \mathbb{E}\left[x_j^2\right] v_j^2 = v_j^2$ to get

$$\left|\frac{v}{4k}\right. \left\{ \sum_{\substack{S\subset[n] \\ |S|=k}} \prod_{j\in S} \mathbb{E}\left[x_j^2 v_j^2\right] \leqslant \|v\|_2^{2k} \right\}, \tag{42}$$

and therefore the first term is at most $\frac{(2k)!}{2^k k!}\|v\|_2^{2k}$. As for the second term, we will use the following argument. By hypercontractivity of the true samples (see Definition 2.18), we have

$$\left|\frac{v}{4r_i}\right. \left\{ \mathbb{E}\left[x_i^{2r} v_i^{2r}\right] = \mathbb{E}\left[x_i^{2r}\right] v_i^{2r} \leqslant c_k^r \cdot \mathbb{E}\left[x_i^2\right]^r v_i^{2r} \leqslant c_k^r v_i^{2r} \right\} \tag{43}$$

and thus

$$\begin{aligned}\left|\frac{v}{4k}\right. \left\{ \sum_{\substack{r_1,\dots r_d \in \mathbb{Z}_{\geqslant 0} \\ r_1+\dots+r_d=k-2}} \frac{(2k)!}{\prod_{i\in[d]}(2r_i)!} \prod_{i\in[d]} \mathbb{E}\left[x_j^4 (x_i v_i)^{2r_i}\right] &\leqslant \sum_{\substack{r_1,\dots r_d \in \mathbb{Z}_{\geqslant 0} \\ r_1+\dots+r_d=k-2}} \frac{(2k)! \cdot c_k^k}{\prod_{i\in[d]}(2r_i)!} \prod_{i\in[d]} v_i^{2r_i} \\ &\leqslant \frac{c_k^k \prod_{i\in[k+1]}(2k-i)}{2^{k-2}} \sum_{\substack{r_1,\dots r_d \in \mathbb{Z}_{\geqslant 0} \\ r_1+\dots+r_d=k-2}} \frac{(k-2)!}{\prod_{i\in[d]} r_i!} \prod_{i\in[d]} v_i^{2r_i} \\ &\leqslant \left(c_k^k k^k\right)\|v\|_2^{2k-4} \Bigg\}.\end{aligned} \tag{44}$$

Finally, we also use $\sum_{j\in[d]} v_j^4 \leqslant \lambda\|v\|_2^4$. Combining the equations above,

$$\left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \leqslant \left(\frac{(2k)!}{2^k}\right)\|v\|_2^{2k} + \left(\lambda c_k^k k^k\right)\|v\|_2^{2k} \right\}, \tag{45}$$

which concludes the proof. □

4.1.2 Proofs of Lemma 4.2 and Lemma 4.3

Using the same techniques from the previous section, we can complete the proof of Lemma 4.2 and Lemma 4.3.

Lemma 4.2 (Lower Bound on $2k$-th Moment). *Let $d/2 \geqslant k \geqslant 2$ and let $x \sim \mathcal{D}$, where $\mathcal{D}$ is a reasonably anticoncentrated distribution. Further, for any set $S \subset [d]$ such that $|S| = k$, denote $\mathbb{E}\left[\prod_{i\in S} x_i^2\right] = \Phi_k$. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda\|v\|_2^4$ for some $\lambda \in (0, 2/k^2)$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda\|v\|_2^4 \right\} \left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \geqslant \left(\frac{(2k)!}{2^k k!}\right) \cdot \Phi_k\left(1 - \frac{(k-1)\lambda}{2}\right)\|v\|_2^{2k} \right\}.$$

Proof of Lemma 4.2. Using the multinomial theorem and substitution rule, we have

$$\begin{aligned} \left|\frac{v}{4k}\right. \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] &= \sum_{\substack{r_1,\dots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\dots+r_d=2k}} \frac{(2k)!}{\prod_{j\in[d]} r_j!}\, \mathbb{E}\left[\prod_{i\in[d]} (v_j x_j)^{r_j}\right] \\ &= \sum_{\substack{r_1,\dots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\dots+r_d=2k}} \frac{(2k)!}{\prod_{j\in[d]} r_j!} \left(\prod_{i\in[d]} v_j^{r_j}\right) \cdot \left(\mathbb{E}\left[\prod_{j\in[d]} x_j^{r_j}\right]\right) \Bigg\}, \end{aligned} \tag{46}$$

Now, consider the monomials that appear with at least 1 odd term above. We can represent them as a product of squares as and a product of literals as follows: $\prod_{i\in\mathcal{T}} x_i \prod_{j\in[d]} x_j^{2\ell_j}$, where $T \subseteq [d]$ such that $|T| \leqslant k$, and $\ell_j \in \mathbb{Z}_{\geqslant 0}$. Since $\mathcal{K}$ is unconditional, we know the expectation of this monomial under any signing of the coordinates remains the same. Let $\zeta \in \{-1, 1\}^d$ be such that for some $i^* \in \mathcal{T}$, $\zeta_{i^*} = -\text{sign}(x_{i^*})$ and for all $i \neq i^* \in \mathcal{T}$, $\zeta_{i^*} = \text{sign}(x_{i^*})$. Then,

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\prod_{j\in[d]} x_j^{r_j}\right] = \mathbb{E}\left[\prod_{i\in\mathcal{T}} x_i \cdot \prod_{j\in[d]} x_j^{2\ell_j}\right] &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\prod_{i\in\mathcal{T}} \zeta_i x_i \cdot \prod_{j\in[d]} (\zeta_j x_j)^{2\ell_j}\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[-x_{i^*} \prod_{i\neq i^*\in\mathcal{T}} x_i \cdot \prod_{j\in[d]} (x_j)^{2\ell_j}\right] \\ &= -\mathbb{E}\left[\prod_{i\in\mathcal{T}} x_i \cdot \prod_{j\in[d]} x_j^{2\ell_j}\right] \end{aligned}$$

Therefore, we can conclude that all terms appearing with at least one odd power must be 0.

Now,

$$\begin{aligned}\left|\frac{v}{4k}\right\{ \mathbb{E}\left[\langle x,v\rangle^{2k}\right] &= \sum_{\substack{\ell_1,\ldots\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{(2k)!}{\prod_{j\in[d]}(2\ell_j)!}\left(\prod_{i\in[d]} v_j^{2\ell_j}\right)\cdot\left(\mathbb{E}\left[\prod_{j\in[d]} x_j^{2\ell_j}\right]\right)\\ &\geqslant \left(\frac{(2k)!}{2^k}\right)\left(\sum_{\substack{S\subset[n]\\|S|=k}}\left(\prod_{j\in S} v_j^2\right)\left(\mathbb{E}\left[\prod_{j\in S} x_j^2\right]\right)\right)\Big\},\end{aligned} \tag{47}$$

Again, using the substitution rule, we have

$$\left|\frac{v}{4k}\right\{ \|v\|_2^{2k} = \left(\sum_{i\in[d]} v_i^2\right)^k = \sum_{\substack{\ell_1,\ldots,\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{k!}{\prod_{j\in[d]}\ell_j!}\prod_{j\in[d]} v_j^{2\ell_j}\Big\}. \tag{48}$$

We again split the term above into terms where all the ℓ_j's are at most 1 and the terms where at one ℓ_j is at least 2 to get,

$$\begin{aligned}\left|\frac{v}{4k}\right\{ \sum_{\substack{\ell_1,\ldots,\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{k!}{\prod_{j\in[d]}\ell_j!}\prod_{j\in[d]} v_j^{2\ell_j} &\leqslant \left(k!\sum_{\substack{S\subset[n]\\|S|=k}}\prod_{j\in S} v_j^2\right) + \frac{k}{2}\sum_{j\in[d]} v_j^2 \sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-1\\ r_j\geqslant 1}} \frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]} v_i^{2r_i}\\ &= \left(k!\sum_{\substack{S\subset[n]\\|S|=k}}\prod_{j\in S} v_j^2\right) + \frac{k}{2}\sum_{j\in[d]} v_j^4 \sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}} \frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]} v_i^{2r_i}\Big\}.\end{aligned} \tag{49}$$

In the second term, the inner summation is independent of j, so

$$\left|\frac{v}{4k}\right\{ \sum_{j\in[d]} v_j^4 \sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}} \frac{(k-1)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]} v_i^{2r_i} = \left(\|v\|_4^4\right)\cdot(k-1)\sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}} \frac{(k-2)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]} v_i^{2r_i}\Big\} \tag{50}$$

To bound the second term above, we use

$$\left|\frac{v}{4k}\right\{ \sum_{\substack{r_1,\ldots r_d\in\mathbb{Z}_{\geqslant 0}\\ r_1+\ldots+r_d=k-2}} \frac{(k-2)!}{\prod_{i\in[d]} r_i!}\prod_{i\in[d]} v_i^{2r_i} = \|v\|_2^{2k-4}\Big\}. \tag{51}$$

to conclude

$$\left\{\ \|v\|_4^4\leqslant\lambda\|v\|_2^4\ \right\}\left|\frac{v}{4k}\right\{ \sum_{\substack{\ell_1,\ldots,\ell_d\in\mathbb{Z}_{\geqslant 0}\\ \ell_1+\ldots+\ell_d=k}} \frac{k!}{\prod_{j\in[d]}\ell_j!}\prod_{j\in[d]} v_j^{2\ell_j} \leqslant k!\sum_{\substack{S\subset[n]\\|S|=k}}\prod_{j\in S} v_j^2 + \frac{k(k-1)}{2}\left(\lambda\|v\|_2^{2k}\right)\Big\}. \tag{52}$$

Therefore,

$$\left\{\ \|v\|_4^4\leqslant\lambda\|v\|_2^4\ \right\}\vdash\left\{\ \sum_{\substack{S\subset[n]\\|S|=k}}\prod_{j\in S} v_j^2 \geqslant \frac{1}{k!}\left(1-\frac{k(k-1)\lambda}{2}\right)\|v\|_2^{2k}\ \right\}.$$

Using the notation $\Phi_k = \mathbb{E}\left[\prod_{j\in S} x_j^2\right]$ for any set $S \subset [d]$ of size k and rearranging,

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash_{\frac{v}{4k}} \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \geqslant \left(\frac{(2k)!}{2^k k!}\right) \cdot \Phi_k \cdot \left(1 - \frac{k(k-1)\lambda}{2}\right) \|v\|_2^{2k} \right\}. \tag{53}$$

□

Lemma 4.3 (Upper Bound on $2k$-th Moment). *Let $k \geqslant 2$ and let $x \sim \mathcal{D}$ where $\mathcal{D}$ is a reasonably anticoncentrated distribution. Let $v \in \mathbb{R}^d$ be such that $\|v\|_4^4 \leqslant \lambda \|v\|_2^4$ for some $\lambda \in (0, 1/(k^k c_k^k))$. Then,*

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash_{\frac{v}{4k}} \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] \leqslant \left(\frac{(2k)!}{2^k k!}\right) \left(1 + \left(\lambda c_k^{k-2} k^k\right)\right) \|v\|_2^{2k} \right\}.$$

Proof of Lemma 4.3. Let $i_1, i_2, \ldots, i_t \subset [d]$ and let $\ell_1, \ell_2, \ldots, \ell_t \in \mathbb{Z}_{\geqslant 1}$ such that $\sum_{j\in[t]} \ell_j = k$. Setting $f(x_{i_1}) = x_{i_1}^{2\ell_1}$ and $g\left(x_{i_j}\right) = x_{i_j}^{2\ell_j}$ for $j \in [2, t]$, and observing that $x^2 = |x|^2$,

$$\mathbb{E}\left[\prod_{j\in[t]} x_{i_j}^{2\ell_j}\right] \leqslant \mathbb{E}\left[x_{i_1}^{2\ell_1}\right] \cdot \mathbb{E}\left[\prod_{j\in[2,t]} x_{i_j}^{2\ell_j}\right].$$

Repeating the argument above t times, we can conclude $\mathbb{E}\left[\prod_{j\in[t]} x_{i_j}^{2\ell_j}\right] \leqslant \prod_{j\in[t]} \mathbb{E}\left[x_{i_j}^{2\ell_j}\right]$.

Then,

$$\vdash_{\frac{v}{4k}} \left\{ \mathbb{E}\left[\langle x, v\rangle^{2k}\right] = \sum_{\substack{\ell_1,\ldots\ell_d\in\mathbb{Z}_{\geqslant 0} \\ \ell_1+\ldots+\ell_d=k}} \frac{(2k)!}{\prod_{j\in[d]}(2\ell_j)!} \mathbb{E}\left[\prod_{i\in[d]} (v_j x_j)^{2\ell_j}\right] \leqslant \sum_{\substack{\ell_1,\ldots\ell_d\in\mathbb{Z}_{\geqslant 0} \\ \ell_1+\ldots+\ell_d=k}} \frac{(2k)!}{\prod_{j\in[d]}(2\ell_j)!} \prod_{i\in[d]} \mathbb{E}\left[(v_j x_j)^{2\ell_j}\right] \right\} \tag{54}$$

The rest of the argument is identical to Lemma 4.5. □

4.2 Anti-Concentration Certificates with shifts.

Next, we show that given a sufficiently ℓ_4-dense direction, we can obtain a certificate of anti-concentration around a point that need not be the origin. Here, we use an explicit construction for the box indicator polynomial, as defined in Section 2.4.

Theorem 4.6 (Certificates with Shifts). *Let v be a vector -valued indeterminate. Given a distribution $\mathcal{D}$ which is reasonably anticoncentrated and given a shift $\zeta > 0$, if $\lambda \leqslant \frac{\delta}{(2d)^{4d}\zeta^d}$,*

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash \left\{ \mathbb{E}\left[p^2(\langle x, v\rangle - \zeta)\right] \leqslant O(\delta)\|v\|_2^2 \right\},$$

where $p^2(z)$ is a degree $t = O\left(\log\left(\frac{1}{\delta} + \frac{\zeta}{\delta}\right) \cdot \left(\frac{1}{1/\delta^2} + \zeta^2\right)\right)$ polynomial.

In order to prove this theorem, we first show that p^2 is anticoncentrated around the shift when the input distribution is Gaussian. Then we show that sufficiently ℓ_4 dense directions have low-degree moments that concentrate around those of a Gaussian, which suffice to prove the aforementioned theorem.

Now, we are ready to provide a proof of the multi-variate inequality. At a high level, we reduce to the univariate inequality and then appeal to the fact that every univariate inequality has a sum-of-squares proof.

Proof of Theorem 4.6. Let $q\big(\langle \Sigma^{\dagger/2}x, v\rangle - \zeta\big) = \|v\|^t p\big(\langle \Sigma^{\dagger/2}x, v\rangle/\|v\| - \zeta\big)$. Since p^2 is an even polynomial, we have

$$q^2\Big(\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle - \zeta\Big) = \|v\|_2^{2t} \cdot \left(\sum_{j\in[t]} c_{2j}\left(\frac{\langle \Sigma^{\dagger/2}x, v\rangle}{\|v\|} - \zeta\right)^{2j}\right)^2$$

Observe, when $x \sim \mathcal{N}(0,\Sigma)$, $\Sigma^{\dagger/2}x$ has variance 1 in every direction that lies in the span of Σ, and further by rotational invariance of Gaussians,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma)}\Big[q^2\Big(\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle - \zeta\Big)\Big] &= \|v\|_2^{2t} \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma)}\left[\left(\sum_{j\in[t]} c_{2j}\left(\frac{\langle \Sigma^{\dagger/2}x, v\rangle}{\|v\|} - \zeta\right)^{2j}\right)^2\right] \\ &= \|v\|^{2t} \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,1)}\left[\left(\sum_{j\in[t]} c_{2j}(g-\zeta)^{2j}\right)^2\right].\end{aligned} \tag{55}$$

Therefore, $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\left[q\big(\langle \Sigma^{\dagger/2}x, v\rangle - \zeta\big)\right]$ is a polynomial in the formal variable $\|v\|^2$ and using Theorem 2.30, we have,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma)}\Big[q^2\Big(\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle - \zeta\Big)\Big] \leqslant O(\delta)\|v\|_2^{2t} \tag{56}$$

Since the aforementioned inequality is a univariate polynomial in $\|v\|^2$, invoking Fact 2.10, we have

$$\Big|\!\!-_{\|v\|^2}^{4t}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma)}\Big[q^2\Big(\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle - \zeta\Big)\Big] \leqslant O(\delta)\|v\|_2^{2t}\right\}$$

Now, we reduce to the Gaussian certificate by bounding each low-degree moment, when the samples are drawn from a distribution $\mathcal{D}$. We begin with invoking Lemma 4.2, for all $k \in [t]$,

$$\left\{\|v\|_4^4 \leqslant \lambda\|v\|_2^4\right\} \Big|\!\!-_{4k}^{v}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle^{2k}\right] \geqslant \left(\frac{(2k)!}{2^k k!}\right)\left(1 - \frac{(k-1)\lambda}{2}\right)\|v\|_2^{2k}\right\}.$$

Similarly, invoking Lemma 4.3, for all $k \in [t]$,

$$\left\{\|v\|_4^4 \leqslant \lambda\|v\|_2^4\right\} \Big|\!\!-_{4k}^{v}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle^{2k}\right] \leqslant \left(\frac{(2k)!}{2^k k!}\right)\left(1 + \left(\lambda c_k^{k-2}k^k\right)\right)\|v\|_2^{2k}\right\}.$$

Since each moment is tightly concentrated around the k-th moment of a Gaussian, we expand q^2 in it's monomial basis and bound each term :

$$\vdash \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[q^2\Big(\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle - \zeta\Big)\Big] = \sum_{i\in[t]}\sum_{j\in[t]} \hat{c}_{i,j} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\Big\langle \Sigma^{\dagger/2}x, v\Big\rangle^i\right]\zeta^j \tag{57}$$

where each $|\hat{c}_{i,j}| \leqslant 2^{O(t)}$. Note that we know the coefficients $c_{i,j}$ apriori and can invoke the appropriate upper bound from above for each of the terms in the expansion. Therefore, we have

$$\left\{\|v\|_4^4 \leqslant \lambda\|v\|_2^4\right\} \Big|\!\!-_{4d}^{v}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[p^2(\langle x, v\rangle - \zeta)\right] \leqslant \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)}\left[p^2(z-\zeta)\right]\|v\|_2^2 + (c_d d)^{4d}\zeta^d\lambda\|v\|_2^2\right\} \tag{58}$$

Recall, $\lambda \leqslant \frac{\delta}{(c_d d)^{4d}\zeta^d}$, and thus both terms above can be upper bounded by $O(\delta)$, completing the proof. □

4.3 Proof of Theorem 4.1

Now, we have all the tools we need to finish the proof of Theorem 4.1, restated for convenience.

Theorem 4.1 (Anti-concentration certificates with bounded shifts). *Let v be a vector such that $0 < \|v\|_2^2 \leqslant 1$. Let $\{ x_i \}_{i\in[n]}$ be n iid samples from a distribution $\mathcal{D}$ which is reasonably anticoncentrated. Let p^2 be a degree $t = O\Big(\log\Big(\frac{1}{\delta} + \frac{\zeta}{\delta}\Big) \cdot \Big(\frac{1}{1/\delta^2} + \Delta^2\Big)\Big)$ polynomial such that for all $\zeta^2 \leqslant \Delta^2$ and $\lambda \leqslant \frac{\delta}{(2t)^{4t}\Delta^t}$,*

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash^{v} \left\{ \mathbb{E}\left[p^2(\langle x_i, v\rangle - \zeta)\right] \leqslant O(\delta)\|v\|_2^2 \right\},$$

Further, let $\mathcal{D}'$ be a distribution such that $\|y\|_2 \leqslant \Delta$ almost surely for all $y \sim \mathcal{D}'$ and let $\{ y_i \}_{i\in[n]}$ be n iid samples from $\mathcal{D}'$. Then,

$$\left\{ \|v\|_4^4 \leqslant \lambda \|v\|_2^4 \right\} \vdash \left\{ \frac{1}{n}\sum_{i\in[n]} p^2(\langle x_i, v\rangle - y_i) \leqslant O(\delta)\|v\|_2^2 \right\}.$$

Proof. We know that with probability at least $99/100$, for all $\zeta \in [-\Delta, \Delta]$, we have

$$\frac{1}{n}\sum_{i\in[n]} p^2(g_i - \zeta) \leqslant O(\delta) \tag{59}$$

Observe,

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]} p^2(g_i - y_i) &= \frac{1}{n}\sum_{i\in[n]} \int_{y\in[-\Delta,\Delta]} p^2(g_i - y)\mathbf{1}y = y_i \mu(y)dy \\
&\leqslant \frac{1}{n}\sum_{i\in[n]} \int_{y\in[-\Delta,\Delta]} p^2(g_i - y)\mu(y)dy \\
&= \int_{y\in[-\Delta,\Delta]} \frac{1}{n}\sum_{i\in[n]} p^2(g_i - y)\mu(y)dy \\
&\leqslant O(\delta)\int_{y\in[-\Delta,\Delta]} \mu(y)dy,
\end{aligned}$$

where the last inequality follows from invoking Equation (59) for $\zeta = y$. To complete the proof, observe, we can repeat the analysis in the proof of Theorem 4.6, where we showed that the precondition $\|v\|_4^4 \leqslant \lambda\|v\|_2^2$ implies the low-degree moments of $\langle x_i, v\rangle$ are within tiny additive error of those of a Gaussian. □

5 Reweighting Pseudo-distributions over the sphere

In this section, we show that given a pseudo-distribution μ over a vector valued indeterminate $v \in \mathbb{R}^d$, there exists a re-weighted pseudo-distribution μ' and a fixed vector u_S supported on a small subset of coordinates $S \subset [d]$ such that either the vector $v - u_S$ is *analytically dense* or all the low-degree moments of $\|v - u_S\|$ are bounded in pseudo-expectation.

Theorem 5.1 (Re-weighting pseudo-distributions over the sphere). *Let $k, d \in \mathbb{N}$ and $\delta, \eta > 0$, let $z \in \mathbb{N} \geqslant \log(d)$ and $t = \mathcal{O}(z/(\eta^2\delta^{4k+2}))$. Let μ be a degree-t pseudo-distribution over the vector valued indeterminate $v \in \mathbb{R}^d$ satisfying $\|v\|_2^2 = 1$. Then, Algorithm 5.2 outputs a re-weighted pseudo-distribution μ' and a fixed constant u_S, supported on a subset $S \subset [d]$ of size at most $\mathcal{O}(1/(\delta^{4k+2}\eta^2))$ and satisfies one of the following:*

1. *Hypercontractivity at non-trivial scale:*

$$\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_{2z}^{2z}\right] \leqslant (2\delta)^{2z}\cdot\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_{2}^{2z}\right] \text{ and } \widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_{2}^{2k}\right] > \eta,$$

2. *Bounded pseudo-moments of the norm:* $\forall y \in [k]$,

$$\widetilde{\mathbb{E}}_{\mu'}\left[\|v-u_{\mathcal{S}}\|_{2}^{2y}\right] \leqslant \eta. \tag{60}$$

Algorithm 5.2 (Re-Weighting Pseudo-Distributions)**.**

Input: Integers $z, k, \in \mathbb{N}$ and $0 < \eta, \delta < 1$. A degree-ϕ pseudo-distribution over a vector valued indeterminate $v \in \mathbb{R}^d$ satisfying the constraint $\|v\|^2 = 1$, such that $z \geqslant \log(d)$ and $\phi = \Omega\left(z/\delta^{4k+2}\eta^2\right)$.

Operation:

1. Let $u_S = 0$, $v^{(0)} = v$ and $\mu_0 = \mu$. Let $L = \mathcal{O}(1/(\delta^{4k+2}\eta^2))$.
2. For $\ell \in [L]$,
   (a) If $\forall y \in [k]$, $\widetilde{\mathbb{E}}_{\mu}\left[\|v^{(\ell-1)}\|^{2y}\right] \leqslant \eta$ or $\widetilde{\mathbb{E}}_{\mu}\left[\|v^{(\ell-1)}\|_{2z}^{2z}\right] \leqslant (2\delta)^{2z}\widetilde{\mathbb{E}}_{\mu}\left[\|v^{(\ell-1}\|_{2}^{2z}\right]$, stop.
   (b) Else, find a coordinate $i_\ell \in [d]$ such that

$$\widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[\left(v_{i_\ell}^{(\ell-1)}\right)^{2z}\right] > \delta^{2z}\cdot\widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[\|v^{(\ell-1)}\|_{2}^{2z}\right].$$

   (c) Compute a re-weighted pseudo-distribution μ_ℓ such that Lemma 5.4 is satisfied.
   (d) Set $v^{(\ell)} = v^{(\ell)} - \widetilde{\mathbb{E}}_{\mu_\ell}[v_{i_\ell}]$ and $u_{\mathcal{S}} = u_{\mathcal{S}} + \widetilde{\mathbb{E}}_{\mu_\ell}[v_{i_\ell}]$.

Output: A re-weighted pseudo-distribution μ' and a fixed constant $u_{\mathcal{S}} \in \mathbb{R}^d$ supported on at most $\mathcal{O}(1/\delta^{4k+2}\eta^2)$ coordinates such that the conclusion of Theorem 5.1 holds.

We will prove this theorem in the rest of the section. But first, we need a few technical ingredients, starting with the following re-weighting lemma from Barak, Kothari and Steurer [BKS17]:

Lemma 5.3 (Scalar Reweighting, Lemmas 6.1, 6.3 [BKS17])**.** *Given $t \in \mathbb{N}$ let μ be a pseudo-distribution over $\mathbb{R}$ and z be an indeterminate such that $\widetilde{\mathbb{E}}_{\mu}\left[z^2\right] \geqslant 1$ and $\left\{ z^2 \leqslant \Delta^2 \right\}$, for some $\Delta \geqslant 1$. Given $\varepsilon > 0$, there is a degree-$\left(ct\log(\Delta)/\varepsilon^2\right)$ re-weighting of μ, for a fixed constant $c \geqslant 1$, denoted by μ' such that*

1. *Fixing the moments: For some $m \in \mathbb{R}$ be such that $|m| \geqslant \sqrt{\widetilde{\mathbb{E}}_{\mu'}[z^2]}$,*

$$\widetilde{\mathbb{E}}_{\mu'}\left[(z-m)^{2t}\right] \leqslant \varepsilon^{2t}m^{2t},$$

2. *Concentration :*

$$(1+\varepsilon)m \geqslant \widetilde{\mathbb{E}}_{\mu'}[z] \geqslant (1-\varepsilon)m.$$

We note that the last inequality above follows from observing that by Holder's inequality for pseudo-distributions,

$$\left(\widetilde{\mathbb{E}}_{\mu'}\left[z-m\right]\right)^{2t} \leqslant \widetilde{\mathbb{E}}_{\mu'}\left[(z-m)^{2t}\right] \leqslant \varepsilon^{2t}m^{2t},$$

where the last inequality follows from property (2) in Lemma 5.3. Taking the $2t$-th root implies $\widetilde{\mathbb{E}}_{\mu'}\left[z\right] = (1\pm\varepsilon)m$. Next, we prove the following lemma that shows that we can fix each coordinate, given that a large enough pseudo-moment of the coordinate is non-trivially large.

Lemma 5.4 (Key Coordinate Re-weighting Lemma)**.** *Let μ be a degree-ϕ pseudo-distribution over $v \in \mathbb{R}^d$ satisfying $\{\, \|v\|^2 \leqslant 1 \,\}$ and let $t \in \mathbb{N}$. Further, assume $\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2k}\right] \geqslant \Delta$ and that there exists an $i \in [d]$ and $k \in \mathbb{N}$ such that $\widetilde{\mathbb{E}}_{\mu}\left[v_i^{2k}\right] \geqslant \delta^k \widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2k}\right]$. Given $\varepsilon > 0$, there exists a re-weighting of μ, denoted by μ' such that*

1. $\left(\widetilde{\mathbb{E}}_{\mu'}\left[v_i\right]\right)^2 \geqslant (1-\varepsilon)^8\delta\left(\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2k}\right]\right)^{1/k}$, *and,*
2. $\widetilde{\mathbb{E}}_{\mu'}\left[\left(v_i - \widetilde{\mathbb{E}}_{\mu'}\left[v_i\right]\right)^{2t}\right] \leqslant \varepsilon^{2t}\left(\widetilde{\mathbb{E}}_{\mu'}\left[v_i\right]\right)^{2t}$,
3. $\forall \ell \in [k]$, *let q be the smallest power of 2 larger than 2ℓ. Then, $\widetilde{\mathbb{E}}_{\mu'}\left[v_i^{2\ell}\right] \geqslant (1-\varepsilon)^8\delta^q\Delta$.*

assuming $\phi \geqslant \left(ck + \left(ct\log\left(\frac{1}{\delta\Delta}\right)\right)/\left(\varepsilon^2\delta^2\right)\right)$, for a large enough fixed constant c.

Proof. First, we consider the case where $k = 1$, i.e. $\widetilde{\mathbb{E}}_{\mu}\left[v_i^2\right] \geqslant \delta\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^2\right] \geqslant \delta\Delta$. Then, consider the indeterminate $z_i = \frac{v_i}{\sqrt{\delta\Delta}}$. We have

$$\widetilde{\mathbb{E}}_{\mu}\left[z_i^2\right] = \frac{\widetilde{\mathbb{E}}_{\mu}\left[v_i^2\right]}{\delta\Delta} \geqslant 1, \tag{61}$$

where the inequality follows from our assumption. Further, since $\{\, \|v\|^2 \leqslant 1 \,\} \big|\!\frac{v}{4}\, \{\, v_i^2 \leqslant 1 \,\}$, for all $i \in [d]$, we obtain the following:

$$\vdash \left\{ z_i^2 \leqslant \frac{v_i^2}{\delta\Delta} \leqslant \frac{1}{\delta\Delta} \right\}. \tag{62}$$

Invoking Property (1) in Lemma 5.3, since μ has degree at least $\Omega(t\log(1/\delta\Delta)/\varepsilon^2)$, we have

$$\widetilde{\mathbb{E}}_{\mu'}\left[\left(v_i - \widetilde{\mathbb{E}}_{\mu}\left[v_i\right]\right)^{2t}\right] \leqslant \varepsilon^{2t}\widetilde{\mathbb{E}}\left[v_i\right]^{2t}, \tag{63}$$

as desired. Further, using Property (2) in Lemma 5.3, we have

$$\left(\widetilde{\mathbb{E}}_{\mu'}\left[v_i\right]\right)^2 \geqslant (1-\varepsilon)^2\widetilde{\mathbb{E}}_{\mu'}\left[v_i^2\right] \geqslant (1-\varepsilon)^2\widetilde{\mathbb{E}}_{\mu}\left[v_i^2\right] \geqslant (1-\varepsilon)^2\delta\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^2\right], \tag{64}$$

where the second inequality follows from Fact 2.16 and the last inequality follows from the definition of the i-th coordinate.

Now, consider the case where $k > 1$ and the indeterminate $\frac{v_i^k}{\delta^{k/2}\sqrt{\Delta}}$. Again, by assumption, we have $\frac{\widetilde{\mathbb{E}}_{\mu}\left[v_i^{2k}\right]}{\delta^k\Delta} \geqslant 1$. Further, $\vdash \left\{ \frac{v_i^{2k}}{\delta^k\Delta} \leqslant \frac{1}{\delta^k\Delta} \right\}$. Using Property (2) in Lemma 5.3, and recalling that k is even, we can set $m = \sqrt{\widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{2k}\right]}$ and obtain a re-weighting μ_1 of μ such that

$$\widetilde{\mathbb{E}}_{\mu_1}\left[v_i^k\right] \geqslant (1-\varepsilon)\sqrt{\widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{2k}\right]} \geqslant (1-\varepsilon)\sqrt{\widetilde{\mathbb{E}}_{\mu}\left[v_i^{2k}\right]} \geqslant (1-\varepsilon)\left(\delta^{k/2}\sqrt{\Delta}\right), \tag{65}$$

where the second inequality follows from invoking Fact 2.16 and the last inequality follows from our assumption about coordinate i. Further, for all even $\ell \in [k, 2k-2]$, we have

$$\widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{\ell}\right] \geqslant \widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{2k}\right] \geqslant \widetilde{\mathbb{E}}_{\mu}\left[v_i^{2k}\right] \geqslant \delta^k \Delta, \tag{66}$$

where the first inequality follows from the fact that $\left\{ \|v\|^2 \leqslant 1 \right\} \Big|_2^v \left\{ v_i^2 \leqslant 1 \right\}$, the second follows from Fact 2.16 and the last one follows from our assumption on coordinate i.

We then repeat this argument by considering the indeterminate $\frac{v_i^{k/2}}{\left(\delta^k \Delta\right)^{1/4}}$, and obtain a pseudo-distribution μ_2 which is a re-weighting of μ_1 such that

$$\widetilde{\mathbb{E}}_{\mu_2}\left[v_i^{k/2}\right] \geqslant (1-\varepsilon)\sqrt{\widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{k}\right]} \geqslant (1-\varepsilon)^{1.5}\delta^{k/4}\Delta^{1/4}, \tag{67}$$

where the last inequality follows from plugging in the lower bound in Equation (65). Further, for all even $\ell \in [k/2, k]$

$$\widetilde{\mathbb{E}}_{\mu_2}\left[v_i^{\ell}\right] \geqslant \widetilde{\mathbb{E}}_{\mu_2}\left[v_i^{k}\right] \geqslant \widetilde{\mathbb{E}}_{\mu_1}\left[v_i^{k}\right] \geqslant (1-\varepsilon)\left(\delta^{k/2}\sqrt{\Delta}\right), \tag{68}$$

where the first inequality follows from $\left\{ \|v\|^2 \leqslant 1 \right\}$, the second follows from Fact 2.16 and the last follows from Equation (65).

Repeating this argument $L = O(\log(k))$ times, we obtain a pseudo-distribution μ_L such that

$$\widetilde{\mathbb{E}}_{\mu_L}\left[v_i^2\right] \geqslant (1-\varepsilon)^4 \delta \Delta^{1/k}. \tag{69}$$

Since we can set $\varepsilon < 0.01$, we can invoke the argument for the case of $k = 1$, and conclude

$$\widetilde{\mathbb{E}}_{\mu_L}\left[\left(v_i - \widetilde{\mathbb{E}}_{\mu}\left[v_i\right]\right)^{2t}\right] \leqslant \varepsilon^{2t}\left(\widetilde{\mathbb{E}}_{\mu}\left[v_i\right]\right)^{2t},$$

and

$$\left(\widetilde{\mathbb{E}}_{\mu_L}\left[v_i\right]\right)^2 \geqslant (1-\varepsilon)^8 \delta \left(\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2k}\right]\right)^{1/k}. \tag{70}$$

Finally, for all $\ell \in [k]$, let q be the next largest power of two. Then,

$$\widetilde{\mathbb{E}}_{\mu_L}\left[v_i^{\ell}\right] \geqslant (1-\varepsilon)^4 \delta^{q/2} \Delta. \tag{71}$$

Setting $\mu' = \mu_L$ concludes the proof. □

Next, we also need the following lemma to show that if the $2z$-th pseudo-moment of each coordinate is bounded, then the ℓ_{2z}^{2z}-norm of the resulting vector must be bounded, as long as $z \geqslant \log(d)$.

Lemma 5.5 (Bounded Coordinates Implies z to 2 Hypercontractivity)**.** *Given $d \in \mathbb{N}$, let $z \geqslant \log(d)$ and let μ be a degree $2z$ pseudo-distribution on a d-dimensional vector valued indeterminate v. If for all $i \in [d]$, $\widetilde{\mathbb{E}}_{\mu}\left[v_i^{2z}\right] \leqslant \delta^{2z}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2z}\right]$, then $\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_{2z}^{2z}\right] \leqslant (2\delta)^{2z}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2z}\right]$.*

Proof. Summing over all $i \in [d]$, using the hypothesis, we have

$$\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_{2k}^{2k}\right] = \sum_{i\in[d]} \widetilde{\mathbb{E}}_{\mu}\left[v_i^{2k}\right] \leqslant d\delta^{2z}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2z}\right] \leqslant (\delta)^{2z}\frac{d}{2^{2z}}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2z}\right] \leqslant (2\delta)^{2z}\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2z}\right],$$

which concludes the proof. □

Next, we show that a re-weighting that fixes the variance of a coordinate continues to remain fixed under further re-weightings, as long as the the resulting re-weighting polynomial has non-trivial pseudo-expectation and is mildly hypercontractive.

Lemma 5.6 (Variance remains small in subsequent re-weightings)**.** *Let μ be a pseudo-distribution of degree $\geqslant t$ a d-dimensional vector valued indeterminate v satisfying $\{ \|v\|_2^2 \leqslant 1 \}$ and for some $i \in [d]$ let $\widetilde{\mathbb{E}}_\mu \left[\left(v_i - \widetilde{\mathbb{E}}_\mu [v_i] \right)^{2t} \right] \leqslant \varepsilon^{2t}$ for all $t \in [k]$. Suppose that for some $\delta > 0$, there exists a coordinate, $j \in [d]$ such that $\widetilde{\mathbb{E}}_\mu \left[v_j^{2k} \right] \geqslant \delta^{2k} \widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \right]$ and let μ' be a re-weighting of μ satisfying the guarantees of Lemma 5.4. Then, for $t = \Omega\left(k \log\left(\widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \right] \right) \right)$, we have*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\left(v_i - \widetilde{\mathbb{E}}_\mu [v_i] \right) \right]^2 \leqslant \varepsilon^2 / \delta^2.$$

Proof. Recall, by definition, the re-weighting polynomial can be assumed to be v_j^{2k}, i.e.

$$\begin{aligned} \widetilde{\mathbb{E}}_{\mu'} \left[\left(v_i - \widetilde{\mathbb{E}}_\mu [v_i] \right) \right]^2 &= \left(\frac{\widetilde{\mathbb{E}}_\mu \left[\left(v_i - \widetilde{\mathbb{E}}_\mu [v_i] \right) v_j^{2k} \right]}{\widetilde{\mathbb{E}}_\mu \left[v_j^{2k} \right]} \right)^2 \\ &\leqslant \frac{\widetilde{\mathbb{E}}_\mu \left[\left(v_i - \widetilde{\mathbb{E}}_\mu [v_i] \right)^{2t} \right]^{\frac{1}{t}} \cdot \widetilde{\mathbb{E}}_\mu \left[v_j^{\frac{2kt}{t-1}} \right]^{2-\frac{2}{t}}}{\widetilde{\mathbb{E}}_\mu \left[v_j^{2k} \right]} \\ &\leqslant \varepsilon^2 \cdot \widetilde{\mathbb{E}}_\mu [v_i]^2 \cdot \widetilde{\mathbb{E}}_\mu \left[v_j^{2k} \right]^{-\frac{2}{t}} \\ &\leqslant \varepsilon^2 \left(\delta^{2k} \widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \right] \right)^{-2/t} \widetilde{\mathbb{E}}_\mu [v_i]^2, \end{aligned} \tag{72}$$

where the first inequality follows from applying Hólder's inequality for psuedo-distributions, and the last inequality follows from out assumption on v_j. Plugging in $t = \Omega\left(k \widetilde{\mathbb{E}}_\mu \left[\|v\|_2^{2k} \right] \right)$ yields the claim. □

Using these lemmas, we are ready to prove our main theorem on reweighting, restated for convenience.

Theorem 5.1 (Re-weighting pseudo-distributions over the sphere)**.** *Let $k, d \in \mathbb{N}$ and $\delta, \eta > 0$, let $z \in \mathbb{N} \geqslant \log(d)$ and $t = \mathcal{O}(z/(\eta^2 \delta^{4k+2}))$. Let μ be a degree-t pseudo-distribution over the vector valued indeterminate $v \in \mathbb{R}^d$ satisfying $\|v\|_2^2 = 1$. Then, Algorithm 5.2 outputs a re-weighted pseudo-distribution μ' and a fixed constant u_S, supported on a subset $S \subset [d]$ of size at most $\mathcal{O}(1/(\delta^{4k+2}\eta^2))$ and satisfies one of the following:*

1. *Hypercontractivity at non-trivial scale:*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_S\|_{2z}^{2z} \right] \leqslant (2\delta)^{2z} \cdot \widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_S\|_2^{2z} \right] \text{ and } \widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_S\|_2^{2k} \right] > \eta,$$

2. *Bounded pseudo-moments of the norm: $\forall y \in [k]$,*

$$\widetilde{\mathbb{E}}_{\mu'} \left[\|v - u_S\|_2^{2y} \right] \leqslant \eta. \tag{60}$$

Proof. We iteratively re-weight the pseudo-distribution till either the norm of the resulting vector is tiny, or we have *fixed* all the large coordinates in v and the resulting vector is $\ell_{2z\to 2}$-hypercontractive. Let $\mu_0 = \mu$ be a pseudo-distribution that is consistent with the program defined in Equation (12). Let $z = 2\log(d)$, and consider the case where for all $i \in [d]$, $\widetilde{\mathbb{E}}_{\mu_0}\left[v_i^{2z}\right] \leqslant \delta^{2z}\widetilde{\mathbb{E}}_{\mu_0}\left[\|v\|_2^{2z}\right]$. Then, it follows from Lemma 5.5, that

$$\widetilde{\mathbb{E}}_{\mu_0}\left[\|v\|_{2z}^{2z}\right] \leqslant (2\delta)^{2z}\widetilde{\mathbb{E}}_{\mu_0}\left[\|v\|_2^{2z}\right], \tag{73}$$

and therefore the claim follows from setting $\mathcal{S} = \{\,\emptyset\,\}$, and $v_{\mathcal{S}} = 0$.

Otherwise, we know there exists some coordinate $i_1 \in [d]$ such that $\widetilde{\mathbb{E}}_{\mu_0}\left[v_{i_1}^{2z}\right] > \delta^{2z}\widetilde{\mathbb{E}}_{\mu_0}\left[\|v\|_2^{2z}\right]$. We can then apply Lemma 5.4 to obtain a re-weighted pseudo-distribution μ_1 such that

$$\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^2 \geqslant (1-\varepsilon)^4\delta\Big(\widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|^{2z}\right]\Big)^{1/z} \geqslant \delta^2\widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|_2^2\right] \tag{74}$$

and the last inequality follows from setting ε to be appropriately small and invoking Fact 2.7.

Now, consider the vector of indeterminates $v^{(1)} = v - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]$. More generally, we use the notation $v^{(\ell)} = v - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right] - \widetilde{\mathbb{E}}_{\mu_2}\left[v_{i_2}^{(1)}\right] - \ldots - \widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_\ell}^{(\ell-1)}\right]$ to denote the residual indeterminate after ℓ iterations.

Throughout the re-weighting process, we assume that for some $y \in [k]$, $v^{(\ell)}$ satisfies $\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell)}\right\|_2^{2y}\right] \geqslant \eta$, otherwise we simply stop since condition (2) holds. We show that this process must terminate since all the pseudo-moments of the ℓ_2 norm are decreasing monotonically.

Consider the first step of this process. For any $y \in [k]$,

$$\begin{aligned}
&\widetilde{\mathbb{E}}_{\mu_1}\left[\left\|v^{(1)}\right\|_2^{2y}\right] \\
&= \widetilde{\mathbb{E}}_{\mu_1}\left[\left(\sum_{i\neq i_1} v_i^2 + \left(v_{i_1} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]\right)^2\right)^y\right] \\
&= \widetilde{\mathbb{E}}_{\mu_1}\left[\left(\|v\|^2 + \left(v_{i_1} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]\right)^2 - v_{i_1}^2\right)^y\right] \\
&\leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[\left(\|v\|^2 - v_{i_1}^2\right)^y\right] + y\underbrace{\widetilde{\mathbb{E}}_{\mu_1}\left[\left(v_{i_1} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]\right)^2\left(\|v\|^2 - v_{i_1}^2\right)^{2y-2}\right]}_{(75).1} + \underbrace{\widetilde{\mathbb{E}}_{\mu_1}\left[\left(v_{i_1} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]\right)^{2y}\right]}_{(75).2} \\
&\leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[\left(\|v\|^2 - v_{i_1}^2\right)^y\right] + y\widetilde{\mathbb{E}}_{\mu_1}\left[\left(v_{i_1} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]\right)^4\right]^{1/2}\widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|^2\right] + \varepsilon^{2y}\widetilde{\mathbb{E}}_{\mu_1}\left[v_i\right]^{2y} \\
&\leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|^{2y}\right] - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}^{2y}\right] + y\varepsilon^2\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^2 + \varepsilon^{2y}\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^{2y}.
\end{aligned} \tag{75}$$

where the first inequality follow from the fact that $\vdash \{\,(a+b)^{2t} \leqslant a^{2t} + 2tba^{2t-1} + b^{2t}\,\}$, the second inequality follows from applying SoS Cauchy-Schwarz to term (75).1 and using property (2) from Lemma 5.4 to bound term (75).2. The last inequality follows from using property (2) again and invoking Fact 2.17.

It follows from the constraint that $\|v\|_2^2 = 1$ and Hölder's inequality for pseudo-distributions that $\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^2 \leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}^2\right] \leqslant 1$ and similarly $\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^{2y} \leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}^{2y}\right] \leqslant \widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|_2^{2y}\right]$. Further, by Lemma 5.4,

$$\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}^{2y}\right] \geqslant (1-\varepsilon)^4\delta^{2y}\left(\widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2y}\right]\right).$$

Combining these bounds and recalling the constraint that $\|v\|_2^2 = 1$, we have

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu_1}\left[\left\|v^{(1)}\right\|_2^{2y}\right] &\leqslant \widetilde{\mathbb{E}}_{\mu}\left[\|v\|_2^{2y}\right] - (0.99\delta^{2y} - \varepsilon^{2y})\widetilde{\mathbb{E}}_{\mu_1}\left[\|v\|_2^{2y}\right] + \varepsilon y \\ &\leqslant 1 - 0.9\delta^{2y}.\end{aligned} \tag{76}$$

where the last inequality follows from setting $\varepsilon \ll \delta^{2y}/y$. Therefore, in a single iteration of the re-weighting we have made progress.

Assuming the iterative process does not halt due to either the norm becoming small or the vector satisfying $\ell_{2z\to 2}$-hypercontractivity, we show that in each subsequent iteration, the norm of the residual vector decreases additively. Consider the ℓ-th iteration: repeating a similar argument to Equation (75), we have

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell)}\|^{2y}\right] &= \widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(\|v^{(\ell-1)}\|^2 + \left(v_{i_\ell} - \widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_\ell}\right]\right)^2 - v_{i_\ell}^2\right)^{2y}\right] \\ &\leqslant \underbrace{\widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell-1)}\|_2^{2y}\right]}_{(77).(1)} - \widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_\ell}^2\right]^{2y} + y\varepsilon^2\widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_1}\right]^2 + \varepsilon^{2y}\widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_\ell}\right]^{2y}\end{aligned} \tag{77}$$

We now unroll the indeterminate $v^{(\ell-1)}$, and show that at each step, the variance (and higher moments) of all the coordinates fixed in previous steps remains bounded. Focusing on Term (77).(1), it suffices to show that for any $m \in [1, \ell]$, we have:

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell-m)}\right\|^{2y}\right] &= \widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(\|v^{(\ell-m-1)}\|^2 + \left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[v_{i_{\ell-m}}\right]\right)^2 - v_{i_{\ell-m}}^2\right)^{y}\right] \\ &\leqslant \widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell-m-1)}\|^{2y}\right] - \widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}^{2y}\right] + \underbrace{\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_{\ell-m}}\right]\right)^{2y}\right]}_{(78).1} \\ &\quad + y\underbrace{\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}\right]\right)^2\left(\|v^{(\ell-m)}\|^2 - v_{i_{\ell-m}}^2\right)^{2y-2}\right]}_{(78).2}\end{aligned} \tag{78}$$

where the inequality follows from the fact that $\vdash \{\ (a+b)^{2t} \leqslant a^{2t} + 2tba^{2t-1} + b^{2t}\ \}$ and Fact 2.17. We can now bound terms (78).1 and (78).2 by repeatedly invoking Lemma 5.6 as follows:

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_{\ell-m}}\right]\right)^{2y}\right] &\leqslant (0.1\delta^2)\widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_{\ell-m}}\right]\right)^{2yt}\right]^{1/t} \\ &\leqslant (0.1\delta^2)^2\widetilde{\mathbb{E}}_{\mu_{\ell-2}}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_{\ell-m}}\right]\right)^{4yt}\right]^{1/2t} \\ &\ \ \vdots \\ &\leqslant (0.1\delta^2)^m\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[\left(v_{i_{\ell-m}} - \widetilde{\mathbb{E}}_{\mu_1}\left[v_{i_{\ell-m}}\right]\right)^{2^m yt}\right]^{2/(2^m t)} \\ &\leqslant (0.1\delta^2)^m\varepsilon^{2y}\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[v_{i_{\ell-m}}\right]^{2y}\end{aligned} \tag{79}$$

and similarly,

$$\begin{aligned}
&\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(v_{i_{\ell-m}}-\widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}\right]\right)^2\left(\|v^{(\ell-m)}\|^2-v_{i_{\ell-m}}^2\right)^{2y-1}\right]\\
&\leqslant \widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(v_{i_{\ell-m}}-\widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}\right]\right)^4\right]^{1/2}\cdot\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left(\|v^{(\ell-m)}\|^2-v_{i_{\ell-m}}^2\right)^{4y-4}\right]^{1/2}\\
&\leqslant \widetilde{\mathbb{E}}_{\mu_{\ell-1}}\left[\left(v_{i_{\ell-m}}-\widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}\right]\right)^{4t}\right]^{1/2t}\\
&\leqslant (0.01\delta^2)^m\varepsilon^4\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[v_{i_{\ell-m}}\right]^4
\end{aligned} \tag{80}$$

Combining the two equations above, and substituting back into Equation (78) we have

$$\begin{aligned}
&\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell-m)}\right\|^{2y}\right]\\
&\leqslant \widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell-m-1)}\|^{2y}\right]-\widetilde{\mathbb{E}}_{\mu_\ell}\left[v_{i_{\ell-m}}^{2y}\right]+\left(0.1\delta^2\right)^m\varepsilon^{2y}\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[v_{i_{\ell-m}}\right]^{2y}+(0.01\delta^2)^m\varepsilon^4\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[v_{i_{\ell-m}}\right]^4\\
&\leqslant \widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell-m-1)}\|^{2y}\right]-\left(0.99\delta^{2y}-\left(0.1\delta^2\right)^m\varepsilon^{2y}\right)\widetilde{\mathbb{E}}_{\mu_{\ell-m}}\left[\|v^{(\ell-m-1)}\|^{2y}\right]+(0.01\delta^2)^m\varepsilon^4\\
&\leqslant \widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell-m-1)}\|^{2y}\right]-0.9\delta^{2y}\eta
\end{aligned} \tag{81}$$

where the last inequality follows from recalling that $\varepsilon \ll \delta^{2k}\eta/k$. Substituting the above back into Equation (77), we have

$$\begin{aligned}
\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell)}\right\|_2^{2y}\right] &\leqslant \left(\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell-1)}\right\|_2^{2y}\right]-0.9\delta^{2y}\eta\right)\\
&\leqslant \left(\widetilde{\mathbb{E}}_{\mu_\ell}\left[\left\|v^{(\ell-2)}\right\|_2^{2y}\right]-2(0.9)\delta^{2y}\eta\right)\\
&\vdots\\
&\leqslant \left(\widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v\|_2^{2y}\right]-\ell(0.9)\delta^{2y}\eta\right)\\
&\leqslant \left(1-0.9\ell\delta^{2y}\eta\right)
\end{aligned} \tag{82}$$

where the subsequent inequalities follow from repeatedly invoking Equation (78), and the last inequality uses that $\widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v\|\right]=1$, since re-weightings satisfy equality constraints. Therefore, setting $\ell=\Omega\big(k/(\delta^{4k+2}\eta^2)\big)$, either the aforementioned process stops and the resulting vector is $\ell_{2z\to 2}$-hypercontractive or for all $y\in[k]$, $\widetilde{\mathbb{E}}_{\mu_\ell}\left[\|v^{(\ell)}\|^{2y}\right]\leqslant\eta$, as desired. □

6 Clustering a Spectrally Separated Mixture

In this section, we demonstrate the effectiveness of our anti-concentration program to the problem of clustering a mixture of two reasonably anti-concentrated distributions with mean 0 and arbitrary covariances. In the next section, we will show how to apply our techniques to k-clustering in general whenever they are clusterable.

In particular, let $\mathcal{M}=\frac{1}{2}\mathcal{D}_1(0,\Sigma_1)+\frac{1}{2}\mathcal{D}_2(0,\Sigma_2)$, where $\mathcal{D}_1(0,I)$ and $\mathcal{D}_2(0,I)$ satisfy Definition 2.23. Let $X=\{x_1,x_2,\ldots,x_n\}$ be a set of n i.i.d samples from $\mathcal{M}$.

Theorem 6.1 (Clustering Mixtures with Spectral Separation). *Given $0<\delta<1$, and $n\geqslant n_0$ samples from an isotropic mixture of two $(\delta,\exp(1/\delta^2),O(1))$-reasonably anti-concentrated distributions $\mathcal{D}_1(0,\Sigma_1)$ and $\mathcal{D}_2(0,\Sigma_2)$, let u be a direction such that $u^\top\Sigma_1 u\leqslant\delta^3 u^\top\Sigma_2 u$. Then, there exists an*

algorithm that runs in $\mathcal{O}(n^{\left(\log^2(d)\exp\left(1/\delta^4\right)^{\exp\left(1/\delta^2\right)}\right)})$ *time and with probability at least* $99/100$ *outputs a direction* z *such that* $z^\top \Sigma_1 z \leqslant \mathcal{O}(\delta)\cdot z^\top \Sigma_2 z$, *whenever* $n_0 \geqslant d^{\log(d)\exp\left(1/\delta^4\right)^{\exp\left(1/\delta^2\right)}}$.

We will prove this theorem in the rest of this section.

Algorithm 6.2 (Clustering Spectrally Separated Mixtures).

Input: Samples $\{ x_i \}_{i\in[n]}$ from an isotropic mixture of distributions and $0 < \delta < 1$. Let $r = \Omega\left(\log(d)/\exp\left(\delta^4\right)^{\exp\left(1/\delta^2\right)}\right)$.

Operation:

1. Let $t = \mathcal{O}(r\log(d))$. Compute a degree-$t$ pseudo-distribution μ consistent with the following constraints:

$$\mathcal{C}_\delta = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i\in[n]} w_i = (1-\delta^2)n/2 \\ \forall i \in [n] & w_i\langle x_i, v\rangle^2 \leqslant c\delta w_i \|v\|^2 \\ & \|v\|_2^2 = 1 \end{Bmatrix} \tag{83}$$

2. Pick x_j uniformly at random from $\{ x_i \}_{i\in[n]}$. Re-weight μ by $\langle x_j, v\rangle^{2t}$.
3. Let $\widetilde{\Sigma} = \widetilde{\mathbb{E}}_\mu \left[vv^\top\right]$. Sample $g \sim \mathcal{N}(0,\Sigma)$.

Output: $z = g/\|g\|$ such that $z^\top\Sigma_1 z < \mathcal{O}(\delta)\left(z^\top\Sigma_2 z\right)$.

Lemma 6.3 (Feasibility). *With at least constant probability, the constraint system* $\mathcal{C}_\delta$ *is feasible.*

Proof. Since the mixture is equi-weighted, there is a constant probability that there are at least $n/2$ samples from $\mathcal{D}_1(0,\Sigma_1)$, the cluster with smaller variance along the separating direction u. Let w_i's denote the indicators for the points sampled from $\mathcal{D}_1(0,\Sigma_1)$, denoted by $\mathcal{C}_1$. By Markov, for any $x_i \in \mathcal{C}_1$,

$$\Pr\left[\langle x_i, u\rangle^2 > \frac{1}{\delta^2}\left(u^\top\Sigma_1 u\right)\right] \leqslant \delta^2,$$

and thus for all but $1-\delta^2$ fraction of points in $\mathcal{C}_1$, $\langle x_i, u\rangle^2 \leqslant \frac{1}{\delta^2}\left(u^\top\Sigma_1 u\right) \leqslant 2\delta\|u\|^2$. □

Next, we prove the following key sum-of-squares identity.

Lemma 6.4 (Key SoS Identity). *Let* r *be the degree of the certificate required in Theorem 3.1. Then,*

$$\mathcal{C}_\delta \left|\frac{w,v}{r}\right. \left\{ \left\langle \Sigma_1, vv^\top\right\rangle^2 \right.$$
$$\left. \leqslant \mathcal{O}\left(c_4^4\left(\delta v^\top\Sigma_2 v - \sum_{i\in\mathcal{C}_2} q_i^2(v,w)\cdot w_i\left(v^\top\Sigma_2 v - \frac{\delta\|v\|^2}{100}\right)\right) + \delta^2\right)\right\},$$

where c_4^4 *is the certifiable hyper-contractivity constant.*

Proof.

$$\mathcal{C}_\delta \vdash \left\{ \left\langle \Sigma_1, vv^\top \right\rangle^2 = \left\langle \frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} x_i x_i^\top, vv^\top \right\rangle^2 \right.$$
$$\left. \leqslant \underbrace{2\left(\frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} w_i \langle x_i, v \rangle^2 \right)^2}_{(84).(1)} + \underbrace{2\left(\frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} (1 - w_i) \langle x_i, v \rangle^2 \right)^2}_{(84).(2)} \right\} \tag{84}$$

Using the constraint that $w_i \langle x_i, v \rangle^2 \leqslant \frac{c w_i \|v\|^2}{\Delta}$, we can bound (84).(1) as follows:

$$\mathcal{C}_\delta \vdash \left\{ \frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} w_i \langle x_i, v \rangle^2 \leqslant \frac{c\|v\|^2}{\Delta} \right\} \tag{85}$$

Using certifiable hyper-contractivity, we can bound (84).(2) as follows:

$$\mathcal{C}_\delta \vdash \left\{ \left(\frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} (1 - w_i) \langle x_i, v \rangle^2 \right)^2 \leqslant \left(\frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} (1 - w_i)^2 \right) \cdot \left(\frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} \langle x_i, v \rangle^4 \right) \right.$$
$$\left. \leqslant c_4^4 \cdot \underbrace{\left(\frac{2}{n} \sum_{i \in \mathcal{C}_2} w_i \right)}_{(86).(1)} \cdot \left(v^\top \Sigma_1 v \right)^2 \right\}, \tag{86}$$

where the last inequality follows from recalling our constraint that $\sum_{i \in \mathcal{C}_1} w_i + \sum_{i \in \mathcal{C}_2} w_i = n/2$. We focus on bounding term (86).(1) using our anti-concentration certification on the cluster $\mathcal{C}_2$. Invoking the certificate from Theorem 3.1 along the direction v,

$$\vdash^{w,v} \left\{ \delta - \frac{2}{n} \sum_{i \in \mathcal{C}_2} w_i - \sum_{i \in \mathcal{C}_2} q_i^2(v, w) w_i \left(\delta v^\top \Sigma_2 v - \langle x_i, v \rangle^2 \right) \geqslant 0 \right\} \tag{87}$$

Further, the constraint system implies

$$\mathcal{C}_\Delta \vdash^{w} \left\{ w_i \langle x_i, v \rangle^2 \leqslant \frac{w_i}{\Delta} \|v\|^2 \leqslant \frac{w_i \delta}{100} \|v\|^2 \right\}$$

Therefore,

$$\mathcal{C}_\Delta \vdash^{w} \left\{ \frac{2}{n} \sum_{i \in \mathcal{C}_2} w_i \leqslant \delta - \sum_{\in \mathcal{C}_2} q_i^2(v, w) \cdot w_i \left(v^\top \Sigma_2 v - \frac{\delta \|v\|^2}{100} \right) \right\} \tag{88}$$

Plugging this back in Eq. (86) and combining it with Eqs. (84) and (85), we can conclude that

$$\mathcal{C}_\Delta \vdash^{w,v} \left\{ \left\langle \Sigma_1, vv^\top \right\rangle^2 \right.$$
$$\left. \leqslant \mathcal{O}\left(c_4^4 v^\top \Sigma_1 v \left(\delta v^\top \Sigma_2 v - \sum_{i \in \mathcal{C}_2} q_i^2(v, w) \cdot w_i \left(v^\top \Sigma_2 v - \frac{\delta \|v\|^2}{100} \right) \right) + \delta \|v\|^4 \right) \right\}$$

The claim follows from recalling that $v^\top \Sigma_1 v \leqslant 2\|v\|_2^2$ and $\|v\|_2^2 = 1$. □

As a thought experiment, assume we know $v^\top \Sigma_2 v$. Then we re-weight the pseudo-distribution μ with $(v^\top \Sigma_2 v)^{2t}$ for large t. Let the resulting pseudo-distribution be μ'. We show that under the re-weighting, we can lower bound expressions of the form $\widetilde{\mathbb{E}}_{\mu'} \left[q_i^2(w, v) v^\top \Sigma_2 v \right]$.

Lemma 6.5 ("Fixing" the quadratic form). *Let $q_i^2(v,w)$ be are degree-r sum-of-squares polynomials such that $\|q_i^2\|_2^2 \leqslant d^{\mathcal{O}(r)}$. Let μ be a degree-$\Omega(r\log(d))$ distribution consistent with $\mathcal{C}_\Delta$. Let μ' be the pseudo-distribution obtained by re-weighting μ by $(v^\top\Sigma_2 v)^{2t}$, for $t = \mathcal{O}(r\log(d))$. Then,*

$$\mathop{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)v^\top\Sigma_2 v\right] \geqslant 0.9\mathop{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)\|v\|^2\right].$$

Proof. Using the scalar fixing lemma (Lemma 5.3), we have that for $\ell \in [t/100]$

$$\widetilde{\mathbb{E}}_{\mu'}\left[\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)^{2\ell}\right] \leqslant \varepsilon^{2\ell}\left(\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)^{2\ell}. \tag{89}$$

Let μ'' be the pseudo-distribution obtained by re-weighting μ' by the unknown sum-of-squares polynomial $q_i^2(v,w)$.

Then, applying Cauchy-Schwarz for pseudo-distributions,

$$\begin{aligned}
\widetilde{\mathbb{E}}_{\mu''}\left[\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)\right]^{4\ell} &\leqslant \left(\widetilde{\mathbb{E}}_{\mu''}\left[\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)^{2\ell}\right]\right)^2 \\
&= \left(\frac{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]^{2\ell}\right)\right]}{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]}\right)^2 \\
&\leqslant \frac{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^4(w,v)\right]\cdot\widetilde{\mathbb{E}}_{\mu'}\left[\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)^{4\ell}\right]}{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]^2} \\
&\leqslant \frac{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^4(w,v)\right]}{\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]^2}\cdot\varepsilon^{4\ell}\cdot\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]^{4\ell} \\
&\leqslant d^{\mathcal{O}(r)}\cdot\varepsilon^{4\ell}\cdot\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]^{4\ell},
\end{aligned} \tag{90}$$

where the last inequality follows from observing that $\widetilde{\mathbb{E}}_{\mu'}\left[q_i^4(w,v)\right] \leqslant \|q_i\|_2^2\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]^2 \leqslant d^{\mathcal{O}(r)}\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]^2$. Setting $\varepsilon = 0.01$ and $\ell = \Omega(r\log(d))$, and taking the 4ℓ-th root, we get that

$$\left|\widetilde{\mathbb{E}}_{\mu''}\left[\left(v^\top\Sigma_2 v - \widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right]\right)\right]\right| \leqslant 0.1\widetilde{\mathbb{E}}\left[v^\top\Sigma_2 v\right]$$

Plugging in the definition of μ'', we have

$$\begin{aligned}
\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\left(v^\top\Sigma_2 v\right)\right] &\geqslant \widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right] - 0.1\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right] \\
&\geqslant 0.9\cdot\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(w,v)\right]\widetilde{\mathbb{E}}_\mu\left[v^\top\Sigma_2 v\right].
\end{aligned}$$

□

Next, we show that picking a random sample x_j and re-weighting with $\langle x_j, v\rangle^{2t}$ instead, we obtain can obtain the same conclusion as Lemma 6.5.

Lemma 6.6 (Simulating "fixing" the quadratic form). *Let $q_i^2(v,w)$ be are degree-r sum-of-squares polynomials such that $\|q_i^2\|_2^2 \leqslant d^{\mathcal{O}(r)}$. Let μ be a degree-$\Omega(r\log(d))$ distribution consistent with $\mathcal{C}_\Delta$. Let $j \in [n]$ be picked uniformly at random and let μ' be the pseudo-distribution obtained by re-weighting μ by $\langle x_j, v\rangle^{2t}$, for $t = \mathcal{O}(r\log(d))$. Then with probability at least $1/100$, we have*

$$\mathop{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)v^\top\Sigma_2 v\right] \geqslant 0.1\mathop{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)\|v\|^2\right].$$

Proof. Consider the following expression:

$$\begin{aligned}\mathop{\mathbb{E}}_{x_j\in\mathcal{C}_2}\left[\widetilde{\mathbb{E}}_\mu\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\cdot\left\langle x_j,v\right\rangle^{2t}\right]\right] &= \widetilde{\mathbb{E}}_\mu\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\cdot\mathop{\mathbb{E}}_{x_j\in\mathcal{C}_2}\left[\left\langle x_jx_j^\top,vv^\top\right\rangle^{2t}\right]\right]\\ &\geqslant \widetilde{\mathbb{E}}_\mu\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\cdot(v^\top\Sigma_2 v)^{2t}\right]\\ &\geqslant 0.9\widetilde{\mathbb{E}}_\mu\left[q_i^2(w,v)\cdot\|v\|_2^2\cdot(v^\top\Sigma_2 v)^{2t}\right],\end{aligned}$$

where the last inequality follows from clearing the denominator in Lemma 6.5. Then, by Markov's inequality for at least $0.05n$ points $x_j\in\mathcal{C}_2$, it follows that

$$\widetilde{\mathbb{E}}_\mu\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\cdot\left\langle x_j,v\right\rangle^{2t}\right]\geqslant 0.1\widetilde{\mathbb{E}}_\mu\left[q_i^2(w,v)\cdot\|v\|_2^2\cdot(v^\top\Sigma_2 v)^{2t}\right].$$

Therefore, with probability at least $1/100$, we sample a good x_j, and therefore,

$$\begin{aligned}\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\right] &= \frac{\widetilde{\mathbb{E}}_\mu\left[q_i^2(v,w)\cdot v^\top\Sigma_2 v\cdot\left\langle x_j,v\right\rangle^{2t}\right]}{\widetilde{\mathbb{E}}_\mu\left[\left\langle x_j,v\right\rangle^{2t}\right]}\\ &\geqslant 0.1\cdot\frac{\widetilde{\mathbb{E}}_\mu\left[q_i^2(w,v)\cdot\|v\|_2^2\cdot(v^\top\Sigma_2 v)^{2t}\right]}{\widetilde{\mathbb{E}}_\mu\left[\left\langle x_j,v\right\rangle^{2t}\right]}\\ &= 0.1\widetilde{\mathbb{E}}_{\mu'}\left[q_i^2(v,w)\cdot\|v\|^2\right].\end{aligned}$$

□

We can now show that can recover a nearly-separating direction via Gaussian rounding.

Lemma 6.7 (Rounding). *Algorithm 6.2 outputs a direction z such that $z^\top\Sigma_1 z\leqslant\sqrt{\delta}$.*

Proof. We round to the pseudo-expected density matrix $\widetilde{\mathbb{E}}_{\mu'}\left[vv^\top\right]$. It is easy to check that $\mathrm{tr}\left(\widetilde{\mathbb{E}}_{\mu'}\left[vv^\top\right]\right)=1$ and $\widetilde{\mathbb{E}}_{\mu'}\left[vv^\top\right]\succeq 0$. Further, it follows from Hölder's inequality for pseudo-distributions and Eq. (84) that with probability at least $1/100$ over the choice of x_j,

$$\begin{aligned}\left\langle\Sigma_1,\widetilde{\mathbb{E}}_{\mu'}\left[vv^\top\right]\right\rangle^2 &\leqslant \widetilde{\mathbb{E}}_{\mu'}\left[\left\langle\Sigma_1,vv^\top\right\rangle^2\right]\\ &\leqslant \mathcal{O}(1)\widetilde{\mathbb{E}}_{\mu'}\left[\left(\delta v^\top\Sigma_2 v-\sum_{i\in\mathcal{C}_2}q_i^2(v,w)\cdot w_i\left(v^\top\Sigma_2 v-\frac{\delta\|v\|^2}{100}\right)\right)+\delta\right] \qquad (91)\\ &\leqslant \mathcal{O}(\delta+\delta).\end{aligned}$$

Taking square-roots, $\left\langle\Sigma_1,\widetilde{\mathbb{E}}_\mu\left[vv^\top\right]\right\rangle\leqslant\mathcal{O}(\sqrt{\delta})$. Let $\widetilde{\Sigma}\sim\mathcal{N}\left(0,\widetilde{\mathbb{E}}_\mu\left[vv^\top\right]\right)$, and observe that $\mathbb{E}\left[gg^\top\right]=\widetilde{\mathbb{E}}_\mu\left[vv^\top\right]$. Therefore,

$$\mathop{\mathbb{E}}_g\left[g^\top\Sigma_1 g\right]=\left\langle\Sigma_1,\mathbb{E}\left[gg^\top\right]\right\rangle=\left\langle\Sigma_1,\widetilde{\mathbb{E}}\left[vv^\top\right]\right\rangle\leqslant\mathcal{O}(\sqrt{\delta}). \qquad (92)$$

The claim follows from applying Markov's inequality. □

We finally complete the proof of Theorem 6.1, restated for convenience.

Theorem 6.1 (Clustering Mixtures with Spectral Separation). *Given $0 < \delta < 1$, and $n \geqslant n_0$ samples from an isotropic mixture of two $(\delta, \exp(1/\delta^2), O(1))$-reasonably anti-concentrated distributions $\mathcal{D}_1(0, \Sigma_1)$ and $\mathcal{D}_2(0, \Sigma_2)$, let u be a direction such that $u^\top \Sigma_1 u \leqslant \delta^3 u^\top \Sigma_2 u$. Then, there exists an algorithm that runs in $\mathcal{O}(n^{\left(\log^2(d)\exp\left(1/\delta^4\right)\right)^{\exp\left(1/\delta^2\right)}})$ time and with probability at least 99/100 outputs a direction z such that $z^\top \Sigma_1 z \leqslant \mathcal{O}(\delta) \cdot z^\top \Sigma_2 z$, whenever $n_0 \geqslant d^{\log(d)\exp\left(1/\delta^4\right)^{\exp\left(1/\delta^2\right)}}$.*

Proof. We use Algorithm 6.2. We compute the degree t pseudo-distribution satisfying C_δ, which by Lemma 6.4 is feasible. The reweighting in step 2 simulates fixing the quadratic form, as per Lemma 6.6. This gives rise to the lower bounds used in the analysis of Lemma 6.7, which will show that Gaussian rounding outputs a valid separating direction (where we use the fact that the distribution is isotropic). □

7 Clustering Mixtures of Reasonably Anti-Concentrated distributions

In this section, we extend our techniques to cluster a mixture of k distributions, whenever they are *clusterable* according to the following notion:

Definition 7.1 (Δ- Clusterable Mixture). Given $\Delta \gg 1$ and an equi-weighted mixture of k *reasonably anti-concentrated* distributions, denoted by $\mathcal{M} = \sum_{i\in[k]} \mathcal{D}(\mu_i, \Sigma_i)$, we call the mixture *clusterable* if for every distinct pair $i, j \in [k]$, at least one of the following holds:

1. Mean Separation: $\exists v$ such that $\langle \mu_i - \mu_j, v\rangle^2 \geqslant \Delta v^\top(\Sigma_1 + \Sigma_2)v$.
2. Spectral Separation: $\exists v$ such that $\Delta v^\top \Sigma_i v \leqslant v^\top \Sigma_j v$.
3. Frobenius Separation: $\|\Sigma_i^{\dagger/2}\Sigma_j\Sigma_i^{\dagger/2} - I\|_F^2 \geqslant \Delta\|\Sigma_i^{\dagger/2}\Sigma_j^{1/2}\|_{\text{op}}^4$, where $M^\dagger$ denotes the Moore-Penrose psuedo-inverse.

The main result we prove in this section is as follows:

Theorem 7.2 (Clustering Mixtures of Reasonably Anti-Concentrated distributions). *Given $0 < \varepsilon < 1$ and $n \geqslant n_0$ samples from a mixture of k $(\mathcal{O}_{k,\varepsilon}(1), \mathcal{O}_{k,\varepsilon}(1), \mathcal{O}(1))$-reasonably anti-concentrated distributions that also has $\mathcal{O}_{k,\varepsilon}(1)$-certifiably hypercontractive quadratic forms (see Definition 2.19) denoted by $\mathcal{M} = \sum_{i\in k} \mathcal{D}(\mu_i, \Sigma_i)$, such that $\mathcal{M}$ is $\mathcal{O}_{k,\varepsilon}(1)$-clusterable, there exists an algorithm that runs in time $n^{\mathcal{O}_{k,\varepsilon}(\log^2(d))}$ and outputs a clustering $\{\hat{\mathcal{C}}_i\}_{i\in[k]}$ such that with probability at least 99/100, for all $i \in [k]$, $|\hat{\mathcal{C}}_i \cap \mathcal{C}_i| \geqslant (1-\varepsilon)n/k$, where $\mathcal{C}_i$ is the set of points from component $\mathcal{D}(\mu_i, \Sigma_i)$, whenever $n_0 = d^{\Omega_{k,\varepsilon}(\log(d))}$.*

We begin by recalling the notion of a partial clustering, where if any pair of components are mean, spectral or Frobenius separated, we can partition the mixture into two sets such that the partition respects the true clustering and is non-trivial.

Definition 7.3 (Partial Clustering). Given n samples from a mixture of k distributions, denoted by $\mathcal{M} = \frac{1}{k}\sum_{i\in[k]} \mathcal{D}(\mu_i, \Sigma_i)$, let $\{\mathcal{C}_i\}_{i\in[k]}$ denote the true clustering, i.e. $\mathcal{C}_i$ corresponds to the samples generated from $\mathcal{D}(\mu_i, \Sigma_i)$. A bi-partition $\mathcal{P}_1, \mathcal{P}_2$ of the points is ε-approximate partial clustering if

1. Partition respects clustering: for all $i \in [k]$

$$\max\left\{ \frac{k}{n}|\mathcal{C}_i \cap \mathcal{P}_1|, \frac{k}{n}|\mathcal{C}_i \cap \mathcal{P}_2| \right\} \geqslant 1 - \varepsilon,$$

2. The partition is non-trivial:

$$\max_{i\in[k]} \frac{k}{n}|\mathcal{C}_i \cap \mathcal{P}_1|, \max_{i\in[k]} \frac{k}{n}|\mathcal{C}_i \cap \mathcal{P}_2| \geqslant 1 - \eta.$$

Algorithm 7.4 (Clustering Spectrally Separated Mixtures).

Input: Target accuracy $0 < \varepsilon < 1$. Samples $\{ x_i \}_{i\in[n]}$ from an isotropic mixture of k distributions $\mathcal{M} = \sum_{i\in[k]} \mathcal{D}(\mu_i, \Sigma_i)$ such that there exists a pair $i \neq j \in [k]$ and a direction u with $\Delta u^\top \Sigma_i u < u^\top \Sigma_j u$, where $\Delta = 1/\varepsilon^4$. Let $r = \Omega_{k,\varepsilon}(\log(d))$.

Operation:

1. Run the clustering sub-routines corresponding to Theorem 4.2 in [BDJKKV22].
2. Guess the threshold at which the resulting mixture is concentrated along the spectrally separating direction, and let this be δ_k. Guess the fraction of points that are concentrated, i.e. $\langle x_i, v\rangle^2 \leqslant \delta_k \|v\|^2$ for the separating direction v and denote this by η_k.
3. Let $t = \mathcal{O}(r\log(d))$. Compute a degree-$t$ pseudo-distribution μ consistent with the following constraints:

$$\mathcal{C}_{\delta_k,\eta_k} = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i\in[n]} w_i = \eta_k n \\ \forall i \in [n] & w_i\langle x_i, v\rangle^2 \leqslant c\delta_k w_i \|v\|^2 \\ & \|v\|_2^2 = 1 \end{Bmatrix} \tag{93}$$

4. Terminate if the system is infeasible and return the current set as the clustering.
5. Pick x_j uniformly at random from $\{ x_i \}_{i\in[n]}$. Re-weight μ by $\langle x_j, v\rangle^{2t}$.
6. Let $\widetilde{\Sigma} = \widetilde{\mathbb{E}}_\mu [vv^\top]$. Sample $g \sim \mathcal{N}(0, \Sigma)$. Consider the projected instance $\langle x_i, g/\|g\|\rangle$, and let $\mathcal{P}_1$ be the set of points such that $\langle x_i, g/\|g\|\rangle \leqslant c\delta_k$.
7. Recurse on $\mathcal{P}_1$ and $\{ x_i \} \setminus \mathcal{P}_1$.

Output: A clustering $\{ \hat{\mathcal{C}}_i \}_{i\in[k]}$.

Next, we recall results from [BDJKKV22] that demonstrate how to obtain a partial clustering if the mixture is mean or Frobenius separated. These results do not rely on certificates of anti-concentration, and therefore can be used black box. Further, in our setting, we do consider corruptions, so the sample complexity and running time only depend on the desired accuracy.

Lemma 7.5 (Partial Clustering under Mean/ Frobenius Separation,Thm 4.2 [BDJKKV22]). *Given $0 < \varepsilon < 1$ and $n \geqslant n_0$ samples $\{ x_i \}_{i\in[n]}$ from a mixture of k reasonably anti-concentrated distributions,*

$\mathcal{M} = \frac{1}{k}\mathcal{D}(\mu_i, \Sigma_i)$, *let* $i \neq j \in [k]$ *be such that* $\mathcal{D}(\mu_i, \Sigma_i)$ *and* $\mathcal{D}(\mu_j, \Sigma_j)$ *are either mean-separated or Frobenius separated with* $\Delta = (k/\varepsilon)^{\mathcal{O}(k)}$. *Let* $\{ \mathcal{C}_i \}_{i\in[k]}$ *be the true clustering. Then, there exists an algorithm that runs in* $\mathcal{O}(n^{(k/\varepsilon)^{\mathcal{O}(k)}})$ *time and with probability at least* 99/100, *outputs an* ε*-partial clustering (see Definition 7.3) whenever* $n_0 = \Omega\left(d^{(k/\varepsilon)^{\mathcal{O}(k)}}\right)$.

Next, we show that the algorithm from Section 6 can be adapted to work whenever there is a multiplicative gap in the second moments, for an equi-weighted mixture of k reasonably anti-concentrated distributions, where the mixture is in isotropic position.

Lemma 7.6 (Anti-Concentration program)**.** *Given* $n \geqslant n_0$ *samples from a mixture of* k *reasonably anti-concentrated distributions,* $\mathcal{M} = \frac{1}{k}\sum_{i\in[k]} \mathcal{D}(\mu_i, \Sigma_i)$, *in isotropic position, such that* $\|\mu_i\|^2 \leqslant \mathrm{poly}(k)$, *let* v *be a direction satisfying* $(v^\top \Sigma_1 v) \leqslant \varepsilon k^{-\mathrm{poly}(k)} v^\top \Sigma_j v$ *for some* $j \in [k]$. *Then, there exists an algorithm that runs in time* $n^{\mathcal{O}_{k,\varepsilon}(\log^2(d))}$ *and with probability at least* 99/100, *outputs an* ε*-partial clustering of the samples.*

Proof. Let $\sigma_1, \sigma_2, \ldots, \sigma_k$ be the variance of the unpaired mixture along v. We know that $\sigma_1 < k^{-k^{100}}$ and $\sigma_k \leqslant k$. Consider the case where $\sigma_k < 1/k^k$. Then, the must be some mean μ_i such that $|\langle \mu_i, v\rangle| \geqslant 1/2$, since the mixture is isotropic and the mixture variance along v is 1 and since the mean of the mixture is 0, there must be some μ_j such that $|\langle \mu_j, v\rangle| \geqslant 1/k$ on the opposite side of μ_i, along v. Therefore, there exists a pair i, j such that $\langle \mu_i - \mu_j, v\rangle^2 \geqslant k^{\mathrm{poly}(k)}\sigma_k$, i.e. the mixture is mean separated and this is a contradiction.

It suffices to consider the case where $\sigma_k \geqslant 1/k^k$. Since $\sigma_1 \leqslant 1/k^{k^{100}}$, observe there exists a variance $\sigma_t \in [k^{-k^{100}}, k]$ such that $\sigma_t < k^{-k^{10}}\sigma_{t+1}$ and guess $\delta_k = \sigma_t \pm \varepsilon k^{-k^{10}}$. Let $\mathcal{S}_1$ denote the effective component corresponding to points from $\mathcal{C}_1 \ldots \mathcal{C}_t$ and $\mathcal{S}_2$ be the complement. The empirical mean of $\mathcal{S}_1$ is bounded, so we can guess it by discretizing the interval $[-k, k]$ at granularity $\varepsilon k^{-\mathrm{poly}(k)}$. Note, this only makes the success probability $\varepsilon k^{-\mathrm{poly}(k)}$, but we can boost this up by repeating the algorithm several times. Assuming we have guessed the mean, $\mu_{\mathcal{S}_1}$, correctly, we can shift the samples by $\mu_{\mathcal{S}_1}$ so that $\mathcal{S}_1$ is mean 0. Further, $\mathcal{S}_1$ and $\mathcal{S}_2$ have a spectral gap of k^{-k^2} and we can essentially reduce to the 2 component case.

Next, we show that we can modify the proof of Theorem 6.1 to work in this setting. Let $k \cdot \eta_k$ be the guess of how many components lie in $\mathcal{S}_1$. Next, we note that since the mixture is isotropic, $\|\mu_i\| \leqslant k$, and thus $\|\mu_{\mathcal{S}_2}\| \leqslant k$. Now, we observe that we can generalize Lemma 6.4 as follows: repeating the previous analysis, we have

$$\mathcal{C}_{\delta_k,\eta_k} \vdash \left\{ \left\langle \Sigma_{\mathcal{S}_1}, vv^\top \right\rangle \leqslant \left(\frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}(1-w_i)\right)\cdot\left(c_4^4 (v^\top \Sigma_{\mathcal{S}_1} v)^2\right) \right.$$
$$\left. \leqslant c_4^4\left(\frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_2} w_i\right)\cdot\left(v^\top \Sigma_{\mathcal{S}_1} v\right)^2 \right\},$$

where the last inequality follows from $\sum_{i\in\mathcal{S}_1} w_i + \sum_{i\in\mathcal{S}_2} w_i = \eta_k n$ and $\sum_{i\in\mathcal{S}_1} 1 = \eta_k n$. Let $\mu_{\mathcal{S}_2}$ be the mean of the components in $\mathcal{S}_2$. Invoking our the shifted anti-concentration certificate from

Corollary 3.6, with $\phi = \Omega(\|\mu_{\mathcal{S}_2}\|^2)$ for the cluster $\mathcal{S}_2$, we have

$$\begin{aligned}\overset{w,v}{\vdash} \left\{ \delta_k - \frac{1}{|\mathcal{S}_2|}\sum_{i\in\mathcal{S}_2} w_i \geqslant \sum_{i\in\mathcal{S}_2} q^2(v,w)w_i\Big(\delta_k v^\top \Sigma_{\mathcal{S}_2} v - \Big(\langle x_i - \mu_{\mathcal{S}_2}, v\rangle^2 - \phi\|v\|^2\Big)\Big) \right. \\ \geqslant \sum_{i\in\mathcal{S}_2} q^2(v,w)w_i\Big(\delta_k v^\top \Sigma_{\mathcal{S}_2} v - \Big(\langle x_i, v\rangle^2 + \|\mu_{\mathcal{S}_2}\|^2\|v\|^2 - \phi\|v\|^2\Big)\Big) \\ \left. \geqslant \sum_{i\in\mathcal{S}_2} q^2(v,w)w_i\Big(\delta_k v^\top \Sigma_{\mathcal{S}_2} v - \delta\|v\|^2/100\Big)\right\},\end{aligned}$$

where the last inequality follows from our constraints, and $\phi = \Omega(\|\mu_{\mathcal{S}_2}\|^2)$. Since $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ are within a factor of k of each other, we can conclude that

$$\mathcal{C}_{\delta_k,\eta_k} \vdash \left\{ \left\langle \Sigma_{\mathcal{S}_1}, vv^\top \right\rangle \leqslant \mathcal{O}(k\delta_k) - \sum_{i\in\mathcal{S}_2} q^2(v,w)w_i\Big(\delta v^\top \Sigma_{\mathcal{S}_2} v - \delta_k\|v\|^2/100\Big) \right\}.$$

Now, with probability at least $1/k$, we sample a point $x_i \in \mathcal{S}_2$ and as before, re-weight by $\langle x_i, v\rangle^{2t}$. In expectation, this fixes $v^\top \Sigma_{\mathcal{S}_2} v$, and repeating the argument in Lemmas 6.6 and 6.7, we can find a direction z such that

$$z\Sigma_{\mathcal{S}_1} z \leqslant \delta_k.$$

We then simply project onto the rounded direction z and get an ε-approximate partial clustering, where components corresponding to $\sigma_1, \ldots, \sigma_t$ are $\mathcal{P}_1$ and the rest are $\mathcal{P}_2$. □

Now, we are ready to combine these algorithms together and complete the proof of Theorem 7.2, restated for convenience.

Theorem 7.2 (Clustering Mixtures of Reasonably Anti-Concentrated distributions)**.** *Given $0 < \varepsilon < 1$ and $n \geqslant n_0$ samples from a mixture of k $(\mathcal{O}_{k,\varepsilon}(1), \mathcal{O}_{k,\varepsilon}(1), \mathcal{O}(1))$-reasonably anti-concentrated distributions that also has $\mathcal{O}_{k,\varepsilon}(1)$-certifiably hypercontractive quadratic forms (see Definition 2.19) denoted by $\mathcal{M} = \sum_{i\in k} \mathcal{D}(\mu_i, \Sigma_i)$, such that $\mathcal{M}$ is $\mathcal{O}_{k,\varepsilon}(1)$-clusterable, there exists an algorithm that runs in time $n^{\mathcal{O}_{k,\varepsilon}(\log^2(d))}$ and outputs a clustering $\{\hat{\mathcal{C}}_i\}_{i\in[k]}$ such that with probability at least 99/100, for all $i \in [k]$, $|\hat{\mathcal{C}}_i \cap \mathcal{C}_i| \geqslant (1-\varepsilon)n/k$, where $\mathcal{C}_i$ is the set of points from component $\mathcal{D}(\mu_i, \Sigma_i)$, whenever $n_0 = d^{\Omega_{k,\varepsilon}(\log(d))}$.*

Proof of Theorem 7.2. Given n samples from a mixture in isotropic position, we run the algorithm corresponding to the case where the mixture is mean or Frobenius separated. It follows from Lemma 7.5 that if algorithm succeeds, we can simply recurse on the resulting partial clustering.

Instead, consider the case where the mixture is not mean or Frobenius separated. Then, the means, μ_i are bounded by $\mathrm{poly}(k)$. Further, we know that for some pair must be spectrally separated and we can invoke Lemma 7.6 on the mixture and again obtain an ε-partial clustering, allowing us to recurse until there's only a single component left. The running time is dominated by the anti-concentration program, and this concludes the proof. □

8 List-Decodable Regression

In this section, we show that our certificate can be used to obtain list-decodable regression algorithms for *reasonably anti-concentrated* distributions.

Model 8.1 (List-Decodable Regression). *Given $0 < \alpha < 1$ and $\ell^* \in \mathbb{R}^d$ such that $\|\ell^*\| = 1$, we define list-decodable learning w.r.t. a distribution $\mathcal{D}$ as, denoted by $\mathsf{Lin}_{\mathcal{D}}(\alpha, \ell^*)$ as follows: generate αn i.i.d. samples, x_i, from $\mathcal{D}$ and set $y_i = \langle x_i, \ell^* \rangle$. Let these $\{ (y_i, x_i) \}_{i \in [\alpha n]}$ be denoted by $\mathcal{I}$. Generate the remaining $(1-\alpha)n$ pairs arbitrarily and potentially adversarially w.r.t. $\mathcal{I}$.*

We show that we can output a short list containing a vector close to the true regressor for any *reasonably anti-concentrated* distribution. This is the task of list-decodable linear regression.

Theorem 8.2 (List-decodable Regression). *For any $0 < \alpha, \eta < 1$, there exists an algorithm that takes $n \geqslant n_0$ samples from $\mathsf{Lin}_{\mathcal{D}}(\alpha, \ell^*)$, where $\mathcal{D}$ is a reasonably anti-concentrated distribution, runs in $\mathcal{O}(n^{\log(d)\exp\left(1/\alpha^8\eta^8\right)^{\exp\left(1/\alpha^4\eta^4\right)}})$ time and outputs a list of $\mathcal{O}(1/\alpha)$ vectors such that with probability at least $99/100$, the list contains at least one vector $\hat{\ell}$ such that $\|\hat{\ell} - \ell\|_2^2 \leqslant \eta$, when $n_0 \geqslant d^{\log(d)\exp\left(1/\alpha^8\eta^8\right)^{\exp\left(1/\alpha^4\eta^4\right)}}$.*

We consider the following constraint system:

$$\mathcal{C}_\alpha = \begin{Bmatrix} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i \in [n]} w_i = \alpha n \\ \forall i \in [n] & w_i(y_i - \langle x_i, \ell \rangle)^2 = 0 \\ & \|v\|_2^2 = 1 \end{Bmatrix} \tag{94}$$

Lemma 8.3 (Key SoS Identity, analog of Lemma 4.1 in [KKK19]). *Given $0 < \alpha, \eta < 1$, let $\delta = \alpha^2\eta^2/4$ and let $r = \Omega\left(\log(d) \cdot \exp\left(1/\delta^4\right)^{\exp\left(1/\delta^2\right)}\right)$. Then,*

$$\mathcal{C}_\alpha \vdash_r^{w,v} \left\{ \frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} w_i \|\ell - \ell^*\|^2 \leqslant \delta \right\}$$

Proof. Observe, for the inliers, our certificate on anti-concentration states

$$\vdash^{w,v} \left\{ \delta - \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} w_i - \sum_{i \in \mathcal{I}} q_i^2(v,w) w_i \left(\delta v^\top \Sigma v - \langle x_i, v \rangle^2 \right) \geqslant 0 \right\}$$

Consider $v = \ell - \ell^*$ and observe that $w_i \langle x_i, \ell - \ell^* \rangle^2 = 0$. Invoking the above identity along $v = \ell - \ell^*$, observe $\delta q_i^2(v,w) w_i (v^\top \Sigma v)$ is a sum-of-squares polynomial and thus $\vdash^w \left\{ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} w_i \leqslant \delta \right\}$. By sum-of-squares triangle inequality, $\{ \|\ell - \ell^*\|^2 \leqslant 2 \}$ since both are unit vectors and therefore

$$\mathcal{C}_\alpha \vdash_r^{w,v} \left\{ \frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} w_i \|\ell - \ell^*\|^2 \leqslant 2\delta \right\},$$

which yields the claim. □

The theorem follows from repeating the remaining arguments in [KKK19].

9 Reasonably anti-concentrated Distributions

In this section, we show that uniform distributions over ℓ_p balls for $p > 1$ are reasonably anti-concentrated.

Theorem 9.1 (Uniform distributions over ℓ_p balls)**.** *For any $p \in [1+\alpha, \infty)$, for a fixed constant α, and any $0 < \delta < 1$, the uniform distribution over the unit ℓ_p ball is δ-reasonably anti-concentrated.*

We need to show that the uniform distribution on ℓ_p balls satisfies anti-concentration, certifiable hypercontractivity, almost product-ness and strictly sub-exponential tails.

First, the uniform distribution over any convex body is log-concave, and therefore, for any univariate projection, the PDF is bounded by $e^{-|\langle x,v\rangle|/c}$, for a fixed universal constant c, along direction v. It is easy to check that such a distribution is anti-concentrated. Next, we show that the distribution is strictly sub-exponential.

Lemma 9.2 (Strictly sub-exponential)**.** *For any $p \in [1+\alpha, \infty)$, where α is a fixed constant, let x be a sample from the uniform distribution on the unit ℓ_p ball. Then, for all unit vectors $v \in \mathbb{R}^d$ and $t > 0$, we have*

$$\Pr[|\langle x, v\rangle| \geqslant t\sqrt{v^T \Sigma v}] \leqslant \exp\left(-t^{1+\alpha}/c\right),$$

where Σ is the covariance of the distribution.

Proof. This result follows from [BGMN05, Proposition 10], where for $p \geqslant 2$, we can choose $\alpha = 1$ because the distribution is sub-Gaussian, i.e. the ψ_2 norm is bounded by a fixed universal constant. For $p \in (1+\alpha, 2]$, the ψ_p norm is bounded by a fixed constant.

□

To show almost-productness, we first show that the uniform distribution over ℓ_p balls satisfy negative association. We invoke a theorem from [PW08], which shows this property for a larger class of uniform distributions over Generalized Orlicz balls.

Definition 9.3 (Generalized Orlicz Ball)**.** A function $f : \mathbb{R}^+ \to \mathbb{R}^+ \cup \{\infty\}$ is a Young function if f is convex, $f(0) = 0$, $\exists x$ such that $f(x) \neq 0$, and $\exists x \neq 0$ such that $f(x) \neq \infty$. Let $\mathcal{F} = \{f_i\}_{i\in[d]}$ be a set of d Young functions. Then,

$$\mathcal{K} = \{x \ : \ \sum_{i\in[d]} f_i(|x_i|) \leqslant 1\}$$

is a Generalized Orlicz Ball.

Note in particular that ℓ_p balls are Generalized Orlicz balls.

Definition 9.4 (Negative Association)**.** A set of random variables $\{x_i\}_{i\in[d]}$ is negatively associated if for any coordinate-wise increasing, bounded, functions f, g and disjoint sets $\mathcal{S}, \mathcal{T} \subset [d]$ such that $\mathcal{S} = \{i_1, i_2, \ldots i_s\}, \mathcal{T} = \{j_1, j_2, \ldots j_t\}$,

$$\mathbb{E}\left[f(x_{i_1}, x_{i_2}, \ldots, x_{i_s}) \cdot g(x_{j_1}, x_{j_2}, \ldots, x_{j_s})\right] \leqslant \mathbb{E}\left[f(x_{i_1}, x_{i_2}, \ldots, x_{i_s})\right] \cdot \mathbb{E}\left[g(x_{j_1}, x_{j_2}, \ldots, x_{j_s})\right].$$

Theorem 9.5 (Generalized Orlicz Ball in the Positive Orthant, Theorem 1.2 in [PW08])**.** *Let $\mathcal{K}$ be a Generalized Orlicz ball in $\mathbb{R}^d$ and let $\mathcal{D}$ be the uniform distribution over $\mathcal{K}$. If $x \sim \mathcal{D}$, then the set of random variables $\{\ |x_i|\ \}_{i\in[d]}$ is negatively associated.*

This establishes one side of almost k-wise independence. We establish the other side via standard integration techniques. In particular, we show that several families of distributions satisfy the following criterion:

Lemma 9.6 (The solid ℓ_p ball is almost-product). *Let $x \sim \mathcal{D}$ where $\mathcal{D}$ is the uniform distribution on the solid ℓ_p ball in $\mathbb{R}^d$, that is, $\mathbb{B}_p = \{x \in \mathbb{R}^d, \|x\|_p^p \leqslant 1\}$, for any $p \geqslant 1$. Then, for any subset $S \subset [d]$, such that $|S| = k$,*

$$\mathbb{E}\left[\prod_{j\in S} x_j^2\right] \geqslant (1 - o_d(1)) \prod_{j\in S} \mathbb{E}\left[x_j^2\right],$$

where $k \ll d$.

Proof. To prove this, we can compute these expectations explicitly as follows:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\prod_{i\in S} x_i^2\right] = \frac{\Gamma(3/p)^k\Gamma(d/p+1)}{\Gamma(1/p)^k\Gamma((d+2k)/p+1)}.$$

The proof of this equality follows [Wan05] so we only give a brief sketch. Let the uniform distribution on the standard ℓ_2 ball B_2^d be $\mathcal{D}'$. Let $\phi(y) = (y_1^{2/p}, \ldots, y_d^{2/p})$ with the corresponding Jacobian determinant $J\phi(y) = (2/p)^d y_1^{2/p-1} \ldots y_d^{2/p-1}$. Then,

$$\int_{\mathcal{D}} f(x)\,dx = \int_{\mathcal{D}'} f(x)|J\phi(y)|\,dy = (2/p)^d \int_{\mathcal{D}'} f(\phi(y))|y_1|^{2/p-1} \ldots |y_d|^{2/p-1}\,dy$$

Therefore,

$$\int_{\mathcal{D}} (\prod_S x_i^2)\,dx = (2/p)^d \int_{\mathcal{D}'} (\prod_S |y_i|^{4/p})|y_1|^{2/p-1} \ldots |y_d|^{2/p-1}\,dy$$

For reals $\alpha_1, \ldots, \alpha_n > -1$, we have

$$\int_{\mathcal{D}'} |y_1|^{\alpha_1} \ldots |y_d|^{\alpha_d}\,dy = \frac{\Gamma(\beta_1)\ldots\Gamma(\beta_d)}{\Gamma(\beta_1+\ldots+\beta_d+1)}$$

where $\beta_i = (\alpha_i+1)/2$ for all i. Using this, we get

$$\int_{\mathcal{D}} (\prod_S x_i^2)\,dx = (2/p)^d \frac{\Gamma(3/p)^k\Gamma(1/p)^{d-k}}{\Gamma((d+2k)/p+1)}$$

Therefore,

$$\mathbb{E}_{\mathcal{D}}[\prod_S x_i^2] = \frac{\int_{\mathcal{D}}(\prod_S x_i^2)\,dx}{\int_{\mathcal{D}} 1\,dx} = \frac{\Gamma(3/p)^k\Gamma(d/p+1)}{\Gamma(1/p)^k\Gamma((d+2k)/p+1)}$$

Next, we use the above characterization to complete the proof.

$$\frac{\mathbb{E}[\prod_{j\in S} x_j^2]}{\prod_{j\in S}\mathbb{E}[x_j^2]} = \frac{\Gamma(d/p+1+2/p)^k}{\Gamma(d/p+1)^{k-1}\Gamma((d/p+1)+(2k/p))}$$

Because the gamma function is strictly logarithmically convex on the positive reals, we have

$$\Gamma(x_1)^t\Gamma(x_2)^{1-t} \geqslant \Gamma(tx_1+(1-t)x_2)$$

for reals $x_1, x_2 > 0$ and $t \in [0,1]$. Setting $x_1 = d/p+1+2/p, y = d/p+1, t = k$, we get

$$\frac{\Gamma(d/p+1+2/p)^k}{\Gamma(d/p+1)^{k-1}} \geqslant \Gamma((d/p+1)+(2k/p)+(1-k))$$

Assuming $A = (d/p+1) + (2k/p) \geqslant k-1$ (follows if $d \gg pk$), we get

$$\begin{aligned}\frac{\mathbb{E}[\prod_{j\in S} x_j^2]}{\prod_{j\in S}\mathbb{E}[x_j^2]} &\geqslant \frac{\Gamma(A+(1-k))}{\Gamma(A)} \\ &= \frac{1}{(A-(k-1))(A-(k-2))\dots A}\end{aligned}$$

If we assume p, k are constants and d is growing, then $A \gg k$ in which case this will be $1 - o_d(1)$. □

Next, we show that we can obtain certifiable hypercontractivity of linear forms.

Lemma 9.7 (Certifiable Hypercontractivity of linear forms)**.** *For any $p \in [1,\infty)$, the uniform distribution on the unit ℓ_p ball has $2k$-certifiably c-hypercontractive linear forms for any $k \in \mathbb{N}$.*

Proof. This proof follows from two known facts. First, Kothari and Steinhardt showed that for any distribution that satisfies a Poincaré inequality with a dimension-independent constant has certifiably-hyper-contractive linear forms [KS18, Theorem 1.1]. Second, for $p \in [1,2]$, Sodin [Sod08] showed that uniform distribution over the ℓ_p ball satisfies a Poincaré inequality with a dimension-independent constant. For $p \in (2,\infty)$, this was shown by Latała and Wojtaszczyk [LW08]. □

Finally, we show that distributions that satisfy a a Poincaré inequality also have certifiably-hypercontractive quadratic forms (recall Definition 2.19).

Lemma 9.8 (Certifiable Hypercontractivity from Poincaré)**.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is σ-Poincaré if for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(f(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(x)]\right)^2\right] \leqslant \sigma^2\|\nabla f(x)\|_2^2,$$

for some fixed constant σ. Any σ-Poincaré distribution has k-certifiably c_k-hyper-contractive quadratic forms.

Proof. For a matrix-valued indeterminate Q, let $f_0(x) = \langle x, Qx\rangle - \mathbb{E}_{x\sim\mathcal{D}}[\langle x, Qx\rangle]$. Then,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[(f_0(x))^2] \leqslant \sigma^2\|\nabla f_0(x)\|_2^2,$$

where

$$\|\nabla f_0(x)\|_2^2 = 4\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Q^\top Qx\right] = 4\left\langle QQ^\top, \Sigma\right\rangle \leqslant 4\|Q\|_F^2,$$

and thus

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\langle x, Qx\rangle - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\langle x, Qx\rangle]\right)^2\right] \leqslant (2\sigma)^2\|Q\|_F^2.$$

Let $f_{2t}(x) = (\langle x, Qx\rangle - \mathbb{E}_{x\sim\mathcal{D}}[\langle x, Qx\rangle])^{2t}$. Next,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(f_{2t}(x))^2\right] &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\left(f_{2t}(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f_{2t}(x)] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f_{2t}(x)]\right)\right)^2\right] \\ &\leqslant 2\underbrace{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(f_{2t}(x) - \mathbb{E}[f_{2t}(x)]\right)^2\right]}_{(95).(1)} + 2\underbrace{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f_{2t}(x)]^2}_{(95).(2)}\end{aligned} \tag{95}$$

We can bound term (95).(1) as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(f_{2t}(x)-\mathbb{E}\left[f_{2t}(x)\right]\right)^2\right] &\leqslant \sigma^2\|\nabla f_{2t}(x)\|_2^2\\ &= \sigma^2(2t)^2\,\mathbb{E}\left[\left(f_{2t-1}(x)^2\cdot\|Qx\|_2^2\right)\right]\end{aligned}$$

□

Acknowledgements We thank Xue Chen for extensive discussions. G. R. was supported in part by NSF grants CCF-1816372 and CCF-2008920.

References

[AA11] Scott Aaronson and Alex Arkhipov. "The computational complexity of linear optics". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 333–342 (page 9).

[AM05] Dimitris Achlioptas and Frank McSherry. "On spectral learning of mixtures of distributions". In: *International Conference on Computational Learning Theory*. Springer. 2005, pp. 458–469 (page 4).

[ARV04] Sanjeev Arora, Satish Rao, and Umesh Vazirani. "Expander flows and a $\sqrt{\log n}$-approximation to sparsest cut". In: 2004 (page 10).

[Bar] Boaz Barak. "Proofs, beliefs, and algorithms through the lens of sum-of-squares". In: () (page 10).

[BBHKSZ12] Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. "Hypercontractivity, sum-of-squares proofs, and their applications". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. 2012, pp. 307–326 (page 10).

[BDHKKK20] Ainesh Bakshi, Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K Kothari. "Outlier-robust clustering of gaussians and other non-spherical mixtures". In: *2020 ieee 61st annual symposium on foundations of computer science (focs)*. IEEE. 2020, pp. 149–159 (page 4).

[BDJKKV22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. "Robustly learning mixtures of k arbitrary gaussians". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 1234–1247 (pages 1, 10, 48).

[BGMN05] Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. "A probabilistic approach to the geometry of the L_p-ball". In: (2005) (page 52).

[BHKKMP16] B. Barak, S. B. Hopkins, J. Kelner, P. Kothari, A. Moitra, and A. Potechin. "A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem". In: 2016, pp. 428–437 (page 10).

[BK20] Ainesh Bakshi and Pravesh Kothari. "Outlier-robust clustering of non-spherical mixtures". In: *arXiv preprint arXiv:2005.02970* (2020) (pages 2, 5, 9, 10).

[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)*. SIAM. 2021, pp. 1279–1297 (pages 1–3, 9, 10).

[BKS17] Boaz Barak, Pravesh K Kothari, and David Steurer. "Quantum entanglement, sum of squares, and the log rank conjecture". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 975–988 (pages 7, 10, 14, 36).

[BLMT23] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. "Learning quantum hamiltonians at any temperature in polynomial time". In: *arXiv preprint arXiv:2310.02243* (2023) (page 10).

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. "Rounding semidefinite programming hierarchies via global correlation". In: *2011 ieee 52nd annual symposium on foundations of computer science*. IEEE. 2011, pp. 472–481 (page 6).

[BRST21] Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. "Continuous lwe". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 694–707 (page 1).

[CCK15] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. "Comparison and anti-concentration bounds for maxima of Gaussian random vectors". In: *Probability Theory and Related Fields* 162 (2015), pp. 47–70 (page 9).

[CGKMS24] Pritam Chandra, Ankit Garg, Neeraj Kayal, Kunal Mittal, and Tanmay Sinha. "Learning Arithmetic Formulas in the Presence of Noise: A General Framework and Applications to Unsupervised Learning". In: *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, January 30 to February 2, 2024, Berkeley, CA, USA*. Ed. by Venkatesan Guruswami. Vol. 287. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, 25:1–25:19. DOI: 10.4230/LIPICS.ITCS.2024.25. URL: https://doi.org/10.4230/LIPIcs.ITCS.2024.25 (page 3).

[CLS20] Sitan Chen, Jerry Li, and Zhao Song. "Learning mixtures of linear regressions in subexponential time via fourier moments". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 587–600 (page 10).

[CR11] Amit Chakrabarti and Oded Regev. "An optimal lower bound on the communication complexity of gap-hamming-distance". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 51–60 (page 9).

[CTV06] Kevin P Costello, Terence Tao, and Van Vu. "Random symmetric matrices are almost surely nonsingular". In: (2006) (page 9).

[DHKK20] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. "Robustly learning any clusterable mixture of gaussians". In: *arXiv preprint arXiv:2005.06417* (2020) (page 9).

[DK23] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023 (page 10).

[DKKLT22] 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*. 2022, pp. 1262–1275 (page 10).

[DKS17a] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. "Learning multivariate log-concave distributions". In: *Conference on Learning Theory*. PMLR. 2017, pp. 711–727 (page 4).

[DKS17b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84 (page 10).

[Erd45] Paul Erdös. "On a lemma of Littlewood and Offord". In: (1945) (page 9).

[FKP+19] Noah Fleming, Pravesh Kothari, Toniann Pitassi, et al. *Semialgebraic Proofs and Efficient Algorithm Design.* now the essence of knowledge, 2019 (pages 4, 13).

[FKP19] N. Fleming, P. Kothari, and T. Pitassi. *Semialgebraic Proofs and Efficient Algorithm Design*. 2019 (page 10).

[GH22] Kristian Georgiev and Samuel Hopkins. "Privacy induces robustness: Information-computation gaps and sparse mean estimation". In: *Advances in neural information processing systems* 35 (2022), pp. 6829–6842 (page 10).

[GJJPR20] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. "Sum-of-Squares lower bounds for Sherrington-Kirkpatrick via Planted Affine Planes". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2020, pp. 954–965 (page 10).

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver. "The ellipsoid method and its consequences in combinatorial optimization". In: *Combinatorica* 1 (1981), pp. 169–197 (page 11).

[Gri01] Dima Grigoriev. "Complexity of Positivstellensatz proofs for the knapsack". In: *computational complexity* 10.2 (2001), pp. 139–154 (pages 10, 11).

[GW95] M.X. Goemans and D.P. Williamson. "Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming". In: 42.6 (1995). Preliminary version in *Proc. of STOC'94*, pp. 1115–1145 (page 10).

[HKM22] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. "Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 1406–1417 (page 10).

[HKMN23] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. "Robustness implies privacy in statistical estimation". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 497–506 (page 10).

[HL18] Samuel B Hopkins and Jerry Li. "Mixture models, robustness, and sum of squares proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1021–1034 (pages 4, 10).

[HM13] Moritz Hardt and Ankur Moitra. "Algorithms and hardness for robust subspace recovery". In: *Conference on Learning Theory*. PMLR. 2013, pp. 354–375 (page 1).

[Hop20] Samuel B Hopkins. "Mean estimation with sub-Gaussian rates in polynomial time". In: *The Annals of Statistics* 48.2 (2020), pp. 1193–1213 (page 10).

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. "Tensor principal component analysis via sum-of-square proofs". In: *Conference on Learning Theory*. PMLR. 2015, pp. 956–1006 (page 10).

[IK22] Misha Ivkov and Pravesh K Kothari. "List-decodable covariance estimation". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 1276–1283 (pages 1, 9).

[Jon22] Christopher Jones. “Symmetrized Fourier Analysis of Convex Relaxations for Combinatorial Optimization Problems”. PhD thesis. The University of Chicago, 2022 (page 10).

[JPRTX22] Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. “Sum-of-squares lower bounds for sparse independent set”. In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 406–416 (page 10).

[JPRX23] Chris Jones, Aaron Potechin, Goutham Rajendran, and Jeff Xu. “Sum-of-squares lower bounds for densest k-subgraph”. In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 84–95 (page 10).

[Kha95] L. Khachiyan. “On the Complexity of Approximating Extremal Determinants in Matrices”. In: *Journal of Complexity* 11.1 (1995), pp. 138–153. ISSN: 0885-064X. DOI: https://doi.org/10.1006/jcom.1995.1005. URL: https://www.sciencedirect.com/science/article/pii/S0885064X85710059 (page 1).

[KK10] Amit Kumar and Ravindran Kannan. “Clustering with Spectral Norm and the k-Means Algorithm”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 2010, pp. 299–308. DOI: 10.1109/FOCS.2010.35 (page 4).

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. “List-decodable linear regression”. In: *Advances in neural information processing systems* 32 (2019) (pages 1, 2, 5, 9, 10, 17, 18, 51).

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. “Efficient algorithms for outlier-robust regression”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 1420–1430 (page 10).

[KMOW17] Pravesh Kothari, Ryuhei Mori, Ryan O’Donnell, and David Witmer. “Sum of squares lower bounds for refuting any CSP”. In: 2017 (page 10).

[KMV22] Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. “Private robust estimation by stabilizing convex relaxations”. In: *Conference on Learning Theory*. PMLR. 2022, pp. 723–777 (page 10).

[KNS10] Guy Kindler, Assaf Naor, and Gideon Schechtman. “The UGC hardness threshold of the Lp Grothendieck problem”. In: *Mathematics of Operations Research* 35.2 (2010), pp. 267–283 (pages 5, 28, 29).

[Kri16] MANJUNATH Krishnapur. “Anti-concentration inequalities”. In: *Lecture notes* (2016) (page 1).

[KS17a] Pravesh K Kothari and Jacob Steinhardt. “Better agnostic clustering via relaxed tensor norms”. In: *arXiv preprint arXiv:1711.07465* (2017) (page 10).

[KS17b] Pravesh K Kothari and David Steurer. “Outlier-robust moment-estimation via sum-of-squares”. In: *arXiv preprint arXiv:1711.11581* (2017) (pages 4, 10).

[KS18] Pravesh Kothari and Jacob Steinhardt. “Better Agnostic Clustering via Relaxed Tensor Norms”. In: *Symposium on Theory of Computing (STOC)*. 2018 (pages 15, 54).

[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*. 2018, pp. 1035–1046 (page 1).

[Kun20] Dmitriy Kunisky. "Positivity-preserving extensions of sum-of-squares pseudomoments over the hypercube". In: *arXiv preprint arXiv:2009.07269* (2020) (page 10).

[Lau09] Monique Laurent. "Sums of squares, moment matrices and optimization over polynomials". In: *Emerging applications of algebraic geometry*. Springer, 2009, pp. 157–270 (page 13).

[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*. 2022, pp. 1248–1261 (page 10).

[LM21] Allen Liu and Ankur Moitra. "Settling the robust learnability of mixtures of gaussians". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 518–531 (pages 1, 10).

[Lov10] Shachar Lovett. "An elementary proof of anti-concentration of polynomials in Gaussian variables". In: *Electronic Colloquium on Computational Complexity (ECCC)*. Vol. 17. 2010, p. 182 (page 9).

[LS01] Wenbo V Li and Q-M Shao. "Gaussian processes: inequalities, small ball probabilities and applications". In: *Handbook of Statistics* 19 (2001), pp. 533–597 (page 9).

[LW08] Rafał Latała and Jakub Onufry Wojtaszczyk. "On the infimum convolution inequality". In: *arXiv preprint arXiv:0801.4036* (2008) (page 54).

[MP16] Dhruv Medarametla and Aaron Potechin. "Bounds on the norms of uniform low degree graph matrices". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2016 (page 9).

[Nes00] Yurii Nesterov. "Squared functional systems and optimization problems". In: *High performance optimization*. Springer, 2000, pp. 405–440 (pages 10, 11).

[Nie22] Zipei Nie. "Matrix anti-concentration inequalities with applications". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 568–581 (page 9).

[Par00] Pablo A Parrilo. "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization". PhD thesis. California Institute of Technology, 2000 (pages 10, 11).

[PR20] Aaron Potechin and Goutham Rajendran. "Machinery for Proving Sum-of-Squares Lower Bounds on Certification Problems". In: *arXiv preprint arXiv:2011.04253* (2020) (page 10).

[PR22] Aaron Potechin and Goutham Rajendran. "Sub-exponential time Sum-of-Squares lower bounds for Principal Components Analysis". In: *Advances in Neural Information Processing Systems*. 2022 (page 10).

[PW08] Marcin Pilipczuk and Jakub Onufry Wojtaszczyk. "The negative association property for the absolute values of random variables equidistributed on a generalized Orlicz ball". In: *Positivity* 12.3 (2008), pp. 421–474 (page 52).

[Rag08] Prasad Raghavendra. "Optimal algorithms and inapproximability results for every CSP?" In: 2008, pp. 245–254 (page 10).

[Raj22] Goutham Rajendran. "Nonlinear Random Matrices and Applications to the Sum of Squares Hierarchy". PhD thesis. University of Chicago, 2022 (page 10).

[RT23] Goutham Rajendran and Madhur Tulsiani. "Concentration of polynomial random matrices via Efron-Stein inequalities". In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2023, pp. 3614–3653 (page 9).

[RY20a] Prasad Raghavendra and Morris Yau. "List decodable learning via sum of squares". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2020, pp. 161–180 (pages 1, 2, 5, 9, 10).

[RY20b] Prasad Raghavendra and Morris Yau. "List decodable subspace recovery". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3206–3226 (pages 1, 2, 9).

[She12] Alexander A Sherstov. "The communication complexity of gap hamming distance". In: *Theory of Computing* 8.1 (2012), pp. 197–208 (page 9).

[Sho87] Naum Zuselevich Shor. "An approach to obtaining global extremums in polynomial mathematical programming problems". In: *Cybernetics* 23.5 (1987), pp. 695–700 (pages 10, 11).

[Sim15] Max Simchowitz. "Dictionary Learning and Anti-Concentration". In: (2015) (page 9).

[Sod08] Sasha Sodin. "An isoperimetric inequality on the L_p balls". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 44. 2. 2008, pp. 362–373 (page 54).

[SS17] Tselil Schramm and David Steurer. "Fast and robust tensor decomposition with applications to dictionary learning". In: *Conference on Learning Theory*. PMLR. 2017, pp. 1760–1793 (page 10).

[Tro15] Joel A Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230 (page 9).

[Ver20] Roman Vershynin. "High-dimensional probability". In: *University of California, Irvine* 10 (2020), p. 11 (page 15).

[Vid03] René Esteban Vidal. *Generalized principal component analysis (gpca): an algebraic geometric approach to subspace clustering and motion segmentation*. University of California, Berkeley, 2003 (page 3).

[Vid12] Thomas Vidick. "A concentration inequality for the overlap of a vector on a large set". In: *Chicago Journal of Theoretical Computer Science* 1 (2012), pp. 1–12 (page 9).

[VW04] Santosh Vempala and Grant Wang. "A spectral algorithm for learning mixture models". In: *Journal of Computer and System Sciences* 68.4 (2004), pp. 841–860 (page 4).

[Wan05] Xianfu Wang. "Volumes of generalized unit balls". In: *Mathematics Magazine* 78.5 (2005), pp. 390–395 (page 53).

A Bounding the coefficients of the certificates

In this section, we argue the polynomials that arise in our SoS certificate have polynomially bounded bit complexity, which is needed for efficient algorithms. The strategy is to instantiate

the SoS identity for various values of v, w and then analyze the identities together to bound the coefficients of q_i.

In the previous sections, we showed the existence of the identity

$$\delta n - \sum_{i=1}^{n} w_i = \left(\sum_j p_{0,j}(v,w)^2\right) + \sum_{i \leqslant n} \left(\sum_j p_{i,j}(v,w)^2\right) (\delta w_i \cdot v^T \Sigma v - w_i \langle v, x_i\rangle^2) + \sum_i q_i(v,w)(w_i^2 - w_i) + r(w,v)(\|v\|^2 - 1)$$

where $p_{0,j}, p_{i,j}, q_i, r$ are polynomials of degree at most t.

Lemma A.1. $\|p_{0,j}\|^2, \|p_{i,j}\|^2, \|q_i\|^2, \|r\|^2 \leqslant d^{2t}$

Proof. We will show the argument for $\|p_{1,j}\|^2$, the rest are similar. Assume without loss of generality that $\Sigma = I$. Fix a subset U of indices in $[n]$ of size at most t. We will consider the coefficients of the terms which depends on $(w_i : i \in U)$ and not on $(w_i : i \notin U)$. To do this, set $w = w^*$ where $w_i^* = 1$ if $i = 1$ or $i \in U$ and 0 otherwise. Also sample v from a Gaussian distribution with $\mathbb{E}\|v\|^2 = 1$ such that $\langle v, x_i\rangle^2 \leqslant \frac{\delta^2}{100}$ w.h.p. for all $i \in U \cup \{1\}$. Note that this can be done because there are only $t+1$ such constraints. Moreover, we can choose the mean of v to be 0 and covariance Θ of the distribution of v to be $\eta^2 = \delta^2/(400 \log t)$ in the space corresponding to the span of $\{x_i : i \in U \cup \{1\}\}$ and 1 in every direction orthogonal to this subspace. Then, we scale the covariance by a constant factor so as to satisfy the norm constraint. With this choice, then the identity simplifies to (after taking expectation)

$$\delta n - |U \cup \{1\}| = \left(\sum_j \mathbb{E}_v p_{0,j}(v,w^*)^2\right) + \sum_{i \in U \cup \{1\}} \left(\sum_j p_{i,j}(v,w^*)^2\right) \left(\frac{99\delta}{100}\right)$$

This implies

$$\delta n \geqslant \frac{99\delta}{100} \sum_{i \in U \cup \{1\}} \mathbb{E}_v \left(\sum_j p_{i,j}(v,w^*)^2\right) \tag{96}$$

where many terms got zeroed out because of our choice of w^* and for the nonzero w_i^*, we invoked the constraint bound on the terms $\delta w_i \cdot v^T \Sigma v - w_i \langle v, x_i\rangle^2$.

It now suffices to prove the following claim.

Claim. *For some constant $c = c(t) > 0$*

$$\mathbb{E}_v[p_{ij}(v,w^*)^2] \geqslant \mathbf{Var}[p_{ij}(v,w^*)] \geqslant \left(\frac{c}{\delta^t}\right)\|p_{ij}\|^2, \tag{97}$$

where $\|p_{ij}\|^2$ denotes the sum of the squares of the coefficients.

Once we establish the claim, (97) together with (96) establishes Lemma A.1. We now prove the above claim.

Let $z = \Theta^{-1/2} v$, so z is a standard Gaussian with identity covariance. For simplicity, let the polynomial $f(v) := p_{ij}(v, w^*)$ and let $f(v) = \sum_{j=0}^t f_j(v)$ where f_j is a homogenous polynomial of degree j. Finally let $\forall j \in \{0, 1, \ldots, t\}$ let $h_j(z) = f_j(\Theta^{1/2} z)$, and let $h(z) := \sum_{j=0}^t h_j(z) = f(\Theta^{1/2} z)$.

From standard facts about standard Gaussians, when $z \sim N(0, I)$ we have

$$\mathbf{Var}_v[p_{ij}(v, w^*)] = \mathbf{Var}_v[f(v)] = \mathbf{Var}_z[h(z)] = \|h\|^2 = \sum_{j=0}^{t} \|h_j\|^2, \tag{98}$$

where $\|h\|^2$ denotes the sum of the squares of the coefficients of h.

To relate $\|h_j\|$ and $\|f_j\|$, let us denote by $T_j^{(f)}$ the symmetric tensor that satisfies $f_j(v) = \langle T_j^{(f)}, v^{\otimes j} \rangle$. We have

$$f_j(v) = \langle T_j^{(f)}, v^{\otimes j} \rangle = \left\langle (\Theta^{1/2})^{\otimes j}\, T_j^{(f)}, z^{\otimes j} \right\rangle = h_j(z).$$

Hence it follows that

$$\|h_j\|^2 \geqslant \left\langle (\Theta^{1/2})^{\otimes j}\, T_j^{(f)}, (\Theta^{1/2})^{\otimes j}\, T_j^{(f)} \right\rangle \geqslant \sigma_{\min}(\Theta)^j \cdot \|T_j^{(f)}\|_F^2 \geqslant \eta^j \cdot \frac{\|f_j\|^2}{j!}.$$

$$\text{Hence } \|f_j\|^2 \leqslant \frac{j!}{\eta^j}\|h_j\|^2 \leqslant \frac{t!}{\eta^t}\|h_j\|^2 \leqslant \frac{c_t}{\delta^t} \cdot \|h_j\|^2 \quad \forall j \in \{0, 1, \ldots, t\},$$

and some constant $c_t > 0$ that depends on t. Summing over all $j \in \{0, 1, \ldots, t\}$ and using (98) establishes the claim. This completes the proof of Lemma A.1. □