Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

On the hardness of learning intersections of two halfspaces

Subhash Khot [a,1], Rishi Saket [b,*,2]

[a] New York University, United States
[b] Georgia Institute of Technology, United States

ARTICLE INFO

Article history:
Received 22 June 2009
Received in revised form 12 October 2009
Accepted 7 June 2010
Available online 11 June 2010

Keywords:
Learning
Hardness
Approximation
Halfspaces

ABSTRACT

We show that unless NP = RP, it is hard to (even) weakly PAC-learn intersection of two halfspaces in $\mathbb{R}^n$ using a hypothesis which is a function of up to ℓ halfspaces (linear threshold functions) for any integer ℓ. Specifically, we show that for every integer ℓ and an arbitrarily small constant $\varepsilon > 0$, unless NP = RP, no polynomial time algorithm can distinguish whether there is an intersection of two halfspaces that correctly classifies a given set of labeled points in $\mathbb{R}^n$, or whether any function of ℓ halfspaces can correctly classify at most $\frac{1}{2} + \varepsilon$ fraction of the points.

1. Introduction

A halfspace in $\mathbb{R}^n$ is given by the set $\{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{r}, \mathbf{x} \rangle \leqslant c\}$ for some non-zero vector $\mathbf{r} \in \mathbb{R}^n$ and a real number c. The problem of learning a halfspace or an intersection of a small number of halfspaces is an extremely well-studied problem in machine learning, with several applications to computer vision [21], artificial intelligence [22] and data mining [24]. In this paper we shall consider the Probably Approximately Correct (PAC) model of learning introduced by Valiant [27]. It is well known that a single halfspace can be PAC-learned efficiently by sampling a polynomial number of data points and finding a separating hyperplane via linear programming [9]. Blum, Frieze, Kannan, and Vempala [6] showed how to learn a single halfspace even in presence of random classification noise, whereas Kalai, Klivans, Mansour and Servedio [15] gave polynomial time algorithm for learning a single halfspace in presence of adversarial noise under certain distributional assumptions.

For learning intersection of halfspaces, algorithms are known for various special cases. When the data points are drawn from the uniform distribution over the unit ball, Blum and Kannan [7] and Vempala [30] gave algorithms to PAC-learn intersection of a constant number of halfspaces. For the uniform distribution over the boolean hypercube, Klivans, O'Donnell and Servedio [18] obtained an algorithm for learning any function of a constant number of halfspaces. Arriaga and Vempala [4] and Klivans and Servedio [19] gave algorithms for learning intersection of halfspaces when no data point is too close to any separating hyperplane (i.e. the problem instance has a good *margin*). However the general problem of learning intersection of halfspaces remains an open problem.

In this paper we study the hardness of learning intersection of *two* halfspaces with a hypothesis that is a function of a constant (but arbitrarily large) number of halfspaces. We prove that this problem is hard to even weakly PAC-learn unless NP = RP. In particular our result holds for hypothesis class of an intersection of a constant number of halfspaces. We state our result formally after reviewing previous work on the hardness side.

[*] Corresponding author.
E-mail addresses: khot@cs.nyu.edu (S. Khot), saket@cc.gatech.edu (R. Saket).

[1] Supported by NSF CAREER award CCF-0643626, Microsoft New Faculty Fellowship, and Sloan Foundation fellowship.

[2] Supported by the same awards as of Subhash Khot.

0022-0000/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2010.06.010

1.1. Previous work

For the problem of proper learning a single halfspace with adversarial noise, Feldman, Gopalan, Khot and Ponnuswami [10] and Guruswami and Raghavendra [13] independently proved that the problem is hard to even weakly PAC-learn. More specifically, they proved:

Theorem 1. *Let $\delta, \varepsilon > 0$ be arbitrarily small constants. Then, given a set of labeled points in $\mathbb{R}^n$ with a guarantee that there is a halfspace that classifies $1 - \delta$ fraction of points correctly, there is no polynomial time algorithm to find a halfspace that classifies $\frac{1}{2} + \varepsilon$ fraction of points correctly, unless* P = NP.

The reduction of Guruswami and Raghavendra also works for points on the hypercube $\{-1, 1\}^n$.

In recent work [11], Feldman, Guruswami, Raghavendra and Wu proved that it is hard (under the same complexity assumptions) to learn even an AND function under adversarial δ-noise by a halfspace to an accuracy of $\frac{1}{2} + \varepsilon$, which is a generalization of Theorem 1 as the class of AND function is contained in halfspaces. Theorem 1 is really a hardness of approximation result for an optimization problem, where the goal is to find a halfspace that maximizes the number of correctly classified points. By considering a distribution that is uniform on all the data points, the theorem implies hardness of weakly PAC-learning a halfspace when the hypothesis is also required to be a halfspace (known as *proper learning*). Note that the imperfect completeness is necessary in the above theorem, since via linear programming, one can always efficiently find a halfspace that correctly classifies *all* the points, if one exists. The theorem is optimal, since one can easily classify $\frac{1}{2}$ fraction of the data points correctly, by taking an arbitrary halfspace or its complement as a hypothesis. From the learning theory perspective, such an optimal hardness result is especially satisfying, since if one could efficiently find $(\frac{1}{2} + \varepsilon)$-consistent hypothesis (i.e. weakly PAC-learn), then one can use boosting techniques [25] to efficiently find a $(1 - \varepsilon)$-consistent hypothesis (i.e. PAC-learn).[3] We note that for boosting techniques to be applicable even a $(\frac{1}{2} + \frac{1}{\text{poly}(n)})$-consistent weak learning algorithm would suffice. The hardness results in [10] and [13] prove only a $\frac{1}{2} + \varepsilon$ inapproximability, where $\varepsilon > 0$ is *not* an inverse polynomial, and therefore do not rule out the existence of such weak learning algorithms. Theorem 1 does, however, provide evidence against the existence of efficient proper learning algorithms for learning halfspaces under adversarial noise.

However, a result such as Theorem 1 was not known for learning intersections of (two) halfspaces. Blum and Rivest [8] showed that it is NP-hard to learn an intersection of two halfspaces with an intersection of two halfspaces, and Alekhnovich, Braverman, Feldman, Klivans and Pitassi [1] proved a similar result even when the hypothesis is an intersection of ℓ halfspaces for any constant ℓ. Both the results are only NP-hardness results and do not prove APX-hardness for the underlying optimization problem.

In a different line of work, under cryptographic assumptions, Klivans and Sherstov [20] showed that there is no polynomial time algorithm to PAC-learn intersection of n^{ϵ} halfspaces, and the result holds without any restriction on the hypothesis class.

1.2. Our results

The following theorem is the main result in this paper.

Theorem 2. *Let ℓ be any fixed integer and $\varepsilon > 0$ be an arbitrarily small constant. Then, given a set of labeled points in $\mathbb{R}^n$ with a guarantee that there is an intersection of two halfspaces that classifies all the points correctly, there is no polynomial time algorithm to find a function f of up to ℓ halfspaces that classifies $\frac{1}{2} + \varepsilon$ fraction of points correctly, unless* NP = RP.

We state the above result in terms of functions of ℓ halfspaces. This encompasses hypotheses such as intersections of ℓ halfspaces. We note that the term *linear threshold function* is commonly used in literature for a *halfspace*. In this paper, though, we shall adhere to the more succinct *halfspace*. Note that the result holds with perfect completeness, i.e. the problem is hard even when an intersection of two halfspaces is guaranteed to classify *every* point correctly. The result is optimal since an arbitrary halfspace or its complement has a success rate of $\frac{1}{2}$ on any given data set. It provides evidence that the approach of weak learning intersection of two halfspaces with a function of a constant number of halfspaces followed by boosting may not work.

1.3. Informal description of the reduction

The reduction starts with an instance of (a variant of) the LabelCover problem on n vertices and label set $[k]$. It produces an instance in nk-dimensional space with a block of k dimensions (coordinates) for each vertex of the LabelCover instance. In the first step we create a set of points for each vertex which are all possible combinations of $\{-1, 1\}$ values in the k coordinates corresponding to the vertex and 0 everywhere else. Essentially, in this block of k coordinates this set of points is the hypercube $\{-1, 1\}^k$ and 0 in all other coordinates. These points simulate a junta test in the following manner. Fix a

[3] After applying the boosting algorithm, the hypothesis is a majority over a set of hypotheses used for the weak PAC-learning.

vertex v of the LabelCover instance, and consider only the corresponding block of k coordinates and the points. Consider any one of the k coordinates, let us say the coordinate x_i. Clearly, all the points corresponding to v lie on one of the two hyperplanes $x_i = 1$ and $x_i = -1$. Therefore, every coordinate (junta) yields a pair of 'good' hyperplanes. On the other hand, suppose that a hyperplane $\langle \mathbf{r}, \mathbf{x} \rangle = c$ passes close to many points corresponding to a particular vertex v. This implies that a random $\{-1, 1\}$ linear combination of the values of $\mathbf{r}$ in the block of k coordinates corresponding to v lies in a particular small range with significant probability. Using standard anti-concentration arguments it can be shown that $\mathbf{r}$ must have the property that whatever mass it has in the k coordinates corresponding to v is concentrated in a few coordinates. This gives a natural way to choose a label for v from among those few coordinates. However, for this test to work, one has to ensure that $\mathbf{r}$ has *some* non-negligible mass to begin with in the k coordinates corresponding to v.

To ensure this, in the second step we replace each point created in the first step with two disjoint small spheres of random points, where points in one of the spheres are labeled '+' and those in the other are labeled '−'. We do this in such a way that all new points corresponding to each of the original points can be grouped together into disjoint *pairs*, each consisting of a '+' and a '−' labeled point, and the displacement between the two points in the pair is the same for all pairs corresponding to the same original point. We also ensure that for any of the k coordinates corresponding to v, say the coordinate x_i, all the points labeled '+' lie inside the intersection of the two halfspaces $x_i \leqslant 1$ and $x_i \geqslant -1$ and all the '−' labeled points lie outside it. On the other hand, if the given hyperplane $\langle \mathbf{r}, \mathbf{x} \rangle = c$ separates a particular pair, then the vector $\mathbf{r}$ must have a projection along the segment joining the two points in that pair. If the spheres are sufficiently 'dense', one can show that if the hyperplane separates significant fraction of such pairs then the projection of $\mathbf{r}$ on that segment must be sufficiently large, and therefore $\mathbf{r}$ has sufficient mass in the corresponding block of k coordinates. Moreover, most of this mass will be concentrated in a small number of coordinates which can be used to assign a label to v. We use the fact that the class of halfspaces has polynomially bounded VC dimension and therefore with high probability a polynomially large set of random points on a sphere is an ϵ-sample for all halfspaces.

Using the above steps along with an averaging argument, one can ensure that a hyperplane $\langle \mathbf{r}, \mathbf{x} \rangle = c$ that separates a significant fraction of the *pairs* of '+' and '−' labeled points yields a labeling to a significant fraction of the vertices of the LabelCover instance. However, we need to ensure that this labeling in fact satisfies a significant fraction of the edges (constraints) of the LabelCover instance. In order to enforce consistency between the labels of different vertices of the instance, the third step involves a folding procedure over a subspace defined by the constraints of the instance. A similar folding technique was used in [12].

Our construction is such that the existence of a 'good' function of ℓ halfspaces gives us one single hyperplane which can be used, as described above, to extract a good labeling to the instance.

Remark. The gap instance that we construct is such that in the YES case, the set of points is classified correctly by the intersection of two parallel halfspaces of the form $\langle \mathbf{r}, \mathbf{x} \rangle \geqslant -c$ and $\langle \mathbf{r}, \mathbf{x} \rangle \leqslant c$, for some $c > 0$. This implies that the points are correctly classified by the degree 2 polynomial given by $(\langle \mathbf{r}, \mathbf{x} \rangle)^2 \leqslant c^2$. So, using linear programming a degree 2 polynomial can be efficiently found that classifies the points in our instance. However, the polynomial obtained may not be factorizable into two parallel hyperplanes.

2. Preliminaries

We start by formally defining the problem.

Definition 1. An instance of Classifier-Halfspace$_\ell$ is a set of points in $\mathbb{R}^n$ each labeled either '+' or '−' and the goal is to find a function of at most ℓ halfspaces (halfspaces) which correctly classifies the maximum number of points, where a '+' point is classified correctly if it lies inside the intersection and a '−' point is classified correctly if it lies outside of it. In particular, the function can be an intersection of ℓ halfspaces.

We show that the problem Classifier-Halfspace$_\ell$ is hard by giving a gap-preserving reduction from a variant of the LabelCover problem. For the purposes of our reduction we require the LabelCover instance to satisfy a certain 'smoothness' property. Similar 'smooth' versions of LabelCover have been used in earlier hardness reductions [16,14,17]. In addition to smoothness we also require that sufficiently large induced subgraphs of the instance have a large number of edges. We define the following version of the LabelCover problem that captures both these additional properties.

Definition 2. An instance $\mathcal{L}$ of Smooth-Label-Cover(t, μ, ν, k, m) consists of a (multi)graph $G(V, E)$ and mappings $\{\pi_{u,e}\}_{e \in E, u \in e}$ where $\pi_{u,e} : [k] \mapsto [m]$. A labeling $\sigma : V \mapsto [k]$ is said to satisfy an edge e between u and v if $\pi_{u,e}(\sigma(u)) = \pi_{v,e}(\sigma(v))$. The instance satisfies:

- (Smoothness:) For any vertex $u \in V$ and any set $S \subseteq [k]$ of size at most t,
$$\Pr_{e: u \in e}\bigl[\exists i, j \in S, i \neq j, \text{ s.t. } \pi_{u,e}(i) = \pi_{u,e}(j)\bigr] \leqslant \mu,$$
where the probability is taken over a random edge incident on u.
- For any $V' \subseteq V$ such that $|V'| = \xi |V|$, the induced subgraph on V' has at least $(\xi^2/2)|E|$ edges, for any $\xi \geqslant \nu$.

The following theorem is proved using the PCP Theorem [3,2] combined with Raz's Parallel Repetition Theorem [23]. We give a proof of the theorem in Section 6 based on the smooth version of LabelCover constructed in [17].

Theorem 3. *For any constant t and arbitrarily small constants $\mu, \nu, \eta > 0$, there exist constants k and m such that given an instance $\mathcal{L}$ of* Smooth-Label-Cover(t, μ, ν, k, m) *it is* NP*-hard to distinguish between the following two cases*:

- YES Case/Completeness: *There is a labeling to the vertices of $\mathcal{L}$ which satisfies all the edges.*
- NO Case/Soundness: *No labeling to the vertices of $\mathcal{L}$ satisfies more than η fraction of the edges.*

We give a gap-preserving reduction from Smooth-Label-Cover to Classifier-Halfspace$_\ell$ stated in the following theorem, which along with Theorem 3 implies Theorem 2.

Theorem 4. *For any constant $\varepsilon > 0$ and integer $\ell > 0$, there is a randomized polynomial time reduction from an instance $\mathcal{L}$ of* Smooth-Label-Cover(t, μ, ν, k, m) *to an instance $\mathcal{I}$ of* Classifier-Halfspace$_\ell$ *for appropriately chosen parameters t, μ, ν and soundness η, such that*

- YES Case/Completeness: *If $\mathcal{L}$ is a* YES *instance, then there is an intersection of two halfspaces which correctly classifies all the points in instance $\mathcal{I}$.*
- NO Case/Soundness: *If $\mathcal{L}$ is a* NO *instance, then with probability at least $\frac{9}{10}$, there is no function of up to ℓ halfspaces that correctly classifies more than $\frac{1}{2} + \varepsilon$ fraction of points in instance $\mathcal{I}$.*

3. Reduction

The reduction proceeds in three steps. In the first step we construct an initial set of unlabeled points, from the instance of Smooth-Label-Cover. In the second step we replace each initial point with two small spheres of points, with points in one sphere labeled '+' and points in the other labeled '−'. The third step consists of reducing the problem into a lower dimensional space via a *folding* over the subspace induced by the consistency constraints of the LabelCover instance. In the end we obtain a set of points each labeled either '+' or '−' as an instance of Classifier-Halfspace$_\ell$ for a given constant ℓ.

3.1. Step 1: Initial unlabeled point set

We start with an instance $\mathcal{L}$ of Smooth-Label-Cover(t, μ, ν, k, m), where we will fix t, μ and ν later in the analysis of the reduction in Section 4.2. Let $|V| = n$. First we define the space in which the points lie. For every vertex $v \in V$, we have a set of k coordinates labeled by $M(v) := \{v(i)\}_{i=1}^{k}$. The complete set of coordinates is the union of these sets over all vertices, say $\mathcal{M} := \bigcup_{v \in V} M(v)$. Therefore, the points we construct lie in nk-dimensional real space $\mathbb{R}^{\mathcal{M}}$. The construction of the points is as follows.

1. For every vertex v, define

$$s(v) := \{\mathbf{x} \in \mathbb{R}^{\mathcal{M}} \mid \forall i \in [k],\ \mathbf{x}(v(i)) \in \{-1, 1\} \text{ and } \mathbf{x}(u(i)) = 0, \forall u \neq v\}.$$

   Thus, $s(v)$ is the set of all vectors that are $\{-1, 1\}$ combinations on the coordinates $M(v)$ and 0 on all other coordinates. Note that $|s(v)| = 2^k$ for all $v \in V$.
2. Let $S = \bigcup_{v \in V} s(v)$. Clearly $|S| = n \cdot 2^k$, since the sets $s(v)$ are disjoint for all $v \in V$.

We output S, as the initial set of points at the end of Step 1. As explained in Section 1.3, one would like to say that if any hyperplane say $\langle \mathbf{r}, \mathbf{x} \rangle = a$ passes close to a significant fraction of the points in S, then $\mathbf{r}$ should be close to a 'junta' i.e. most of the mass of the vector $\mathbf{r}$ should be concentrated in a few coordinates of $M(v)$ for a significant fraction of vertices $v \in V$. However, it is possible that $\mathbf{r}$ has zero mass in the coordinates $M(v)$ for almost all $v \in V$ and yet the hyperplane $\langle \mathbf{r}, \mathbf{x} \rangle = 0$ passes through most of the points in S. To overcome this problem, we replace each point in S with two spheres of points as described in Step 2.

3.2. Step 2: Constructing spheres of labeled points

We start with the set of points $S = \bigcup_{v \in V} s(v)$ constructed in Step 1. For every point in S, we create two 'spheres' of points separated by a small distance. This step is randomized as it requires sampling a suitable number of points from a unit sphere. The following lemma is proved in Section 5.

Lemma 5. *Let $\varepsilon' > 0$ be any arbitrarily small constant and k be any constant positive integer. Given these values let n be a sufficiently large integer. Let R be a set of $(nk)^2$ unit vectors chosen uniformly at random in nk-dimensional real space. Then with probability at least $1 - 1/n$, the set R satisfies the following property: for any subset $T \subseteq R$ such that $|T| = \varepsilon'|R|$ and for any unit vector $\mathbf{r}$, there exist $\mathbf{z}', \mathbf{z}'' \in T$, such that $|\langle \mathbf{r}, \mathbf{z}' \rangle - \langle \mathbf{r}, \mathbf{z}'' \rangle| \geqslant \varepsilon'/(100\sqrt{nk})$.*

Now we describe our construction in Step 2.

1. Set parameters[4] $\delta = 1/((nk)^{10})$ and $\gamma = 1/(100\sqrt{n})$.
2. Let R be the set of $(nk)^2$ unit vectors in $\mathbb{R}^{\mathcal{M}}$ as in Lemma 5 (ε' will be chosen later and is related to the soundness parameter of the reduction).
3. Let $\mathbf{x}$ be a point in S. Construct two sets $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ as follows,

   $$\alpha(\mathbf{x}) := \{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z} \mid \mathbf{z} \in R\}$$

   and

   $$\beta(\mathbf{x}) := \{(1+\delta)\mathbf{x} + \delta\gamma\mathbf{z} \mid \mathbf{z} \in R\}.$$
4. For every vertex $v \in V$, let $A(v) := \bigcup_{\mathbf{x} \in S(v)} \alpha(\mathbf{x})$ and $B(v) := \bigcup_{\mathbf{x} \in S(v)} \beta(\mathbf{x})$.
5. Output the sets $A := \bigcup_{v \in V} A(v)$ and $B := \bigcup_{v \in V} B(v)$.

The points created have the property that any hyperplane that separates the sets $\alpha(\mathbf{x})$ from $\beta(\mathbf{x})$ for a significant fraction of points $\mathbf{x} \in S$, must essentially be a 'junta' in the coordinates $M(v)$ for a significant fraction of vertices $v \in V$. This property will be formally stated and used in the soundness analysis to decode a labeling for the instance $\mathcal{L}$. In conjunction with this property, one needs to enforce the consistency constraints of the instance $\mathcal{L}$. We achieve this in the third step of the reduction, by folding over a subspace defined by these constraints.

3.3. Step 3: Folding and final labeled point set

For the sake of convenience, let $<$ be any arbitrary total order on V. Let e be an edge between u and v in G, with $u < v$. Let $\mathbf{h}_e^j \in \mathbb{R}^{\mathcal{M}}$, for $j \in [m]$, be defined in the following manner: set $\mathbf{h}_e^j(u(i)) = 1$ for all $i \in \pi_{u,e}^{-1}(j)$, set $\mathbf{h}_e^j(v(i)) = -1$ for all $i \in \pi_{v,e}^{-1}(j)$ and set all the other coordinates to 0. Note that for any vector $\mathbf{r} \in \mathbb{R}^{\mathcal{M}}$,

$$\forall u, v \in e,\ e \in E,\ \forall j \in [m],$$
$$\langle \mathbf{r}, \mathbf{h}_e^j \rangle = 0 \iff \sum_{i \in \pi_{u,e}^{-1}(j)} \mathbf{r}(u(i)) = \sum_{i \in \pi_{v,e}^{-1}(j)} \mathbf{r}(v(i)). \tag{1}$$

The folding is done as follows.

1. Let $T = \bigcup_{e \in E, j \in [m]} \{\mathbf{h}_e^j\}$. Let $H \subset \mathbb{R}^{\mathcal{M}}$, where $H = \operatorname{span}(T)$, and F be the subspace of $\mathbb{R}^{\mathcal{M}}$ orthogonal to H such that $\mathbb{R}^{\mathcal{M}} = F \oplus H$ and $F \perp H$.
2. Let $\{\lambda_i\}_{i=1}^{nk}$ be an orthonormal basis for $\mathbb{R}^{\mathcal{M}}$ such that $F = \operatorname{span}(\{\lambda_i\}_{i=1}^{g})$ for some $g \leqslant nk$.
3. Write down all the points in the sets A and B in the basis $\{\lambda_i\}_{i=1}^{nk}$ and only consider the coordinates corresponding to the basis $\{\lambda_i\}_{i=1}^{g}$ of the g-dimensional space F. Ignoring the rest of the coordinates, obtain sets A' and B', which are essentially projections (with multiplicities) of sets A and B respectively onto the subspace F.

We label all the points in A' as '+', and all the points in B' as '−' and output these points as an instance of Classifier-Halfspace$_\ell$.

Intuitively, the idea behind the above folding procedure is as follows. Let $\mathbf{r} \in \mathbb{R}^{\mathcal{M}}$ be a vector which lies entirely in the subspace F. Suppose u and v are two vertices connected by an edge e in the instance $\mathcal{L}$. Moreover, suppose that the mass of $\mathbf{r}$ in the coordinates $M(u)$ is concentrated in coordinates of a small subset of labels for u, say $T_u \subseteq [k]$; and similarly the mass in the coordinates $M(v)$ is concentrated in coordinates of a small subset of labels $T_v \subseteq [k]$. Then, Eq. (1) suggests that $\pi_{u,e}(T_u)$ and $\pi_{v,e}(T_v)$ might have a non-empty intersection, and choosing labels uniformly at random from T_u for u and from T_v for v would satisfy the edge e with significant probability. The process of folding guarantees that the entire set of points of the instance of Classifier-Halfspace$_\ell$ lies in the subspace F.

For the above to work, however, we need to make use of certain properties of the instance $\mathcal{L}$ in addition to ensuring that there is non-negligible mass in the coordinates $M(u)$ and $M(v)$ to begin with.

[4] The parameter δ here is essentially the *margin* in the YES case (up to a polynomial factor). Our reduction goes through even if δ is taken to be $O(\frac{1}{(nk)^2})$ as can be seen in the analysis in Section 4.2.

4. Analysis

4.1. YES case

If the instance $\mathcal{L}$ of SMOOTH-LABEL-COVER(t, μ, ν, k, m) is a YES instance, then there is a labeling, say σ to the vertices in V that satisfies all the edges in E. We need to exhibit two halfspaces such that the points in A' lie inside their intersection while the points in B' lie outside their intersection, where A' and B' are the sets obtained through the reduction given above.

Let us consider the vector $\mathbf{r}^* \in \mathbb{R}^{\mathcal{M}}$, where $\mathbf{r}^*(v(\sigma(v))) = 1/\sqrt{n}$ for all $v \in V$, and all other coordinates are set to 0. So, $\mathbf{r}^*$ has exactly n non-zero coordinates, each set to $1/\sqrt{n}$. Clearly $\|\mathbf{r}^*\| = 1$. We prove the following lemma.

Lemma 6. *For every point $\mathbf{y} \in A$, $|\langle \mathbf{r}^*, \mathbf{y}\rangle| < 1/\sqrt{n}$, and for every point $\mathbf{w} \in B$, $|\langle \mathbf{r}^*, \mathbf{w}\rangle| > 1/\sqrt{n}$.*

Proof. Since $\mathbf{r}^*$ is a unit vector, for any unit vector $\mathbf{z}$, we have

$$\bigl|\langle \mathbf{r}^*, \gamma\delta\mathbf{z}\rangle\bigr| \leqslant \gamma\delta = \delta\left(\frac{1}{100\sqrt{n}}\right). \tag{2}$$

Consider a point $\mathbf{y} \in A$, such that $\mathbf{y} = (1-\delta)\mathbf{x} + \gamma\delta\mathbf{z}$, where $\mathbf{x}$ is a $\{-1, 1\}$ vector in the coordinates $M(v)$ for some $v \in V$, and 0 on all other coordinates, and $\mathbf{z}$ is a unit vector. Now since $\mathbf{r}^*$ has a single non-zero coordinate set to $1/\sqrt{n}$ in each set $M(v)$, clearly $|\langle \mathbf{r}^*, \mathbf{x}\rangle| = 1/\sqrt{n}$ and therefore $|\langle \mathbf{r}^*, (1-\delta)\mathbf{x}\rangle| = (1-\delta)(1/\sqrt{n})$. Combining with (2), we get $|\langle \mathbf{r}^*, ((1-\delta)\mathbf{x} + \gamma\delta\mathbf{z})\rangle| = |\langle \mathbf{r}^*, \mathbf{y}\rangle| < 1/\sqrt{n}$. Similarly, for any $\mathbf{w} \in B$, we obtain $|\langle \mathbf{r}^*, \mathbf{w}\rangle| > 1/\sqrt{n}$. □

Consider any edge $e \in E$ between two vertices u and v in V. Since $\mathbf{r}^*$ is $1/\sqrt{n}$ in exactly one coordinate $u(\sigma(u))$ in $M(u)$ and 0 on all others, for any $j \in [m]$, we have that the quantity $\sum_{i\in\pi_{u,e}^{-1}(j)} \mathbf{r}^*(u(i))$ is $1/\sqrt{n}$ iff $\sigma(u) \in \pi_{u,e}^{-1}(j)$ and 0 otherwise. And similarly, $\sum_{i\in\pi_{v,e}^{-1}(j)} \mathbf{r}^*(v(i))$ is $1/\sqrt{n}$ iff $\sigma(v) \in \pi_{v,e}^{-1}(j)$ and 0 otherwise. Since σ is a satisfying assignment, $\exists j' \in [m]$ such that $\sigma(u) \in \pi_{u,e}^{-1}(j')$ and $\sigma(v) \in \pi_{v,e}^{-1}(j')$. Therefore we have

$$\sum_{i\in\pi_{u,e}^{-1}(j)} \mathbf{r}^*\bigl(u(i)\bigr) = \sum_{i\in\pi_{v,e}^{-1}(j)} \mathbf{r}^*\bigl(v(i)\bigr), \quad \forall u, v \in e,\ e \in E,\ \forall j \in [m]. \tag{3}$$

Combining the above with (1) we obtain

$$\bigl\langle \mathbf{r}^*, \mathbf{h}_e^j\bigr\rangle = 0, \quad \forall e \in E,\ \forall j \in [m]. \tag{4}$$

Since H was defined to be the span of $\{\mathbf{h}_e^j\}_{e\in E, j\in[m]}$, Eq. (4) implies that $\mathbf{r}^* \perp H$. Let $\bar{\mathbf{r}}^*$ be the projection of $\mathbf{r}^*$ onto F, where $F \perp H$ and $\mathbb{R}^{\mathcal{M}} = F \oplus H$. For any $\mathbf{y} \in \mathbb{R}^{\mathcal{M}}$, let $\bar{\mathbf{y}}$ be the projection of $\mathbf{y}$ onto F. Then, since $\mathbf{r}^*$ lies entirely in F we have $\langle \mathbf{r}^*, \mathbf{y}\rangle = \langle \bar{\mathbf{r}}^*, \bar{\mathbf{y}}\rangle$. Combined with Lemma 6 this implies that for every point $\bar{\mathbf{y}} \in A'$, $|\langle \bar{\mathbf{r}}^*, \bar{\mathbf{y}}\rangle| < 1/\sqrt{n}$ and for every point $\bar{\mathbf{w}} \in B'$, $|\langle \bar{\mathbf{r}}^*, \bar{\mathbf{w}}\rangle| > 1/\sqrt{n}$. Therefore the intersection of the two halfspaces in F, namely $\{\mathbf{y} \mid \langle \bar{\mathbf{r}}^*, \mathbf{y}\rangle \leqslant 1/\sqrt{n}\}$ and $\{\mathbf{y} \mid \langle \bar{\mathbf{r}}^*, \mathbf{y}\rangle \geqslant -1/\sqrt{n}\}$, classifies all the points in A' and B' correctly. Note that the intersection of halfspaces that we obtain is the region between two parallel hyperplanes.

4.2. NO case

In this case, we assume that the instance $\mathcal{L}$ of SMOOTH-LABEL-COVER(t, μ, ν, k, m) has soundness η. The parameters t, μ, ν and η will be fixed later in this section, and given these values the parameters k and m will be chosen appropriately as mentioned in Theorem 3. For a contradiction, we assume that we have a boolean function f on points in F, which depends on the outputs of ℓ halfspaces in F, and that f classifies $\frac{1}{2} + \varepsilon$ fraction of the points in A' and B' correctly, where A' is the set of '+' points and B' is the set of '−' points. Let the halfspaces on which f depends be given by the equations

$$\langle \bar{\mathbf{r}}_i, \bar{\mathbf{y}}\rangle \leqslant c_i \quad \text{for } i = 1, \ldots, \ell,$$

where $\bar{\mathbf{r}}_i \in F$ and $\|\bar{\mathbf{r}}_i\| = 1$ for all $i = 1, \ldots, \ell$. The output of the function f can be thought of as that of a ℓ-ary boolean function $\tilde{f} : \{0,1\}^\ell \mapsto \{0,1\}$ which takes as input in the ith coordinate the boolean value given by the ith halfspace.

Let $\mathbf{r}_i$ be the vector in $\mathbb{R}^{\mathcal{M}}$ obtained from $\bar{\mathbf{r}}_i \in F$, by adding zeros on the coordinates corresponding to the basis of H, and rewriting it in the coordinates $\mathcal{M} = \bigcup_{v\in V} M(v)$. Clearly $\|\mathbf{r}_i\| = 1$ for $i = 1, \ldots, \ell$. Let f' be the boolean function in $\mathbb{R}^{\mathcal{M}}$ given by $\tilde{f}$ applied on the outputs of the halfspaces $\{\langle \mathbf{r}_i, \mathbf{y}\rangle \leqslant c_i\}$ for $i = 1, \ldots, \ell$, where the ith boolean coordinate in the input of $\tilde{f}$ is given by the halfspace $\{\langle \mathbf{r}_i, \mathbf{y}\rangle \leqslant c_i\}$ in $\mathbb{R}^{\mathcal{M}}$. Note that f' is exactly the function f applied on points in $\mathbb{R}^{\mathcal{M}}$ after projection onto F. We have the following simple lemma.

Lemma 7. *The function f' of the halfspaces $\{\langle \mathbf{r}_i, \mathbf{y}\rangle \leqslant c_i\}$ for $i = 1, \ldots, \ell$ classifies $\frac{1}{2} + \varepsilon$ fraction of the points in $A \cup B$ correctly.*

Proof. We observe that since $\bar{\mathbf{r}}_i \in F$, if the projection of a point $\mathbf{y} \in \mathbb{R}^{\mathcal{M}}$ onto the subspace F is the point $\bar{\mathbf{y}} \in F$, then $\langle \mathbf{r}_i, \mathbf{y} \rangle = \langle \bar{\mathbf{r}}_i, \bar{\mathbf{y}} \rangle$. Now, since A' and B' are (multi)sets of points in F and are projections of the sets A and B respectively of points in $\mathbb{R}^{\mathcal{M}}$, the lemma follows. □

For the rest of the analysis, we will consider only the sets of points A and B and the halfspaces $\{\langle \mathbf{r}_i, \mathbf{y} \rangle \leqslant c_i\}$ for $i = 1, \dots, \ell$ in $\mathbb{R}^{\mathcal{M}}$. For every vertex v, and every $\mathbf{x} \in s(v)$, there are $|R|$ pairs of points given by the sets $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$. In total there are $|V|2^k|R|$ such pairs, which partition the set $A \cup B$, where each pair has one point from A and one point from B. We say that a pair $\{\mathbf{y}_1, \mathbf{y}_2\}$ where $\mathbf{y}_1 \in A$ and $\mathbf{y}_2 \in B$ is correctly classified by f' if both $\mathbf{y}_1$ and $\mathbf{y}_2$ are correctly classified by f'. Since the function f' of halfspaces $\{\langle \mathbf{r}_i, \mathbf{y} \rangle \leqslant c_i\}$ for $i = 1, \dots, \ell$ correctly classifies $\frac{1}{2} + \varepsilon$ fraction of the points in $A \cup B$, it follows that it correctly classifies 2ε fraction of the pairs $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$. For a pair to be classified correctly by f', it must be separated by at least one of the ℓ halfspaces on which f' depends. Thus, there must be at least one out of the ℓ halfspaces that separates $2\varepsilon/\ell$ fraction of the pairs. Without loss of generality, we can assume that the halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$ separates $2\varepsilon/\ell$ fraction of the pairs. The rest of the analysis uses $\mathbf{r}_1$ to deduce a labeling to the vertices of $\mathcal{L}$ that satisfies a significant fraction of the edges. We first the following simple lemma which is based on standard averaging arguments.

Lemma 8. *Let $\varepsilon' := \varepsilon/(2\ell)$. Define a vertex $v \in V$ to be 'good' if for ε' fraction of vectors $\mathbf{x} \in s(v)$, for $2\varepsilon'$ fraction of vectors $\mathbf{z} \in R$, the pair $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$ is separated by the halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$. Then the number of 'good' vertices is at least $\varepsilon'|V|$.*

Proof. The observation in the previous paragraph can be restated as follows: for $4\varepsilon'$ fraction of the triples $(v, \mathbf{x}, \mathbf{z}) \in (\bigcup_{v' \in V} \bigcup_{\mathbf{x}' \in s(v')} \bigcup_{\mathbf{z}' \in R} (v', \mathbf{x}', \mathbf{z}'))$ the pair $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$ is separated by the halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$. An initial averaging yields that for $2\varepsilon'$ fraction of tuples $(v, \mathbf{x}) \in (\bigcup_{v' \in V} \bigcup_{\mathbf{x}' \in s(v')} (v', \mathbf{x}'))$ for $2\varepsilon'$ fraction of vectors $\mathbf{z} \in R$ the pair $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$ is separated. A further averaging yields that for ε' fraction of the vertices $v \in V$, for ε' fraction of vectors $\mathbf{x} \in s(v)$, for $2\varepsilon'$ fraction of vectors $\mathbf{z} \in R$, the pair $\{(1-\delta)\mathbf{x} + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x} + \delta\gamma\mathbf{z}\}$ is separated by the halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$. Thus the set of 'good' vertices is of size at least $\varepsilon'|V|$. □

In the following lemma we will show that the vector $\mathbf{r}_1$ must have non-negligible mass in the coordinates $M(u)$, for any 'good' vertex u.

Lemma 9. *Let u be a 'good' vertex. Then the vector $\mathbf{r}_1$ satisfies the following property*:

$$\sum_{i \in k} \left| \mathbf{r}_1\big(u(i)\big) \right| \geqslant \frac{\varepsilon'}{2 \cdot 10^4 n\sqrt{k}}. \tag{5}$$

Proof. We know from above that for a 'good' vertex u, there is a vector $\mathbf{x}' \in s(u)$ such that for $2\varepsilon'$ fraction of $\mathbf{z} \in R$ the pair $\{(1-\delta)\mathbf{x}' + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x}' + \delta\gamma\mathbf{z}\}$ is separated by $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$. Let us this fix one such $\mathbf{x}' \in s(u)$. Let us say that a pair is separated 'correctly' by the halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$ if the '+' point is inside the halfspace and the '−' point is outside, and otherwise we say that the pair is separated 'incorrectly'. Based on the above, for our choice of $\mathbf{x}' \in s(u)$ we have the following two cases.

Case 1. The halfspace $\{\langle \mathbf{r}_1, \mathbf{y} \rangle \leqslant c_1\}$ separates 'correctly' the pair $\{(1-\delta)\mathbf{x}' + \delta\gamma\mathbf{z}, (1+\delta)\mathbf{x}' + \delta\gamma\mathbf{z}\}$ for ε' fraction of $\mathbf{z} \in R$. Let T be this set of vectors $\mathbf{z} \in R$, for which the corresponding pairs are separated correctly. Clearly $|T| \geqslant \varepsilon'|R|$. Since R satisfies the property stated in Lemma 5, there exist $\mathbf{z}', \mathbf{z}'' \in T$ such that

$$\left| \langle \mathbf{r}_1, \mathbf{z}' \rangle - \langle \mathbf{r}_1, \mathbf{z}'' \rangle \right| \geqslant \varepsilon' / \big(100\sqrt{nk}\big). \tag{6}$$

Moreover, since the pairs are separated 'correctly' we have that

$$\big\langle \mathbf{r}_1, \big((1-\delta)\mathbf{x}' + \gamma\delta\mathbf{z}'\big)\big\rangle - c_1 \leqslant 0, \tag{7}$$

$$\big\langle \mathbf{r}_1, \big((1+\delta)\mathbf{x}' + \gamma\delta\mathbf{z}'\big)\big\rangle - c_1 \geqslant 0, \tag{8}$$

$$\big\langle \mathbf{r}_1, \big((1-\delta)\mathbf{x}' + \gamma\delta\mathbf{z}''\big)\big\rangle - c_1 \leqslant 0, \tag{9}$$

$$\big\langle \mathbf{r}_1, \big((1+\delta)\mathbf{x}' + \gamma\delta\mathbf{z}''\big)\big\rangle - c_1 \geqslant 0. \tag{10}$$

Subtracting Eq. (7) from (10), and (9) from (8), we obtain

$$2\delta\langle \mathbf{r}_1, \mathbf{x}' \rangle - \delta\gamma\big\langle \mathbf{r}_1, \big(\mathbf{z}' - \mathbf{z}''\big)\big\rangle \geqslant 0,$$
$$2\delta\langle \mathbf{r}_1, \mathbf{x}' \rangle + \delta\gamma\big\langle \mathbf{r}_1, \big(\mathbf{z}' - \mathbf{z}''\big)\big\rangle \geqslant 0.$$

Combining the above with Eq. (6) we get that

$$\left|2\delta\langle\mathbf{r}_1,\mathbf{x}'\rangle\right| \geqslant \gamma\delta\varepsilon'/\left(100\sqrt{nk}\right).$$

Substituting the value of γ and simplifying we have $|\langle\mathbf{r}_1,\mathbf{x}'\rangle| \geqslant \varepsilon'/(2\cdot 10^4 n\sqrt{k})$. Since $\mathbf{x}'$ takes values 1 or -1 on coordinates in $M(u)$ and is 0 on all other coordinates, this implies

$$\sum_{i\in k}\left|\mathbf{r}_1\left(u(i)\right)\right| \geqslant \frac{\varepsilon'}{2\cdot 10^4 n\sqrt{k}},$$

which is what we require.

Case 2. In this case we have that the halfspace $\{\langle\mathbf{r}_1,\mathbf{y}\rangle \leqslant c_1\}$ separates 'incorrectly' the pair $\{(1-\delta)\mathbf{x}'+\delta\gamma\mathbf{z}, (1+\delta)\mathbf{x}'+\delta\gamma\mathbf{z}\}$ for ε' fraction of $\mathbf{z}\in R$. The analysis is identical to Case 1, replacing the halfspace $\{\langle\mathbf{r}_1,\mathbf{y}\rangle \leqslant c_1\}$ by $\{\langle-\mathbf{r}_1,\mathbf{y}\rangle \leqslant -c_1\}$. We omit the details.

This completes the proof of Lemma 9. □

The above analysis shows that for a 'good' vertex u the vector $\mathbf{r}_1$ has non-negligible mass in the coordinates $M(u)$. To show it is concentrated in a small number of coordinates, we need the following lemma.

Lemma 10. *Let u be a 'good' vertex. There is a set $Q \subseteq s(u)$, s.t. $|Q| \geqslant \varepsilon'|s(u)|$ and for every $\mathbf{x}\in Q$, $\langle\mathbf{r}_1,\mathbf{x}\rangle \in [c_1-2\delta\sqrt{k}, c_1+2\delta\sqrt{k}]$.*

Proof. We consider the set Q of points $\mathbf{x}\in s(u)$, such that a pair $\{(1-\delta)\mathbf{x}+\delta\gamma\mathbf{z}, (1+\delta)\mathbf{x}+\delta\gamma\mathbf{z}\}$ for some $\mathbf{z}\in R$ is separated by $\langle\mathbf{r}_1,\mathbf{y}\rangle \leqslant c_1$. From Lemma 8, $|Q| \geqslant \varepsilon'|s(u)| = \varepsilon' 2^k$. Now, for any given $\mathbf{x}\in s(u)$, all the points of the form $(1-\delta)\mathbf{x}+\delta\gamma\mathbf{z}$ and $(1+\delta)\mathbf{x}+\delta\gamma\mathbf{z}$ for any $\mathbf{z}\in R$ lie in a ball of radius $2\delta\sqrt{k}$ around $\mathbf{x}$. Therefore, for all $\mathbf{x}\in Q$, the hyperplane $\langle\mathbf{r}_1,\mathbf{y}\rangle = c_1$ passes at a perpendicular distance of at most $2\delta\sqrt{k}$ from $\mathbf{x}$. The lemma follows. □

The following is a well-known lemma (see Lemma 7.3 of [13]). We state a version based on Lemma 3.5 proved in [5].

Lemma 11. *Let $X_1,\ldots,X_p$ be i.i.d. $\{-1,1\}$ valued Bernoulli random variables, with $\Pr[1]=\frac{1}{2}$, and let $\omega_1,\ldots,\omega_p$ be positive real numbers. Then there is a universal constant b such that, for any $c\in\mathbb{R}$ and $\zeta>0$, if*

$$\Pr\left[\sum_{i=1}^{p}\omega_i X_i \in [c-\zeta, c+\zeta]\right] \geqslant \frac{b}{p^{\frac{1}{2}}},$$

then $\exists i\in[p]$ such that $\omega_i \leqslant \zeta$.

Let $X_1,\ldots,X_k$ be i.i.d. $\{-1,1\}$ valued Bernoulli random variables with $\Pr[1]=\frac{1}{2}$. For convenience, we let $\delta'=\delta\sqrt{k}$. Observe that Lemma 10 implies that for a 'good' vertex u,

$$\Pr\left[\sum_{i=1}^{k}\left|\mathbf{r}_1\left(u(i)\right)\right|X_i \in \left[c_1-2\delta', c_1+2\delta'\right]\right] \geqslant \varepsilon'. \tag{11}$$

Suppose we apply Lemma 11 to the above. Then it gives us a coordinate in $M(u)$ such that on that coordinate $\mathbf{r}_1$ has very small mass. Removing that coordinate, we can again apply the lemma to the remaining coordinates and do this until a small number of coordinates remain. If we ensure that at each step we remove a coordinate from $M(u)$ on which $\mathbf{r}_1$ has small mass, then the total mass of the coordinates removed is small. Combining this with the lower bound given by Eq. (5), this would imply that most of the mass of $\mathbf{r}_1$ in $M(u)$ is concentrated in a small number of coordinates. This would enable us to select a labeling for the vertex u from among those 'large' coordinates. We formalize the above line of argument in the following lemma.

Lemma 12. *Let u be a 'good' vertex. Then there exists a set of indices $I^u \subseteq [k]$ satisfying*

$$\left|I^u\right| \leqslant b^2/\left(\varepsilon'\right)^2, \tag{12}$$

where b is the universal constant from Lemma 11, such that

$$\sum_{i\in I^u}\left|\mathbf{r}_1\left(u(i)\right)\right| \geqslant \frac{\varepsilon'}{4\cdot 10^4 n\sqrt{k}} \tag{13}$$

and

$$\sum_{i\in[k]\setminus I^u}\left|\mathbf{r}_1\big(u(i)\big)\right| \leqslant \frac{\varepsilon'}{16\cdot 10^4 n\sqrt{k}}. \tag{14}$$

Proof. As discussed above, we apply Lemma 11 iteratively starting with Eq. (11), until the set of coordinates is of size at most b^2/ε'^2, in the following manner. Initialize $I_0=[k]$.

1. At step j, we have a set of indices I_j of size $k-j$, with the following inequality satisfied

$$\Pr\Bigg[\sum_{i\in I_j}\left|\mathbf{r}_1\big(u(i)\big)\right|X_i\in\big[c_1-2^{j+1}\delta',c_1+2^{j+1}\delta'\big]\Bigg]\geqslant\varepsilon'.$$

2. If $k-j<b^2/\varepsilon'^2$, then we stop and obtain a set $I^u=I_j$ of indices, such that $|I^u|<b^2/\varepsilon'^2$.
3. If $k-j\geqslant b^2/\varepsilon'^2$, then we apply Lemma 11 to obtain $i'\in I_j$ such that $|\mathbf{r}_1(u(i'))|\leqslant 2^{j+1}\delta'$. Now for any setting $x_i\in\{-1,1\}$ of variables X_i for $i\in I_j$,

$$\sum_{i\in I_j}\left|\mathbf{r}_1\big(u(i)\big)\right|x_i\in\big[c_1-2^{j+1}\delta',c_1+2^{j+1}\delta'\big]\quad\Longrightarrow\quad\sum_{i\in I_j\setminus\{i'\}}\left|\mathbf{r}_1\big(u(i)\big)\right|x_i\in\big[c_1-2^{j+2}\delta',c_1+2^{j+2}\delta'\big].$$

Therefore we have

$$\Pr\Bigg[\sum_{i\in I_j\setminus\{i'\}}\left|\mathbf{r}_1\big(u(i)\big)\right|X_i\in\big[c_1-2^{j+2}\delta',c_1+2^{j+2}\delta'\big]\Bigg]\geqslant\varepsilon'.$$

So, we set $I_{j+1}=I_j\setminus\{i'\}$ and proceed to step $j+1$.

At the jth step, an index corresponding to a coordinate of mass at most $2^{j+1}\delta'$ is removed. There are at most k steps for $j=0,\dots,k-1$. Therefore, the total mass of the coordinates removed is at most $2^{k+1}\delta'$. Combining this with (5) and with our small enough choice of δ (as fixed in Section 3.2), we have a set $I^u\subseteq[k]$ such that $|I^u|\leqslant b^2/(\varepsilon')^2$ and

$$\sum_{i\in I^u}\left|\mathbf{r}_1\big(u(i)\big)\right|\geqslant\frac{\varepsilon'}{2\cdot 10^4 n\sqrt{k}}-2^{k+1}\delta'\geqslant\frac{\varepsilon'}{4\cdot 10^4 n\sqrt{k}}$$

and

$$\sum_{i\in[k]\setminus I^u}\left|\mathbf{r}_1\big(u(i)\big)\right|\leqslant 2^{k+1}\delta'\leqslant\frac{\varepsilon'}{16\cdot 10^4 n\sqrt{k}}.$$

This completes the proof of Lemma 12. □

Since an ε' fraction of the vertices of V are 'good', using we can obtain sets of indices I^v satisfying the properties of Lemma 12, for all 'good' vertices v. Construct the labeling σ^* to these vertices by choosing a label independently for every 'good' vertex $v\in V$ uniformly at random from I^v. We first choose the constants $\nu<\varepsilon'$ and $t>b^2/(\varepsilon')^2$. From the properties of the instance $\mathcal{L}$ of Smooth-Label-Cover(t,μ,ν,k,m) and the choice of ν, we obtain that the set of 'good' vertices induces an $(\varepsilon')^2/2$ fraction of edges in E. Let e be a random edge in E, and say e is between vertices v_1 and v_2 in V. Then with probability $(\varepsilon')^2/2$, both v_1 and v_2 are 'good'. Furthermore, by our choice of t, we obtain that except with probability 2μ, $\pi_{v_1,e}$ maps the elements of I^{v_1} to distinct elements $J^{v_1}\subseteq[m]$ and $\pi_{v_2,e}$ maps the elements of I^{v_2} to distinct elements in $J^{v_2}\subseteq[m]$.

Suppose for a contradiction that J^{v_1} and J^{v_2} are disjoint. This implies that $\pi^{-1}_{v_2,e}(J^{v_1})$ and I^{v_2} are disjoint. Since $\mathbf{r}_1$ is orthogonal to the subspace H, from (1) we have that for every $j\in J^{v_1}$,

$$\sum_{i\in\pi^{-1}_{v_1,e}(j)}\mathbf{r}_1\big(v_1(i)\big)=\sum_{i\in\pi^{-1}_{v_2,e}(j)}\mathbf{r}_1\big(v_2(i)\big)$$

and taking the absolute values and summing over all $j\in J^{v_1}$, we have

$$\sum_{j\in J^{v_1}}\Bigg|\sum_{i\in\pi^{-1}_{v_1,e}(j)}\mathbf{r}_1\big(v_1(i)\big)\Bigg|=\sum_{j\in J^{v_1}}\Bigg|\sum_{i\in\pi^{-1}_{v_2,e}(j)}\mathbf{r}_1\big(v_2(i)\big)\Bigg|. \tag{15}$$

Now, since $\pi_{v_1,e}$ maps elements of I^{v_1} to distinct elements $J^{v_1} \subseteq [m]$,

$$\sum_{j\in J^{v_1}}\left|\sum_{i\in\pi^{-1}_{v_1,e}(j)}\mathbf{r}_1\big(v_1(i)\big)\right| \geqslant \sum_{i\in I^{v_1}}\big|\mathbf{r}_1\big(v_1(i)\big)\big| - \sum_{i\in[k]\setminus I^{v_1}}\big|\mathbf{r}_1\big(v_1(i)\big)\big|.$$

From Eq. (15) and the above we have

$$\begin{aligned}\sum_{j\in J^{v_1}}\left|\sum_{i\in\pi^{-1}_{v_2,e}(j)}\mathbf{r}_1\big(v_2(i)\big)\right| &\geqslant \sum_{i\in I^{v_1}}\big|\mathbf{r}_1\big(v_1(i)\big)\big| - \sum_{i\in[k]\setminus I^{v_1}}\big|\mathbf{r}_1\big(v_1(i)\big)\big| \\ &\geqslant \frac{\varepsilon'}{4\cdot 10^4 n\sqrt{k}} - \frac{\varepsilon'}{16\cdot 10^4 n\sqrt{k}} \\ &= \frac{3\varepsilon'}{16\cdot 10^4 n\sqrt{k}} \qquad (16)\end{aligned}$$

where we used (13) and (14) applied to v_1. But since, $\pi^{-1}_{v_2,e}(J^{v_1}) \subseteq [k]\setminus I^{v_2}$, Eq. (16) is a contradiction to Eq. (14) applied to v_2. Therefore, J^{v_1} and J^{v_2} are not disjoint. So, with probability $1/(|I^{v_1}||I^{v_2}|) \geqslant (\varepsilon')^4/b^4$, the labeling σ^* satisfies the edge e.

Combining everything, we obtain that there is a labeling to the vertices of V that satisfies $((\varepsilon')^2/2 - 2\mu)((\varepsilon')^4/b^4)$ fraction of the edges in E. By choosing the smoothness parameter μ and the soundness parameter η of the instance $\mathcal{L}$ to be arbitrarily small, we obtain a contradiction. Thus, if the instance $\mathcal{L}$ of Smooth-Label-Cover(t,μ,ν,k,m) is a NO instance, then with high probability, there is no function of up to ℓ halfspaces that correctly classifies $\frac{1}{2}+\varepsilon$ fraction of the points in $A'\cup B'$. This, along with the analysis of the YES case proves Theorem 4, and hence Theorem 2.

5. Sampling from the unit sphere

In this section we prove Lemma 5. First we need some definitions.

Definition 3. A range space is a pair $(X,\mathcal{F})$, where X is a set and $\mathcal{F}$ is a family of subsets of X, i.e. $\mathcal{F}\subseteq 2^X$.

Definition 4. For any set $A\subseteq X$, define $P_{\mathcal{F}}(A)$ the projection of $\mathcal{F}$ onto A, as $P_{\mathcal{F}}(A) := \{F\cap A\colon\ F\in\mathcal{F}\}$.

Definition 5. We say that a set $A\subseteq X$ is shattered by $(X,\mathcal{F})$ if $P_{\mathcal{F}}(A) = 2^A$.

The VC dimension of a range space is defined as follows.

Definition 6. The VC dimension of $(X,\mathcal{F})$ is the cardinality of the maximum set it shatters, i.e. VC dim $= \sup\{|A|\colon A \text{ is shattered}\}$. It may be infinite.

We use the following restatement of Theorem 4.4 from [28], first proved in [29], regarding sampling from range spaces of bounded VC dimension.

Theorem 13. *Let $(X,\mathcal{F})$ a range space of VC dimension d, and let ϕ be a uniform measure on X. There is a universal constant C_{VC} such that with probability at least $1-\delta$, a random set $S\subseteq X$ of size,*

$$C_{VC}\left(\frac{d}{\tau^2}\log\left(\frac{d}{\tau}\right)+\frac{1}{\tau^2}\log\left(\frac{1}{\delta}\right)\right)$$

satisfies

$$\left|\frac{|S\cap F|}{|S|}-\phi(F)\right|\leqslant\tau,$$

for all $F\in\mathcal{F}$.

Let $N = nk$, let $\mathbb{S}^{N-1}$ be the unit sphere in N dimensions, and let ϕ be a uniform measure over $\mathbb{S}^{N-1}$. Define $P(\mathbf{r},[a,b]) := \{\mathbf{z}\in\mathbb{S}^{N-1} \mid \langle\mathbf{r},\mathbf{z}\rangle\in[a,b]\}$. The set $P(\mathbf{r},[a,b])$ is exactly the set of unit vectors whose dot product with $\mathbf{r}$ lies in the interval $[a,b]$. Let $\mathcal{P} := \{P(\mathbf{r},[a,b]) \mid \|\mathbf{r}\|=1,\ b-a=\varepsilon'/(100\sqrt{N})\}$. We know that the surface area of $\mathbb{S}^{N-1}$ is given by $\frac{N\pi^{N/2}}{\Gamma(N/2+1)}$ where $\Gamma(.)$ is the Gamma function. Stirling's approximation gives us that for large enough $z>0$, $\Gamma(z)=\sqrt{\frac{2\pi}{z}}(\frac{z}{e})^z(1+O(\frac{1}{z}))$. Combining the above, we obtain that for large enough N, the surface area of the $\mathbb{S}^{N-1}$ is at least $1/\sqrt{N}$ times the surface area of $\mathbb{S}^{N-2}$. As a result the following fact is easy to derive.

Fact 14. *For large enough N, for any $P \in \mathcal{P}$, $\phi(P) \leqslant \varepsilon'/10$.*

Moreover, we observe that every element $P \in \mathcal{P}$ is a set of points in $\mathbb{S}^{N-1}$ that lie in an intersection of two halfspaces. Since, in $\mathbb{R}^N$, the VC dimension of the class of all N-dimensional halfspaces is $N+1$, the VC dimension of the class of N-dimensional halfspaces for the set $\mathbb{S}^{N-1}$ is at most $N+1$. Using the above along with Lemma 2 of [26] we obtain the following simple bound.

Lemma 15. *The VC dimension of the range space $(\mathbb{S}^{N-1}, \mathcal{P})$ is at most $10(N+1)$.*

Now suppose R is a set of N^2 random unit vectors from $\mathbb{S}^{N-1}$. Then for large enough $N = nk$, we apply Theorem 13, choosing $\delta = 1/n$, along with the above lemma, to obtain that with probability at least $1 - 1/n$,

$$\left| \frac{|R \cap P|}{|R|} - \phi(P) \right| \leqslant \varepsilon'/10,$$

for any $P \in \mathcal{P}$. The above, coupled with Fact 14 implies that with probability $1 - 1/n$ over the choice of R,

$$\frac{|R \cap P|}{|R|} \leqslant \varepsilon'/5,$$

for all $P \in \mathcal{P}$. In other words, with probability $1 - 1/n$ over the choice of R, for any unit vector $\mathbf{r}$, at most $\varepsilon'/5$ fraction of points in R are contained in the set $\{\mathbf{z} \mid \langle \mathbf{r}, \mathbf{z} \rangle \in [a, b]\}$ for any a, b s.t. $b - a = \varepsilon'/(100\sqrt{N})$. This implies that with probability $1 - 1/n$ over the choice of R, for any set $T \subseteq R$, such that $|T| = \varepsilon'|R|$, for any unit vector $\mathbf{r}$, there exist $\mathbf{z}', \mathbf{z}'' \in T$, such that $|\langle \mathbf{r}, \mathbf{z}' \rangle - \langle \mathbf{r}, \mathbf{z}'' \rangle| \geqslant \varepsilon'/(100\sqrt{N})$. This proves Lemma 5.

6. Inapproximability of Smooth-Label-Cover

In this section we prove Theorem 3. Let us first define the ‘smooth’ version of the bipartite LabelCover problem.

Definition 7. An instance of Smooth-Bipartite-Label-Cover(k, m, T) consists of a bipartite graph $G(U, V, E)$, where the vertices in U have the same degree, and a set of projections $\pi^{vu} : [k] \mapsto [m]$ for all $\{u, v\} \in E$ such that $u \in U, v \in V$. A labeling σ to the vertices in G satisfies an edge $e = \{u, v\}$ s.t. $u \in U, v \in V$ iff $\pi^{vu}(\sigma(v)) = \sigma(u)$. Moreover, for any vertex $v \in V$ for any $i, j \in [k], i \neq j$,

$$\Pr_{e=\{u,v\} \in E}\left[\pi^{vu}(i) = \pi^{vu}(j)\right] \leqslant \frac{1}{T}, \tag{17}$$

where the probability is taken over a random edge incident on v.

The following theorem was proved in [17], using the PCP Theorem [2,3] and Raz’s Parallel Repetition Theorem [23].

Theorem 16. *For any constant $\delta > 0$, for any constant $T > 0$, there exist k and m such that given an instance $\mathcal{L}'$ of* Smooth-Bipartite-Label-Cover(k, m, T) *it is* NP*-hard to distinguish between the following two cases,*

- YES Case/Completeness: *There is a labeling to the vertices of $\mathcal{L}'$ that satisfies all the edges.*
- NO Case/Soundness: *No labeling to the vertices of $\mathcal{L}'$ satisfies more than δ fraction of the edges.*

The construction of an instance $\mathcal{L}$ of Smooth-Label-Cover(t, μ, ν, k, m) is as follows. We start with an instance $\mathcal{L}'$ of Smooth-Bipartite-Label-Cover(k, m, T) where we will fix T later. The vertex set of $\mathcal{L}$ is the V side of $\mathcal{L}'$. An edge of $\mathcal{L}$ is constructed as follows: select a vertex u from U and for every two neighbors v_1 and v_2 of u in $\mathcal{L}'$, add an edge e between them in $\mathcal{L}$. Set $\pi_{v_1,e} = \pi^{v_1 u}$ and $\pi_{v_2,e} = \pi^{v_2 u}$ in E. Let $E(u)$ be the set of such edges added in $\mathcal{L}$ corresponding to a vertex $u \in U$. Note that we are constructing a multigraph, since two vertices v_1 and v_2 in V might share two different neighbors in U, in which case there will be multiple edges between v_1 and v_2 in $\mathcal{L}$. Clearly the sets $E(u)$ for $u \in U$ are a partition of edges in $\mathcal{L}$, and since U side is regular, the sets $E(u)$ are of equal size. Essentially, we are adding a clique of edges $E(u)$ corresponding to u on its neighborhood $N(u) \subseteq V$ for every $u \in U$. Let v be a vertex in V and let $S \subseteq [k]$ be a set of size t, then applying Eq. (17) to all pairs in S and taking union bound, we have

$$\Pr_{e:\, v \in e}\left[\exists x, y \in S, x \neq y \colon \pi_{v,e}(x) = \pi_{v,e}(y)\right] \leqslant \frac{t^2}{T}$$

where the probability is over the edges incident on v in $\mathcal{L}$. Note that we have used the fact the vertices in U have the same degree. Now, taking T to be large enough, we can reduce this probability to at most μ. To verify the second property, let V' be a subset of V such that $|V'| = \xi|V|$, for some $\xi > 0$. Now, consider a vertex u in U, and let p_u be the probability that a random neighbor of u falls in V'. From the proof of Theorem 16 one can see that vertices on U side have the same degree, and therefore $\mathrm{E}_u[p_u] = \xi$. The degree of each vertex on the U side is the same fixed value d which can be increased to any arbitrary constant by parallel repetition. Moreover, the fraction of edges in $\mathcal{L}$ that lies inside V' is the probability for a random $u \in U$, a random pair of its neighbors lies in V'. For a particular u this is $p_u^2 - 1/d$, where $1/d$ is the probability of selecting the same vertex twice out of d neighbors of u. Hence, we have that the fraction of edges induced by V' in $\mathcal{L}$ is $\mathrm{E}_u[p_u^2 - 1/d] \geqslant (\mathrm{E}_u[p_u])^2 - 1/d = \xi^2 - 1/d$. This fraction is at least $\xi^2/2$ if $\xi \geqslant \sqrt{\frac{2}{d}}$, and so we are done by taking $\nu = \sqrt{\frac{2}{d}}$ which can be made arbitrarily small by taking d to be large enough.

Now, if $\mathcal{L}'$ was a YES instance, then there is a labeling σ to vertices $U \cup V$ that satisfies all the edges of $\mathcal{L}'$. This implies,

$$\pi^{v_1 u}\big(\sigma(v_1)\big) = \sigma(u) = \pi^{v_2 u}\big(\sigma(u)\big),$$

for all edges $e_1 = \{u, v_1\}, e_2 = \{u, v_2\}$ of $\mathcal{L}'$, $u \in U$, $v_1, v_2 \in N(u) \subseteq V$, where $N(u)$ is the neighborhood of $u \in U$ in $\mathcal{L}'$. Consider the edge $e \in E(u)$ between v_1 and v_2 in $\mathcal{L}$. Clearly, $\pi_{v_1,e}(\sigma(v_1)) = \pi_{v_2,e}(\sigma(v_2))$. Therefore, the labeling σ restricted to V satisfies all the edges of $\mathcal{L}$.

Now consider a labeling σ' to V that satisfies ε fraction of the edges in $\mathcal{L}$. Consider any vertex $u \in U$. For $j \in [m]$, let $S_u^j \subseteq N(u)$ the set of vertices $v \in N(u)$ such that $\pi^{vu}(\sigma'(v)) = j$. It can be seen that the sets S_u^j ($j \in [m]$) form a partition of $N(u)$ and the disjoint union of edges (corresponding to u) induced by each S_u^j in $E(u)$ is exactly the subset of edges of $E(u)$ that are satisfied by σ'. Let $l_u = \operatorname{argmax}_j |S_u^j|$ for each $u \in U$. Observe that seen that any subset S of $N(u)$ containing c ($c < 1$) fraction of vertices of $N(u)$ induces in $E(u)$ at most c^2 fraction of the total edges of $E(u)$. Suppose σ' satisfies ε_u fraction of the edges of $E(u)$, then a simple argument shows that $S_u^{l_u}$ must contain at least ε_u fraction of vertices in $N(u)$. Now σ' satisfies ε fraction of all the edges of $\mathcal{L}$, and since the U side is regular in $\mathcal{L}'$, we have that $\mathrm{E}_u[|S_u^{l_u}|/|N(u)|] \geqslant \mathrm{E}_u[\varepsilon_u] \geqslant \varepsilon$. Therefore, by extending the labeling σ' to U by setting $\sigma'(u) = l_u$ for $u \in U$, we can satisfy the edges of $\mathcal{L}'$ between vertices of $S_u^{l_u}$ and u for all $u \in U$. This would satisfy ε fraction of the edges in $\mathcal{L}'$. So, if the instance of $\mathcal{L}'$ is a NO instance with soundness η then there is no labeling to the vertices of $\mathcal{L}$ which satisfies more than η fraction of the edges of $\mathcal{L}$. This completes the proof of Theorem 3.

7. Conclusion

We proved a tight hardness result for learning intersection of two halfspaces using functions of up to ℓ halfspaces for any constant ℓ. An interesting open question is whether a similar hardness result holds for learning intersection of halfspaces by more general classes of hypotheses such as (functions of) low degree polynomials. As noted in the remark in Section 1.3 our reduction does not extend even to degree 2 polynomials.

In addition a limitation of our result is that the parameter ε in the $\frac{1}{2} + \varepsilon$ inapproximability factor is only an arbitrarily small constant. It is an important open problem to obtain a similar result with ε being an inverse *polynomial* in the dimension of the problem.

References

[1] M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, T. Pitassi, The complexity of properly learning simple concept classes, J. Comput. System Sci. 74 (1) (2008) 16–34.

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy, Proof verification and the hardness of approximation problems, J. ACM 45 (3) (1998) 501–555.

[3] S. Arora, S. Safra, Probabilistic checking of proofs: A new characterization of NP, J. ACM 45 (1) (1998) 70–122.

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

[5] I. Benjamin, G. Kalai, O. Schramm, Noise sensitivity of boolean functions and applications to percolation, Inst. Hautes Études Sci. Publ. Math. 90 (1999) 5–43.

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

[7] A. Blum, R. Kannan, Learning an intersection of a constant number of halfspaces over a uniform distribution, J. Comput. System Sci. 54 (2) (1997) 371–380.

[8] A. Blum, R. Rivest, Training a 3-node neural network is NP-complete, in: Proc. Machine Learning: From Theory to Applications, 1993, pp. 9–28.

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

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

[11] V. Feldman, V. Guruswami, P. Raghavendra, Y. Wu, Agnostic learning of monomials by halfspaces is hard, in: Proc. 50th IEEE FOCS, 2009.

[12] P. Gopalan, S. Khot, R. Saket, Hardness of reconstructing multivariate polynomials over finite fields, in: Proc. 48th IEEE FOCS, 2007, pp. 349–359.

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

[14] J. Holmerin, S. Khot, A new PCP outer verifier with applications to homogeneous linear equations and max-bisection, in: Proc. 36th ACM STOC, 2004, pp. 11–20.

[15] A.T. Kalai, A.R. Klivans, Y. Mansour, R.A. Servedio, Agnostically learning halfspaces, SIAM J. Comput. 37 (6) (2008) 1777–1805.

[16] S. Khot, Hardness results for coloring 3-colorable 3-uniform hypergraphs, in: Proc. 43rd IEEE FOCS, 2002, pp. 23–32.

[17] S. Khot, R. Saket, A 3-query non-adaptive PCP with perfect completeness, in: Proc. IEEE CCC, 2006, pp. 159–169.

[18] A.R. Klivans, R. O'Donnell, R.A. Servedio, Learning intersections and thresholds of halfspaces, J. Comput. System Sci. 68 (4) (2004) 808–840.

[19] A.R. Klivans, R.A. Servedio, Learning intersections of halfspaces with a margin, J. Comput. System Sci. 74 (1) (2008) 35–48.

[20] A.R. Klivans, A.A. Sherstov, Cryptographic hardness for learning intersections of halfspaces, J. Comput. System Sci. 75 (1) (2009) 2–12.

[21] A. Levin, A. Shashua, Principal component analysis over continuous subspaces and intersection of half-spaces, in: Proc. ECCCV(3), 2002, pp. 635–650.

[22] D. Murphy, Nearest neighbor pattern classification perceptrons, Proc. IEEE 78 (10) (1990) 1595–1598.

[23] R. Raz, A parallel repetition theorem, SIAM J. Comput. 27 (3) (1998) 763–803.

[24] U. Rückert, L. Richter, S. Kramer, Quantitative association rules based on half-spaces: An optimization approach, in: Proc. ICDM, 2004, pp. 507–510.

[25] R. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.

[26] E.D. Sontag, VC dimension of neural networks, in: Neural Networks and Machine Learning, Springer, Berlin, 1998, pp. 69–95.

[27] L.G. Valiant, A theory of the learnable, Commun. ACM 27 (11) (1984) 1134–1142.

[28] V.N. Vapnik, Statistical Learning Theory, Wiley–Interscience, 1998.

[29] V.N. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280.

[30] S. Vempala, A random sampling based algorithm for learning the intersection of half-spaces, in: Proc. 38th IEEE FOCS, 1997, pp. 508–513.