Hardness of learning noisy halfspaces using polynomial thresholds

Arnab Bhattacharyya[*]   Suprovat Ghoshal[†]   Rishi Saket[‡]

Abstract

We prove the hardness of weakly learning halfspaces in the presence of adversarial noise using polynomial threshold functions (PTFs). In particular, we prove that for any constants $d \in \mathbb{Z}^+$ and $\varepsilon > 0$, it is NP-hard to decide: given a set of $\{-1, 1\}$-labeled points in $\mathbb{R}^n$ whether (YES Case) there exists a halfspace that classifies $(1-\varepsilon)$-fraction of the points correctly, or (NO Case) any degree-d PTF classifies at most $(1/2 + \varepsilon)$-fraction of the points correctly. This strengthens to all constant degrees the previous NP-hardness of learning using degree-2 PTFs shown by Diakonikolas et al. (2011). The latter result had remained the only progress over the works of Feldman et al. (2006) and Guruswami et al. (2006) ruling out weakly proper learning adversarially noisy halfspaces.

1 Introduction

Given a distribution $\mathcal{D}$ over $\{-1, 1\}$-labeled points in $\mathbb{R}^n$, the accuracy of a classifier function $f : \mathbb{R}^n \to \{-1, 1\}$ is the probability that $f(x) = \ell$ for a random point-label pair (x, ℓ) sampled from $\mathcal{D}$. A concept class $\mathcal{C}$ is said to be *learnable* by hypothesis class $\mathcal{H}$ if there is an efficient procedure which, given access to samples from any distribution $\mathcal{D}$ consistent with some $f \in \mathcal{C}$, generates with high probability a classifier $h \in \mathcal{H}$ of accuracy approaching that of f for $\mathcal{D}$. When $\mathcal{H}$ can be taken as $\mathcal{C}$ itself, the latter is said to be *properly* learnable. The focus of this work is one of the simplest and most well-studied concept classes: the *halfspace* which maps $x \in \mathbb{R}^n$ to $\text{sign}(\langle v, x \rangle - c)$ for some $v \in \mathbb{R}^n$ and $c \in \mathbb{R}$. The study of halfspaces goes back several decades to the development of various algorithms in artificial intelligence and machine learning such as the Perceptron [Ros62, MP69] and SVM [CV95]. Since then, halfspace-based classification has found applications in many other areas, such as computer vision [Mur90] and data-mining [RRK04].

It is known that a halfspace can be properly learnt by using linear programming along with a polynomial number of samples to compute a separating hyperplane [BEHW89]. In noisy data however, it is not always possible to find a hyperplane separating the differently labeled points. Indeed, in the presence of (adversarial) noise, i.e. the *agnostic* setting, proper learning of a halfspace to optimal accuracy with no distributional assumptions was shown to be NP-hard by Johnson and Preparata [JP78]. Subsequent results showed the hardness of approximating the accuracy of properly learning a noisy halfspace to constant factors: $\frac{262}{261} - \varepsilon$ by Amaldi and Kann [AK98], $\frac{418}{415} - \varepsilon$ by Ben-David et al. [BDEL03], and $\frac{85}{84} - \varepsilon$ by Bshouty and Burroughs [BB06]. These results were considerably strengthened independently by Feldman, Gopalan, Khot, and Ponnuswami [FGKP09] and by Guruswami and Raghavendra [GR09][1] who proved hardness of even

---

[*]Indian Institute of Science, Bangalore, India. Supported in part by DST Ramanujan Grant DSTO1358. Research partly done while visiting the Simons Institute for Theory of Computing, Berkeley, USA. Email: `arnabb@csa.iisc.ernet.in`

[†]Indian Institute of Science, Bangalore, India. Email: `suprovat.ghoshal@csa.iisc.ernet.in`

[‡]IBM Research, Bangalore, India. Email: `rissaket@in.ibm.com`

[1]The reduction of Guruswami and Raghavendra [GR09] works even for the special case when the points are over the boolean hypercube.

weakly proper learning a noisy halfspace, i.e. to an accuracy beyond the random threshold of $1/2$. This implies an optimal $(2 - \varepsilon)$-inapproximability in terms of the learning accuracy. Building upon these works Feldman, Guruswami, Raghavendra, and Wu [FGRW12] showed that the same hardness holds for learning noisy monomials (OR functions over the boolean hypercube) using halfspaces.

At this point, it is natural to ask whether the halfspace learning problem remains hard if the classifier is allowed to be from a larger class of functions, i.e., *non-proper* learning. In particular, consider the class of degree-d *polynomial threshold functions* (PTF) which are given by mapping $x \in \mathbb{R}^n$ to $\text{sign}(P(x))$ where P is a degree-d polynomial. They generalize halfspaces a.k.a. *linear threshold functions* (LTFs) which are degree-1 PTFs and are very common hypotheses in machine learning because they are output by kernelized models (e.g., perceptrons, SVM's, kernel k-means, kernel PCA, etc.) when instantiated with the polynomial kernel. From a complexity viewpoint, PTFs were studied by Diakonikolas, O'Donnell, Servedio, and Wu [DOSW11] who showed the hardness of weakly proper learning a noisy degree-d PTF for any constant $d \in \mathbb{Z}^+$, assuming Khot's Unique Games Conjecture (UGC) [Kho02]. On the other hand, proving the hardness of weakly learning noisy halfspaces using degree-d PTFs has turned out to be quite challenging. Indeed, the only such result is by Diakonikolas et al. [DOSW11] who showed the corresponding hardness of learning using a degree-2 PTF. With no further progress till now, the situation remained unsatisfactory.

In this work, we significantly advance our understanding by proving the hardness of weakly learning an ε-noisy halfspace by a degree-d PTF *for any constant $d \in \mathbb{Z}^+$*. Our main result is formally stated as follows.

Theorem 1.1. (This work) *For any constants $\delta > 0$, and $d \in \mathbb{Z}^+$, it is* NP*-hard to decide whether a given set of $\{-1, 1\}$-labeled points in $\mathbb{R}^n$ satisfies:*

> **YES Case.** *There exists a halfspace that correctly classifies $(1 - \delta)$-fraction of the points, or*

> **NO Case.** *Any degree-d PTF classifies at most $(1/2 + \delta)$-fraction of the points correctly.*

The NO *case can be strengthened to rule out any function of constantly many degree-d PTFs.*

To place our results in context, we note that algorithmic results for learning noisy halfspaces are known under assumptions on the distribution of the noise or the pointset. In the presence of *random classification noise*, Blum, Frieze, Kannan, and Vempala [BFKV98] gave an efficient learning algorithm approaching optimal accuracy, which was improved by Cohen [Coh97] who showed that in this case the halfspace can in fact be properly learnt. For certain well behaved distributions, Kalai, Klivans, Mansour, and Servedio [KKMS05] showed that halfspaces can be learnt even in the presence of adversarial noise. Subsequent works by Klivans, Long, and Servedio [KLS09], and Awasthi, Balcan, and Long [ABL17] improved the noise tolerance and introduced new algorithmic techniques. Building upon them, Daniely [Dan15] recently obtained a PTAS for minimizing the hypothesis error with respect to the uniform distribution over a sphere. Several of these learning algorithms use halfspaces and low degree PTFs (or simple combinations thereof) as their hypotheses, and one could conceivably apply their techniques to the setting without any distributional assumptions. Our work provides evidence to the contrary.

1.1 Previous related work

Hypothesis-independent intractability results for learning for halfspaces are also known, but they make average-case or cryptographic hardness assumptions which seem considerably stronger than P $\neq$ NP. Specifically, for exactly learning noisy halfspaces, such results have been shown in the works of Feldman et al. [FGKP09], Kalai et al. [KKMS05], Kothari and Klivans [KK14], and Daniely and Shalev-Shwartz [DS16]. In a recent work, Daniely [Dan16] rules out weakly learning noisy halfspaces assuming the intractability of strongly refuting random K-XOR formulas. On the other hand, Applebaum, Barak, and Xiao [ABX08]

have shown that hypothesis-independent hardness results under standard complexity assumptions would imply a major leap in our current understanding of complexity theory and are unlikely to be obtained for the time being. Therefore, any study (such as ours) of the standard complexity-theoretic hardness of learning halfspaces would probably need to constrain the hypothesis.

A natural generalization of the learning halfspaces problem is that of learning intersections of two or more halfspaces. Observe that unlike the single halfspace, properly learning the intersection of two halfspaces without noise does not in general admit a separating hyperplane based solution. Indeed, this problem was shown to be NP-hard by Blum and Rivest [BR93], later strengthened by Alekhnovich, Braverman, Feldman, Klivans, and Pitassi [ABF+08] to rule out intersections of constantly many halfspaces as hypotheses. The corresponding hardness of even weak learning was established by Khot and Saket [KS11], while Klivans and Sherstov [KS09] proved under a cryptographic hardness assumption the intractability of learning the intersection of n^ε halfspaces. Algorithms for learning intersections of constantly many halfspaces have been given in the works of Blum and Kannan [BK97] and Vempala [Vem97] for the uniform distribution over the unit ball, Klivans, O'Donnell, and Servedio [KOS04] for the uniform distribution over the boolean hypercube, and by Arriaga and Vempala [AV06] and Klivans and Servedio [KS08] for instances with good *margin*, i.e. the points being well separated from the hyperplanes.

As was the case for learning a single noisy halfspace, there is no known NP-hardness for learning intersections of two halfspaces using (intersections of) degree-d PTFs. This cannot, however, be said of the finite field analog of learning halfspaces, i.e. the problem of learning noisy parities over $\mathbb{F}[2]$. While Håstad's [Hås01] seminal work itself rules out weakly proper learning a noisy parity over $\mathbb{F}[2]$, later work of Gopalan, Khot, and Saket [GKS10] showed the hardness of learning an ε-noisy parity by a degree-d PTF to within $(1 - 1/2^d + \varepsilon)$-accuracy – which, however, is not optimal for $d > 1$. Shortly thereafter, Khot [Kho09] observed[2] that Viola's [Vio09] pseudo-random generator fooling degree-d PTFs can be combined with coding-theoretic inapproximability results to yield optimal lower bounds for all constant degrees d. From the algorithmic perspective, one can learn an ε-noisy parity over the uniform distribution in $2^{O(n/\log n)}$-time as shown by Feldman et al. [FGKP09] and Blum et al. [BKW03]. For general distributions, Kalai, Mansour, and Verbin [KMV08] gave a non-proper $2^{O(n/\log n)}$-time algorithm achieving an accuracy close to optimal.

Several of the inapproximability results mentioned above, e.g. those of [GR09], [GKS10], [KS11], [FGRW12] and [DOSW11], follow the *probabilistically checkable proof (PCP) test* based approach for their hardness reductions. While our result builds upon these methods, in the remainder of this section, we give an overview of our techniques and describe the key enhancements which allow us to overcome some of the technical limitations of previous hardness reductions.

1.2 Overview of Techniques

For hardness reductions, due to the uniform convergence results of [Hau92, KSS94], it is sufficient to take the optimization version of the learning halfspaces problem which consists of a set of coordinates and a finite set of labeled points, the latter replacing a random distribution. A typical reduction (including ours) given a hard instance of a constraint satisfaction problem (CSP) $\mathcal{L}$ over vertex set V and label set $[k]$, defines $\mathcal{C} := V \times [k]$ to be the set of coordinates over $\mathbb{R}$. We let the formal variables $Y_{(w,i)}$ be associated with the coordinate $(w, i) \in \mathcal{C}$. The hypothesis H (the *proof* in PCP terminology) is defined over these variables. In our case, the proof will be a degree-d PTF. The PCP test chooses randomly a small set of vertices S of $\mathcal{L}$,

---

[2]Khot's observations remained unpublished for while, before they were included with his permission by Bhattacharyya et al. [BGGS16] in their paper which made a similar use of Viola's [Vio09] pseudo-random generator.

$\mathfrak{P}_0(\mathbb{R}^k, \eta, \varepsilon)$ tests halfspace $\text{sign}(f(Y))$.

1. Sample $b \in \{-1, 1\}$ uniformly at random.
2. Choose a random "noise" subset $\mathcal{I} \subseteq [k]$ by including each i independently with probability ε.
3. For $i \in [k] \setminus \mathcal{I}$, set $y_i = b\eta$,
4. For $i \in \mathcal{I}$, sample y_i i.i.d. at random from $N(0, 1)$.
5. Accept iff $\text{sign}(f(y)) = b$.

Figure 1: Dictatorship Test $\mathfrak{P}_0$

and runs a *dictatorship* test on S: it tests H on a set of labeled points $P_S \subseteq \mathbb{R}^\mathcal{C}$ generated by the dictatorship test. We desire the following two properties from the test:

- **(completeness)** if H "encodes" a good labeling for S, then it is a good classifier for P_S,
- **(soundness)** a good classifier H for P_S can be "decoded" into a good labeling for S.

The soundness property is leveraged to show that if H classifies P_S for a significant fraction of the choices S, it can be used to define a good global labeling for $\mathcal{L}$. The CSP of choice in the above template is usually the Label Cover or the Unique Games problem. While the NP-hardness of Label Cover is unconditional, its projective constraints seem to present technical roadblocks – also faced by Diakonikolas et al. [DOSW11] – in analyzing learnability by degree-d ($d > 2$) PTFs.

Our work overcomes these issues and gives a hardness reduction from Label Cover. The key ingredient to incorporate the Label Cover projective constraints is a *folding* over an appropriate subspace defined by them. This amounts to restricting the entire instance to the corresponding orthogonal subspace. Similar folding for analyzing linear forms has been used earlier in the works of Khot and Saket [KS11], Feldman, Guruswami, Raghavendra, and Wu [FGRW12], and Guruswami, Raghavendra, Saket, and Wu [GRSW16]. We are able to extend it over degree-d polynomials leveraging the linear-like structure decoded by an appropriate dictatorship test. This uses a *smoothness* property of the constraints (analogous to [KS11, FGRW12, GRSW16]) of the Label Cover instance which is combined with the dictatorship test – along with folding – to yield the PCP test.

In the rest of this section, we informally describe our dictatorship test, the motivation behind its design and the key ingredients involved in its analysis. To begin, we present a simple preliminary dictatorship test $\mathfrak{P}_0$ over $\mathbb{R}^k$ which works for linear thresholds. Of course, the NP-hardness of properly learning noisy halfspaces is already known [FGKP09, GR09], so this test does not yield anything new. Our purpose is illustrative and we include a sketch of the arguments of its analysis. Taking $\varepsilon > 0$ as a small constant and $\eta > 0$ a small parameter (to be defined later), the description of $\mathfrak{P}_0$ is given in Figure 1.

Observe that the linear threshold $\text{sign}(Y_i)$ for each $i \in [k]$ correctly classifies (y, b) with probability $(1 - \varepsilon)$. In other words, every *dictator* corresponds to a good solution.

1.2.1 Soundness analysis of $\mathfrak{P}_0$

Suppose there exists a linear form $f = \sum_{i \in [k]} \widehat{f}_i Y_i$ (assuming for simplicity f has no constant term) such that $\text{sign}(f)$ passes $\mathfrak{P}_0$ with probability $1/2 + 2\xi$ for some $\xi = \Omega(1)$. Using (by now) standard analytical

arguments, we show that there exists $i^* \in [k]$ such that

$$\widehat{f}_{i^*}^2 \geq \Omega(1) \cdot \sum_{i \in [k]} \widehat{f}_i^2 > 0. \tag{1}$$

In other words, every good solution f can be decoded into a dictator.

It is not particularly challenging to obtain (1). However, we sketch a systematic proof which shall be useful when analyzing a more complicated dictatorship test for PTFs.

Call a setting of $\mathcal{I}$ *good* if $\mathrm{sign}(f)$ passes the test conditioned on $\mathcal{I}$ with probability $1/2 + \xi$. By averaging, it is easy to see that $\mathrm{Pr}_{\mathcal{I}}[\mathcal{I} \text{ is good}] \geq \xi/2$. Let us fix such a good $\mathcal{I}$. Without loss of generality, we may assume that $\mathcal{I} = \{k^* + 1, \ldots, k\}$ and further that $k^* \geq k/2$ by the Chernoff bound. We now define $\{W_1, \ldots, W_{k^*}\}$ as a basis for $\{Y_i \mid i \in [k^*]\}$ where $W_1 := (1/k^*) \sum_{i \in [k^*]} Y_i$, such that $\{W_1, \ldots, W_{k^*}\}$ is an orthogonal transformation of $\{Y_i \mid i \in [k^*]\}$ of the same $1/\sqrt{k^*}$ norm. Thus, we may rewrite f as:

$$f = \sum_{i \in [k] \setminus [k^*]} \tilde{f}_i Y_i + \sum_{\ell \in [k^*]} \overline{f}_\ell W_\ell. \tag{2}$$

The variables in the first sum in the RHS of the above are all i.i.d. $N(0, 1)$. Further, it can be seen that under the test distribution, $W_1 = b\eta$, and $W_\ell = 0$ ($\ell = 2, \ldots, k^*$). Therefore, we may assume that,

$$\overline{f}_1^2 > 0. \tag{3}$$

Since the sign of f must flip with that of b with probability $\Omega(\xi) = \Omega(1)$, one can apply Carbery-Wright's Gaussian anti-concentration theorem to show that,

$$\sum_{i \in [k] \setminus [k^*]} \tilde{f}_i^2 \leq O(\eta^2) \overline{f}_1^2, \tag{4}$$

since otherwise, contributions from the first sum of (2) will overwhelm the contribution of W_1 to f. Further, from the definition of $\{W_\ell\}_{\ell=1}^{k^*}$, we obtain

$$\sum_{i \in [k^*]} \tilde{f}_{i1}^2 = \frac{1}{k^*} \sum_{\ell \in [k^*]} \overline{f}_\ell^2 \geq \overline{f}_1^2 / k^*. \tag{5}$$

Let us now revert to the notation with $\mathcal{I} = [k] \setminus [k^*]$. Using (5) along with (4), and taking $\eta = o(\varepsilon^3/\sqrt{k})$ one can ensure that,

$$\sum_{i \in \mathcal{I}} \tilde{f}_i^2 \leq \frac{\varepsilon}{10} \sum_{i \in [k]} \tilde{f}_i^2, \tag{6}$$

and from (3) we obtain

$$\sum_{i \in [k]} \tilde{f}_i^2 > 0. \tag{7}$$

Note that (6) holds for every good $\mathcal{I}$ which is at least $\xi/2$ fraction of the choices of $\mathcal{I}$. Randomizing over $\mathcal{I}$, an application of the Chernoff-Hoeffding bound shows that (6) holds only with substantially smaller probability unless there exists $i^* \in [k]$ such that:

$$\tilde{f}_{i^*}^2 \geq \frac{\varepsilon^3}{8} \sum_{i \in [k]} \tilde{f}_i^2. \tag{8}$$

4

The desired bound in (1) now easily follow from (7) and (8). The details are omitted.

The main idea of the above methodical analysis is a natural definition of the W variables using which we isolate the sign-perturbation $b\eta$ into a single variable W_1! Gaussian anti-concentration directly lower bounds the squared mass corresponding to W_1. Moreover, when transforming back to the squared mass of Y_i ($i \in [k] \setminus \mathcal{I}$), the presence of the heretofore ignored W_ℓ ($\ell > 1$) terms can only increase this quantity, as shown in (5). Lastly, the the "decoding list size" does not depend on the sign-perturbation parameter η which can be taken to be small enough to makes sure that this size is a constant depending only on the noise parameter ε and the marginal acceptance probability ξ of the test.

1.2.2 Enhancing the Dictatorship Test for degree-d PTFs

Our goal is a reduction proving the hardness of weakly learning noisy halfspaces using degree-d PTFs. One could hope to utilize the dictatorship test $\mathfrak{P}_0$ itself for this purpose. Unfortunately, this presents problems even for $d = 5$. To see this consider the degree-5 polynomial,

$$f(Y) = Y_{i^*}^3 \left(\sum_{i \in [k] \setminus \{i^*\}} Y_i^2 \right),$$

for some distinguished $i^* \in [k]$. It is easy to see that $\text{sign}(f)$ passes the test with probability close to 1. However, the distinguished variable Y_{i^*} appears with a cubic power in f, whereas the folding approach works well only when Y_{i^*} occurs as a linear factor of some sub-polynomial. This is due to the inherently linear nature of the folding constraints. Consequently, when $\mathfrak{P}_0$ is combined with a Label Cover instance the analysis becomes infeasible.

Our approach to overcome this bottleneck is for the PCP to test several independently and randomly chosen vertices. For this, the dictatorship test would be on the domain $\mathbb{R}^{[k] \times [T]}$ where T is chosen much larger than the degree d of the PTF to be tested. The space $\mathbb{R}^{[k] \times [T]}$ is thought of as real space spanned by T blocks of k dimensions each. In this case, if the test passes with probability $> 1/2$, then there is a way to decode a good label to at least one out of the T blocks. A key step in our analysis crucially leverages the choice of T to extract out a specific sub-polynomial which is linear in the variables of one of the T blocks. This is done via an application of the following lemma which is proved in Section 7.

Lemma 1.2. *Given a degree-d polynomial of the form $(Y_1 + \cdots + Y_T) \cdot S(Y_1, \ldots, Y_T)$, where $T > 2d$ and S is a degree-$(d - 1)$ polynomial, there exist at least $T/2$ indices $j \in [T]$ such that: for each such j, the sum of squares of the coefficients corresponding to the terms (in the monomial representation) linear in Y_j is at least c times the sum of squares of coefficients of S, where $c := c(T, d) > 0$.*

In Figure 2, we give a formal description of the Dictatorship test $\mathfrak{P}_1$ employed by our reduction. Its analysis builds upon that of $\mathfrak{P}_0$ above, so we provide a short sketch. Let $T = 10d$ and $\varepsilon > 0$ be a constant, and $\eta > 0$ be parameter to be defined later. Consider the linear threshold given by,

$$\text{sign} \left(\sum_{j=1}^{T} Y_{i_j j} \right),$$

for any $i_j \in [k]$ ($1 \leq j \leq T$). It is easy to see that this passes the test with probability at least $(1 - \varepsilon T)$. Thus, choosing a dictator for each block yields a good solution for the test.

5

> **PCP Test $\mathfrak{P}_1(\mathbb{R}^{[k]\times[T]}, \eta, \varepsilon)$ tests degree-d PTF sign$(P(Y))$**
>
> 1. Sample $\{\delta_j \mid j \in [T]\}$ from the joint Gaussian distribution where the marginals are $N(0,1)$, $\mathbb{E}[\delta_j \delta_{j'}] = -1/(T-1)$ for all $j \neq j'$, and $\sum_{j=1}^{T} \delta_j = 0$.
> 2. Sample $b \in \{-1, 1\}$ uniformly at random.
> 3. Sample $\mathcal{I} \subseteq [k] \times [T]$ to be a random subset where each $(i,j) \in [k] \times [T]$ is added to $\mathcal{I}$ independently with probability ε.
> 4. For each $(i,j) \in ([k] \times [T]) \setminus \mathcal{I}$, set $y_{ij} = (\sqrt{(T-1)/T})\delta_j + b\eta$.
> 5. Independently for each $(i,j) \in \mathcal{I}$, sample $y_{ij} \sim N(0,1)$.
> 6. Accept iff sign$(P(y)) = b$.

Figure 2: Dictatorship Test $\mathfrak{P}_1$

For the soundness analysis, as in Section 1.2.1 we fix a good noise set $\mathcal{I}$ conditioned on which the test accepts P with probability at least $1/2 + \xi$, and $\Pr[\mathcal{I} \text{ is good }] \geq \xi/2$. Further, without loss of generality, we assume that $\mathcal{I} = \cup_{j=1}^{T}(\{k_j + 1, \ldots k\} \times \{j\})$, where (by Chernoff bound) $k_j \geq k/2$ for $1 \leq j \leq T$. For each j, $\{W_{1j}, \ldots, W_{k_j j}\}$ is defined to be an orthogonal transformation of $\{Y_{1j}, \ldots, Y_{k_j j}\}$ of the same $1/\sqrt{k_j}$ norm, where $W_{1j} = (1/k_j) \sum_{i=1}^{k_j} Y_{ij}$. It is easy to see that $W_{1j} = (\sqrt{(T-1)/T})\delta_j + b\eta$, while $W_{\ell j} = 0$ under the test distribution for $\ell > 1$.

Additionally, we also define $\{U_1, \ldots, U_T\}$ to be an orthonormal transformation of $\{W_{11}, \ldots, W_{1T}\}$ where $U_1 = (1/\sqrt{T}) \sum_{j=1}^{T} W_{1j}$. Again, it can observed that $U_1 = (\sqrt{T})b\eta$ and $U_2, \ldots, U_T$ are independent $N(0,1)$. Using this we write the polynomial $P = P' + Q_0 + U_1 Q_1$, where P' consists of all the terms which have any $W_{\ell j}, \ell > 1$ as a factor. Further, Q_0 is independent of U_1. Since $P' = 0$ under the distribution we ignore it for now, noting that $\|Q_1\|_2^2 = \mathbb{E}[Q_1^2] > 0$, since the test accepts with probability $> 1/2$. The first step is to show, via Gaussian anti-concentration on Q_0 and Chebyshev's inequality on Q_1, that

$$\|Q_0\|_2^2 \leq O(\eta^2)\|Q_1\|_2^2. \tag{9}$$

Let us write $Q_1 = \sum_{H \in \mathscr{H}} H \cdot Q_{1,H}(U_1, \ldots, U_T)$, where the sum is over the set $\mathscr{H}$ of normalized Hermite monomials[3] over the independent $N(0,1)$ variables $\cup_{j=1}^{T}\{Y_{ij}\}_{i=k_j+1}^{k}$. Moreover, let $Q_1^{(D)} = \sum_{H \in \mathscr{H}_D} H \cdot Q_{1,H}(U_1, \ldots, U_T)$ for $0 \leq D \leq d-1 \geq \deg(Q_1)$, where $\mathscr{H}_D$ is the subset of $\mathscr{H}$ of degree exactly D. Thus, $\|Q_1\|_2^2 = \sum_{H \in \mathscr{H}} \|Q_{1,H}\|_2^2$. Writing $Q_{1,H} = Q_{1,H}(W_{11}, \ldots, W_{1T})$ we also define $\|Q_{1,H}\|_{\mathrm{mon}}^2$ as sum of squares of the coefficients in the standard monomial basis $\mathscr{M}$ of $\{W_{11}, \ldots, W_{1T}\}$. A straightforward calculation shows that:

$$\|Q_{1,H}\|_2^2 \leq O(1)\|Q_{1,H}\|_{\mathrm{mon}}^2, \tag{10}$$

where the constants depending on T and d are absorbed in the $O(1)$ notation. On the other hand, since Q_0 is independent of U_1, using similar definition of $Q_{0,H}$, we can establish the reverse bound for it:

$$\|Q_{0,H}\|_{\mathrm{mon}}^2 \leq O(1)\|Q_{0,H}\|_2^2. \tag{11}$$

The rest of the arguments significantly build upon those in Section 1.2.1. We present a semi-formal description, omitting much of the technical details. For reasons made clear later, we first carefully select $d^* \in \{0, \ldots, d-1\}$ to be the largest $D \in \{0, \ldots, d-1\}$ such that $\|Q_1^{(D)}\|_2^2 \geq \frac{1}{4}\rho^D \|Q_1\|_2^2$ for a small

---

[3]By Hermite *monomials*, we mean elements of the polynomial Hermite basis over the corresponding variables.

enough constant depending on k, T, d, and ε. It is easily observed that such a d^* must exist satisfying the properties: (i) $\|Q_1^{(d^*+1)}\|_2^2 \leq \frac{1}{4}\rho^{d^*+1}\|Q_1\|_2^2$, and (ii) $\|Q_1^{(d^*)}\|_2^2 \geq \frac{1}{4}\rho^{d^*}\|Q_1\|_2^2$.

Now we focus our attention on $U_1 Q_1^{(d^*)}$ writing it as

$$U_1 Q_1^{(d^*)} = \sum_{H \in \mathscr{H}_{d^*}} H U_1 Q_{1,H}(W_{11}, \ldots, W_{1T}) = \sum_{H \in \mathscr{H}_{d^*}} \sum_{M \in \mathscr{M}} c_{H,M} H M. \tag{12}$$

Let $\mathscr{H}_{-j^*D} \subseteq \mathscr{H}_D$ (resp. $\mathscr{M}_{-j^*} \subseteq \mathscr{M}$) be the subset of basis elements not containing any variable from the j^*th block, i.e. $\{Y_{ij^*}\}_{k_{j^*} < i \leq k}$ (resp. W_{1j^*}). Now with $U_1 = (1/\sqrt{T}) \sum_{j=1} W_{1j}$, we apply Lemma 1.2 to each $U_1 Q_{1,H}(W_{11}, \ldots, W_{1T})$ in the first expansion of (12). Using the fact that each H has at most d variables along with our choice of $T = 10d$ yields a $j^* \in [T]$ such that

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c_{H,MW_{1j^*}}^2 \geq \Omega(1)\left(\sum_{H \in \mathscr{H}_{d^*}} \sum_{M \in \mathscr{M}} c_{H,M}^2 \right) \tag{13}$$

$$\geq \Omega(1)\|Q_1^{(d^*)}\|_2^2 \geq \Omega(1)\rho^{d^*}\|Q_1\|_2^2 \tag{14}$$

where the last two inequalities use (10) along with property (ii) above.

The next component of the analysis is to relate the bounds above with the coefficients of a suitable sub-polynomial of P which is linear in the variables $Y_{ij^*}, 1 \leq i \leq k_{j^*}$. For this, let us first define $\tilde{Q}$ to be exactly the sub-polynomial of P which does not contain any term with W_{ij} where $i \neq 1$ and $j \neq j^*$. Rewriting the variables $\{W_{ij^*} \mid i \in [k_{j^*}]\}$ in terms of $\{Y_{ij^*} \mid i \in [k_{j^*}]\}$, we consider the sub-polynomial $\tilde{Q}_{\text{lin}}$ (of $\tilde{Q}$) which is linear in the variables $\{Y_{ij^*} \mid 1 \leq i \leq k\}$. Note that $\left(\cup_{D=0}^{d-1} \mathscr{H}_{-j^*D}\right) \circ \mathscr{M}_{-j^*} \circ \{Y_{ij^*}\}_{i=1}^k$ is a basis in which $\tilde{Q}_{\text{lin}}$ can be written with coefficients $\tilde{c}_{H,M,i}$ corresponding to the basis element HMY_{ij^*}. Using the orthonormal transformation between $\{W_{ij^*}\}_{i \in [k_{j^*}]}$ and $\{Y_{ij^*}\}_{i=1}^{k_{j^*}}$ we obtain

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k_{j^*}]} \tilde{c}_{H,M,i}^2 \geq \frac{1}{2k_{j^*}} \left(\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c_{H,MW_{1j^*}}^2 \right), \tag{15}$$

neglecting any contribution to the LHS of the above from Q_0 by our a small enough choice of $\eta \ll \rho$ along with (9) and (11). The loss of k_{j^*} factor in (15) is compensated by the dependence of ρ on k as we shall see later. Combining (15) with (13)-(14) yields

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k_{j^*}]} \tilde{c}_{H,M,i}^2 \geq \Omega\left(1/k_{j^*}\right)\rho^{d^*}\|Q_1\|_2^2. \tag{16}$$

Consider now the sum

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{k_{j^*} < i \leq k} \tilde{c}_{H,M,i}^2.$$

Contribution to the above can be from Q_0 or from $U_1 Q_1^{(d^*+1)}$ – the latter due to the presence of Y_{ij^*} ($k_{j^*} < i \leq k$) which increases the degree of $H \in \mathscr{H}_{-j^*d^*}$ to (d^*+1) in the representation of Q_1 over the basis $\mathscr{H} \circ \mathscr{M}$. Property (i) from our careful selection of d^* is leveraged along with our small enough choice of η in (9) along with (11) to yield

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{k_{j^*} < i \leq k} \tilde{c}_{H,M,i}^2 \leq O(1)\rho^{d^*+1}\|Q_1\|_2^2. \tag{17}$$

Using a choice $\rho \ll \varepsilon/k$ we can combine the above with (16) to obtain the following analog of (6):

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in \mathcal{I}_{j^*}} \tilde{c}_{H,M,i}^2 \leq \frac{\varepsilon}{10} \sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k]} \tilde{c}_{H,M,i}^2, \tag{18}$$

where $\mathcal{I}_{j^*} := \mathcal{I} \cap ([k] \times \{j^*\})$. Of course, since $\|Q_1\|_2 > 0$, we also obtain

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k]} \tilde{c}_{H,M,i}^2 > 0. \tag{19}$$

The analysis above shows that for every good choice of $\mathcal{I}$ there exist (d^*, j^*) satisfying (18)-(19). What remains is a probabilistic concentration argument. Since $\Pr[\mathcal{I} \text{ is good}] \geq \xi/2$, by averaging we get that there exist (d^*, j^*) and a fixing of $\mathcal{I} \setminus \mathcal{I}_{j^*}$ such that with probability at least $\xi/4Td$ over the choice $\mathcal{I}_{j^*}$, (18)-(19) hold. Since each i is added to $\mathcal{I}_{j^*}$ independently with probability ε, an application of Chernoff-Hoeffding shows that the large deviation observed in (18) cannot occur with probability $\xi/4Td$ (which is significant) unless the squared mass on the LHS of (19) is concentrated on a small number of $i \in [k]$. This yields the desired decoding completing our sketch of the analysis. The formal proof appearing in this work – while following the approach given above – employs additional notation and definitions for handling a few technicalities and ease of presentation.

Combining $\mathfrak{P}_1$ with Label Cover and Folding. The test $\mathfrak{P}_1$ is executed on the T blocks of coordinates corresponding to T randomly chosen vertices of a Smooth Label Cover instance (as used in [GRSW16]). The resulting instance is then folded, i.e. the distribution on the point-label pairs is projected onto a subspace $\mathcal{F}$ orthogonal to the span of all the linear constraints implied by the edges of the Label Cover. These linear constraints ensure that any vector in $\mathcal{F}$ has equal mass sum in the coordinates of the two pre-images of a label given by an edge's projections. This property can be extended to polynomials P residing in $\mathcal{F}$. This fits with our decoding of $\mathfrak{P}_1$ which is via a sub-polynomial $\tilde{Q}_{\text{lin}}$ linear in the variables $\{Y_{ij^*}\}_{i=1}^k$ of the j^*th block. More specifically, we may fix the vertices corresponding to all the blocks except the j^*th and also the restriction of $\mathcal{I}$ to all the blocks except the j^*th. This fixes $\mathscr{H}_{-j^*d^*} \circ M_{-j^*}$ used in (18)-(19). For a vertex v let $\tilde{c}_{H,M,i,v} = \tilde{c}_{H,M,i}$ when v is chosen as the j^*th vertex. Suppose for an edge between u and v (not among the fixed vertices) the respective pre-images of a common label are A and B. Then, the folding constraints imply

$$\sum_{i \in A} \tilde{c}_{H,M,i,u} = \sum_{i \in B} \tilde{c}_{H,M,i,v}. \tag{20}$$

We combine the above with the decoding obtained from the analysis of $\mathfrak{P}_1$ using appropriately set smoothness parameters to prevent masses in the pre-images containing the decoded coordinates from cancelling out. The constraints (20) then imply that the decoded labels define a labeling satisfying a significant fraction of edges of the Label Cover instance.

Organization. Section 2 presents some preliminaries. Section 3 describes the reduction from Label Cover in the form of a PCP test. Section 3.1.1 gives the constraints implied by folding extended to polynomials. In Section 4, we show the soundness of the reduction assuming a lemma (essentially restating (18)-(19)) about the structure of polynomials passing the test. The rest of the paper is devoted to proving this lemma. In Section 5, we apply Gaussian anti-concentration to prove the analog of (9). In Section 6, we prove the structural lemma using Lemma 1.2 as a key ingredient. Lemma 1.2 is proved in Section 7.

2 Preliminaries

2.1 The SMOOTH LABEL COVER Problem

Definition 2.1 (Smooth Label Cover)**.** *A* SMOOTH LABEL COVER *instance $\mathcal{L}(G(V,E), k, L, \{\pi_{e,v}\}_{e \in E, v \in e})$ consists of a regular connected graph with vertex set V and edge set E, along with projection maps $\pi_{e,v} : [k] \to [L]$ for all $e \in E, v \in e$. The goal is to find an assignment $\sigma : V \to [k]$ such that $\forall e = (u, w) \in E, \pi_{e,u}(\sigma(u)) = \pi_{e,w}(\sigma(w))$. The optimum for a* SMOOTH LABEL COVER *instance is the maximum fraction of edges satisfied by an assignment.*

The following Theorem from [GRSW16] states the hardness of SMOOTH LABEL COVER problem:

Theorem 2.2. *There exists a constant $c_0 > 0$ such that for any constant integer parameters $J, R \geq 1$, it is* NP-*hard to distinguish between the following cases for a* SMOOTH LABEL COVER *instance $\mathcal{L}(G(V,E), k, L, \{\pi_{e,v}\}_{e \in E, v \in e})$ with parameters $k = 7^{(J+1)R}, L = 2^R 7^{JR}$.*

- **YES***: There is a labeling that satisfies every edge.*

- **NO***: Every labeling satisfies less than $2^{-c_0 R}$-fraction of edges.*

Additionally, the instance $\mathcal{L}$ satisfies the following properties:

- **Smoothness***: For any $v \in V$, and labels $i, j \in [k], i \neq j$, $\mathrm{Pr}_{e \sim v}[\pi_{e,v}(i) = \pi_{e,v}(j)] \leq 1/J$. In particular, for a subset $S \subseteq [k]$, $\mathrm{Pr}_{e \sim v}[|\pi_{e,v}(S)| = |S|] \leq |S|^2/(2J)$.*

- *The degree $d_{\mathcal{L}}$ of the graph G is a constant dependent only on J and R.*

- *For any vertex $v \in V$, edge $e \in E$ incident on vertex v, and $j \in [L]$, we have $\left|(\pi_{e,v})^{-1}(j)\right| \leq t_{\mathcal{L}} := 4^R$.*

- **Weak Expansion***: For any $V' \subseteq V$, the number of edges induced in V' is at least $\frac{\delta^2}{2}|E|$ where $\delta = |V'|/|V|$.*

2.2 Hermite Bases for Multivariate Polynomials

For integer $d \geq 0$, the *Hermite polynomials $H_d(x)$* are degree-d univariate polynomials such that $\mathbb{E}_{X \sim N(0,1)}[H_d(X)^2] = 1$ and $\mathbb{E}_{X \sim N(0,1)}[H_d(X)H_{d'}(X)] = 0$ for any $d \neq d'$. For example, $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1)$, and $H_3(x) = \frac{1}{\sqrt{6}}(x^3 - x)$.

For $\mathbf{d} \in \mathbb{N}^n$, we define $H_{\mathbf{d}}(x_1, \ldots, x_n) = \prod_{i \in [n]} H_{d_i}(x_i)$. For $D \geq 0$, let $\mathcal{H}_D = \{H_{\mathbf{d}} : \mathbf{d} \in \mathbb{N}^n, \sum_{i \in [n]} d_i \leq D\}$ denote the *Hermite basis for degree-D polynomials*. The following is immediate.

Fact 2.3. *The set $\mathcal{H}_D$ forms an orthonormal basis for n-variate degree-D polynomials whose inputs are drawn from $N(0,1)^n$. In particular, for any $P : \mathbb{R}^n \to \mathbb{R}$ of degree $\leq D$, we can write:*

$$P(x) = \sum_{\mathbf{d} \in \mathbb{N}^n : \sum_i d_i \leq D} \hat{f}(\mathbf{d}) \cdot H_{\mathbf{d}}(x)$$

and moreover, $\mathbb{E}_x P(x) = \hat{f}(\mathbf{0})$ and $\mathbb{E}_x[P(x)^2] = \sum_{\mathbf{d}} \hat{f}^2(\mathbf{d})$.

The Basic PCP Test given instance $\mathcal{L}$ of SMOOTH LABEL COVER

1. For each $j \in [T]$, the test chooses T random vertices $v_1, v_2, \ldots, v_T \overset{u.a.r.}{\sim} V$. Let $Y_{ij} := Y_i^{v_j}$.
2. Sample $\{\delta_j \mid j \in [T]\}$ from the joint Gaussian distribution where the marginals are $N(0,1)$, $\mathbb{E}[\delta_j \delta_{j'}] = -1/(T-1)$ for all $j \neq j'$, and $\sum_{j=1}^{T} \delta_j = 0$.
3. Sample $b \in \{-1, 1\}$ uniformly at random.
4. Sample $\mathcal{I} \subseteq [k] \times [T]$ to be a random subset where each $(i,j) \in [k] \times [T]$ is added to $\mathcal{I}$ independently with probability ε.
5. For each $(i,j) \in ([k] \times [T]) \setminus \mathcal{I}$, set $Y_{ij} := \sqrt{(T-1)/T} \cdot \delta_j + b\eta$.
6. Independently for each $(i,j) \in \mathcal{I}$, sample Y_{ij} from $N(0,1)$.
7. Set the variables of all other vertices (except $\{v_j \mid j \in [T]\}$) to be 0. Let this setting of the variables be the point $\mathbf{y} \in \mathbb{R}^{\mathcal{Y}}$.
8. Output the point-sign pair $(\mathbf{y}, b)$.

Figure 3: Basic PCP Test

2.3 Concentration and Anti-Concentration

The magnitude of polynomials in our analysis is controlled using the following standard bound.

Chebyshev's Inequality. For any random variable X and $t > 0$, $\Pr\left[|X| > t\right] \leq \mathbb{E}[X^2]/t^2$.

The above is used in conjunction with Carbery and Wright's [CW01] powerful anti-concentration bound for polynomials over independent Gaussian variables.

Theorem 2.4. (Carbery-Wright [CW01]) *Suppose $P : \mathbb{R}^\ell \to \mathbb{R}$ is a degree-d polynomial over independent $N(0,1)$ random variables. Then,*

$$\Pr\left[|P| \leq \varepsilon \|P\|_2\right] = O(d\varepsilon^{1/d}).$$

In addition, we also use following Chernoff-Hoeffding bound.

Theorem 2.5 (Chernoff-Hoeffding)**.** *Let $X_1, \ldots, X_n$ be independent random variables, each bounded as $a_i \leq X_i \leq b_i$ with $\Delta_i = b_i - a_i$ for $i = 1, \ldots, n$. Then, for any $t > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| > t\right] \leq 2 \cdot \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} \Delta_i^2}\right).$$

3 Hardness Reduction

The following reduction from SMOOTH LABEL COVER directly implies our main theorem.

Theorem 3.1. *For any $\xi > 0$ and $d \in \mathbb{Z}^+$, there exists a choice of R and J in Theorem 2.2 and a polynomial-time reduction from the corresponding SMOOTH LABEL COVER instance $\mathcal{L}$ to a set of point-sign pairs $\mathcal{Q} \subseteq \mathbb{R}^N \times \{-1, 1\}$ such that:*

- **YES Case.** *If $\mathcal{L}$ is a YES instance, then there exists a linear form L satisfying*

$$\Pr_{(\mathbf{x}, s) \in \mathcal{Q}} \left[\text{sign}\left(L(\mathbf{x})\right) = s\right] \geq 1 - \xi.$$

- *NO Case. If $\mathcal{L}$ is a NO instance, then for any degree-d polynomial P*

$$\Pr_{(\mathbf{x},s)\in\mathcal{Q}}\left[\text{sign}\left(P(\mathbf{x})\right)=s\right]\leq\frac{1}{2}+\xi.$$

The last sentence of Theorem 1.1 is justified in Section 4.6.

3.1 The Basic PCP Test

We begin with a Basic PCP Test given an instance $\mathcal{L}(G(V,E),k,L,\{\pi_{e,v}\}_{e\in E,v\in e})$ of SMOOTH LABEL COVER. For each vertex $v\in V$, there is a set of variables $\{Y_i^v\}_{i=1}^k$, and the set of all the variables $\mathcal{Y}$ is a union over all vertices $v\in V$ of these variable sets. The test is described by the sampling procedure in Figure 3, and yields a distribution over point-sign pairs which is independent of the constraints in $\mathcal{L}$. It uses some additional parameters set as follows: $T:=10d$, $\varepsilon:=(\xi/32Td)$, $\eta:=\left(\frac{\varepsilon\xi}{20kdT}\right)^{d6^{3d}}$, where d is from the statement of Theorem 3.1.

3.1.1 Folding over constraints of $\mathcal{L}$

To ensure consistency across the edges of $\mathcal{L}$, the points generated by the Basic PCP Test are *folded* over a specific subspace. The points generated by the Basic PCP Test reside in the space $\mathbb{R}^{\mathcal{Y}}$. Now, for a fixed $e=(u,w)\in E$ and $j\in[L]$, we define the vector $\mathbf{h}_j^e\in\mathbb{R}^{\mathcal{Y}}$ as

$$\mathbf{h}_j^e(Y_i^v)=\begin{cases}1 & \text{if }v=u\text{ and }i\in(\pi_{e,u})^{-1}(j),\\-1 & \text{if }v=w\text{ and }i\in(\pi_{e,w})^{-1}(j),\\0 & \text{otherwise.}\end{cases}\tag{21}$$

Let $\mathcal{H}\subseteq\mathbb{R}^{\mathcal{Y}}$ be the subspace formed by the linear span of the vectors $\{\mathbf{h}_j^e\}_{e\in E,j\in[L]}$, and let $\mathcal{F}$ be the orthogonal complement of $\mathcal{H}$ in $\mathbb{R}^{\mathcal{Y}}$, i.e. $\mathbb{R}^{\mathcal{Y}}=\mathcal{H}\oplus\mathcal{F}$ and $\mathcal{H}\perp\mathcal{F}$. For each point-sign pair $(\mathbf{y},b)$ generated by the Basic PCP Test, construct $(\overline{\mathbf{y}},b)$ where $\overline{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the subspace $\mathcal{F}$, represented in some (fixed) orthogonal basis for $\mathcal{F}$.

Conversely, for any vector $\overline{\mathbf{z}}\in\mathcal{F}$, let $\mathbf{z}$ be its representation in $\mathbb{R}^{\mathcal{Y}}$. It is easy to see that such a $\mathbf{z}$ satisfies: for every $e=(u,w)\in E$ and $j\in[L]$, $\langle\mathbf{z},\mathbf{h}_j^e\rangle=0$ which is equivalent to

$$\text{Constraint }\ \mathcal{C}_{e,j}:\qquad\sum_{i\in(\pi_{e,u})^{-1}(j)}\mathbf{z}(Y_i^u)=\sum_{i\in(\pi_{e,w})^{-1}(j)}\mathbf{z}(Y_i^w).\tag{22}$$

For our purpose we shall extend the above constraint to polynomials as well. Consider a polynomial Q in $\mathbb{R}^{\mathcal{Y}}$. For any monomial M over the variables $\mathcal{Y}$, let $c_{Q,M}$ be its coefficient in Q. Fix an edge $e=(u,w)$ and $j\in[L]$, and a monomial M such that M does not contain any variable from the set $\{Y_i^u\mid i\in(\pi_{e,u})^{-1}(j)\}\cup\{Y_i^w\mid i\in(\pi_{e,w})^{-1}(j)\}$. For such a choice of e,j, and M we say that $\mathcal{C}_{e,j,M}$ is a *valid* constraint where:

$$\text{Constraint }\ \mathcal{C}_{e,j,M}:\qquad\sum_{i\in(\pi_{e,u})^{-1}(j)}c_{Q,M\cdot Y_i^u}=\sum_{i\in(\pi_{e,w})^{-1}(j)}c_{Q,M\cdot Y_i^w}.\tag{23}$$

We have the following lemma.

11

Lemma 3.2. *Let $\overline{Q}$ be a polynomial that resides in $\mathcal{F}$, i.e. is represented in an orthogonal basis[4] for $\mathcal{F}$, and let Q be its representation in $\mathbb{R}^{\mathcal{Y}}$. Then, Q satisfies all valid constraints $\mathcal{C}_{e,j,M}$.*

Proof. Suppose for a contradiction Q does not satisfy a valid constraint $\mathcal{C}_{e,j,M}$. Consider the vector $\mathbf{r}$ where,

$$\mathbf{r}(Y_i^v) = \begin{cases} c_{Q,M \cdot Y_i^u} & \text{if } v = u, i \in (\pi_{e,u})^{-1}(j) \\ c_{Q,M \cdot Y_i^w} & \text{if } v = w, i \in (\pi_{e,w})^{-1}(j) \\ 0 & \text{otherwise.} \end{cases}$$

Since Equation (23) is not satisfied, it is easy to see that $\langle \mathbf{r}, \mathbf{h}_j^e \rangle \neq 0$, and thus $\mathbf{r} = \mathbf{r}_0 + \mathbf{r}_1$ where $\mathbf{r}_0 \in \mathcal{F}$ and $\mathbf{r}_1 \in \mathcal{H}$. On the other hand, consider an orthogonal basis $\mathcal{B}$ for $\mathbb{R}^{\mathcal{Y}}$ that is an extension of $\{\mathbf{r}_1\}$, i.e. $\mathbf{r}_1$ is an element of $\mathcal{B}$. P can now be represented as:

$$P \equiv \mathbf{r}_1[\mathcal{Y}] \cdot P_1 + P_0,$$

where P_1 is a polynomial represented in $\mathcal{B}$, P_0 is represented in $\mathcal{B} \setminus \{\mathbf{r}_1\}$, and $\mathbf{r}_1[\mathcal{Y}]$ is the $\mathcal{Y}$-linear form $\sum_{Y \in \mathcal{Y}} \mathbf{r}_1(Y) \cdot Y$. Note that P_1 is not identically zero, in particular it contains the monomial M. This implies that P cannot be represented over any basis for $\mathcal{F}$, which is a contradiction. $\square$

Remark 3.3. *Instead of monomials M, the constraints in (23) analogously hold for elements B of a basis $\mathcal{B}$ for polynomials over any set of variables not containing $\{Y_i^u \mid i \in (\pi_{e,u})^{-1}(j)\} \cup \{Y_i^w \mid i \in (\pi_{e,w})^{-1}(j)\}$.*

3.2 The Final PCP Test

Given a degree-d polynomial $\overline{P}_{\text{global}}$ over the space $\mathcal{F}$, the test samples $(\mathbf{y}, b)$ from the Basic PCP Test (as described in Figure 3), and constructs $(\overline{\mathbf{y}}, b)$ as described in Section 3.1.1. The test accepts *iff* $\text{sign}\left(\overline{P}_{\text{global}}(\overline{\mathbf{y}})\right) = b$.

Remark 3.4. *The Basic PCP Test generates a distribution over $\mathbb{R}^{\mathcal{Y}} \times \{-1, 1\}$ using various independently Gaussian random variables. Therefore, the support set of this distribution is not finite. In Section 4.6, using techniques from [DOSW11], we discretize the Basic PCP Test. Building upon the discretized Basic PCP Test, the Final PCP Test yields the desired finite subset $\mathcal{Q}$ in polynomial time.*

3.3 Completeness Analysis

Suppose there is a labeling $\sigma : V \rightarrow [k]$ which satisfies all the edges of $\mathcal{L}$. Define $L^*(\mathcal{Y}) = \sum_{v \in V} Y_{\sigma(v)}^v$ to be a linear form. Note that $L^*(\mathbf{y}) := \langle \mathbf{r}^*, \mathbf{y} \rangle$ for some $\mathbf{r}^* \in \mathcal{F}$, and so L^* can be represented in an orthogonal basis for $\mathcal{F}$. Thus, for any point $\mathbf{y} \in \mathbb{R}^{\mathcal{Y}}$, $L^*(y) = L^*(\overline{\mathbf{y}})$ where $\overline{\mathbf{y}}$ is the projection of $\mathbf{y}$ on to $\mathcal{F}$ as defined in Section 3.1.1.

Now consider $(\mathbf{y}, b)$ generated by the Basic PCP Test. By a union bound over the randomness of the test, with probability at least $(1 - \varepsilon T)$: $(\sigma(v_j), j) \notin \mathcal{I}$ for each $j \in [T]$. Given this, it is easy to see that $L^*(\mathbf{y}) = b$, and by the above reasoning $L^*(\overline{\mathbf{y}}) = b$. Thus, L^* satisfies the Final PCP Test with probability at least $(1 - \varepsilon T)$. Our choice of ε yields the desired accuracy.

---

[4]A polynomial $\overline{Q}$ being represented in an orthogonal basis for a subspace $\mathcal{F}$ means $\overline{Q}$ can be written as a polynomial over the linear forms corresponding to an orthogonal basis for $\mathcal{F}$.

4 Soundness Analysis

Given the SMOOTH LABEL COVER instance $\mathcal{L}$, suppose that there is a degree-d polynomial (over $\mathcal{F}$) $\overline{P}_{\text{global}}$ such that the Final PCP Test accepts with probability $1/2+\xi$. Our goal in the rest of this paper is to show that in this case there exists a labeling that satisfies at least $2^{-c_0 R}$-fraction of the edges of $\mathcal{L}$, for an appropriate choice of constants R and J in Theorem 2.2 and because of its NO Case we would be done.

Let P_{global} be the representation of $\overline{P}_{\text{global}}$ in $\mathbb{R}^{\mathcal{Y}}$, so that $\overline{P}_{\text{global}}(\overline{\mathbf{y}}) = P_{\text{global}}(\mathbf{y})$ where $\overline{\mathbf{y}} \in \mathcal{F}$ is a point generated by the Final PCP Test from a point $\mathbf{y}$ generated by the Basic PCP Test as given in Section 3.2. Therefore, $P_{\text{global}}(\mathbf{y}) = b$ with probability at least $1/2 + \xi$ over the pairs $(\mathbf{y}, b)$ output by the Basic PCP Test. Using this, we focus on analyzing the structure of P_{global}.

To begin the analysis note that with probability at least 2ξ over the choices of the verifier other than b, P_{global} flips its sign on flipping b. Call a choice of $\{v_j \mid j \in [T]\}$ *good* if conditioned on this, the same holds with probability at least ξ over the rest of the choices (other than b) of the verifier. By averaging, with probability at least ξ, the verifier makes a good choice. We now fix such a good choice $\{v_j \mid j \in [T]\}$.

For convenience, we shall use P to denote the restriction of P_{global} to $\mathbf{Y} := \{Y_{ij} \mid i \in [k], j \in [T]\}$. Let $\mathcal{D}$ be the distribution on $(\mathbf{Y}, b)$ generated by the steps of the verifier. Our analysis shall first show that in terms of this basis P must have a certain structure which will then be used to determine a good labeling for $\mathcal{L}$.

4.1 Basis Transformations

For the purpose of the analysis, we shall rewrite the variables $\mathbf{Y}$ in different bases. Before we do that, we shall isolate the noisy set $\mathcal{I}$ of the Basic PCP Test.

4.1.1 Choice of set $\mathcal{I}$

The distribution $\mathcal{D}$ involves choosing the set $\mathcal{I}$ in which each (i, j) is added independently at random with probability ε. Let us call a setting of $\mathcal{I}$ as *nice* if it satisfies:

1. For each j, $|\{i \mid (i, j) \in \mathcal{I}\}| \leq k/2$.

2. With probability $\xi/2$ over the rest of the choices of the verifier (except b), P flips its sign on flipping b.

By our setting of ε and T, for a large enough value of k, and applying the Chernoff Bound, a union bound and an averaging argument, we have:

$$\Pr_{\mathcal{D}}[\mathcal{I} \text{ is nice}] \geq \xi/4. \tag{24}$$

Going forward, we shall fix a nice choice of $\mathcal{I}$. By relabeling, we may assume that there exist $k/2 \leq k_j \leq k$ for $j \in [T]$ such that

$$\mathcal{I} = \bigcup_{j=1}^{T} \{(i, j) \mid i = k_j + 1, \ldots, k\}. \tag{25}$$

Based on this nice choice of $\mathcal{I}$, we now define new bases for the $\mathbf{Y}$ variables. Let $\mathcal{D}_{\mathcal{I}}$ denote the distribution of the variables after fixing a nice $\mathcal{I}$.

4.1.2 Bases W and U

For each $j \in [T]$, we define $(W_{1j}, W_{2j}, \ldots, W_{k_j j})$ as a fixed orthogonal transformation of $(Y_{1j}, Y_{2j}, \ldots, Y_{k_j j})$ so that

$$W_{1j} = \frac{1}{k_j} \sum_{i=1}^{k_j} Y_{ij}, \quad \text{and} \quad W_{ij} = \sum_{\ell \in [k_j]} c_{i\ell} Y_{\ell j} \text{ for all } i \in [2, k_j] \tag{26}$$

where the vectors $\left\{ \mathbf{c}_i = [c_{i1}, c_{i2}, \ldots, c_{ik_j}]^\mathsf{T} \right\}_{i=2}^{k_j}$ satisfy

- For all $i, i' \in [k_j] \setminus \{1\}$ we have $\langle \mathbf{c}_i, \mathbf{c}_{i'} \rangle = 0$
- Each vector $\mathbf{c}_i$ satisfies $\|\mathbf{c}_i\|^2 = 1/k_j$ and $\mathbf{c}_i \perp \quad$ where $\quad$ is the all ones vector in $\mathbb{R}^{k_j}$.

We shall also define the vector $\mathbf{c}_1 = \frac{1}{k_j} \cdot \quad$ where $\quad \in \mathbb{R}^{k_j}$ is the the vector of all ones. The above along with the distribution of $\{Y_{ij} \mid i = 1, \ldots, k_j\}_{j=1}^T$ in $\mathcal{D}_\mathcal{I}$ directly implies the following.

Lemma 4.1. *Under the distribution $\mathcal{D}_\mathcal{I}$:*

(i) $W_{1j} = \sqrt{(T-1)/T} \cdot \delta_j + b\eta$
(ii) *For $i \neq 1$, $W_{ij} = 0$.*

Let $U_1, \ldots, U_T$ be a fixed orthonormal transformation of $(W_{11}, \ldots, W_{1T})$, where

$$U_1 = \frac{1}{\sqrt{T}} \sum_{j=1}^T W_{1j}, \quad \text{and } U_t = \sum_{j \in [T]} a_{tj} W_{1j} \text{ for all } t \in [2, T] \tag{27}$$

where vectors $\mathbf{a}_2, \ldots, \mathbf{a}_T$ are orthonormal and each vector $\mathbf{a}_t = [a_{t1}, a_{t2}, \ldots, a_{tT}]^\mathsf{T}$ satisfies $\sum_{j \in T} a_{tj} = 0$ (i.e., they are orthogonal to the all ones vector).

Lemma 4.2. *Under the distribution $\mathcal{D}_\mathcal{I}$,*

(i) $U_1 = b\eta\sqrt{T}$
(ii) *For each $1 < t \leq T$, $U_t \sim N(0, 1)$ i.i.d.*

Proof. Lemma 4.1 along with the definition of U_1 yields the first part. The second part follows from an application of Lemma A.2. $\qquad\square$

Before we proceed, we briefly summarize the variables and their distribution under $\mathcal{D}_\mathcal{I}$.

- **Noisy Indices** For a fixed $j \in [T]$, $[k_j]$ is the set of non-noisy i's where $k_j \geq k/2$.

- **The Y-variables** . For each $(i, j) \in [k] \times [T] \setminus \mathcal{I}$, $Y_{ij} = \sqrt{(T-1)/T} \cdot \delta_j + b\eta$. For $(i, j) \in \mathcal{I}$, Y_{ij}'s are independent $N(0, 1)$ random variables.

- **The W-variables** For a fixed j, we define variables $W_{1j}, \ldots, W_{k_j j}$ with $W_{1j} = \sqrt{(T-1)/T} \cdot \delta_j + b\eta$ and $W_{2j}, \ldots, W_{k_j j}$ are 0.

- **U-variables** We define $U_1 = \frac{1}{\sqrt{T}} \sum_{j \in [T]} W_{1j}$ which is $b\eta\sqrt{T}$ and is independent of the variables $U_2, \ldots, U_T$ where each U_t is i.i.d. $N(0, 1)$ for $t > 1$.

4.2 A Hybrid Basis Relative to j^* and d^*

Recall that we have fixed a nice $\mathcal{I}$. In this section, we define a basis for polynomials using a fixed choice of $j^* \in [T]$ and $d^* \in [d]$. For convenience let $[T_{-j^*}] := [T] \setminus \{j^*\}$.

Definition 4.3. *Let $\mathscr{H}_{-j^*}$ be the Hermite basis for all polynomials over the independent Gaussian variables* $\{Y_{ij} \mid i \in [k]\setminus[k_j], j \in [T_{-j^*}]\}$. *In particular,* $\mathbb{E}[H^2] = \mathbb{E}[G^2] = 1$ *and* $\mathbb{E}[HG] = 0$ *for each* $H, G \in \mathscr{H}_{-j^*}$, $H \neq G$. *Let $\mathscr{H}_{-j^*d^*}$ be the set of basis elements of $\mathscr{H}_{-j^*}$ of degree exactly d^*.*

Definition 4.4. *Let $\mathscr{M}_{-j^*}$ be the* standard monomial basis *for polynomials over the variables* $\{W_{1j} \mid j \in [T_{-j^*}]\}$. *In particular, each element of $\mathscr{M}_{-j^*}$ is of the form* $\prod_{j \in [T_{-j^*}]} W_{1j}^{a_j}$ *for some non-negative integers* a_j $(j \in [T_{-j^*}])$.

Definition 4.5. *Let $\mathscr{B}_{-j^*} := \mathscr{H}_{-j^*} \circ \mathscr{M}_{-j^*}$ be the combined basis for polynomials over the variables of $\mathscr{H}_{-j^*}$ and $\mathscr{M}_{-j^*}$, where each element B is of the form HM for some $H \in \mathscr{H}_{-j^*}$ and $M \in \mathscr{M}_{-j^*}$ and $\deg(B) = \deg(H) + \deg(M)$. For convenience we also define the subset $\mathscr{B}_{-j^*d^*} := \mathscr{H}_{-j^*d^*} \circ \mathscr{M}_{-j^*}$, i.e. each element of $\mathscr{B}_{-j^*d^*}$ is of the form HM where $H \in \mathscr{H}_{-j^*d^*}$ and $M \in \mathscr{M}_{-j^*}$.*

Lastly, let $\mathscr{S}_{j^*}$ be the set of all multisets of $R_{j^*} = \{(i, j^*) \mid i \in [k]\}$. For an element $S \in \mathscr{S}_{j^*}$, let $S(i, j^*)$ denote the number of occurrences of (i, j^*) in S. Using this, we define $Y_S := \prod_{(i,j^*) \in R_{j^*}} Y_{ij^*}^{S(i,j^*)}$.

Writing the polynomial P in the basis given by products of $\mathscr{B}_{-j^*}$, $\{W_{ij} : j \in [T_{-j^*}], i \in [k_j] \setminus \{1\}\}$ and $\{Y_{ij^*} : i \in [k]\}$, the polynomial P can be represented as:

$$P = P_{\text{omit}} + \sum_{\substack{S \in \mathscr{S}_{j^*} \\ B \in \mathscr{B}_{-j^*}}} c_{S,B} Y_S B, \tag{28}$$

where $c_{S,B}$ are constants and[5] P_{omit} is the sub-polynomial of P consisting of all monomials containing a variable from $\{W_{ij} : j \in [T_{-j^*}], i \in [k_j] \setminus \{1\}\}$. Of course, since P is of degree at most d, the only terms that occur in the above sum satisfy $\deg(B) + |S| \leq d$.

For a fixed $0 \leq d^* \leq d - 1$ we will be interested in capturing the the mass of P linear in Y_{ij^*} and the subset $\mathscr{B}_{-j^*d^*}$. Abusing notation to let $c_{(i,j^*),B} = c_{(S,B)}$ where $S = \{(i, j^*)\}$ is the singleton multiset, define

$$c_{i,j^*,d^*} = \sqrt{\sum_{B \in \mathscr{B}_{-j^*d^*}} c_{(i,j^*),B}^2} \tag{29}$$

for each $(i, j^*) \in R_{j^*}$ and $0 \leq d^* \leq d - 1$.

4.3 Main Structural Lemma

We are now ready to describe the structure that P must exhibit in order to pass the Basic PCP test. Let us first define a *distinguished pair* (j^*, d^*) for a fixed setting of $\mathcal{I}$.

Definition 4.6. *A pair $(j^*, d^*) \in [T] \times \{0, \ldots, d-1\}$ is said to be* distinguished *for $\mathcal{I}$ if,*

$$\sum_{(i,j^*) \in \mathcal{I}} c_{i,j^*,d^*}^2 \leq \frac{\varepsilon^4}{4} \cdot \left(\sum_{(i,j^*) \in ([k] \times \{j^*\}) \setminus \mathcal{I}} c_{i,j^*,d^*}^2 \right), \tag{30}$$

---

[5]The reason for treating P_{omit} separately is that it vanishes under the distribution $\mathcal{D}_{\mathcal{I}}$.

and,

$$\sum_{(i,j^*)\in([k]\times\{j^*\})\setminus\mathcal{I}} c^2_{i,j^*,d^*} > 0. \tag{31}$$

Here, ε is the noise parameter used in the PCP test.

The main lemma that we prove is the following.

Lemma 4.7 (Main Structural Lemma). *For every* nice *choice of $\mathcal{I}$, there exists $j^* \in [T]$ and $d^* \in \{0, 1, \ldots, d-1\}$ such that (j^*, d^*) is distinguished for $\mathcal{I}$.*

The proof of the above lemma is given in Section 6 building upon analysis in Section 5. Both Sections 5 and 6 assume a setting of nice $\mathcal{I}$.

Using (24) and a simple averaging, the above lemma implies that there exists (j^*, d^*) such that:

$$\Pr_{\mathcal{I}}\left[(j^*, d^*) \text{ is distinguished for } \mathcal{I}\right] \geq \frac{\xi}{4Td}. \tag{32}$$

4.4 Implications of the Structural Lemma

We now fix (j^*, d^*) satisfying (32). Let us consider the random choice of $\mathcal{I}$ as first picking $\mathcal{I}_{-j^*} := \mathcal{I} \cap ([k] \times ([T] \setminus \{j^*\}))$, and then picking $\mathcal{I}_{j^*} := \mathcal{I} \cap ([k] \times \{j^*\})$. Note that the choice of $\mathcal{I}_{j^*}$ is independent of $\mathcal{I}_{-j^*}$. Call a choice of $\mathcal{I}_{-j^*}$ as *shared-heavy* if,

$$\Pr_{\mathcal{I}_{j^*}}\left[(j^*, d^*) \text{ is distinguished for } \mathcal{I}_{j^*} \cup \mathcal{I}_{-j^*}\right] \geq \frac{\xi}{8Td}. \tag{33}$$

From (32) and an averaging argument we have:

$$\Pr_{\mathcal{I}_{-j^*}}\left[\mathcal{I}_{-j^*} \text{ is shared-heavy}\right] \geq \frac{\xi}{8Td}. \tag{34}$$

Let us fix a shared-heavy $\mathcal{I}_{-j^*}$. Note that with this fixing, the bases given in Section 4.2 are well defined, and in particular P can be represented as in (28). Since there is at least one choice of $\mathcal{I}_{j^*}$ such that (j^*, d^*) is distinguished for $\mathcal{I}_{j^*} \cup \mathcal{I}_{-j^*}$, using (31) this implies

$$\sum_{i\in[k]} c^2_{i,j^*,d^*} > 0. \tag{35}$$

Further we have the following lemma. (This is where we are finally randomizing over $\mathcal{I}_{j^*}$.)

Lemma 4.8. *There exists $i^* \in [k]$ such that,*

$$c^2_{i^*,j^*,d^*} \geq \nu^2 \left(\sum_{i\in[k]} c^2_{i,j^*,d^*}\right),$$

for $\nu = \varepsilon^2/2$.

Proof. Assume that there is no such i^* as in the lemma. Over the choice of $\mathcal{I}_{j^*}$, consider the random variable $\sum_{(i,j^*)\in\mathcal{I}_{j^*}} c_{i,j^*,d^*}^2$. The contribution from each i to this sum is independently 0 with probability $(1-\varepsilon)$ and c_{i,j^*,d^*}^2 with probability ε. Thus,

$$\mathbb{E}_{\mathcal{I}_{j^*}}\left[\sum_{(i,j^*)\in\mathcal{I}_{j^*}} c_{i,j^*,d^*}^2\right] = \varepsilon\left(\sum_{i\in[k]} c_{i,j^*,d^*}^2\right).$$

Now,

$$\Pr\left[\sum_{(i,j^*)\in\mathcal{I}} c_{i,j^*,d^*}^2 \le (\varepsilon/2)\left(\sum_{(i,j^*)\in[k]\times\{j^*\}\setminus\mathcal{I}} c_{i,j^*,d^*}^2\right)\right]$$

$$\le \Pr\left[\sum_{(i,j^*)\in\mathcal{I}_{j^*}} c_{i,j^*,d^*}^2 \le (\varepsilon/2)\left(\sum_{i\in[k]} c_{i,j^*,d^*}^2\right)\right]$$

$$\le \Pr\left[\left|\sum_{(i,j^*)\in\mathcal{I}_{j^*}} c_{i,j^*,d^*}^2 - \mathbb{E}\left[\sum_{(i,j^*)\in\mathcal{I}_{j^*}} c_{i,j^*,d^*}^2\right]\right| \ge (\varepsilon/2)\left(\sum_{i\in[k]} c_{i,j^*,d^*}^2\right)\right]$$

$$\overset{1}{\le} 2\cdot\exp\left(-\frac{2(\varepsilon/2)^2\cdot\left(\sum_{i\in[k]} c_{i,j^*,d^*}^2\right)^2}{\sum_{i\in[k]} c_{i,j^*,d^*}^4}\right)$$

$$\le 2\cdot\exp\left(-\frac{(\varepsilon^2/2)\cdot\left(\sum_{i\in[k]} c_{i,j^*,d^*}^2\right)^2}{\max_{i\in[k]} c_{i,j^*,d^*}^2 \sum_{i\in[k]} c_{i,j^*,d^*}^2}\right) \le 2\cdot\exp\left(-\varepsilon^2/2\nu^2\right) \le \varepsilon,$$

for $\nu^2 = \varepsilon^4/4 \le \varepsilon^2/(2\log(2/\varepsilon))$. Here, step 1 follows from the Chernoff-Hoeffding inequality (Theorem 2.5). Since our choice of $\varepsilon < \xi/(8Td)$, this yields a contradiction to our choice of $\mathcal{I}_{-j^*}$, (30), and (33).

$\square$

4.5 Decoding a Labeling for $\mathcal{L}$

In Figure 4 we define a randomized (partial) labeling σ for the vertices V of $\mathcal{L}$. To analyze σ, we first define the following random subsets of vertices and edges, where the randomness is over the choices made in the above procedure of labeling.

Vertex subset $V_0 \subseteq V$: Consists of all $v \in V$ such that:

- Setting $v_{j^*} = v$, the choice of $\{v_j \mid j \in [T]\}$ is good,

- The choice of (j^*, d^*) satisfied (32) and,

- The choice of $\mathcal{I}_{-j^*}$ is shared-heavy.

Over the randomness of the labeling procedure and a random choice of v, the above happens with probability at least:

$$\Delta_0 := \xi \cdot \frac{1}{Td} \cdot \frac{\xi}{8Td}. \tag{37}$$

17

<div style="border:1px solid">

Randomized Partial Labeling σ

1. Choose $j^* \in [T]$ and $d^* \in \{0, \ldots, d-1\}$ independently and u.a.r.
2. Choose $v_j \in V$ independently and u.a.r. for each $j \in [T] \setminus \{j^*\}$.
3. Choose the random subset $\mathcal{I}_{-j^*}$ of $[k] \times ([T] \setminus \{j^*\})$ by independently adding each element with probability ε.
4. For each $v \in V$,
   (a) Set $v_{j^*} = v$.
   (b) Letting P be the restriction of P_{global} to $\mathbf{Y} = \{Y_{ij} \mid i \in [k], j \in [T]\}$, define the set:

$$\Gamma_0(v) \quad := \quad \left\{ i' \in [k] \mid c^2_{i',j^*,d^*} > \frac{\nu^2}{4} \left(\sum_{i \in [k]} c^2_{i,j^*,d^*} \right) \right\}, \tag{36}$$

   where $\nu = \varepsilon^2/4$ (as in Lemma 4.8).
   (c) If $\Gamma_0(v)$ is non-empty, assign v a label chosen uniformly at random from $\Gamma_0(v)$.

</div>

Figure 4: Randomized Partial Labeling

Thus,

$$\mathbb{E}\left[|V_0|\right] \geq \Delta_0 |V|.$$

Moreover, by the weak expansion property in Theorem 2.2,

$$\mathbb{E}\left[|E(V_0)|\right] \geq \mathbb{E}\left[(|V_0|/|V|)^2\right] \cdot (|E|/2) \geq (\mathbb{E}\left[|V_0|/|V|\right])^2 \cdot (|E|/2) \geq \left(\Delta_0^2/2\right)|E|. \tag{38}$$

Edge Set $E' \subseteq E(V_0)$: Let us first define for each $v \in V$

$$\Gamma_1(v) \quad := \quad \left\{ i' \in [k] \mid c^2_{i',j^*,d^*} > \frac{\nu^2}{100 \cdot 4^{2R}} \left(\sum_{i \in [k]} c^2_{i,j^*,d^*} \right) \right\}, \tag{39}$$

when v_{j^*} is set to v in Step 4a of Figure 4. Here, R is the parameter (to be set) from Theorem 2.2. From (36) and (39), we have $\Gamma_0(v) \subseteq \Gamma_1(v)$ along with

$$|\Gamma_0(v)| \leq 4/\nu^2, \quad \text{and} \quad |\Gamma_1(v)| \leq (100 \cdot 4^{2R})/\nu^2. \tag{40}$$

The set E' is defined as:

$$E' := \left\{ e = (u,w) \in E(V_0) \mid |\pi_{e,u}(\Gamma_1(u))| = |\Gamma_1(u)| \text{ and } |\pi_{e,w}(\Gamma_1(w))| = |\Gamma_1(w)| \right\}. \tag{41}$$

Since the graph G of the instance $\mathcal{L}$ is regular, using second bound in (40) along with the smoothness property of Theorem 2.2, the fraction of edges $e = (u,w) \in E$ that do not satisfy

$$(|\pi_{e,u}(\Gamma_1(u))| = |\Gamma_1(u)| \text{ and } |\pi_{e,w}(\Gamma_1(w))| = |\Gamma_1(w)|)$$

is at most,

$$\Delta_1 := \left(\frac{10^4 \cdot 4^{4R}}{\nu^4 J} \right).$$

Thus,
$$\mathbb{E}\left[|E'|\right] \geq \left(\Delta_0^2/2 - \Delta_1\right)|E|. \tag{42}$$

The following lemma gives the desired property of edges in E'.

Lemma 4.9. *For every edge* $e = (u, w) \in E'$,
$$\pi_{e,u}\left(\Gamma_0(u)\right) \cap \pi_{e,w}\left(\Gamma_0(w)\right) \neq \emptyset. \tag{43}$$

Proof. Suppose for a contradiction that (43) does not hold for an edge $e = (u, w) \in E'$, i.e.
$$\pi_{e,u}\left(\Gamma_0(u)\right) \cap \pi_{e,w}\left(\Gamma_0(w)\right) = \emptyset. \tag{44}$$

Let us now define for $v \in \{u, w\}$, and $i \in [k]$, vector $\mathbf{C}_{v,i} \in \mathbb{R}^{\mathscr{B}_{j^*d^*}}$ where for any $B \in \mathscr{B}_{j^*d^*}$
$$\mathbf{C}_{v,i}(B) = c_{(i,j^*),B} \quad \text{when } v_{j^*} \text{ is set to } v. \tag{45}$$

Without loss of generality, we may assume that
$$\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2 \geq \sum_{i \in [k]} \|\mathbf{C}_{w,i}\|_2^2. \tag{46}$$

Since both $u \in V_0$, (35) and Lemma 4.8 imply that there exists $i_u \in [k]$ such that
$$\|\mathbf{C}_{u,i_u}\|_2 \geq \nu \left(\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2\right)^{\frac{1}{2}} > 0. \tag{47}$$

This implies that $i_u \in \Gamma_0(u)$. Now, let $\ell^* := \pi_{e,u}(i_u)$. Since P is a restriction of P_{global} which is a representation of the folded polynomial $\overline{P}_{\text{global}}$, Lemma 3.2 along with Remark 3.3 (applied to elements B of $\mathscr{B}_{-j^*d^*}$) implies
$$\sum_{i \in \pi_{e,u}^{-1}(\ell^*)} \mathbf{C}_{u,i} = \sum_{i \in \pi_{e,w}^{-1}(\ell^*)} \mathbf{C}_{w,i}. \tag{48}$$

On the other hand, since $e \in E'$, (41) along with our supposition (44) and the construction of $\{\Gamma_r(v) \mid r \in \{0,1\}, v \in \{u, w\}\}$ implies that

- For all $i \in \pi_{e,u}^{-1}(\ell^*) \setminus \{i_u\}$
$$\|\mathbf{C}_{u,i}\|_2 \leq \frac{\nu}{10 \cdot 4^R} \left(\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2\right)^{\frac{1}{2}}. \tag{49}$$

- For all $i \in \pi_{e,w}^{-1}(\ell^*)$
$$\|\mathbf{C}_{w,i}\|_2 \leq \frac{\nu}{2} \left(\sum_{i \in [k]} \|\mathbf{C}_{w,i}\|_2^2\right)^{\frac{1}{2}}. \tag{50}$$

- There exists at most one $i' \in [k]$ such that,

$$\|\mathbf{C}_{w,i}\|_2 > \frac{\nu}{10 \cdot 4^R} \left(\sum_{i \in [k]} \|\mathbf{C}_{w,i}\|_2^2\right)^{\frac{1}{2}}. \tag{51}$$

The above implications along with (48) and (46) yields

$$
\begin{aligned}
\|\mathbf{C}_{u,i_u}\|_2 &\leq \sum_{\substack{i \in \pi_{e,u}^{-1}(\ell^*) \\ i \neq i_u}} \|\mathbf{C}_{u,i}\|_2 + \sum_{i \in \pi_{e,w}^{-1}(\ell^*)} \|\mathbf{C}_{w,i}\|_2 \\
&\leq \frac{\nu \left|\pi_{e,u}^{-1}(\ell^*)\right|}{10 \cdot 4^R} \left(\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2\right)^{\frac{1}{2}} + \left(\frac{\nu}{2} + \frac{\nu \left|\pi_{e,w}^{-1}(\ell^*)\right|}{10 \cdot 4^R}\right) \left(\sum_{i \in [k]} \|\mathbf{C}_{w,i}\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{\nu}{10} \left(\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2\right)^{\frac{1}{2}} + \left(\frac{\nu}{2} + \frac{\nu}{10}\right) \left(\sum_{i \in [k]} \|\mathbf{C}_{w,i}\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{7\nu}{10} \left(\sum_{i \in [k]} \|\mathbf{C}_{u,i}\|_2^2\right)^{\frac{1}{2}},
\end{aligned}
\tag{52}
$$

where we used the property (from Theorem 2.2) that $\left|\pi_{e,u}^{-1}(\ell^*)\right|, \left|\pi_{e,w}^{-1}(\ell^*)\right| \leq 4^R$. Clearly, (52) is a contradiction to (47) which completes the proof of the lemma. $\square$

Note that the set E' is determined by Step 3 of the randomized labeling procedure. Lemma 4.9 implies that in the subsequent steps of the procedure, each edge $e = (u, w) \in E'$ is satisfied with probability at least

$$\frac{1}{|\Gamma_0(u)| \, |\Gamma_0(w)|} \geq \frac{\nu^4}{16},$$

using the first bound in (40). The above along with (42) lower bounds the expected fraction of edges σ satisfies by

$$\Delta_2 := \left(\Delta_0^2/2 - \Delta_1\right) \left(\frac{\nu^4}{16}\right).$$

Choosing R to be large enough and $J \gg 4^{4R}$ we can ensure that $\Delta_2 > 2^{-c_0 R}$ which yields a contradiction to the soundness of Theorem 2.2, completing the NO case analysis.

4.6 Loose Ends

Discretization of the Basic PCP Test Distribution. Let $\mathcal{H}_N$ be the distribution of $\left(\sum_{i=1}^N B_i\right)/\sqrt{N}$ where each B_i is an independent $\{-1, 1\}$-valued balanced Bernoulli random variable. The following theorem was proved in [DOSW11].

Theorem 4.10. *Fix any constant $D \geq 1$, and let $f(x_1, \ldots, x_m)$ be any degree-D polynomial over $\mathbb{R}^m$. Let $(y, z) \in \mathbb{R}^m \times \mathbb{R}^m$ be generated by sampling each (y_i, z_i) from $(N(0,1), \mathcal{H}_N)$ where $N = m^{24D^2}$. Then,*

$$\Pr\left[\mathrm{sign}(f(y)) \neq \mathrm{sign}(f(z))\right] \leq O(1/m).$$

In our Basic PCP Test distribution (for a fixed choice of the vertices of the SMOOTH LABEL COVER instance) we have $m = \Theta(kT)$ Gaussian random variables. Choosing $D = d$ and $N = m^{24D^2}$, we can completely discretize the test distribution using $\exp((kT)^{O(d^2)})$ points. Note that this also incorporates the possible $2^{O(kT)}$ choices of the noise set $\mathcal{I}$. From the above theorem, this discretization results in an at most $O(1/kT)$ loss in the acceptance probability of the test. This discretization is done for all possible choices by the test of the vertices of the instance.

Ruling out functions of constantly many degree-d PTFs.   Analogous to the argument in [KS11], consider any function $\overline{h}$ of K degree-d PTFs (over $\mathcal{F}$) that passes the Final PCP test with probability $1/2 + \xi$. Let h be the function $\overline{h}$ with the PTFs represented over $\mathbb{R}^{\mathcal{Y}}$. By averaging, h flips its sign with respect to flipping b for at least ξ fraction of the rest of the choices made by the Basic PCP Test. Again by averaging, there must be a degree-d PTF sign $\left(P'_{\text{global}}\right)$ satisfying the same for at least ξ/K fraction of the choices. The entire analysis can then be repeated using P'_{global}.

5   Relative bounds for mass in P

Let $\mathbf{Z}$ denote the set of variables $\{Y_{ij} : j \in [T], k_j < i \leq k\}$. As shown in Section 4, the $\mathbf{Z}$ variables are all i.i.d. $N(0,1)$ under the test distribution. We begin by expressing P as

$$P\Big(\mathbf{Z}, \{U_i\}_{i \in [T]}, \{W_{ij}\}_{i \neq 1}\Big) \;\; = \;\; P_{\text{omit}}\Big(\mathbf{Z}, \{U_i\}_{i \in [T]}, \{W_{ij}\}_{i \in [2,k_j], j \in [T]}\Big) + Q_0(\mathbf{Z}, U_2, \ldots, U_T)$$
$$+ U_1 Q_1(\mathbf{Z}, U_1, \ldots, U_T) \tag{53}$$

where P_{omit} consists of all the terms that contain some $\{W_{ij} \mid i \in [2, k_j], j \in [T]\}$ as a factor, and Q_0 is the part in the remaining polynomial independent of U_1. From the nice setting of $\mathcal{I}$, we have that with probability at least $\xi/2$ over the rest of the choices of the verifier, P flips its sign on flipping b. Since P_{omit} evaluates to zero under the test distribution and Q_0 is independent of $b\eta$ by construction, we obtain that Q_1 is not identically zero. For the time being, our analysis ignores P_{omit}. Extending Definitions 4.3 and 4.4, let $\mathcal{H}$ be the Hermite basis over all the $\mathbf{Z}$ variables, and $\mathcal{M}$ be the monomial basis over the variables $\{W_{1j} : j \in [T]\}$. Using these we define two norms to quantify the relevant mass of polynomials. For convenience, let $\mathbf{U}$ denote the variables $U_1, \ldots, U_T$, $\widetilde{\mathbf{U}}$ denote the set $\mathbf{U} \setminus \{U_1\}$, and $\mathbf{W}$ denote the set of variables $W_{11}, \ldots, W_{1T}$.

Definition 5.1 ($\| \cdot \|_2$-norm). *Given a polynomial Q over the variables defined in the PCP test, define its $\| \cdot \|_2$-norm as*

$$\|Q\|_2 = \sqrt{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_\mathcal{I}}\Big[|Q(\mathbf{x})|^2\Big]}.$$

Definition 5.2 ($\| \cdot \|_{\text{mon},1}, \| \cdot \|_{\text{mon},2}$-norms). *Given a polynomial $Q(\mathbf{W}) = \sum_{W_S \in \mathcal{M}} c_S W_S$ represented in the monomial basis $\mathcal{M} = \{W_S\}$, for any $p \geq 1$ define its $\| \cdot \|_{\text{mon},p}$-norm as*

$$\|Q\|_{\text{mon},p} = \left(\sum_{W_S \in \mathcal{M}} |c_S|^p\right)^{1/p}.$$

In particular, $\| \cdot \|_{\text{mon},1}$ is the absolute sum of the coefficients, and $\| \cdot \|_{\text{mon},2}^2$ is the squared sum of the coefficients in Q,

As pointed out above, Q_1 is not identically zero and therefore by definition it satisfies.

$$\|Q_1\|_2 > 0 \tag{54}$$

Our goal in this section is to prove the following lemma lower bounding $\|Q_1\|_2$ relative to $\|Q_0\|_2$.

Lemma 5.3. *Using the definitions given above,*

$$\|Q_0\|_2 \leq \left(\frac{8\eta\sqrt{T}}{(\xi/4d)^d\sqrt{\xi}} \right) \|Q_1\|_2 \tag{55}$$

Proof. From Lemma 4.2, we know that $U_1 = b\eta\sqrt{T}$ under the distribution $\mathcal{D}_{\mathcal{I}}$. Since Q_1 is dependent on U_1, its distribution can be dependent on b. Let $Q_1^+ := Q_1|_{b=1}$, and and $Q_1^- := Q_1|_{b=-1}$. Thus,

$$
\begin{aligned}
\|Q_1\|_2^2 = \mathbb{E}_{b,Z,\mathbf{U}}\big[|Q_1|^2\big] &= \frac{1}{2}\mathbb{E}_{Z,\mathbf{U}}\big[|Q_1|^2\big|b=1\big] + \frac{1}{2}\mathbb{E}_{Z,\mathbf{U}}\big[|Q_1|^2\big|b=-1\big] \\
&= \frac{1}{2}\|Q_1^+\|_2^2 + \frac{1}{2}\|Q_1^-\|_2^2.
\end{aligned} \tag{56}
$$

Using the above along with Chebyshev's inequality (see Section 2.3) we obtain for any $a > 0$

$$
\begin{aligned}
\Pr_{\mathbf{Z},\widetilde{\mathbf{U}}}\big[|Q_1^+|,|Q_1^-| \leq a\|Q_1\|_2\big] &\geq 1 - \Pr\big[|Q_1^+| \geq a\|Q_1\|_2\big] - \Pr\big[|Q_1^-| \geq a\|Q_1\|_2\big] \\
&\geq 1 - \left(\frac{\|Q_1^+\|_2^2 + \|Q_1^-\|_2^2}{a^2\|Q_1\|_2^2} \right) = 1 - 2/a^2,
\end{aligned} \tag{57}
$$

where the last step follows from (56). On the other hand note that Q_0 is a polynomial over standard Gaussian variables and is independent of b. Applying the bound of Carbery-Wright (Theorem 2.4) we obtain the following.

$$\Pr\left[|Q_0| \leq (\xi/4d)^d\|Q_0\|_2\right] \leq \frac{\xi}{4} \tag{58}$$

Setting $a = 4/\sqrt{\xi}$ in (57) and using the above we obtain that with probability at least $1 - \xi/4 - \xi/8 = 1 - 3\xi/8$ over the choice of the variables $\mathbf{Z}$ and $U_2, \ldots, U_T$

$$(\eta\sqrt{T})|Q_1^+|, \ (\eta\sqrt{T})|Q_1^-| \ \leq \ (4\eta\sqrt{T/\xi})\|Q_1\|_2, \quad \text{and,} \quad |Q_0| \ > \ (\xi/4d)^d\|Q_0\|_2.$$

When $\eta\sqrt{T}(|Q_1^+| + |Q_1^-|) < |Q_0|$ then flipping b does not change the sign of P. Since the sign of P must flip with b with probability at least $\xi/2$ over the choice of $\mathbf{Z}$ and $U_2, \ldots, U_T$, the above is a contradiction unless,

$$\|Q_0\|_2 \leq \left(\frac{8\eta\sqrt{T}}{(\xi/4d)^d\sqrt{\xi}} \right) \|Q_1\|_2,$$

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

6   Proof of Main Structural Lemma 4.7

As in the previous section, we have $\mathbf{U}$ denote the variables $U_1, \ldots, U_T$, $\widetilde{\mathbf{U}}$ denote the set $\mathbf{U} \setminus \{U_1\}$, and $\mathbf{W}$ denote the set of variables $W_{11}, \ldots, W_{1T}$. Similarly, we use $\mathbf{Y} = \{Y_{ij} : i \in [k], j \in [T]\}$ to denote the set

of all the Y variables. We use $\mathbf{Z}$ to denote the set of variables $\{Y_{ij} : j \in [T], k_j < i \le k\}$. The $\mathbf{Z}$ variables are all $N(0,1)$ under the test distribution. For a particular $j^* \in [T]$, let $\mathbf{Z}_{j^*} = \mathbf{Z} \cap \{Y_{ij^*} : i \in [k]\}$, and let $\mathbf{Z}_{-j^*} = \mathbf{Z} \setminus \mathbf{Z}_{j^*}$. Also, for given $j^* \in [T]$, define $\mathbf{Y}_{j^*} = \mathbf{Y} \cap \{Y_{ij^*} : i \in [k]\}$ and $\mathbf{Y}_{-j^*} = \mathbf{Y} \setminus \mathbf{Y}_{j^*}$. Finally, for given $j^* \in [T]$, we define $\mathbf{W}_{j^*}$ and $\mathbf{W}_{-j^*}$ similarly.

Recall the definitions of the bases in Definitions 4.3, 4.4 and 4.5. Extending these as in the previous section, let $\mathscr{H}$ be the Hermite basis for polynomials in the variables $\mathbf{Z}$ and $\mathscr{M}$ the monomial basis for polynomials in the variables $\mathbf{W}$. For any $D \in [d]$, we also define $\mathscr{H}_D$ to be the set of all Hermite monomials of degree exactly D.

For convenience of measuring the monomial mass, we use Definition 5.2 to define two different norms as follows:

Definition 6.1 ($\| \cdot \|_{\mathscr{B}}$-Norm)**.** *For a polynomial* $L(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}} H(\mathbf{Z}) \cdot L_H(\mathbf{W})$, *let*

$$\|L(\mathbf{Z}, \mathbf{W})\|_{\mathscr{B}}^2 = \sum_{H \in \mathscr{H}} \|L_H(\mathbf{W})\|_{\mathrm{mon},2}^2 \tag{59}$$

Definition 6.2 ($\| \cdot \|_{\mathscr{B}_{-j^*,d^*,J}}$-Norm)**.** [6] *Suppose* $j^* \in [T], d^* \in [d-1]$ *and* $J \subseteq [k]$ *are given. Then, for any polynomial* $M(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*})$ *of the form*

$$M(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*}) = \sum_{H \in \mathscr{H}_{-j^*}} \sum_{S \in \mathscr{S}_{j^*}} H(\mathbf{Z}_{-j^*}) \cdot Y_S \cdot M_{H,S}(\mathbf{W}_{-j^*}),$$

we define:

$$\left\| M(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*}) \right\|_{\mathscr{B}_{-j^*,d^*,J}}^2 = \sum_{H \in \mathscr{H}_{-j^*,d^*}} \sum_{i \in J} \|M_{H,\{(i,j^*)\}}(\mathbf{W}_{-j^*})\|_{\mathrm{mon},2}^2 \tag{60}$$

Finally, for $j^* \in [T]$, we shall find it convenient to define the sets $\mathcal{A}_1^{j^*} = \{i : (i,j^*) \in \mathcal{I}\}$, and $\mathcal{A}_0^{j^*} = [k] \setminus \mathcal{A}_1^{j^*}$.

6.1 An intermediate Lemma

We start by writing the polynomial P in the variables $\mathbf{Z}, \{W_{ij} : j \in [T], 1 < i \le k_j\}, \mathbf{U}$:

$$P = P_{\mathrm{omit}} + P_{\mathrm{rel}} = P_{\mathrm{omit}} + \overline{Q}_0(\mathbf{Z}, \mathbf{U} \setminus \{U_1\}) + U_1 \cdot \overline{Q}_1(\mathbf{Z}, \mathbf{U})$$

where P_{omit} contains all monomials depending on variables in $\{W_{ij} : j \in [T], 1 < i \le k_j\}$.

Let $Q_0(\mathbf{Z}, \mathbf{W})$ and $Q_1(\mathbf{Z}, \mathbf{W})$ be $\overline{Q}_0$ and $\overline{Q}_1$ respectively after a change of variables from $\mathbf{U}$ to $\mathbf{W}$. For $a = 0, 1$, we write $Q_a(\mathbf{Z}, \mathbf{W})$ in the $\mathscr{H} \circ \mathscr{M}$ basis: $Q_a(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}} H(\mathbf{Z}) \cdot Q_{a,H}(\mathbf{W})$. For a fixed $d^* \in \{0\} \cup [d-1]$, we let

$$Q_a^{(d^*)}(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}_{d^*}} H(\mathbf{Z}) \cdot Q_{a,H}(\mathbf{W}).$$

For a fixed $j^* \in [T]$, we define P_{omit,j^*} as the sub-polynomial of P containing all the monomials containing at least one variable from $\{W_{ij} : j \ne j^*, i \ne 1\}$, and let P_{rel,j^*} be the rest of the polynomial.

We shall prove Lemma 4.7 using the following intermediate result:

---

[6]Note that although we call it so, $\| \cdot \|_{\mathscr{B}_{-j^*,d^*,J}}$ is not an actual norm, as it may vanish even for non-zero polynomials.

Lemma 6.3. *There exists choice of $d^* \in \{0, 1, \ldots, d-1\}$ and $j^* \in [T]$ such that the following properties hold simultaneously:*

1. $\|Q_0\|_{\mathscr{B}}^2 \leq \rho^{2d} \|Q_1\|_{\mathscr{B}}^2$

2. $\|Q_1^{(d^*+1)}\|_{\mathscr{B}}^2 \leq \frac{1}{4} \rho^{d^*+1} \|Q_1\|_{\mathscr{B}}^2$

3. $\left\| \widetilde{Q} \right\|_{\mathscr{B}_{-j^*,d^*,\mathcal{A}_0^{j^*}}}^2 \geq \frac{1}{8kT^2} (20dT)^{-4^d} \rho^{d^*} \|Q_1\|_{\mathscr{B}}^2$

where $\rho = (20dkT^3/\epsilon^4)^{-6^d} (kT)^{-1}$ and $\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*})$ is the polynomial obtained by rewriting the $\mathbf{W}_{j^*}$ variables in P_{rel,j^*} in terms of the $\mathbf{Y}_{j^*}$ variables.

Using this, we give a proof of Lemma 4.7.

Proof of Lemma 4.7. Let d^* and j^* be as given in Lemma 6.3. Let $\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*},)$ be as in the Lemma 6.3. We can express $\widetilde{Q}$ as :

$$\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{W}_{-j^*}, \mathbf{Y}_{j^*}) = \sum_{D=0}^{d-1} \sum_{H \in \mathscr{H}_{-j^*D}} \sum_{S \in \mathscr{S}_{j^*}} HY_S \widetilde{Q}_{H,S}(\mathbf{W}_{-j^*}) \tag{61}$$

where $\mathscr{H}_{-j^*D}$ is the set of Hermite monomials which are of degree D and do not contain $\mathbf{Z}_{j^*}$ variables. By construction we have

$$\sum_{(i,j^*) \in \mathcal{I}} c_{i,j^*,d^*}^2 = \|\widetilde{Q}\|_{\mathscr{B}_{-j^*,d^*,\mathcal{A}_1^{j^*}}}^2 \tag{62}$$

Consider a term that contributes to the RHS of (62) (as defined in 6.2). Since the additional Y_{ij^*} (for $(i,j^*) \in \mathcal{I}$ variable adds to the degree of H, the corresponding term appears in the $\mathscr{B}$-representation of P_{rel} as HM where the degree of H is of degree $d^* + 1$. Therefore it must be a part of $Q_0^{(d^*+1)}$ or $Q_1^{(d^*+1)}$. Hence,

$$\|\widetilde{Q}\|_{\mathscr{B}_{-j^*,d^*,\mathcal{A}_1^{j^*}}}^2 \leq \|U_1 Q_1^{(d^*+1)}\|_{\mathscr{B}}^2 + \|Q_0\|_{\mathscr{B}}^2 \overset{1}{\leq} T\|Q_1^{(d^*+1)}\|_{\mathscr{B}}^2 + \rho^{2d}\|Q_1\|_{\mathscr{B}}^2 \leq 2T\rho^{d^*+1}\|Q_1\|_{\mathscr{B}}^2 \tag{63}$$

where the upper bound on the first term in step 1 follows from

$$\begin{aligned}
\|U_1(\mathbf{W})Q_1^{(d^*+1)}(\mathbf{W})\|_{\mathscr{B}}^2 &= \sum_{H \in \mathscr{H}_{d^*+1}} \|U_1(\mathbf{W})Q_H(\mathbf{W})\|_{\mathrm{mon},2}^2 \\
&\leq \sum_{H \in \mathscr{H}_{d^*+1}} \|U_1(\mathbf{W})\|_{\mathrm{mon},1}^2 \|Q_H(\mathbf{W})\|_{\mathrm{mon},2}^2 \qquad \left(\text{Claim } C.1\right) \\
&= T\|Q_1^{(d^*+1)}\|_{\mathscr{B}}^2
\end{aligned}$$

and the upper bound on the second term in step 1 follows from Lemma 6.3 (part 1). The last inequality uses Part 2. of Lemma 6.3. On the other hand we have,

$$\sum_{(i,j^*) \in ([k] \times \{j^*\}) \backslash \mathcal{I}} c_{i,j^*,d^*}^2 = \|\widetilde{Q}\|_{\mathscr{B}_{-j^*,d^*,\mathcal{A}_0^{j^*}}}^2 \tag{64}$$

24

From Lemma 6.3 (part 3) and the choice of ρ in Lemma 6.3 we have

$$\|\widetilde{Q}\|^2_{\mathscr{B}_{-j^*,d^*,\mathcal{A}_0^{j^*}}} \geq \frac{1}{8kT^2}(20dT)^{-4^d}\rho^{d^*}\|Q_1\|^2_{\mathscr{B}} \geq \frac{16T}{\epsilon^4}\rho^{d^*+1}\|Q_1\|^2_{\mathscr{B}} \tag{65}$$

Combining (63),(64) and (65), we get an upper bound on LHS of (62) which gives us

$$\sum_{(i,j^*)\in\mathcal{I}} c^2_{i,j^*,d^*} \leq \frac{\epsilon^4}{8}\left(\sum_{(i,j^*)\in([k]\times\{j^*\})\setminus\mathcal{I}} c^2_{i,j^*,d^*}\right) \tag{66}$$

thus implying inequality (30). Furthermore, from (54), we know that $\|Q_1\|^2_2 > 0$, which along with Lemma B.1(part 1) implies that $\|Q_1\|^2_{\mathscr{B}} > 0$. Therefore, combining (65) and (64), we get that the LHS of (64) is strictly positive, thus implying (31). Hence, the choice of (d^*, j^*) satisfy (30) and (31). $\qquad\square$

6.2 Proof of Lemma 6.3

6.2.1 Upper bounding $\|Q_0\|_{\mathscr{B}}$ in terms $\|Q_1\|_{\mathscr{B}}$

In this section, we show that $\|Q_0\|_{\mathscr{B}}$ is small compared terms $\|Q_1\|_{\mathscr{B}}$ due to our choice of η.

Lemma 6.4. *Let ρ be chosen as in Lemma 6.3. Then $\|Q_0\|^2_{\mathscr{B}} \leq \rho^{2d}\|Q_1\|^2_{\mathscr{B}}$*

Proof. We express Q_0 as

$$Q_0(\mathbf{Z},\mathbf{W}) = \sum_{H\in\mathscr{H}} HQ_{0,H}(\mathbf{W})$$

where $H \in \mathscr{H}$ are the Hermite monomials. Then by definition of $\|\cdot\|^2_{\mathscr{B}}$ we have,

$$\begin{aligned}
\|Q_0(\mathbf{Z},\mathbf{W})\|^2_{\mathscr{B}} &= \sum_{H\in\mathscr{H}}\|Q_{0,H}(\mathbf{W})\|^2_{\text{mon},2} \\
&\overset{1}{\leq} (10dT)^{14d}\sum_{H\in\mathscr{H}}\|Q_{0,H}(\widetilde{\mathbf{U}})\|^2_2 \\
&= (10dT)^{14d}\|Q_0(\widetilde{\mathbf{U}})\|^2_2 \\
&\overset{2}{\leq} \frac{\rho^{4d}}{4}\|Q_1\|^2_2
\end{aligned}$$

where step 1 follows from Lemma B.1 (part 2), and step 2 follows from Claim 5.3 and our choice of η in Section 3. Furthermore, we can relate the $\|Q_1\|^2_2$ to $\|Q_1\|^2_{\mathscr{B}}$ as follows

$$\|Q_1\|^2_{\mathscr{B}} = \sum_{H\in\mathscr{H}}\|Q_{1,H}(\mathbf{W})\|^2_{\text{mon},2} \overset{1}{\geq} (20dT)^{-10d}\sum_H\|Q_{1,H}(\mathbf{U})\|^2_2 = (20dT)^{-10d}\|Q_1\|^2_2$$

where step 1 follows from Lemma B.1 (part 1). Combining the bounds, we get $\|Q_0\|^2_{\mathscr{B}} \leq \rho^{2d}\|Q_1\|^2_{\mathscr{B}}$. $\qquad\square$

6.2.2 Finding a heavy $d^* \in \{0,1,\ldots,d-1\}$

We begin by finding a $d^* \in \{0\}\cup[d-1]$ such that Q_1 restricted to Hermite monomials in $\mathscr{H}_{d^*}$ has large mass compared to those from $\mathscr{H}_{d^*+1}$.

25

Lemma 6.5. *There exists* $d^* \in \{0\} \cup [d-1]$ *such that*

1. $\|Q_1^{(d^*+1)}\|_{\mathscr{B}}^2 \leq \frac{1}{4}\rho^{d^*+1}\|Q_1\|_{\mathscr{B}}^2$

2. $\|Q_1^{(d^*)}\|_{\mathscr{B}}^2 \geq \frac{1}{4}\rho^{d^*}\|Q_1\|_{\mathscr{B}}^2$

Proof. We claim that there exists $D \in \{0\} \cup [d-1]$ such that $\|Q_1^{(D)}\|_{\mathscr{B}}^2 \geq \frac{1}{4}\rho^D\|Q_1\|_{\mathscr{B}}^2$. If not, then for all $D \in \{0\} \cup [d-1]$ we have $\|Q_1^{(D)}\|_{\mathscr{B}}^2 < \frac{1}{4}\rho^D\|Q_1\|_{\mathscr{B}}^2$. Then,

$$\|Q_1\|_{\mathscr{B}}^2 = \sum_{D=0}^{d-1}\|Q^{(D)}\|_{\mathscr{B}}^2 \leq \sum_{D=0}^{d-1}\frac{\rho^D}{4}\|Q_1\|_{\mathscr{B}}^2 < \frac{1}{2}\|Q_1\|_{\mathscr{B}}^2$$

which is a contradiction.

Now we set d^* to be the largest such $D \in \{0\} \cup [d-1]$ such that $\|Q_1^{(D)}\|_{\mathscr{B}}^2 \geq \frac{1}{4}\rho^D\|Q_1\|_{\mathscr{B}}^2$. If $d^* < d-1$, then by construction we know that $\|Q_1^{(d^*+1)}\|_{\mathscr{B}}^2 < \frac{1}{4}\rho^{d^*+1}\|Q_1\|_{\mathscr{B}}^2$. On the other hand if $d^* = d-1$, then by construction $Q_1^{(d^*+1)}$ is identically 0 (since Q_1 is of degree at most $d-1$) and hence the claim is vacuously true. $\qquad\square$

6.2.3 Locating a good $j^* \in [T]$

Let $d^* \in \{0\} \cup [d-1]$ be as in Lemma 6.5. Now, we shall find a good $j^* \in [T]$ in the sub-polynomial $U_1Q_1^{(d^*)}$ which contains a sub-polynomial linear in W_{1j^*} with significant $\|\cdot\|_{\mathscr{B}}$-mass.

Lemma 6.6. *Let the polynomial* $U_1Q^{(d^*)}(\mathbf{Z}, \mathbf{W})$ *be expressed in the basis* $\mathscr{B}$ *as*

$$U_1Q^{(d^*)}(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}_{d^*}}\sum_{M \in \mathscr{M}} c_{H,M} HM$$

Then there exists $j^* \in [T]$ *such that*

$$\sum_{H \in \mathscr{H}_{-j^*d^*}}\sum_{M \in \mathscr{M}_{-j^*}} c_{H,MW_{1j^*}}^2 \geq \frac{1}{T^2}(20dT)^{-4^d}\left(\sum_{H \in \mathscr{H}_{d^*}}\sum_{M \in \mathscr{M}} c_{H,M}^2\right) \qquad (67)$$

Proof. Consider the following representation of $U_1Q_1^{(d^*)}$:

$$U_1Q_1^{(d^*)}(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}_{d^*}} HU_1Q_{1,H}(\mathbf{W}) \qquad (68)$$

Using the fact that $U_1 = (1/\sqrt{T})\sum_{j=1}^{T} W_{1j}$ and $T = 10d$, the following lemma is directly implied by Lemma 7.1.

Lemma 6.7. *Fix* $H \in \mathscr{H}_{d^*}$. *Let* $U_1Q_{1,H}(\mathbf{W})$ *(as defined in* (68)*) be expressed in the basis* $\mathscr{B}$ *as*

$$U_1Q_{1,H}(\mathbf{W}) = \sum_{M \in \mathscr{M}} c_{H,M}M$$

Then there exists at least $T/2$ *choices of* $j^* \in [T]$ *such that*

$$\sum_{M \in \mathscr{M}_{-j^*}} c_{H,MW_{1j^*}}^2 \geq \frac{1}{T}(20dT)^{-4^d}\sum_{M \in \mathscr{M}} c_{H,M}^2 \qquad (69)$$

26

For a fixed Hermite monomial $H \in \mathscr{H}_{d^*}$, we call a $j^* \in [T]$ to be *good* for H if the following conditions hold:

1. The Hermite monomial H does not contain $\mathbf{Z}_{j^*}$-variables.

2. The index j^* satisfies (69) with respect to H

Now for a fixed Hermite monomial $H \in \mathscr{H}_{d^*}$, out of T values of j, at most $d-1$ can appear in H. Furthermore, Lemma 6.7 guarantees that for at least $T/2$-values of $j \in [T]$, (69) is satisfied. Since $T = 10d$, for each Hermite monomial H there exists at least some $j^*(H)$ which is good for H. Therefore by averaging over all $H \in \mathscr{H}_{d^*}$, there exists $j^* \in [T]$ such that

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,MW_{1j^*}} \geq \frac{1}{T} \sum_{H \in \mathscr{H}_{-d^*}} \sum_{M \in \mathscr{M}_{-j^*(H)}} c^2_{H,MW_{1j^*(H)}} \geq \frac{1}{T^2}(20dT)^{-4d}\left(\sum_{H \in \mathscr{H}_{d^*}} \sum_{M \in \mathscr{M}} c^2_{H,M}\right)$$

$\square$

6.2.4 Substituting $\mathbf{W}_{j^*}$ with Y_{j^*}-variables

For the $j^* \in [T]$ chosen in the previous section, P_{rel,j^*} can be rewritten by expanding $\mathbf{W}_{j^*}$ in the $\mathbf{Y}_{j^*}$-variables as $\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{Y}_{j^*}, \mathbf{W}_{-j^*})$ which can be expressed in the basis $\mathscr{B}_{-j^*}$ as follows:

$$\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{W}_{-j^*}, \mathbf{Y}_{j^*}) = \sum_{D=0}^{d-1} \sum_{H \in \mathscr{H}_{-j^*D}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{S \in \mathscr{S}_{j^*}} \tilde{c}_{H,M,S} HMY_S \tag{70}$$

where $\mathscr{H}_{-j^*D}$, $\mathscr{M}_{-j^*}$ and $\mathscr{S}_{j^*}$ are as defined in Section 4.2. Now we show that the squared sum of coefficients in the above expression, restricted to factors to terms of the form HMY_{ij^*} capture a significant fraction of mass.

Claim 6.8. *Let $\widetilde{Q}(\mathbf{Z}_{-j^*}, \mathbf{W}_{-j^*}, \mathbf{Y}_{j^*})$ be as in (70). Then,*

$$\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k_{j^*}]} \tilde{c}^2_{H,M,(ij^*)} \geq \frac{1}{2k_{j^*}}\left(\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,MW_{1j^*}}\right) \tag{71}$$

Proof. Consider the polynomial P_{lin} defined as follows:

$$P_{\text{lin}}(\mathbf{Z}, \mathbf{W}) = \sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \sum_{i \in [k_{j^*}]} \alpha_{H,M,i} HMW_{ij^*} \tag{72}$$

which is the sub-polynomial in P consisting of monomials containing exactly one $\mathbf{W}_{j^*}$-variable. Note that terms on the RHS of (72) for $i > 1$ are contained in P_{omit}.

Fix a $HM \in \mathscr{H}_{-j^*d^*} \circ \mathscr{M}_{-j^*}$ and $i \in [k_{j^*}]$. Under the linear transformation $\mathbf{W}_{j^*} \mapsto \mathbf{Y}_{j^*}$ we have

$$\tilde{c}_{H,M,(ij^*)} = \sum_{l \in [k_{j^*}]} \alpha_{H,M,l} c_{l,i} \tag{73}$$

27

where the $c_{1,l}, \ldots, c_{T,l}$ are the lth coordinates of vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$ (as in Section 4). Recall that $\langle \mathbf{c}_i, \mathbf{c}_{i'} \rangle = 0$ for all $i \neq i'$. Therefore

$$\sum_{i \in [k_{j^*}]} \tilde{c}^2_{H,M,Y_{ij^*}} = \left\| \sum_{l \in [k_{j^*}]} \alpha_{H,M,l} \mathbf{c}_l \right\|^2 \tag{74}$$

$$= \sum_{l \in [k_{j^*}]} \left\| \alpha_{H,M,l} \mathbf{c}_l \right\|^2 \tag{75}$$

$$\geq \alpha^2_{H,M,1} \|\mathbf{c}_1\|^2 = \frac{\alpha^2_{H,M,1}}{k_{j^*}} \tag{76}$$

To finish the proof, we note that for $i = 1$ the RHS of (72) has contribution either from terms in $U_1 Q_1^{(d^*)}$ or Q_0. Summing over all pairs $HM \in \mathscr{B}_{-j^*}$ and using the triangle inequality we obtain

$$\sqrt{\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} \alpha^2_{H,M,1}} \geq \sqrt{\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,M,W_{1j^*}}} - \|Q_0\|_{\mathscr{B}} \tag{77}$$

$$\geq \frac{1}{\sqrt{2}} \sqrt{\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,M,W_{1j^*}}} \tag{78}$$

where we upper bound $\|Q_0\|_{\mathscr{B}}$ as follows:

$$\|Q_0\|^2_{\mathscr{B}} \overset{1}{\leq} \rho^{2d} \|Q_1\|^2_{\mathscr{B}} \overset{2}{\leq} \rho^d \|Q_1^{(d^*)}\|^2_{\mathscr{B}} = \rho^d \sum_{H \in \mathscr{H}_{d^*}} \sum_{M \in \mathscr{M}} c^2_{H,M} \tag{79}$$

$$\overset{3}{\leq} \frac{1}{16} \sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,M,W_{1j^*}} \tag{80}$$

where inequality 1 follows from Lemma 6.4, inequality 2 follows from Lemma 6.5 and the last inequality follows from Lemma 6.6 and our choice of ρ. $\qquad \square$

6.2.5   Completing the proof of Lemma 6.3

Part 1 follows from Lemma 6.4 and Part 2 follows directly from Lemma 6.5. For Part 3, observe that the LHS of Part 3 (in Lemma 6.3) is equal to the LHS of (71), which can be lower bounded using Claim 6.8, Lemma 6.6 and Lemma 6.5 as follows

$$\frac{1}{2k_{j^*}} \left(\sum_{H \in \mathscr{H}_{-j^*d^*}} \sum_{M \in \mathscr{M}_{-j^*}} c^2_{H,MW_{1j^*}} \right) \geq \frac{1}{2k_{j^*}T^2} (20dT)^{-4^d} \left(\sum_{H \in \mathscr{H}_{d^*}} \sum_{M \in \mathscr{M}} c^2_{H,M} \right) \tag{81}$$

$$= \frac{1}{2T^2 k_{j^*}} (20dT)^{-4^d} \|Q_1^{(d^*)}\|^2_{\mathscr{B}} \tag{82}$$

$$\geq \frac{1}{8T^2 k_{j^*}} (20dT)^{-4^d} \rho^{d^*} \|Q_1\|^2_{\mathscr{B}} \tag{83}$$

which completes the proof.

7 A Linear Mass Bound for Low Degree Polynomials

In this section we study the structure of polynomials over the variable set $\{W_1, \ldots, W_T\}$. For a polynomial $P(W_1, \ldots, W_T)$, dropping the subscript we use $\|P\|$ to denote the ℓ_2-norm of the coefficients of P in the monomial basis. Let $U := \sum_{j=1}^{T} W_j$. Define $Q(W_1, \ldots, W_T) = U \cdot S(W_1, \ldots, W_T)$, a polynomial of degree $d+1$. For any $j \in [T]$, write:

$$S(W_1, \ldots, W_T) = \sum_{\ell=0}^{d} W_j^{\ell} \cdot S_{j,\ell}(\mathbf{W}_{\neq j}) \tag{84}$$

$$Q(W_1, \ldots, W_T) = \sum_{\ell=1}^{d+1} W_j^{\ell} \cdot Q_{j,\ell}(\mathbf{W}_{\neq j}) \tag{85}$$

where $\mathbf{W}_{\neq \sigma} = \{W_i\}_{i \notin \sigma}$ for any list σ of indices. The main result of this section is the following lemma showing that for many $j \in [T]$, the W_j-linear sub-polynomial $Q_{j,1}$ has significant mass:

Lemma 7.1. *For polynomials S and Q as above, if $T > 2d$, there are at least $T/2$ choices of $j \in [T]$ such that $\|Q_{j,1}\| \geq (20dT)^{-3^d} \|S\|$.*

The rest of this section is devoted to proving Lemma 7.1.

7.1 The Variable Removal Lemma

The key ingredient that is needed to prove this is the following lemma that will be iteratively applied while reducing the number of variables and the degree at each iteration:

Lemma 7.2 (Variable Removal). *Let $d \geq 1$. For variables X, Y, Z, suppose there are polynomials S_1, S_2 of degree $d-1$, polynomials R_1, R_2 of degree $d-2$, and error polynomials Δ^X, Δ^Y of degree d satisfying:*

$$(aX - Y - Z)S_1(Y, Z) + \Delta^X(X, Y, Z) + X^2 R_1(X, Y, Z)$$
$$= (aY - X - Z)S_2(X, Z) + \Delta^Y(X, Y, Z) + Y^2 R_2(X, Y, Z). \tag{86}$$

Then,

$$S_1(Y, Z) = \left((a+1)Y - Z\right)C(Z) + Y^2 A_1(Y, Z) + \Delta(Y, Z)$$

where Δ is such that $\|\Delta\| \leq 20a \max(\|\Delta^X\|, \|\Delta^Y\|)$. Furthermore, we have $\deg(C(Z)) \leq d - 2$, $\deg(A_1(Y, Z)) \leq d - 3$, and $\deg(\Delta(Y, Z)) \leq d - 1$.

Proof. We write the polynomials S_1 and S_2 in the following way[7]:

$$S_1(Y, Z) = Y^2 \cdot A_1(Y, Z) + Y \cdot B_1(Z) + Z \cdot C_1(Z) + D_1$$
$$S_2(X, Z) = X^2 \cdot A_2(X, Z) + X \cdot B_2(Z) + Z \cdot C_2(Z) + D_2$$

Note that $C_1(Z)$ and $A_1(Y, Z)$ can be of degree at most $d-2$ and $d-3$ respectively. Additionally, we write the error polynomials as:

$$\Delta^X = X \cdot \Delta_X^X + Z \cdot \Delta_Z^X + Z^2 \cdot \Delta_{Z^2}^X(Z) + YZ \cdot \Delta_{YZ}^X(Z) + \tilde{\Delta}^X(X, Y, Z)$$
$$\Delta^Y = X \cdot \Delta_X^Y + Z \cdot \Delta_Z^Y + Z^2 \cdot \Delta_{Z^2}^Y(Z) + YZ \cdot \Delta_{YZ}^Y(Z) + \tilde{\Delta}^Y(X, Y, Z)$$

To be clear, the functions without any arguments, such as Δ_X^X or Δ_Z^Y, are constants. The above decomposition is unique. Now we match coefficients in (86).

---

[7]If $d \leq 2$, then some of the polynomials below are automatically 0.

29

1. Matching terms of the form $X^0Y^0Z^{\geq 2}$, we get $-C_1(Z) + \Delta^X_{Z^2} = -C_2(Z) + \Delta^Y_{Z^2} \Rightarrow C_2(Z) = C_1(Z) + \Delta^Y_{Z^2} - \Delta^X_{Z^2}$

2. Matching terms of the form $X^1Y^0Z^0$, we get $aD_1 + \Delta^X_X = -D_2 + \Delta^Y_X \Rightarrow D_2 = -aD_1 + \Delta^Y_X - \Delta^X_X$

3. Matching terms of the form $X^0Y^0Z^1$, we get $-D_1 + \Delta^X_Z = -D_2 + \Delta^Y_Z$. Substituting D_2 from above:

$$-D_1 = -D_2 + \Delta^Y_Z - \Delta^X_Z = aD_1 - (\Delta^Y_X - \Delta^X_X) + (\Delta^Y_Z - \Delta^X_Z)$$

which on rearranging gives us $D_1 = -\frac{1}{a+1}\left[\Delta^Y_Z - \Delta^X_Z - \Delta^Y_X + \Delta^X_X\right]$

4. Matching $X^0Y^1Z^{\geq 1}$ we get $-B_1(Z) - C_1(Z) + \Delta^X_{YZ} = aC_2(Z) + \Delta^Y_{YZ}$. Substituting $C_2(Z)$ from above,

$$
\begin{aligned}
-B_1(Z) &= aC_2(Z) + C_1(Z) + \Delta^Y_{YZ} - \Delta^X_{YZ} \\
&= a\left(C_1(Z) + \Delta^Y_{Z^2} - \Delta^X_{Z^2}\right) + C_1(Z) + \Delta^Y_{YZ} - \Delta^X_{YZ} \\
&= (a+1)C_1(Z) + a\left(\Delta^Y_{Z^2} - \Delta^X_{Z^2}\right) + \Delta^Y_{YZ} - \Delta^X_{YZ}
\end{aligned}
$$

Finally by substituting $B_1(Z)$ and D_1 in the expression for $S_1(Y,Z)$ and collecting the error terms, we get

$$
\begin{aligned}
S_1(Y,Z) &= Y^2 A_1(Y,Z) - Y\left[(a+1)C_1(Z) + a\left(\Delta^Y_{Z^2} - \Delta^X_{Z^2}\right) + \Delta^Y_{YZ} - \Delta^X_{YZ}\right] + ZC_1(Z) + D_1 \\
&= Y^2 A_1(Y,Z) - C_1(Z)\left[(a+1)Y - Z\right] - Y\left[a\left(\Delta^Y_{Z^2} - \Delta^X_{Z^2}\right) + \Delta^Y_{YZ} - \Delta^X_{YZ}\right] \\
&\quad - \frac{1}{a+1}\left[\Delta^Y_Z - \Delta^X_Z - (\Delta^Y_X - \Delta^X_X)\right] \\
&= Y^2 A_1(Y,Z) - C_1(Z)\left[(a+1)Y - Z\right] + \Delta(Y,Z)
\end{aligned}
$$

We obtain the lemma setting $C(Z) = -C_1(Z)$ and $\Delta(Y,Z) = -Y\left[a\left(\Delta^Y_{Z^2} - \Delta^X_{Z^2}\right) + \Delta^Y_{YZ} - \Delta^X_{YZ}\right]$. The upper bound on $\|\Delta\|$ follows by triangle inequality. $\qquad\square$

7.2 Proof of Lemma 7.1

Fix $j \in [T]$. Comparing the coefficients of the sub-polynomial that are degree 1 in W_j in the expansion of Q (see (85)) and US (see (84)), we get

$$Q_{j,1}(\mathbf{W}_{\neq j}) = S_{j,0} + S_{j,1}(\mathbf{W}_{\neq j})\sum \mathbf{W}_{\neq j} \tag{87}$$

where $\sum \mathbf{W}_{\neq \sigma}$ is the sum of all variables in $\mathbf{W}_{\neq \sigma}$ for any list σ of indices. Denote $Q_{j,1}(\mathbf{W}_{\neq j})$ as $\Delta^{(1)}_j$.

The proof is by contradiction i.e., we assume that more than $T/2$ of the $\Delta^{(1)}_j$ polynomials have small mass. We show first that there exist many j's such that the sub-polynomial of S not divisible by W_j^2 retains significant mass. This is achieved using Lemma 7.3. Next, we apply Lemmas 7.2 and 7.3 as well as the degree bound on S to obtain a contradiction.

7.2.1 Finding a non-quadratic sub-polynomial with significant mass

Lemma 7.3. *Given a polynomial P on variables $W_1, \ldots, W_T$ of degree d such that $\|P\| = 1$, let $P = W_j^2 P_j(W_1, \ldots, W_T) + R_j(W_1, \ldots, W_T)$ for every $j \in [T]$ where $R_j(\cdot)$ is the sub-polynomial which does not contain a W_j^2 factor. Then, if $T > d$, there exists $j \in [T]$ such that $\|R_j\| > 4^{-2^d} \|P\|$.*

Proof. Without loss of generality, assume $\|P\| = 1$ by rescaling. Suppose that for all $j \in [d]$, $\|R_j\| \leq \eta := 4^{-2^d}$. We show that this violates the degree bound on P using the following claim.

Claim 7.4. *For every $j \in [d]$, if polynomials H_j and L_j are defined such that $P = W_1^2 \cdots W_j^2 \cdot H_j + L_j$ and L_j is not divisible by $W_1^2 \cdots W_j^2$, then $\|L_j\| \leq 4 \cdot \eta^{1/2^{j-1}}$.*

This claim proves the lemma because it shows $\|L_d\| \leq 4 \cdot \eta^{1/2^{d-1}} < 1/2$, so $\|H_d\| > 0$ (since they contribute disjoint monomials to P), and therefore P contains a monomial of degree $2d$, a contradiction. $\square$

Proof of Claim 7.4. The proof is by induction on j. The base case $j = 1$ is clear, since $L_1 = R_1$.

For the inductive step, suppose the claim is true for $j - 1$. Then, we have that $W_j^2 P_j + R_j = P = W_1^2 \cdots W_{j-1}^2 H_{j-1} + L_{j-1}$ with $\|L_{j-1}\| \leq 4\eta^{1/2^{j-2}}$. Write $H_{j-1} = W_j^2 H_j' + L_j'$ where L_j' is not divisible by W_j^2. Now, $P = W_1^2 \cdots W_j^2 H_j' + W_1^2 \cdots W_{j-1}^2 L_j' + L_{j-1}$.

By looking at the terms divisible by W_j^2, we have that $\|W_j^2 P_j\| = \|P_j\| \leq \|H_j'\| + \|L_{j-1}\|$. Since $\|P_j\| \geq 1 - \eta$ and $\|L_{j-1}\| \leq 4\eta^{1/2^{j-2}}$, we get that $\|H_j'\| \geq 1 - 8\eta^{1/2^{j-2}}$.

Let $H_j = H_j'$ and $L_j = W_1^2 \cdots W_{j-1}^2 L_j' + L_{j-1}$. Then,

$$\|L_j\|^2 = 1 - \|H_j\|^2 = 1 - \|H_j'\|^2 \leq 1 - (1 - 8\eta^{1/2^{j-2}})^2 \leq 16\eta^{1/2^{j-2}}$$

$\square$

7.2.2 Iterative expansion of S

We are now ready to prove Lemma 7.1. For contradiction, suppose that $\max_{j \in [T/2]} \|\Delta_j^{(1)}\| \leq C_{\max} := (20dT)^{-3^d}$. By rescaling, we can assume $\|S\| = 1$. We expand the polynomial S iteratively using Lemma 7.2. At each step, we shall use Lemma 7.3 to find a W_j variable such that S contains a sub-polynomial of significant mass which is not divisible by W_j^2.

As a first step, using (87) and the definition of $\Delta_1^{(1)}$, for every $j \in [T/2]$, we can write:

$$S(\mathbf{W}) = \left(W_j - \sum \mathbf{W}_{\neq j}\right) \cdot S_j^{(1)}(\mathbf{W}_{\neq j}) + W_j^2 \cdot R_j^{(1)}(\mathbf{W}) + \Delta_j^{(1)}(\mathbf{W}) \tag{88}$$

where $S_i^{(1)}$, $R_i^{(1)}$ and $\Delta_j^{(1)}$ are polynomials of degrees at most $d-1$, $d-2$ and d respectively and $\|\Delta_j^{(1)}\| \leq C_{\max}$. Because $T/2 > d$, using Lemma 7.3 and re-indexing, we can assume that the sub-polynomial of S not divisible by W_1^2 has ℓ_2-norm at least $\eta := 4^{-2^d}$.

Now, applying the variable reduction lemma (Lemma 7.2) for every $j \in [2, T/2]$, with $a = 1, X = W_1, Y = W_j$, and $Z = \sum \mathbf{W}_{\neq 1,j}$, we obtain that there exist polynomials $S_j^{(2)}$, $R_j^{(2)}$ and $\Delta_j^{(2)}$ of degrees $d-2$, $d-3$ and $d-1$ respectively such that

$$S_1^{(1)}(\mathbf{W}_{\neq 1}) = \left(2W_j - \sum \mathbf{W}_{\neq 1,j}\right) \cdot S_j^{(2)}(\mathbf{W}_{\neq 1,j}) + W_j^2 \cdot R_j^{(2)}(\mathbf{W}_{\neq 1}) + \Delta_j^{(2)}(\mathbf{W}_{\neq 1})$$

and $\|\Delta_j^{(2)}\| \leq 20C_{\max}$. Again, by Lemma 7.3 and re-indexing, we can ensure that the sub-polynomial of $S_1^{(1)}$ not divisible by W_2^2 has ℓ_2-norm at least $\eta\|S_1^{(1)}\|$.

Applying the variable reduction lemma again with $a = 2$, we obtain polynomials $S_j^{(3)}, R_j^{(3)}$ and $\Delta_j^{(3)}$ of degrees $d - 3, d - 4$, and $d - 2$ respectively such that for any $j \in [3, T/2]$:

$$S_2^{(2)}(\mathbf{W}_{\neq 1,2}) = \left(3W_j - \sum \mathbf{W}_{\neq 1,2,j}\right) \cdot S_j^{(3)}(\mathbf{W}_{\neq 1,2,j}) + W_j^2 \cdot R_j^{(3)}(\mathbf{W}_{\neq 1,2}) + \Delta_j^{(3)}(\mathbf{W}_{\neq 1,2})$$

and $\|\Delta_j^{(3)}\| \leq 20^2 \cdot 2 \cdot C_{\max}$. Continuing this way, we get that for every $1 \leq \ell < j \leq T/2$, there exist polynomials $S_j^{(\ell)}, R_j^{(\ell)}$ and $\Delta_j^{(\ell)}$ of degrees $d - \ell, d - \ell - 1$, and $d - \ell + 1$ such that:

$$\begin{aligned}
S_{\ell-1}^{(\ell-1)}(\mathbf{W}_{\neq[\ell-1]}) &= \left(\ell W_j - \sum \mathbf{W}_{\neq[\ell-1]\cup\{j\}}\right) \cdot S_j^{(\ell)}(\mathbf{W}_{\neq[\ell-1]\cup\{j\}}) \\
&\quad + W_j^2 \cdot R_j^{(\ell)}(\mathbf{W}_{\neq[\ell-1]}) + \Delta_j^{(\ell)}(\mathbf{W}_{\neq[\ell-1]})
\end{aligned} \tag{89}$$

and $\|\Delta_j^{(\ell)}\| \leq (20\ell)^{\ell-1}C_{\max}$. Here, $S_0^{(0)} = S$. Moreover, using Lemma 7.3, we can assume that the sub-polynomial of $S_{\ell-1}^{(\ell-1)}$ not divisible by W_ℓ^2 has ℓ_2-mass at least $\eta\|S_{\ell-1}^{(\ell-1)}\|$.

For $\ell = d$, we obtain a linear polynomial $S_{d-1}^{(d-1)}(\mathbf{W}_{\neq 1,\ldots,d-1})$ such that for every $j \in [d, T/2]$, there exists constant $S_j^{(d)}$ and linear polynomial $\Delta_j^{(d)}$ such that:

$$S_{d-1}^{(d-1)}(\mathbf{W}_{\neq 1,\ldots,d-1}) = \left(dW_j - \sum \mathbf{W}_{\neq 1,\ldots,d-1,j}\right) \cdot S_j^{(d)} + \Delta_j^{(d)}(\mathbf{W}_{\neq 1,\ldots,d-1})$$

Note that $R_j^{(d)} = 0$ because $S_{d-1}^{(d-1)}$ is not divisible by W_j^2 being a linear polynomial.

Applying Lemma 7.2 one final time, we get that $|S_d^{(d)}| \leq (40d)^d C_{\max}$. On the other hand, we have the following claim:

Claim 7.5. *For any $0 \leq \ell \leq T/2$, $\|S_\ell^{(\ell)}\| \geq \left(\frac{\eta}{T}\right)^\ell - 2\frac{(20\ell)^\ell C_{\max}}{T}$.*

Proof. The proof is by induction. For $\ell = 0$, the claim is true because $\|S_0^{(0)}\| = \|S\| = 1$. For the induction, note that by our choice of the index ℓ above, the sub-polynomial of $S_{\ell-1}^{(\ell-1)}$ not divisible by W_ℓ^2 has ℓ_2-mass at least $\eta\|S_{\ell-1}^{(\ell-1)}\|$. Moreover, from (89) and triangle inequality this mass is at most

$$\|(\ell W_\ell - \sum \mathbf{W}_{\neq[\ell]})S_\ell^{(\ell)}\| + \|\Delta_\ell^{(\ell)}\|$$

So:

$$\begin{aligned}
\|(\ell W_\ell - \sum \mathbf{W}_{\neq[\ell]})S_\ell^{(\ell)}\| &\geq \eta\|S_{\ell-1}^{(\ell-1)}\| - \|\Delta_\ell^{(\ell)}\| \\
&\geq \eta\|S_{\ell-1}^{(\ell-1)}\| - (20\ell)^{\ell-1}C_{\max} \\
&\geq \eta^\ell/T^{\ell-1} - 2\eta(20\ell)^{\ell-1}C_{\max}/T - (20\ell)^{\ell-1}C_{\max} \\
&\geq \eta^\ell/T^{\ell-1} - 2(20\ell)^\ell C_{\max}
\end{aligned}$$

The claim follows by observing $\|(\ell W_\ell - \sum \mathbf{W}_{\neq[\ell]})S_\ell^{(\ell)}\| \leq T \cdot \|S_\ell^{(\ell)}\|$. $\qquad\square$

Therefore, $|S_d^{(d)}| \geq (\eta/T)^d - 2(20d)^d C_{\max}/T$. But by our choice of η and $C_{\max}$, $(\eta/T)^d - 2(20d)^d C_{\max}/T > (40d)^d C_{\max}$, since $C_{\max}((40d)^d + 2(20d)^d/T) < C_{\max}(80d)^d < (1/4T)^{2^d} = (\eta/T)^d$. This is a contradiction.

References

[ABF⁺08] M. Alekhnovich, M. Braverman, V. Feldman, A. R. Klivans, and T. Pitassi. The complexity of properly learning simple concept classes. *J. Comp. Sys. Sci.*, 74(1):16–34, 2008.

[ABL17] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

[ABX08] B. Applebaum, B. Barak, and D. Xiao. On basing lower-bounds for learning on worst-case assumptions. In *Proc. 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 211–220, 2008.

[AK98] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoret. Comput. Sci.*, 209(1-2):237–260, 1998.

[AV06] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[BB06] N. H. Bshouty and L. Burroughs. Maximizing agreements and coagnostic learning. *Theoret. Comput. Sci.*, 350(1):24–39, 2006.

[BDEL03] S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *J. Comp. Sys. Sci.*, 66(3):496 – 514, 2003.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[BFKV98] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998.

[BGGS16] A. Bhattacharyya, A. Gadekar, S. Ghoshal, and R. Saket. On the hardness of learning sparse parities. In *Proc. 24th European Symposium on Algorithms*, pages 11:1–11:17, 2016.

[BK97] A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution. *J. Comp. Sys. Sci.*, 54(2):371–380, 1997.

[BKW03] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

[BR93] A. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. In *Machine Learning: From Theory to Applications - Cooperative Research at Siemens and MIT*, pages 9–28, 1993.

[Coh97] E. Cohen. Learning noisy perceptrons by a perceptron in polynomial time. In *Proc. 38th Annual IEEE Symposium on Foundations of Computer Science*, pages 514–523, 1997.

[CV95] C. Cortes and V. Vapnik. Support-Vector networks. *Machine Learning*, 20(3):273–297, 1995.

[CW01] A. Carbery and J. Wright. Distributional and l_q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.*, 8(3):233–248, 2001.

[Dan15]   A. Daniely. A PTAS for agnostically learning halfspaces. In *Proc. 28th Annual ACM Workshop on Computational Learning Theory*, pages 484–502, 2015.

[Dan16]   A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proc. 48th Annual ACM Symposium on the Theory of Computing*, pages 105–117, 2016.

[DOSW11]   I. Diakonikolas, R. O'Donnell, R. A. Servedio, and Y. Wu. Hardness results for agnostically learning low-degree polynomial threshold functions. In *Proc. 22nd ACM-SIAM Symposium on Discrete Algorithms*, pages 1590–1606, 2011.

[DS16]   A. Daniely and S. Shalev-Shwartz. Complexity theoretic limitations on learning DNF's. In *Proc. 29th Annual ACM Workshop on Computational Learning Theory*, pages 815–830, 2016.

[FGKP09]   V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM J. Comput.*, 39(2):606–645, 2009.

[FGRW12]   V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM J. Comput.*, 41(6):1558–1590, 2012.

[GKS10]   P. Gopalan, S. Khot, and R. Saket. Hardness of reconstructing multivariate polynomials over finite fields. *SIAM J. Comput.*, 39(6):2598–2621, 2010.

[GR09]   V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39(2):742–765, 2009.

[GRSW16]   V. Guruswami, P. Raghavendra, R. Saket, and Yi Wu. Bypassing UGC from some optimal geometric inapproximability results. *ACM Trans. Algorithms*, 12(1):6:1–6:25, 2016.

[Hås01]   J. Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, 2001.

[Hau92]   D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Inform. and Comp.*, 100(1):78–150, 1992.

[JP78]   D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoret. Comput. Sci.*, 6:93–107, 1978.

[Kho02]   S. Khot. On the power of unique 2-prover 1-round games. In *Proc. 34th Annual ACM Symposium on the Theory of Computing*, pages 767–775, 2002.

[Kho09]   S. Khot. personal communication, 2009.

[KK14]   A. R. Klivans and P. Kothari. Embedding hard learning problems into gaussian space. In *Proc. 18th International Workshop on Randomization and Computation (RANDOM)*, pages 793–809, 2014.

[KKMS05]   A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[KLS09]   A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *J. Mach. Learn. Res.*, 10:2715–2740, 2009.

[KMV08] A. T. Kalai, Y. Mansour, and E. Verbin. On agnostic boosting and parity learning. In *Proc. 40th Annual ACM Symposium on the Theory of Computing*, pages 629–638, 2008.

[KOS04] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comp. Sys. Sci.*, 68(4):808–840, 2004.

[KS08] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. *J. Comp. Sys. Sci.*, 74(1):35–48, 2008.

[KS09] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. *J. Comp. Sys. Sci.*, 75(1):2–12, 2009.

[KS11] S. Khot and R. Saket. On the hardness of learning intersections of two halfspaces. *J. Comp. Sys. Sci.*, 77(1):129–141, 2011.

[KSS94] M.J. Kearns, R.E. Schapire, and L.M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[MP69] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1969.

[Mur90] O.J. Murphy. Nearest neighbor pattern classification perceptrons. *Proceedings of the IEEE*, 78(10):1595–1598, 1990.

[Ros62] F. Rosenblatt. *Principles of Neurodynamics*. Spartan, New York, 1962.

[RRK04] U. Rükert, L. Richter, and S. Kramer. Quantitative association rules based on half-spaces. In *Proc. 4th IEEE International Conference on Data Mining*, pages 507–510, 2004.

[Vem97] S. Vempala. A random sampling based algorithm for learning the intersection of half-spaces. In *Proc. 38th Annual IEEE Symposium on Foundations of Computer Science*, pages 508–513, 1997.

[Vio09] E. Viola. The sum of D small-bias generators fools polynomials of degree D. *Comput. Complexity*, 18(2):209–217, 2009.

A Useful Tools and Results

Fact A.1. *There exists a distribution of random variables $g_1, \ldots, g_R$ such that each g_i is marginally $N(0,1)$, $\mathbb{E}[g_i g_j] = -1/(R-1)$ for all $i \neq j$, and $\sum_{i=1}^{R} g_i = 0$.*

Lemma A.2. *Let $\mathbf{g} = (g_1 \ldots g_R)^{\mathrm{T}}$ where $\{g_i\}_{i=1}^{R}$ are as given in Fact A.1, and suppose $\mathbf{x} = (x_1 \ldots, x_R)^{\mathrm{T}}$, $\mathbf{y} = (y_1 \ldots y_R)^{\mathrm{T}} \in \mathbb{R}^R$ are orthogonal unit vectors such that $\langle \mathbf{1}, \mathbf{x} \rangle = 0$ and $\langle \mathbf{1}, \mathbf{y} \rangle = 0$. Define, $f := \langle \mathbf{x}, \mathbf{g} \rangle$ and $h := \langle \mathbf{y}, \mathbf{g} \rangle$. Then, f and g are independent $N(0, R/(R-1))$ random variables.*

Proof. We have,

$$
\begin{aligned}
\mathbb{E}[f^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{R} x_i g_i\right)^2\right] \\
&= \sum_{i=1}^{R} x_i^2 \mathbb{E}[g_i^2] + \sum_{\substack{i,j \in [R] \\ i \neq j}} x_i x_j \mathbb{E}[g_i g_j] \\
&= \sum_{i=1}^{R} x_i^2 - \left(\frac{1}{R-1}\right) \sum_{\substack{i,j \in [R] \\ i \neq j}} x_i x_j \\
&= \left(1 + \frac{1}{R-1}\right) \sum_{i=1}^{R} x_i^2 - \left(\frac{1}{R-1}\right) \left(\sum_{i=1}^{R} x_i\right)^2 \\
&= \frac{R}{R-1}.
\end{aligned} \tag{90}
$$

The same holds for $\mathbb{E}[h^2]$. For the second part of the lemma observe that,

$$
\begin{aligned}
\mathbb{E}[fh] &= \sum_{i=1}^{R} \left[x_i y_i \mathbb{E}[g_i^2] + \sum_{\substack{j \in [R] \\ j \neq i}} \mathbb{E}[g_i g_j] x_i y_j \right] \\
&= \sum_{i=1}^{R} \left[x_i y_i - \left(\frac{1}{R-1}\right) \sum_{\substack{j \in [R] \\ j \neq i}} x_i y_j \right] \\
&= \left(1 + \frac{1}{R-1}\right) \langle \mathbf{x}, \mathbf{y} \rangle - \left(\frac{1}{R-1}\right) \langle \mathbf{x}, \mathbf{1} \rangle \langle \mathbf{y}, \mathbf{1} \rangle = 0.
\end{aligned} \tag{91}
$$

$\square$

Fact A.3 (Fact 3.4 in [DOSW11]). *Let $P : \mathbb{R}^\ell \to \mathbb{R}$ be a degree-d polynomial over independent standard normal variables which has at least one coefficient of magnitude at least α. Then, $\|P\|_2 \equiv \sqrt{\mathbb{E}[|P(\mathbf{x})|^2]}$ is at least $\frac{\alpha}{d^d \binom{\ell+d}{d}}$.*

B Comparing monomial and ℓ_2-masses

In this section, we relate the monomial mass of the polynomials with their ℓ_2-mass under the distribution $\mathcal{D}$.

Lemma B.1. *Let $Q(U_1, \ldots, U_T)$ be a polynomial of degree $d \geq 1$. Let $\tilde{Q}(W_{1,1}, \ldots, W_{1,T})$ be the polynomial obtained from $Q(U_1, \ldots, U_T)$ by the orthonormal transformation. With η and $T = 10d$ chosen as in Section 3, the following bounds hold:*

  1. $\|Q(U_1, \ldots, U_T)\|_2 \leq (20dT)^{5d} \|\tilde{Q}(W_{1,1}, \ldots, W_{1,T})\|_{\mathrm{mon},2}$

2. If Q depends only on variables $U_2, \ldots, U_T$ then $\|\tilde{Q}(W_{1,1}, \ldots, W_{1,T})\|_{\mathrm{mon},2} \leq (10dT)^{7d}\|Q(U_2, \ldots, U_T)\|_2$

Proof. For ease of notation, we shall denote variables $W_{11}, \ldots, W_{1T}$ by $W_1, \ldots, W_T$. Let $\mathscr{S}_{T,d}$ be the set of all multi-sets on $[T]$ of size at most d. Using the fact that $\binom{T}{d} \leq \left(\frac{Te}{d}\right)^d \leq (eT)^d$ we have $|\mathscr{S}_{T,d}| \leq (10T)^{2d}$

Proof of Part 1.: For the first direction let $Q(U_1, \ldots, U_T) = \sum_{S \in \mathscr{S}_{T,d}} c_S U_S$, where the monomial U_S is defined as $U_S = \prod_{i \in S} U_i^{S(i)}$. Therefore,

$$\|Q\|_2^2 \;=\; \mathbb{E}_{\mathcal{D}_\mathcal{I}}\left[\left(\sum_{S \in \mathscr{S}_{T,d}} c_S U_S\right)^2\right] \tag{92}$$

$$\leq\; \mathbb{E}_{\mathcal{D}_\mathcal{I}}\left[\left(\sum_{S \in \mathscr{S}_{T,d}} c_S^2\right)\left(\sum_{S \in \mathscr{S}_{T,d}} U_S^2\right)\right] \tag{93}$$

$$=\; \|Q(U_1, \ldots, U_T)\|_{\mathrm{mon},2}^2\left(\mathbb{E}_{\mathcal{D}_\mathcal{I}}\left[\sum_{S \in \mathscr{S}_{T,d}} U_S^2\right]\right) \tag{94}$$

For the first term, we claim that

$$\begin{aligned}
\|Q(U_1, \ldots, U_T)\|_{\mathrm{mon},2} \;&\leq\; \|Q(U_1, \ldots, U_T)\|_{\mathrm{mon},1}\\
&\leq\; (10T)^{3d}\|\tilde{Q}(W_1, \ldots, W_T)\|_{\mathrm{mon},1}\\
&\leq\; (10T)^{4d}\|\tilde{Q}(W_1, \ldots, W_T)\|_{\mathrm{mon},2}
\end{aligned}$$

where the first inequality follows the fact that ℓ_2-norm is upper bounded by the ℓ_1-norm, and the third inequality follows from *Cauchy-Schwarz* and $|\mathscr{S}_{T,d}| \leq (10T)^{2d}$. The middle inequality can be argued as follows. Consider $U_S = \prod_{i \in S} U_i^{S(i)}$. Then it can be expressed as in terms of $W_1, \ldots, W_T$ as

$$\prod_{i \in S}\left(\sum_{l \in [T]} a_{i,l} W_l\right)^{S(i)}$$

By construction, the linear transformation $\{U_1, \ldots, U_T\} \mapsto \{W_1, \ldots, W_T\}$ is *orthonormal* (See section 4.1.2). Therefore each coefficient satisfies $|a_{i,l}| \leq 1$. Furthermore, there can be at most T^d distinct terms in the expansion of U_S. Therefore, the total contribution to the coefficient of a fixed monomial from U_S can be at most $|c_S|T^d$. Repeating the argument across all $S \in \mathscr{S}_{T,d}$ completes the argument.

For upper bounding the expectation term in (94), fix a $S \in \mathscr{S}_{T,d}$. Then,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\mathcal{I}}}\Big[U_S^2\Big] &= \mathbb{E}_{\mathcal{D}_{\mathcal{I}}}\Big[\prod_{i \in S} U_i^{2S(i)}\Big] \\
&= \prod_{i \in S} \mathbb{E}_{\mathcal{D}_{\mathcal{I}}}\Big[U_i^{2S(i)}\Big] && \Big(\text{Since } U_1, \ldots, U_T \text{ are independent}\Big) \\
&\leq \prod_{i \in S \setminus \{1\}} \mathbb{E}_{\mathcal{D}_{\mathcal{I}}}\Big[U_i^{2S(i)}\Big] && \Big(\text{Since } \eta\sqrt{T} < 1\Big) \\
&\overset{1}{\leq} \prod_{i \in S \setminus \{1\}} (2S(i))! \\
&\leq (2|S|)!
\end{aligned}
$$

where step 1 follows from the well known fact that for $g \sim N(0,1)$, $\mathbb{E}[g^k] \leq k!$ for all $k \in \mathbb{Z}_+$. Therefore, plugging in the upper bounds in (94) we get

$$
\|Q(U_1, \ldots, U_T)\|_{\mathrm{mon},2}^2\Big(\mathbb{E}_{\mathcal{D}_{\mathcal{I}}}\Big[\sum_{S \in \mathscr{S}_{T,d}} U_S^2\Big]\Big) \leq (10T)^{10d}(2d)^{(2d)}\|\tilde{Q}(W_1, \ldots, W_T)\|_{\mathrm{mon},2}^2
$$

Proof of Part 2: For the second direction, we observe that

$$
\begin{aligned}
\|\tilde{Q}(W_1, \ldots, W_T)\|_{\mathrm{mon},2} &\leq \|\tilde{Q}(W_1, \ldots, W_T)\|_{\mathrm{mon},1} && (95) \\
&\overset{1}{\leq} (10T)^{3d}\|Q(U_2, \ldots, U_T)\|_{\mathrm{mon},1} && (96) \\
&\overset{2}{\leq} (10dT)^{7d}\|Q(U_2, \ldots, U_T)\|_2 && (97)
\end{aligned}
$$

where inequality 1 again can be argued similarly to the previous direction (using the fact that $\{W_1, \ldots, W_T\} \mapsto \{U_1, \ldots, U_T\}$ is again an orthonormal linear transformation).

For step 2, we write $Q(U_2, \ldots, U_T)$ in the monomial basis of U i.e., $Q(U_2, \ldots, U_T) = \sum_S c_S U_S$ and see that

$$
\Big\|\sum_S c_S U_S\Big\|_{\mathrm{mon},1} = \sum_{S \in \mathscr{S}_{T-1,d}} |c_S| \overset{1}{\leq} \sum_{S \in \mathscr{S}_{T-1,d}} (6Td)^{2d}\|Q(U_2, \ldots, U_T)\|_2 \leq (10dT)^{4d}\|Q(U_2, \ldots, U_T)\|_2 \quad (98)
$$

with step 1 following from Fact A.3, and the last inequality uses the upper bound on $|\mathscr{S}_{T,d}|$.

$\square$

C   Comparison inequalities between Norms

Claim C.1. *Given polynomials* $P_1(\mathbf{W}), P_2(\mathbf{W})$ *over variables* $\mathbf{W} = (W_{11}, \ldots, W_{1T})$, *we have*

$$
\|P_1(\mathbf{W})P_2(\mathbf{W})\|_{\mathrm{mon},2} \leq \|P_1(\mathbf{W})\|_{\mathrm{mon},1}\|P_2(\mathbf{W})\|_{\mathrm{mon},2}.
$$

Proof. Let $P_1(\mathbf{W}) = \sum_{W_S \in \mathscr{M}} c_S W_S$. Then,

$$
\begin{aligned}
\|P_1(\mathbf{W})P_2(\mathbf{W})\|_{\mathrm{mon},2} &= \left\| \sum_{W_S \in \mathscr{M}} c_S W_S P_2(\mathbf{W}) \right\|_{\mathrm{mon},2} \\
&\leq \sum_{W_S \in \mathscr{M}} |c_S| \|W_S P_2(\mathbf{W})\|_{\mathrm{mon},2} \\
&= \|P_1(\mathbf{W})\|_{\mathrm{mon},1} \|P_2(\mathbf{W})\|_{\mathrm{mon},2}
\end{aligned}
$$

$\square$