Agnostic Learning of Monomials by Halfspaces is Hard

Vitaly Feldman* Venkatesan Guruswami† Prasad Raghavendra‡ Yi Wu†
*IBM Almaden Research Center
San Jose, CA
Email: vitaly@post.harvard.edu
† Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: yiwu,guruswami@cmu.edu
‡Microsoft Research New England.
Cambridge, MA.
Email:prasad@cs.washington.edu

Abstract— We prove the following strong hardness result for learning: Given a distribution on labeled examples from the hypercube such that there exists a monomial (or conjunction) consistent with $(1-\epsilon)$-fraction of the examples, it is NP-hard to find a halfspace that is correct on $(\frac{1}{2}+\epsilon)$-fraction of the examples, for arbitrary constant $\epsilon > 0$. In learning theory terms, weak agnostic learning of monomials by halfspaces is NP-hard. This hardness result bridges between and subsumes two previous results which showed similar hardness results for the proper learning of monomials and halfspaces. As immediate corollaries of our result, we give the first optimal hardness results for weak agnostic learning of decision lists and majorities.

Our techniques are quite different from previous hardness proofs for learning. We use an invariance principle and sparse approximation of halfspaces from recent work on fooling halfspaces to give a new natural list decoding of a halfspace in the context of dictatorship tests/label cover reductions. In addition, unlike previous invariance principle based proofs which are only known to give Unique Games hardness, we give a reduction from a smooth version of Label Cover that is known to be NP-hard.

Keywords-Hardness of Learning, PCPs, Agnostic Learning, Dictatorship Tests.

1. Introduction

Monomials (conjunctions), 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 [41], [7], [38]. However, in practice data is often noisy or too complex to be consistently explained by a simple concept. A general model for learning that addresses this scenario is the *agnostic* learning model [20], [25]. Under agnostic learning, a learning algorithm for a class of functions $\mathcal{C}$ is required to classify the examples drawn from some unknown distribution nearly as well as is possible by a hypothesis from $\mathcal{C}$. Learning algorithms are referred to as *proper* when they output a hypothesis in $\mathcal{C}$.

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 [39], Winnow [33], and Support Vector Machines [42] 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.

Uniform convergence results in Haussler's work [20] (see also [25]) imply that proper agnostic learning is equivalent to the ability to come up with a function in a concept class $\mathcal{C}$ that has the optimal agreement rate with the given set of examples. This problem is referred to as the *Maximum Agreement* problem for $\mathcal{C}$. The Maximum Agreement problem for halfspaces is long known to be NP-complete [22] (see also Hemisphere problem in [17]). Two natural ways to relax the problem are: (i) allow the learning algorithm to output a hypothesis with any non-trivial (and not necessarily close to optimal) performance, and (ii) strengthen the assumptions on the examples that are given to the learning algorithm. In the main result of this work, we prove a strong hardness result for agnostic learning of halfspaces with both of these relaxations.

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.

Since the class of monomials is a subset of the class of decision lists which in turn is a subset of the class of halfspaces, our result implies an optimal hardness result for proper agnostic learning of decision lists. In addition, a similar hardness result for proper agnostic learning of

majority functions can be obtained via a simple reduction.

1.1. Previous work

A number of hardness results for proper agnostic learning of monomials, decision lists and halfspaces have appeared in the literature. 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 [26]. The hardness of approximating the problem within some constant factor (i.e., APX-hardness) was first shown by Ben-David *et al.* [5]. The factor was improved to $58/59$ by Bshouty and Burroughs [9]. Finally, Feldman showed a tight inapproximability result [14] (see also [15]), 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 the instances where every monomials is consistent with at most $(1/2+\epsilon)$-fraction of the examples. Recently, Khot and Saket [31] 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).

The Maximum Agreement problem for halfspaces was shown to be NP-hard to approximate by Amaldi and Kann [1], Ben-David *et al.* [5], and Bshouty and Burroughs [9] for approximation factors $\frac{261}{262}$, $\frac{415}{418}$, and $\frac{84}{85}$, respectively. An optimal inapproximability result was established independently by Guruswami and Raghavendra [19] 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] produced examples with real-valued coordinates, whereas the proof in [19] worked also for examples drawn from the Boolean hypercube.

For the concept class of decisions lists, APX-hardness of the Maximum Agreement problem was shown by Bshouty and Burroughs [9].

We note that our result *subsumes all these results* except [31] since we obtain the optimal inapproximability factor *and* allow learning of monomials by halfspaces.

A number of hardness of approximation results are also known for the symmetric problem of minimizing disagreement for each of the above concept classes [25], [21], [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 [32]. 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 — another long-standing open problem in learning theory and coding [23].

Monomials, decision lists and halfspaces are known to be efficiently learnable in the presence of more benign *random* classification noise [2], [24], [27], [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 suggests 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 $2^{\tilde{O}(\sqrt{n})}$ [23]. 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 linear thresholds of single variables).

2. Proof Overview

We show 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 $k \geqslant 2$, a instance of k-Label Cover $\mathcal{L}(G(V,E), d, R, \{\pi^{v,e} | e \in E, v \in e\})$ consists of a k-uniform connected hypergraph $G(V,E)$ with vertex set V, an edge set E, and finally a set of labels $[dR] = \{1, 2, \ldots, dR\}$ (Here d and R are positive integers). 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 $\pi^{v_i,e} : [dR] \to [R]$.*

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

The goal in Label Cover is to find a labelling that satisfies as many consistency checks (projection constraints) as possible.

For the sake of exposition, we first present the proof of Theorem 1.1 assuming the Unique Games Conjecture. In this light, we will be interested in the k-Unique Label Cover problem which is a special case of k-Label Cover where $d = 1$, and all the functions $\{\pi^{v,e} | v \in e, e \in E\}$ are bijections. The following strong inapproximability result for k-Unique Label Cover can be easily shown to be equivalent to the Unique Games Conjecture of Khot [29].

Conjecture 2.2. *For every $\eta > 0$ and a positive integer k, there exists R_0 such that for all positive integers $R > R_0$, given an instance $\mathcal{L}(G(V,E), 1, R, \{\pi^{v,e} | e \in E, v \in e\})$ it is* NP*-hard to distinguish whether,*

- *(Strongly satisfiable instances) There exists a labelling $\mathcal{A} : V \to [R]$ that strongly satisfies $1 - k\eta$ fraction of the edges E.*
- *(Near 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 set of labelled examples such that the following holds: If $\mathcal{L}$ is strongly satisfiable instance, then there is a monomial (an OR function) that agrees with $1-\epsilon$ fraction of the examples, while if $\mathcal{L}$ is a near unsatisfiable instance then no halfspace has agreement more than $\frac{1}{2} + \epsilon$. Clearly, a reduction of this nature immediately implies Theorem 1.1 under the Unique Games Conjecture.

Let $\mathcal{L}$ be an instance of k-UNIQUE LABEL COVER with an associated 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 $\boldsymbol{x} = (x_v^{(\ell)})_{v\in V, \ell\in[R]}$.

For every labelling $\mathcal{A} : V \to [R]$ of the instance, there is a corresponding OR function (a monomial) over $\{0,1\}^{|V|\times R}$ given by,

$$\mathcal{A} \leftrightarrow h(\boldsymbol{x}) = \bigvee_v x_v^{(\mathcal{A}(v))}.$$

Thus, using a label ℓ for a vertex v is encoded as including the literal $x_v^{(\ell)}$ in to the disjunction. Notice that an arbitrary halfspace over $\{0,1\}^{|V|\times R}$ need not correspond to any labelling at all. The idea would be to construct examples which ensure that any halfspace $\frac{1}{2} + \epsilon$ agreement somehow corresponds to a labelling of $\mathcal{L}$ 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 require a set of examples $\mathcal{D}_e$ with the following properties:

- All coordinates $x_v^{(\ell)}$ for a vertex $v \notin e$ are fixed to be zero. Restricted to these examples, the halfspace h can be written as $h(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i\in[k]} \langle \boldsymbol{w}_{v_i}, \boldsymbol{x}_{v_i}\rangle - \theta)$.
- For any label $\ell \in [R]$, the labelling $L(v_1) = \ldots = L(v_k) = \ell$ *strongly* satisfies the edge e. Hence, the corresponding disjunction $\vee_{i\in[k]} x_{v_i}^{(\ell)}$ has agreement $1-\epsilon$ with the examples $\mathcal{D}_e$.
- There exists a decoding procedure that given a halfspace h outputs a labelling L_h for $\mathcal{L}$ such that, if h has agreement $\frac{1}{2} + \epsilon$ on the set of examples $\mathcal{D}_e$, then L_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 labelling. Further, let us suppose the halfspace h is given by $h(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{v\in V} \langle \boldsymbol{w}_v, \boldsymbol{x}_v\rangle - \theta\right)$. Define the linear function $l_v : \{0,1\}^R \to \mathbb{R}$ as $l_v(\boldsymbol{x}_v) = \langle \boldsymbol{w}_v, \boldsymbol{x}_v\rangle$. Then, we have $h(\boldsymbol{x}) = \mathrm{sgn}(\sum_{v\in V} l_v(\boldsymbol{x}_v) - \theta)$.

For a halfspace h corresponding to a labelling L, we will have $l_v(\boldsymbol{x}_v) = x_v^{(L(v))}$ – a *dictator* function. Formally, the ℓ'th dictator function on $\{0,1\}^R$ is given by $F(\boldsymbol{x}) = x^{(\ell)}$. Thus, in the intended solution every linear function l_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 l_i for the linear function l_{v_i}. We need a randomized procedure that does the following:

> Given k linear functions $l_1, \ldots, l_k : \{0,1\}^R \to \mathbb{R}$, queries the functions at one point each (say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ respectively), and accepts if $\mathrm{sgn}(\sum_{i=1}^k l_i(\boldsymbol{x}_i) - \theta) = b$.

The procedure must satisfy,

- (Completeness) If each of the linear functions l_i is the ℓ'th dictator function for some $\ell \in [R]$, 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 $\mathrm{sgn}(10^{100}x_1 + x_2 - 0.5)$, although the coordinate x_2 has little influence on the linear function, it has the significant influence on the halfspace.

In this light, we make use of the notion of *critical index* (Definition 3.1) that was defined in [40] and has found numerous applications in the context of halfspaces [34], [37], [13]. Roughly speaking, given a linear function l, 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 l. Let $l_\tau(\boldsymbol{w}_i)$ denote the critical index of $\boldsymbol{w}_i$, and let $L_\tau(\boldsymbol{w}_i)$ denote the set of $l_\tau(\boldsymbol{w}_i)$ 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 $L_\tau(\boldsymbol{w}_i)$. A function is *far* from every dictator if it has critical index $= 0$.

An important issue is that unlike the number of influential coordinates, the critical index of a linear function is not 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 performance of the halfspace h only changes by a negligible amount! Thus, we first carry out the zeroing operation for all rows with large critical index. This doesn't affect the performance of the halfspace h by much.

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 its rows and columns.

We construct two distributions $\mathcal{D}_0, \mathcal{D}_1$ on $\{0,1\}^k$ such that for $s = 0, 1$, we have $\Pr_{x \in \mathcal{D}_s}\left[\vee_{i=1}^k x_i = s\right] \geqslant 1 - \epsilon$ for $\epsilon = o_k(1)$ (this will ensure the completeness of the reduction, i.e., certain monomials pass with high probability). Further, the distributions 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 random noise. We define the following test (or equivalently, example-label pairs): 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 $h(\boldsymbol{x}) = \mathrm{sgn}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \theta)$ is a halfspace that passes the test with probability $1/2 + \delta$. 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., $h(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i=1}^k \langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle - \theta) = \mathrm{sgn}(\sum_{i=1}^R \langle \boldsymbol{w}^{(i)}, \boldsymbol{x}^{(i)} \rangle - \theta)$. Let us denote $l_i(\boldsymbol{x}) = \langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle$.

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 l_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, $h(\boldsymbol{x}) = \mathrm{sgn}(\sum_{j \in [R]} \langle \boldsymbol{w}^{(j)}, \boldsymbol{x}^{(j)} \rangle - \theta)$. Here $l(\boldsymbol{x}) = \sum_{j \in [R]} \langle \boldsymbol{w}^{(j)}, \boldsymbol{x}^{(j)} \rangle - \theta$ is a degree 1 polynomial in to 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 2 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.

Suppose no two linear functions l_i are *close* to the same dictator, i.e., $L_\tau(\boldsymbol{w}_i) \cap L_\tau(\boldsymbol{w}_j) = \emptyset$. In this case, we condition on the values of $x_i^{(j)}$ for $j \in L_\tau(\boldsymbol{w}_i)$ (note that we condition on at most *one* value in each column so the conditional distribution on each column still has matching first three moments), and then apply the invariance principle using the fact on deleting the coordinates in $L_\tau(\boldsymbol{w}_i)$, all the remaining coefficients of the weight vector $\boldsymbol{w}$ are small (by definition of critical index). This implies that $L_\tau(\boldsymbol{w}_i) \cap L_\tau(\boldsymbol{w}_j) \neq \phi$ for some two rows i, j. This finishes the proof of the soundness claim.

The above consistency-enforcing test almost immediately yields the Unique Games-hardness of weak learning monomials by halfspaces. To prove NP-hardness, 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 $L = \{\ell_1, \ldots, \ell_d\}$ corresponding to a vertex u to a single label ℓ for the vertex v. While composing the label cover constraint (u, v) with the dictatorship test, all labels in L have to be necessarily *equivalent*. In several settings including this work, this requires the coordinates corresponding to labels in L to be mostly identical! However, on making the coordinates corresponding to L 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 [30] 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 [28]) is one in which for every vertex u, and a pair of labels ℓ, ℓ', the labels $\{\ell, \ell'\}$ 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 halfspace, and use the smoothness to bound this error term for most edges of the label cover. It is of great interest to find other applications where a *weak uniqueness* property like the

smoothness condition can be used to convert a Unique games hardness result to an unconditional NP-hardness result.

Due to space constraints, we only present the details of the Unique games based hardness result here, and the details of the unconditional hardness result will appear in the full version.

3. Preliminaries

In this section, we define two important tools in our analysis: i) Critical Index, ii) Invariance Principle.

3.1. Critical Index

The critical index which was first introduced in [40] and played an important role in the dealing with halfspaces in [34], [37] and very recently in [13].

Definition 3.1. *Given any real vector $\boldsymbol{w} = (w^{(1)}, w^{(2)}, ..w^{(n)}) \in \mathbb{R}^n$. Reorder the coordinates by decreasing absolute value, i.e., $|w^{(i_1)}| \geqslant |w^{(i_2)}| ... \geqslant |w^{(i_n)}|$ 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 important observation from [13] is that if the critical index of a sequence is big, it must contain some geometric decreasing subsequence.

Lemma 3.2. *(Lemma 5.5 [13]) Given a vector $\boldsymbol{w} = (w^{(i)})_{i=1}^n$ such that $|w^{(1)}| \geqslant |w^{(2)}| ... \geqslant |w^{(n)}|$, 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 vector, following lemma bound the probability that its weighted sum falls into a small interval under some distribution of the weights. The proof of the following lemma will appear in the full version.

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 $\tilde{\mathcal{D}}$, sample $\boldsymbol{x}$ from $\mathcal{D}$,*

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

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

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

Intuitively, $\langle \boldsymbol{w}, \boldsymbol{y}\rangle$ is τ close to the Gaussian distribution if each $y^{(i)}$ is a random bit and therefore we can bound the probability that $\langle \boldsymbol{w}, \boldsymbol{y}\rangle$ falls into $[a,b]$. 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 therefore bound the probability that $\langle \boldsymbol{w}, \boldsymbol{y}\rangle$ falls into $[a,b]$.

Definition 3.4. *For a vector $\boldsymbol{w} \in \mathbb{R}^n$, define set of indices $S_t(\boldsymbol{w}) \subseteq [n]$ as the set of indices containing the t largest coordinates of $\boldsymbol{w}$ by absolute value. Suppose its τ-critical index is l_τ, define set of indices $L_\tau(\boldsymbol{w}) = S_{l_\tau}(\boldsymbol{w})$. In other words, $L_\tau(\boldsymbol{w})$ 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 is suggested by Lemma 3.2, vector with large critical index has some geometric decreasing subsequence. Following two lemmas are about bounding the probability that the weighted sum of a geometric decreasing sequence falls into a small interval. First, we restate Claim 5.7 from [13] here.

Lemma 3.6. *[Claim 5.7, [13]] Let $|w^{(1)}| \geqslant |w^{(2)}| \ldots \geqslant |w^{(T)}| \geqslant 0$ be a sequence of numbers so that $|w^{(i+1)}| \leqslant |\frac{w^{(i)}}{3}|$ 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 points $\boldsymbol{x} \in \{0,1\}^T$ such that $\langle \boldsymbol{w}, \boldsymbol{x}\rangle \in I$.*

Lemma 3.7. *Let $|w^{(1)}| \geqslant |w^{(2)}| \ldots \geqslant |w^{(T)}| \geqslant 0$ be a sequence of numbers so that $|w^{(i+1)}| \leqslant |\frac{w^{(i)}}{3}|$ for $1 \leqslant i \leqslant T-1$. $\mathcal{D}$ is 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

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

Proof: By Lemma 3.6, we know that for the interval $J = \left[\theta - \frac{|w^T|}{6}, \theta + \frac{|w^T|}{6}\right]$, 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 $(y_1^{(1)}, y_1^{(2)}, ...y_1^{(T)}) = (r^{(1)}, ..r^{(T)})$ 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, $\Pr[y^{(1)} = r^{(1)}, y^{(2)} = r^{(2)}, ...y^{(T)} = r^{(T)}] \leqslant \left(1 - \frac{\gamma}{2}\right)^T$. ∎

3.2. Invariance Principle

While invariance principles have been shown in various settings by [36], [11], [35], we restate a version of the principle well suited for our application. We present a self-contained proof for the it in the full version.

Definition 3.8. *A C^4-function $\Psi(x) : \mathbb{R} \to \mathbb{R}$ in is said to be B-nice if $|\Psi''''(t)| \leqslant B$ for all $t \in \mathbb{R}$.*

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

Theorem 3.10. *(Invariance Theorem) 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 $\mathcal{A}, \mathcal{B}$ 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 $\boldsymbol{l} : \mathbb{R}^{k_1} \times \ldots \times \mathbb{R}^{k_R} \to \mathbb{R}$ as

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

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

$$\left| \mathop{\mathbf{E}}_{\mathcal{A}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{A}) - \theta \Big) \right] - \mathop{\mathbf{E}}_{\mathcal{B}} \left[\Psi\Big(\boldsymbol{l}(\mathcal{B}) - \theta \Big) \right] \right| \leqslant B \sum_{i\in[R]} \|\boldsymbol{l}^{\{i\}}\|_1^4 \tag{2}$$

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 $\alpha > 0$, Let*

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

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

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

4. Construction of Dictator Test

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

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

The dictator 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 $\Pr_{\boldsymbol{x}\sim\mathcal{D}_0}\{$every x_i is $0\}$ $\geqslant$ $1 - \frac{2}{\sqrt{k}}$, $\Pr_{\boldsymbol{x}\sim\mathcal{D}_1}\{$every x_i is $0\} \leqslant \frac{1}{\sqrt{k}}$, while matching moments up to degree 4, i.e., $\forall a, b, c, d \in [k]$*

$$\mathop{\mathbf{E}}_{\mathcal{D}_0}[x_a] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_a] \quad \mathop{\mathbf{E}}_{\mathcal{D}_0}[x_a x_b x_c x_d] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_a x_b x_c x_d]$$
$$\mathop{\mathbf{E}}_{\mathcal{D}_0}[x_a x_b] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_a x_b] \qquad \mathop{\mathbf{E}}_{\mathcal{D}_0}[x_a x_b x_c] = \mathop{\mathbf{E}}_{\mathcal{D}_1}[x_a x_b x_c]$$

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 $\Pr_{\boldsymbol{x}\sim\mathcal{D}_0}\{$every x_i is $0\}$ $\geqslant$ $1 - (\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4)$ and $\Pr_{\boldsymbol{x}\sim\mathcal{D}_1}\{$every x_i is $0\} \leqslant \epsilon = \frac{1}{\sqrt{k}}$

It remains to decide 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:

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

$$\sum_{i=1}^4 \epsilon_i \left(\frac{i}{k^{1/3}}\right)^2 = \sum_{i=1}^4 \frac{\epsilon}{4} \left(\frac{i}{k^{\frac{1}{3}}}\right)^2 \tag{4}$$

$$\sum_{i=1}^4 \epsilon_i \left(\frac{i}{k^{\frac{1}{3}}}\right)^3 = \sum_{i=1}^4 \frac{\epsilon}{4} \left(\frac{i}{k^{\frac{1}{3}}}\right)^3 \tag{5}$$

$$\sum_{i=1}^4 \epsilon_i \left(\frac{i}{k^{\frac{1}{3}}}\right)^4 = \sum_{i=1}^4 \frac{\epsilon}{4} \left(\frac{i}{k^{\frac{1}{3}}}\right)^4 \tag{6}$$

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 big 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 $\tilde{D}_b$ on $\{0,1\}^k$ as follows:*

- *First generate $x \in \{0,1\}^k$ according to 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. *Since the noise is independent uniform random bits, when calculating some moments of y, such as $\mathbf{E}_{\tilde{D}_b}[y_{i_1} y_{i_2} ... y_{i_d}]$, we can substitute y_i by $(1-\gamma)x_i + \frac{1}{2}\gamma$. Therefore, degree d moment of y can be expressed as by weighted sum of moments of x of degree up to d. Since $\mathcal{D}_0$ and $\mathcal{D}_1$ has matching moments up to degree 4, $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ also have matching moments up to degree 4.*

The following Lemma asserts that conditioning the two distributions $\tilde{D}_0$ and $\tilde{D}_1$ on the same coordinate x_j being fixed to have a value b, the resulting conditional distribution of $\tilde{D}_0$ and $\tilde{D}_1$ still have matching moments up to degree 3. The proof of is fairly straightforward and is deferred to the full version.

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 $S \subseteq [k], |S| \leqslant d-1$, $j \in [k]$ and $c \in \{0,1\}$.*

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

4.2. The Dictator Test

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

1) Generate a random bit $b \in \{0,1\}$.
2) Generate $\boldsymbol{x} \in \{0,1\}^{kR}$ 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 example-label pair $(\boldsymbol{y}, b)$. Equivalently, 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 $\tilde{\mathcal{D}_1}^R$.

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

Proof: 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 $\{x_1^{(j)}, x_2^{(j)} ... x_k^{(j)}\}$ are set to 0. By union bound, with probability at least $1 - \frac{2}{\sqrt{k}} - \frac{1}{k}$, $\{y_1^{(j)}, y_2^{(j)} ... y_k^{j}\}$ 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 bit in $\{x_1^{(j)}, x_2^{(j)} ... x_k^{(j)}\}$ is set to 1 and by union bound one of $\{y_1^{(j)}, y_{(j)}^2 ... y_k^{(j)}\}$ 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. Soundness Analysis

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

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

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

The soundness Theorem (formally stated in Theorem 4.8) claims that if some $h(y)$ passes above dictator test with high probability, then we can decode each $\boldsymbol{w}_i$ ($i \in [k]$) in to a small list and at least two of the list will intersect. The proof of the soundness theorem is by two important lemmas (Lemma 4.6, 4.7). Briefly speaking, the first lemma states that if a halfspace passes the test with good probability,

then two of its critical index sets $L_\tau(\boldsymbol{w}_i), L_\tau(\boldsymbol{w}_j)$ (see Definition 3.1) must intersect; the second Lemma states that every halfspace can be approximated by another halfspace with small critical index.

Lemma 4.6. *(Common Influential Coordinate) For $\tau = \frac{1}{k^6}$, let $h(\boldsymbol{y})$ be a halfspace such that for all $i \neq j \in [k]$, we have $L_\tau(\boldsymbol{w}_i) \cap L_\tau(\boldsymbol{w}_j) = \emptyset$. Then*

$$\Big| \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})] \Big| \leqslant O\Big(\frac{1}{k}\Big) \tag{7}$$

Proof: Fix the following notation,

$$\boldsymbol{s}_i = \mathsf{Truncate}(\boldsymbol{w}_i, L_\tau(\boldsymbol{w}_i)) \qquad \boldsymbol{l}_i = \boldsymbol{w}_i - \boldsymbol{s}_i$$
$$\boldsymbol{y}_i^L = \mathsf{Truncate}(\boldsymbol{y}_i, L_\tau(\boldsymbol{w}_i)) \qquad \boldsymbol{y}^L = \cup_{i=1}^k \boldsymbol{y}_i^L$$

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

$$\boldsymbol{A}^{\{j\}} = \{y_i^{(j)} |\ \forall i \in [k] \text{ such that } j \notin L_\tau(\boldsymbol{w}_i)\}$$

Similarly define the ensemble $\mathcal{B} = \boldsymbol{B}^{\{1\}}, \ldots, \boldsymbol{B}^{\{R\}}$ with the distribution $\tilde{\mathcal{D}}_1^R$. Further let us denote $\boldsymbol{l}^{\{j\}} = (l_1^{(j)}, \ldots, l_k^{(j)})$. Now we shall apply the invariance 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 $j \in L_\tau(\boldsymbol{w}_i)$. Thus, conditioning on $\boldsymbol{y}^L$ amounts to fixing of at most one variable $y_j^{(i)}$ in each $\{y_j^{(i)}\}$. By Lemma 4.4, since $\tilde{\mathcal{D}}_0$ and $\tilde{\mathcal{D}}_1$ has matching moments of degree 4, we know $\boldsymbol{A}^{\{j\}}$ and $\boldsymbol{B}^{\{j\}}$ has matching moments up to degree 3. Also notice that $\max_{j\in[R], i\in[k]} |l_i^{(j)}| \leqslant \tau\|\boldsymbol{l}_i\|_2 \leqslant \tau\|l\|_2$ (as l_i is a τ-regular) and each $y_i^{(j)}$ is setting to be random bit with probability $\frac{1}{k^2}$; by Lemma 3.3, the linear functions $\boldsymbol{l}$ and the ensembles $\mathcal{A}$, $\mathcal{B}$ satisfy the following spread property:

$$\Pr_{\mathcal{A}}\Big[\boldsymbol{l}(\mathcal{A}) \in [\theta-\alpha, \theta+\alpha]\Big] \leqslant c(\alpha)$$
$$\Pr_{\mathcal{B}}\Big[\boldsymbol{l}(\mathcal{B}) \in [\theta-\alpha, \theta+\alpha]\Big] \leqslant c(\alpha)$$

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 theorem 3.10, this implies:

$$\Big| \mathop{\mathbf{E}}_{\mathcal{A}}\Big[\mathrm{sgn}\Big(\langle \boldsymbol{s}, \boldsymbol{y}^L\rangle + \sum_{j\in[R]} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{A}^{\{j\}}\rangle - \tilde{\theta}\Big)\Big|\boldsymbol{y}^L\Big] - \mathop{\mathbf{E}}_{\mathcal{B}}\Big[\mathrm{sgn}\Big(\langle \boldsymbol{s}, \boldsymbol{y}^L\rangle + \sum_{j\in[R]} \langle \boldsymbol{l}^{\{j\}}, \boldsymbol{B}^{\{j\}}\rangle - \tilde{\theta}\Big)\Big|\boldsymbol{y}^L\Big]\Big| \leqslant O\Big(\frac{1}{\alpha^4}\Big) \sum_{i\in[R]} \|\boldsymbol{l}^{\{i\}}\|_1^4 + 2c(\alpha) \tag{8}$$

By definition of 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: $\sum_{j\in[R]} \|\boldsymbol{l}^{\{j\}}\|_1^4 \leqslant k^4 \sum_{i\in[k]} \sum_{j\in[R]} \|l_i^{(j)}\|^4 \leqslant k^4 \sum_{i\in[k]} \Big(\max_{j\in[R]} |l_i^{(j)}|^2\Big)\|\boldsymbol{l}_i\|_2^2 \leqslant k^4\tau^2 \sum_{i\in[k]} \|\boldsymbol{l}_i\|_2^2 \leqslant k^4\tau^2\|\boldsymbol{l}\|_2^2 \leqslant \frac{1}{k^8}$. In the final step, we used the fact that $\tau = \frac{1}{k^6}$ and $\|\boldsymbol{l}\|_2 = 1$ by normalization. Let us fix $\alpha = \frac{1}{k^3}$. The inequality (8) holds for all settings of $\boldsymbol{y}^L$. Averaging over all settings of $\boldsymbol{y}^L$ we get that (8) can be bounded by $O(\frac{1}{k})$. ∎

The set $L_\tau(\boldsymbol{w}_i)$ can be thought of as the set of *influential* coordinates of $\boldsymbol{w}_i$. In this light, the above lemma asserts that the 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 the traditional notion of influence, it is unclear whether the number of coordinates in $L_\tau(\boldsymbol{w}_i)$ is small. The following lemma yields a way to get around this.

Lemma 4.7. *(Bounding the number of influential coordinates) Fix $t = \frac{4}{\tau^2}(3\log(1/\tau) + \log R) + 4k^2\log(1/k)$. Given a halfspace $h(\boldsymbol{y})$ and $\ell \in [k]$ such that $|L_\tau(\boldsymbol{w}_\ell)| > t$, define $\tilde{h}(\boldsymbol{y}) = \mathrm{sgn}(\sum_{i\in[k]} \langle \tilde{\boldsymbol{w}}_i, \boldsymbol{y}_i\rangle - \theta)$ as follows: $\tilde{\boldsymbol{w}}_\ell = \mathsf{Truncate}(\boldsymbol{w}_\ell, S_t(\boldsymbol{w}_\ell))$ and $\tilde{\boldsymbol{w}}_i = \boldsymbol{w}_i$ for all $i \neq \ell$. Then,*

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

Proof: Without loss of generality, we assume $\ell = 1$ and $|w_1^{(1)}| \geqslant |w_1^{(2)}| \ldots \geqslant |w_1^{(R)}|$. In particular, this implies $S_t(\boldsymbol{w}_1) = \{1, \ldots, t\}$. Set $T = 4k^2\log(1/k)$. Define the subset of $S_t(\boldsymbol{w}_1)$ 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, g_i is a geometrically decreasing sequence such that $g_{i+1} \leqslant g_i/3$. Let $H = S - G$. Fix the following notation:

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

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

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

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

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

$$\Big|\sum_{i=2}^k \langle \boldsymbol{w}_i, \boldsymbol{y}_i\rangle + \langle \boldsymbol{w}_1^G, \boldsymbol{y}_1^G\rangle + \langle \boldsymbol{w}_1^H, \boldsymbol{y}_1^H\rangle - \theta\Big| \leqslant |\langle \boldsymbol{w}_1^{>t}, \boldsymbol{y}_1^{>t}\rangle| \tag{9}$$

By Lemma 3.2, we know that

$$|w_1^{(g_T)}|^2 \geqslant \frac{\tau^2}{((\sqrt{1-\tau^2})^{t-T})}\|w_1^{>t}\|_2^2$$
$$\geqslant \frac{\tau^2}{(1-\tau^2)^{\frac{4}{\tau^2}(3\log(1/\tau)+\log R)}}\|w_1^{>t}\|_2^2 \geqslant \frac{R^2}{\tau}|w_1^{>t}|_2^2 \quad (10)$$

Using the fact $R\|w_1^{>t}\|_2^2 \geqslant \|w_1^{>t}\|_1^2$, we can get that $\|w_1^{>t}\|_1 \leqslant \sqrt{\tau}|w_1^{(g_T)}| \leqslant \frac{1}{6}|w_1^{g_T}|$. Combining above inequality with (9) we see that,

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

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 $\theta' \in \mathbb{R}$, induces some arbitrary distribution on $\boldsymbol{y}^G$. However, the $\frac{1}{k^2}$ noise introduced in $\boldsymbol{y}^{\boldsymbol{G}}$ is completely independent. This corresponds to the setting in Lemma 3.7, and hence we can bound the above probability by $\left(1-\frac{1}{2k^2}\right)^T \leqslant \frac{1}{k^2}$. Averaging over all values for θ', the result follows. ∎

Theorem 4.8. *(Soundness) Fix $\tau = \frac{1}{k^6}$ and $t = \frac{4}{\tau^2}(3\log(1/\tau) + \log R) + k^2 log(1/k)$. Let $h(\boldsymbol{x}) = \mathrm{sgn}(\langle \boldsymbol{w}, \boldsymbol{y}\rangle - \theta)$ be a halfspace such that $S_t(\boldsymbol{w}_i) \cap S_t(\boldsymbol{w}_j) = \emptyset$ 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: 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 suffice to show $\Big|\mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant O(\frac{1}{k})$.

Define $K = \{l \mid L_\tau(\boldsymbol{w}_l) \geqslant t\}$. We discuss the following two cases.

1. $K = \emptyset$; i.e., $\forall i \in [k]$, $\boldsymbol{w}_i$, $L_\tau(\boldsymbol{w}_i) \leqslant t$. Then for any i, j, $S_t(\boldsymbol{w}_i) \cap S_t(\boldsymbol{w}_j) = \emptyset$ implies $L_\tau(\boldsymbol{w}_i) \cap L_\tau(\boldsymbol{w}_j) = \emptyset$. By Lemma 4.6, we thus have $\Big|\mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant O(\frac{1}{k})$.

2. $K \neq \emptyset$. Then for all $l \in K$, we set $\tilde{\boldsymbol{w}}_\ell = \mathsf{Truncate}(\boldsymbol{w}_\ell, S_t(\boldsymbol{w}_\ell))$ and replace $\boldsymbol{w}_\ell$ with $\tilde{\boldsymbol{w}}_\ell$ in h to get a new halfspace h'. Since such replacement occur at most k times and by Lemma 4.7 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

$$\Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_0^R}[h(\boldsymbol{y})]\Big|, \Big|\mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h'(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\tilde{\mathcal{D}}_1^R}[h(\boldsymbol{y})]\Big| \leqslant \frac{1}{k} \quad (11)$$

Also notice that for h', for all $\ell \in [k]$, the critical index of $|L_\tau(\tilde{w}_\ell)|$ is less than t. This reduces the problem to Case 1, and we conclude $\Big|\mathbf{E}_{\tilde{\mathcal{D}}_0^R}[h'(\boldsymbol{y})] - \mathbf{E}_{\tilde{\mathcal{D}}_1^R}[h'(\boldsymbol{y})]\Big| \leqslant O(1/k)$. Along with (11) this finishes the proof. ∎

5. Reduction from k-Unique Label Cover

In this section, we describe a reduction from k-Unique Label Cover problem to agnostic learning monomials, thus showing Theorem 1.1 under the Unique Games Conjecture (Conjecture 2.2).

Let $\mathcal{L}(G(V,E), 1, R, \{\pi^{v,e} | v \in V, e \in E\})$ be an instance of k-Unique Label Cover. The reduction will produce a distribution over examples and label pairs: $(\boldsymbol{y}, b)$ where y lies 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 $(y_w^{(1)}, y_w^{(2)}, ...y_w^{(R)})$.

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 $\mathcal{D}_b^R$.
4) Generate $\boldsymbol{y} \in \{0,1\}^{|V|\times R}$ as follows:
   a) For each $v \notin \{v_1, ..v_k\}$, $\boldsymbol{y}_v = \mathbf{0}$.
   b) For each $i \in [k]$, set $\boldsymbol{y}_{v_i} \in \{0,1\}^R$ as follows:

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

5) Output the example-label pair $(\boldsymbol{y}, b)$.

Proof of Theorem 1.1 assuming Unique Games Conjecture: Fix $k = \frac{10}{\epsilon^2}$, $\eta = \frac{\epsilon^3}{100}$ and a positive integer $R > \lceil (2k^{29})^{\frac{1}{\eta^2}} \rceil$ for which Conjecture 2.2 holds.

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

Soundness: Suppose there is some halfspace $h(y) = \sum_{v\in V}\langle \boldsymbol{w}_v, \boldsymbol{y}_v\rangle$ that agrees more than $\frac{1}{2} + \epsilon \geqslant \frac{1}{2} + \frac{1}{\sqrt{k}}$ fraction of the examples. Set $t = k^{12}(3\log(k^6) + \log R) + 4k^2\log(1/k)) = O\big(k^{13}\log(R)\big)$ (same as in Theorem 4.8). Define the labelling $\mathcal{A}$ using the following strategy : for each vertex $v \in V$, if $S_t(\boldsymbol{w}_v)$ is nonempty, randomly pick a label from $S_t(\boldsymbol{w}_v)$ else pick a uniformly random label.

By an averaging argument, for at least $\frac{\epsilon}{2}$ fraction of the edges $e \in E$ generated in step 1 of the reduction, $h(y)$

agrees with more than $\frac{1}{2}+\frac{\epsilon}{2}$ of the corresponding examples. We will refer to these edges as *good*. By Theorem 4.8 for each *good* edge $e \in E$, there exists $i, j \in [k]$, such that $\pi^{v_i,e}\big(S_t(w_{v_i})\big) \cap \pi^{v_j,e}\big(S_t(w_{v_j})\big) \neq \emptyset$. Therefore the edge $e \in E$ is *weakly* satisfied by the labelling $\mathcal{A}$ with probability at least $\frac{1}{t^2}$. Hence, in expectation the labelling $\mathcal{A}$ *weakly* satisfies at least $\frac{\epsilon}{2} \cdot \frac{1}{t^2} = \Omega(\frac{1}{k^{27}\log^2 R}) \geqslant \frac{2k^2}{R^{\eta/4}}$ (by the choice of R and t).

REFERENCES

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

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

[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.*, vol. 54, no. 2, pp. 317–331, 1997.

[4] P. Auer and M. K. Warmuth, "Tracking the best disjunction," *Machine Learning*, vol. 32, no. 2, pp. 127–150, 1998.

[5] S. Ben-David, N. Eiron, and P. M. Long, "On the difficulty of approximately maximizing agreements." *J. Comput. Syst. Sci.*, vol. 66, no. 3, pp. 496–514, 2003.

[6] A. Blum, A. Frieze, R. Kannan, and S. Vempala, "A polynomial-time algorithm for learning noisy linear threshold functions," *Algorithmica*, vol. 22, no. 1-2, pp. 35–52, 1998.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Inf. Process. Lett.*, vol. 24, no. 6, pp. 377–380, 1987.

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

[9] ——, "Maximizing agreements and coagnostic learning," *Theoretical Computer Science*, vol. 350, no. 1, pp. 24–39, 2006.

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

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

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

[13] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola, "Bounded independence fools halfspaces," *ECCC*, vol. TR09-016, 2009.

[14] V. Feldman, "Optimal hardness results for maximizing agreements with monomials," in *IEEE CCC*, 2006, pp. 226–236.

[15] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami, "New results for learning noisy parities and halfspaces," in *IEEE FOCS*, 2006, pp. 563–574.

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

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

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

[19] V. Guruswami and P. Raghavendra, "Hardness of learning halfspaces with noise," in *IEEE FOCS*, 2006, pp. 543–552.

[20] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, no. 1, pp. 78–150, 1992.

[21] K. Hoffgen, K. van Horn, and H. U. Simon, "Robust trainability of single neurons," *J. Comput. Syst. Sci.*, vol. 50, no. 1, pp. 114–125, 1995.

[22] D. S. Johnson and F. P. Preparata, "The densest hemisphere problem." *Theoretical Computer Science*, vol. 6, pp. 93–107, 1978.

[23] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio, "Agnostically learning halfspaces," *SIAM Journal on Computing*, vol. 37, no. 6, pp. 1777–1805, 2008.

[24] M. Kearns, "Efficient noise-tolerant learning from statistical queries," *Journal of the ACM*, vol. 45, no. 6, pp. 983–1006, 1998.

[25] M. Kearns, R. Schapire, and L. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, pp. 115–141, 1994.

[26] M. J. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM J. Comput.*, vol. 22, no. 4, pp. 807–837, 1993.

[27] M. J. Kearns and R. E. Schapire, "Efficient distribution-free learning of probabilistic concepts," *J. Comput. Syst. Sci.*, vol. 48, no. 3, pp. 464–497, 1994.

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

[29] ——, "On the power of unique 2-Prover 1-Round games," in *ACM STOC*, May 19–21 2002, pp. 767–775.

[30] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell, "Optimal inapproximability results for MAX-CUT and other 2-variable CSPs?" *SIAM J. Comput*, vol. 37, no. 1, pp. 319–357, 2007.

[31] S. Khot and R. Saket, "Hardness of minimizing and learning DNF expressions," in *IEEE FOCS*, 2008, pp. 231–240.

[32] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "On efficient agnostic learning of linear combinations of basis functions," in *Proceedings of COLT*, 1995, pp. 369–376.

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

[34] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. A. Servedio, "Testing halfspaces," in *SODA*, 2009, pp. 256–264.

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

[36] E. Mossel, R. O'Donnell, and K. Oleszkiewicz, "Noise stability of functions with low influences: Invariance and optimality," in *IEEE FOCS*, 2005.

[37] R. O'Donnell and R. A. Servedio, "The chow parameters problem," in *ACM STOC*, 2008, pp. 517–526.

[38] R. Rivest, "Learning decision lists," *Machine Learning*, vol. 2, no. 3, pp. 229–246, 1987.

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

[40] R. A. Servedio, "Every linear threshold function has a low-weight approximator," *Comput. Complex.*, vol. 16, no. 2, pp. 180–209, 2007.

[41] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[42] V. Vapnik, *Statistical Learning Theory*, 1998.