Electronic Colloquium on Computational Complexity, Report No. 185 (2010)

Agnostic Learning of Monomials by Halfspaces is Hard*

Vitaly Feldman† Venkatesan Guruswami‡ Prasad Raghavendra§ Yi Wu¶

December 1, 2010

Abstract

We prove the following strong hardness result for learning: Given a distribution of labeled examples from the hypercube such that there exists a *monomial* consistent with $(1-\epsilon)$ of the examples, it is NP-hard to find a *halfspace* that is correct on $(1/2+\epsilon)$ of the examples, for arbitrary constants $\epsilon > 0$. In learning theory terms, weak agnostic learning of monomials is hard, even if one is allowed to output a hypothesis from the much bigger concept class of halfspaces. This hardness result subsumes a long line of previous results, including two recent hardness results for the proper learning of monomials and halfspaces. As an immediate corollary of our result we show that weak agnostic learning of *decision lists* is NP-hard.

Our techniques are quite different from previous hardness proofs for learning. We define distributions on positive and negative examples for monomials whose first few moments match. We use the *invariance principle* to argue that *regular* halfspaces (all of whose coefficients have small absolute value relative to the total ℓ_2 norm) cannot distinguish between distributions whose first few moments match. For highly non-regular subspaces, we use a structural lemma from recent work on fooling halfspaces to argue that they are "junta-like" and one can zero out all but the top few coefficients without affecting the performance of the halfspace. The top few coefficients form the natural list decoding of a halfspace in the context of dictatorship tests/Label Cover reductions.

We note that unlike previous invariance principle based proofs which are only known to give Unique-Games hardness, we are able to reduce from a version of Label Cover problem that is known to be NP-hard. This has inspired follow-up work on bypassing the Unique Games conjecture in some optimal geometric inapproximability results.

*An extended abstract appeared in the Proceedings of the 50th IEEE Symposium on Foundations of Computer Science, 2009.

†IBM Almaden Research Center, San Jose, CA. vitaly@post.harvard.edu.

‡Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. guruswami@cmu.edu.

§College of Computing, Georgia Institute of Technology, Atlanta, GA. praghave@cc.gatech.edu. Some of this work was done when visiting Carnegie Mellon University.

¶IBM Almaden Research Center, San Jose, CA. wuyi@us.ibm.com. Most of this work was done when the author was at Carnegie Mellon University.

ISSN 1433-8092

1 Introduction

Boolean conjunctions (or *monomials*), decision lists, and halfspaces are among the most basic concept classes in learning theory. They are all long-known to be efficiently PAC learnable, when the given examples are guaranteed to be consistent with a function from any of these concept classes [42, 7, 39]. However, in practice data is often noisy or too complex to be consistently explained by a simple concept. A common practical approach to such problems is to find a predictor in a certain space of hypotheses that best fits the given examples. A general model for learning that addresses this scenario is the *agnostic* learning model [21, 26]. An *agnostic* learning algorithm for a class of functions $\mathcal{C}$ using a hypothesis space $\mathcal{H}$ is required to perform the following task: Given examples drawn from some unknown distribution, the algorithm must find a hypothesis in $\mathcal{H}$ that classifies the examples nearly as well as is possible by a hypothesis from $\mathcal{C}$. The algorithm is said to be a *proper* learning algorithm if $\mathcal{C} = \mathcal{H}$.

In this work we address the complexity of agnostic learning of monomials by algorithms that output a halfspace as a hypothesis. Learning methods that output a halfspace as a hypothesis such as Perceptron [40], Winnow [34], Support Vector Machines [43] as well as most boosting algorithms are well-studied in theory and widely used in practical prediction systems. These classifiers are often applied to labeled data sets which are not linearly separable. Hence it is of great interest to determine the classes of problems that can be solved by such methods in the agnostic setting. In this work we demonstrate a strong negative result on agnostic learning by halfspaces. We prove that non-trivial agnostic learning of even the relatively simple class of monomials by halfspaces is an NP-hard problem.

Theorem 1.1. *For any constant $\epsilon > 0$, it is* NP*-hard to find a halfspace that correctly labels $(1/2 + \epsilon)$-fraction of given examples over $\{0,1\}^n$ even when there exists a monomial that agrees with a $(1 - \epsilon)$-fraction of the examples.*

Note that this hardness result is essentially optimal since it is trivial to find a hypothesis with agreement rate $1/2$ — output either the function that is always 0 or the function that is always 1. Also note that Theorem 1.1 measures agreement of a halfspace and a monomial with the given set of examples rather than the probability of agreement of h with an example drawn randomly from an unknown distribution. Uniform convergence results based on the VC dimension imply that these settings are essentially equivalent (see for example [21, 26]).

The class of monomials is a subset of the class of decision lists which in turn is a subset of the class of halfspaces. Therefore our result immediately implies an optimal hardness result for proper agnostic learning of decision lists.

Previous work

Before describing the details of the prior body of work on hardness results for learning, we note that our result *subsumes all these results* with just one exception (the hardness of learning monomials by t-CNFs [32]). This is because we obtain the optimal inapproximability factor *and* allow learning of monomials by the much richer class of halfspaces.

The results of the paper are noteworthy in the broader context of hardness of approximation. Previously, hardness proofs based on the invariance principle were only known to give Unique-Games

hardness. In this work, we are able to harness invariance principles to show NP-hardness result by working with a version of Label Cover whose projection functions are only required to be *unique-on-average*. This could be one potential approach to revisit the many strong inapproximability results conditioned on the Unique Games conjecture (UGC), with an eye towards bypassing the UGC assumption. Such a goal was achieved for some geometric problems recently [**?**]; see Section 2.3.

Agnostic learning of monomials, decision lists and halfspaces has been studied in a number of previous works. Proper agnostic learning of a class of functions $\mathcal{C}$ is equivalent to the ability to come up with a function in $\mathcal{C}$ which has the optimal agreement rate with the given set of examples and is also referred to as the *Maximum Agreement* problem for a class of function $\mathcal{C}$.

The Maximum Agreement problem for halfspaces is equivalent to the so-called Hemisphere problem and is long known to be NP-complete [23, 17]. Amaldi and Kann [1] showed that Maximum Agreement for halfspaces is NP-hard to approximate within $\frac{261}{262}$ factor. This was later improved by Ben-David *et al.* [5], and Bshouty and Burroughs [9] to approximation factors $\frac{415}{418}$, and $\frac{84}{85}$, respectively. An optimal inapproximability result was established independently by Guruswami and Raghavendra [20] and Feldman *et al.* [15] showing NP-hardness of approximating the Maximum Agreement problem for halfspaces within $(1/2 + \epsilon)$ for every constant $\epsilon > 0$. The reduction in [15] requires examples with real-valued coordinates, whereas the proof in [20] also works for examples drawn from the Boolean hypercube.

The Maximum Agreement problem for monotone monomials was shown to be NP-hard by Angluin and Laird [2], and NP-hardness for general monomials was shown by Kearns and Li [27]. The hardness of approximating the maximum agreement within $\frac{767}{770}$ was shown by Ben-David *et al.* [5]. The factor was subsequently improved to 58/59 by Bshouty and Burroughs [9]. Finally, Feldman *et al.* [14, 15] showed a tight inapproximability result, namely that it is NP-hard to distinguish between the instances where $(1-\epsilon)$-fraction of the labeled examples are consistent with some monomial and instances where every monomial is consistent with at most $(1/2+\epsilon)$-fraction of the examples. Recently, Khot and Saket [32] proved a similar hardness result even when a t-CNF is allowed as output hypothesis for an arbitrary constant t (a t-CNF is the conjunction of several clauses, each of which has at most t literals; a monomial is thus a 1-CNF).

For the concept class of decisions lists, APX-hardness (or hardness to approximate within some constant factor) of the Maximum Agreement problem was shown by Bshouty and Burroughs [9]. As mentioned above, our result subsumes all these results with the exception of [32].

A number of hardness of approximation results are also known for the complementary problem of minimizing disagreement for each of the above concept classes [26, 22, 3, 8, 14, 15]. Another well-known evidence of the hardness of agnostic learning of monomials is that even a non-proper agnostic learning of monomials would give an algorithm for learning DNF — a major open problem in learning theory [33]. Further, Kalai *et al.* proved that even agnostic learning of halfspaces with respect to the uniform distribution implies learning of parities with random classification noise — a long-standing open problem in learning theory and coding [24].

Monomials, decision lists and halfspaces are known to be efficiently learnable in the presence of more benign *random* classification noise [2, 25, 28, 10, 6, 12]. Simple online algorithms like Perceptron and Winnow learn halfspaces when the examples can be separated with a significant *margin* (as is the case if the examples are consistent with a monomial) and are known to be robust

to a very mild amount of adversarial noise [16, 4, 18]. Our result implies that these positive results will not hold when the adversarial noise rate is ϵ for any constant $\epsilon > 0$.

Kalai *et al.* gave the first non-trivial algorithm for agnostic learning monomials in time $2^{\tilde{O}(\sqrt{n})}$ [24]. They also gave a breakthrough result for agnostic learning of halfspaces with respect to the uniform distribution on the hypercube up to any constant accuracy (and analogous results for a number of other settings). Their algorithms output linear thresholds of parities as hypotheses. In contrast, our hardness result is for algorithms that output a halfspace (which is a linear threshold of single variables).

Organization of the paper: We sketch the idea of our proof in Section 2. We define some probability and analytical tools in Section 3. In Section 4 we define the *dictatorship test*, which is an important gadget for the hardness reduction. For the purpose of illustration, we also show why this dictatorship test already suffices to prove Theorem 1.1 assuming the Unique Games Conjecture [29]. In Section 5, we describe a reduction from a variant of the Label Cover problem to prove Theorem 1.1 under the assumption that P $\neq$ NP.

Notation: We use 0 to encode "False" and 1 to encode "True". We denote $\mathrm{pos}(t) : \mathbb{R} \to \{0,1\}$ as the indicator function of whether $t \geqslant 0$; i.e., $\mathrm{pos}(t) = 1$ when $t \geqslant 0$ and $\mathrm{pos}(t) = 0$ when $t < 0$.

For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, $\boldsymbol{w} \in \mathbb{R}^n$, and $\theta \in \mathbb{R}$, a halfspace $h(\boldsymbol{x})$ is a Boolean function of the form $\mathrm{pos}(\boldsymbol{w} \cdot \boldsymbol{x} - \theta)$; a monomial (conjunction) is a function of the form $\bigwedge_{i \in S} s_i$, where $S \subseteq [n]$ and s_i is the literal of x_i which can represent either x_i or $\neg x_i$; a disjunction is a function of the form $\bigvee_{i \in S} s_i$. One special case of monomials is the function $f(x) = x_i$ for some $i \in [n]$, also referred to as the i-th *dictator* function.

2 Proof Overview

We prove Theorem 1.1 by exhibiting a reduction from the k-Label Cover problem, which is a particular variant of the Label Cover problem. The k-Label Cover problem is defined as follows:

Definition 2.1. *For positive integer M, N that $M \geqslant N$ and $k \geqslant 2$, an instance of k-*Label Cover *$\mathcal{L}(G(V,E), M, N, \{\pi^{v,e} | e \in E, v \in e\})$ consists of a k-uniform connected (multi-)hypergraph $G(V,E)$ with vertex set V and an edge multiset E; a set of functions $\{\pi^{v_i,e}\}_{i=1}^k$. Every hyperedge $e = (v_1, \ldots, v_k)$ is associated with a k-tuple of projection functions $\{\pi^{v_i,e}\}_{i=1}^k$ where $\pi^{v_i,e} : [M] \to [N]$.*

A vertex labeling Λ is an assignment of labels to vertices $\Lambda : V \to [M]$. A labeling Λ is said to strongly *satisfy an edge e if $\pi^{v_i,e}(\Lambda(v_i)) = \pi^{v_j,e}(\Lambda(v_j)))$ for every $v_i, v_j \in e$. A labeling Λ* weakly *satisfies edge e if $\pi^{v_i,e}(\Lambda(v_i)) = \pi^{v_j,e}(\Lambda(v_j)))$ for some $v_i, v_j \in e$, $v_i \neq v_j$.*

The goal in Label Cover is to find a vertex labeling that satisfies as many edges (projection constraints) as possible.

2.1 Hardness assuming the Unique Games conjecture

For the sake of clarity, we first sketch the proof of Theorem 1.1 with a reduction from the k-Unique Label Cover problem which is a special case of k-Label Cover where $M = N$ and all the projection functions $\{\pi^{v,e} | v \in e, e \in E\}$ are bijections. The following inapproximability result [**?**] for k-Unique Label Cover is equivalent to the Unique Games Conjecture of Khot [29].

Conjecture 2.2. *For every constant $\eta > 0$ and a positive integer k, there exists an integer R_0 such that for all positive integers $R > R_0$, given an instance $\mathcal{L}(G(V,E), R, R, \{\pi^{v,e} | e \in E, v \in e\})$ it is* NP*-hard to distinguish between,*

- *strongly satisfiable instances: there exists a labeling $\Lambda : V \to [R]$ that strongly satisfies $1 - k\eta$ fraction of the edges E.*
- *almost unsatisfiable instances: there is no labeling that weakly satisfies $\frac{2k^2}{R^{\eta/4}}$ fraction of the edges.*

Given an instance $\mathcal{L}$ of k-Unique Label Cover, we will produce a distribution $\mathcal{D}$ over labeled examples such that the following holds: if $\mathcal{L}$ is a strongly satisfiable instance, then there is a disjunction that agrees with the label on a randomly chosen example with probability at least $1 - \epsilon$, while if $\mathcal{L}$ is an almost unsatisfiable instance then no halfspace agrees with the label on a random example from $\mathcal{D}$ with probability more than $\frac{1}{2} + \epsilon$. Clearly, such a reduction implies Theorem 1.1 assuming the Unique Games Conjecture but with disjunctions in place of conjunctions. De Morgan's law and the fact that a negation of a halfspace is a halfspace then imply that the statement is also true for monomials (we use disjunctions only for convenience).

Let $\mathcal{L}$ be an instance of k-Unique Label Cover on hypergraph $G = (V, E)$ and a set of labels $[R]$. The examples we generate will have $|V| \times R$ coordinates, i.e., belong to $\{0,1\}^{|V| \times R}$. These coordinates are to be thought of as one block of R coordinates for every vertex $v \in V$. We will index the coordinates of $\boldsymbol{x} \in \{0,1\}^{|V| \times R}$ as $\boldsymbol{x} = (x_v^{(r)})_{v \in V, r \in [R]}$.

For every labeling $\Lambda : V \to [R]$ of the instance, there is a corresponding disjunction over $\{0,1\}^{|V| \times R}$ given by,

$$h(\boldsymbol{x}) = \bigvee_v x_v^{(\Lambda(v))}.$$

Thus, using a label r for a vertex v is encoded as including the literal $x_v^{(r)}$ in the disjunction. Notice that an arbitrary halfspace over $\{0,1\}^{|V| \times R}$ need not correspond to any labeling at all. The idea would be to construct a distribution on examples which ensures that any halfspace agreeing with at least $\frac{1}{2} + \epsilon$ fraction of random examples somehow corresponds to a labeling of Λ weakly satisfying a constant fraction of the edges in $\mathcal{L}$.

Fix an edge $e = (v_1, \ldots, v_k)$. For the sake of exposition, let us assume $\pi^{v_i,e}$ is the identity permutation for every $i \in [k]$. The general case is not anymore complicated.

For the edge e, we will construct a distribution on examples $\mathcal{D}_e$ with the following properties:

- All coordinates $x_v^{(r)}$ for a vertex $v \notin e$ are fixed to be zero. Restricted to these examples, the halfspace h can be written as $h(\boldsymbol{x}) = \mathrm{pos}(\sum_{i \in [k]} \langle \boldsymbol{w}_{v_i}, \boldsymbol{x}_{v_i} \rangle - \theta)$.

- For any label $r \in [R]$, the labeling $\Lambda(v_1) = \ldots = \Lambda(v_k) = r$ *strongly* satisfies the edge e. Hence, the corresponding disjunction $\vee_{i\in[k]} x_{v_i}^{(r)}$ needs to have agreement $\geqslant 1-\epsilon$ with the examples from $\mathcal{D}_e$.
- There exists a decoding procedure that given a halfspace h outputs a labeling Λ_h for $\mathcal{L}$ such that, if h has agreement $\geqslant \frac{1}{2}+\epsilon$ with the examples from $\mathcal{D}_e$, then Λ_h *weakly* satisfies the edge e with non-negligible probability.

For conceptual clarity, let us rephrase the above requirement as a testing problem. Given a halfspace h, consider a randomized procedure that samples an example $(\boldsymbol{x}, b)$ from the distribution $\mathcal{D}_e$, and accepts if $h(\boldsymbol{x}) = b$. This amounts to a test that checks if the function h corresponds to a consistent labeling. Further, let us suppose the halfspace h is given by $h(\boldsymbol{x}) = \mathrm{pos}\left(\sum_{v\in V}\langle \boldsymbol{w}_v, \boldsymbol{x}_v\rangle - \theta\right)$. Define the linear function $f_v : \{0,1\}^R \to \mathbb{R}$ as $f_v(\boldsymbol{x}_v) = \langle \boldsymbol{w}_v, \boldsymbol{x}_v\rangle$. Then, we have $h(\boldsymbol{x}) = \mathrm{pos}(\sum_{v\in V} f_v(\boldsymbol{x}_v) - \theta)$.

For a halfspace h corresponding to a labeling Λ, we will have $f_v(\boldsymbol{x}_v) = x_v^{(\Lambda(v))}$ – a dictator function. Thus, in the intended solution every linear function f_v associated with the halfspace h is a dictator function.

Now, let us again restate the above testing problem in terms of these linear functions. For succinctness, we write f_i for the linear function f_{v_i}. We need a randomized procedure that does the following:

> Given k linear functions $f_1, \ldots, f_k : \{0,1\}^R \to \mathbb{R}$, queries the functions at one point each (say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ respectively), and accepts if $\mathrm{pos}(\sum_{i=1}^k f_i(\boldsymbol{x}_i) - \theta) = b$.

The procedure must satisfy,

- (Completeness) If each of the linear functions f_i is the r'th dictator function for some $r \in [R]$, then the test accepts with probability $1-\epsilon$.
- (Soundness) If the test accepts with probability $\frac{1}{2}+\epsilon$, then at least *two* of the linear functions are *close* to the same dictator function.

A testing problem of the above nature is referred to as a *Dictatorship Testing* and is a recurring theme in hardness of approximation.

Notice that the notion of a linear function being *close* to a dictator function is not formally defined yet. In most applications, a function is said to be close to a dictator if it has *influential* coordinates. It is easy to see that this notion is not sufficient by itself here. For example, in the linear function $\mathrm{pos}(10^{100}x_1 + x_2 - 0.5)$, although the coordinate x_2 has little influence on the linear function, it has significant influence on the halfspace.

We resolve this problem by using the notion of *critical index* (Definition 3.1) that was introduced in [41] and has found numerous applications in the analysis of halfspaces [35, 38, 13]. Roughly speaking, given a linear function f, the idea is to recursively delete its influential coordinates until there are none left. The total number of coordinates so deleted is referred to as the critical index of f. Let $c_\tau(\boldsymbol{w}_i)$ denote the critical index of $\boldsymbol{w}_i$, and let $C_\tau(\boldsymbol{w}_i)$ denote the set of $c_\tau(\boldsymbol{w}_i)$ largest coordinates of $\boldsymbol{w}_i$. The linear function l is said to be *close* to the i'th dictator function for every i

in $C_\tau(\boldsymbol{w}_i)$. A function is *far* from every dictator if it has critical index 0 – no influential coordinate to delete.

An important issue is that the critical index of a linear function can be much larger than the number of influential coordinates and cannot be appropriately bounded. In other words, a linear function can be close to a large number of dictator functions, as per the definition above. To counter this, we employ a structural lemma about halfspaces that was used in the recent work on fooling halfspaces with limited independence [13]. Using this lemma, we are able to prove that if the critical index is large, then one can in fact zero out the coordinates of $\boldsymbol{w}_i$ outside the t largest coordinates for some large enough t, and the agreement of the halfspace h only changes by a negligible amount! Thus, we first carry out the zeroing operation for all linear functions with large critical index.

We now describe the above construction and analysis of the dictatorship test in some more detail. It is convenient to think of the k queries $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ as the rows of a $k \times R$ matrix with $\{0,1\}$ entries. Henceforth, we will refer to matrices $\{0,1\}^{k \times R}$ and their rows and columns.

We construct two distributions $\mathcal{D}_0, \mathcal{D}_1$ on $\{0,1\}^k$ such that for $s \in \{0,1\}$, we have $\mathbf{Pr}_{x \in \mathcal{D}_s}\left[\vee_{i=1}^k x_i = s\right] \geqslant 1 - \epsilon/2$ for $\epsilon = o_k(1)$ (this will ensure the completeness of the reduction, i.e., certain disjunctions pass with high probability). Further, the distributions $\mathcal{D}_0, \mathcal{D}_1$ will be carefully chosen to have matching first four moments. This will be used in the soundness analysis where we will use an *invariance principle* to infer structural properties of halfspaces that pass the test with probability noticeably greater than $1/2$.

We define the distribution $\tilde{\mathcal{D}}_s^R$ on matrices $\{0,1\}^{k \times R}$ by sampling R columns independently according to $\mathcal{D}_s$, and then perturbing each bit with a small probability $\epsilon/2$. We define the following test (or equivalently, distribution on examples): given a halfspace h on $\{0,1\}^{k \times R}$, with probability $1/2$ we check $h(\boldsymbol{x}) = 0$ for a sample $\boldsymbol{x} \in \tilde{\mathcal{D}}_0^R$, and with probability $1/2$ we check $h(\boldsymbol{x}) = 1$ for a sample $\boldsymbol{x} \in \tilde{\mathcal{D}}_1^R$.

Completeness: By construction, each of the R disjunctions $\mathsf{OR}_j(\boldsymbol{x}) = \vee_{i=1}^k x_i^{(j)}$ passes the test with probability at least $1 - \epsilon$ (here $x_i^{(j)}$ denotes the entry in the i'th row and j'th column of $\boldsymbol{x}$).

Soundness: For the soundness analysis, suppose $h(\boldsymbol{x}) = \mathrm{pos}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \theta)$ is a halfspace that passes the test with probability at least $1/2 + \epsilon$. The halfspace h can be written in two ways by expanding the inner product $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ along rows and columns, i.e., $h(\boldsymbol{x}) = \mathrm{pos}(\sum_{i=1}^k \langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle - \theta) = \mathrm{pos}(\sum_{i=1}^R \langle \boldsymbol{w}^{(i)}, \boldsymbol{x}^{(i)} \rangle - \theta)$. Let us denote $f_i(\boldsymbol{x}) = \langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle$.

First, let us see why the linear functions $\langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle$ must be close to *some* dictator. Note that we need to show that two of the linear functions are close to the *same* dictator.

Suppose each of the linear functions f_i is not *close* to any dictator. In other words, for each i, no single coordinate of the vector $\boldsymbol{w}_i$ is too large (contains more than τ-fraction of the ℓ_2 mass $\|\boldsymbol{w}_i\|_2$ of vector $\boldsymbol{w}_i$). Clearly, this implies that no single column of the matrix $\boldsymbol{w}$ is too *large*.

Recall that the halfspace is given by $h(\boldsymbol{x}) = \mathrm{pos}(\sum_{j \in [R]} \langle \boldsymbol{w}^{(j)}, \boldsymbol{x}^{(j)} \rangle - \theta)$. Here $l(\boldsymbol{x}) = \sum_{j \in [R]} \langle \boldsymbol{w}^{(j)}, \boldsymbol{x}^{(j)} \rangle - \theta$ is a degree 1 polynomial into which we are substituting values from two product distributions $\mathcal{D}_0^R$ and $\mathcal{D}_1^R$. Further, the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ have matching moments up to order 4 by design. Using the invariance principle, the distribution of $l(\boldsymbol{x})$ is roughly the same, whether $\boldsymbol{x}$ is from $\mathcal{D}_0^R$ or $\mathcal{D}_1^R$. Thus, by the invariance principle, the halfspace h is unable to distinguish between the distributions $\mathcal{D}_0^R$ and $\mathcal{D}_1^R$ with a noticeable advantage.

Further, suppose no two linear functions f_i are *close* to the same dictator, i.e., $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) = \emptyset$. In this case, we condition on the values of $x_i^{(j)}$ for $j \in C_\tau(\boldsymbol{w}_i)$. Since $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) = \emptyset$, this conditions at most *one* value in each column. Therefore, the conditional distribution on each column in cases $\mathcal{D}_0$ and $\mathcal{D}_1$ still have matching first three moments. We thus apply the invariance principle using the fact that after deleting the coordinates in $C_\tau(\boldsymbol{w}_i)$, all the remaining coefficients of the weight vector $\boldsymbol{w}$ are small (by definition of critical index). This implies that $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) \neq \emptyset$ for some two rows i, j and finishes the proof of the soundness claim.

The above consistency-enforcing test almost immediately yields the Unique Games hardness of weak learning disjunctions by halfspaces via standard methods.

2.2 Extending to NP-hardness

To prove NP-hardness as opposed to hardness assuming the Unique Games conjecture, we reduce a version of Label Cover to our problem. This requires a more complicated consistency check, and we have to overcome several additional technical obstacles in the proof.

The main obstacle encountered in transferring the dictatorship test to a Label Cover-based hardness is one that commonly arises for several other problems. Specifically, the projection constraint on an edge $e = (u, v)$ maps a large set of labels $\mathcal{R} = \{r_1, \ldots, r_d\}$ corresponding to a vertex u to a single label r for the vertex v. While composing the Label Cover constraint (u, v) with the dictatorship test, all labels in $\mathcal{R}$ have to be necessarily *equivalent*. In several settings including this work, this requires the coordinates corresponding to labels in $\mathcal{R}$ to be mostly identical! However, on making the coordinates corresponding to $\mathcal{R}$ identical, the prover corresponding to u can determine the identity of edge (u, v), thus completely destroying the soundness of the composition. In fact, the natural extension of the Unique Games-based reduction for MaxCut [31] to a corresponding Label Cover hardness fails primarily for this reason.

Unlike MaxCut or other Unique Games-based reductions, in our case, the soundness of the dictatorship test is required to hold against a specific class of functions, i.e, halfspaces. Harnessing this fact, we execute the reduction starting from a Label Cover instance whose projections are *unique on average*. More precisely, a *smooth* Label Cover (introduced in [30]) is one in which for every vertex u, and a pair of labels r, r', the labels $\{r, r'\}$ project to the same label with a tiny probability over the choice of the edge $e = (u, v)$. Technically, we express the error term in the invariance principle as a certain fourth moment of the coefficients of the halfspace, and use the smoothness to bound this error term for most edges of the Label Cover instance.

2.3 Bypassing the Unique Games conjecture

Unlike previous invariance principle based proofs which are only known to give Unique-Games hardness, we are able to reduce from a version of the Label Cover problem, based on *unique on average* projections, that can be shown to be NP-hard. It is of great interest to find other applications where a *weak uniqueness* property like the smoothness condition mentioned above can be used to convert a Unique-Games hardness result to an unconditional NP-hardness result. Indeed, inspired by the success of this work in avoiding the UGC assumption and using some of our methods, follow-up work has managed to bypass the Unique Games conjecture in some optimal geometric inapproximability results [**?**]. To the best of our knowledge, the results of [**?**] are the first NP-hardness proofs showing a tight inapproximability factor that is related to fundamental

parameters of Gaussian space, and among the small handful of results where optimality of a non-trivial semidefinite programming based algorithm is shown under the assumption P $\neq$ NP. We hope that this paper has thus opened the avenue to convert at least some of the many tight Unique-Games hardness results to NP-hardness results.

3 Preliminaries

In this section, we define two important tools in our analysis: i) critical index, ii) invariance principle.

3.1 Critical Index

The notion of critical index was first introduced by Servedio [41] and plays an important role in the analysis of halfspaces in [35, 38, 13].

Definition 3.1. *Given any real vector $\boldsymbol{w} = (w^{(1)}, w^{(2)}, \ldots, w^{(n)}) \in \mathbb{R}^n$. Reorder the coordinates by decreasing absolute value, i.e., $|w^{(i_1)}| \geqslant |w^{(i_2)}| \geqslant \ldots \geqslant |w^{(i_n)}|$ and denote $\sigma_t^2 = \sum_{j=t}^n |w^{(i_j)}|^2$. For $0 \leqslant \tau \leqslant 1$, the τ-critical index of the vector $\boldsymbol{w}$ is defined to be the smallest index k such $|w^{(i_k)}| \leqslant \tau\sigma_k$. If no such k exists ($\forall k$, $|w^{(i_k)}| > \tau\sigma_k$), the τ-critical index is defined to be $+\infty$. The vector $\boldsymbol{w}$ is said to be τ-regular if the τ-critical index is 1.*

A simple observation from [13] is that if the critical index of a sequence is large then the sequence must contain a geometrically decreasing subsequence.

Lemma 3.2. *(Lemma 5.5 in [13]) Given a vector $\boldsymbol{w} = (w^{(i)})_{i=1}^n$ such that $|w^{(1)}| \geqslant |w^{(2)}| \geqslant \ldots \geqslant |w^{(n)}|$, if the τ-critical index of the vector $\boldsymbol{w}$ is larger than l, then for any $1 \leqslant i \leqslant j \leqslant l+1$,*

$$|w^{(j)}| \leqslant \sigma_j \leqslant (\sqrt{1-\tau^2})^{j-i}\sigma_i \leqslant (\sqrt{1-\tau^2})^{j-i}|w^{(i)}|/\tau.$$

In particular, if $j > i + (4/\tau^2)\ln(1/\tau)$ then $|w^{(j)}| \leqslant |w^{(i)}|/3$.

For a τ-regular weight vector, the following lemma bounds the probability that its weighted sum falls into a small interval under certain distributions on the points. The proof is in Appendix B.

Lemma 3.3. *Let $\boldsymbol{w} \in \mathbb{R}^n$ be a τ-regular vector $\boldsymbol{w}$, and $\sum |w^{(i)}|^2 = 1$. $\mathcal{D}$ is a distribution over $\{0,1\}^n$. Define a distribution $\tilde{\mathcal{D}}$ on $\{0,1\}^n$ as follows: to generate $\boldsymbol{y}$ from $\tilde{\mathcal{D}}$, first sample $\boldsymbol{x}$ from $\mathcal{D}$ and then define,*

$$y^{(i)} = \begin{cases} x^{(i)} & \text{with probability } 1-\gamma \\ \text{random bit} & \text{with probability } \gamma. \end{cases}$$

Then for any interval $[a,b]$, we have

$$\mathbf{Pr}\Big[\langle \boldsymbol{w}, \boldsymbol{y}\rangle \in [a,b]\Big] \leqslant \frac{4|b-a|}{\sqrt{\gamma}} + \frac{4\tau}{\sqrt{\gamma}} + 2e^{-\frac{\gamma^2}{2\tau^2}}.$$

Intuitively, by the Berry-Esseen Theorem, $\langle \boldsymbol{w}, \boldsymbol{y} \rangle$ is τ close to the Gaussian distribution if each $y^{(i)}$ is a random bit; therefore we can bound the probability that $\langle \boldsymbol{w}, \boldsymbol{y} \rangle$ falls into the interval $[a, b]$. In above lemma, each $y^{(i)}$ has probability γ to be a random bit, then γ fraction of $y^{(i)}$ is set to be a random bit and we can similarly bound the probability that $\langle \boldsymbol{w}, \boldsymbol{y} \rangle$ falls into the interval $[a, b]$.

Definition 3.4. *For a vector $\boldsymbol{w} \in \mathbb{R}^n$, define set of indices $H_t(\boldsymbol{w}) \subseteq [n]$ as the set of indices containing the t biggest coordinates of $\boldsymbol{w}$ by absolute value. Suppose its τ-critical index is c_τ, define set of indices $C_\tau(\boldsymbol{w}) = H_{c_\tau}(\boldsymbol{w})$. In other words, $C_\tau(\boldsymbol{w})$ is the set of indices whose deletion makes the vector w to be τ-regular.*

Definition 3.5. *For a vector $\boldsymbol{w} \in \mathbb{R}^n$ and a subset of indices $S \subseteq [n]$, define the vector $\mathsf{Truncate}(\boldsymbol{w}, S) \in \mathbb{R}^n$ as:*

$$(\mathsf{Truncate}(\boldsymbol{w}, S))^{(i)} = \begin{cases} w^{(i)} & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases}$$

As suggested by Lemma 3.2, a weight vector with a large critical index has a geometrically decreasing subsequence. The following two lemmas use this fact to bound the probability that the weighted sum of a geometrically decreasing sequence of weights falls into a small interval. First, we restate Claim 5.7 from [13] here.

Lemma 3.6. *[Claim 5.7, [13]] Let $\boldsymbol{w} = (w^{(1)}, \ldots, w^{(T)})$ be such that $|w^{(1)}| \geqslant |w^{(2)}| \ldots \geqslant |w^{(T)}| \geqslant 0$ and $|w^{(i+1)}| \leqslant |\frac{w^{(i)}}{3}|$ for $1 \leqslant i \leqslant T-1$. Then for any interval $I = [\alpha - \frac{w^{(T)}}{6}, \alpha + \frac{w^{(T)}}{6}]$ of length $\frac{|w^{(T)}|}{3}$, there is at most one point $\boldsymbol{x} \in \{0,1\}^T$ such that $\langle \boldsymbol{w}, \boldsymbol{x} \rangle \in I$.*

Lemma 3.7. *Let $\boldsymbol{w} = (w^{(1)}, \ldots, w^{(T)})$ be such that $|w^{(1)}| \geqslant |w^{(2)}| \ldots \geqslant |w^{(T)}| \geqslant 0$ and $|w^{(i+1)}| \leqslant |\frac{w^{(i)}}{3}|$ for $1 \leqslant i \leqslant T-1$. Let $\mathcal{D}$ be a distribution over $\{0,1\}^T$. Define a distribution $\tilde{\mathcal{D}}$ on $\{0,1\}^T$ as follows: To generate $\boldsymbol{y}$ from $\tilde{\mathcal{D}}$, sample $\boldsymbol{x}$ from $\mathcal{D}$ and set*

$$y^{(i)} = \begin{cases} x^{(i)} & \text{with probability } 1-\gamma \\ \text{random bit} & \text{with probability } \gamma. \end{cases}$$

Then for any $\theta \in \mathbb{R}$ we have

$$\mathbf{Pr}\Big[\langle \boldsymbol{w}, \boldsymbol{y} \rangle \in [\theta - \frac{w^{(T)}}{6}, \theta + \frac{w^{(T)}}{6}]\Big] \leqslant \Big(1 - \frac{\gamma}{2}\Big)^T.$$

Proof. By Lemma 3.6, we know that for the interval $J = \Big[\theta - \frac{|w^T|}{6}, \theta + \frac{|w^T|}{6}\Big]$, there is at most one point $\boldsymbol{r} \in \{0,1\}^T$ such that $\langle \boldsymbol{w}, \boldsymbol{r} \rangle \in J$. If no such $\boldsymbol{r}$ exists then clearly the probability is zero. On the other hand, suppose there exists such an $\boldsymbol{r}$, then $\langle \boldsymbol{w}, \boldsymbol{y} \rangle \in J$ only if $(y_1^{(1)}, y_1^{(2)}, \ldots, y_1^{(T)}) = (r^{(1)}, \ldots, r^{(T)})$ holds.

Conditioned on any fixing of the bits $\boldsymbol{x}$, every bit $y^{(j)}$ is an independent random bit with probability γ. Therefore, for every fixing of $\boldsymbol{x}$, for each $i \in [T]$, with probability at least $\gamma/2$, $y^{(i)}$ is not equal to $r^{(i)}$. Therefore, $\mathbf{Pr}[y^{(1)} = r^{(1)}, y^{(2)} = r^{(2)}, \ldots, y^{(T)} = r^{(T)}] \leqslant \big(1 - \frac{\gamma}{2}\big)^T$. □

3.2 Invariance Principle

While invariance principles have been shown in various settings by [37, 11, 36], we restate a version of the principle well suited for our application. We present a self-contained proof for it in Appendix C.

Definition 3.8. *A function $\Psi(x) : \mathbb{R} \to \mathbb{R}$ for which fourth-order derivatives exist everywhere on $\mathbb{R}$ is said to be K-bounded if $|\Psi''''(t)| \leqslant K$ for all $t \in \mathbb{R}$.*

Definition 3.9. *Two ensembles of random variables $\mathcal{P} = (p_1, \ldots, p_k)$ and $\mathcal{Q} = (q_1, \ldots, q_k)$ are said to have matching moments up to degree d if for every multi-set S of elements from $[k]$, $|S| \leqslant d$, we have $\mathbf{E}[\prod_{i \in S} p_i] = \mathbf{E}[\prod_{i \in S} q_i]$.*

Theorem 3.10. *(Invariance Principle) Let $\mathcal{A} = \{\boldsymbol{A}^{\{1\}}, \ldots, \boldsymbol{A}^{\{R\}}\}, \mathcal{B} = \{\boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{R\}}\}$ be families of ensembles of random variables with $\boldsymbol{A}^{\{i\}} = \{a_1^{(i)}, \ldots, a_{k_i}^{(i)}\}$ and $\boldsymbol{B}^{\{i\}} = \{b_1^{(i)}, \ldots, b_{k_i}^{(i)}\}$, satisfying the following properties:*

- *For each $i \in [R]$, the random variables in ensembles $(\boldsymbol{A}^{\{i\}}, \boldsymbol{B}^{\{i\}})$ have matching moments up to degree 3. Further all the random variables in $\mathcal{A}$ and $\mathcal{B}$ are bounded by 1.*
- *The ensembles $\boldsymbol{A}^{\{i\}}$ are all independent of each other, similarly the ensembles $\boldsymbol{B}^{\{i\}}$ are independent of each other.*

Given a set of vectors $\boldsymbol{l} = \{\boldsymbol{l}^{\{1\}}, \ldots, \boldsymbol{l}^{\{R\}}\} (\boldsymbol{l}^{\{i\}} \in \mathbb{R}^{k_i})$, define the linear function $\boldsymbol{l} : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_R} \to \mathbb{R}$ as

$$\boldsymbol{l}(\boldsymbol{x}) = \sum_{i \in [R]} \langle \boldsymbol{l}^{\{i\}}, \boldsymbol{x}^{\{i\}} \rangle$$

Then for a K-bounded function $\Psi : \mathbb{R} \to \mathbb{R}$ we have

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{A}) - \theta \Big) \right] - \mathop{\mathbf{E}}_{\mathcal{B}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{B}) - \theta \Big) \right] \right| \leqslant K \sum_{i \in [R]} \|\boldsymbol{l}^{\{i\}}\|_1^4$$

for all $\theta > 0$. Further, define the spread function $c(\alpha)$ corresponding to the ensembles $\mathcal{A}, \mathcal{B}$ and the linear function $\boldsymbol{l}$ as follows,

*(**Spread Function:**)For $1/2 > \alpha > 0$, let*

$$c(\alpha) = \max\Big(\sup_{\theta} \mathbf{Pr}_{\mathcal{A}} \Big[\boldsymbol{l}(\mathcal{A}) \in [\theta - \alpha, \theta + \alpha] \Big], \quad \sup_{\theta} \mathbf{Pr}_{\mathcal{B}} \Big[\boldsymbol{l}(\mathcal{B}) \in [\theta - \alpha, \theta + \alpha] \Big] \Big)$$

then for all θ,

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\mathrm{pos}\left(\boldsymbol{l}(\mathcal{A}) - \theta \right) \right] - \mathop{\mathbf{E}}_{\mathcal{B}} \left[\mathrm{pos}\left(\boldsymbol{l}(\mathcal{B}) - \theta \right) \right] \right| \leqslant O\left(\frac{1}{\alpha^4} \right) \sum_{i \in [R]} \|\boldsymbol{l}^{\{i\}}\|_1^4 + 2c(\alpha).$$

Roughly speaking, the second part of the theorem states that pos function can be thought of as $\frac{1}{\alpha^4}$-bounded with error parameter $c(\alpha)$.

4 Construction of the Dictatorship Test

In this section we describe the construction of the dictatorship test which will be the key ingredient in the hardness reduction from k-Unique Label Cover.

4.1 Distributions $\mathcal{D}_0$ and $\mathcal{D}_1$

The dictatorship test is based on following two distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ defined on $\{0,1\}^k$.

Lemma 4.1. *For $k \in \mathbb{N}$, there exists two probability distributions $\mathcal{D}_0$, $\mathcal{D}_1$ on $\{0,1\}^k$ such that for $x = (x_1, \ldots, x_k)$,*

$$\mathbf{Pr}_{x\sim\mathcal{D}_0}\{\textit{every } x_l \textit{ is } 0\} \geqslant 1 - \frac{2}{\sqrt{k}} \textit{ and } \mathbf{Pr}_{x\sim\mathcal{D}_1}\{\textit{every } x_l \textit{ is } 0\} \leqslant \frac{1}{\sqrt{k}},$$

while matching moments up to degree 4, i.e., $\forall i,j,m,n \in [k]$

$$\mathop{\mathbf{E}}_{\mathcal{D}_0}[x_i] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_i] \qquad \mathop{\mathbf{E}}_{\mathcal{D}_0}[x_i x_j x_m x_n] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_i x_j x_m x_n]$$
$$\mathop{\mathbf{E}}_{\mathcal{D}_0}[x_i x_j] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_i x_j] \qquad \mathop{\mathbf{E}}_{\mathcal{D}_0}[x_i x_j x_m] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_i x_j x_m]$$

Proof. For $\epsilon = \frac{1}{\sqrt{k}}$, take $\mathcal{D}_1$ to be the following distribution:

1. with probability $(1-\epsilon)$, randomly set exactly one of the bit to be 1 and all the other to be 0;
2. with probability $\frac{\epsilon}{4}$, independently set every bit to be 1 with probability $\frac{1}{k^{1/3}}$;
3. with probability $\frac{\epsilon}{4}$, independently set every bit to be 1 with probability $\frac{2}{k^{1/3}}$;
4. with probability $\frac{\epsilon}{4}$, independently set every bit to be 1 with probability $\frac{3}{k^{1/3}}$;
5. with probability $\frac{\epsilon}{4}$, independently set every bit to be 1 with probability $\frac{4}{k^{1/3}}$.

The distribution $\mathcal{D}_0$ is defined to be the following distribution with parameter $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ to be specified later:

1. with probability $1 - (\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4)$, set every bit to be zero;
2. with probability ϵ_1, independently set every bit to be 1 with probability $\frac{1}{k^{1/3}}$;
3. with probability ϵ_2, independently set every bit to be 1 with probability $\frac{2}{k^{1/3}}$;
4. with probability ϵ_3, independently set every bit to be 1 with probability $\frac{3}{k^{1/3}}$;
5. with probability ϵ_4, independently set every bit to be 1 with probability $\frac{4}{k^{1/3}}$.

From the definition of $\mathcal{D}_0, \mathcal{D}_1$, we know that $\mathbf{Pr}_{x\sim\mathcal{D}_0}[\text{every } x_i \text{ is } 0] \geqslant 1-(\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4)$ and $\mathbf{Pr}_{x\sim\mathcal{D}_1}[\text{every } x_i \text{ is } 0] \leqslant \epsilon = \frac{1}{\sqrt{k}}$.

It remains to determine each ϵ_i. Notice that the moment matching conditions can be expressed as a linear system over the parameters $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ as follows:

$$\sum_{i=1}^{4} \epsilon_i (\frac{i}{k^{1/3}}) = (1-\epsilon)/k + \sum_{i=1}^{4} \frac{\epsilon}{4} (\frac{i}{k^{\frac{1}{3}}})$$
$$\sum_{i=1}^{4} \epsilon_i (\frac{i}{k^{1/3}})^2 = \sum_{i=1}^{4} \frac{\epsilon}{4} (\frac{i}{k^{\frac{1}{3}}})^2$$
$$\sum_{i=1}^{4} \epsilon_i (\frac{i}{k^{\frac{1}{3}}})^3 = \sum_{i=1}^{4} \frac{\epsilon}{4} (\frac{i}{k^{\frac{1}{3}}})^3$$
$$\sum_{i=1}^{4} \epsilon_i (\frac{i}{k^{\frac{1}{3}}})^4 = \sum_{i=1}^{4} \frac{\epsilon}{4} (\frac{i}{k^{\frac{1}{3}}})^4.$$

We then show that such a linear system has a feasible solution $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0$ and $\sum_{i=1}^{4} \epsilon_i \leqslant 2/\sqrt{k}$.

To prove this, by applying Cramer's rule,

$$\epsilon_1 = \frac{\begin{vmatrix} (1-\epsilon)/k + \sum_{i=1}^{4} \frac{\epsilon}{4}(\frac{i}{k^{\frac{1}{3}}}) & \frac{2}{k^{\frac{1}{3}}} & \frac{3}{k^{\frac{1}{3}}} & \frac{4}{k^{\frac{1}{3}}} \\ \sum_{i=1}^{4} \frac{\epsilon}{4}(\frac{i}{k^{\frac{1}{3}}})^2 & \frac{4}{k^{\frac{2}{3}}} & \frac{9}{k^{\frac{2}{3}}} & \frac{16}{k^{\frac{2}{3}}} \\ \sum_{i=1}^{4} \frac{\epsilon}{4}(\frac{i}{k^{\frac{1}{3}}})^3 & \frac{8}{k^{\frac{3}{3}}} & \frac{27}{k^{\frac{3}{3}}} & \frac{64}{k^{\frac{3}{3}}} \\ \sum_{i=1}^{4} \frac{\epsilon}{4}(\frac{i}{k^{\frac{1}{3}}})^4 & \frac{16}{k^{\frac{4}{3}}} & \frac{81}{k^{\frac{4}{3}}} & \frac{256}{k^{\frac{4}{3}}} \end{vmatrix}}{\begin{vmatrix} \frac{1}{k^{\frac{1}{3}}} & \frac{2}{k^{\frac{1}{3}}} & \frac{3}{k^{\frac{1}{3}}} & \frac{4}{k^{\frac{1}{3}}} \\ \frac{1}{k^{\frac{2}{3}}} & \frac{4}{k^{\frac{2}{3}}} & \frac{9}{k^{\frac{2}{3}}} & \frac{16}{k^{\frac{2}{3}}} \\ \frac{1}{k^{\frac{3}{3}}} & \frac{8}{k^{\frac{3}{3}}} & \frac{27}{k^{\frac{3}{3}}} & \frac{64}{k^{\frac{3}{3}}} \\ \frac{1}{k^{\frac{4}{3}}} & \frac{16}{k^{\frac{4}{3}}} & \frac{81}{k^{\frac{4}{3}}} & \frac{256}{k^{\frac{4}{3}}} \end{vmatrix}}$$

With some calculation using basic linear algebra, we get

$$\epsilon_1 = \epsilon/4 + \frac{\begin{vmatrix} (1-\epsilon)/k & \frac{2}{k^{\frac{1}{3}}} & \frac{3}{k^{\frac{1}{3}}} & \frac{4}{k^{\frac{1}{3}}} \\ 0 & \frac{4}{k^{\frac{2}{3}}} & \frac{9}{k^{\frac{2}{3}}} & \frac{16}{k^{\frac{2}{3}}} \\ 0 & \frac{8}{k^{\frac{3}{3}}} & \frac{3}{k^{\frac{3}{3}}} & \frac{64}{k^{\frac{3}{3}}} \\ 0 & \frac{16}{k^{\frac{4}{3}}} & \frac{3}{k^{\frac{4}{3}}} & \frac{256}{k^{\frac{4}{3}}} \end{vmatrix}}{\begin{vmatrix} \frac{1}{k^{\frac{1}{3}}} & \frac{2}{k^{\frac{1}{3}}} & \frac{3}{k^{\frac{1}{3}}} & \frac{4}{k^{\frac{1}{3}}} \\ \frac{1}{k^{\frac{2}{3}}} & \frac{4}{k^{\frac{2}{3}}} & \frac{9}{k^{\frac{2}{3}}} & \frac{16}{k^{\frac{2}{3}}} \\ \frac{1}{k^{\frac{3}{3}}} & \frac{8}{k^{\frac{3}{3}}} & \frac{27}{k^{\frac{3}{3}}} & \frac{64}{k^{\frac{3}{3}}} \\ \frac{1}{k^{\frac{4}{3}}} & \frac{16}{k^{\frac{4}{3}}} & \frac{81}{k^{\frac{4}{3}}} & \frac{256}{k^{\frac{4}{3}}} \end{vmatrix}} = \frac{1}{4\sqrt{k}} + O(\frac{1}{k^{\frac{2}{3}}}).$$

For large enough k, we have $0 \leqslant \epsilon_1 \leqslant \frac{1}{2\sqrt{k}}$. By similar calculation, we can bound $\epsilon_2, \epsilon_3, \epsilon_4$ by $\frac{1}{2\sqrt{k}}$. Overall, we have $\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 \leqslant 2/\sqrt{k}$

□

We define a "noisy" version of $\mathcal{D}_b$ ($b \in \{0,1\}$) below.

Definition 4.2. *For $b \in \{0,1\}$, define the distribution $\tilde{\mathcal{D}}_b$ on $\{0,1\}^k$ as follows:*

- *First generate $x \in \{0,1\}^k$ according to $\mathcal{D}_b$.*
- *For each $i \in [k]$,*

$$y_i = \begin{cases} x_i & \text{with probability } 1 - \frac{1}{k^2} \\ \text{uniform random bit } u_i & \text{with probability } \frac{1}{k^2} \end{cases}$$

Observation 4.3. *$\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ also have matching moments up to degree 4.*

Proof. Since the noise is defined to be an independent uniform random bit, when calculating moments of y, such as $\mathbf{E}_{\tilde{D}_b}[y_{i_1} y_{i_2} \cdots y_{i_d}]$, we can substitute y_i by $(1-\gamma)x_i + \frac{1}{2}\gamma$. Therefore, a degree d moment of y can be expressed as a weighted sum of moments of x of degree up to d. Since $\mathcal{D}_0$ and $\mathcal{D}_1$ have matching moments up to degree 4, it follows that $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ also have the same property. □

The following simple lemma asserts that conditioning the two distributions $\tilde{D}_0$ and $\tilde{D}_1$ on the same coordinate x_j being fixed to value b results in conditional distributions that still have matching moments up to degree 3.

Lemma 4.4. *Given two distributions $\mathcal{P}_0, \mathcal{P}_1$ on $\{0,1\}^k$ with matching moments up to degree d, for any multi-set S of elements from $[k]$, $|S| \leqslant d-1$, $j \in [k]$ and $c \in \{0,1\}$.*

$$\mathop{\mathbf{E}}_{\mathcal{P}_0}[\prod_{i \in S} x_i \mid x_j = c] = \mathop{\mathbf{E}}_{\mathcal{P}_1}[\prod_{i \in S} x_i \mid x_j = c].$$

Proof. For the case $c = 1$ and any $b \in \{0,1\}$,

$$\mathop{\mathbf{E}}_{\mathcal{P}_b}[x_j \prod_{i \in S} x_i] = \mathop{\mathbf{E}}_{\mathcal{P}_b}[\prod_{i \in S} x_i \mid x_j = 1]\mathbf{Pr}_{\mathcal{P}_0}[x_j = 1] = \mathop{\mathbf{E}}_{\mathcal{P}_b}[\prod_{i \in S} x_i \mid x_j = 1] \mathop{\mathbf{E}}_{\mathcal{P}_0}[x_j].$$

Therefore,

$$\mathop{\mathbf{E}}_{\mathcal{P}_0}[\prod_{i \in S} x_i \mid x_j = 1] = \frac{\mathbf{E}_{\mathcal{P}_0}[x_j \prod_{i \in S} x_i]}{\mathbf{E}_{\mathcal{P}_0}[x_j]} = \frac{\mathbf{E}_{\mathcal{P}_1}[x_j \prod_{i \in S} x_i]}{\mathbf{E}_{\mathcal{P}_1}[x_j]} = \mathop{\mathbf{E}}_{\mathcal{P}_1}[\prod_{i \in S} x_i \mid x_j = 1].$$

For the case $c = 0$, replace x_j with $x'_j = 1 - x_j$. It is easy to see that $\mathcal{P}_0$ and $\mathcal{P}_1$ still have matching moments and conditioning on $x_j = 0$ is the same as conditioning on $x'_j = 1$. Hence we can reduce to the case $c = 1$. □

4.2 The Dictatorship Test

Let R be a positive integer. Based on the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$, we define the dictatorship test as follows:

1. Generate a random bit $b \in \{0,1\}$.
2. Generate $\boldsymbol{x} \in \{0,1\}^{kR}$ (which is also written as $\{x_i^{(j)}\}_{i\in[k],j\in[R]}$) from $\mathcal{D}_b^R$.
3. For each $i \in [k], j \in [R]$,
$$y_i^{(j)} = \begin{cases} x_i^{(j)} & \text{with probability } 1-\frac{1}{k^2}; \\ \text{random bit} & \text{with probability } \frac{1}{k^2}. \end{cases}$$
4. Output the labelled example $(\boldsymbol{y}, b)$. Equivalently, if h denotes the halfspace, ACCEPT if $h(\boldsymbol{y}) = b$.

We can also view y as being generated as follows: i) With probability $\frac{1}{2}$, generate a negative sample from distribution $\tilde{\mathcal{D}}_0^R$; ii) With probability $\frac{1}{2}$, generate a positive sample from distribution $\tilde{\mathcal{D}}_1^R$.

The dictatorship test has the following completeness and soundness properties.

Theorem 4.5. *(completeness) For any $j \in [R]$, $h(\boldsymbol{y}) = \vee_{i=1}^k y_i^{(j)}$ passes with probability $\geqslant 1 - \frac{3}{\sqrt{k}}$.*

Theorem 4.6. *(soundness) Fix $\tau = \frac{1}{k^7}$ and $t = \frac{1}{\tau^2}(3\ln(1/\tau) + \ln R) + \lceil 4k^2 \ln k \rceil \lceil \frac{4}{\tau^2}\ln(1/\tau)\rceil$. Let $h(\boldsymbol{x}) = \mathrm{pos}(\langle \boldsymbol{w}, \boldsymbol{y}\rangle - \theta)$ be a halfspace such that $H_t(\boldsymbol{w}_i) \cap H_t(\boldsymbol{w}_j) = \emptyset$ for all $i, j \in [k]$. Then the halfspace $h(\boldsymbol{y})$ passes the dictatorship test with probability at most $\frac{1}{2} + O(\frac{1}{k})$.*

Proof. (Theorem 4.5) If x is generated from $\mathcal{D}_0^R$, we know that with probability at least $1 - \frac{2}{\sqrt{k}}$, all the bits in $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_k^{(j)}\}$ are set to 0. By union bound, with probability at least $1 - \frac{2}{\sqrt{k}} - \frac{1}{k}$, $\{y_1^{(j)}, y_2^{(j)}, \ldots, y_k^{(j)}\}$ are all set to 0, in which case the test passes as $\vee_{i=1}^k y_i^{(j)} = 0$. If x is generated from $\mathcal{D}_1^R$, we know that with probability at least $1 - \frac{1}{\sqrt{k}}$, one of the bits in $\{x_1^{(j)}, x_2^{(j)}, \ldots, x_k^{(j)}\}$ is set to 1 and by union bound one of $\{y_1^{(j)}, y_2^{(j)}, \ldots, y_k^{(j)}\}$ is set to 1 with probability at least $1 - \frac{1}{\sqrt{k}} - \frac{1}{k}$, in which case the test passes since $\vee_{i=1}^k y_i^{(j)} = 1$. Overall, the test passes with probability at least $1 - \frac{3}{\sqrt{k}}$. □

4.3 Proof of Soundness (Theorem 4.6)

We will prove the contrapositive statement of Theorem 4.6: if some $h(\boldsymbol{y})$ passes the above dictatorship test with high probability, then we can decode for each $\boldsymbol{w}_i$ ($i \in [k]$), a small list of coordinates and at least two of the lists will intersect.

The proof is based on two key lemmas (Lemmas 4.7, 4.8). The first lemma states that if a halfspace passes the test with good probability, then two of its critical index sets $C_\tau(\boldsymbol{w}_i), C_\tau(\boldsymbol{w}_j)$

must intersect. This would immediately imply Theorem 4.6 if c_τ is less than t. The second lemma states that every halfspace can be approximated by another halfspace with critical index less than t; so we can assume that c_τ is small without loss of generality.

Let $h(\boldsymbol{y})$ be a halfspace function on $\{0,1\}^{kR}$ given by $h(\boldsymbol{y}) = \mathrm{pos}(\langle \boldsymbol{w}, \boldsymbol{y}\rangle - \theta)$. Equivalently, $h(\boldsymbol{y})$ can be written as

$$h(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{j\in[R]} \langle \boldsymbol{w}^{(j)}, \boldsymbol{y}^{(j)}\rangle - \theta\Big) = \mathrm{pos}\Big(\sum_{i\in[k]} \langle \boldsymbol{w}_i, \boldsymbol{y}_i\rangle - \theta\Big)\ ,$$

where $\boldsymbol{w}^{(j)} \in \mathbb{R}^k$ and $\boldsymbol{w}_i \in \mathbb{R}^R$.

Lemma 4.7. *(Common Influential Coordinates) For $\tau = \frac{1}{k^7}$, let $h(\boldsymbol{y})$ be a halfspace such that for all $i \neq j \in [k]$, we have $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) = \emptyset$. Then*

$$\Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant O\Big(\frac{1}{k}\Big)\ .$$

Proof. Fix the following notation,

$$\begin{aligned} \boldsymbol{l}_i &= \mathsf{Truncate}(\boldsymbol{w}_i, C_\tau(\boldsymbol{w}_i)) & \boldsymbol{s}_i &= \boldsymbol{w}_i - \boldsymbol{w}_i^C \\ \boldsymbol{y}_i^C &= \mathsf{Truncate}(\boldsymbol{y}_i, C_\tau(\boldsymbol{w}_i)) & \boldsymbol{y}^C &= \boldsymbol{y}_1^C, \boldsymbol{y}_2^C, \ldots, \boldsymbol{y}_k^C \\ \boldsymbol{s} &= \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k & \boldsymbol{l} &= \boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_k. \end{aligned}$$

We can rewrite the halfspace $h(\boldsymbol{y})$ as $h(\boldsymbol{y}) = \mathrm{pos}\Big(\langle \boldsymbol{l}, \boldsymbol{y}^C\rangle + \langle \boldsymbol{s}, \boldsymbol{y}\rangle - \theta\Big)$. Let us first normalize the halfspace $h(\boldsymbol{y})$ so that $\sum_{i\in[k]} \|\boldsymbol{l}_i\|^2 = 1$. We now condition on a possible fixing of the vector $\boldsymbol{y}^C$. Under this conditioning and for $\boldsymbol{y}$ chosen randomly from the distribution $\tilde{\mathcal{D}}_0^R$, define the family of ensembles $\mathcal{A} = \boldsymbol{A}^{\{1\}}, \ldots, \boldsymbol{A}^{\{R\}}$ as follows:

$$\boldsymbol{A}^{\{j\}} = \{y_i^{(j)} |\ i \in [k] \text{ for which } j \notin C_\tau(\boldsymbol{w}_i)\}$$

Similarly define the ensemble $\mathcal{B} = \boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{R\}}$ using $\boldsymbol{y}$ chosen randomly from the distribution $\tilde{\mathcal{D}}_1^R$. Further let us denote $\boldsymbol{l}^{\{j\}} = (l_1^{(j)}, \ldots, l_k^{(j)})$. Now we apply the invariance principle (Theorem 3.10) to the ensembles $\mathcal{A}, \mathcal{B}$ and the linear function $\boldsymbol{l}$. For each $j \in [R]$, there is at most one coordinate $i \in [k]$ such that $j \in C_\tau(\boldsymbol{w}_i)$. Thus, conditioning on $\boldsymbol{y}^C$ amounts to fixing of at most one variable $y_i^{(j)}$ in each column $\{y_i^{(j)}\}_{i\in[k]}$. By Lemma 4.4, since $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ have matching moments up to degree 4, we get that $\boldsymbol{A}^{\{j\}}$ and $\boldsymbol{B}^{\{j\}}$ have matching moments up to degree 3. Also notice that $\max_{j\in[R],i\in[k]} |l_i^{(j)}| \leqslant \tau\|\boldsymbol{l}_i\|_2 \leqslant \tau\|\boldsymbol{l}\|_2$ (as $\boldsymbol{l}_i$ is a τ-regular) and each $y_i^{(j)}$ is set to be a random unbiased bit with probability $\frac{1}{k^2}$; by Lemma 3.3, the linear function $\boldsymbol{l}$ and the ensembles $\mathcal{A}$, $\mathcal{B}$ satisfy the following spread property for every $\theta' \in \mathbb{R}$:

$$\begin{aligned} \mathbf{Pr}_{\mathcal{A}}\Big[\boldsymbol{l}(\mathcal{A}) \in [\theta' - \alpha, \theta' + \alpha]\Big] &\leqslant c(\alpha) \\ \mathbf{Pr}_{\mathcal{B}}\Big[\boldsymbol{l}(\mathcal{B}) \in [\theta' - \alpha, \theta' + \alpha]\Big] &\leqslant c(\alpha), \end{aligned}$$

where $c(\alpha) \leqslant 8\alpha k + 4\tau k + 2e^{-\frac{1}{2\tau^2 k^4}}$ (by setting $\gamma = \frac{1}{k^2}$ and $|b-a| = 2\alpha$ in Lemma 3.3). Using the invariance principle (Theorem 3.10) this implies:

$$\Big|\mathop{\mathbf{E}}_{\mathcal{A}}\Big[\mathrm{pos}\Big(\langle \boldsymbol{s}, \boldsymbol{y}^C\rangle + \sum_{j\in[R]}\langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}}\rangle - \theta\Big)\Big|\boldsymbol{y}^C\Big] - \mathop{\mathbf{E}}_{\mathcal{B}}\Big[\mathrm{pos}\Big(\langle \boldsymbol{s}, \boldsymbol{y}^C\rangle + \sum_{j\in[R]}\langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}}\rangle - \theta\Big)\Big|\boldsymbol{y}^C\Big]\Big| \leqslant O\Big(\frac{1}{\alpha^4}\Big)\sum_{i\in[R]}\|\boldsymbol{l}^{\{i\}}\|_1^4 + 2c(\alpha) \quad (1)$$

By definition of the critical index, we have $\max_{j\in[R]} l_i^j \leqslant \tau\|\boldsymbol{l}_i\|_2$. Using this, we can bound $\sum_{i\in[R]}\|\boldsymbol{l}^{\{i\}}\|_1^4$ as follows:

$$\sum_{j\in[R]}\|\boldsymbol{l}^{\{j\}}\|_1^4 \leqslant k^4\sum_{i\in[k]}\sum_{j\in[R]}\|l_i^{(j)}\|^4 \leqslant k^4\sum_{i\in[k]}\Big(\max_{j\in[R]}|l_i^{(j)}|^2\Big)\|\boldsymbol{l}_i\|_2^2 \leqslant k^4\tau^2\sum_{i\in[k]}\|\boldsymbol{l}_i\|_2^2 \leqslant k^4\tau^2\|\boldsymbol{l}\|_2^2 \leqslant \frac{1}{k^{10}}.$$

In the final inequality in above calculation, we used the fact that $\tau = \frac{1}{k^7}$ and $\|\boldsymbol{l}\|_2 = 1$. Let us choose $\alpha = \frac{1}{k^2}$ and (1) is therefore bounded by $O(1/k)$ for all settings of $\boldsymbol{y}^C$. Averaging over all settings of $\boldsymbol{y}^C$ we get that

$$\Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant O\left(\frac{1}{k}\right).$$

□

The above lemma asserts that unless some two vectors $\boldsymbol{w}_i, \boldsymbol{w}_j$ have a *common influential coordinate*, the halfspace $h(\boldsymbol{y})$ cannot distinguish between $\tilde{\mathcal{D}}_0^R$ and $\tilde{\mathcal{D}}_1^R$. Unlike with the traditional notion of influence, it is unclear whether the number of coordinates in $C_\tau(\boldsymbol{w}_i)$ is small. The following lemma yields a way to get around this.

Lemma 4.8. *(Bounding the number of influential coordinates) Let t be set as in Theorem 4.6. Given a halfspace $h(\boldsymbol{y})$ and $r \in [k]$ such that $|C_\tau(\boldsymbol{w}_r)| > t$, define $\tilde{h}(\boldsymbol{y}) = \mathrm{pos}(\sum_{i\in[k]}\langle\tilde{\boldsymbol{w}}_i, \boldsymbol{y}_i\rangle - \theta)$ as follows: $\tilde{\boldsymbol{w}}_r = \mathsf{Truncate}(\boldsymbol{w}_r, H_t(\boldsymbol{w}_r))$ and $\tilde{\boldsymbol{w}}_i = \boldsymbol{w}_i$ for all $i \neq r$. Then,*

$$\Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[\tilde{h}(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})]\Big| \leqslant \frac{1}{k^2} \text{ and } \Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[\tilde{h}(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant \frac{1}{k^2}.$$

Proof. Without loss of generality, we assume $r = 1$ and $|w_1^{(1)}| \geqslant |w_1^{(2)}| \geqslant \cdots \geqslant |w_1^{(R)}|$. In particular, this implies $H_t(\boldsymbol{w}_1) = \{1, \ldots, t\}$. Set $T = \lceil 4k^2 \ln k\rceil$. Define the subset G of $H_t(\boldsymbol{w}_1)$ as

$$G = \{g_i \mid g_i = 1 + i\lceil(4/\tau^2)\ln(1/\tau)\rceil, 0 \leqslant i \leqslant T\}.$$

Therefore, by Lemma 3.2, $|w_1^{(g_i)}|$ is a geometrically decreasing sequence such that $|w_1^{(g_{i+1})}| \leqslant |w_1^{(g_i)}|/3$. Let $H = H_t(\boldsymbol{w}_1) \setminus G$. Fix the following notation:

$$\boldsymbol{w}_1^G = \mathsf{Truncate}(\boldsymbol{w}_1, G), \quad \boldsymbol{w}_1^H = \mathsf{Truncate}(\boldsymbol{w}_1, H), \quad \boldsymbol{w}_1^{>t} = \mathsf{Truncate}(\boldsymbol{w}_1, \{t+1, \ldots, n\}).$$

Similarly, define the vectors $\boldsymbol{y}_1^G, \boldsymbol{y}_1^H, \boldsymbol{y}_1^{>t}$. We now rewrite the halfspace functions $h(\boldsymbol{y})$ and $\tilde{h}(\boldsymbol{y})$ as:

$$h(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle + \langle \boldsymbol{w}_1^{>t}, \boldsymbol{y}_1^{>t} \rangle - \theta \Big)$$

$$\tilde{h}(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle - \theta \Big) \ .$$

Notice that for any $\boldsymbol{y}$, $h(\boldsymbol{y}) \neq \tilde{h}(\boldsymbol{y})$ implies

$$\Big| \sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle - \theta \Big| \leqslant |\langle \boldsymbol{w}_1^{>t}, \boldsymbol{y}_1^{>t} \rangle|. \tag{2}$$

By Lemma 3.2, we know that

$$|w_1^{(g_T)}|^2 \quad \geqslant \quad \frac{\tau^2}{(1-\tau^2)^{t-g_T}} \|\boldsymbol{w}_1^{>t}\|_2^2 \quad \geqslant \quad \frac{\tau^2}{(1-\tau^2)^{\frac{1}{\tau^2}(3\ln(1/\tau)+\ln R)}} \|\boldsymbol{w}_1^{>t}\|_2^2 \quad \geqslant \quad \frac{R}{\tau} \|\boldsymbol{w}_1^{>t}\|_2^2.$$

Using the fact that $R\|\boldsymbol{w}_1^{>t}\|_2^2 \geqslant \|\boldsymbol{w}_1^{>t}\|_1^2$, we can get that $\|\boldsymbol{w}_1^{>t}\|_1 \leqslant \sqrt{\tau}|w_1^{(g_T)}| \leqslant \frac{1}{6}|w_1^{(g_T)}|$. Combining the above inequality with (2) we see that,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\tilde{\mathcal{D}}_0^R}\Big[h(\boldsymbol{y}) \neq \tilde{h}(\boldsymbol{y})\Big] &\leqslant \mathop{\mathbf{Pr}}_{\tilde{\mathcal{D}}_0^R}\Big[\Big| \sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle - \theta \Big| \leqslant |\langle \boldsymbol{w}_1^{>t}, \boldsymbol{y}_1^{>t} \rangle|\Big] \\
&\leqslant \mathop{\mathbf{Pr}}_{\tilde{\mathcal{D}}_0^R}\Big[\Big| \sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle - \theta \Big| \leqslant \frac{|w_1^{(g_T)}|}{6}\Big] \\
&= \mathop{\mathbf{Pr}}_{\tilde{\mathcal{D}}_0^R}\Big[\langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle \in [\theta' - \frac{1}{6}|w_1^{(g_T)}|, \theta' + \frac{1}{6}|w_1^{(g_T)}|]\Big]
\end{aligned}$$

where $\theta' = -\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle - \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle + \theta$. For any fixing of the value of $\theta' \in \mathbb{R}$, it induces a certain distribution on $\boldsymbol{y}_1^G$. However, the $\frac{1}{k^2}$ noise introduced in $\boldsymbol{y}_1^G$ is completely independent. This corresponds to the setting of Lemma 3.7, and hence we can bound the above probability by $\big(1 - \frac{1}{2k^2}\big)^T \leqslant \frac{1}{k^2}$. The result follows from averaging over all values of θ'. □

With the two lemmas above, we now prove the soundness property.

Proof. (Theorem 4.6) The probability of success of $h(\boldsymbol{y})$ is given by $\frac{1}{2} + \frac{1}{2}\big(\mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})]\big)$. Therefore, it suffices to show that $\Big|\mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| = O(\frac{1}{k})$.

Define $I = \{r \mid C_\tau(\boldsymbol{w}_r) \geqslant t\}$. We discuss the following two cases.

1. $I = \emptyset$; i.e., $\forall i \in [k]$, $C_\tau(\boldsymbol{w}_i) \leqslant t$. Then for all i, j, $H_t(\boldsymbol{w}_i) \cap H_t(\boldsymbol{w}_j) = \emptyset$ implies $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) = \emptyset$. By Lemma 4.7, we thus have $\Big|\mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| = O(\frac{1}{k})$.

1. Sample an edge $e = (v_1, \ldots, v_k) \in E$.
2. Generate a random bit $b \in \{0,1\}$.
3. Sample $\boldsymbol{x} \in \{0,1\}^{kR}$ from $\tilde{\mathcal{D}}_b^R$.
4. Define $\boldsymbol{y} \in \{0,1\}^{|V| \times R}$ as follows:
   (a) For each $v \notin \{v_1, \ldots, v_k\}$, $\boldsymbol{y}_v = \mathbf{0}$.
   (b) For each $i \in [k]$ and $j \in [R]$, $y_{v_i}^{(j)} = x_i^{(\pi^{v_i,e}(j))}$.
5. Output the example $(\boldsymbol{y}, b)$.

Figure 1: Reduction from k-Unique Label Cover

2. $I \neq \emptyset$. Then for all $r \in I$, we set $\tilde{\boldsymbol{w}}_r = \mathsf{Truncate}(\boldsymbol{w}_r, H_t(\boldsymbol{w}_r))$ and replace $\boldsymbol{w}_r$ with $\tilde{\boldsymbol{w}}_r$ in h to get a new halfspace h'. Since such replacements occur at most k times and by Lemma 4.8 every replacement changes the output of the halfspace on at most $\frac{1}{k^2}$ fraction of examples, we can bound the overall change by $k \times \frac{1}{k^2} = \frac{1}{k}$. That is

$$\left| \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] \right| \leqslant \frac{1}{k}, \qquad \left| \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})] \right| \leqslant \frac{1}{k}. \tag{3}$$

Also notice that for h' and all $r \in [k]$, the critical index of $\tilde{\boldsymbol{w}}_r$ (i.e., $|C_\tau(\tilde{\boldsymbol{w}}_r)|$) is less than t. This reduces the problem to Case 1, and we conclude $\left| \mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h'(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h'(\boldsymbol{y})] \right| = O(1/k)$. Along with (3) this finishes the proof of Theorem 4.6.

□

4.4 Reduction from k-Unique Label Cover

With the dictatorship test defined, we now describe briefly a reduction from k-Unique Label Cover problem to agnostic learning of monomials, thus showing Theorem 1.1 under the Unique Games Conjecture (Conjecture 2.2). Although our final hardness result only assumes P $\neq$ NP, we describe the reduction to k-Unique Label Cover for the purpose of illustrating the main idea of our proof.

Let $\mathcal{L}(G(V,E), R, R, \{\pi^{v,e} | v \in V, e \in E\})$ be an instance of k-Unique Label Cover. The reduction is defined in Figure 4.4. It will produce a distribution over labeled examples: $(\boldsymbol{y}, b)$ where $\boldsymbol{y} \in \{0,1\}^{|V| \times R}$ and label $b \in \{0,1\}$. We will index the coordinates of $\boldsymbol{y} \in \{0,1\}^{|V| \times R}$ by $y_w^{(i)}$ (for $w \in V, i \in R$) and denote $\boldsymbol{y}_w$ (for $w \in V$) to be the vector $(y_w^{(1)}, y_w^{(2)}, \ldots, y_w^{(R)})$.

Proof of Theorem 1.1 assuming Unique Games Conjecture Fix $k = \frac{10}{\epsilon^2}$, $\eta = \frac{\epsilon^3}{100}$ and a positive integer $R > \lceil (2k)^{\frac{1}{\eta^2}} \rceil$ for which Conjecture 2.2 holds.

Completeness: Suppose that $\Lambda : V \to [R]$ is a labeling that *strongly* satisfies $1 - k\eta$ fraction of the edges. Consider disjunction $h(\boldsymbol{y}) = \bigvee_{v \in V} y_v^{(\Lambda(v))}$. For at least $1 - k\eta$ fraction of edges $e = (v_1, v_2, \ldots, v_k) \in E$, $\pi^{v_1,e}(\Lambda(v_1)) = \cdots = \pi^{v_k,e}(\Lambda(v_k)) = r$. Let us fix such a choice of edge e in step 1. As all coordinates of $\boldsymbol{y}$ outside of $\{\boldsymbol{y}_{v_1}, \ldots, \boldsymbol{y}_{v_k}\}$ are set to 0 in step $4(a)$, the disjunction reduces to $\vee_{i \in [k]} y_{v_i}^{(\Lambda(v_i))} = \vee_{i \in [k]} x_i^{(r)}$. By Theorem 4.5, such a disjunction agrees with every $(\boldsymbol{y}, b)$ with probability at least $1 - \frac{3}{\sqrt{k}}$. Therefore $h(\boldsymbol{y})$ agrees with a random example with probability at least $(1 - \frac{3}{\sqrt{k}})(1 - k\eta) \geqslant 1 - \frac{3}{\sqrt{k}} - k\eta \geqslant 1 - \epsilon$.

Soundness: Suppose there exists a halfspace $h(\boldsymbol{y}) = \sum_{v \in V} \langle \boldsymbol{w}_v, \boldsymbol{y}_v \rangle$ that agrees with more than $\frac{1}{2} + \epsilon \geqslant \frac{1}{2} + \frac{1}{\sqrt{k}}$ fraction of the examples. Set $t = k^{14}(3\ln(k^7) + \ln R) + \lceil 4k^{14} \ln k^7 \rceil \cdot \lceil 4k^2 \ln k \rceil = O\big(k^{16} \ln R\big)$ (same as in Theorem 4.6). Define the labeling Λ using the following strategy : for each vertex $v \in V$ randomly pick a label from $H_t(\boldsymbol{w}_v)$.

By an averaging argument, for at least $\frac{\epsilon}{2}$ fraction of the edges $e \in E$ generated in step 1 of the reduction, $h(\boldsymbol{y})$ agrees with the examples corresponding to e with probability at least $\frac{1}{2} + \frac{\epsilon}{2}$. We will refer to such edges as *good*. By Theorem 4.6 for each *good* edge $e \in E$, there exists $i, j \in [k]$, such that $\pi^{v_i,e}\big(H_t(\boldsymbol{w}_{v_i})\big) \cap \pi^{v_j,e}\big(H_t(\boldsymbol{w}_{v_j})\big) \neq \emptyset$. Therefore the edge $e \in E$ is *weakly* satisfied by the labeling Λ with probability at least $\frac{1}{t^2}$. Hence, in expectation the labeling Λ *weakly* satisfies at least $\frac{\epsilon}{2} \cdot \frac{1}{t^2} = \Omega(\frac{1}{k^{33} \ln^2 R}) \geqslant \frac{2k^2}{R^{\eta/4}}$ fraction of the edges (by the choice of R and t).

5 Reduction from Label Cover

In this section, we describe a reduction from a k-Label Cover with an additional *smoothness* property to the problem of agnostic learning of disjunctions by halfspaces. This will give us Theorem 1.1 without assuming the Unique Games Conjecture.

5.1 Smooth k-Label Cover

Our reduction use the following hardness result for k-Label Cover (Definition 2.1) with the additional smoothness property.

Theorem 5.1. *There exists a constant $\gamma > 0$ such that for any integer parameter $J, u \geqslant 1$, it is* NP*-hard to distinguish between the following two types of k-Label Cover $\mathcal{L}(G(V, E), M, N, \{\pi^{v,e} | e \in E, v \in e\})$ instances with $M = 7^{(J+1)u}$ and $N = 2^u 7^{Ju}$:*

1. *(Strongly satisfiable instances) There is some labeling that strongly satisfies every hyperedge.*
2. *(Instances that are not $2k^2 2^{-\gamma u}$-weakly satisfiable) There is no labeling that weakly satisfies at least $2k^2 2^{-\gamma u}$ fraction of the hyperedges.*

In addition, the k-Label Cover instances have the following properties:

- *(Smoothness) for a fixed vertex v and a randomly picked hyperedge containing v,*

$$\forall i, j \in [M], \mathbf{Pr}[\pi^{v,e}(i) = \pi^{v,e}(j)] \leqslant 1/J.$$

- Pick a hyperedge $e = (v_1, v_2, \ldots, v_k) \in E$ with corresponding projections $\pi^{v_1,e}, \ldots, \pi^{v_k,e} : [M] \to [N]$.
- Generate a random bit $b \in \{0,1\}$.
- Sample $\boldsymbol{x} \in \{0,1\}^{kN}$ from $\mathcal{D}_b^N$.
- Generate $\boldsymbol{y} \in \{0,1\}^{|V| \times M}$ as follows:
  1. For each $v \notin e$, $\boldsymbol{y}_v = \mathbf{0}$.
  2. For each $i \in [k]$, set $\boldsymbol{y}_{v_i} \in \{0,1\}^M$ as follows:

$$y_{v_i}^{(j)} = \begin{cases} x_i^{(\pi^{v_i,e}(j))} & \text{with probability } 1 - \frac{1}{k^2} \\ \text{random bit} & \text{with probability } \frac{1}{k^2} \end{cases}$$

- Output the example $(\boldsymbol{y}, b)$ or equivalently ACCEPT if $h(\boldsymbol{y}) = b$.

Figure 2: Reduction from k-Label Cover

- *For any mapping $\pi^{v,e}$ and any number $i \in [N]$, we have $|(\pi^{v,e})^{-1}(i)| \leqslant d = 4^u$; i.e., there are at most $d = 4^u$ elements in $[M]$ that are mapped to the same number in $[N]$.*

The proof of the above theorem can be found in Appendix D.

In the rest of the paper, we will set $u = k$ and therefore $d = 4^k$. Also we set the smoothness parameter $J = d^{17} = 4^{17k}$.

5.2 Reduction from Smooth k-Label Cover

The starting point is a smooth k-Label Cover $\mathcal{L}(G(V,E), M, N, \{\pi^{v,e} | e \in E, v \in e\})$ with $M = 7^{(J+1)u}$ and $N = 2^u 7^{Ju}$ as described in Theorem 5.1. Figure 5.2 illustrates the reduction from k-Label Cover $\mathcal{L}(G(V,E), N, M, \{\pi^{v,e} | e \in E, v \in e\})$ that given an instance of k-Label Cover $\mathcal{L}$ produces a random labeled example. We refer to the obtained distribution on examples as $\mathcal{E}$.

5.3 Proof of Theorem 1.1

We claim that our reduction has the following completeness and soundness properties.

Theorem 5.2.
- Completeness: *If $\mathcal{L}$ is a strongly-satisfiable instance of smooth k-Label Cover, then there is a disjunction that agrees with a random example from $\mathcal{E}$ with probability at least $1 - O(\frac{1}{\sqrt{k}})$.*
- Soundness: *If $\mathcal{L}$ is not $2k^2 2^{-\gamma k}$-weakly satisfiable and is smooth with parameters $J = 4^{17k}$ and $d = 4^k$, then there is no halfspace that agrees with a random example from $\mathcal{E}$ with probability more than $\frac{1}{2} + O(\frac{1}{\sqrt{k}})$.*

Combining the above theorem with Theorem 5.1 we get that for $k = O(1/\epsilon^2)$, we obtain our main result: Theorem 1.1.

It remains to check the correctness of the completeness and soundness claims in Theorem 5.2. First let us prove the completeness property.

Proof. (Proof of Completeness) Let Λ be the labeling that strongly satisfies $\mathcal{L}$. Consider disjunction $h(\boldsymbol{y}) = \bigvee_{v \in V} y_v^{(\Lambda(v))}$. Let $e = (v_1, v_2, \ldots, v_k)$ be any hyperedge and let $\mathcal{E}_e$ be the distribution $\mathcal{E}$ restricted to the examples generated for e. With probability at least $1 - 1/k$, $y_{v_i}^{\Lambda(v_i)} = x_i^{\pi^{v_i,e}(\Lambda(v_i))}$ for every $i \in [k]$. As e is strongly satisfied by Λ, for all $i, j \in [k]$, $\pi^{v_i,e}(\Lambda(v_i)) = \pi^{v_j,e}(\Lambda(v_j))$. Therefore, as in the proof of Theorem 4.5, we obtain that $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}_e$ with probability at least $1 - O(1/\sqrt{k})$. Labeling Λ strongly satisfies all edges and therefore we obtain that $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}$ with probability at least $1 - O(1/\sqrt{k})$. □

The more complicated part is the soundness property which we prove in Section 5.4.

5.4 Soundness Analysis

Proof Idea The main idea is similar to the proof of Theorem 4.6 although it is more technically involved. Notice that the reduction in Figure 5.2 produces examples such that $y_{v_i}^{j_1}$, $y_{v_i}^{j_2}$ are "almost identical" copies when $\pi^{v_i,e}(j_1) = \pi^{v_i,e}(j_2)$. Further for different edges e, the coordinates of y will be grouped in different ways, such that each group will have almost identical copies.

To handle these additional complications, the first step of the proof is to show that almost all the hyperedges in smooth k-Label Cover satisfy a certain "niceness" property. After that we generalize the proofs of Lemma 4.7 and Lemma 4.8 under the weaker assumption that most of the hyperedges are "nice".

The formal definition of "niceness" and the proof that most of the edges are "nice" appear in Section 5.4.1. The generalization of Lemma 4.7 appears in Section 5.4.2. The generalization of Lemma 4.8 appears in Section 5.4.3. All these results are put together into a proof of Theorem 5.2 in Section 5.4.4.

5.4.1 Most of the edges are "nice"

Let $h(\boldsymbol{y})$ be a halfspace that agrees with more than $\frac{1}{2} + \frac{1}{\sqrt{k}}$-fraction of the examples. Suppose,

$$h(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{v \in V} \langle \boldsymbol{w}_v, \boldsymbol{y}_v \rangle - \theta\Big).$$

Let $\tau = \frac{1}{k^{13}}$ and let

$$\boldsymbol{s}_v = \mathsf{Truncate}(\boldsymbol{w}_v, C_\tau(\boldsymbol{w}_v)), \quad \boldsymbol{l}_v = \boldsymbol{w}_v - \boldsymbol{s}_v.$$

Definition 5.3. *A vertex $v \in V$ is said to be β-nice with respect to a hyperedge $e \in E$ containing it if*

$$\sum_{i \in [N]} \Big(\sum_{j \in \pi^{-1}(i)} |l_v^{(j)}| \Big)^4 \leqslant \beta \|\boldsymbol{l}_v\|_2^4,$$

where $\pi : [M] \to [N]$ is the projection associated with vertex v and hyperedge e. A hyperedge $e = (v_1, v_2, \ldots, v_k)$ is β-nice, if for every $i \in [k]$, the vertex v_i is β-nice with respect to e.

Lemma 5.4. *The fraction of 2τ-nice hyperedges in E is at least $1 - O(1/k)$.*

Proof. By definition, we know that $\boldsymbol{l}_v$ is τ-regular vector. Denote $I_v = \{i \mid \frac{(l_v^{(i)})^2}{\|l_v\|_2^2} \geqslant \frac{1}{d^8}\}$. By definition $|I| \leqslant d^8$. Notice there are at most d^{16} pairs of values in $I \times I$. By the smoothness property of the k-Label Cover instance, for any vertex v, at least $1 - \frac{d^{16}}{J}$ fraction of the hyperedges incident on v have the following property: for any $i, j \in I_v$, $\pi^{v,e}(i) \neq \pi^{v,e}(j)$. If all the vertices in a hyperedge have this property we call it a *good* hyperedge. By an averaging argument, we know that among all hyperedges at least $1 - \frac{kd^{16}}{J} = 1 - \frac{k}{4^k} \geqslant 1 - O(\frac{1}{k})$ fraction is *good*.

We will show all these *good* hyperedges are also 2τ-nice. For a given *good* hyperedge e, a vertex $v \in e$, $\pi = \pi^{v,e}$ and $i \in [N]$, there is at most one $j \in \pi^{-1}(i)$ such that $\frac{(l_v^{(i)})^2}{\|l_v\|_2^2} \geqslant \frac{1}{d^8}$.

Based on the above property, we will show

$$\sum_{i \in [N]} \Big(\sum_{j \in \pi^{-1}(i)} |l_v^{(j)}| \Big)^4 \leqslant 2\tau \|\boldsymbol{l}_v\|_2^4 \, .$$

Notice that

$$\sum_{i \in [N]} \big(\sum_{j \in \pi^{-1}(i)} |l_v^{(j)}| \big)^4 = \sum_{i \in [N]} \sum_{j_1, j_2, j_3, j_4 \in \pi^{-1}(i)} \big| l_v^{(j_1)} l_v^{(j_2)} l_v^{(j_3)} l_v^{(j_4)} \big| \tag{4}$$

and the sum of all the terms with $j_1 = j_2 = j_3 = j_4$ is $\|\boldsymbol{l}_v\|_4^4$.

For all other terms $|l_v^{(j_1)} l_v^{(j_2)} l_v^{(j_3)} l_v^{(j_4)}|$ with j_1, j_2, j_3, j_4 that are not all equal, there is at least one $|l_v^{(j_r)}|$ ($r \in [4]$) smaller than $\frac{\|l_v\|_2}{d^4}$. Therefore, $|l_v^{(j_1)} l_v^{(j_2)} l_v^{(j_3)} l_v^{(j_4)}|$ can be bounded by

$$\frac{\|\boldsymbol{l}_v\|_2}{d^4} \big(\sum_{j_1, j_2, j_3, j_4} |l_v^{(j_1)}|^3 + |l_v^{(j_2)}|^3 + |l_v^{(j_3)}|^3 + |l_v^{(j_4)}|^3 \big).$$

Overall, expression (4) can be bounded by

$$\begin{aligned}
&\|\boldsymbol{l}_v\|_4^4 + \frac{\|\boldsymbol{l}_v\|_2}{d^4} \sum_{i \in [N]} \sum_{j_1, j_2, j_3, j_4 \in \pi^{-1}(i)} |l_v^{(j_1)}|^3 + |l_v^{(j_2)}|^3 + |l_v^{(j_3)}|^3 + |l_v^{(j_4)}|^3 \\
\leqslant &\tau^2 \|\boldsymbol{l}_v\|_2^4 + \frac{\|l_v\|_2}{d^4} 4d^3 \sum_{j \in [M]} |l_v^{(j)}|^3 && \text{(since } |\pi^{-1}(i)| \leqslant d, \text{ each } l_v^{(j)} \text{ appears at most } 4d^3 \text{ times)} \\
\leqslant &(\tau^2 + 4\frac{\tau}{d}) \|\boldsymbol{l}_v\|_2^4 && (\boldsymbol{l}_v \text{ is } \tau\text{-regular vector, so } |l_v^j| \leqslant \tau \|\boldsymbol{l}_v\|_2 \text{ for all } j \in [M] \text{)} \\
\leqslant &2\tau \|\boldsymbol{l}_v\|_2^4.
\end{aligned}$$

□

Let us fix a 2τ-nice hyperedge $e = (v_1, \ldots, v_k)$. As before let $\mathcal{E}_e$ denote the distribution on examples restricted to those generated for hyperedge e. We will analyze the probability that the halfspace $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}_e$.

Let $\pi^{v_1,e}, \pi^{v_2,e}, \ldots, \pi^{v_k,e} : [M] \to [N]$ denote the projections associated with the hyperedge e. For the sake of brevity, we shall write $\boldsymbol{w}_i, \boldsymbol{y}_i, \boldsymbol{l}_i$ instead of $\boldsymbol{w}_{v_i}, \boldsymbol{y}_{v_i}, \boldsymbol{l}_{v_i}$. For all $j \in [N]$ and $i \in [k]$, define

$$\boldsymbol{y}_i^{\{j\}} = \mathsf{Truncate}(\boldsymbol{y}_i, (\pi^{v_i,e})^{-1}(j)).$$

Similarly, define vectors $\boldsymbol{w}_i^{\{j\}}, \boldsymbol{l}_i^{\{j\}}$ and $\boldsymbol{s}_i^{\{j\}}$.

Notice that for every example $(\boldsymbol{y}, b)$ in the support of $\mathcal{E}_e$, $\boldsymbol{y}_v = \boldsymbol{0}$ for every vertex $v \notin e$. Therefore, on restricting to examples from $\mathcal{E}_e$ we can write:

$$h(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{i\in[k]} \langle \boldsymbol{w}_i, \boldsymbol{y}_i\rangle - \theta\Big).$$

5.4.2 Common Influential Variables (generalization of Lemma 4.7)

Lemma 5.5. *Let $h(\boldsymbol{y})$ be a halfspace such that for all $i \neq j \in [k]$, we have $\pi^{v_i,e}(C_\tau(\boldsymbol{w}_i)) \cap \pi^{v_j,e}(C_\tau(\boldsymbol{w}_j)) = \emptyset$. Then*

$$\Big|\mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=0] - \mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=1]\Big| \leqslant O\Big(\frac{1}{k}\Big). \tag{5}$$

Proof. Fix the following notation:

$$\boldsymbol{y}_i^C = \mathsf{Truncate}(\boldsymbol{y}_i, C_\tau(\boldsymbol{w}_i)) \qquad \boldsymbol{y}^C = \boldsymbol{y}_1^C, \boldsymbol{y}_2^C, \ldots, \boldsymbol{y}_k^C$$
$$\boldsymbol{s} = \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k \qquad \boldsymbol{l} = \boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_k.$$

We can rewrite the halfspace $h(\boldsymbol{y})$ as $h(\boldsymbol{y}) = \mathrm{pos}\Big(\langle \boldsymbol{s}, \boldsymbol{y}^C\rangle + \langle \boldsymbol{l}, \boldsymbol{y}\rangle - \theta\Big)$. Let us first normalize the weights of $h(\boldsymbol{y})$ so that $\sum_{i\in[k]} \|\boldsymbol{l}_i\|_2^2 = 1$. Let us condition on a possible fixing of the vector $\boldsymbol{y}^C$. Under this conditioning and also for $b = 0$, define the family of ensembles $\mathcal{A} = \boldsymbol{A}^{\{1\}}, \ldots, \boldsymbol{A}^{\{N\}}$ as follows:

$$\boldsymbol{A}^{\{j\}} = \Big\{ y_i^{(r)} \mid i \in [k], r \in [M] \text{ such that } \pi^{v_i,e}(r) = j \text{ and } r \notin C_\tau(\boldsymbol{w}_i)\Big\}$$

Similarly define the ensemble $\mathcal{B} = \boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{N\}}$ for the conditioning $b = 1$. Now we shall apply the invariance principle (Theorem 3.10) to the ensembles $\mathcal{A}, \mathcal{B}$ and the linear function $\boldsymbol{l}(\boldsymbol{y})$:

$$\boldsymbol{l}(\boldsymbol{y}) = \sum_{j\in[N]} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{y}^{\{j\}}\rangle.$$

As we prove in Claim 5.6 below, the ensembles $\mathcal{A}, \mathcal{B}$ have matching moments up to degree 3. Furthermore, by Lemma 3.3, the linear function $\boldsymbol{l}$ and the ensembles $\mathcal{A}$, $\mathcal{B}$ satisfy the following spread property:

$$\mathbf{Pr}_{\mathcal{A}}\Big[l(\mathcal{A}) \in [\theta' - \alpha, \theta' + \alpha]\Big] \leqslant c(\alpha) \qquad \mathbf{Pr}_{\mathcal{B}}\Big[l(\mathcal{B}) \in [\theta' - \alpha, \theta' + \alpha]\Big] \leqslant c(\alpha)$$

for all $\theta' \in \mathbb{R}$, where $c(\alpha) = 8\alpha k + 4\tau k + 2e^{-\frac{1}{2k^4\tau^2}}$ (by setting $\gamma = \frac{1}{k^2}$ and $|b-a| = 2\alpha$ in Lemma 3.3).

Using the invariance principle (Th. 3.10), this implies:

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\mathrm{pos}\left(\langle \boldsymbol{s}, \boldsymbol{y}^C \rangle + \sum_{j\in[N]} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}} \rangle - \theta \right) |\boldsymbol{y}^C \right] - \mathop{\mathbf{E}}_{\mathcal{B}} \left[\mathrm{pos}\left(\langle \boldsymbol{s}, \boldsymbol{y}^C \rangle + \sum_{j\in[N]} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}} \rangle - \theta \right) |\boldsymbol{y}^C \right] \right| \leqslant O(\frac{1}{\alpha^4}) \sum_{j\in[N]} \|\boldsymbol{l}^{\{j\}}\|_1^4 + 2c(\alpha). \quad (6)$$

Take α to be $\frac{1}{k^2}$ and recall that $\tau = \frac{1}{k^{13}}$. In Claim 5.7 below we show that

$$\sum_{j\in[N]} \|\boldsymbol{l}^{\{j\}}\|_1^4 \leqslant 2\tau k^4.$$

The above inequality holds for an arbitrary conditioning of the values of $\boldsymbol{y}^C$. Hence, by averaging over all settings of $\boldsymbol{y}^C$ we prove (5). □

Claim 5.6. *The ensembles $\mathcal{A}$ and $\mathcal{B}$ have matching moments up to degree 3.*

Let us suppose for a moment that $\boldsymbol{y}$ was generated by setting $y_{v_i}^{(j)} = x_i^{(\pi^{v_i,e}(j))}$, that is without adding any noise. By Lemma 4.1, the first four moments of random variable $\boldsymbol{y}$ conditioned on $b = 0$ agree with the first moments of random variable $\boldsymbol{y}$ conditioned on $b = 1$. As we showed in Observation 4.3, even with noise, the first four moments of $\boldsymbol{y}$ remain the same when conditioned on $b = 0$ and $b = 1$. Finally, $\pi^{v_i,e}(C_\tau(\boldsymbol{w}_i)) \cap \pi^{v_j,e}(C_\tau(\boldsymbol{w}_j)) = \emptyset$ for all $i \neq j \in [k]$. Hence for each $j \in [N]$, conditioning on $\boldsymbol{y}^C$ fixes bits in at most one row of $\boldsymbol{A}^{\{j\}}$. Formally, for every $j \in [N]$, there exists at most one $i \in [k]$ such that $\boldsymbol{y}_i^{\{j\}}$ and $\boldsymbol{y}^C$ have shared variables. Therefore, by Lemma 4.4, $\mathcal{A}$ and $\mathcal{B}$ have matching moments up to degree 3.

Claim 5.7.

$$\sum_{j\in[N]} \|\boldsymbol{l}^{\{j\}}\|_1^4 \leqslant 2\tau k^4 \ .$$

Proof. Since $\|\boldsymbol{l}^{\{j\}}\|_1 = \sum_{i\in[k]} \|\boldsymbol{l}_i^{\{j\}}\|_1$, we can write

$$\sum_{j\in[N]} \|\boldsymbol{l}^{\{j\}}\|_1^4 \leqslant \sum_{j\in N} k^4 \Big(\sum_{i\in[k]} \|\boldsymbol{l}_i^{\{j\}}\|_1^4 \Big) = k^4 \sum_{i\in[k]} \Big(\sum_{j\in[N]} \|\boldsymbol{l}_i^{\{j\}}\|_1^4 \Big). \quad (7)$$

As $e = (v_1, \ldots, v_k)$ is a 2τ-nice hyperedge, we have $\sum_{j\in[N]} \|\boldsymbol{l}_i^{\{j\}}\|_1^4 \leqslant 2\tau \|\boldsymbol{l}_i\|_2^4$. By normalization of $\boldsymbol{l}$, we know $\sum_{i\in[k]} \|\boldsymbol{l}_i\|_2^2 = 1$. Substituting this into inequality (7) we get the claimed bound. □

5.4.3 Bounding the Number of Influential Coordinates (generalization of Lemma 4.8)

Lemma 5.8. *Given a halfspace $h(\boldsymbol{y}) = \mathrm{pos}(\sum_{i\in[k]} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle - \theta)$ and $r \in [k]$ such that $|C_\tau(\boldsymbol{w}_r)| \geqslant t$ for $t = \frac{1}{\tau^2}(\lceil 4k^2 \ln(2k) \rceil \lceil 4 \ln(1/\tau) \rceil + \ln(1/\tau) + 10 \ln d) = O(k^{29})$, define $\tilde{h}(\boldsymbol{y}) = \mathrm{pos}(\sum_{i\in[k]} \langle \tilde{\boldsymbol{w}}_i, \boldsymbol{y}_i \rangle - \tilde{\theta})$ as follows:*

- $\tilde{\boldsymbol{w}}_r = \mathsf{Truncate}(\boldsymbol{w}_r, H_t(\boldsymbol{w}_r))$ *and* $\tilde{\boldsymbol{w}}_i = \boldsymbol{w}_i$ *for all* $i \neq r$.
- $\tilde{\theta} = \theta - \mathbf{E}[\langle \boldsymbol{a}_r, \boldsymbol{y}_r \rangle | b = 0]$, *for* $\boldsymbol{a} = \boldsymbol{w} - \tilde{\boldsymbol{w}}$.

Then,

$$\Big| \mathop{\mathbf{E}}_{\mathcal{E}_e}[\tilde{h}(\boldsymbol{y})|b=0] - \mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=0] \Big| \leqslant \frac{1}{k^2}, \qquad \Big| \mathop{\mathbf{E}}_{\mathcal{E}_e}[\tilde{h}(\boldsymbol{y})|b=1] - \mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=1] \Big| \leqslant \frac{1}{k^2}.$$

Proof. It is easy to see that the matching moments condition implies that

$$\mathbf{E}_{\mathcal{E}_e}[\langle \boldsymbol{a}_r, \boldsymbol{y}_r \rangle | b=0] = \mathbf{E}_{\mathcal{E}_e}[\langle \boldsymbol{a}_r, \boldsymbol{y}_r \rangle | b=1].$$

Let us show the inequality for the case $b = 0$, the other inequality can be derived in an identical way. Let $\mathcal{E}_{e,0}$ denote distribution $\mathcal{E}_e$ conditioned on $b = 0$. Without loss of generality, we may assume that $r = 1$ and $|w_1^{(1)}| \geqslant |w_1^{(2)}| \ldots \geqslant |w_1^{(M)}|$. In particular, this implies $H_t(\boldsymbol{w}_1) = \{1, \ldots, t\}$. Define

$$\mu_r = \mathbf{E}_{\mathcal{E}_{e,0}}[\langle \boldsymbol{a}_r, \boldsymbol{y}_r \rangle], \qquad \mu_r^{\{i\}} = \mathbf{E}_{\mathcal{E}_{e,0}}[\langle \boldsymbol{a}_r^{\{i\}}, \boldsymbol{y}_r^{\{i\}} \rangle].$$

Let us set $T = \lceil 4k^2 \ln(2k) \rceil$ and define the subset $G = \{g_1, \ldots, g_T\}$ of $H_t(\boldsymbol{w}_1)$ as follows:

$$G = \{g_i \mid g_i = 1 + i\lceil (4/\tau^2)\ln(1/\tau) \rceil, 0 \leqslant i \leqslant T\}.$$

Therefore, by Lemma 3.2, $|w_1^{(g_i)}|$ is a geometrically decreasing sequence such that $|w_1^{(g_{i+1})}| \leqslant |w_1^{(g_i)}|/3$. Let $H = H_t(\boldsymbol{w}_1) \setminus G$. Fix the following notation:

$$\boldsymbol{w}_1^G = \mathsf{Truncate}(\boldsymbol{w}_1, G), \qquad \boldsymbol{w}_1^H = \mathsf{Truncate}(\boldsymbol{w}_1, H), \qquad \boldsymbol{w}_1^{>t} = \mathsf{Truncate}(\boldsymbol{w}_1, \{t+1, \ldots, n\}).$$

Similarly, define the vectors $\boldsymbol{y}_1^G, \boldsymbol{y}_1^H, \boldsymbol{y}_1^{>t}$. By definition, we have $\boldsymbol{a}_1 = \boldsymbol{w}_1^{>t}$. Rewriting the halfspace functions $h(\boldsymbol{y}), \tilde{h}(\boldsymbol{y})$:

$$h(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle + \langle \boldsymbol{a}_1, \boldsymbol{y}_1^{>t} \rangle - \theta \Big),$$
$$\tilde{h}(\boldsymbol{y}) = \mathrm{pos}\Big(\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle + \mu_1 - \theta \Big).$$

By Claim 5.9 below, with probability at most $\frac{1}{d} = \frac{1}{4^k}$, we have $|\langle \boldsymbol{a}_1, \boldsymbol{y}_1 \rangle - \mu_1| \geqslant d^4 \|\boldsymbol{a}_1\|_2$. Suppose $|\langle \boldsymbol{a}_1, \boldsymbol{y}_1 \rangle - \mu_1| < d^4 \|\boldsymbol{a}_1\|_2$, then Claim 5.10 below gives $|\langle \boldsymbol{a}_1, \boldsymbol{y}_1 \rangle - \mu_1| < 1/d^6 |w_1^{(g_T)}| < \frac{1}{3}|w_1^{(g_T)}|$. Thus, we can write

$$\mathbf{Pr}_{\mathcal{E}_{e,0}}\Big[h(\boldsymbol{y}) \neq \tilde{h}(\boldsymbol{y}) \Big] \leqslant \mathbf{Pr}_{\mathcal{E}_{e,0}}\Big[\langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G \rangle \in [\theta' - \frac{1}{3}|w_1^{(g_T)}|, \theta' + \frac{1}{3}|w_1^{(g_T)}|] \Big] + \frac{1}{4^k}.$$

where $\theta' = -\sum_{i=2}^{k} \langle \boldsymbol{w}_i, \boldsymbol{y}_i \rangle - \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H \rangle - \mu_1 + \theta$. For any fixing of the value of $\theta' \in \mathbb{R}$, induces a certain distribution on $\boldsymbol{y}_1^G$. However, the $\frac{1}{k^2}$ noise introduced in $\boldsymbol{y}_1^G$ is completely independent. This corresponds to the setting of Lemma 3.7, and hence we can bound the above probability by $(1 - 1/(2k^2))^T + 1/4^k \leqslant (1 - 1/(2k^2))^{4k^2 \ln(2k)} + 1/4^k \leqslant 1/k^2$. □

Claim 5.9.

$$\mathbf{Pr}_{\mathcal{E}_{e,0}}\Big[|\langle \boldsymbol{a}_1, \boldsymbol{y}_1\rangle - \mu_1| \geqslant d^4\|\boldsymbol{a}_1\|_2\Big] \leqslant \frac{1}{d}.$$

Proof. Write $[M]$ as the union of disjoint sets $R_1 \cup R_2 \cup \cdots \cup R_N$ where $R_i = (\pi^{v_1,e})^{-1}(i)$. Notice every R_i has size at most d, therefore

$$\mathbf{Var}_{\mathcal{E}_{e,0}}\big(\langle \boldsymbol{a}_1, \boldsymbol{y}_1\rangle\big) = \sum_{i\in[N]} \mathbf{Var}_{\mathcal{E}_{e,0}}\big(\langle \boldsymbol{a}_1^{R_i}, \boldsymbol{y}_1^{R_i}\rangle\big) \leqslant \sum_{i\in[N]} d\|\boldsymbol{a}_1^{R_i}\|_2^2 = d\|\boldsymbol{a}_1\|_2^2.$$

By applying Chebyshev's inequality (Th. A.3), we have

$$\mathbf{Pr}_{\mathcal{E}_{e,0}}\left[|\langle \boldsymbol{a}_1, \boldsymbol{y}_1\rangle - \mu_1| \geqslant d^4\|\boldsymbol{a}_1\|_2\right] \leqslant \frac{2d}{d^8} \leqslant \frac{1}{d}.$$

□

Claim 5.10. *By the choice of the parameters T and t,*

$$\|\boldsymbol{a}_1\|_2 \leqslant \frac{1}{d^{10}}|w_1^{(g_T)}|.$$

Proof. By Lemma 3.2,

$$|w_1^{(g_T)}|^2 \geqslant \frac{\tau}{(1-\tau^2)^{t-g_T}}\|a_1\|_2^2 \geqslant \frac{\tau}{(1-\tau^2)^{\frac{1}{\tau^2}(\ln(1/\tau)+10\ln d)}}\|\boldsymbol{a}_1\|_2^2 \geqslant d^{10}\|\boldsymbol{a}_1\|_2^2.$$

□

5.4.4 Proof of Soundness

Recall that we chose $\tau = 1/k^{13}$ and $t = O(k^{29})$.

Lemma 5.11. *Fix a hyperedge e which is 2τ-nice. If for all $i \neq j \in [k]$, $\pi^{v_i,e}\big(H_t(\boldsymbol{w}_i)\big) \cap \pi^{v_j,e}\big(H_t(\boldsymbol{w}_j)\big) = \emptyset$ then the probability that halfspace $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}_e$ is at most $\frac{1}{2} + O(\frac{1}{k})$.*

Proof. The proof is similar to the proof of Theorem 4.6. Define $I = \{r \mid C_\tau(\boldsymbol{w}_r) > t\}$. We divide the problem into the following two cases.

1. $I = \emptyset$; i.e., for all $i \in [k]$, $C_\tau(\boldsymbol{w}_i) \leqslant t$. Then for any $i \neq j \in [k]$, $H_t(\boldsymbol{w}_i) \cap H_t(\boldsymbol{w}_j) = \emptyset$ implies $C_\tau(\boldsymbol{w}_i) \cap C_\tau(\boldsymbol{w}_j) = \emptyset$. By Lemma 5.5, we have

$$\Big|\mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=0] - \mathop{\mathbf{E}}_{\mathcal{E}_e}[h(\boldsymbol{y})|b=1]\Big| \leqslant O\Big(\frac{1}{k}\Big).$$

2. $I \neq \emptyset$. Then for all $r \in I$, we set $\tilde{\boldsymbol{w}}_r = \mathsf{Truncate}(\boldsymbol{w}_r, H_t(\boldsymbol{w}_r))$ and define a new halfspace h' by replacing $\boldsymbol{w}_r$ with $\tilde{\boldsymbol{w}}_r$ in h. Since such replacements occur at most k times and, by Lemma

5.8, every replacement changes the output of the halfspace on at most $\frac{1}{k^2}$ fraction of examples from $\mathcal{E}_e$, we can bound the overall change by $k \times \frac{1}{k^2} = \frac{1}{k}$. That is

$$\left| \mathop{\mathbf{E}}_{\mathcal{E}_{e,0}} [h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\mathcal{E}_{e,0}} [h(\boldsymbol{y})] \right| \leqslant \frac{1}{k}, \qquad \left| \mathop{\mathbf{E}}_{\mathcal{E}_{e,1}} [h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\mathcal{E}_{e,1}} [h(\boldsymbol{y})] \right| \leqslant \frac{1}{k}. \tag{8}$$

For the halfspace h' and for all $r \in [k]$, we have $|C_\tau(\tilde{\boldsymbol{w}}_r)| \leqslant t$, thus reducing to Case 1. Therefore

$$\left| \mathop{\mathbf{E}}_{\mathcal{E}_{e,o}} [h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\mathcal{E}_{e,1}} [h'(\boldsymbol{y})] \right| \leqslant O\Big(\frac{1}{k}\Big). \tag{9}$$

Combining (8) and (9), we get

$$\left| \mathop{\mathbf{E}}_{\mathcal{E}_{e,0}} [h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\mathcal{E}_{e,1}} [h(\boldsymbol{y})] \right| \leqslant O\Big(\frac{1}{k}\Big).$$

In other words, the probability that halfspace $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}_e$ is at most $\frac{1}{2} + O(\frac{1}{k})$. □

We first recall the soundness statement:

Proposition 5.12. *If $\mathcal{L}$ is not a $2k^2 2^{-\gamma k}$-weakly satisfiable instance of smooth k-*Label Cover*, then there is no halfspace that agrees with a random example from $\mathcal{E}$ with probability more than $\frac{1}{2} + \frac{1}{\sqrt{k}}$.*

Proof. The proof is by contradiction. We can define the following labeling strategy: for each vertex v, uniformly randomly pick a label from $H_t(\boldsymbol{w}_v)$. We know that the size of $H_t(\boldsymbol{w}_{v_i})$ is $t = O(k^{29})$.

Suppose there exists a halfspace that agrees with a random example from $\mathcal{E}$ with probability more than $\frac{1}{2} + \frac{1}{\sqrt{k}}$. Then by an averaging argument, for at least $\frac{1}{2\sqrt{k}}$-fraction of the hyperedges e, $h(\boldsymbol{y})$ agrees with a random example from $\mathcal{E}_e$ with probability at least $\frac{1}{2} + \frac{1}{2\sqrt{k}}$. We refer to these edges as *good*.

Since there is at most $O(1/k)$-fraction of the hyperedges that are *not* 2τ-nice we know that at least $\frac{1}{4\sqrt{k}}$-fraction of the hyperedges are 2τ-nice and *good*. By Lemma 5.11, for each 2τ-nice and *good* hyperedge e there exist two vertices $v_i, v_j \in e$ such that $\pi^{v_i,e}(H_t(\boldsymbol{w}_i))$ and $\pi^{v_j,e}(H_t(\boldsymbol{w}_j))$ intersect. Then there is a $\frac{1}{t^2}$ probability that the labeling strategy we defined will weakly satisfy hyperedge e.

Overall this strategy is expected to weakly satisfy at least $\frac{1}{4\sqrt{k}} \frac{1}{t^2} = \Omega(\frac{1}{k^{59}})$ fraction of the hyperedges. This is a contradiction since $\mathcal{L}$ is not $\frac{2k^2}{2^{\gamma k}}$-weakly satisfiable. □

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. 3

[2] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988. 3

[3] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *J. Comput. Syst. Sci.*, 54(2):317–331, 1997. 3

[4] P. Auer and M. K. Warmuth. Tracking the best disjunction. *Machine Learning*, 32(2):127–150, 1998. 4

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

[6] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1-2):35–52, 1998. 3

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Inf. Process. Lett.*, 24(6):377–380, 1987. 2

[8] N. Bshouty and L. Burroughs. Bounds for the minimum disagreement problem with applications to learning theory. In *Proceedings of COLT*, pages 271–286, 2002. 3

[9] N. Bshouty and L. Burroughs. Maximizing agreements and coagnostic learning. *Theoretical Computer Science*, 350(1):24–39, 2006. 3

[10] T. Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of COLT*, pages 340–347, 1994. 3

[11] S. Chatterjee. A simple invariance theorem. *arxiv:math/0508213v1.*, 2005. 11

[12] E. Cohen. Learning noisy perceptrons by a perceptron in polynomial time. In *IEEE FOCS*, pages 514–523, 1997. 3

[13] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. A. Servedio, and E. Viola. Bounded independence fools halfspaces. In *FOCS*, pages 171–180, 2009. 6, 7, 9, 10

[14] V. Feldman. Optimal hardness results for maximizing agreements with monomials. In *IEEE CCC*, pages 226–236, 2006. 3

[15] 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. 3

[16] S. Galant. Perceptron based learning algorithms. *IEEE Trans. on Neural Networks*, 1(2), 1990. 4

[17] M. Garey and D. S. Johnson. *Computers and Intractability.* 1979. 3

[18] C. Gentile and M. K. Warmuth. Linear hinge loss and average margin. In *Proceedings of NIPS*, pages 225–231, 1998. 4

[19] P. Gopalan, S. Khot, and R. Saket. Hardness of reconstructing multivariate polynomials over finite fields. *SIAM J. Comput.*, 39(6):2598–2621, 2010. 36

[20] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39(2):742–765, 2009. 3

[21] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. 2

[22] K. Hoffgen, K. van Horn, and H. U. Simon. Robust trainability of single neurons. *J. Comput. Syst. Sci.*, 50(1):114–125, 1995. 3

[23] D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978. 3

[24] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. 3, 4

[25] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998. 3

[26] M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994. 2, 3

[27] M. J. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4):807–837, 1993. 3

[28] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994. 3

[29] S. Khot. On the power of unique 2-Prover 1-Round games. In *ACM STOC*, pages 767–775, May 19–21 2002. 4, 5

[30] S. Khot. New techniques for probabilistically checkable proofs and inapproximability results (thesis). *Princeton University Technical Reports*, TR-673-03, 2003. 8, 36

[31] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM J. Comput*, 37(1):319–357, 2007. 8

[32] S. Khot and R. Saket. Hardness of minimizing and learning DNF expressions. In *IEEE FOCS*, pages 231–240, 2008. 2, 3

[33] W. S. Lee, P. L. Bartlett, and R. C. Williamson. On efficient agnostic learning of linear combinations of basis functions. In *Proceedings of COLT*, pages 369–376, 1995. 3

[34] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987. 2

[35] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. A. Servedio. Testing halfspaces. In *SODA*, pages 256–264, 2009. 6, 9

[36] E. Mossel. Gaussian bounds for noise correlation of functions. *IEEE FOCS*, 2008. 11

[37] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. In *IEEE FOCS*, 2005. 11, 35

[38] R. O'Donnell and R. A. Servedio. The chow parameters problem. In *ACM STOC*, pages 517–526, 2008. 6, 9

[39] R. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987. 2

[40] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958. 2

[41] R. A. Servedio. Every linear threshold function has a low-weight approximator. *Comput. Complex.*, 16(2):180–209, 2007. 6, 9

[42] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 2

[43] V. Vapnik. *Statistical Learning Theory*. 1998. 2

Appendix

A Probabilistic Inequalities

In the discussion below we will make use of the following well-known inequalities.

Theorem A.1. *(Hoeffding's Inequality) Let $x^{(1)}, \ldots, x^{(n)}$ be independent real random variables such that $x^{(i)} \in [a^{(i)}, b^{(i)}]$. Then the sum of these variables $S = \sum_{i=1}^{n} x^{(i)}$ satisfies*

$$\mathbf{Pr}[|S - \mathbf{E}[S]| \geqslant nt] \leqslant 2e^{-\frac{n^2t^2}{\sum_{i=1}^{n}(b^{(i)}-a^{(i)})^2}}.$$

Theorem A.2. *(Berry-Esseen Theorem) Let $x_1, x_2, \ldots, x_n$ be i.i.d. random unbiased $\{-1, 1\}$ variables. Also assume that $\sum_{i=1}^{n} c_i^2 = 1$ and $\max_i\{|c_i|\} \leqslant \alpha$. Let g denote a unit Gaussian variable $N(0, 1)$. Then for any $t \in \mathbb{R}$,*

$$\left|\mathbf{Pr}\left[\sum c_i x_i \leqslant t\right] - \mathbf{Pr}[g \leqslant t]\right| \leqslant \alpha.$$

Theorem A.3. *(Chebyshev's Inequality) Let X be a random variable with expected value u and variance σ^2. Then for any real number $t > 0$,*

$$\mathbf{Pr}[|X - \mu| \geqslant t \cdot \sigma] \leqslant 1/t^2.$$

B Proof of Lemma 3.3

Recall that each $y^{(i)}$ is generated by the following manner:

$$y^{(i)} = \begin{cases} x^{(i)} & \text{with probability } 1 - \gamma \\ \text{random bit} & \text{with probability } \gamma. \end{cases} \tag{10}$$

Let us define a random vector $\boldsymbol{z} \in \{0, 1\}^n$ based on $\boldsymbol{y}$. For $\boldsymbol{y}$ generated, if $y^{(i)}$ is generated as a copy of $x^{(i)}$ in (10), then $\boldsymbol{z}^{(i)} = 0$; if $y^{(i)}$ is generated as a random bit in (10), then $z^{(i)} = 1$. Let us write $S = \sum_{i=1}^{n} w^{(i)} y^{(i)}$. Our proof is based on two claims.

Claim B.1. *For a τ-regular vector $\boldsymbol{w}$,* $\mathbf{Pr}[\sum_{i=1}^{n} |w^{(i)}|^2 z^{(i)} \geqslant \gamma/2] \geqslant 1 - 2e^{-\frac{\gamma^2}{2\tau^2}}$.

Claim B.2. *For a τ-regular vector $\boldsymbol{w}$, given any $a' < b' \in \mathbb{R}$ and any fixing of $z^{(1)}, z^{(2)}, \ldots, z^{(n)}$, if $\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)} = \sigma^2 > 0$, then* $\mathbf{Pr}[S \in [a', b']] \leqslant \frac{2|b'-a'|}{\sigma} + \frac{2\tau}{\sigma}$.

Given the above two claims are correct, define event V to be $\{\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)} \geqslant \frac{\gamma}{2}\}$ and use $\mathbf{1}_{[a,b]}(x) : \mathbb{R} \to \{0, 1\}$ to denote the indicator function of whether x falls into interval $[a, b]$.

$$\mathbf{Pr}[S \in [a, b]] = \mathbf{E}[\mathbf{1}_{[a,b]}(S)] = \mathbf{Pr}[V]\,\mathbf{E}[\mathbf{1}_{[a,b]}(S) \mid V] + \mathbf{Pr}[\neg V]\,\mathbf{E}[\mathbf{1}_{[a,b]}(S) \mid \neg V]$$

By Claim B.1,

$$\mathbf{Pr}[\neg V]\,\mathbf{E}[\mathbf{1}_{[a,b]}(S) \mid \neg V] \leqslant \mathbf{Pr}[\neg V] \leqslant 2e^{-\frac{\gamma^2}{2\tau^2}}.$$

By Claim B.2,

$$\mathbf{Pr}[V]\,\mathbf{E}[\mathbf{1}_{[a,b]}(S) \mid V] \leqslant \frac{4(b-a)}{\sqrt{\gamma}} + \frac{4\tau}{\sqrt{\gamma}}.$$

Overall,

$$\mathbf{Pr}\left[S \in [a, b]\right] \leqslant \frac{4(b-a)}{\sqrt{\gamma}} + \frac{4\tau}{\sqrt{\gamma}} + 2e^{-\frac{\gamma^2}{2\tau^2}}.$$

It remains to verify Claim B.1 and Claim B.2.

To prove Claim B.1, we need to apply the Hoeffding's inequality (see Theorem A.1).

Notice that $(w^{(i)})^2 z^{(i)} \in [0, (w^{(i)})^2]$ and applying Hoeffding's Inequality, we know

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)} - \mathbf{E}\left[\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)}\right]\right| \geqslant nt\right] \leqslant 2e^{\frac{-2n^2t^2}{\sum_{i=1}^{n} (w^{(i)})^4}}.$$

We know $\mathbf{E}[\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)}] = \gamma$ and $\sum_{i=1}^{n} ((w^{(i)})^2)^2 \leqslant \max_i \left\{(w^{(i)})^2\right\} \sum_{i=1}^{n} (w^{(i)})^2 \leqslant \tau^2$. If we take $nt = \gamma/2$, we have

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)} - \gamma\right| \geqslant \frac{\gamma}{2}\right] \leqslant 2e^{-\frac{\gamma^2}{2\tau^2}}.$$

Therefore, with probability at least $1 - 2e^{-\frac{\gamma^2}{2\tau^2}}$, $\sum_{i=1}^{n} (w^{(i)})^2 z^{(i)} \geqslant \frac{\gamma}{2}$.

To prove Claim B.2, we need use Berry-Esseen Theorem (See Theorem A.2). Let us split S into two parts: $S' = \sum_{z_i=1} w_i y_i$ and $S'' = \sum_{z_i=0} w_i y_i$. Since $S = S' + S''$ and S' is independent of S'', it suffices to show that $\mathbf{Pr}\left[S' \in [a', b']\right] \leqslant \frac{2|b'-a'|}{\sqrt{\sigma}} + \frac{2\tau}{\sigma}$ for any $a', b' \in \mathbb{R}$. Define $y'^{(i)} = 2y^{(i)} - 1$ and note that $y'^{(i)}$ a $\{-1, 1\}$ variable. By rewriting S' using this definition, we have

$$S' = \sum_{z^{(i)}=1} w^{(i)} y^{(i)} = \sum_{z^{(i)}=1} w^{(i)} \frac{1 + y'^{(i)}}{2}.$$

Then

$$\mathbf{Pr}\left[S' \in [a', b']\right] = \mathbf{Pr}\left[\sum_{z^{(i)}=1} w^{(i)} y'^{(i)} \in [a'', b'']\right], \tag{11}$$

where $a'' = 2a' - \sum_{z^{(i)}=1} w^{(i)}$ and $b'' = 2b' - \sum_{z^{(i)}=1} w^{(i)}$. We can further rewrite the above term as

$$\mathbf{Pr}\left[\sum_{z^{(i)}=1} w^{(i)} y'^{(i)} \leqslant b''\right] - \mathbf{Pr}\left[\sum_{z^{(i)}=1} w^{(i)} y'^{(i)} \leqslant a''\right]$$
$$= \mathbf{Pr}\left[\sum_{z^{(i)}=1} \frac{w^{(i)} y'^{(i)}}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}} \leqslant \frac{b''}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}\right] - \mathbf{Pr}\left[\sum_{z^{(i)}=1} \frac{w^{(i)} y'^{(i)}}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}} \leqslant \frac{a''}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}\right].$$

We can now apply Berry-Esseen's theorem. Notice that for all the i such that $z^{(i)} = 1$, $y'^{(i)}$ is distributed as an independent unbiased random $\{-1, 1\}$ variable. Also $\max_{z^{(i)}=1} \frac{|w^{(i)}|}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}} \leqslant \frac{\tau}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}$.

By Berry-Esseen's theorem, we know that expression (11) is bounded by

$$\mathbf{Pr}\left[N(0,1) \leqslant \frac{b''}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}\right] - \mathbf{Pr}\left[N(0,1) \leqslant \frac{a''}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}\right] + \frac{2\tau}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}}.$$

Using the fact that a unit Gaussian variable falls in any interval of length λ with probability at most λ and noticing that $b'' - a'' = 2(b' - a')$, we can bound the above quantity by

$$\frac{2|b' - a'|}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}} + \frac{2\tau}{\sqrt{\sum_{z^{(i)}=1}(w^{(i)})^2}} = \frac{2|b-a|}{\sigma} + \frac{2\tau}{\sigma}.$$

C Proof of Invariance Principle (Th. 3.10)

We restate our version of the invariance principle here for convenience.

Theorem 3.10 restated (Invariance Principle) Let $\mathcal{A} = \{\boldsymbol{A}^{\{1\}}, \ldots, \boldsymbol{A}^{\{R\}}\}, \mathcal{B} = \{\boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{R\}}\}$ be families of ensembles of random variables with $\boldsymbol{A}^{\{i\}} = \{a_1^{(i)}, \ldots, a_{k_i}^{(i)}\}$ and $\boldsymbol{B}^{\{i\}} = \{b_1^{(i)}, \ldots, b_{k_i}^{(i)}\}$, satisfying the following properties:

- For each $i \in [R]$, the random variables in ensembles $(\boldsymbol{A}^{\{i\}}, \boldsymbol{B}^{\{i\}})$ have matching moments up to degree 3. Further all the random variables in $\mathcal{A}$ and $\mathcal{B}$ are bounded by 1.
- The ensembles $\boldsymbol{A}^{\{i\}}$ are all independent of each other, similarly the ensembles $\boldsymbol{B}^{\{i\}}$ are independent of each other.

Given a set of vectors $\boldsymbol{l} = \{\boldsymbol{l}^{\{1\}}, \ldots, \boldsymbol{l}^{\{R\}}\}(\boldsymbol{l}^{\{i\}} \in \mathbb{R}^{k_i})$, define the linear function $\boldsymbol{l} : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_R} \to \mathbb{R}$ as

$$\boldsymbol{l}(\boldsymbol{x}) = \sum_{i \in [R]} \langle \boldsymbol{l}^{\{i\}}, \boldsymbol{x}^{\{i\}} \rangle$$

Then for a K-bounded function $\Psi : \mathbb{R} \to \mathbb{R}$ we have

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{A}) - \theta \Big) \right] - \mathop{\mathbf{E}}_{\mathcal{B}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{B}) - \theta \Big) \right] \right| \leqslant K \sum_{i \in [R]} \|\boldsymbol{l}^{\{i\}}\|_1^4. \tag{12}$$

for all $\theta > 0$. Further, define the spread function $c(\alpha)$ corresponding to the ensembles $\mathcal{A}, \mathcal{B}$ and the linear function $\boldsymbol{l}$ as follows,

(**Spread Function:**)For $1/2 > \alpha > 0$, let

$$c(\alpha) = \max \big(\sup_{\theta} \mathbf{Pr}_{\mathcal{A}} \Big[\boldsymbol{l}(\mathcal{A}) \in [\theta - \alpha, \theta + \alpha] \Big], \quad \sup_{\theta} \mathbf{Pr}_{\mathcal{B}} \Big[\boldsymbol{l}(\mathcal{B}) \in [\theta - \alpha, \theta + \alpha] \Big] \big)$$

then for all $\tilde{\theta}$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\mathrm{pos}\left(\boldsymbol{l}(\mathcal{A}) - \tilde{\theta} \right) \right] - \quad \mathbf{E}_{\mathcal{B}} \left[\mathrm{pos}\left(\boldsymbol{l}(\mathcal{B}) - \tilde{\theta} \right) \right] \right| \leqslant O\Big(\tfrac{1}{\alpha^4} \Big) \textstyle\sum_{i \in [R]} \|\boldsymbol{l}^{\{i\}}\|_1^4 + 2c(\alpha). \tag{13}$$

Proof. Let us prove equation (12) first. Let $\mathcal{X}_i = \{\boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{i-1\}}, \boldsymbol{B}^{\{i\}}, \boldsymbol{A}^{\{i+1\}}, \ldots, \boldsymbol{A}^{\{R\}}\}$.

We know that

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathcal{A}}[\Psi(\boldsymbol{l}(\mathcal{A}) - \theta)] - \mathop{\mathbf{E}}_{\mathcal{B}}[\Psi(\boldsymbol{l}(\mathcal{B}) - \theta)] &= \mathop{\mathbf{E}}_{\mathcal{X}_0}[\Psi(\boldsymbol{l}(\mathcal{X}_0) - \theta)] - \mathop{\mathbf{E}}_{\mathcal{X}_R}[\Psi(\boldsymbol{l}(\mathcal{X}_R) - \theta)] \\ &= \sum_{i=1}^{R} \mathop{\mathbf{E}}_{\mathcal{X}_{i-1}}[\Psi(\boldsymbol{l}(\mathcal{X}_{i-1}) - \theta)] - \mathop{\mathbf{E}}_{\mathcal{X}_i}[\Psi(\boldsymbol{l}(\mathcal{X}_i) - \theta)]. \end{aligned}$$

Therefore, it suffices to prove

$$\Big| \mathop{\mathbf{E}}_{\mathcal{X}_{i-1}}[\Psi(\boldsymbol{l}(\mathcal{X}_{i-1}) - \theta)] - \mathop{\mathbf{E}}_{\mathcal{X}_i}[\Psi(\boldsymbol{l}(\mathcal{X}_i) - \theta)] \Big| \leqslant K \|\boldsymbol{l}^{\{i\}}\|_1^4. \tag{14}$$

Let $\mathcal{Y}_i = \{\boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{i-1\}}, \boldsymbol{A}^{\{i+1\}}, \ldots, \boldsymbol{A}^{\{R\}}\}$ and we have $\mathcal{X}_i = \{\mathcal{Y}_i, \boldsymbol{B}^{\{i\}}\}$ and $\mathcal{X}_{i-1} = \{\mathcal{Y}_i, \boldsymbol{A}^{\{i\}}\}$. Then

$$\mathop{\mathbf{E}}_{\mathcal{X}_{i-1}}[\Psi(\boldsymbol{l}(\mathcal{X}_{i-1}) - \theta)] - \mathop{\mathbf{E}}_{\mathcal{X}_i}[\Psi(\boldsymbol{l}(\mathcal{X}_i) - \theta)] = \mathop{\mathbf{E}}_{\mathcal{Y}_i} \left[\mathop{\mathbf{E}}_{A^{\{i\}}}[\Psi(\boldsymbol{l}(\mathcal{X}_{i-1}) - \theta)] - \mathop{\mathbf{E}}_{B^{\{i\}}}[\Psi(\boldsymbol{l}(\mathcal{X}_i) - \theta)] \right]. \tag{15}$$

Notice that

$$\boldsymbol{l}(\mathcal{X}_{i-1}) - \theta = \langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}} \rangle + \sum_{1 \leqslant j \leqslant i-1} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}} \rangle + \sum_{i+1 \leqslant j \leqslant R} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}} \rangle - \theta$$

and

$$\boldsymbol{l}(\mathcal{X}_i) - \theta = \langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}} \rangle + \sum_{1 \leqslant j \leqslant i-1} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}} \rangle + \sum_{i+1 \leqslant j \leqslant R} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}} \rangle - \theta.$$

Take $\theta' = \sum_{1\leqslant j\leqslant i-1}\langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}}\rangle + \sum_{i+1\leqslant j\leqslant R}\langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}}\rangle - \theta$, We can further rewrite equation (15) as

$$\mathop{\mathbf{E}}_{\mathcal{Y}_i}\Big[\mathop{\mathbf{E}}_{A^{\{i\}}}[\Psi(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle + \theta')] - \mathop{\mathbf{E}}_{B^{\{i\}}}[\Psi(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle + \theta')]\Big]. \tag{16}$$

Using the Taylor expansion of Ψ, we have that the inner expectation of equation (16) is equal to

$$\begin{aligned}\Big|\mathop{\mathbf{E}}_{A^{\{i\}}}[\Psi(\theta') + \Psi'(\theta')\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle + \frac{\Psi''(\theta')}{2}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle)^2 + \frac{\Psi'''(\theta')}{6}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle)^3 + \frac{\Psi''''(\delta_1)}{24}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle)^4] \\ - \mathop{\mathbf{E}}_{B^{\{i\}}}[\Psi(\theta') + \Psi'(\theta')\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle + \frac{\Psi''(\theta')}{2}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle)^2 + \frac{\Psi'''(\theta')}{6}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle)^3 + \frac{\Psi''''(\delta_2)}{24}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle)^4]\Big|.\end{aligned} \tag{17}$$

for some $\delta_1, \delta_2 \in \mathbb{R}$.

Using the fact that $\boldsymbol{A}^{\{i\}}$ and $\boldsymbol{B}^{\{i\}}$ have matching moments up to degree 3, we can upper bound equation (17) by

$$\left|\mathop{\mathbf{E}}_{A^{\{i\}}}[\frac{\Psi''''(\delta_1)}{24}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle)^4] - \mathop{\mathbf{E}}_{B^{\{i\}}}[\frac{\Psi''''(\delta_2)}{24}(\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle)^4]\right| \leqslant \frac{K}{12}|\boldsymbol{l}^{\{i\}}|_1^4.$$

In the last inequality, we use the fact that Ψ is K-bounded and $\langle \boldsymbol{l}^{\{i\}}, \boldsymbol{A}^{\{i\}}\rangle \leqslant \|\boldsymbol{l}^{\{i\}}\|_1, \langle \boldsymbol{l}^{\{i\}}, \boldsymbol{B}^{\{i\}}\rangle \leqslant \|\boldsymbol{l}^{\{i\}}\|_1$ since all random variables in $\mathcal{A}, \mathcal{B}$ are bounded by 1.

Overall, we bound the inner expectation of equation (16) by $\frac{K}{12}\|\boldsymbol{l}^{\{i\}}\|_1^4$. This implies equation (16) and therefore equation (14) is bounded by $\frac{K}{12}\|\boldsymbol{l}^{\{i\}}\|_1^4$, establishing equation (12).

To prove equation (13), we need to use the following lemma.

Lemma C.1. *([37], Lemma 3.21) There exists an absolute constant C such that $\forall 0 < \lambda < \frac{1}{2}$, there exists $\frac{C}{\lambda^4}$-bounded function $\Phi_\lambda : \mathbb{R} \to [0,1]$ which approximates the $\mathrm{pos}(x)$ function in the following sense: $\Phi_\lambda(t) = 1$ for all $t > \lambda$; $\Phi_\lambda(t) = 0$ for $t < -\lambda$.*

By the above lemma, we can find a $\frac{C}{\alpha^4}$-bounded function Φ_α such that $\Phi_\alpha(\boldsymbol{l}(\mathcal{A}) - \theta)$ is equal to $\mathrm{pos}(\boldsymbol{l}(\mathcal{A}) - \theta)$ except when $\boldsymbol{l}(\mathcal{A}) \in [\theta - \alpha, \theta + \alpha]$ and $\Phi_\alpha(\boldsymbol{l}(\mathcal{B}) - \theta)$ is equal to $\mathrm{pos}(\boldsymbol{l}(\mathcal{B}) - \theta)$ except when $\boldsymbol{l}(\mathcal{B}) \in [\theta - \alpha, \theta + \alpha]$. Also for any $x \in \mathbb{R}$, $|\mathrm{pos}(x) - \Phi_\alpha(x)| \leqslant 1$ as $\mathrm{pos}(x)$ and $\Phi_\alpha(x)$ are both in $[0,1]$.

Overall, we have

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\mathcal{A}}\Big[\mathrm{pos}\Big(\boldsymbol{l}(\mathcal{A}) - \theta\Big)\Big] - \mathop{\mathbf{E}}_{\mathcal{B}}\Big[\mathrm{pos}\Big(\boldsymbol{l}(\mathcal{B}) - \theta\Big)\Big]\right| &\leqslant \left|\mathop{\mathbf{E}}_{\mathcal{A}}\Big[\mathrm{pos}\Big(\boldsymbol{l}(\mathcal{A}) - \theta\Big)\Big] - \mathop{\mathbf{E}}_{\mathcal{A}}\Big[\Phi_\alpha\Big(\boldsymbol{l}(\mathcal{A}) - \theta\Big)\Big]\right| \\ &+ \left|\mathop{\mathbf{E}}_{\mathcal{A}}\Big[\Phi_\alpha\Big(\boldsymbol{l}(\mathcal{A}) - \theta\Big)\Big] - \mathop{\mathbf{E}}_{\mathcal{B}}\Big[\Phi_\alpha\Big(\boldsymbol{l}(\mathcal{B}) - \theta\Big)\Big]\right| + \left|\mathop{\mathbf{E}}_{\mathcal{B}}\Big[\Phi_\alpha\Big(\boldsymbol{l}(\mathcal{B}) - \theta\Big)\Big] - \mathop{\mathbf{E}}_{\mathcal{B}}\Big[\mathrm{pos}\Big(\boldsymbol{l}(\mathcal{B}) - \theta\Big)\Big]\right| \\ &\leqslant \frac{C}{\alpha^4}\sum_{i\in[R]}\|\boldsymbol{l}^{\{i\}}\|_1^4 + 2c(\alpha).\end{aligned}$$

□

D Hardness of Smooth k-Label Cover

First we state the bipartite smooth Label Cover given by Khot [30]. Our reduction is similar to the one in [19] but in addition requires proving the smoothness property.

Definition D.1. *A Label Cover problem $\mathcal{L}(G(W,V,E),M,N,\{\pi^{v,w}|(w,w)\in E\})$ consists of a bipartite graph $G(V,W,E)$ with bipartition V and W. M,N are two positive integers such that $M>N$. There are projection functions $\pi^{v,w}:[M]\to[N]$ associated with each edge $(w,v)\in E$ where $v\in V,w\in W$. All vertices in W have the same degree (i.e., W-side regular). For any labeling $\Lambda:V\to[M]$ and $\Lambda:W\to[N]$, an edge is said to be satisfied if $\pi^{v,w}(\Lambda(v))=\Lambda(w)$. We define $Opt(\mathcal{L})$ to be the maximum fraction of edges satisfied by any labeling.*

Theorem D.2. *There is an absolute constant $\gamma>0$ such that for all integer parameters u and J, it is NP-hard to distinguish the following two cases: A Label Cover problem $\mathcal{L}(G(W,V,E),N,M,\{\pi^{v,w}|(w,v)\in E\})$ with $M=7^{(J+1)u}$ and $N=2^u7^{Ju}$ having*

- *$Opt(\mathcal{L})=1$ or*
- *$Opt(\mathcal{L})\leqslant 2^{-2\gamma u}$.*

In addition, the Label Cover has the following properties:

- *for each $\pi^{v,w}$ and any $i\in[N]$, we have $|(\pi^{v,w})^{-1}(i)|\leqslant 4^u$;*
- *for a fixed vertex w and a randomly picked neighbor v of w,*

$$\forall i,j\in[M],\mathbf{Pr}[\pi^{v,w}(i)=\pi^{v,w}(j)]\leqslant 1/J.$$

Below we prove Theorem 5.1.

Proof. Given an instance of bipartite Label Cover $\mathcal{L}(G(V,W,E),M,N,\{\pi^{v,w}|(w,v)\in E\})$, we can convert it to a smooth k-Label Cover instance $\mathcal{L}'$ as follows. The vertex set of $\mathcal{L}'$ is V and we generate the hyperedge set E' and projections associated with the hyperedges in the following way:

1. pick a vertex $w\in W$;
2. pick a k-tuple of w's neighbors $v_1,\ldots,v_k$ and add a hyperedge $e=(v_1,\ldots,v_k)$ to E' with projections $\pi^{v_i,e}=\pi^{v_i,w}$ for each $i\in[k]$.

Completeness: If $Opt(\mathcal{L})=1$, then there exists a labeling Λ such that for every edge $(w,v)\in E$, $\pi^{v,w}(\Lambda(v))=\Lambda(w)$. We can simply take the restriction of labeling Λ on V for the smooth k-Label Cover instance $\mathcal{L}'$. For any hyperedge $e=(v_1,v_2,\ldots,v_k)$ generated by $w\in W$, we know $\pi^{v_i,e}(\Lambda(v_i))=\Lambda(w)=\pi^{v_j,e}(\Lambda(v_j))$ for any $i,j\in[k]$.

Soundness: If $Opt(\mathcal{L}) \leqslant 2^{-2\gamma u}$, then we can weakly satisfy at most $2k^2 2^{-\gamma u}$-fraction of the hyperedges in $\mathcal{L}'$. This can be proved via contrapositive argument. Suppose there is a labeling strategy Λ (defined on V) for the smooth k-Label Cover that weakly satisfies $\alpha \geqslant 2k^2 2^{-\gamma u}$ fraction of the hyperedges. Extend the labelling to W as follows: For each vertex $w \in W$ and a neighbor $v \in V$, let $\pi_{v,w}(\Lambda(v))$ be the label *recommended* by v to w. Simply assign for every vertex $w \in W$, the label most *recommended* by its neighbours.

By the fact that Λ weakly satisfies α-fraction of hyperedges in $\mathcal{L}'$, we know that if we pick a vertex w and randomly pick two of its neighbors v_1, v_2 then

$$\mathbf{Pr}\left[\pi^{v_1,w}(\Lambda(v_1)) = \pi^{v_2,w}(\Lambda(v_2))\right] \geqslant \frac{\alpha}{\binom{k}{2}} \geqslant \frac{2\alpha}{k^2}.$$

By an averaging argument, at least $\frac{\alpha}{k^2}$-fraction of the vertices $w \in W$, will have the following property: among all the possible pairs of w's neighbors, at least $\frac{\alpha}{k^2}$-fraction of pairs *recommend* the same label for w. Let us call such a w to be a *nice*. It is easy to see that for every *nice* w, the most *recommended* label is actually recommended by at least $\frac{\alpha}{k^2}$ fraction of its neighbours. Hence, the extended labelling satisfies at least α/k^2 fraction of edges incident at each *nice* $w \in W$. Using W-side regularity, we conclude that the extended labelling satisfies $\frac{\alpha^2}{k^4} = 4 \cdot 2^{-2\gamma u}$-fraction the edges of $\mathcal{L}$ – a contradiction.

Smoothness of $\mathcal{L}'$: For any given vertex v in $\mathcal{L}'$, we want so show that if we randomly pick an hyperedge e' containing v, then for the projection $\pi^{v,e}$ as defined in $\mathcal{L}'$,

$$\forall i, j \in [M], \mathbf{Pr}[\pi^{v,e'}(i) = \pi^{v,e'}(j)] \leqslant \frac{1}{J}.$$

To see this, notice that all vertices in W have the same degree; picking a projection $\pi^{v,e'}$ using the above procedure is the same as randomly picking a neighbor w of v and using the projection $\pi^{v,w}$ defined in $\mathcal{L}$. Therefore,

$$\forall i, j \in [M], \mathbf{Pr}[\pi^{v,e'}(i) = \pi^{v,e'}(j) = \mathbf{Pr}[\pi^{v,w}(i) = \pi^{v,w}(j)] \leqslant \frac{1}{J}.$$

□

ECCC ISSN 1433-8092

http://eccc.hpi-web.de