7-2010

Hardness Results for Agnostically Learning Low-Degree Polynomial Threshold Functions

Ilias Diakonikolas
Columbia University

Ryan O'Donnell
Carnegie Mellon University

Rocco A. Servedio
Columbia University

Yi Wu
Carnegie Mellon University

Hardness Results for Agnostically Learning Low-Degree Polynomial Threshold Functions

Ilias Diakonikolas*
Columbia University
ilias@cs.columbia.edu

Ryan O'Donnell†
CMU
odonnell@cs.cmu.edu

Rocco A. Servedio‡
Columbia University
rocco@cs.columbia.edu

Yi Wu§
CMU
wuyish@gmail.com

July 18, 2010

Abstract

Hardness results for maximum agreement problems have close connections to hardness results for proper learning in computational learning theory. In this paper we prove two hardness results for the problem of finding a low degree polynomial threshold function (PTF) which has the maximum possible agreement with a given set of labeled examples in $\mathbb{R}^n \times \{-1, 1\}$. We prove that for any constants $d \geqslant 1, \epsilon > 0$,

- Assuming the Unique Games Conjecture, no polynomial-time algorithm can find a degree-d PTF that is consistent with a $(\frac{1}{2} + \epsilon)$ fraction of a given set of labeled examples in $\mathbb{R}^n \times \{-1, 1\}$, even if there exists a degree-d PTF that is consistent with a $1 - \epsilon$ fraction of the examples.
- It is NP-hard to find a degree-2 PTF that is consistent with a $(\frac{1}{2}+\epsilon)$ fraction of a given set of labeled examples in $\mathbb{R}^n \times \{-1, 1\}$, even if there exists a halfspace (degree-1 PTF) that is consistent with a $1 - \epsilon$ fraction of the examples.

These results immediately imply the following hardness of learning results: (i) Assuming the Unique Games Conjecture, there is no better-than-trivial proper learning algorithm that agnostically learns degree-d PTFs under arbitrary distributions; (ii) There is no better-than-trivial learning algorithm that outputs degree-2 PTFs and agnostically learns halfspaces (i.e. degree-1 PTFs) under arbitrary distributions.

*Research supported by NSF grants CCF-0728736, CCF-0525260, and by an Alexander S. Onassis Foundation Fellowship.
†Supported by NSF grants CCF-0747250 and CCF-0915893, BSF grant 2008477, and Sloan and Okawa fellowships.
‡Supported by NSF grants CCF-0347282, CCF-0523664 and CNS-0716245, and by DARPA award HR0011-08-1-0069.
§Supported by NSF grants CCF-0747250 and CCF-0915893, BSF grant 2008477, and Sloan and Okawa fellowships.

1 Introduction

A *polynomial threshold function* (PTF) of degree d is a function $f : \mathbb{R}^n \to \{-1, +1\}$ of the form $f(x) = \text{sign}(p(x))$, where

$$p(x) = \sum_{\text{multiset } S \subseteq [n], |S| \leqslant d} c_S \prod_{i \in S} x_i$$

is a degree-d multivariate polynomial with real coefficients. Degree-1 PTFs are commonly known as *halfspaces* or *linear threshold functions*, and have been intensively studied for decades in fields as diverse as theoretical neuroscience, social choice theory and Boolean circuit complexity.

The last few years have witnessed a surge of research interest and results in theoretical computer science on halfspaces and low-degree PTFs, see e.g. [25, 23, 7, 8, 10, 6, 15]. One reason for this interest is the central role played by low-degree PTFs (and halfspaces in particular) in both practical and theoretical aspects of *machine learning*, where many learning algorithms either implicitly or explicitly use low-degree PTFs as their hypotheses. More specifically, several widely used linear separator learning algorithms such as the Perceptron algorithm and the "maximum margin" algorithm at the heart of Support Vector Machines output halfspaces as their hypotheses. These and other halfspace-based learning methods are commonly augmented in practice with the "kernel trick," which makes it possible to efficiently run these algorithms over an expanded feature space and thus potentially learn from labeled data that is not linearly separable in $\mathbb{R}^n$. The "polynomial kernel" is a popular kernel to use in this way; when, as is usually the case, the degree parameter in the polynomial kernel is set to be a small constant, these algorithms output hypotheses that are equivalent to low-degree PTFs. Low-degree PTFs are also used as hypotheses in several important learning algorithms with a more complexity-theoretic flavor, such as the low-degree algorithm of Linial *et al.* [21] and its variants [12, 22], including some algorithms for distribution-specific agnostic learning [14, 20, 3, 6].

Given the importance of learning algorithms that construct low-degree PTF hypotheses, it is a natural goal to study the limitations of learning algorithms that work in this way. On the positive side, it is well known that if there is a PTF (of constant degree d) that is consistent with *all* the examples in a data set, then a consistent hypothesis can be found in polynomial time simply by using linear programming (with the $\Theta(n^d)$ monomials of degree at most d as the variables in the LP). However, the assumption that some low-degree PTF correctly labels all examples seems quite strong; in practice data is often noisy or too complex to be consistent with a simple concept. Thus we are led to ask: if no low-degree PTF classifies an entire data set perfectly, to what extent can the data be learned using low-degree PTF hypoptheses?

In this paper, we address this question under the agnostic learning framework [11, 16]. Roughly speaking, a function class $\mathcal{C}$ is agnostically learnable if we can efficiently find a hypothesis that has accuracy arbitrarily close to the accuracy of the best hypothesis in $\mathcal{C}$. Uniform convergence results [11] imply that learnability in this model is essentially equivalent to the ability to come up with a hypothesis that correctly classifies almost as many examples as the optimal hypothesis in the function class. This problem is sometimes referred to as a "Maximum Agreement" problem for $\mathcal{C}$. As we now describe, this problem has previously been well studied for the class $\mathcal{C}$ of halfspaces.

Related Work. The Maximum Agreement problem for halfspaces over $\mathbb{R}^n$ was shown to be NP-hard to approximate within some constant factor in [1, 2]. The inapproximability factor was improved to $84/85 + \epsilon$ in [4], which showed that this hardness result applies even if the examples must lie on the n-dimensional Boolean hypercube. Finally, a tight inapproximability result was established independently in [10] and [7]; these works showed that for any constant $\epsilon > 0$, it is NP-hard to find a halfspace consistent with $(\frac{1}{2} + \epsilon)$ of the examples even if there exists a halfspace consistent with $(1 - \epsilon)$ of the examples. (It is trivial to find a halfspace consistent with half of the examples since either the constant-0 or constant-1 halfspace will suffice.) The reduction in [7] produced examples with real-valued coordinates, whereas the proof in [10] yielded examples that lie on the Boolean hypercube.

Thanks to these results the Maximum Agreement problem is well-understood for halfspaces, but the situation is very different for low-degree PTFs. Even for degree-2 PTFs no hardness results were previously known, and recent work [6] has in fact given efficient agnostic learning algorithms for low-degree PTFs under specific distributions on examples such as Gaussian distributions or the uniform distribution over $\{-1,1\}^n$ (though it should be noted that these distribution-specific agnostic learning algorithms for degree-d PTFs are not proper – they output PTF hypotheses of degree $\gg d$). In this paper we make the first progress on this problem, by establishing strong hardness of approximation results for the Maximum Agreement problem for low-degree PTFs. Our results directly imply corresponding hardness results for agnostically learning low degree PTFs under arbitrary distributions; we present all these results below.

Main Results. Our main results are the following two theorems. The first result establishes UGC-hardness of finding a nontrivial degree-d PTF hypothesis even if some degree-d PTF has almost perfect accuracy:

Theorem 1.1. *Fix $\epsilon > 0$, $d \geqslant 1$. Assuming the Unique Games Conjecture, no polynomial-time algorithm can find a degree-d PTF that is consistent with $(\frac{1}{2} + \epsilon)$ fraction of a given set of labeled examples in $\mathbb{R}^n \times \{-1,1\}$, even if there exists a degree-d PTF that is consistent with a $1-\epsilon$ fraction of the examples.*

The second result shows that it is NP-hard to find a degree-2 PTF hypothesis that has nontrivial accuracy even if some halfspace has almost perfect accuracy:

Theorem 1.2. *Fix $\epsilon > 0$. It is NP-hard to find a degree-2 PTF that is consistent with $(\frac{1}{2} + \epsilon)$ fraction of a given set of labeled examples in $\mathbb{R}^n \times \{-1,1\}$, even if there exists a halfspace (degree-1 PTF) that is consistent with a $1-\epsilon$ fraction of the examples.*

As noted above, both problems become easy (using linear programming) if the best hypothesis is assumed to have perfect agreement with the data set rather than agreement $1-\epsilon$, and it is trivial to find a (constant-valued) hypothesis with agreement rate $1/2$ for any data set. Thus the parameters in both hardness results are essentially the best possible.

These results can be rephrased as hardness of agnostic learning results in the following way: (i) Assuming the Unique Games Conjecture, even if there exists a degree-d PTF that is consistent with $1-\epsilon$ fraction of the examples, there is no efficient *proper* agnostic learning algorithm that can output a degree-d PTF correctly labeling more than $\frac{1}{2} + \epsilon$ fraction of the examples; (ii) Assuming $\mathrm{P} \neq \mathrm{NP}$, even if there exists a halfspace that is consistent with $1-\epsilon$ fraction of the examples, there is no efficient agnostic learning algorithm that can find a degree-2 PTF correctly labeling more than $\frac{1}{2} + \epsilon$ fraction of the examples.

Organization. In Appendix A we present the complexity-theoretic basis (the Unique Games conjecture and the NP-hardness of Label Cover) of our hardness results. In Section 2 we sketch a new proof of the hardness of the Maximum Agreement problem for halfspaces, and give an overview of how the proofs of Theorems 1.1 and 1.2 build on this basic argument. In Sections 3 and 4 we prove Theorems 1.1 and 1.2.

Notational Preliminaries: For $n \in \mathbb{Z}_+$ we denote by $[n]$ the set $\{1,\ldots,n\}$. For $i,j \in \mathbb{Z}_+$, $i \leqslant j$, we denote by $[i,j]$ the set $\{i, i+1, \ldots, j\}$. We write $\{j:m\}$ to denote the multi-set that contains m copies of the element j. We write $\chi_S(x)$ to denote $\prod_{i \in S} x_i$, the monomial corresponding to the multiset S.

2 Overview of our arguments

To illustrate the structure of our arguments, let us begin by sketching a proof of the following hardness result for the Maximum Agreement problem for halfspaces:

Proposition 2.1. *Assuming the Unique Games Conjecture, no polynomial-time algorithm can find a halfspace (degree-1 PTF) that is consistent with $(\frac{1}{2} + \epsilon)$ fraction of a given set of labeled examples in $\mathbb{R}^n \times \{-1,1\}$, even if there exists a halfspace that is consistent with a $1-\epsilon$ fraction of the examples.*

As mentioned above, the same hardness result (based only on the assumption that $\mathrm{P} \neq \mathrm{NP}$) has already been established in [7, 10]; indeed, we do not claim Proposition 2.1 as a new result. However, the argument sketched below is is different from (and, we believe, simpler than) the other proofs; it helps to illustrate how we eventually achieve the more general hardness results Theorems 1.1 and 1.2.

Proof Sketch for Proposition 2.1: We describe a reduction that maps any instance $\mathcal{L}$ of Unique Games to a set of labeled examples with the following guarantee: if $\mathrm{Opt}(\mathcal{L})$ is very close to 1 then there is a halfspace that agrees with $1-\epsilon$ fraction of the examples, while if $\mathrm{Opt}(\mathcal{L})$ is very close to 0 then no halfspace agrees with more than $\frac{1}{2}+\epsilon$ fraction of the examples. A reduction of this sort directly yields Proposition 2.1.

Let $\mathcal{L} = (U, V, E, k, \Pi)$ be a Unique Games instance. Each example generated by the reduction has $(|V|+|U|)k$ coordinates, i.e. the examples lie in $\mathbb{R}^{(|U|+|V|)k}$. The coordinates should be viewed as being grouped together in the following way: there is a block of k coordinates for each vertex w in $U \cup V$. We index the coordinates of $x \in \mathbb{R}^{(|U|+|V|)k}$ as $x = (x_w^{(i)})$ where $w \in U \cup V$ and $i \in [k]$.

Given any function $f : \mathbb{R}^{(|U|+|V|)k} \to \{-1, 1\}$ and vertex $w \in U \cup V$, we write f_w to denote the restriction of f to the k coordinates $(x_w^{(i)})_{i\in[k]}$ that is obtained by setting all other coordinates $(x_{w'}^{(j)})_{w'\neq w}$ to 0. Similarly, for $e = \{u, v\}$ an edge in $U \times V$, we write f_e for the restriction that fixes all coordinates $(x_{w'}^{(i)})_{w'\notin e}$ to 0 and leaves the $2k$ coordinates $x_u^{(i)}, x_v^{(i)}$ unrestricted.

For every labeling $\ell : U \cup V \to [k]$ of the instance, there is a corresponding halfspace over $\mathbb{R}^{(|V|+|U|)k}$

$$\mathrm{sign}\Big(\sum_{u\in U} x_u^{(\ell(u))} - \sum_{v\in V} x_v^{(\ell(v))}\Big).$$

Given a Unique Games instance $\mathcal{L}$, the reduction constructs a distribution $\mathcal{D}$ over labeled examples such that if $\mathrm{Opt}(\mathcal{L})$ is almost 1 then the above halfspace has very high accuracy w.r.t. $\mathcal{D}$, and any halfspace that has accuracy at least $\frac{1}{2}+\epsilon$ yields a labeling that satisfies a constant fraction of edges in $\mathcal{L}$. A draw from $\mathcal{D}$ is obtained by first selecting a uniform random edge $e = \{u, v\}$ from E, and then making a draw from $\mathcal{D}_e$, where $\mathcal{D}_e$ is a distribution over labeled examples that we describe below.

Fix an edge $e = (u, v)$. For the sake of exposition, let us assume the mapping $\pi^e \in \Pi$ associated with e is the identity permutation, i.e. $\pi^e(i) = i$ for every $i \in [k]$. The distribution $\mathcal{D}_e$ will have the following properties:

(i) For every (y, b) in the support of $\mathcal{D}_e$, all coordinates $y_w^{(i)}$ for every vertex $w \notin e$ are zero.

(ii) For every label $i \in [k]$, the halfspace $\mathrm{sign}(x_u^{(i)} - x_v^{(i)})$ has accuracy $1-\epsilon$ w.r.t.$\mathcal{D}_e$.

(iii) If $\mathrm{sign}(f_e)$ is a halfspace that has accuracy at least $\frac{1}{2}+\epsilon$ w.r.t. $\mathcal{D}_e$, then the functions f_u, f_v can each be individually "decoded" to a "small" (constant-sized) set $S_u, S_v \subseteq [k]$ of labels such that $S_u \cap S_v \neq \emptyset$ (so a labeling that satisfies a nonnegligible fraction of edges in expectation can be obtained simply by choosing a random label from S_w for each w – such a random choice will satisfy each edge's bijection with constant probability, so in expectation will satisfy a constant fraction of constraints).

Let us explain item (iii) in more detail. Since the distribution $\mathcal{D}_e$ is supported on vectors y that have the $(y_w^{(i)})_{w\notin e}$ coordinates all 0, the distribution $\mathcal{D}_e$ only "looks at" the restriction f_e of f, which is a halfspace on $\mathbb{R}^{2k}$. Thus achieving (iii) can be viewed as solving a kind of property testing problem which may loosely be described as "Matching dictatorship testing for halfspaces." To be more precise, what is required is a distribution $\mathcal{D}_e$ over $2k$-dimensional labeled examples and a "decoding" algorithm A which takes as input a k-variable halfspace and outputs a set of coordinates. Together these must have the following properties:

- (Completeness) If $f_e(x) = x_u^{(i)} - x_v^{(i)}$ then $\mathrm{sign}(f_e(y)) = b$ with probability $1-\epsilon$ for $(y, b) \sim \mathcal{D}_e$;

- (Soundness) If f_e is such that $\mathrm{sign}(f_e(y)) = b$ with probability at least $1/2 + \epsilon$ for (y, b) drawn from $\mathcal{D}_e$, then the output sets $A(f_u)$, $A(f_v)$ of the decoding algorithm (when it is run on f_u and f_v respectively) are two small sets that intersect each other.

Testing problems of this general form are often referred to as *Dictatorship Testing*; the design and analysis of such tests is a recurring theme in hardness of approximation.

We give a "matching dictatorship test for halfspaces" below. More precisely, in the following figure we describe the distribution $\mathcal{D}_e$ over examples (the decoding algorithm A is described later).

$\mathcal{T}_1$: Matching Dictatorship Test for Halfspaces

Input: A halfspace $f_e : \mathbb{R}^{2k} \to \mathbb{R}$.

Set $\epsilon := \frac{1}{\log k}$, $\delta := 1/2^k$.

1. Generate independent 0/1 bits $a_1, a_2, \ldots, a_k$ each with $\mathbf{E}[a_i] = \epsilon$. Generate $2k$ independent $N(0,1)$ Gaussian random variables: $h_1, h_2 \ldots, h_k, g_1, g_2 \ldots, g_k$. Generate a random bit $b \in \{-1, 1\}$.
2. Set $r = (a_1 h_1 + g_1, \ldots, a_k h_k + g_k, g_1, \ldots, g_k)$ and $\omega = (1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{2k}$ to be the vector whose first k coordinates are 1 and last k coordinates are 0.
3. Set $y = r + b\delta\omega$. The result of a draw from $\mathcal{D}_e$ is the labeled example (y, b).

The test checks whether $\mathrm{sign}(f_e(y))$ equals b.

It is useful to view the test in the following light: Let us write $f_e(x)$ as $\theta + \sum_{i=1}^k w_u^{(i)} x_u^{(i)} + \sum_{i=1}^k w_v^{(i)} x_v^{(i)}$, and let us suppose that $\sum_{i=1}^k |w_u^{(i)}| = 1$ (as long as some $w_u^{(i)}$ is nonzero this is easily achieved by rescaling; for this intuitive sketch we ignore the case that all $w_u^{(i)}$ are 0, which is not difficult to handle). Then we have $f_e(y) = f_e(r) + b\delta$, and we may view the test as randomly choosing one of the two inequalities $f_e(r) - \delta < 0$, $f_e(r) - \delta > 0$ and checking that it holds. Since at least one of these inequalities must hold for every f_e, the probability that f_e passes the test is is $\frac{1}{2} + \frac{1}{2}\mathbf{Pr}_r[f_e(r) \in [-\delta, \delta)]$. This interpretation will be useful both for analyzing completeness and soundness of the test.

For completeness, it is easy to see that the "matching dictator" function $f_e(x) = x_u^{(i)} - x_v^{(i)}$ has $f_e(r) = a_i h_i$ and thus $\mathbf{Pr}[f_e(r) = 0] = 1 - \epsilon$, so this function indeed passes the test with probability $1 - \epsilon$.

The soundness analysis, which we now sketch, is more involved. Let f be such that $\mathbf{Pr}_r[f_e(r) \in [-\delta, \delta)] \geqslant 2\epsilon$. Since $f_e(r) = \sum_i (w_u^{(i)} + w_v^{(i)}) g_i + \sum w_u^{(i)} a_i h_i$ and g_i, h_i are i.i.d. Gaussians, conditioned on a given outcome of the a_i-bits the value $f_e(r)$ follows the Gaussian distribution with mean 0 and variance $\sum (w_u^{(i)} + w_v^{(i)})^2 + \sum (a_i w_u^{(i)})^2$. Now recall that an $N(0, \sigma)$ Gaussian random variable lands in the interval $[-t, t]$ with probability at most $O(t/\sigma)$. So any a-vector for which the variance $\sum (w_u^{(i)} + w_v^{(i)})^2 + \sum (a_i w_u^{(i)})^2$ is not "tiny" can contribute only a negligible amount to the overall probability that $f_e(r)$ lies in $[-\delta, \delta)$ (recall that δ is extremely tiny). Since by assumption $\mathbf{Pr}_r[f_e(r) \in [-\delta, \delta)]$ is non-negligible (at least 2ϵ), there must be a non-negligible fraction of a-vector outcomes that make the variance $\sum (w_u^{(i)} + w_v^{(i)})^2 + \sum (a_i w_u^{(i)})^2$ be "tiny." This implies that there must be only a "few" coordinates $w_u^{(j)}$ for which $|w_u^{(j)}|$ is not tiny (for if there were many non-tiny $w_u^{(j)}$ coordinates, then $\sum_i (w_u^{(i)} a_i)^2$ would be non-tiny with probability nearly 1 over the choice of the a-vector). Moreover, $w_u^{(i)} + w_{(i)}^v$ must be ≈ 0 for each i, so for each i the magnitudes $|w_u^{(i)}|$ and $|w_v^{(i)}|$ must be nearly equal; and in particular, each $|w_u^{(i)}|$ is large if and only if $|w_v^{(i)}|$ is large. Finally, since $\sum_i |w_u^{(i)}|$ equals 1 some $w_u^{(i)}$'s must be large (at least $1/k$).

With these facts in place, the appropriate decoding algorithm A is rather obvious: given $f_u = \theta + \sum_{i=1}^{k} w_u^{(i)} x_u^{(i)}$ as input, A outputs the set S_u of those coordinates i for which $|w_u^{(i)}|$ is large (and similarly for f_v). This set cannot be too large since $\sum_{i=1}^{k} |w_u^{(i)}|$ equals 1. Now a labeling that satisfies edge e with non-negligible probability can be obtained by outputing a random element from S_u and a random element from S_v; since these sets are small there is a non-negligible probability that the labels will match as required. This concludes the proof sketch of Proposition 2.1. □

Overview of the proofs of Theorems 1.1 and 1.2. For Theorem 1.1 (hardness of properly learning degree-d PTFs), we must deal with the additional complication of handling the cross-terms such as $x_u^{(i)} x_v^{(j)}$ between u-variables and v-variables that may be present in degree-d PTFs. As an example of how such cross-terms can cause problems, observe that the degree-3 polynomial $f_e = (x_u^{(i)} - x_v^{(i)}) \sum (x_u^{(i)})^2$ would pass the test $\mathcal{T}_1$ with high probability, but this polynomial has $f_v = 0$ so there is no way to successfully "decode" a good label for v. To get around this, we modify the test $\mathcal{T}_1$ to set $y = (a_1 h_1 + g_1^d + b\delta, a_2 h_2 + g_2^d + b\delta, \ldots, a_k h_k + g_k^d + b\delta, g_1, \ldots, g_k)$; intuitively this modified test checks whether the polynomial f_e is of the form $x_u^{(i)} - (x_v^{(i)})^d$. The bulk of our work is in analyzing the soundness of this test; we show that any polynomial f_e that passes the modified test with probability significantly better than $1/2$ must have almost no coefficient weight on cross-terms, and that in fact the restricted polynomials f_u, f_v can each be decoded to a small set in such a way that there is a matching pair as desired. We give a complete description and analysis of our Dictator Test and prove Theorem 1.1 in Section 3.

For Theorem 1.2, a first observation is that the test $\mathcal{T}_1$ in fact already has soundness $3/4 + \epsilon$ for degree-2 PTFs. To see this, we begin by writing the degree-2 polynomial $f_e(x)$ as $\theta + f_1(x) + f_2(x)$ where $f_1(x)$ is the linear (degree 1) part and $f_2(x)$ is the quadratic (degree 2) part (note that f_1 is an odd function and f_2 is an even function). We next observe that since any vector r is generated with the same probability as $-r$, the test may be viewed as randomly selecting one of the following 4 inequalities to verify: $f_e(r + \delta\omega) > 0$, $f_e(r - \delta\omega) < 0$, $f_e(-r + \delta\omega) > 0$, $f_e(-r - \delta\omega) < 0$. If all four inequalities hold, then combining $f_e(r + \delta\omega) > 0$ with $f_e(-r - \delta\omega) < 0$ we get that $f_1(r + \delta\omega) > 0$ and combining $f_e(r - \delta\omega) < 0$ with $f_e(-r + \delta\omega) > 0$ we get $f_1(r - \delta\omega) < 0$. Consequently, if a degree-2 polynomial f_e passes the test with probability $3/4 + \epsilon$, then by an averaging argument, for at least an ϵ fraction of the r-outcomes all four of the inequalities must hold. This implies that for an ϵ fraction of the r's we must have $f_1(r + \delta\omega) > 0$ and $f_1(r - \delta\omega) < 0$, and so the degree-1 PFT f_1 must pass the Dictator Test $\mathcal{T}_1$ with probability at least $1/2 + \epsilon$. This essentially reduces to the problem of testing degree-1 PTFs, whose analysis is sketched above.

To get the soundness down to $1/2$ more work has to be done. Roughly speaking, we modify the test by checking that $\mathrm{sign}(f(k_1 r + k_2 \delta\omega)) = \mathrm{sign}(k_2)$ for k_1, k_2 generated from a carefully constructed distribution in which k_1, k_2 can assume many different possible orders of magnitude. Using these many different possibilities for the magnitudes of k_1, k_2, a careful analysis (based on carefully combining inequalities in a way that is similar to the previous paragraph, though significantly more complicated) shows that if a polynomial passes the test with probability $1/2 + \epsilon$ fraction then it can be "decoded" to a small set of coordinates. In addition to this modification, to avoid using the Unique Games Conjecture we employ the "folding trick" that is proposed in [9, 19] to ensure consistency across different vertices. One benefit of using this trick is that with it, we only need to design a test on one vertex instead of an edge.[1] The complete proof of Theorem 1.2 appears in Section 4.

[1] The reason that we can not use "folding" for our first result on low-degree PTFs, roughly speaking, is that such a folding does not seem able to handle cross-terms of degree greater than 2.

3 Hardness of proper learning noisy degree-d PTFs: Proof of Theorem 1.1

3.1 Dictator Test

Let $f : \mathbb{R}^{2n} \to \mathbb{R}$ be a $2n$-variable degree-d polynomial over the reals. The key gadget in our UG–hardness reduction is a *dictator test* of whether f is of the form $\mathrm{sign}(x_i - x_{n+i}^d)$ for some $i \in [n]$. More concretely, our dictator test queries the value of f on a *single* point $y \in \mathbb{R}^{2n}$ and decides to accept or reject based on the value $\mathrm{sign}(f(y))$.

$\mathcal{T}_d$: Matching Dictator Test for degree-d PTFs

Input: A degree-d real polynomial $f : \mathbb{R}^{2n} \to \mathbb{R}$.
Set $\beta := 1/\log n$ and $\delta := 2^{-n^2}$.

1. Generate n i.i.d. bits $a_i \in \{0,1\}$ with $\mathbf{Pr}[a_i = 1] = \beta$, $i \in [n]$. Generate $2n$ i.i.d. $N(0,1)$ Gaussians $\{h_i, g_i\}_{i=1}^n$. Generate a uniform random bit $b \in \{-1, 1\}$.
2. Set $y = (y_i)_{i=1}^{2n}$ where $y_i = a_i h_i + g_i^d + b\delta$ and $y_{n+i} = g_i$, $i \in [n]$.
3. Accept iff $\mathrm{sign}(f(y)) = b$.

We can now state and prove the properties of our test. The completeness is straightforward.

Lemma 3.1 (Completeness)**.** *The polynomial $f(x) = x_i - x_{n+i}^d$ passes the test with probability at least $1 - \beta$.*

Proof. Note that $f(y) = a_i h_i + b\delta$. Hence if $a_i = 0$ we have $\mathrm{sign}(f(y)) = b$ and this happens with probability $1 - \beta$. □

To state the soundness lemma we need some more notation. For a degree-d polynomial $f(x) = \sum_{S \subseteq [n], |S| \leqslant d} c_S \cdot \chi_S(x)$ we denote $\mathrm{wt}(f) = \sum_{S \neq \emptyset} |c_S|$. For $\theta > 0$, we define $I_\theta(f) := \{i \in [n] \mid \exists S \ni i \text{ s.t. } |c_S| \geqslant \theta \cdot \mathrm{wt}(f)/\binom{n+d}{d}\}$. Note that for $\theta \in [0,1]$ we have that $I_\theta(f) \neq \emptyset$, since there are $\binom{n+d}{d}$ nonempty monomials of degree at most d over $x_1, \ldots, x_n$.

Let $f : \mathbb{R}^{2n} \to \mathbb{R}$ be a $2n$-variable polynomial $f(x) = \sum_{S \subseteq [2n], |S| \leqslant d} c_S \cdot \chi_S(x)$ fed as input to our test. We will consider the restrictions obtained from f by setting the first (resp. second) half of the variables to 0. In particular, for $x = (x_1, \ldots, x_{2n})$ we shall denote $f_1(x_1, \ldots, x_n) = f(x_1, \ldots, x_n, \mathbf{0}_n)$ and $f_2(x_{n+1}, \ldots, x_{2n}) = f(\mathbf{0}_n, x_{n+1}, \ldots, x_{2n})$.

We are now ready to state our soundness lemma. The proof of this lemma poses significant complications and constitutes the bulk of the analysis in this section.

Lemma 3.2 (Soundness)**.** *Suppose that $f(x) = \sum_{S \subseteq [2n], |S| \leqslant d} c_S \cdot \chi_S(x)$ passes the test with probability at least $1/2 + \beta$. Then for f_1, f_2 as defined above, we have $|I_{0.5}(f_1)| \leqslant 1/\beta^2, |I_1(f_2)| \leqslant 1/\beta^2$. In addition, every $i \in [n]$ such that $n + i \in I_1(f_2)$ also satisfies $i \in I_{0.5}(f_1)$.*

Proof. We can assume that $\mathrm{wt}(f) > 0$, since otherwise f is a constant function, hence passes the test with probability exactly $\frac{1}{2}$. Since our test is invariant under scaling, we can further assume that $\mathrm{wt}(f) = 1$.

Let $x \in \mathbb{R}^{2n}$. By definition, $f_1(x) = \sum_{S \subseteq [n]} c_S \cdot \chi_S(x)$ and $f_2(x) = \sum_{S \subseteq [n+1, 2n]} c_S \cdot \chi_S(x)$. We can write

$$f(x) = f_1(x) + f_2(x) + f_{12}(x)$$

where $f_{12}(x) = \sum_{S\subseteq[2n],S\cap[n]\neq\emptyset,S\cap[n+1,2n]\neq\emptyset} c_S \cdot \chi_S(x)$.

Let us start by giving a very brief overview of the argument. The proof proceeds by carefully analyzing the structure of the coefficients c_S for the subfunctions f_1, f_2, f_{12}. In particular, we show that the total weight of the cross terms (i.e. $\mathrm{wt}(f_{12})$) is negligible, and that the weight of f is roughly equally spread among f_1 and f_2. Moreover, the coefficients of f_1, f_2 are either themselves negligible or "matching" (see equations (i)-(iv) on p. 9). Once these facts have been established, it is not hard to complete the proof.

The main step towards achieving this goal is to relate the coefficients c_S with the coefficients of an appropriately chosen restriction of f, obtained by carefully choosing an appropriate value of $a \in \{0,1\}^n$. We start with the following crucial claim:

Claim 3.3. *Suppose f passes the test with probability at least $1/2+\beta$. Then there exists $\alpha' \in \{0,1\}^n$ such that*

$$\|f_{\alpha'}\|_2 \leqslant 2^{-n} \cdot \log^{d^2} n.$$

Proof of Claim 3.3. Let us start be giving an equivalent description of the test. Denote $\omega = (\mathbf{1}_n, \mathbf{0}_n) \in \mathbb{R}^{2n}$, $r = (r_i)_{i=1}^{2n}$ with $r_i = a_i h_i + g_i^d$ and $r_{n+i} = g_i$, $i \in [n]$. Note that $y = r + (b\delta)\omega$. Then the Dictator Test $\mathcal{T}_d$ is as follows:

- Generate r, and with probability $1/2$, test whether $f(r+\delta\omega) \geqslant 0$; otherwise test $f(r-\delta\omega) < 0$.

Hence, since f passes with probability $1/2+\beta$, with probability at least 2β over the choice of r, the following inequalities are simultaneously satisfied:

$$f(r+\delta\omega) \geqslant 0; f(r-\delta\omega) < 0.$$

We now upper bound $|f(r+\delta\omega) - f(r)|$:

$$\begin{aligned}\big|f(r+\delta\omega) - f(r)\big| &= \Big|\textstyle\sum_{|S|\leqslant d} c_S \cdot \Big(\prod_{i\in S\cap[n]}(r_i+\delta)\cdot\prod_{j\in S\cap[n+1,2n]} r_j - \prod_{i\in S} r_i\Big)\Big| \\ &\leqslant \textstyle\sum_{1\leqslant|S|\leqslant d}|c_S|\cdot\Big(\sum_{\emptyset\neq T\subseteq S\cap[n]}\delta^{|T|}\cdot\prod_{i\in S\setminus T}|r_i|\Big) \leqslant \sum_{1\leqslant|S|\leqslant d}|c_S|\cdot 2^{|S|}\cdot\Big(\delta\cdot\prod_{i\in S: r_i\geqslant 1}|r_i|\Big)\end{aligned}$$

The last inequality follows from the fact that there are at most $2^{|S|}$ terms in the second summation each bounded from above by $\delta\cdot\prod_{i\in S:r_i\geqslant 1}|r_i|$.

We now claim that with probability at least $1-n^{-1}$ over the choice of r it holds $M := \max_{i\in[2n]}|r_i| \leqslant \log^d n$. To see this note that if $\max_{i\in[n]}\{|g_i|,|h_i|\} \leqslant c$ then $M \leqslant 2c^d$. Now recall that for $g \sim N(0,1)$ and $c > 2$ we have $\mathbf{Pr}[|g| > c] \leqslant e^{-c^2/2}$. The claim follows by fixing $c = \Theta(\log^{1/2} n)$ and taking a union bound over the corresponding $2n$ events.

Therefore, with probability $1-n^{-1}$ over the choice of r, we have

$$\big|f(r+\delta\omega) - f(r)\big| \leqslant \delta\cdot 2^d\cdot(\log n)^{d^2}\cdot\mathrm{wt}(f) \leqslant 2^{-n}.$$

Analogously we obtain that $|f(r) - f(r-\delta\omega)| \leqslant 2^{-n}$. We conclude that with probability $2\beta - n^{-1} \geqslant \beta$ over r

$$\big|f(r)\big| \leqslant 2^{-n}. \tag{1}$$

Recall that r is a random vector that depends on a, g, h. For every realization of $a \in \{0,1\}^n$, we denote the corresponding restriction of f as $f_a(g,h)$; note that $f_a(g,h)$ is a degree d^2 real polynomial over Gaussian random variables. Let us denote $\|f_a\|_2 := \mathbf{E}_{g,h}[f_a(g,h)^2]^{1/2}$.

At this point we appeal to an analytic fact from [5]: low degree polynomials over independent Gaussian inputs have good anti-concentration. In particular, an application of Theorem B.2 for $f_a(g,h)$ yields that for all $a \in \{0,1\}^n$ it holds

$$\mathbf{Pr}_{g,h}[|f_a(g,h)| \leqslant 2^{-n}] \leqslant d^2\cdot(2^{-n}/\|f_a\|_2)^{1/d^2}.$$

Combined with (1) this gives

$$\beta \leqslant \mathbf{Pr}_{a,g,h}\left[|f_a(g,h)| \leqslant 1/2^n\right] \leqslant \mathop{\mathbf{E}}_{a}\left[d^2 \cdot (2^{-n}/\|f_a\|_2)^{1/d^2}\right].$$

Now let us fix $a' := \arg\min_{a\in\{0,1\}^n} \|f_a\|_2$; the above relation implies $\left(2^{-n}/\|f_{a'}\|_2\right)^{1/d^2} \geqslant \beta$ or $\|f_{a'}\|_2 \leqslant 2^{-n}(1/\beta)^{d^2}$ as desired. This completes the proof of Claim 3.3. □

Since a' is fixed, we can express $f_{a'}$ as a degree-d^2 polynomial over the g_i's and h_i's. Let us write

$$f_{a'} = \textstyle\sum_{T,T'} w_{T,T'} \cdot \prod_{i\in T} g_i \cdot \prod_{i\in T'} h_i$$

where $T, T' \subseteq [n]$ are multi-sets satisfying $|T| + |T'| \leqslant d^2$ and $w_{T,T'} = w_{T,T'}(a')$. Since $f_{a'}$ has small variance, intuitively each of its coefficients should also be small. The following simple fact establishes such a relationship:

Fact 3.4. *Let $f : \mathbb{R}^l \to R$ be a degree-d polynomial $f(x) = \sum_{|S|\leqslant d} c_S \cdot \chi_S(x)$ and $\mathcal{G} \sim N(0,1)^l$. For all $T \subseteq [l]$ we have $\|f(\mathcal{G})\|_2 \geqslant d^{-d} \cdot |c_T|/\binom{l+d}{d}$.*

Proof of Fact 3.4. The fact follows by expressing f in an appropriate orthonormal basis. Let $\{H_S\}_{S\subseteq[l],|S|\leqslant d}$ be the set of Hermite polynomials of degree at most d over l variables, let and $f(x) = \sum_{|S|\leqslant d} \hat{f}(S) H_S(x)$ be the Hermite expansion of f. Then, $\|f(\mathcal{G})\|_2^2 = \sum \hat{f}(S)^2$ which clearly implies that $\|f(\mathcal{G})\|_2 \geqslant \max_S |\hat{f}(S)|$.

Fix an $S \subseteq [l]$ with $|S| \leqslant d$. By basic properties of the Hermite polynomials (see e.g. [13]) we have that $H_S(x) = \sum_{U\subseteq S} h_S^U \cdot \chi_U(x)$ with $|h_S^U| \leqslant d^d$. Hence, for a fixed $T \subseteq [l]$, c_T can be written as $\sum_{S\supseteq T} h_S^T \hat{f}(S)$. Since $S \subseteq [l]$ and $|S| \leqslant d$, there are at most $\binom{l+d}{d}$ terms in the summation. Therefore, it must be the case that there exists some S such that $|\hat{f}(S)| \geqslant d^{-d} \cdot |c_T|/\binom{l+d}{d}$. This completes the proof. □

Notation: For the remaining of this proof we will be interested in the coefficients $w_{T,T'}$ for $T' = \emptyset$. For notational convenience we shall denote $w_T := w_{T,\emptyset}$.

We now claim that for all T we have

$$|w_T| \leqslant n^{-10d}. \tag{2}$$

Using Fact 3.4, if this were not the case we would get a contradiction with Claim 3.3.

At this point we establish the relationship between the w_T's and the coefficients c_S of f in our original basis $\{\chi_S\}$.

By definition, the restriction obtained from $f_{a'}(g,h)$ by setting the h_i variables to 0 is identical to the function $f(g_1^d, \ldots, g_n^d, g_1, \ldots, g_n)$. Therefore we have

$$\textstyle\sum_{T\subseteq[n]} w_T \cdot \prod_{i\in T} g_i = \sum_{S\subseteq[2n]} c_S \cdot \prod_{i\in S\cap[n]} g_i^d \cdot \prod_{(n+i)\in S} g_i \tag{3}$$

For any fixed T in the LHS of (3) there is an equivalence class of sets S in the RHS such that the monomial $\prod_{i\in S\cap[n]} g_i^d \cdot \prod_{(n+i)\in S} g_i$ equals $\prod_{i\in T} g_i$. It is clear that w_T equals $\sum_S c_S$, where the sum is over all S in the equivalence class. In fact, the structure of the equivalence classes is quite simple, as established by the following claim:

Claim 3.5. *For any $S_0 \neq S_1 \subseteq [2n]$ of size at most d, if*

$$\textstyle\prod_{i\in S_0\cap[n]} g_i^d \cdot \prod_{n+j\in S_0, j\in[n]} g_j = \prod_{i\in S_1\cap[n]} g_i^d \cdot \prod_{n+j\in S_1, j\in[n]} g_j, \tag{4}$$

then there exists some $\ell \in [n]$ such that $S_0 = \{\ell\}$ and $S_1 = \{n+\ell : d\}$ or vice versa.

*P*roof of Claim 3.5. Consider the following two complementary cases.

- $S_0 \cap [n] \neq S_1 \cap [n]$. Without loss of generality, we can assume that there is some $\ell \in S_0 \cap [n]$ with $\ell \notin S_1$. (Otherwise the role of S_0, S_1 can be reversed.) Then to make (4) hold, it must be the case that S_1 contains d copies of $n+\ell$. Now, since $|S_1| \leqslant d$, it can only be the case that $S_1 = \{n+\ell : d\}$, which implies that $S_0 = \{\ell\}$.
- $S_0 \cap [n+1, 2n] \neq S_1 \cap [n+1, 2n]$. We may assume that there is some $\ell \in [n]$ such that $(n+\ell) \in S_0$. Then, for (4) to hold, it must be the case that $\ell \in S_1$. Hence, it must be the case that $S_1 = \{n+\ell : d\}$ (since g_ℓ is raised to the dth power in the RHS of (4)); this in turns enforces $S_0 = \{\ell\}$. □

Claim 3.5 implies the following relation between the coefficients c_S and w_T:

(A) If $T = \{i : d\}$, for some $i \in [n]$, then we have $w_T = c_{S_1} + c_{S_2}$ with $S_1 = \{i\}$ and $S_2 = \{n+i : d\}$.

(B) If T is not of the above form, then there exists a multi-set $S \subseteq [2n]$, $|S| \leqslant d$, where $S \neq \{i\}$ and $S \neq \{n+i : d\}$ for any $i \in [n]$, such that T equals $\{i : d \mid i \in S\} \cup \{i \mid n+i \in S\}$. In this case, we have $w_T = c_S$.

We are now ready to establish the desired bounds on the coefficients of the subfunctions f_1, f_2, f_{12}.

(i) For all $S \subseteq [n]$ with $|S| \geqslant 2$, (2) and (B) yield $|c_S| \leqslant n^{-10d}$.

(ii) For all $S \subseteq [n+1, 2n]$ with $S \neq \{n+i : d\}$ for some $i \in [n]$, (2) and (B) yield $|c_S| \leqslant n^{-10d}$.

(iii) For all $i \in [n]$, by (2) and (A) we obtain $\big||c_{\{i\}}| - |c_{\{n+i:d\}}|\big| \leqslant |c_{\{i\}} + c_{\{n+i:d\}}| \leqslant n^{-10d}$.

(iv) For all S such that $S \cap [n] \neq \emptyset$ and $S \cap [n+1, 2n] \neq \emptyset$, (2) and (B) yield $|c_S| \leqslant n^{-10d}$.

Since the coefficients of f_1, f_2 are either very small (cases (i), (ii) above) or matching (case (iii)), we get $\big|\mathrm{wt}(f_1) - \mathrm{wt}(f_2)\big| \leqslant n^{-10d} \cdot \binom{n+d}{d} \leqslant n^{-1}$. Moreover, since every efficient of f_{12} is small (case (iv)), we deduce that $\mathrm{wt}(f_{12}) \leqslant n^{-10d} \cdot \binom{2n+d}{d} \leqslant n^{-1}$. Recalling that $\mathrm{wt}(f_1) + \mathrm{wt}(f_2) + \mathrm{wt}(f_{12}) = \mathrm{wt}(f) = 1$, we get $\mathrm{wt}(f_1) + \mathrm{wt}(f_2) \geqslant 1 - \frac{1}{n}$. Combining these bounds, we get that

$$0.51 \geqslant \mathrm{wt}(f_1), \mathrm{wt}(f_2) \geqslant 0.49 \tag{5}$$

Now fix an $i \in [n]$ with $(n+i) \in I_1(f_2)$. The above inequality implies that there must exist some $S \ni (n+i)$ such that $|c_S| \geqslant 0.49/\binom{n+d}{d}$. By (ii), we deduce that it can only be the case that S equals $\{n+i : d\}$ (as all other coefficients in f_2 are very small). Moreover, (iii) implies that $|c_i| \geqslant 0.48\binom{n+d}{d}^{-1}$, hence $i \in I_{0.5}(f_1)$ (recalling that $\mathrm{wt}(f_1) \leqslant 0.51$). So we have $|I_1(f_2)| \leqslant |I_{0.5}(f_1)|$ and it remains to bound from above the size of $I_{0.5}(f_1)$ by β^{-2}.

Suppose (for the sake of contradiction) that $|I_{0.5}(f_1)| \geqslant \beta^{-2}$. Since $\mathrm{wt}(f_1) \geqslant 0.49$, every $j \in I_{0.5}(f_1)$ comes from the set $S = \{j\}$ (as all the other coefficients of f_1 are too small). Consider all possible realizations of $a \in \{0,1\}^n$. With probability $1 - (1-\beta)^{|I_{0.5}(f_1)|} \geqslant 1 - n^{-1}$ over the choice of a, there exists $i \in I_{0.5}(f_1)$ with $a_i = 1$. Fix such an i. By the definition of $I_{0.5}(f_1)$, we must have $|c_{\{i\}}| \geqslant 0.5 \cdot 0.49\binom{n+d}{d}^{-1} \geqslant 0.2 \cdot \binom{n+d}{d}^{-1}$. Hence, there will be a degree-1 monomial in the expansion of f_a as a polynomial over g and h whose coefficient has absolute value at least $0.2 \cdot \binom{n+d}{d}^{-1}$.

The aforementioned and Fact 3.4 imply that with probability $1 - n^{-1}$ over a it holds

$$\|f_a\|_2 \geqslant \frac{0.2}{\binom{n+d}{d}} \cdot \frac{1}{\binom{2n+d^2}{d^2}(d^2)^{d^2}} \geqslant \Omega(\frac{1}{n^{2d^2}}).$$

By Theorem B.2 and the fact that $\mathrm{wt}(f) = 1$ we get

$$\mathbf{Pr}_{a,g,h}\left[|f_a(g,h)| \leqslant 2^{-n}\right] \leqslant n^{-1} + O(d^2 \cdot n^2 \cdot 2^{-n/d^2}) = o(\beta)$$

which contradicts (1). This completes the proof of Lemma 3.2. □

Because of space constraints, we give the hardness reduction from Unique Games and the rest of the proof of Theorem 1.1 in Appendix C.

4 Hardness of learning noisy halfspaces with degree 2 PTF hypotheses: Proof of Theorem 1.2

Similar to Section 3, the proof has two parts; first (Section 4.1) we construct a dictator test for degree 2 PTFs, and then (Section D.1) we compose the dictator test with the Label Cover instance to prove NP-hardness.

4.1 The Dictator Test

The key gadget in the hardness reduction is a Dictator Test that is designed to check whether a degree-2 PTF is of the form $\mathrm{sign}(x_i)$ for some $i \in [n]$. Suppose f is a degree 2 polynomial

$$f(x) = \theta + f_1(x) + f_2(x), \quad \text{where } f_1(x) = \sum_{i\in[n]} c_i x_i \quad \text{and } f_2(x) = \sum_{i,j\in[n],i\leqslant j} c_{ij} x_i x_j.$$

Below we give a one-query Dictator Test $\mathcal{T}_2$ for $\mathrm{sign}(f(x))$.

$\mathcal{T}_2$: Dictator Test for Degree-2 Polynomials

Input: A degree-2 real polynomial $f : \mathbb{R}^n \to \mathbb{R}$
Fix $\beta := \frac{1}{\log n}$ and $\delta := 2^{-n}$.

1. Generate independent bits $a_1, a_2, \ldots, a_n \in \{0, 1\}$ each with expected value β. Generate n independent $N(0, 1)$ Gaussian variables $g_1, \ldots, g_n$. Set $r = (a_1 g_1, a_2 g_2, \ldots, a_n g_n)$.
2. Generate t by randomly picking a number $i \in \{1, 2, \ldots, (\log n)^2\}$ and set $t = n^i$. Generate a random bit $b \in \{-1, 1\}$.
3. Set $\omega \in \mathbb{R}^n$ to be the all-1s vector $(1, \ldots, 1)$ and set $y = t^3 r + bt^2\delta\omega$.
4. Accept iff $\mathrm{sign}(f(y)) = b$.

We show that $\mathcal{T}_2$ has the following completeness and soundness properties.

Lemma 4.1. *(Completeness) For $i \in [n]$, the polynomial $f(x) = x_i$ passes $\mathcal{T}_2$ with probability at least $1 - \beta$.*

Proof. If $f(x) = x_i$ for some $i \in [n]$, then as long as a_i is set to zero in step 1 we have that $f(x) = b\delta t^2$ and f passes the test. By definition of the test a_i is 0 with probability $1 - \beta$. □

Lemma 4.2. *(Soundness) Let A denote $\sum_{i=1}^n c_i$ and let $I(f)$ be the set $\{i \mid c_i > A/n^2\}$. If a degree-2 polynomial f passes the test with probability at least $1/2 + \beta$, then $|I(f)| \leqslant 1/\beta^2$ and $A > 0$.*

Because of space constraints we prove Lemma 4.2 and give the hardness reduction from Label Cover in Appendix D.

5 Conclusion

We have established two hardness results for proper agnostic learning of low-degree PTFs. Our results show that even if there exist low-degree PTFs that are almost perfect hypotheses, it is computationally hard to find low-degree PTF hypotheses that perform even slightly better than random guessing; in this sense our hardness are rather strong. However, our results do not rule out the possibility of efficient learning algorithms when ϵ is sub-constant, or if unrestricted hypotheses may be used. Strengthening the hardness results along these lines is an important goal for future work, but may require significantly new ideas.

Another natural goal for future work is the following technical strengthening of our results: show that for any constant d, it is hard to construct a degree-d PTF that is consistent with $(\frac{1}{2}+\epsilon)$ fraction of a given set of labeled examples, even if there exists a halfspace that is consistent with a $1-\epsilon$ fraction of the data. Such a hardness result would subsume both of the results of this paper as well as much prior work, and would serve as strong evidence that agnostically learning halfspaces under arbitrary distributions is a computationally hard problem.

References

[1] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 109:237–260, 1998.

[2] S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[3] E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. In *Proc. 21st Annual Conference on Learning Theory (COLT)*, pages 193–204, 2008.

[4] N. Bshouty and L. Burroughs. Maximizing agreements and coagnostic learning. *Theoretical Computer Science*, 350(1):24–39, January 2006.

[5] A. Carbery and J. Wright. Distributional and L_q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[6] I. Diakonikolas, P. Harsha, A. Klivans, R. Meka, P. Raghavendra, R. A. Servedio, and L.-Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *STOC*, pages 533–542, 2010.

[7] 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.

[8] V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. In *FOCS*, pages 385–394, 2009.

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

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

[11] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[12] J. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55:414–440, 1997.

[13] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, UK, 1997.

[14] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 11–20, 2005.

[15] D. Kane. The Gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. CCC 2010, to appear, 2010.

[16] M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

[17] S. Khot. On the power of unique 2-prover 1-round games. In *Proc. 34th STOC*, pages 767–775, 2002.

[18] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

[19] S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *STOC '08: Proceedings of the 40th annual ACM Symposium on Theory of Computing*, pages 345–354, 2008.

[20] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008.

[21] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[22] R. O'Donnell and R. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.

[23] Y. Rabani and A. Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. In *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 649–658, 2009.

[24] R. Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, 1998.

[25] R. Servedio. Every linear threshold function has a low-weight approximator. *Computational Complexity*, 16(2):180–209, 2007.

APPENDIX

A Complexity-theoretic preliminaries

We recall the Unique Games problem that was introduced by Khot [17]:

Definition A.1. *A Unique Games instance $\mathcal{L}$ is defined by a tuple (U, V, E, k, Π). Here U and V are the two vertex sets of a regular bipartite graph and E is the set of edges between U and V. Π is a collection of bijections, one for each edge: $\Pi = \{\pi_e : [k] \to [k]\}_{e \in E}$ where each π_e is a bijection on $[k]$. A* labeling ℓ *is a function that maps $U \to [k]$ and $V \to [k]$. We say that an edge $e = (u, v)$ is* satisfied *by labeling ℓ if $\pi_e(\ell(v)) = \ell(u)$. We define the* value *of the Unique Games instance $\mathcal{L}$, denoted $\mathrm{Opt}(\mathcal{L})$, to be the maximum fraction of edges that can be satisfied by any labeling.*

The Unique Games Conjecture (UGC) was proposed by Khot in [17] and has led to many improved hardness of approximation results over those which can be achieved assuming only $\mathrm{P} \neq \mathrm{NP}$:

Conjecture A.2 (Unique Games Conjecture). [2] *Fix any constant $\eta > 0$. For sufficiently large $k = k(\eta)$, given a Unique Games instance $\mathcal{L} = (U, V, E, k, \Pi)$ that is guaranteed to satisfy one of the following two conditions, it is* NP*-hard to determine which condition is satisfied:* $\mathrm{Opt}(\mathcal{L}) \geqslant 1 - \eta$, *or* $\mathrm{Opt}(\mathcal{L}) \leqslant \frac{1}{k^\eta}$.

Our first hardness result, Theorem 1.1, is proved under the the Unique Games Conjecture. Our second hardness result, Theorem 1.2, uses only the assumption that $\mathrm{P} \neq \mathrm{NP}$; the proof employs a reduction from the Label Cover problem, defined below.

Definition A.3. *A Label Cover instance $\mathcal{L}$ is defined by a tuple (U, V, E, k, m, Π). Here U and V are the two vertex sets of a regular bipartite graph and E is the set of edges between U and V. Π is a collection of "projections", one for each edge: $\Pi = \{\pi_e : [m] \to [k]\}_{e \in E}$ and m, k are positive integers. A* labeling *ℓ is a function that maps $U \to [k]$ and $V \to [m]$. We say that an edge $e = (u, v)$ is* satisfied *by labeling ℓ if $\pi_e(\ell(v)) = \ell(u)$. We define the* value *of the Label Cover instance, denoted $\mathrm{Opt}(\mathcal{L})$, to be the maximum fraction of edges that can be satisfied by any labeling.*

We use the following theorem [24] which establishes NP-hardness of a "gap" version of Label Cover:

Theorem A.4. *Fix any constant $\eta > 0$. Given a Label Cover instance $\mathcal{L} = (U, V, E, k, m, \Pi)$ that is guaranteed to satisfy one of the following two conditions, it is* NP*-hard to determine which condition is satisfied:* $\mathrm{Opt}(\mathcal{L}) = 1$, *or* $\mathrm{Opt}(\mathcal{L}) \leqslant 1/m^\eta$.

B Probability inequalities

We will use the Berry-Esséen Theorem, which is a quantitative version of the Central Limit Theorem:

Theorem B.1. *(Berry-Esséen Theorem) Let $x_1, x_2, \ldots, x_n$ be i.i.d. uniform $\{-1, 1\}$-valued random variables. Let $c_1, \ldots, c_n \in \mathbb{R}$ be such that $\sum_{i=1}^n c_i^2 = 1$ and $\max_i |c_i| \leqslant \tau$. Let g denote a unit Gaussian variable drawn from $N(0, 1)$. Then for any $\theta \in \mathbb{R}$, we have*

$$|\mathbf{Pr}[\sum_{i=1}^n c_i x_i \leqslant \theta] - \mathbf{Pr}[g \leqslant \theta]| \leqslant \tau.$$

We will also use the following anti-concentration result for low-degree polynomials over Gaussian random variables, due to Carbery and Wright:

Theorem B.2 ([5])**.** *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a nonzero degree-d polynomial over the reals. Then for all $\tau > 0$, we have*

$$\mathbf{Pr}_{x \sim \mathbb{N}^n}[|p(x)| \leqslant \tau \|p\|_2] \leqslant O(d\tau^{1/d}).$$

C Omitted Proofs from Section 3: Proof of Theorem 1.1 (degree-d PTFs)

C.1 Hardness reduction from Unique Games

With the completeness and soundness lemmas in place, we are ready to prove Theorem 1.1. The hardness reduction is from a Unique Games Instance $\mathcal{L}(U, V, E, \Pi, k)$ to a distribution of positive and negative examples. The examples lie in $\mathbb{R}^{(|U|+|V|)k}$ and are labeled with either $(+1)$ or (-1). Denote $\dim = (|U| + |V|)k$.

[2] We use the statement from [18] which is equivalent to the original Unique Games Conjecture.

For $w \in U \cup V$ and $x \in \mathbb{R}^{\dim}$, we use $x_w^{(i)}$ to denote the coordinate corresponding to the vertex w's i-th label. We use x_w to indicate the collection of coordinates corresponding to vertex w; i.e., $(x_w^{(1)}, x_w^{(2)}, \ldots, x_w^{(k)})$. For a function $f(x) : \mathbb{R}^{\dim} \to \mathbb{R}$, we use f_u to denote the restriction of f obtained by setting all the coordinates except x_u to 0. Similarly, $f_{u,v}$ denotes the restriction of f obtained by setting all the coordinates except x_u, x_v to 0.

In the reduction that follows, starting from an instance $\mathcal{L}$ of Unique Games, we construct a distribution $\mathcal{D}$ over labeled examples. Let us denote by $\mathrm{Opt}(\mathcal{D})$ the agreement of the best degree-d PTF on $\mathcal{D}$; our constructed distribution has the following properties:

- If $\mathrm{Opt}(\mathcal{L}) = 1 - \eta$, then $\mathrm{Opt}(\mathcal{D}) = 1 - \eta - \frac{1}{\log k}$; and
- If $\mathrm{Opt}(\mathcal{L}) \leqslant 1/k^{\theta(\eta)}$, then $\mathrm{Opt}(\mathcal{D}) \leqslant \frac{1}{2} + \frac{2}{\log k}$.

This immediately yields the desired hardness result. We now describe and analyze our reduction.

Reduction from Unique Games

Input: Unique Games Instance $\mathcal{L}(U, V, E, \Pi, k)$.
Set $\beta = \frac{1}{\log k}$ and $\delta = 2^{-k^2}$.

1. Randomly choose an edge $(u, v) \in E$.
2. Set $y_w = 0$ for any $w \in U \cup V$ such that $w \neq u, w \neq v$.
3. Generate k i.i.d. bits $a_i \in \{0, 1\}$ with $\mathbf{Pr}[a_i = 1] = \beta$, $2k$ independent standard Gaussians $\{h_i, g_i\}_{i=1}^k$ and a uniform random sign $b \in \{-1, 1\}$.
4. For all $i \in [k]$, set $y_v^{(i)} := g_i$ and $y_u^{(i)} := a_i h_i + (g_{\pi^e(i)})^d + \delta b$.
5. Output the labeled example (y, b).

Lemma C.1 (Completeness)**.** *If $\mathrm{Opt}(\mathcal{L}) = 1 - \eta$, then there is a degree-d PTF that is consistent with $1 - \eta - \beta$ fraction of the examples.*

Proof. Suppose that there is a labeling L that satisfies $1 - \eta$ fraction of the edges. Then it is easy to verify that the degree-d PTF

$$\mathrm{sign}(\textstyle\sum_{u \in U} x_u^{(L(u))} - \sum_{v \in V} (x_v^{(L(v))})^d)$$

agrees with $1 - \eta - \beta$ fraction of the examples. □

Lemma C.2 (Soundness)**.** *If $\mathrm{Opt}(\mathcal{L}) \leqslant 1/k^{\Theta(\eta)}$, then no degree-$d$ PTF agrees with more than $1/2 + 2\beta$ fraction of the examples.*

Proof. Suppose (for the sake of contradiction) that some degree-d polynomial f satisfies $1/2+2\beta$ fraction of examples. Then by an averaging argument, for β fraction of the edges (u, v) picked in the first step, we have that $f(x)$ agrees with the labeled example (y, b) with probability $1/2 + \beta$. Let us call these edges"good".

Fix a "good" edge $e = (u, v)$ and let us assume for notational convenience that π^e is the identity mapping. Essentially, we are conducting the test $\mathcal{T}_d$ for the restriction $f_{u,v}$ with parameter $n := k$. Since $f_{u,v}$ passes the test with probability $1/2+\beta$, Lemma 3.2 implies that we must have that $I_{0.5}(f_u), I_1(f_v) \neq \emptyset$ and $|I_1(f_v)|, |I_{0.5}(f_u)| \leqslant 1/\beta^2$.

We are now ready to give our randomized labeling strategy (based on f). For every $u \in U$, randomly pick its label from $I_{0.5}(f_u)$ and for every $v \in V$ randomly pick its label from $I_1(f_v)$. It is clear that each good edge is satisfied with probability β^2. Since at least β fraction of the edges is good, such a labeling satisfies at least $\beta^3 = 1/(\log k)^3$ fraction of the edges in expectation. Hence, there exists a labeling that satisfies such a fraction of the edges, which contradicts the assumption that $\mathrm{Opt}(\mathcal{L}) \leqslant 1/k^\eta$, for k sufficiently large. □

C.2 A technical point: Discretizing the Gaussian Distribution

Lemmas C.1 and C.2 do not quite suffice to prove Theorem 1.1, because the reduction described above is not computable in polynomial time. This is because the distribution $\mathcal{D}$ has infinite support; recall that for each edge e, sampling from the corresponding distribution $\mathcal{D}_e$ requires generating $2k$ independent Gaussian random variables $h = (h_1, \ldots, h_k), g = (g_1, \ldots g_k)$.

To discretize the reduction we replace h by h' and g by g', where each of the $2k$ random variables h'_i, g'_i is independently generated as a sum of N uniform $\{-1, 1\}$ bits divided by $\sqrt{N}$. In Theorem C.4 of Appendix C.2.1, we argue that for sufficiently large N (in particular any $N \geqslant (2k)^{24(d^2)^2}$ suffices), there is a way to couple the distribution of (g, h) with that of (g', h') such that every degree-d^2 polynomial takes the same sign on (g, h) as on (g', h') except with probability at most $1/k$. Since every outcome of $a \in \{0, 1\}^k$ results in the polynomial $f_a(g, h)$ being a degree-d^2 polynomial, if we replace (g, h) with (g', h') in the reduction then the discretized reduction will almost preserve the soundness and completeness guarantees of Section C.1, with only a loss of $\frac{1}{k}$: writing $\mathcal{D}'$ for the discretized distribution, we have

- If $\mathrm{Opt}(\mathcal{L}) \geqslant 1 - \eta$, then $\mathrm{Opt}(\mathcal{D}') \geqslant 1 - \eta - \frac{1}{\log k} - 1/k$; and
- If $\mathrm{Opt}(\mathcal{L}) \leqslant 1/k^\eta$, then $\mathrm{Opt}(\mathcal{D}') \leqslant \frac{1}{2} + \frac{2}{\log k} + 1/k$.

Finally, we observe that the distribution of (g', h') has support of size $(N+1)^{2k} \leqslant (2N)^{2k} \leqslant (4k)^{48d^4k}$; since the label size k is regarded as constant in a Unique Games instance, this is a (large) constant for constant d. Thus it is possible to simply enumerate the entire support of $\mathcal{D}$ in polynomial time (since there are $|E|$ distributions $\mathcal{D}_e$, the overall size of the support of $\mathcal{D}$ is polynomial in the size of the Unique Games instance) and consequently there is no need for randomness – the entire overall reduction is deterministic. Theorem 1.1 now follows by choosing appropriate settings of η and k (e.g., $\eta = \epsilon/2$ and $k = e^{1/\epsilon^2}$ suffices).

Finally, we note that the above remarks imply that Theorem 1.1 holds not only for constant d, but for d as large as $O((\log n)^{1/4})$ – since k is constant, for such d the support size $(4k)^{48d^4k}$ is still polynomial in n.

C.2.1 Discretizing the Gaussian distribution

The following theorem shows that there exists a distribution $\mathcal{H}_N/\sqrt{N}$ that is point-wise close to a Gaussian distribution $\mathcal{G}$ with high probability:

Theorem C.3. *There is a probability distribution $(\mathcal{G}, \mathcal{H}_N)$ on $\mathbb{R}^2$ such that the marginal distribution $\mathcal{G}$ of the first coordinate follows the standard $N(0, 1)$ Gaussian distribution, and the marginal distribution $\mathcal{H}_N$ of the second coordinate is distributed as a sum of N random bits, i.e., $\mathcal{H}_N = \sum_{i=1}^N b_i$ where each b_i is an independent random bit from $\{-1, 1\}$. In addition, $\mathcal{H}_N$ and $\mathcal{G}$ are pointwise close in the following sense: $\mathbf{Pr}[|\mathcal{G} - \frac{\mathcal{H}_N}{\sqrt{N}}| \leqslant O(N^{-1/4})] \geqslant 1 - O(N^{-1/4})$.*

Proof. Let Φ be the CDF (cumulative distribution function) of $\mathcal{H}_N$, and let Ψ be the CDF of $\mathcal{G}$ (the standard Gaussian Distribution).

We couple the random variables $\mathcal{G}, \mathcal{H}_N$ in the following way: to obtain a draw (g_0, h_0) from the joint distribution, first we sample h_0 from the marginal distribution on $\mathcal{H}_N$. We know that

$$\mathbf{Pr}[\mathcal{H}_N = h_0] = \Phi(h_0) - \Phi(h_0 - 2),$$

since if h_0 is a feasible outcome of summing N bits then $h_0 - 2$ is the largest feasible outcome that is less than h_0 (if any feasible outcome less than h_0 exists). Then we generate g_0 by drawing random samples from the standard Gaussian distribution until we obtain a sample that lies in the interval $(\Psi^{-1}(\Phi(h_0 - 2)), \Psi^{-1}(\Phi(h_0)]$; when we obtain such a sample, we set g_0 to this value.

It is not difficult to see that the random variable $\mathcal{G}$ defined in this way follows the standard Gaussian distribution; essentially we are using the value of h_0 as a indicator of whether $\mathcal{G}$ is in the interval $(\Psi^{-1}(\Phi(h_0 - 2)), \Psi^{-1}(\Phi(h_0)]$. We also need to check that $\mathbf{Pr}[\mathcal{H} = h_0]$ is equal to $\mathbf{Pr}[\mathcal{G} \in (\Psi^{-1}(\Phi(h_0 - 2)), \Psi^{-1}(\Phi(h_0))]$. This is true because

$$\begin{aligned}\mathbf{Pr}[\mathcal{H} = h_0] &= \mathbf{Pr}[h \in (h_0 - 2, h_0]] \\ &= \Phi(h_0) - \Phi(h_0 - 2) \\ &= \mathbf{Pr}[\mathcal{G} \in (\Psi^{-1}(\Phi(h_0 - 2)), \Psi^{-1}(\Phi(h_0))]].\end{aligned}$$

With the above coupling of $\mathcal{G}$ and $\mathcal{H}$, it remains to prove that every value in the interval $(\Psi^{-1}(\Phi(h_0 - 2)), \Psi^{-1}(\Phi(h_0))]$ is close to $h_0/\sqrt{N}$, with high probability over a random choice of h_0 as described above. It suffices to verify that the following two inequalities each hold with probability at least $1 - O(N^{-1/4})$:

$$\left|\Psi^{-1}(\Phi(h_0)) - \frac{h_0}{\sqrt{N}}\right| \leqslant O(N^{-1/4}) \qquad \text{and} \qquad \left|\Psi^{-1}(\Phi(h_0 - 2)) - \frac{h_0}{\sqrt{N}}\right| \leqslant O(N^{-1/4}).$$

We consider the first inequality; the first one is entirely similar. We show that $\Psi^{-1}(\Phi(h_0)) - \frac{h_0}{\sqrt{N}} \leqslant O(N^{-1/4})$; the other direction $\Psi^{-1}(\Phi(h_0)) - \frac{h_0}{\sqrt{N}} \geqslant -O(N^{-1/4})$ is similar.

By the Berry-Esséen Theorem (Theorem B.1 in Appendix B), we have that $|\Phi(h_0) - \Psi(\frac{h_0}{\sqrt{N}})| \leqslant \frac{1}{\sqrt{N}}$. Therefore, we have that

$$\Psi^{-1}(\Phi(h_0)) \leqslant \Psi^{-1}(\Psi(\frac{h_0}{\sqrt{N}}) + \frac{1}{\sqrt{N}}) \leqslant \frac{h_0}{\sqrt{N}} + E_{h_0}, \tag{6}$$

where the "error term" E_{h_0} is the value for which $\Psi(h_0/\sqrt{N} + E_{h_0}) - \Psi(h_0/\sqrt{N}) = 1/\sqrt{N}$.

If $|h_0| \leqslant \sqrt{\frac{N \ln N}{2}}$, then in an interval of width $N^{1/4}$ around h_0 the PDF of the standard Gaussian is everywhere at least $\Omega(N^{-1/4})$; consequently, if $|h_0| \leqslant \sqrt{\frac{N \ln N}{2}}$ then the error term E_{h_0} is at most $O(N^{-1/4})$ as required. A standard Chernoff Bound implies that $\mathbf{Pr}[|h_0| < \sqrt{\frac{N \ln N}{2}}]$ is at most $O(N^{-1/4})$, and the argument is complete. □

Now we use the joint distribution constructed in Theorem C.3 to discretize the standard n-dimensional Gaussian space for low-degree PTFs.

Theorem C.4. *Fix any constant $D \geqslant 1$, and let $f(x_1, \ldots, x_n) = \sum_{|S| \leqslant D} \hat{f}(S) \prod_{i \in S} x_i$ be a degree-D polynomial over $\mathbb{R}^n$. Let $(y, z) \in \mathbb{R}^n \times \mathbb{R}^n$ be generated by taking each pair (y_i, z_i) to be an i.i.d. draw from the distribution $(\mathcal{G}, \mathcal{H}_N)$ of Theorem C.3, where we take $N = n^{24D^2}$. Then we have*

$$\mathbf{Pr}[\mathrm{sign}(f(y)) \neq \mathrm{sign}(f(z))] \leqslant O(1/n).$$

Proof. First, we may assume without loss of generality that the polynomial f is normalized so that $\sum_{S \neq \emptyset} |\hat{f}(S)|$ equals 1. Since there are at most $\binom{n+D}{D}$ coefficients in f, one of these coefficients $\hat{f}(S)$ must satisfy $|\hat{f}(S)| \geqslant \frac{1}{\binom{n+D}{D}}$; now Lemma 3.4 implies that $\|f\|_2 \geqslant \frac{1}{\binom{n+D}{D}^2 D^D}$.

We have

$$\mathbf{Pr}[\mathrm{sign}(f(y)) \neq \mathrm{sign}(f(z)] \leqslant \mathbf{Pr}[|f(y)| \leqslant |f(z) - f(y)|].$$

To bound the latter probability by $O(1/n)$, we show that $|f(y)| \geqslant n^{-3D^2}$ with probability $1 - O(1/n)$, and that $|f(z) - f(y)| < n^{-3D^2}$ with probability $1 - O(1/n)$.

The first desired bound, $\mathbf{Pr}[|f(y)| \leqslant n^{-3D^2}] \leqslant O(1/n)$, is an immediate consequence of Theorem B.2.

For the second, we note that by a union bound and Theorem C.3, with probability at least $1-O(n/N^{1/4}) \geqslant 1 - O(\frac{1}{n})$ every $i \in [n]$ satisfies $|y_i - z_i| \leqslant O(N^{-1/4})$. Standard Chernoff bounds and Gaussian tail bounds give that the probability any $|y_i|$ or $|z_i|$ exceeds $n^{1/d}$ is much less than $1/n$. Now similar to the calculation used to bound $f(r + \delta\omega) - f(r)|$ in the proof of Claim 3.3, when y and z are $O(N^{-1/4})$-close in each coordinate and each coordinate is at most $n^{1/d}$, we have that

$$|f(y) - f(z)| \leqslant O(N^{-1/4}) \cdot O(n) < n^{-3D^2}.$$

This concludes the proof. □

D Omitted Proofs from Section 4: Proof of Theorem 1.2 (degree-2 PTFs)

Recall Lemma 4.2:

Lemma 4.2. *(Soundness) Let A denote $\sum_{i=1}^{n} c_i$ and let $I(f)$ be the set $\{i \mid c_i > A/n^2\}$. If a degree-2 polynomial f passes the test with probability at least $1/2 + \beta$, then $|I(f)| \leqslant 1/\beta^2$ and $A > 0$.*

Proof. The proof is by contradiction. Let f be a degree-2 polynomial with $|I(f)| > 1/\beta^2$ or $A \leqslant 0$, and suppose that f passes the test with probability at least $\frac{1}{2} + \beta$.

First we show the following lemma.

Lemma D.1. $\mathbf{Pr}_r[f_1(r) \in (-\delta A, \delta A)] \leqslant \frac{2}{n}$.

Proof. The inequality obviously holds for $A \leqslant 0$ since the interval has measure 0. Thus we may assume that $A > 0$ and $|I(f)| \geqslant 1/\beta^2$. We know that in step 1 when generating the bit-vector a, with probability at least $1 - (1 - \beta)^{|I(f)|} \geqslant 1 - \frac{1}{n}$ at least one of the coordinates in $I(f)$ has its bit a_i nonzero. Fix any such outcome for the bit-vector a; now considering the random choice of the Gaussians $g_1, \ldots, g_n$, we have that the resulting $f_1(r)$ is a Gaussian variable with variance at least A^2/n^4 (as one of the weights is at least A/n^2). Using the standard fact that an $N(\sigma, \mu)$ Gaussian random variable puts probability mass at most t/σ on any interval of length t, we have that for such an outcome of the a-vector,

$$\mathbf{Pr}_g[f_1(r) \in (-\delta A, \delta A)] \leqslant \frac{2\delta A}{A/n^2} \leqslant \frac{n^3}{2^n} \leqslant \frac{1}{n}.$$

Now a union bound gives that for at most $\frac{2}{n}$ of the r generated, $f(r)$ is inside the interval $(-\delta A, \delta A)$. □

Now we observe that for any outcome r, the vectors r and $-r$ are generated with equal probability. Thus an equivalent test to $\mathcal{T}_2$ would be to generate r, t as described by the test and then check a randomly selected one of the following four inequalities:

$$f(t^3 r + t^2 \delta\omega) \geqslant 0 \tag{7}$$
$$f(t^3 r - t^2 \delta\omega) < 0 \tag{8}$$
$$f(-t^3 r + t^2 \delta\omega) \geqslant 0 \tag{9}$$
$$f(-t^3 r - t^2 \delta\omega) < 0. \tag{10}$$

Since f is assumed to pass the test with probability $\frac{1}{2}+\beta$ an averaging argument gives that for a $\beta/2$ fraction of the possible outcomes of r, at least a $(\frac{1}{2}+\beta/2)$ fraction of all the constraints involving that r outcome are satisfied. (Note that for any fixed outcome of r there are $4(\log n)^2$ constraints, corresponding to inequalities (7)–(10) for each of the $(\log n)^2$ possible values of t.) For this $\beta/2$ fraction of r, let us remove those outcomes r such that $p_1(r) \in (-\delta A, \delta A)$ (recall that this is at most a $2/n$ fraction of all r-outcomes). Recalling that $\beta = \frac{1}{\log n}$, we know there are at least $\beta/4$ fraction of r-outcomes remaining; we call these "good" r's.

Let us fix a good r. By an averaging argument again, for any "good" r, for at least a $\beta/4$ fraction of the possible outcomes of t, at least 3 out of the 4 of the inequalities that contain t and r are satisfied. There are 4 different ways of choosing 3 out of the 4 constraints. Without loss of generality, let us assume that for a $\beta/16$ fraction of the t-outcomes, the first, second, and fourth constraints (7), (8) and (10) are satisfied. That is:

$$f(t^3 r + t^2 \delta\omega) > 0 \tag{11}$$

$$f(t^3 r - t^2 \delta\omega) < 0 \tag{12}$$

$$f(-t^3 r - t^2 \delta\omega) < 0. \tag{13}$$

Let us call these t "good" for the corresponding r, and let us denote the set that contains all the "good" t for a given "good" r by T_r. Since the possible choice of $t = n^i$ ranges over all $i \in [\log^2 n]$, we therefore obtain $|T_r| \geqslant (\log n)^2 \cdot \beta/16 = \Theta(\log n)$.

Since $f(x)$ is a degree 2 polynomial, we can express $f(r+\delta\omega)$ as:

$$f(r+\delta\omega) = \theta + f_1(r) + f_2(r) + \delta \sum_{i=1}^{n} c_i + \delta^2 \sum_{1 \leqslant i \leqslant j \leqslant n} c_{ij} + \delta \sum_{1 \leqslant i \leqslant j \leqslant n} c_{ij}(r_i + r_j).$$

Denote $B = \sum_{1 \leqslant i \leqslant j \leqslant n} c_{ij}$ and $f_2'(r) = \sum_{1 \leqslant i \leqslant j \leqslant n} c_{ij}(r_i + r_j)$. We can rewrite (11), (12), (13) as:

$$t^3 f_1(r) + t^2 \delta A + t^6 f_2(r) + t^5 \delta f_2'(r) + t^4 \delta^2 B + \theta \geqslant 0 \tag{14}$$

$$t^3 f_1(r) - t^2 \delta A + t^6 f_2(r) - t^5 \delta f_2'(r) + t^4 \delta^2 B + \theta < 0 \tag{15}$$

$$t^3 f_1(r) + t^2 \delta A - t^6 f_2(r) - t^5 \delta f_2'(r) - t^4 \delta^2 B - \theta > 0 \tag{16}$$

Notice that (14) and (16) yield

$$f_1(r) \geqslant -\delta A/t + \left| t^3 f_2(r) + \delta t^2 f_2'(r) + \delta^2 t B + \theta/t^3 \right|.$$

Since we already know that $f_1(r) \notin (-\delta A, \delta A)$ and t is at least 1, we get that

$$f_1(r) \geqslant \delta A.$$

Also for (15), we can rewrite it as

$$f_1(r) \leqslant \delta A/t - (t^3 f_2(r) - \delta t^2 f_2'(r) + \delta^2 t B + \theta/t^3).$$

Let us further simplify the notation by writing C for $f_2(r)$, D for $\delta f_2'(r)$ and E for $\delta^2 B$. Then we may rewrite the above constraints as follows:

$$f_1(r) \geqslant -\delta A/t + \left| t^3 C + t^2 D + t E + \theta/t^3 \right|$$

and

$$\delta A \leqslant f_1(r) \leqslant \delta A/t - (t^3 C - t^2 D + t E + \theta/t^3). \tag{17}$$

Notice that above (upper and lower) bound hold for any t in T_r. Therefore, we know that for any $t_1, t_2 \in T_r$,

$$\delta A/t_1 - (t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3) \geqslant -\delta A/t_2 + \left|t_2^3 C + t_2^2 D + t_2 E + \theta/t_2^3\right|$$

which is equivalent to

$$-(t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3) + \delta A(\frac{1}{t_1} + \frac{1}{t_2}) \geqslant \left|t_2^3 C + t_2^2 D + t_2 E + \theta/t_2^3\right|. \tag{18}$$

Using the fact that $f_1(r) > \delta A$, the inequality (17) gives $-(t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3) > (1 - \frac{1}{t_1})\delta A$, which may be rewritten as $\delta A \leqslant \frac{-(t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3)}{1 - 1/t_1}$. Combining this with (18), we know that for any $t_1, t_2 \in T_r$, we have

$$-(t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3)\left(1 + \frac{(\frac{1}{t_1} + \frac{1}{t_2})}{1 - \frac{1}{t_1}}\right) \geqslant \left|t_2^3 C + t_2^2 D + t_2 E + \theta/t_2^3\right|.$$

By definition, $t_i \geqslant n$ for any i, so we have $\frac{(\frac{1}{t_1} + \frac{1}{t_2})}{1 - \frac{1}{t_1}} \leqslant 3/n$. Therefore, for any t_1, t_2 in T_r, the following inequality holds:

$$\frac{-(t_1^3 C + t_1^2 D - t_1 E + \theta/t_1^3)}{\left|t_2^3 C + t_2^2 D + t_2 E + \theta/t_2^3\right|} \geqslant \frac{1}{1 + \frac{(\frac{1}{t_1} + \frac{1}{t_2})}{1 - \frac{1}{t_1}}} \geqslant 1 - 3/n. \tag{19}$$

Note that the denominator of the LHS of (19) can be zero for at most 6 values of t_2; we eliminate any such values from T_r, and we still have $|T_r| \geqslant \Theta(\log n)$. (Actually, we will only need $|T_r| \geqslant 5$ for the remainder of the argument to establish the required contradiction.) Let us pick $t_0 < t_1 < t_2 < t_3 < t_4$ from T_r, and let us write G to denote $-(t_1^3 C - t_1^2 D + t_1 E + \theta/t_1^3)$. We know that

$$G \leqslant t_1^3|C| + t_1^2|D| + t_1|E| + |\theta|/t_1^3.$$

Also for t_0, t_2, t_3, t_4, we write:

$$F_0 := t_0^3 C - t_0^2 D + t_0 E + \theta/t_0^3 \tag{20}$$
$$F_2 := t_2^3 C - t_2^2 D + t_2 E + \theta/t_2^3 \tag{21}$$
$$F_3 := t_3^3 C - t_3^2 D + t_3 E + \theta/t_3^3 \tag{22}$$
$$F_4 := t_4^3 C - t_4^2 D + t_4 E + \theta/t_4^3. \tag{23}$$

Let F denote $\max_{i=0,2,3,4} |F_i|$. By (19) we know that

$$\frac{G}{F} \geqslant 1 - 3/n. \tag{24}$$

Viewing C, D, E, θ as unknowns, we may solve the above linear system consisting of equations (20),(21),(22),(23) using Cramer's rule. We find that

$$C = \frac{\begin{vmatrix} F_0 & -t_0^2 & t_0 & 1/t_0^3 \\ F_2 & -t_2^2 & t_2 & 1/t_2^3 \\ F_3 & -t_3^2 & t_3 & 1/t_3^3 \\ F_4 & -t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}}{\begin{vmatrix} t_0^3 & -t_0^2 & t_0 & 1/t_0^3 \\ t_2^3 & -t_2^2 & t_2 & 1/t_2^3 \\ t_3^3 & -t_3^2 & t_3 & 1/t_3^3 \\ t_4^3 & -t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}} = \frac{\begin{vmatrix} F_0 & t_0^2 & t_0 & 1/t_0^3 \\ F_2 & t_2^2 & t_2 & 1/t_2^3 \\ F_3 & t_3^2 & t_3 & 1/t_3^3 \\ F_4 & t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}}{\begin{vmatrix} t_0^3 & t_0^2 & t_0 & 1/t_0^3 \\ t_2^3 & t_2^2 & t_2 & 1/t_2^3 \\ t_3^3 & t_3^2 & t_3 & 1/t_3^3 \\ t_4^3 & t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}}.$$

Since $0 < t_0 < t_2 < t_3 < t_4$ and these values are at least a factor of n apart from each other, we have that

$$\begin{vmatrix} t_0^3 & t_0^2 & t_0 & 1/t_0^3 \\ t_2^3 & t_2^2 & t_2 & 1/t_2^3 \\ t_3^3 & t_3^2 & t_3 & 1/t_3^3 \\ t_4^3 & t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}$$

is $\Omega(t_4^3 t_3^2 t_2 t_0^{-3})$.

Since $F = \max_{i=0,2,3,4} |F_i|$, we know that the absolute value of

$$\begin{vmatrix} F_0 & t_0^2 & t_0 & 1/t_0^3 \\ F_2 & t_2^2 & t_2 & 1/t_2^3 \\ F_3 & t_3^2 & t_3 & 1/t_3^3 \\ F_4 & t_4^2 & t_4 & 1/t_4^3 \end{vmatrix}$$

is at most $O(F t_4^2 t_3 t_0^{-3})$. Thus we have $|C| = O(\frac{F}{t_4 t_3 t_2})$.

Similar analysis shows that

$$|D| = O(F/t_3 t_2); \quad |E| = O(F/t_2); \quad \text{and} \quad |\theta| = O(F t_0^3).$$

Therefore, we have

$$G \leqslant |C| t_1^3 + t_1^2 |D| + t_1 |E| + |\theta|/t_1^3 \leqslant F \cdot O(t_1^3/t_4 t_3 t_2 + + t_1^2/t_2 t_3 + t_1/t_2 + t_0^3/t_1^3).$$

Recalling that $t_{i+1}/t_i \geqslant n$ as they are different powers of n, we have that

$$\frac{G}{F} \leqslant O(1/n).$$

This contradicts (24) and concludes the proof of the soundness Lemma, Lemma 4.2. □

D.1 Hardness reduction from Label Cover

Recall that our reduction is from a Label Cover instance $\mathcal{L}$ specified by (U, V, E, k, m, Π) . For notational convenience let us write $F(q)$ to denote the space of possible labels for vertex $q \in U \cup V$, for $u \in U$, $F(u)$ denotes $[k]$ and for $v \in V$, $F(v)$ denotes $[m]$.

We reduce to a learning problem with labeled examples in $\mathbb{R}^{|U|k+|V|m} \times \{-1, 1\}$. Let $\dim$ denote $|U|k + |V|m$. For $y \in \mathbb{R}^{\dim}$ and $q \in U \cup V$, we write $y_q^{(i)}$ to denote the vector consisting of all coordinates that correspond to vertex q, i.e. y_u denotes $(y_u^{(i)})_{i \in [k]}$ for $u \in U$ and y_v denotes $(y_v^{(i)})_{i \in [m]}$ for $v \in V$.

We give the reduction from Label Cover to the learning problem below. The high level idea is that the Dictator Test $\mathcal{T}_2$ is performed on the restricted function $p_v(y)$ for a random $v \in V$.

Reduction from Label-Cover $\mathcal{L}$

Input: Label Cover Instance (U, V, E, k, m, Π).

1. Randomly pick a vertex $v \in V$.
2. For each $w \neq v$, $w \in U \cup V$, set $y_w = 0$.
3. Let $a_1, \ldots, a_m$ be independent $\{0,1\}$ bits each with $\mathbf{E}[a_i] = \beta$. Let $g_1, \ldots, g_m$ be independent $N(0,1)$ Gaussian random variables. Let i be chosen uniformly from $[(\log m)^2]$ and set $t = m^i$. Let b be a random uniform bit from $\{-1, 1\}$.
4. Set $r = (a_1 g_1, a_2 g_2, \ldots, a_m g_m)$.
5. Let $\omega \in \mathbb{R}^m$ be $\omega = (1, \ldots, 1)$, and set $y_v := t^3 r + b t^2 \delta \omega$.
6. Output the labeled example $(\mathrm{Fold}(y_v), b)$ (we describe the folding procedure $\mathrm{Fold}(\cdot)$ later).

The learning problem is to find a degree 2 polynomial $p : \mathbb{R}^{dim} \to \{-1, 1\}$ such that $\mathrm{sign}(p(y)) = b$ for the largest possible fraction of labeled examples generated as described above. Let us denote

$$p(y) = \theta + \sum_{q \in U \cup V, i \in F(q)} c_q^{(i)} y_q^{(i)} + \sum_{q_1, q_2 \in U \cup V, i \in F(q_1, j \in F(q_2)} c_{(q_1,q_2)}^{(i,j)} y_{q_1}^{(i)} y_{q_2}^{(j)}.$$

Notice that in the reduction, when vertex v is picked we set all the coordinates to zero except y_v. Essentially we are performing the test $\mathcal{T}_2$ on the function

$$p_v = \theta + \sum_{i \in [m]} c_v^{(i)} y_v^{(i)} + \sum_{i,j \in [m]} c_{(v(i),v(j))} y_v^{(i)} y_v^{(j)}$$

which is the restriction of $p(y)$ obtained by setting all the coordinates to zero except those coordinates corresponding to vertex v. The overall fraction of agreement of $p(y)$ on all examples is the average probability, over all $v \in V$, that p_v passes $\mathcal{T}_2$.

Folding Trick: We use the "folding " technique that was first introduced in [9, 19]. The trick essentially amounts to the following: instead of outputting the labeled example (y, b) in the last step of the reduction, we output $(\mathrm{Fold}(y), b)$ where $\mathrm{Fold}(y)$ is the projection of y into a subspace $H^\perp$ (defined below). Folding enables us to enforce that p takes the same value on different points in $\mathbb{R}^{\mathrm{dim}}$ as long as they project to the same point in $H^\perp$.

We define the subspaces $H, H^\perp$ for our folding as follows:

Definition D.2. *For every $e = \{u, v\} \in E, i \in [k]$, we define $b(e,i) \in \mathbb{R}^{\mathrm{dim}}$ to be the vector that has 0 at every coordinate except that $b(e,i)_u^{(i)}$ is 1 and for every $j \in (\pi^e)^{-1}(i)$, $b(e,j)_v^{(j)}$ is -1. Let B be the collection of all such $b(e,i)$, i.e. $B = \{b(e,i) \mid e = \{u,v\} \in E, i \in [k]\}$. We define H to be $span(B)$ and $H^\perp$ to be the orthogonal complement of H in R^{dim}.*

We define $\mathrm{Fold}(y)$ to be the projection of y onto $H^\perp$. It is easy to see that the mapping $\mathrm{Fold}(\cdot)$ can be performed in polynomial time.

After folding, we can further enforce $p(x)$ to have following "folding" property:

$$\text{For any } h \in H \text{ and } x \in \mathbb{R}^{\mathrm{dim}}, p(x + h) = p(x).$$

We call functions that have the above property "folded". In particular for $e = \{u, v\} \in E$, $c \in \mathbb{R}$, and $i \in [k]$, a folded function p satisfies $p(x + cb(e, i)) = p(x)$. If we view $p(y)$ as a polynomial only on $y_u^{(i)}$ and $y_v^{(j)}$ for $j \in (\pi^e)^{-1}(i)$, then Lemma D.5 shows that we have the following folding property of p:

$$c_u^{(i)} = \sum_{j \in (\pi^e)^{-1}} c_v^{(j)}.$$

If we sum over all possible i, this implies for any edge $\{u, v\}$, we have

$$\sum_{i \in [k]} c_u^{(i)} = \sum_{i \in [m]} c_v^{(i)}.$$

Now we are ready to prove Theorem 1.2. We will show the following two properties of the reduction to complete the proof.

Lemma D.3 (Completeness)**.** *If $\mathrm{Opt}(\mathcal{L}) = 1$, then there is a folded function $p(x)$ that is consistent with $1 - 1/\log m$ fraction of the labeled examples generated by the reduction.*

Lemma D.4 (Soundness)**.** *If $\mathrm{Opt}(\mathcal{L}) \leqslant 1/m^\eta$, then there is no folded degree-2 polynomial that is consistent with $1/2 + \frac{2}{\log^2 m}$ fraction of the labeled examples generated by the reduction.*

Combining Lemmas D.3 and D.4 and noticing that m can be an arbitrarily large constant (such as e^{1/ϵ^2}),we obtain Theorem 1.2. (A discretization similar to that of Section C.2 is also required, and can be obtained in a routine way by slightly modifying the parameters of that section's construction.)

Proof of Theorem D.3: Suppose that $\mathrm{Opt}(\mathcal{L}) = 1$, so there is a labeling l satisfying all the edges. Then consider the following function

$$p(x) = \sum_{w \in U \cup V} x_w^{(l(w))}.$$

For every $v \in V$, the function p_v is a dictator and passes $\mathcal{T}_m$ with probability at least $1 - \frac{1}{\log m}$ by Lemma 4.1. Consequently the overall probability that p passes the test is at least $1 - 1/\log m$. Finally, it is easy to check that thus function $p(x)$ is folded. □

Proof of Theorem D.4: Suppose that there is some folded degree-2 polynomial $p(x)$ such that $\mathrm{sign}(p(x))$ agrees with more than $\frac{1}{2} + \frac{2}{\log m}$ fraction of the example, i.e., the averaging passing probability of p_v on $\mathcal{T}_m$ is $\frac{1}{2} + \frac{2}{\log m}$. We will show that $\mathrm{Opt}(\mathcal{L}) > 1/m^\eta$ and thus prove the theorem.

By an averaging argument, we know for a $\frac{1}{\log m}$ fraction of the vertices $v \in V$, the restricted polynomial p_v passes the test $\mathcal{T}_k$ with probability at laest $\frac{1}{2} + \frac{1}{\log m}$; we refer to any such v as a "good" vertex. We say that an edge is "good" if the V-endpoint of the edge is a good vertex. Since the graph is regular, we know that at least a $\frac{1}{\log m}$ fraction of all edges are "good".

For a "good" vertex v, let us define I_v to be

$$I_v = \{j |\ j \in [m], c_v^{(j)} > \sum_{i=1}^{m} c_v^{(i)}/m^2\}.$$

By Lemma 4.2, we have $|I_v| \leqslant (\log m)^2$ and $\sum_{i \in [m]} c_v^{(i)} > 0$. For every $u \in U$, we define $J_u = \{j |\ j \in [k], c_u^{(j)} \geqslant \sum_{i \in [k]} c_u^{(i)}/k\}$. We note that J_u is not empty as

$$\max_j c_u^{(j)} \geqslant \sum_{i \in [k]} c_{u[i]}/k.$$

We define the following labeling strategy for $\mathcal{L}$. For $u \in U$, randomly assign it a label from J_u; for $v \in V$, randomly assign it a label from I_v (if I_v is empty, we assign a random label to v).

For every good edge $e = (u, v)$ and any $j \in J_u$, since p is folded, we have that

$$\sum_{i \in \pi_e^{-1}(j)} c_v^{(i)} = c_u^{(j)} \geqslant \sum_{i \in [k]} c_u^{(i)}/k = \sum_{i \in [m]} c_v^{(i)}/k.$$

There is at least one label i in $\pi_e^{-1}(j)$ such that $\sum_{i \in [m]} c_v^{(i)}/km \geqslant \sum_{i \in [m]} c_v^{(i)}/m^2$, and this label is therefore in I_v. As noted earlier we have $|I_v| \leqslant (\log m)^2$, and so by our randomized labeling strategy there is at least a $1/(\log m)^2$ probability that edge $\{u, v\}$ is satisfied.

Therefore the above labeling strategy satisfies (in expectation) at least $1/(\log(m)^2)$ fraction of the good edges and consequently at least $1/(\log m)^3$ fraction of all edges. This means that $\mathrm{Opt}(\mathcal{L}) > 1/m^\eta$ and the proof is complete. □

D.1.1 Folding Lemma

Lemma D.5. *Let*

$$f(x) = \theta + \sum_{i=0}^{n} w_i x_i + \sum_{0 \leqslant i \leqslant j \leqslant n} w_{ij} x_i x_j$$

be a degree 2 function. Suppose that for every $x \in \mathbb{R}^n, c \in \mathbb{R}$ we have $f(x + c(1, -1, \ldots, -1)) = f(x)$. Then $w_0 = \sum_{i=1}^n w_i$.

Proof. Expanding the equality $f(x + c(1, -1, \ldots, -1)) = f(x)$, we get that

$$\theta + w_0(x_0 + c) + \sum_{i=1}^{n} w_i(x_i - c) + w_{00}(x_0 + c)^2 + \sum_{j=1}^{n} w_{0j}(x_0 + c)(x_j - c) + \sum_{1 \leqslant i \leqslant j \leqslant n} w_{ij}(x_i - c)(x_j - c)$$
$$= \theta + \sum_{i=0}^{n} w_i x_i + \sum_{0 \leqslant i \leqslant j \leqslant n} w_{ij} x_i x_j.$$

Since this equation holds for all c, x, if we express the LHS and RHS as polynomials in the variables $c, x_0, x_1, \ldots, x_n$, the corresponding coefficients must be the same. If we look at the coefficients of the degree-1 monomial c, we have that $w_0 - \sum_{i=1}^n w_i = 0$, and the lemma is proved. □