Proceedings of Machine Learning Research vol 65:1–29, 2017

Inapproximability of VC Dimension and Littlestone's Dimension

Pasin Manurangsi PASIN@BERKELEY.EDU and **Aviad Rubinstein** AVIAD@BERKELEY.EDU
University of California, Berkeley

Abstract

We study the complexity of computing the VC Dimension and Littlestone's Dimension. Given an explicit description of a finite universe and a concept class (a binary matrix whose (x, C)-th entry is 1 iff element x belongs to concept C), both can be computed exactly in quasi-polynomial time ($n^{O(\log n)}$). Assuming the randomized Exponential Time Hypothesis (ETH), we prove nearly matching lower bounds on the running time, that hold even for *approximation* algorithms.

Keywords: VC Dimension; Littlestone's Dimension; Hardness of Approximation.

1. Introduction

A common and essential assumption in learning theory is that the concepts we want to learn come from a nice, simple concept class, or (in the agnostic case) they can at least be approximated by a concept from a simple class. When the concept class is sufficiently simple, there is hope for good (i.e. sample-efficient and low-error) learning algorithms.

There are many different ways to measure the *simplicity* of a concept class. The most influential measure of simplicity is the VC Dimension, which captures learning in the PAC model. We also consider Littlestone's Dimension (Littlestone, 1988), which corresponds to minimizing mistakes in online learning (see Section 2 for definitions). When either dimension is small, there are algorithms that exploit the simplicity of the class, to obtain good learning guarantees.

Two decades ago, it was shown (under appropriate computational complexity assumptions) that neither dimension can be computed in polynomial time (Papadimitriou and Yannakakis, 1996; Frances and Litman, 1998); and these impossibility results hold even in the most optimistic setting where the entire universe and concept class are given as explicit input (a binary matrix whose (x, C)-th entry is 1 iff element x belongs to concept C). The computational intractability of computing the (VC, Littlestone's) dimension of a concept class suggests that even in cases where a simple structure exists, it may be inaccessible to computationally bounded algorithms (see Discussion below).

In this work we extend the results of (Papadimitriou and Yannakakis, 1996; Frances and Litman, 1998) to show that the VC and Littlestone's Dimensions cannot even be *approximately* computed in polynomial time. We don't quite prove that those problems are NP-hard:

both dimensions can be computed (exactly) in quasi-polynomial ($n^{O(\log n)}$) time, hence it is very unlikely that either problem is NP-hard. Nevertheless, assuming the randomized Exponential Time Hypothesis (ETH)[1] (Impagliazzo et al., 2001; Impagliazzo and Paturi, 2001), we prove essentially tight quasi-polynomial lower bounds on the running time - that hold even against approximation algorithms.

Theorem 1 (Hardness of Approximating VC Dimension) *Assuming Randomized ETH, approximating VC Dimension to within a $(1/2 + o(1))$-factor requires $n^{\log^{1-o(1)} n}$ time.*

Theorem 2 (Hardness of Approximating Littlestone's Dimension) *There exists an absolute constant $\varepsilon > 0$ such that, assuming Randomized ETH, approximating Littlestone's Dimension to within a $(1 - \varepsilon)$-factor requires $n^{\log^{1-o(1)} n}$ time.*

1.1. Discussion

As we mentioned before, the computational intractability of computing the (VC, Littlestone's) dimension of a concept class suggests that even in cases where a simple structure exists, it may be inaccessible to computationally bounded algorithms. We note however that it is not at all clear that any particular algorithmic applications are immediately intractable as a consequence of our results.

Consider for example the adversarial online learning zero-sum game corresponding to Littlestone's Dimension: At each iteration, Nature presents the learner with an element from the universe; the learner attempts to classify the element, and loses a point for every wrong classification; at the end of the iteration, the correct (binary) classification is revealed. The Littlestone's Dimension is equal to the worst case loss of the Learner before learning the exact concept. (see Section 2 for a more detailed definition.)

What can we learn from the fact that the Littlestone's Dimension is hard to compute? The first observation is that there is no efficient learner that can commit to a concrete mistake bound. But this does not rule out a computationally-efficient learner that plays optimal strategy and makes at most as many mistakes as the unbounded learner. We can, however, conclude that Nature's task is computationally intractable! Otherwise, we could efficiently construct an entire worst-case mistake tree (for a concept class $\mathcal{C}$, any mistake tree has at most $|\mathcal{C}|$ leaves, requiring $|\mathcal{C}| - 1$ oracle calls to Nature).

On a philosophical level, we think it is interesting to understand the implications of an intractable, adversarial Nature. Perhaps this is another evidence that the mistake bound model is too pessimistic?

Also, the only algorithm we know for computing the optimal learner's decision requires computing the Littlestone's Dimension. We think that it is an interesting open question whether an approximately optimal computationally-efficient learner exists.

[1] The randomized ETH (rETH) postulates that there is no $2^{o(n)}$-time Monte Carlo algorithms that solves $\mathsf{3SAT}$ on n variables correctly with probability at least $2/3$ (i.e. $\mathsf{3SAT} \notin \mathsf{BPTIME}(2^{o(n)})$).

In addition, let us note that in the other direction, computing Littlestone's Dimension exactly implies an exactly optimal learner. However, since the learner has to compute Littlestone's Dimension many times, we have no evidence that an approximation algorithm for Littlestone's Dimension would imply any guarantee for the learner.

Finally, we remark that for either problem (VC or Littlestone's Dimension), we are not aware of any non-trivial approximation algorithms.

1.2. Techniques

The starting point of our reduction is the framework of "birthday repetition" (Aaronson et al., 2014). This framework has seen many variations in the last few years, but the high level approach is as follows: begin with a hard-to-approximate instance of a 2CSP (such as 3-Color), and partition the vertices into $\sqrt{n}$-tuples. On one hand, by the birthday paradox, even if the original graph is sparse, we expect each pair of random $\sqrt{n}$-tuples to share an edge; this is crucial for showing hardness of approximation in many applications. On the other hand our reduction size is now approximately $N \approx 2^{\sqrt{n}}$ (there are $3^{\sqrt{n}}$ ways to color each $\sqrt{n}$-tuple), whereas by ETH solving 3-Color requires approximately $T(n) \approx 2^n$ time, so solving the larger problem also takes at least $T(n) \approx N^{\log N}$ time.

VC Dimension The first challenge we have to overcome in order to adapt this framework to hardness of approximation of VC Dimension is that the number of concepts involved in shattering a subset S is $2^{|S|}$. Therefore any inapproximability factor we prove on the size of the shattered set of elements, "goes in the exponent" of the size of the shattering set of concepts. Even a small constant factor gap in the VC Dimension requires proving a polynomial factor gap in the number of shattering concepts (obtaining polynomial gaps via "birthday repetition" for simpler problems is an interesting open problem (Manurangsi and Raghavendra, 2016; Manurangsi, 2017)). Fortunately, having a large number of concepts is also an advantage: we use each concept to test a different set of 3-Color constraints chosen independently at random; if the original instance is far from satisfied, the probability of passing all $2^{\Theta(|S|)}$ tests should now be doubly-exponentially small ($2^{-2^{\Theta(|S|)}}$)! More concretely, we think of half of the elements in the shattered set as encoding an assignment, and the other half as encoding which tests to run on the assignments.

Littlestone's Dimension Our starting point is the reduction for VC Dimension outlined in the previous paragraph. While we haven't yet formally introduced Littlestone's Dimension, recall that it corresponds to an online learning model. If the test-selection elements arrive before the assignment-encoding elements, the adversary can adaptively tailor his assignment to pass the specific test selected in the previous steps. To overcome this obstacle, we introduce a special gadget that forces the assignment-encoding elements to arrive first; this makes the reduction to Littlestone's Dimension somewhat more involved. Note that there is a reduction by (Frances and Litman, 1998) from VC Dimension to Littlestone's Dimension. Unfortunately, their reduction is not (approximately) gap-preserving, so we cannot use it directly to obtain Theorem 2 from Theorem 1.

1.3. Related Work

The study of the computational complexity of the VC Dimension was initiated by Linial, Mansour, and Rivest (Linial et al., 1991), who observed that it can be computed in quasi-polynomial time. (Papadimitriou and Yannakakis, 1996) proved that it is complete for the class LOGNP which they define in the same paper. (Frances and Litman, 1998) reduced the problem of computing the VC dimension to that of computing Littlestone's Dimension, hence the latter is also LOGNP-hard. (It follows as a corollary of our Theorem 1 that, assuming ETH, solving any LOGNP-hard problem requires quasi-polynomial time.)

Both problems were also studied in an implicit model, where the concept class is given in the form of a Boolean circuit that takes as input an element x and a concept c and returns 1 iff $x \in c$. Observe that in this model even computing whether either dimension is 0 or not is already NP-hard. Schafer proved that the VC Dimension is Σ_3^{P}-complete (Schaefer, 1999), while the Littlestone's Dimension is PSPACE-complete (Schaefer, 2000). (Mossel and Umans, 2002) proved that VC Dimension is Σ_3^{P}-hard to approximate to within a factor of almost 2; can be approximated to within a factor slightly better than 2 in AM; and is AM-hard to approximate to within $n^{1-\varepsilon}$.

Another line of related work in the implicit model proves computational intractability of PAC learning (which corresponds to the VC Dimension). Such intractability has been proved either from cryptographic assumptions, e.g. (Kearns and Valiant, 1994; Kharitonov, 1993, 1995; Feldman et al., 2006; Kalai et al., 2008; Klivans and Sherstov, 2009; Klivans, 2016) or from average case assumptions, e.g. (Daniely and Shalev-Shwartz, 2016; Daniely, 2016). (Blum, 1994) showed a "computational" separation between PAC learning and online mistake bound (which correspond to the VC Dimension and Littlestone's Dimension, respectively): if one-way function exist, then there is a concept class that can be learned by a computationally-bounded learner in the PAC model, but not in the mistake-bound model.

Recently, (Bazgan et al., 2016) introduced a generalization of VC Dimension which they call Partial VC Dimension, and proved that it is NP-hard to approximate (even when given an explicit description of the universe and concept class).

Our work is also related to many other quasi-polynomial lower bounds from recent years, which were also inspired by "birthday repetition"; these include problems like Densest k-Subgraph (Braverman et al., 2017; Manurangsi, 2017), Nash Equilibrium and related problems (Braverman et al., 2015; Rubinstein, 2015; Babichenko et al., 2016; Rubinstein, 2016a; Bhaskar et al., 2016; Deligkas et al., 2016) and Community Detection (Rubinstein, 2016b). It is interesting to note that so far "birthday repetition" has found very different applications, but they all share essentially the same quasi-polynomial *algorithm*: The bottleneck in those problem is a bilinear optimization problem $\max_{u,v} u^\top Av$, which we want to approximate to within a (small) constant additive factor. It suffices to find an $O(\log n)$-sparse sample $\hat{v}$ of the optimal v^*; the algorithm enumerates over all sparse $\hat{v}$'s (Lipton et al., 2003; Arora et al., 2012; Barman, 2015; Cheng et al., 2015). In contrast, the problems we consider in this paper have completely different quasi-polynomial time algorithms: For VC Dimension, it suffices to simply enumerate over all $\log |\mathcal{C}|$-tuples of elements (where $\mathcal{C}$ denotes the concept class and $\log |\mathcal{C}|$ is the trivial upper bound on the VC dimension) (Linial et al., 1991). Littlestone's Dimension can be computed in quasi-polynomial time via a recursive "divide and conquer" algorithm (See Appendix A).

2. Preliminaries

For a universe (or ground set) $\mathcal{U}$, a concept C is simply a subset of $\mathcal{U}$ and a concept class $\mathcal{C}$ is a collection of concepts. For convenience, we sometimes relax the definition and allow the concepts to not be subsets of $\mathcal{U}$; all definitions here extend naturally to this case.

The VC and Littlestone's Dimensions can be defined as follows.

Definition 3 (VC Dimension Vapnik and Chervonenkis (1971)) *A subset $S \subseteq \mathcal{U}$ is said to be* shattered *by a concept class $\mathcal{C}$ if, for every $T \subseteq S$, there exists a concept $C \in \mathcal{C}$ such that $T = S \cap C$.*

The VC Dimension VC-dim$(\mathcal{C},\mathcal{U})$ *of a concept class $\mathcal{C}$ with respect to the universe $\mathcal{U}$ is the largest d such that there exists a subset $S \subseteq \mathcal{U}$ of size d that is shattered by $\mathcal{C}$.*

Definition 4 (Mistake Tree and Littlestone's Dimension Littlestone (1988)) *A depth-d instance-labeled tree of $\mathcal{U}$ is a full binary tree of depth d such that every internal node of the tree is assigned an element of $\mathcal{U}$. For convenience, we will identify each node in the tree canonically by a binary string s of length at most d.*

A depth-d mistake tree (aka shattered tree Ben-David et al. (2009)) for a universe $\mathcal{U}$ and a concept class $\mathcal{C}$ is a depth-d instance-labeled tree of $\mathcal{U}$ such that, if we let $v_s \in \mathcal{U}$ denote the element assigned to the vertex s for every $s \in \{0,1\}^{<d}$, then, for every leaf $\ell \in \{0,1\}^d$, there exists a concept $C \in \mathcal{C}$ that agrees with the path from root to it, i.e., that, for every $i < d$, $v_{\ell_{\leqslant i}} \in C$ iff $\ell_{i+1} = 1$ where $\ell_{\leqslant i}$ denote the prefix of ℓ of length i.

The Littlestone's Dimension L-dim$(\mathcal{C},\mathcal{U})$ *of a concept class $\mathcal{C}$ with respect to the universe $\mathcal{U}$ is defined as the maximum d such that there exists a depth-d mistake tree for $\mathcal{U}, \mathcal{C}$.*

An equivalent formulation of Littlestone's Dimension is through mistakes made in online learning, as stated below. This interpretation will be useful in our proof.

Definition 5 (Mistake Bound) *An online algorithm $\mathcal{A}$ is an algorithm that, at time step i, is given an element $x_i \in \mathcal{U}$ and the algorithm outputs a prediction $p_i \in \{0,1\}$ whether x is in the class. After the prediction, the algorithm is told the correct answer $h_i \in \{0,1\}$. For a sequence $(x_1, h_1), \dots, (x_n, h_n)$,* prediction mistake *of $\mathcal{A}$ is defined as the number of incorect predictions, i.e., $\sum_{i \in n} \mathbb{1}[p_i \neq h_i]$. The* mistake bound *of $\mathcal{A}$ for a concept class $\mathcal{C}$ is defined as the maximum prediction mistake of $\mathcal{A}$ over all the sequences $(x_1, h_1), \dots, (x_n, h_n)$ which corresponds to a concept $C \in \mathcal{C}$ (i.e. $h_i = \mathbb{1}[x_i \in C]$ for all $i \in [n]$).*

Theorem 6 (Littlestone (1988)) *For any universe $\mathcal{U}$ and any concept class $\mathcal{C}$,* L-dim$(\mathcal{C},\mathcal{U})$ *is equal to the minimum mistake bound of $\mathcal{C}, \mathcal{U}$ over all online algorithms.*

The following facts are well-know and follow easily from the above definitions.

Fact 7 *For any universe $\mathcal{U}$ and concept class $\mathcal{C}$, we have*

$$\text{VC-dim}(\mathcal{C},\mathcal{U}) \leqslant \text{L-dim}(\mathcal{C},\mathcal{U}) \leqslant \log|\mathcal{C}|.$$

Fact 8 *For any two universes $\mathcal{U}_1,\mathcal{U}_2$ and any concept class $\mathcal{C}$,*

$$\text{L-dim}(\mathcal{C},\mathcal{U}_1 \cup \mathcal{U}_2) \leqslant \text{L-dim}(\mathcal{C},\mathcal{U}_1) + \text{L-dim}(\mathcal{C},\mathcal{U}_2).$$

2.1. Label Cover and PCP

As is standard in hardness of approximation, the starting point for our reductions will be the following problem called Label Cover.

Definition 9 (Label Cover) *A Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$ consists of a bipartite graph (A, B, E), an alphabet Σ, and, for every edge $(a, b) \in E$, a projection constraint $\pi_{(a,b)} : \Sigma \to \Sigma$.*

An assignment (aka labeling) for $\mathcal{L}$ is a function $\phi : A \cup B \to \Sigma$. The value of ϕ, $\mathrm{val}_{\mathcal{L}}(\phi)$ is defined as the fraction of edges $(a, b) \in E$ such that $\pi_{(a,b)}(\phi(a)) = \phi(b)$; these edges are called satisfied *edges. The value of the instance $\mathcal{L}$, $\mathrm{val}(\mathcal{L})$, is defined as the maximum value among all assignments $\phi : A \cup B \to \Sigma$.*

Throughout the paper, we often encounter an assignment that only labels a subset of $A \cup B$ but leaves the rest unlabeled. We refer to such assignment as a *partial assignment* to an instance; more specifically, for any $V \subseteq A \cup B$, a V-partial assignment (or partial assignment on V) is a function $\phi : V \to \Sigma$. For notational convenience, we sometimes write Σ^V to denote the set of all functions from V to Σ.

We will use the following version of the PCP Theorem by Moshkovitz and Raz, which reduces 3SAT to the gap version of Label Cover while preserves the size to be almost linear.

Theorem 10 (Moshkovitz-Raz PCP Moshkovitz and Raz (2010)) *For every n and every $\nu = \nu(n) > 0$, solving 3SAT on n variables can be reduced to distinguishing between the case that a bi-regular instance of Label Cover with $|A|, |B|, |E| = n^{1+o(1)}\mathsf{poly}(1/\nu)$ and $|\Sigma| = 2^{\mathsf{poly}(1/\nu)}$ is satisfiable and the case that its value is at most ν.*

2.2. Useful Lemmata

We end this section by listing a couple of lemmata that will be useful in our proofs.

Lemma 11 (Chernoff Bound) *Let $X_1, \dots, X_n$ be i.i.d. random variables taking value from $\{0, 1\}$ and let p be the probability that $X_i = 1$, then, for any $\delta > 0$, we have*

$$\Pr\left[\sum_{i=1}^{n} X_i \geqslant (1+\delta)np\right] \leqslant \begin{cases} 2^{-\delta^2 np/3} & \text{if } \delta < 1, \\ 2^{-\delta np/3} & \text{otherwise.} \end{cases}$$

Lemma 12 (Partitioning Lemma (Rubinstein, 2016b, Lemma 2.5)) *For any bi-regular bipartite graph $G = (A, B, E)$, let $n = |A| + |B|$ and $r = \sqrt{n}/\log n$. When n is sufficiently large, there exists a partition of $A \cup B$ into $U_1, \dots, U_r$ such that*

$$\forall i \in [r], \frac{n}{2r} \leqslant |U_i| \leqslant \frac{2n}{r}$$

and

$$\forall i, j \in [r], \frac{|E|}{2r^2} \leqslant |(U_i \times U_j) \cap E|, |(U_j \times U_i) \cap E| \leqslant \frac{2|E|}{r^2}.$$

Moreover, such partition can be found in randomized linear time (alternatively, deterministic $n^{O(\log n)}$ time).

3. Inapproximability of VC Dimension

In this section, we present our reduction from Label Cover to VC Dimension, stated more formally below. We note that this reduction, together with Moshkovitz-Raz PCP (Theorem 10), with parameter $\delta = 1/\log n$ gives a reduction from 3SAT on n variables to VC Dimension of size $2^{n^{1/2+o(1)}}$ with gap $1/2 + o(1)$, which immediately implies Theorem 1.

Theorem 13 *For every $\delta > 0$, there exists a randomized reduction from a bi-regular Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$ such that $|\Sigma| = O_\delta(1)$ to a ground set $\mathcal{U}$ and a concept class $\mathcal{C}$ such that, if $n \triangleq |A| + |B|$ and $r \triangleq \sqrt{n}/\log n$, then the following conditions hold for every sufficiently large n.*

- *(Size) The reduction runs in time $|\Sigma|^{O(|E|\mathsf{poly}(1/\delta)/r)}$ and $|\mathcal{C}|, |\mathcal{U}| \leqslant |\Sigma|^{O(|E|\mathsf{poly}(1/\delta)/r)}$.*
- *(Completeness) If $\mathcal{L}$ is satisfiable, then $\text{VC-dim}(\mathcal{C}, \mathcal{U}) \geqslant 2r$.*
- *(Soundness) If $\text{val}(\mathcal{L}) \leqslant \delta^2/100$, then $\text{VC-dim}(\mathcal{C}, \mathcal{U}) \leqslant (1 + \delta)r$ with high probability.*

In fact, the above properties hold with high probability even when δ and $|\Sigma|$ are not constants, as long as $\delta \geqslant \log(1000n \log |\Sigma|)/r$.

We remark here that when $\delta = 1/\log n$, Moshkovitz-Raz PCP produces a Label Cover instance with $|A| = n^{1+o(1)}, |B| = n^{1+o(1)}$ and $|\Sigma| = 2^{\mathsf{polylog}(n)}$. For such parameters, the condition $\delta \geqslant \log(1000n \log |\Sigma|)/r$ holds for every sufficiently large n.

3.1. A Candidate Reduction (and Why It Fails)

To best understand the intuition behind our reduction, we first describe a simpler candidate reduction and explain why it fails, which will lead us to the eventual construction. In this candidate reduction, we start by evoking Lemma 12 to partition the vertices $A \cup B$ of the Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$ into $U_1, \dots, U_r$ where $r = \sqrt{n}/\log n$. We then create the universe $\mathcal{U}$ and the concept class $\mathcal{C}$ as follows:

- We make each element in $\mathcal{U}$ correspond to a partial assignment to U_i for some $i \in [r]$, i.e., we let $\mathcal{U} = \{x_{i,\sigma_i} \mid i \in [r], \sigma_i \in \Sigma^{U_i}\}$. In the completeness case, we expect to shatter the set of size r that corresponds to a satisfying assignment $\sigma^* \in \Sigma^{A \cup B}$ of the Label Cover instance $\mathcal{L}$, i.e., $\{x_{i,\sigma^*|_{U_i}} \mid i \in [r]\}$. As for the soundness, our hope is that, if a large set $S \subseteq \mathcal{U}$ gets shattered, then we will be able to decode an assignment for $\mathcal{L}$ that satisfies many constraints, which contradicts with our assumption that val$(\mathcal{L})$ is small. Note that the number of elements of $\mathcal{U}$ in this candidate reduction is at most $r \cdot |\Sigma|^{O(|E|\mathsf{poly}(1/\delta)r)} = 2^{\tilde{O}(\sqrt{n})}$ as desired.

- As stated above, the intended solution for the completeness case is $\{x_{i,\sigma^*|_{U_i}} \mid i \in [r]\}$, meaning that we must have at least one concept corresponding to each subset $I \subseteq [r]$. We will try to make our concepts "test" the assignment; for each $I \subseteq [r]$, we will choose a set $T_I \subseteq A \cup B$ of $\tilde{O}(\sqrt{n})$ vertices and "test" all the constraints within T_I. Before we specify how T_I is picked, let us elaborate what "test" means: for each T_I-partial assignment ϕ_I that does not violate any constraints within T_I, we create a concept C_{I,ϕ_I}. This concept contains x_{i,σ_i} if and only if $i \in I$ and σ_i agrees with ϕ_I (i.e. $\phi_I|_{T_I \cap U_i} = \sigma_i|_{T_I \cap U_i}$). Recall that, if a set $S \subseteq \mathcal{U}$ is shattered, then each $\tilde{S} \subseteq S$ is an intersection between S and C_{I,ϕ_I} for some I, ϕ_I. We hope that the I's are different for different $\tilde{S}$ so that many different tests have been performed on S.

  Finally, let us specify how we pick T_I. Assume without loss of generality that r is even. We randomly pick a perfect matching between r, i.e., we pick a random permutation $\pi_I : [r] \to [r]$ and let $\left(\pi_I(1), \pi_I(2)\right), \dots, \left(\pi_I(r-1), \pi_I(r)\right)$ be the chosen matching. We pick T_I such that all the constraints in the matchings, i.e., constraints between $U_{\pi_I(2i-1)}$ and $U_{\pi_I(2i)}$ for every $i \in [r/2]$, are included. More specifically, for every $i \in [r]$, we include each vertex $v \in U_{\pi_I(2i-1)}$ if at least one of its neighbors lie in $U_{\pi_I(2i)}$ and we include each vertex $u \in U_{\pi_I(2i)}$ if at least one of its neighbors lie in $U_{\pi_I(2i-1)}$. By Lemma 12, for every pair in the matching the size of the intersection is at most $\frac{2|E|}{r^2}$, so each concept contains assignments to at most $\frac{2|E|}{r}$ variables; so the total size of the concept class is at most $2^r \cdot |\Sigma|^{\frac{2|E|}{r}}$.

Even though the above reduction has the desired size and completeness, it unfortunately fails in the soundness. Let us now sketch a counterexample. For simplicity, let us assume that each vertex in $T_{[r]}$ has a unique neighbor in $T_{[r]}$. Note that, since $T_{[r]}$ has quite small size (only $\tilde{O}(\sqrt{n})$), almost all the vertices in $T_{[r]}$ satisfy this property w.h.p., but assuming that all of them satisfy this property makes our life easier.

Pick an assignment $\tilde{\sigma} \in \Sigma^V$ such that none of the constraints in $T_{[r]}$ is violated. From our unique neighbor assumption, there is always such an assignment. Now, we claim that the set $S_{\tilde{\sigma}} \triangleq \{x_{i,\tilde{\sigma}|_{U_i}} \mid i \in [r]\}$ gets shattered. This is because, for every subset $I \subseteq [r]$, we can pick another assignment σ' such that σ' does not violate any constraint in $T_{[r]}$ and $\sigma'|_{U_i} = \tilde{\sigma}|_{U_i}$ if and only if $i \in I$. This implies that $\{x_{i,\tilde{\sigma}|_{U_i}} \mid i \in I\} = S \cap C_{[r],\sigma'}$ as desired. Note here that such σ' exists because, for every $i \notin I$, if there is a constraint from a vertex $a \in U_i \cap A$ to another vertex $b \in T_{[r]} \cap B$, then we can change the assignment to a in such a way that the constraint is not violated[2]; by doing this for every $i \notin I$, we have created the desired σ'. As a result, VC-dim$(\mathcal{C}, \mathcal{U})$ can still be as large as r even when the value of $\mathcal{L}$ is small.

[2] Here we assume that $|\pi^{-1}_{(a,b)}(\tilde{\sigma}(b))| > 1$; note that this always holds for Label Cover instances produced by Moshkovitz-Raz construction.

3.2. The Final Reduction

In this subsection, we will describe the actual reduction. To do so, let us first take a closer look at the issue with the above candidate reduction. In the candidate reduction, we can view each $I \subseteq [r]$ as being a seed used to pick a matching. Our hope was that many seeds participate in shattering some set S, and that this means that S corresponds to an assignment of high value. However, the counterexample showed that in fact only one seed ($I = [r]$) is enough to shatter a set. To circumvent this issue, we will not use the subset I as our seed anymore. Instead, we create r new elements $y_1, \dots, y_r$, which we will call *test selection elements* to act as seeds; namely, each subset $H \subseteq \mathcal{Y}$ will now be a seed. The benefit of this is that, if $S \subseteq \mathcal{Y}$ is shattered and contains test selection elements $y_{i_1}, \dots, y_{i_t}$, then at least 2^t seeds must participate in the shattering of S. This is because, for each $H \subseteq \mathcal{Y}$, the intersection of S with any concept corresponding to H, when restricted to $\mathcal{Y}$, is always $H \cap \{y_{i_1}, \dots, y_{i_t}\}$. Hence, each subset of $\{y_{i_1}, \dots, y_{i_t}\}$ must come a from different seed.

The only other change from the candidate reduction is that each H will test multiple matchings rather than one matching. This is due to a technical reason: we need the number of matchings, ℓ, to be large in order get the approximation ratio down to $1/2 + o(1)$; in our proof, if $\ell = 1$, then we can only achieve a factor of $1 - \varepsilon$ to some $\varepsilon > 0$. The full details of the reduction are shown in Figure 1.

Before we proceed to the proof, let us define some additional notation that will be used throughout.

- Every assignment element of the form x_{i,σ_i} is called an *i-assignment element*; we denote the set of all i-assignment elements by $\mathcal{X}_i$, i.e., $\mathcal{X}_i = \{x_{i,\sigma_i} \mid \sigma_i \in \Sigma^{U_i}\}$. Let $\mathcal{X}$ denote all the assignment elements, i.e., $\mathcal{X} = \bigcup_i \mathcal{X}_i$.
- For every $S \subseteq \mathcal{U}$, let $I(S)$ denote the set of all $i \in [r]$ such that S contains an i-assignment element, i.e., $I(S) = \{i \in [r] \mid S \cap \mathcal{X}_i \neq \emptyset\}$.
- We call a set $S \subseteq \mathcal{X}$ *non-repetitive* if, for each $i \in [r]$, S contains at most one i-assignment element, i.e., $|S \cap \mathcal{X}_i| \leqslant 1$. Each non-repetitive set S canonically induces a partial assignment $\phi(S) : \bigcup_{i \in I(S)} U_i \to \Sigma$. This is the unique partial assignment that satisfies $\phi(S)|_{U_i} = \sigma_i$ for every $x_{i,\sigma_i} \in S$
- Even though we define each concept as C_{I,H,σ_H} where σ_H is a partial assignment to a subset $T_H \subseteq A \cup B$, it will be more convenient to view each concept as $C_{I,H,\sigma}$ where $\sigma \in \Sigma^V$ is the assignment to the entire Label Cover instance. This is just a notational change: the actual definition of the concept does not depend on the assignment outside T_H.
- For each $I \subseteq [r]$, let U_I denote $\bigcup_{i \in I} U_i$. For each $\sigma_I \in \Sigma^{U_I}$, we say that (I, σ_I) *passes* $H \subseteq \mathcal{Y}$ if σ_I does not violate any constraint within T_H. Denote the collection of H's that (I, σ_I) passes by $\mathcal{H}(I, \sigma_I)$.
- Finally, for any non-repetitive set $S \subseteq \mathcal{X}$ and any $H \subseteq \mathcal{Y}$, we say that S *passes* H if $(I(S), \phi(S))$ passes H. We write $\mathcal{H}(S)$ as a shorthand for $\mathcal{H}(I(S), \phi(S))$.

The output size of the reduction and the completeness follow almost immediately from definition.

Input: A bi-regular Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$ and a parameter $\delta > 0$.

Output: A ground set $\mathcal{U}$ and a concept class $\mathcal{C}$.

The procedure to generate $(\mathcal{U}, \mathcal{C})$ works as follows:

- Let r be $\sqrt{n}/\log n$ where $n = |A| + |B|$. Use Lemma 12 to partition $A \cup B$ into r blocks $U_1, \dots, U_r$.
- For convenience, we assume that r is even. Moreover, for $i \neq j \in [r]$, let $\mathcal{N}_i(j) \subseteq U_i$ denote the set of all vertices in U_i with at least one neighbor in U_j (w.r.t. the graph (A, B, E)). We also extend this notation naturally to a set of j's; for $J \subseteq [r]$, $\mathcal{N}_i(J)$ denotes $\bigcup_{j \in J} \mathcal{N}_i(j)$.
- The universe $\mathcal{U}$ consists of two types of elements, as described below.
    - *Assignment elements*: for every $i \in [r]$ and every partial assignment $\sigma_i \in \Sigma^{U_i}$, there is an assignment element x_{i,σ_i} corresponding to it. Let $\mathcal{X}$ denote all the assignment elements, i.e., $\mathcal{X} = \{x_{i,\sigma_i} \mid i \in [r], \sigma_i \in \Sigma^{U_i}\}$.
    - *Test selection elements*: there are r test selection elements, which we will call $y_1, \dots, y_r$. Let $\mathcal{Y}$ denote the set of all test selection elements.
- The concepts in $\mathcal{C}$ are defined by the following procedure.
    - Let $\ell \triangleq 80/\delta^3$ be the number of matchings to be tested.
    - For each $H \subseteq \mathcal{Y}$, we randomly select ℓ permutations $\pi_H^{(1)}, \dots, \pi_H^{(\ell)} : [r] \to [r]$; this gives us ℓ matchings (i.e. the t-th matching is $\left(\pi_H^{(t)}(1), \pi_H^{(t)}(2)\right), \dots, \left(\pi_H^{(t)}(r-1), \pi_H^{(t)}(r)\right)$). For brevity, let us denote the set of (up to ℓ) elements that i is matched with in the matchings by $M_H(i)$. Let $T_H = \bigcup_i \mathcal{N}_i(M_H(i))$
    - For every $I \subseteq [r], H \subseteq \mathcal{Y}$ and for every partial assignment $\sigma_H \in \Sigma^{T_H}$ that does not violate any constraints, we create a concept C_{I,H,σ_H} such that each $x_{i,\sigma_i} \in \mathcal{X}$ is included in C_{I,H,σ_H} if and only if $i \in I$ and σ_i is consistent with σ_H, i.e., $\sigma_i|_{\mathcal{N}_i(M_H(i))} = \sigma_H|_{\mathcal{N}_i(M_H(i))}$ whereas $y_i \in \mathcal{Y}$ in included in C_{I,H,σ_H} if and only if $y \in H$.

Figure 1: Reduction from Label Cover to VC Dimension

Output Size of the Reduction. Clearly, the size of $\mathcal{U}$ is $\sum_{i \in [r]} |\Sigma|^{|U_i|} \leqslant r \cdot |\Sigma|^{n/r} \leqslant |\Sigma|^{O(|E|\mathsf{poly}(1/\delta)/r)}$. As for $|\mathcal{C}|$, note first that the number of choices for I and H are both 2^r. For fixed I and H, Lemma 12 implies that, for each matching $\pi_H^{(t)}$, the number of vertices from each U_i with at least one constraint to the matched partition in $\pi_H^{(t)}$ is at most $O(|E|/r^2)$. Since there are ℓ matchings, the number of vertices in $T_H = \mathcal{N}_1(M_H(1)) \cup \dots \cup \mathcal{N}_r(M_H(r))$ is at most $O(|E|\ell/r)$. Hence, the number of choices for the partial assignment σ_H is at most $|\Sigma|^{O(|E|\mathsf{poly}(1/\delta)/r)}$. In total, we can conclude that $\mathcal{C}$ contains at most $|\Sigma|^{O(|E|\mathsf{poly}(1/\delta)/r)}$ concepts.

Completeness. If $\mathcal{L}$ has a satisfying assignment $\sigma^* \in \Sigma^V$, then the set $S_{\sigma^*} = \{x_{i,\sigma^*|_{U_i}} \mid i \in [r]\} \cup \mathcal{Y}$ is shattered because, for any $S \subseteq S_{\sigma^*}$, we have $S = S_{\sigma^*} \cap C_{I(S), S \cap \mathcal{Y}, \sigma^*}$. Hence, VC-dim$(\mathcal{C},\mathcal{U}) \geqslant 2r$.

The rest of this section is devoted to the soundness analysis.

3.3. Soundness

In this subsection, we will prove the following lemma, which, combined with the completeness and output size arguments above, imply Theorem 13.

Lemma 14 *Let $(\mathcal{C},\mathcal{U})$ be the output from the reduction in Figure 1 on input $\mathcal{L}$. If* $\text{val}(\mathcal{L}) \leqslant \delta^2/100$ *and* $\delta \geqslant \log(1000n \log|\Sigma|)/r$*, then* $\text{VC-dim}(\mathcal{C},\mathcal{U}) \leqslant (1+\delta)r$ *w.h.p.*

At a high level, the proof of Lemma 14 has two steps:

1. Given a shattered set $S \subseteq \mathcal{U}$, we extract a maximal non-repetitive set $S^{\text{NO-REP}} \subseteq S$ such that $S^{\text{NO-REP}}$ passes many ($\geqslant 2^{|S|-|S^{\text{NO-REP}}|}$) H's. If $|S^{\text{NO-REP}}|$ is small, the trivial upper bound of 2^r on the number of different H's implies that $|S|$ is also small. As a result, we are left to deal with the case that $|S^{\text{NO-REP}}|$ is large.
2. When $|S^{\text{NO-REP}}|$ is large, $S^{\text{NO-REP}}$ induces a partial assignment on a large fraction of vertices of $\mathcal{L}$. Since we assume that $\text{val}(\mathcal{L})$ is small, this partial assignment must violate many constraints. We will use this fact to argue that, with high probability, $S^{\text{NO-REP}}$ only passes very few H's, which implies that $|S|$ must be small.

The two parts of the proof are presented in Subsection 3.3.1 and 3.3.2 respectively. We then combine them in Subsection 3.3.3 to prove Lemma 14.

3.3.1. Part I: Finding a Non-Repetitive Set That Passes Many Tests

The goal of this subsection is to prove the following lemma, which allows us to, given a shattered set $S \subseteq \mathcal{U}$, find a non-repetitive set $S^{\text{NO-REP}}$ that passes many H's.

Lemma 15 *For any shattered $S \subseteq \mathcal{U}$, there is a non-repetitive set $S^{\text{NO-REP}}$ of size $|I(S)|$ s.t. $|\mathcal{H}(S^{\text{NO-REP}})| \geqslant 2^{|S|-|I(S)|}$.*

We will start by proving the following lemma, which will be a basis for the proof of Lemma 15.

Lemma 16 *Let $C, C' \in \mathcal{C}$ correspond to the same H (i.e. $C = C_{I,H,\sigma}$ and $C' = C_{I',H,\sigma'}$ for some $H \subseteq \mathcal{Y}, I, I' \subseteq [r], \sigma, \sigma' \in \Sigma^V$).*

For any subset $S \subseteq \mathcal{U}$ and any maximal non-repetitive subset $S^{\text{NO-REP}} \subseteq S$, if $S^{\text{NO-REP}} \subseteq C$ and $S^{\text{NO-REP}} \subseteq C'$, then $S \cap C = S \cap C'$.

The most intuitive interpretation of this lemma is as follows. Recall that if S is shattered, then, for each $\tilde{S} \subseteq S$, there must be a concept $C_{I_{\tilde{S}},H_{\tilde{S}},\sigma_{\tilde{S}}}$ such that $\tilde{S} = S \cap C_{I_{\tilde{S}},H_{\tilde{S}},\sigma_{\tilde{S}}}$. The above lemma implies that, for each $\tilde{S} \supseteq S^{\text{NO-REP}}$, $H_{\tilde{S}}$ must be different. This means that at least $2^{|S|-|S^{\text{NO-REP}}|}$ different H's must be involved in shattering S. Indeed, this will be the argument we use when we prove Lemma 15.

Proof of Lemma 16 Let $S, S^{\text{NO-REP}}$ be as in the lemma statement. Suppose for the sake of contradiction that there exists $H \subseteq \mathcal{Y}, I, I' \subseteq [r], \sigma, \sigma' \in \Sigma^V$ such that $S^{\text{NO-REP}} \subseteq C_{I,H,\sigma}, S^{\text{NO-REP}} \subseteq C_{I',H,\sigma'}$ and $S \cap C_{I,H,\sigma} \neq S \cap C_{I',H,\sigma'}$.

First, note that $S \cap C_{I,H,\sigma} \cap \mathcal{Y} = S \cap H \cap \mathcal{Y} = S \cap C_{I',H,\sigma'} \cap \mathcal{Y}$. Since $S \cap C_{I,H,\sigma} \neq S \cap C_{I',H,\sigma'}$, we must have $S \cap C_{I,H,\sigma} \cap \mathcal{X} \neq S \cap C_{I',H,\sigma'} \cap \mathcal{X}$. Assume w.l.o.g. that there exists $x_{i,\sigma_i} \in (S \cap C_{I,H,\sigma}) \setminus (S \cap C_{I',H,\sigma'})$.

Note that $i \in I(S) = I(S^{\text{NO-REP}})$ (where the equality follows from maximality of $S^{\text{NO-REP}}$). Thus there exists $\sigma'_i \in \Sigma^{U_i}$ such that $x_{i,\sigma'_i} \in S^{\text{NO-REP}} \subseteq C_{I,H,\sigma} \cap C_{I',H,\sigma'}$. Since x_{i,σ'_i} is in both $C_{I,H,\sigma}$ and $C_{I',H,\sigma'}$, we have $i \in I \cap I'$ and

$$\sigma|_{\mathcal{N}_i(M_H(i))} = \sigma'_i|_{\mathcal{N}_i(M_H(i))} = \sigma'|_{\mathcal{N}_i(M_H(i))}. \tag{1}$$

However, since $x_{i,\sigma_i} \in (S \cap C_{I,H,\sigma}) \setminus (S \cap C_{I',H,\sigma'})$, we have $x_{i,\sigma_i} \in C_{I,H,\sigma} \setminus C_{I',H,\sigma'}$. This implies that

$$\sigma|_{\mathcal{N}_i(M_H(i))} = \sigma_i|_{\mathcal{N}_i(M_H(i))} \neq \sigma'|_{\mathcal{N}_i(M_H(i))},$$

which contradicts to (1). ■

In addition to the above lemma, we will also need the following observation, which states that, if a non-repetitive $S^{\text{NO-REP}}$ is contained in a concept C_{I,H,σ_H}, then $S^{\text{NO-REP}}$ must pass H. This observation follows definitions.

Observation 17 *If a non-repetitive set $S^{\text{NO-REP}}$ is a subset of some concept C_{I,H,σ_H}, then $H \in \mathcal{H}(S^{\text{NO-REP}})$.*

With Lemma 16 and Observation 17 ready, it is now easy to prove Lemma 15.

Proof of Lemma 15 Pick $S^{\text{NO-REP}}$ to be any maximal non-repetitive subset of S. Clearly, $|S^{\text{NO-REP}}| = |I(S)|$. To see that $|\mathcal{H}(S^{\text{NO-REP}})| \geqslant 2^{|S|-|I(S)|}$, consider any $\tilde{S}$ such that $S^{\text{NO-REP}} \subseteq \tilde{S} \subseteq S$. Since S is shattered, there exists $I_{\tilde{S}}, H_{\tilde{S}}, \sigma_{\tilde{S}}$ such that $S \cap C_{I_{\tilde{S}},H_{\tilde{S}},\sigma_{\tilde{S}}} = \tilde{S}$. Since $\tilde{S} \supseteq S^{\text{NO-REP}}$, Observation 17 implies that $H_{\tilde{S}} \in \mathcal{H}(S^{\text{NO-REP}})$. Moreover, from Lemma 16, $H_{\tilde{S}}$ is distinct for every $\tilde{S}$. As a result, $|\mathcal{H}(S^{\text{NO-REP}})| \geqslant 2^{|S|-|I(S)|}$ as desired. ■

3.3.2. Part II: No Large Non-Repetitive Set Passes Many Tests

The goal of this subsection is to show that, if $\text{val}(\mathcal{L})$ is small, then w.h.p. (over the randomness in the construction) every large non-repetitive set passes only few H's. This is formalized as Lemma 18 below.

Lemma 18 *If $\text{val}(\mathcal{L}) \leqslant \delta^2/100$ and $\delta \geqslant 8/r$, then, with high probability, for every non-repetitive set $S^{\text{NO-REP}}$ of size at least δr, $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log |\Sigma|$.*

Note that the mapping $S^{\text{NO-REP}} \mapsto (I(S^{\text{NO-REP}}), \phi(S^{\text{NO-REP}}))$ is a bijection from the collection of all non-repetitive sets to $\{(I, \sigma_I) \mid I \subseteq [r], \sigma_I \in \Sigma^{U_I}\}$. Hence, the above lemma is equivalent to the following.

Lemma 19 *If* $\mathrm{val}(\mathcal{L}) \leqslant \delta^2/100$ *and* $\delta \geqslant 8/r$*, then, with high probability, for every* $I \subseteq [r]$ *of size at least* δr *and every* $\sigma_I \in \Sigma^{U_I}$*,* $|\mathcal{H}(I, \sigma_I)| \leqslant 100n \log |\Sigma|$*.*

Here we use the language in Lemma 19 instead of Lemma 18 as it will be easier for us to reuse this lemma later. To prove the lemma, we first need to bound the probability that each assignment σ_I does not violate any constraint induced by a random matching. More precisely, we will prove the following lemma.

Lemma 20 *For any* $I \subseteq [r]$ *of size at least* δr *and any* $\sigma_I \in \Sigma^{U_I}$*, if* $\pi : [r] \to [r]$ *is a random permutation of* $[r]$*, then the probability that* σ_I *does not violate any constraint in* $\bigcup_{i \in [r]} \mathcal{N}_i(M(i))$ *is at most* $(1 - 0.1\delta^2)^{\delta r/8}$ *where* $M(i)$ *denote the index that* i *is matched with in the matching* $\Big(\pi(1), \pi(2)\Big), \dots, \Big(\pi(r-1), \pi(r)\Big)$*.*

Proof Let p be any positive odd integer such that $p \leqslant \delta r/2$ and let $i_1, \dots, i_{p-1} \in [r]$ be any $p-1$ distinct elements of $[r]$. We will first show that conditioned on $\pi(1) = i_1, \dots, \pi(p-1) = i_{p-1}$, the probability that σ_I violates a constraint induced by $\pi(p), \pi(p+1)$ (i.e. in $\mathcal{N}_{\pi(p)}(\pi(p+1)) \cup \mathcal{N}_{\pi(p+1)}(\pi(p))$) is at least $0.1\delta^2$.

To see that this is true, let $I_{\geqslant p} = I \setminus \{i_1, \dots, i_{p-1}\}$. Since $|I| \geqslant \delta r$, we have $|I_{\geqslant p}| = |I| - p + 1 \geqslant \delta r/2 + 1$. Consider the partial assignment $\sigma_{\geqslant p} = \sigma_I|_{U_{I_{\geqslant p}}}$. Since $\mathrm{val}(\mathcal{L}) \leqslant 0.01\delta^2$, $\sigma_{\geqslant p}$ can satisfy at most $0.01\delta^2|E|$ constraints. From Lemma 12, we have, for every $i \neq j \in I_{\geqslant p}$, the number of constraints between U_i and U_j are at least $|E|/r^2$. Hence, there are at most $0.01\delta^2 r^2$ pairs of $i < j \in I_{\geqslant p}$ such that $\sigma_{\geqslant p}$ does not violate any constraint between U_i and U_j. In other words, there are at least $\binom{|I_{\geqslant p}|}{2} - 0.01\delta^2 r^2 \geqslant 0.1\delta^2 r^2$ pairs $i < j \in I_{\geqslant p}$ such that $\sigma_{\geqslant p}$ violates some constraints between U_i and U_j. Now, if $\pi(p) = i$ and $\pi(p+1) = j$ for some such pair i, j, then $\phi(S^{\text{NO-REP}})$ violates a constraint induced by $\pi(p), \pi(p+1)$. Thus, we have

$$\Pr\left[\sigma_I \text{ does not violate a constraint induced by } \pi(p), \pi(p+1) \;\middle|\; \bigwedge_{t=1}^{p-1} \pi(t) = i_t\right] \leqslant 1 - 0.1\delta^2. \tag{2}$$

Let E_p denote the event that σ_I does not violate any constraints induced by $\pi(p)$ and $\pi(p+1)$. We can now bound the desired probability as follows.

$$\begin{aligned}
\Pr\left[\sigma_I \text{ does not violate any constraint in } \bigcup_{i \in [r]} \mathcal{N}_i(M(i))\right] &\leqslant \Pr\left[\bigwedge_{\text{odd } p \in [\delta r/2+1]} E_p\right] \\
&= \prod_{\text{odd } p \in [\delta r/2+1]} \Pr\left[E_p \;\middle|\; \bigwedge_{\text{odd } t \in [p-1]} E_t\right] \\
\text{(From (2))} &\leqslant \prod_{\text{odd } p \in [\delta r/2+1]} (1 - 0.1\delta^2) \\
&\leqslant (1 - 0.1\delta^2)^{\delta r/4 - 1},
\end{aligned}$$

which is at most $(1 - 0.1\delta^2)^{\delta r/8}$ since $\delta \geqslant 8/r$. ■

We can now prove our main lemma.

Proof of Lemma 19 For a fixed $I \subseteq [r]$ of size at least δr and a fixed $\sigma_I \in \Sigma^{U_I}$, Lemma 20 tells us that the probability that σ_I does not violate any constraint induced by a single matching is at most $(1-0.1\delta^2)^{\delta r/8}$. Since for each $H \subseteq \mathcal{Y}$ the construction picks ℓ matchings at random, the probability that (I, σ_I) passes each H is at most $(1-0.1\delta^2)^{\delta \ell r/8}$. Recall that we pick $\ell = 80/\delta^3$; this gives the following upper bound on the probability:

$$\Pr[(I,\sigma_I) \text{ passes } H] \leq (1-0.1\delta^2)^{\delta \ell r/8} = (1-0.1\delta^2)^{10r/\delta^2} \leqslant \left(\frac{1}{1+0.1\delta^2}\right)^{10r/\delta^2} \leqslant 2^{-r} \quad (3)$$

where the last inequality comes from Bernoulli's inequality.

Inequality (3) implies that the expected number of H's that (I, σ_I) passes is less than 1. Since the matchings M_H are independent for all H's, we can apply Chernoff bound which implies that

$$\Pr[|\mathcal{H}(I,\sigma_I)| \geqslant 100n \log |\Sigma|] \leqslant 2^{-10n \log |\Sigma|} = |\Sigma|^{-10n}.$$

Finally, note that there are at most $2^r|\Sigma|^n$ different (I, σ_I)'s. By union bound, we have

$$\begin{aligned}\Pr\left[\exists I \subseteq [r], \sigma_I \in \Sigma^{U_I} \text{ s.t. } |I| \geqslant \delta r \text{ AND } |\mathcal{H}(I,\sigma_I)| \geqslant 100n\log|\Sigma|\right] &\leqslant (2^r|\Sigma|^n)\left(|\Sigma|^{-10n}\right) \\ &\leqslant |\Sigma|^{-8n},\end{aligned}$$

which concludes the proof. ■

3.3.3. Putting Things Together

Proof of Lemma 14 From Lemma 18, every non-repetitive set $S^{\text{NO-REP}}$ of size at least δr, $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log |\Sigma|$. Conditioned on this event happening, we will show that VC-dim$(\mathcal{U}, \mathcal{C}) \leqslant (1+\delta)r$.

Consider any shattered set $S \subseteq \mathcal{U}$. Lemma 15 implies that there is a non-repetitive set $S^{\text{NO-REP}}$ of size $|I(S)|$ such that $|\mathcal{H}(S^{\text{NO-REP}})| \geqslant 2^{|S|-|I(S)|}$. Let us consider two cases:

1. $|I(S)| \leqslant \delta r$. Since $\mathcal{H}(S^{\text{NO-REP}}) \subseteq \mathcal{P}(\mathcal{Y})$, we have $|S| - |I(S)| \leqslant |\mathcal{Y}| = r$. This implies that $|S| \leqslant (1+\delta)r$.
2. $|I(S)| > \delta r$. From our assumption, $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log |\Sigma|$. Thus, $|S| \leqslant |I(S)| + \log(100n \log |\Sigma|) \leqslant (1+\delta)r$ where the second inequality comes from our assumption that $\delta \geqslant \log(1000n \log |\Sigma|)/r$.

Hence, VC-dim$(\mathcal{U}, \mathcal{C}) \leqslant (1+\delta)r$ with high probability. ■

4. Inapproximability of Littlestone's Dimension

We next proceed to Littlestone's Dimension. The main theorem of this section is stated below. Again, note that this theorem and Theorem 10 implies Theorem 2.

Theorem 21 *There exists $\varepsilon > 0$ such that there is a randomized reduction from any bi-regular Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$ with $|\Sigma| = O(1)$ to a ground set $\mathcal{U}$ and a concept classes $\mathcal{C}$ such that, if $n \triangleq |A| + |B|, r \triangleq \sqrt{n}/\log n$ and $k \triangleq 10^{10}|E|\log|\Sigma|/r^2$, then the following conditions hold for every sufficiently large n.*

- *(Size) The reduction runs in time $2^{rk} \cdot |\Sigma|^{O(|E|/r)}$ and $|\mathcal{C}|, |\mathcal{U}| \leqslant 2^{rk} \cdot |\Sigma|^{O(|E|/r)}$.*
- *(Completeness) If $\mathcal{L}$ is satisfiable, then $\text{L-dim}(\mathcal{C},\mathcal{U}) \geqslant 2rk$.*
- *(Soundness) If $\text{val}(\mathcal{L}) \leqslant 0.001$, then $\text{L-dim}(\mathcal{C},\mathcal{U}) \leqslant (2-\varepsilon)rk$ with high probability.*

4.1. Why the VC Dimension Reduction Fails for Littlestone's Dimension

It is tempting to think that, since our reduction from the previous section works for VC Dimension, it may also work for Littlestone's Dimension. In fact, thanks to Fact 7, completeness for that reduction even translates for free to Littlestone's Dimension. Alas, the soundness property does not hold. To see this, let us build a depth-$2r$ mistake tree for $\mathcal{C},\mathcal{U}$, even when $\text{val}(\mathcal{L})$ is small, as follows.

- We assign the test-selection elements to the first r levels of the tree, one element per level. More specifically, for each $s \in \{0,1\}^{<r}$, we assign $y_{|s|+1}$ to s.
- For every string $s \in \{0,1\}^r$, the previous step of the construction gives us a subset of $\mathcal{Y}$ corresponding to the path from root to s; this subset is simply $H_s = \{y_i \in \mathcal{Y} \mid s_i = 1\}$. Let T_{H_s} denote the set of vertices tested by this seed H_s. Let $\phi_s \in \Sigma^V$ denote an assignment that satisfies all the constraints in T_{H_s}. Note that, since T_{H_s} is of small size (only $\tilde{O}(\sqrt{n})$), even if $\text{val}(\mathcal{L})$ is small, ϕ_s is still likely to exist (and we can decide whether it exists or not in time $2^{\tilde{O}(\sqrt{n})}$).
  We then construct the subtree rooted at s that corresponds to ϕ_s by assigning each level of the subtree $x_{i,\phi_s|_{U_i}}$. Specifically, for each $t \in \{0,1\}^{\geqslant r}$, we assign $x_{|t|-r+1,\phi_{t_{\leqslant r}}|_{U_{|t|-r+1}}}$ to node t of the tree.

It is not hard to see that the constructed tree is indeed a valid mistake tree. This is because the path from root to each leaf $l \in \{0,1\}^{2r}$ agrees with $C_{I(l),H_{l_{\leqslant r}},\phi_{l\leqslant r}}$ (where $I(l) = \{i \in [r] \mid l_i = 1\})$.

4.2. The Final Reduction

The above counterexample demonstrates the main difference between the two dimensions: order does not matter in VC Dimension, but it does in Littlestone's Dimension. By moving the test-selection elements up the tree, the tests are chosen before the assignments, which allows an adversary to "cheat" by picking different assignments for different tests. We would like to prevent this, i.e., we would like to make sure that, in the mistake tree, the upper levels of the tree are occupied with the assignment elements whereas the lower levels are assigned test-selection elements. As in the VC Dimension argument, our hope here is that, given such a tree, we should be able to decode an assignment that passes tests on many different tests. Indeed we will tailor our construction to achieve such property.

Recall that, if we use the same reduction as VC Dimension, then, in the completeness case, we can construct a mistake tree in which the first r layers consist solely of assignment elements and the rest of the layers consist of only test-selection elements. Observe that there is no need for different nodes on the r-th layer to have subtrees composed of the same set of elements; the tree would still be valid if we make each test-selection element only work with a specific $s \in \{0,1\}^r$ and create concepts accordingly. In other words, we can modify our construction so that our test-selection elements are $\mathcal{Y} = \{y_{I,i} \mid I \subseteq [r], i \in [r]\}$ and the concept class is $\{C_{I,H,\sigma_H} \mid I \subseteq [r], H \subseteq \mathcal{Y}, \sigma_H \in \Sigma^{T_H}\}$ where the condition that an assignment element lies in C_{I,H,σ_H} is the same as in the VC Dimension reduction, whereas for $y_{I',i}$ to be in C_{I,H,σ_H}, we require not only that $i \in H$ but also that $I = I'$. Intuitively, this should help us, since each $y_{I,i}$ is now only in a small fraction ($\leqslant 2^{-r}$) of concepts; hence, one would hope that any subtree rooted at any $y_{I,i}$ cannot be too deep, which would indeed implies that the test-selection elements cannot appear in the first few layers of the tree.

Alas, for this modified reduction, it is not true that a subtree rooted at any $y_{I,i}$ has small depth; specifically, we can bound the depth of a subtree $y_{I,i}$ by the log of the number of concepts containing $y_{I,i}$ plus one (for the first layer). Now, note that $y_{I,i} \in C_{I',H,\sigma_H}$ means that $I' = I$ and $i \in H$, but there can be still as many as $2^{r-1} \cdot |\Sigma|^{|T_H|} = |\Sigma|^{O(|E|/r)}$ such concepts. This gives an upper bound of $r + O(|E| \log |\Sigma|/r)$ on the depth of the subtree rooted at $y_{I,i}$. However, $|E| \log |\Sigma|/r = \Theta(\sqrt{n} \log n) = \omega(r)$; this bound is meaningless here since, even in the completeness case, the depth of the mistake tree is only $2r$.

Fortunately, this bound is not useless after all: if we can keep this bound but make the intended tree depth much larger than $|E| \log |\Sigma|/r$, then the bound will indeed imply that no $y_{I,i}$-rooted tree is deep. To this end, our reduction will have one more parameter $k = \Theta(|E| \log |\Sigma|/r)$ where $\Theta(\cdot)$ hides a large constant and the intended tree will have depth $2rk$ in the completeness case; the top half of the tree (first rk layers) will again consist of assignment elements and the rest of the tree composes of the test-selection elements. The rough idea is to make k "copies" of each element: the assignment elements will now be $\{x_{i,\sigma_i,j} \mid i \in [r], \sigma_i \in \Sigma^{U_i}, j \in [k]\}$ and the test-selection elements will be $\{y_{I,i,j} \mid I \subseteq [r] \times [k], j \in [k]\}$. The concept class can then be defined as $\{C_{I,H,\sigma_H} \mid I \subseteq [r] \times [k], H \subseteq [r] \times [k], \sigma_H \in \Sigma^{T_H}\}$ naturally, i.e., H is used as the seed to pick the test set T_H, $y_{I',i,j} \in C_{I,H,\sigma_H}$ iff $I' = I$ and $(i,j) \in H$ whereas $x_{i,\sigma_i,j} \in C_{I,H,\sigma_H}$ iff $(i,j) \in I$ and $\sigma_i|_{(I,\sigma_I)} = \sigma_H|_{(I,\sigma_I)}$. For this concept class, we can again bound the depth of $y_{I,i}$-rooted tree to be $rk + O(|E| \log |\Sigma|/r)$; this time, however, rk is much larger than $|E| \log |\Sigma|/r$, so this bound is no more than, say, $1.001rk$. This is indeed the desired bound, since this means that, for any depth-$1.999rk$ mistake tree, the first $0.998rk$ layers must consist solely of assignment elements.

Unfortunately, the introduction of copies in turn introduces another technical challenge: it is not true any more that a partial assignment to a large set only passes a few tests w.h.p. (i.e. an analogue of Lemma 19 does not hold). By Inequality (3), each H is passed with probability at most 2^{-r}, but now we want to take a union bound there are $2^{rk} \gg 2^r$ different H's. To circumvent this, we will define a map $\tau : \mathcal{P}([r] \times [k]) \to \mathcal{P}([r])$ and use $\tau(H)$ to select the test instead of H itself. The map τ we use in the construction is the *threshold projection* where i is included in H if and only if, for at least half of $j \in [k]$, H contains (i, j). To motivate our choice of τ, recall that our overall proof approach is to first find a node that corresponds to an assignment to a large subset of the Label Cover instance; then argue that it can pass only a few tests, which we hope would imply that the subtree rooted there cannot be too deep. For this implication to be true, we need the following to also hold: for any small subset $\mathcal{H} \subseteq \mathcal{P}([r])$ of $\tau(H)$'s, we have that L-dim$(\tau^{-1}(\mathcal{H}), [r] \times [k])$ is small. This property indeed holds for our choice of τ (see Lemma 29).

With all the moving parts explained, we state the full reduction formally in Figure 2.

Input: A bi-regular Label Cover instance $\mathcal{L} = (A, B, E, \Sigma, \{\pi_e\}_{e \in E})$.
Output: A ground set $\mathcal{U}$ and a concept class $\mathcal{C}$.
The procedure to generate $(\mathcal{U}, \mathcal{C})$ works as follows:

- Let $r, U_1, \dots, U_r, \mathcal{N}$ be defined in the same manner as in Reduction 1 and let $k \triangleq 10^{10}|E| \log |\Sigma|/r^2$.
- The universe $\mathcal{U}$ consists of two types of elements, as described below.
  - *Assignment elements*: for every $i \in [r]$, every partial assignment $\sigma_i \in \Sigma^{U_i}$ and every $j \in [k]$, there is an assignment element $x_{i,\sigma_i,j}$ corresponding to it. Let $\mathcal{X}$ denote all the assignment elements, i.e., $\mathcal{X} = \{x_{i,\sigma_i,j} \mid i \in [r], \sigma_i \in \Sigma^{U_i}, j \in [k]\}$.
  - *Test-selection elements*: there are $rk(2^{rk})$ test-selection elements, which we will call $y_{I,i,j}$ for every $i \in [r], j \in [k], I \subseteq [r] \times [k]$. Let $\mathcal{Y}$ denote the set of all test-selection elements. Let $\mathcal{Y}_i$ denote $\{y_{I,i,j} \mid I \subseteq [r] \times [k], j \in [k]\}$. We call the elements of $\mathcal{Y}_i$ *i-test-selection elements*.
- The concepts in $\mathcal{C}$ are defined by the following procedure.
  - Let $\ell \triangleq 1000$ be the number of matchings to be tested.
  - For each $\tilde{H} \subseteq [r]$, we randomly select ℓ permutations $\pi_{\tilde{H}}^{(1)}, \dots, \pi_{\tilde{H}}^{(\ell)} : [r] \to [r]$; this gives us ℓ matchings (i.e. the t-th matching is $\left(\pi_{\tilde{H}}^{(t)}(1), \pi_{\tilde{H}}^{(t)}(2)\right), \dots, \left(\pi_{\tilde{H}}^{(t)}(r-1), \pi_{\tilde{H}}^{(t)}(r)\right)$). Denote the set of elements that i is matched with in the matchings by $M_{\tilde{H}}(i)$. Let $T_{\tilde{H}} = \bigcup_i \mathcal{N}_i(M_{\tilde{H}}(i))$
  - Let $\tau : \mathcal{P}([r] \times [k]) \to \mathcal{P}([r])$ denote the *threshold projection* operation where each $i \in [r]$ is included in $\tau(H)$ if and only if H contains at least half of the i-test-selection elements, i.e., $\tau(H) = \{i \in [r] \mid |H \cap \mathcal{Y}_i| \geqslant k/2\}$.
  - For every $I \subseteq [r] \times [k], H \subseteq [r] \times [k]$ and for every partial assignment $\sigma_{\tau(H)} \in \Sigma^{T_{\tau(H)}}$ that does not violate any constraints, we create a concept $C_{I,H,\sigma_{\tau(H)}}$ such that each $x_{i,\sigma_i,j} \in \mathcal{X}$ is included in $C_{I,H,\sigma_{\tau(H)}}$ if and only if $(i, j) \in I$ and σ_i is consistent with $\sigma_{\tau(H)}$, i.e., $\sigma_i|_{\mathcal{N}_i(M_{\tau(H)}(i))} = \sigma_{\tau(H)}|_{\mathcal{N}_i(M_{\tau(H)}(i))}$ whereas each $y_{I',i,j} \in \mathcal{Y}$ in included in $C_{I,H,\sigma_{\tau(H)}}$ if and only if $(i, j) \in H$ and $I' = I$.

Figure 2: Reduction from Label Cover to Littlestone's Dimension

Similar to our VC Dimension proof, we will use the following notation:

- For every $i \in [r]$, let $\mathcal{X}_i \triangleq \{x_{i,\sigma_i,j} \mid \sigma_i \in \Sigma^{U_i}, j \in [k]\}$; we refer to these elements as the i-assignment elements. Moreover, for every $(i, j) \in [r] \times [k]$, let $\mathcal{X}_{i,j} \triangleq \{x_{i,\sigma_i,j} \mid \sigma_i \in \Sigma^{U_i}\}$; we refer to these elements as the (i, j)-assignment elements.
- For every $S \subseteq \mathcal{U}$, let $I(S) = \{i \in [r] \mid S \cap \mathcal{X}_i \neq \emptyset\}$ and $IJ(S) = \{(i, j) \in [r] \times [k] \mid S \cap \mathcal{X}_{i,j} \neq \emptyset\}$.
- A set $S \subseteq \mathcal{X}$ is *non-repetitive* if $|S \cap \mathcal{X}_{i,j}| \leqslant 1$ for all $(i, j) \in [r] \times [k]$.
- We say that S *passes* $\tilde{H}$ if the following two conditions hold:
  - For every $i \in [r]$ such that $S \cap \mathcal{X}_i \neq \emptyset$, all i-assignment elements of S are consistent on $T_{\tilde{H}}|_{U_i}$, i.e., for every $(i, \sigma_i, j), (i, \sigma'_i, j') \in S$, we have $\sigma_i|_{U_i} = \sigma'_i|_{U_i}$.
  - The canonically induced assignment on $T_{\tilde{H}}$ does not violate any constraint (note that the previous condition implies that such assignment is unique).

  We use $\mathcal{H}(S)$ to denote the collection of all seeds $\tilde{H} \subseteq [r]$ that S passes.

We also use the following notation for mistake trees:

- For any subset $S \subseteq \mathcal{U}$ and any function $\rho : S \to \{0, 1\}$, let $\mathcal{C}[\rho] \triangleq \{C \in \mathcal{C} \mid \ \forall a \in S, a \in C \Leftrightarrow \rho(a) = 1\}$ be the collections of all concept that agree with ρ on S. We sometimes abuse the notation and write $\mathcal{C}[S]$ to denote the collection of all the concepts that contain S, i.e., $\mathcal{C}[S] = \{C \in \mathcal{C} \mid S \subseteq C\}$.
- For any binary string s, let $\text{pre}(s) \triangleq \{\emptyset, s_{\leqslant 1}, \dots, s_{\leqslant |s|-1}\}$ denote the set of all proper prefixes of s.
- For any depth-d mistake tree $\mathcal{T}$, let $v_{\mathcal{T},s}$ denote the element assigned to the node $s \in \{0, 1\}^{\leqslant d}$, and let $P_{\mathcal{T},s} \triangleq \{v_{\mathcal{T},s'} \mid s' \in \text{pre}(s)\}$ denote the set of all elements appearing from the path from root to s (excluding s itself). Moreover, let $\rho_{\mathcal{T},s} : P_{\mathcal{T},s} \to \{0, 1\}$ be the function corresponding to the path from root to s, i.e., $\rho_{\mathcal{T},s}(v_{\mathcal{T},s'}) = s_{|s'|+1}$ for every $s' \in \text{pre}(s)$.

Output Size of the Reduction The output size of the reduction follows immediately from a similar argument as in the VC Dimension reduction. The only different here is that there are 2^{rk} choices for I and H, instead of 2^r choices as in the previous construction.

Completeness. If $\mathcal{L}$ has a satisfying assignment $\sigma^* \in \Sigma^V$, we can construct a depth-rk mistake tree $\mathcal{T}$ as follows. For $i \in [r], j \in [k]$, we assign $x_{i,\sigma^*|_{U_i},j}$ to every node in the $((i-1)k+j)$-th layer of $\mathcal{T}$. Note that we have so far assigned every node in the first rk layers. For the rest of the vertices s's, if s lies in layer $rk + (i-1)k + j$, then we assign $y_{I(\rho^{-1}_{\mathcal{T},s}(1)),i,j}$ to it. It is clear that, for a leaf $s \in \{0, 1\}^{rk}$, the concept $C_{I(\rho^{-1}_{\mathcal{T},s}(1)),H_{\mathcal{T},s},\sigma^*}$ agrees with the path from root to s where $H_{\mathcal{T},s}$ is defined as $\{(i, j) \in [r] \times [k] \mid y_{I(\rho^{-1}_{\mathcal{T},s}(1)),i,j} \in \rho^{-1}_{\mathcal{T},s}(1)\}$. Hence, $\text{L-dim}(\mathcal{C}, \mathcal{U}) \geqslant 2rk$.

4.3. Soundness

Next, we will prove the soundness of our reduction, stated more precisely below. For brevity, we will assume throughout this subsection that r is sufficiently large, and leave it out of the lemmas' statements. Note that this lemma, together with completeness and output size properties we argue above, implies Theorem 21 with $\varepsilon = 0.001$.

Lemma 22 *Let $(\mathcal{C}, \mathcal{U})$ be the output from the reduction in Figure 2 on input $\mathcal{L}$. If $\text{val}(\mathcal{L}) \leqslant 0.001$, then $\text{L-dim}(\mathcal{C}, \mathcal{U}) \leqslant 1.999rk$ with high probability.*

Roughly speaking, the overall strategy of our proof of Lemma 22 is as follows:

1. First, we will argue that any subtree rooted at any test-selection element must be shallow (of depth $\leqslant 1.001rk$). This means that, if we have a depth-$1.999rk$ mistake tree, then the first $0.998rk$ levels must be assigned solely assignment elements.
2. We then argue that, in this $0.998rk$-level mistake tree of assignment elements, we can always extract a leaf s such that the path from root to s indicates inclusion of a large non-repetitive set. In other words, the path to s can be decoded into a (partial) assignment for the Label Cover instance $\mathcal{L}$.
3. Let the leaf from the previous step be s and the non-repetitive set be $S^{\text{NO-REP}}$. Our goal now is to show that the subtree rooted as s must have small depth. We start working towards this by showing that, with high probability, there are few tests that agree with $S^{\text{NO-REP}}$. This is analogous to Part II of the VC Dimension proof.
4. With the previous steps in mind, we only need to argue that, when $|\mathcal{H}(S^{\text{NO-REP}})|$ is small, the Littlestone's dimension of all the concepts that contains $S^{\text{NO-REP}}$ (i.e. $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{U})$) is small. Thanks to Fact 8, it is enough for us to bound $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{X})$ and $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{Y})$ separately. For the former, our technique from the second step also gives us the desired bound; for the latter, we prove that $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{Y})$ is small by designing an algorithm that provides correct predictions on a constant fraction of the elements in $\mathcal{Y}$.

Let us now proceed to the details of the proofs.

4.3.1. Part I: Subtree of a Test-Selection Assignment is Shallow

Lemma 23 *For any $y_{I,i,j} \in \mathcal{Y}$, $\text{L-dim}(\mathcal{C}[\{y_{I,i,j}\}], \mathcal{U}) \leqslant rk + (4|E|\ell/r)\log|\Sigma| \leqslant 1.001rk$.*

Note that the above lemma implies that, in any mistake tree, the depth of the subtree rooted at any vertex s assigned to some $y_{I,i,j} \in \mathcal{Y}$ is at most $1 + 1.001rk$. This is because every concept that agrees with the path from the root to s must be in $\mathcal{C}[\{y_{I,i,j}\}]$, which has depth at most $1.001rk$.

Proof of Lemma 23 Consider any $C_{I',H,\sigma_{\tau(H)}} \in \mathcal{C}[\{y_{I,i,j}\}], \mathcal{U})$. Since $y_{I,i,j} \in C_{I',H,\sigma_{\tau(H)}}$, we have $I = I'$. Moreover, from Lemma 12, we know that $\left|\mathcal{N}_i\left(M_{\tau(H)}(i)\right)\right| \leqslant 4|E|\ell/r^2$, which implies that $|T_{\tau(H)}| \leqslant 4|E|\ell/r$. This means that there are only at most $|\Sigma|^{4|E|\ell/r}$ choices of $\sigma_{\tau(H)}$. Combined with the fact that there are only 2^{rk} choices of H, we have $|\mathcal{C}[\{y_{I,i,j}\}]| \leqslant 2^{rk} \cdot |\Sigma|^{4|E|\ell/r}$. Fact 7 then implies the lemma. ■

4.3.2. Part II: Deep Mistake Tree Contains a Large Non-Repetitive Set

The goal of this part of the proof is to show that, for mistake tree of $\mathcal{X}, \mathcal{C}$ of depth slightly less than rk, there exists a leaf s such that the corresponding path from root to s indicates an inclusion of a large non-repetitive set; in our notation, this means that we would like to identify a leaf s such that $IJ(\rho^{-1}_{\mathcal{T},s}(1))$ is large. Since we will also need a similar bound later in the proof, we will prove the following lemma, which is a generalization of the stated goal that works even for the concept class $\mathcal{C}[S^{\text{NO-REP}}]$ for any non-repetitive $S^{\text{NO-REP}}$. To get back the desired bound, we can simply set $S^{\text{NO-REP}} = \emptyset$.

Lemma 24 *For any non-repetitive set $S^{\text{NO-REP}}$ and any depth-d mistake tree $\mathcal{T}$ of $\mathcal{X}, \mathcal{C}[S^{\text{NO-REP}}]$, there exists a leaf $s \in \{0,1\}^d$ such that $|IJ(\rho^{-1}_{\mathcal{T},s}(1)) \setminus IJ(S^{\text{NO-REP}})| \geqslant d - r$.*

The proof of this lemma is a double counting argument where we count a specific class of leaves in two ways, which ultimately leads to the above bound. The leaves that we focus on are the leaves $s \in \{0,1\}^d$ such that, for every (i,j) such that an (i,j)-assignment element appears in the path from root to s but not in $S^{\text{NO-REP}}$, the first appearance of (i,j)-assignment element in the path is included. In other words, for every $(i,j) \in IJ(P_{\mathcal{T},s}) \setminus IJ(S^{\text{NO-REP}})$, if we define $u_{i,j} \triangleq \inf_{s' \in \text{pre}(s), v_{\mathcal{T},s'} \in \mathcal{X}_{i,j}} |s'|$, then $s_{u_{i,j}+1}$ must be equal to 1. We call these leaves the *good* leaves. Denote the set of good leaves of $\mathcal{T}$ by $\mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$.

Our first way of counting is the following lemma. Informally, it asserts that different good leaves agree with different sets $\tilde{H} \subseteq [r]$. This can be thought of as an analogue of Lemma 16 in our proof for VC Dimension. Note that this lemma immediately gives an upper bound of 2^r on $|\mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}|$.

Lemma 25 *For any depth-d mistake tree $\mathcal{T}$ of $\mathcal{X}, \mathcal{C}[S^{\text{NO-REP}}]$ and any different good leaves $s_1, s_2 \in \mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$, if C_{I_1,H_1,σ_1} agrees with s_1 and C_{I_2,H_2,σ_2} agrees with s_2 for some $I_1, I_2, H_1, H_2, \sigma_1, \sigma_2$, then $\tau(H_1) \neq \tau(H_2)$.*

Proof Suppose for the sake of contradiction that there exist $s_1 \neq s_2 \in \mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$, $H_1, H_2, I_1, I_2, \sigma_1, \sigma_2$ such that C_{I_1,H_1,σ_1} and C_{I_2,H_2,σ_2} agree with s_1 and s_2 respectively, and $\tau(H_1) = \tau(H_2)$. Let s be the common ancestor of s_1, s_2, i.e., s is the longest string in $\text{pre}(s_1) \cap \text{pre}(s_2)$. Assume w.l.o.g. that $(s_1)_{|s|+1} = 0$ and $(s_2)_{|s|+1} = 1$. Consider the node $v_{\mathcal{T},s}$ in tree $\mathcal{T}$ where the paths to s_1, s_2 split; suppose that this is $x_{i,\sigma_i,j}$. Therefore $x_{i,\sigma_i,j} \in C_{I_2,H_2,\sigma_2} \setminus C_{I_1,H_1,\sigma_1}$.

We now argue that there is some $x_{i,\sigma'_i,j}$ (with the same i,j but a different assignment σ'_i) that is in both concepts, i.e. $x_{i,\sigma'_i,j} \in C_{I_2,H_2,\sigma_2} \cap C_{I_1,H_1,\sigma_1}$. We do this by considering two cases:

- If $(i,j) \in IJ(S^{\text{NO-REP}})$, then there is $x_{i,\sigma'_i,j} \in S^{\text{NO-REP}} \subseteq C_{I_1,H_1,\sigma_1}, C_{I_2,H_2,\sigma_2}$ for some $\sigma'_i \in \Sigma^{U_i}$.
- Suppose that $(i,j) \notin IJ(S^{\text{NO-REP}})$. Since s_1 is a good leaf, there is some $t \in \text{pre}(s)$ such that $v_{\mathcal{T},t} = x_{i,\sigma'_i,j}$ for some $\sigma'_i \in \Sigma^{U_i}$ and t is included by the path (i.e. $s_{|t|+1} = 1$). This also implies that $x_{i,\sigma'_i,j}$ is in both C_{I_1,H_1,σ_1} and C_{I_2,H_2,σ_2}.

Now, since both $x_{i,\sigma_i,j}$ and $x_{i,\sigma'_i,j}$ are in the concept C_{I_2,H_2,σ_2}, we have $(i,j) \in I_2$ and

$$\sigma_i|_{\mathcal{N}_i(M_{\tau(H_1)})} = \sigma_2|_{\mathcal{N}_i(M_{\tau(H_1)})} = \sigma'_i|_{\mathcal{N}_i(M_{\tau(H_1)})}. \tag{4}$$

On the other hand, since C_{I_1,H_1,σ_1} contains $x_{i,\sigma'_i,j}$ but not $x_{i,\sigma_i,j}$, we have $(i,j) \in I_1$ and

$$\sigma_i|_{\mathcal{N}_i(M_{\tau(H_2)})} \neq \sigma_1|_{\mathcal{N}_i(M_{\tau(H_2)})} = \sigma'_i|_{\mathcal{N}_i(M_{\tau(H_2)})}. \tag{5}$$

which contradicts (4) since $\tau(H_1) = \tau(H_2)$. ■

Next, we will present another counting argument which gives a lower bound on the number of good leaves, which, together with Lemma 25, yields the desired bound.

Proof of Lemma 24 For any depth-d mistake tree $\mathcal{T}$ of $\mathcal{C}[S^{\text{NO-REP}}], \mathcal{X}$, let us consider the following procedure which recursively assigns a weight λ_s to each node s in the tree. At the end of the procedure, all the weight will be propagated from the root to good leaves.

1. For every non-root node $s \in \{0,1\}^{\geqslant 1}$, set $\lambda_s \leftarrow 0$. For root $s = \emptyset$, let $\lambda_\emptyset \leftarrow 2^d$.
2. While there is an internal node $s \in \{0,1\}^{<d}$ such that $\lambda_s > 0$, do the following:
   (a) Suppose that $v_s = x_{i,\sigma_i,j}$ for some $i \in [r], \sigma_i \in \Sigma^{U_i}$ and $j \in [k]$.
   (b) If so far no (i,j)-element has appeared in the path or in $S^{\text{NO-REP}}$, i.e., $(i,j) \notin IJ(P_{\mathcal{T},s}) \cup IJ(S^{\text{NO-REP}})$, then $\lambda_{s1} \leftarrow \lambda_s$. Otherwise, set $\lambda_{s0} = \lambda_{s1} = \lambda_s/2$.
   (c) Set $\lambda_s \leftarrow 0$.

The following observations are immediate from the construction:

- The total of λ's over all the tree, $\sum_{s\in\{0,1\}^{\leqslant d}} \lambda_d$ always remain 2^d.
- At the end of the procedure, for every $s \in \{0,1\}^{\leqslant d}$, $\lambda_s \neq 0$ if and only if $s \in \mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$.
- If $s \in \mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$, then $\lambda_s = 2^{|IJ(\rho^{-1}_{\mathcal{T},s}(1)) \setminus IJ(S^{\text{NO-REP}})|}$ at the end of the execution.

Note that the last observation comes from the fact that λ always get divides in half when moving down one level of the tree unless we encounter an (i,j)-assignment element for some i,j that never appears in the path or in $S^{\text{NO-REP}}$ before. For any good leaf s, the set of such (i,j) is exactly the set $IJ(\rho^{-1}_{\mathcal{T},s}(1)) \setminus IJ(S^{\text{NO-REP}})$.

As a result, we have $2^d = \sum_{s\in\mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}} 2^{|IJ(\rho^{-1}_{\mathcal{T},s}(1)) \setminus IJ(S^{\text{NO-REP}})|}$. Since Lemma 25 implies that $|\mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}| \leqslant 2^r$, we can conclude that there exists $s \in \mathcal{G}_{\mathcal{T},S^{\text{NO-REP}}}$ such that $|IJ(\rho^{-1}_{\mathcal{T},s}(1)) \setminus IJ(S^{\text{NO-REP}})| \geqslant d - r$ as desired. ■

4.3.3. Part III: No Large Non-Repetitive Set Passes Many Test

The main lemma of this subsection is the following, which is analogous to Lemma 18

Lemma 26 *If* $\text{val}(\mathcal{L}) \leqslant 0.001$*, then, with high probability, for every non-repetitive set* $S^{\text{NO-REP}}$ *of size at least* $0.99rk$*,* $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log|\Sigma|$.

Proof For every $I \subseteq [r]$, let $U_I \triangleq \bigcup_{i \in I} U_i$. For every $\sigma_I \in \Sigma^{U_I}$ and every $\tilde{H} \subseteq \mathcal{Y}$, we say that (I, σ_I) *passes* $\tilde{H}$ if σ_I does not violate any constraint in $T_{\tilde{H}}$. Note that this definition and the way the test is generated in the reduction is the same as that of the VC Dimension reduction. Hence, we can apply Lemma 19 with $\delta = 0.99$, which implies the following: with high probability, for every $I \subseteq [r]$ of size at least $0.99r$ and every $\sigma_I \in \Sigma^{U_I}$, $|\mathcal{H}(I, \sigma_I)| \leqslant 100n \log |\Sigma|$ where $\mathcal{H}(I, \sigma_I)$ denote the set of all $\mathcal{H}$'s passed by (I, σ_I). Conditioned on this event happening, we will show that, for every non-repetitive set $S^{\text{NO-REP}}$ of size at least $0.99rk$, $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log |\Sigma|$.

Consider any non-repetitive set $S^{\text{NO-REP}}$ of size $0.99rk$. Let $\sigma_{I(S^{\text{NO-REP}})}$ be an assignment on $U_{I(S^{\text{NO-REP}})}$ such that, for each $i \in I(S^{\text{NO-REP}})$, we pick one $x_{i,\sigma_i,j} \in S^{\text{NO-REP}}$ (if there are more than one such x's, pick one arbitrarily) and let $\sigma_{I(S^{\text{NO-REP}})}|_{U_i} = \sigma_i$. It is obvious that $\mathcal{H}(S^{\text{NO-REP}}) \subseteq \mathcal{H}(I(S^{\text{NO-REP}}), \sigma_{I(S^{\text{NO-REP}})})$. Since $S^{\text{NO-REP}}$ is non-repetitive and of size at least $0.99rk$, we have $|I(S^{\text{NO-REP}})| \geqslant 0.99r$, which means that $|\mathcal{H}(I(S^{\text{NO-REP}}), \sigma_{I(S^{\text{NO-REP}})})| \leqslant 100n \log |\Sigma|$ as desired. ■

4.3.4. Part IV: A Subtree Containing $S^{\text{NO-REP}}$ Must be Shallow

In this part, we will show that, if we restrict ourselves to only concepts that contain some non-repetitive set $S^{\text{NO-REP}}$ that passes few tests, then the Littlestone's Dimension of this restrictied concept class is small. Therefore when we build a tree for the whole concept class $\mathcal{C}$, if a path from root to some node indicates an inclusion of a non-repetitive set that passes few tests, then the subtree rooted at this node must be shallow.

Lemma 27 *For every non-repetitive set* $S^{\text{NO-REP}}$,

$$\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{U}) \leqslant 1.75rk - |S^{\text{NO-REP}}| + r + 1000k\sqrt{r}\log(|\mathcal{H}(S^{\text{NO-REP}})| + 1).$$

We prove the above lemma by bounding $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{X})$ and $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{Y})$ separately, and combining them via Fact 8. First, we can bound $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{X})$ easily by applying Lemma 24 coupled with the fact that $|IJ(S^{\text{NO-REP}})| = |S^{\text{NO-REP}}|$ for every non-repetitive $S^{\text{NO-REP}}$. This immediately gives the following corollary.

Corollary 28 *For every non-repetitive set* $S^{\text{NO-REP}}$,

$$\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{X}) \leqslant rk - |S^{\text{NO-REP}}| + r.$$

We will next prove the following bound on $\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{Y})$. Note that Corollary 28, Lemma 29, and Fact 8 immediately imply Lemma 27.

Lemma 29 *For every non-repetitive set* $S^{\text{NO-REP}}$,

$$\text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{Y}) \leqslant 0.75rk + 500k\sqrt{r}\log(|\mathcal{H}(S^{\text{NO-REP}})| + 1).$$

The overall outline of the proof of Lemma 29 is that we will design a prediction algorithm whose mistake bound is at most $0.75rk + 1000k\sqrt{r}\log|\mathcal{H}(S^{\text{NO-REP}})|$. Once we design this algorithm, Lemma 6 immediately implies Lemma 29. To define our algorithm, we will need the following lemma, which is a general statement that says that, for a small collection of H's, there is a some $\tilde{H}^* \subseteq [r]$ that agrees with almost half of every H in the collection.

Lemma 30 *Let $\mathcal{H} \subseteq \mathcal{P}([r])$ be any collections of subsets of $[r]$, there exists $\tilde{H}^* \subseteq [r]$ such that, for every $\tilde{H} \in \mathcal{H}$, $|\tilde{H}^* \Delta \tilde{H}| \leqslant 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)$ where Δ denotes the symmetric difference between two sets.*

Proof We use a simple probabilistic method to prove this lemma. Let $\tilde{H}^r$ be a random subset of $[r]$ (i.e. each $i \in [r]$ is included independently with probability 0.5). We will show that, with non-zero probability, $|\tilde{H}^r \Delta \tilde{H}| \leqslant 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)$ for all $\tilde{H} \in \mathcal{H}$, which immediately implies that a desired $\tilde{H}^*$ exists.

Fix $\tilde{H} \in \mathcal{H}$. Observe that $|\tilde{H}^r \Delta \tilde{H}|$ can be written as $\sum_{i \in [r]} \mathbb{1}[i \in (\tilde{H}^r \Delta \tilde{H})]$. For each i, $\mathbb{1}[i \in (\tilde{H}^r \Delta \tilde{H})]$ is a $0,1$ random variable with mean 0.5 independent of other $i' \in [r]$. Applying Chernoff bound here yields

$$\Pr[|\tilde{H}^r \Delta \tilde{H}| > 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)] \leqslant 2^{-\log^2(|\mathcal{H}|+1)} \leqslant \frac{1}{|\mathcal{H}|+1}.$$

Hence, by union bound, we have

$$\Pr[\exists \tilde{H} \in \mathcal{H}, |\tilde{H}^r \Delta \tilde{H}| > 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)] \leqslant \frac{|\mathcal{H}|}{|\mathcal{H}|+1} < 1.$$

In other words, $|\tilde{H}^r \Delta \tilde{H}| \leqslant 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)$ for all $\tilde{H} \in \mathcal{H}$ with non-zero probability as desired. ■

We also need the following observation, which is an analogue of Observation 17 in the VC Dimension proof; it follows immediately from definition of $\mathcal{H}(S)$.

Observation 31 *If a non-repetitive set $S^{\text{NO-REP}}$ is a subset of some concept $C_{I,H,\sigma_{\tau(H)}}$, then $\tau(H) \in \mathcal{H}(S^{\text{NO-REP}})$.*

With Lemma 30 and Observation 31 in place, we are now ready to prove Lemma 29.

Proof of Lemma 29 Let $\tilde{H}^* \subseteq [r]$ be the set guaranteed by applying Lemma 30 with $\mathcal{H} = \mathcal{H}(S^{\text{NO-REP}})$. Let $H^* \triangleq \tilde{H}^* \times [k]$.

Our prediction algorithm will be very simple: it always predicts according to H^*; i.e., on an input[3] $y \in \mathcal{Y}$, it outputs $\mathbb{1}[y \in H^*]$. Consider any sequence $(y_1, h_1), \dots, (y_w, h_w)$ that agrees with a concept $C_{I,H,\sigma_{\tau(H)}} \in \mathcal{C}[S^{\text{NO-REP}}]$. Observe that the number of incorrect predictions of our algorithm is at most $|H^* \Delta H|$.

Since $C_{I,H,\sigma_{\tau(H)}} \in \mathcal{C}[S^{\text{NO-REP}}]$, Observation 31 implies that $\tau(H) \in \mathcal{H}(S^{\text{NO-REP}})$. This means that $|\tau(H) \Delta \tilde{H}^*| \leqslant 0.5r + 1000\sqrt{r}\log(|\mathcal{H}|+1)$. Now, let us consider each $i \in [r] \setminus (\tau(H) \Delta \tilde{H}^*)$.

[3] We assume w.l.o.g. that input elements are distinct; if an element appears multiple times, we know the correct answer from its first appearance and can always correctly predict it afterwards.

Suppose that $i \in \tau(H) \cap \tilde{H}^*$. Since $i \in \tau(H)$, at least $k/2$ elements of $\mathcal{Y}_i$ are in H and, since $i \in \tilde{H}^*$, we have $\mathcal{Y}_i \subseteq H^*$. This implies that $|(H^* \Delta H) \cap Y_i| \leqslant k/2$. A similar bound can also be derived when $i \notin \tau(H) \cap \tilde{H}^*$. As a result, we have

$$\begin{aligned}
|H^* \Delta H| &= \sum_{i \in [r]} |(H^* \Delta H) \cap Y_i| \\
&= \sum_{i \in \tau(H) \Delta \tilde{H}^*} |(H^* \Delta H) \cap Y_i| + \sum_{i \in [r] \setminus (\tau(H) \Delta \tilde{H}^*)} |(H^* \Delta H) \cap Y_i| \\
&\leqslant (|\tau(H) \Delta \tilde{H}^*|)(k) + (r - |\tau(H) \Delta \tilde{H}^*|)(k/2) \\
&\leqslant 0.75rk + 500k\sqrt{r} \log(|\mathcal{H}| + 1),
\end{aligned}$$

concluding our proof of Lemma 29. ■

4.3.5. Putting Things Together

Proof of Lemma 22 Assume that $\text{val}(\mathcal{L}) \leqslant 0.001$. From Lemma 26, we know that, with high probability, $|\mathcal{H}(S^{\text{NO-REP}})| \leqslant 100n \log |\Sigma|$ for every non-repetitive set $S^{\text{NO-REP}}$ of size at least $0.99rk$. Conditioned on this event, we will show that $\text{L-dim}(\mathcal{C}, \mathcal{U}) \leqslant 1.999rk$.

Suppose for the sake of contradiction that $\text{L-dim}(\mathcal{C}, \mathcal{U}) > 1.999rk$. Consider any depth-$1.999rk$ mistake tree $\mathcal{T}$ of $\mathcal{C}, \mathcal{U}$. From Lemma 23, no test-selection element is assigned to any node in the first $1.999rk - 1.001rk - 1 \geqslant 0.997rk$ levels. In other words, the tree induced by the first $0.997rk$ levels is simply a mistake tree of $\mathcal{C}, \mathcal{X}$. By Lemma 24 with $S^{\text{NO-REP}} = \emptyset$, there exists $s \in \{0,1\}^{0.997rk}$ such that $|IJ(\rho^{-1}_{\mathcal{T},s}(1))| \geqslant 0.997rk - r \geqslant 0.996rk$.

Since $|IJ(\rho^{-1}_{\mathcal{T},s}(1))| \geqslant 0.996rk$, there exists a non-repetitive set $S^{\text{NO-REP}} \subseteq \rho^{-1}_{\mathcal{T},s}(1)$ of size $0.996rk$. Consider the subtree rooted at s. This is a mistake tree of $\mathcal{C}[\rho_{\mathcal{T},s}], \mathcal{U}$ of depth $1.002rk$. Since $S^{\text{NO-REP}} \subseteq \rho^{-1}_{\mathcal{T},s}(1)$, we have $\mathcal{C}[\rho_{\mathcal{T},s}] \subseteq \mathcal{C}[S^{\text{NO-REP}}]$. However, this implies

$$\begin{aligned}
1.002rk &\leqslant \text{L-dim}(\mathcal{C}[\rho_{\mathcal{T},s}], \mathcal{U}) \\
&\leqslant \text{L-dim}(\mathcal{C}[S^{\text{NO-REP}}], \mathcal{U}) \\
\text{(From Lemma 27)} &\leqslant 1.75rk - 0.996rk + r + 100k\sqrt{r} \log(|\mathcal{H}(S^{\text{NO-REP}})| + 1) \\
\text{(From Lemma 26)} &\leqslant 0.754rk + r + 100k\sqrt{r} \log(100n \log |\Sigma| + 1) \\
&= 0.754rk + o(rk),
\end{aligned}$$

which is a contradiction when r is sufficiently large. ■

5. Conclusion and Open Questions

In this work, we prove inapproximability results for VC Dimension and Littlestone's Dimension based on the randomized exponential time hypothesis. Our results provide an almost matching running time lower bound of $n^{\log^{1-o(1)} n}$ for both problems while ruling out approximation ratios of $1/2 + o(1)$ and $1 - \varepsilon$ for some $\varepsilon > 0$ for VC Dimension and Littlestone's Dimension respectively. Even though our results help us gain more insights on approximability of both problems, it is not yet completely resolved. More specifically, we are not aware of any constant factor $n^{o(\log n)}$-time approximation algorithm for either problem; it is an intriguing open question whether such algorithm exists and, if not, whether our reduction can be extended to rule out such algorithm. Another potentially interesting research direction is to derandomize our construction; note that the only place in the proof in which the randomness is used is in Lemma 19.

A related question which remains open, originally posed by Ben-David and Eiron Ben-David and Eiron (1998), is that of computing the *self-directed learning*[4] mistake bound. Similarly, it may be interesting to understand the complexity of computing (approximating) the recursive teaching dimension Doliwa et al. (2014); Moran et al. (2015).

Acknowledgement

We thank Shai Ben-David for suggesting the question of approximability of Littlestone's dimension, and several other fascinating discussions. We also thank Yishay Mansour and COLT anonymous reviewers for their useful comments.

Pasin Manurangsi is supported by NSF Grants No. CCF 1540685 and CCF 1655215.

Aviad Rubinstein was supported by a Microsoft Research PhD Fellowship, as well as NSF grant CCF1408635 and Templeton Foundation grant 3966. This work was done in part at the Simons Institute for the Theory of Computing.

References

Scott Aaronson, Russell Impagliazzo, and Dana Moshkovitz. AM with multiple Merlins. In *IEEE 29th Conference on Computational Complexity, CCC 2014, Vancouver, BC, Canada, June 11-13, 2014*, pages 44–55, 2014. doi: 10.1109/CCC.2014.13. URL http://dx.doi.org/10.1109/CCC.2014.13.

Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *ACM Conference on Electronic Commerce, EC '12, Valencia, Spain, June 4-8, 2012*, pages 37–54, 2012. doi: 10.1145/2229012.2229020. URL http://doi.acm.org/10.1145/2229012.2229020.

Yakov Babichenko, Christos H. Papadimitriou, and Aviad Rubinstein. Can almost everybody be almost happy? In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 1–9, 2016. doi: 10.1145/2840728.2840731. URL http://doi.acm.org/10.1145/2840728.2840731.

[4] Roughly, self-directed learning is similar to the online learning model corresponding to Littlestone's dimension, but where the learner chooses the order elements; see Ben-David and Eiron (1998) for details.

Siddharth Barman. Approximating Nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 361–369, 2015. doi: 10.1145/2746539.2746566. URL http://doi.acm.org/10.1145/2746539.2746566.

Cristina Bazgan, Florent Foucaud, and Florian Sikora. On the approximability of partial VC dimension. In *Combinatorial Optimization and Applications - 10th International Conference, COCOA 2016, Hong Kong, China, December 16-18, 2016, Proceedings*, pages 92–106, 2016. doi: 10.1007/978-3-319-48749-6_7. URL http://dx.doi.org/10.1007/978-3-319-48749-6_7.

Shai Ben-David and Nadav Eiron. Self-directed learning and its relation to the vc-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998. doi: 10.1023/A:1007510732151. URL http://dx.doi.org/10.1023/A:1007510732151.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/~colt2009/papers/032.pdf#page=1.

Umang Bhaskar, Yu Cheng, Young Kun Ko, and Chaitanya Swamy. Hardness results for signaling in Bayesian zero-sum and network routing games. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, pages 479–496, 2016. doi: 10.1145/2940716.2940753. URL http://doi.acm.org/10.1145/2940716.2940753.

Avrim Blum. Separating distribution-free and mistake-bound learning models over the boolean domain. *SIAM J. Comput.*, 23(5):990–1000, 1994. doi: 10.1137/S009753979223455X. URL http://dx.doi.org/10.1137/S009753979223455X.

Mark Braverman, Young Kun-Ko, and Omri Weinstein. Approximating the best Nash equilibrium in $n^{o(\log n)}$-time breaks the exponential time hypothesis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 970–982, 2015. doi: 10.1137/1.9781611973730.66. URL http://dx.doi.org/10.1137/1.9781611973730.66.

Mark Braverman, Young Kun-Ko, Aviad Rubinstein, and Omri Weinstein. ETH hardness for densest-k-subgraph with perfect completeness. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1326–1341, 2017. doi: 10.1137/1.9781611974782.86. URL http://dx.doi.org/10.1137/1.9781611974782.86.

Yu Cheng, Ho Yee Cheung, Shaddin Dughmi, Ehsan Emamjomeh-Zadeh, Li Han, and Shang-Hua Teng. Mixture selection, mechanism design, and signaling. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1426–1445, 2015. doi: 10.1109/FOCS.2015.91. URL http://dx.doi.org/10.1109/FOCS.2015.91.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 105–117, 2016. doi: 10.1145/2897518.2897520. URL http://doi.acm.org/10.1145/2897518.2897520.

Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning DNF's. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 815–830, 2016. URL http://jmlr.org/proceedings/papers/v49/daniely16.html.

Argyrios Deligkas, John Fearnley, and Rahul Savani. Inapproximability results for approximate Nash equilibria. *CoRR*, abs/1608.03574, 2016. URL http://arxiv.org/abs/1608.03574.

Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, vc-dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014. URL http://dl.acm.org/citation.cfm?id=2697064.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 563–574, 2006. doi: 10.1109/FOCS.2006.51. URL http://dx.doi.org/10.1109/FOCS.2006.51.

Moti Frances and Ami Litman. Optimal mistake bound learning is hard. *Inf. Comput.*, 144 (1):66–82, 1998. doi: 10.1006/inco.1998.2709. URL http://dx.doi.org/10.1006/inco.1998.2709.

Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-SAT. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001. doi: 10.1006/jcss.2000.1727. URL http://dx.doi.org/10.1006/jcss.2000.1727.

Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001. doi: 10.1006/jcss.2001.1774. URL http://dx.doi.org/10.1006/jcss.2001.1774.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008. doi: 10.1137/060649057. URL http://dx.doi.org/10.1137/060649057.

Michael J. Kearns and Leslie G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *J. ACM*, 41(1):67–95, 1994. doi: 10.1145/174644.174647. URL http://doi.acm.org/10.1145/174644.174647.

Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 372–381, 1993. doi: 10.1145/167088.167197. URL http://doi.acm.org/10.1145/167088.167197.

Michael Kharitonov. Cryptographic lower bounds for learnability of boolean functions on the uniform distribution. *J. Comput. Syst. Sci.*, 50(3):600–610, 1995. doi: 10.1006/jcss.1995.1046. URL http://dx.doi.org/10.1006/jcss.1995.1046.

Adam R. Klivans. Cryptographic hardness of learning. In *Encyclopedia of Algorithms*, pages 475–477. 2016. doi: 10.1007/978-1-4939-2864-4_96. URL http://dx.doi.org/10.1007/978-1-4939-2864-4_96.

Adam R. Klivans and Alexander A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. *J. Comput. Syst. Sci.*, 75(1):2–12, 2009. doi: 10.1016/j.jcss.2008.07.008. URL http://dx.doi.org/10.1016/j.jcss.2008.07.008.

Nathan Linial, Yishay Mansour, and Ronald L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Inf. Comput.*, 90(1):33–49, 1991. doi: 10.1016/0890-5401(91)90058-A. URL http://dx.doi.org/10.1016/0890-5401(91)90058-A.

Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proceedings 4th ACM Conference on Electronic Commerce (EC-2003), San Diego, California, USA, June 9-12, 2003*, pages 36–41, 2003. doi: 10.1145/779928.779933. URL http://doi.acm.org/10.1145/779928.779933.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, April 1988. ISSN 0885-6125. doi: 10.1023/A:1022869011914. URL http://dx.doi.org/10.1023/A:1022869011914.

Pasin Manurangsi. Almost-polynomial ratio ETH-hardness of approximating densest k-subgraph. In *Proceedings of the Fortieth-ninth Annual ACM Symposium on Theory of Computing*, STOC '17, 2017. To appear.

Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense CSPs. *CoRR*, abs/1607.02986, 2016. URL http://arxiv.org/abs/1607.02986.

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Compressing and teaching for low vc-dimension. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 40–51, 2015. doi: 10.1109/FOCS.2015.12. URL https://doi.org/10.1109/FOCS.2015.12.

Dana Moshkovitz and Ran Raz. Two-query PCP with subconstant error. *J. ACM*, 57(5):29:1–29:29, 2010. doi: 10.1145/1754399.1754402. URL http://doi.acm.org/10.1145/1754399.1754402.

Elchanan Mossel and Christopher Umans. On the complexity of approximating the VC dimension. *J. Comput. Syst. Sci.*, 65(4):660–671, 2002. doi: 10.1016/S0022-0000(02)00022-3. URL http://dx.doi.org/10.1016/S0022-0000(02)00022-3.

Christos H. Papadimitriou and Mihalis Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. *J. Comput. Syst. Sci.*, 53(2):161–170, 1996. doi: 10.1006/jcss.1996.0058. URL http://dx.doi.org/10.1006/jcss.1996.0058.

Aviad Rubinstein. ETH-hardness for signaling in symmetric zero-sum games. *CoRR*, abs/1510.04991, 2015. URL http://arxiv.org/abs/1510.04991.

Aviad Rubinstein. Settling the complexity of computing approximate two-player Nash equilibria. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 258–265, 2016a. doi: 10.1109/FOCS.2016.35. URL http://dx.doi.org/10.1109/FOCS.2016.35.

Aviad Rubinstein. Detecting communities is hard, and counting them is even harder. *CoRR*, abs/1611.08326, 2016b. URL http://arxiv.org/abs/1611.08326.

Marcus Schaefer. Deciding the Vapnik-Cervonenkis dimension in Σ_3^P-complete. *J. Comput. Syst. Sci.*, 58(1):177–182, 1999. doi: 10.1006/jcss.1998.1602. URL http://dx.doi.org/10.1006/jcss.1998.1602.

Marcus Schaefer. Deciding the k-dimension is PSPACE-complete. In *Proceedings of the 15th Annual IEEE Conference on Computational Complexity, Florence, Italy, July 4-7, 2000*, pages 198–203, 2000. doi: 10.1109/CCC.2000.856750. URL http://dx.doi.org/10.1109/CCC.2000.856750.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16 (2):264–280, 1971. doi: 10.1137/1116025. URL http://dx.doi.org/10.1137/1116025.

Appendix A. Quasi-polynomial Algorithm for Littlestone's Dimension

In this section, we provides the following algorithm which decides whether $\text{L-dim}(\mathcal{C},\mathcal{U}) \leqslant d$ in time $O(|\mathcal{C}| \cdot (2|\mathcal{U}|)^d)$. Since we know that $\text{L-dim}(\mathcal{C},\mathcal{U}) \leqslant \log |\mathcal{C}|$, we can run this algorithm for all $d \leqslant \log |\mathcal{C}|$ and compute Littlestone's Dimension of $\mathcal{C},\mathcal{U}$ in quasi-polynomial time.

Theorem 32 (Quasi-polynomial Time Algorithm for Littlestone's Dimension) *There is an algorithm that, given a universe $\mathcal{U}$, a concept class $\mathcal{C}$ and a non-negative integer d, decides whether $\text{L-dim}(\mathcal{C},\mathcal{U}) \leqslant d$ in time $O(|\mathcal{C}| \cdot (2|\mathcal{U}|)^d)$.*

Proof Our algorithm is based on a simple observation: if an element x belongs to at least one concept and does not belong to at least one concept, the maximum depth of mistake trees rooted at x is exactly $1 + \min\{\text{L-dim}(\mathcal{C}[x \to 0],\mathcal{U}), \text{L-dim}(\mathcal{C}[x \to 1],\mathcal{U})\}$. Recall from Section 4 that $\mathcal{C}[x \to 0]$ and $\mathcal{C}[x \to 1]$ denote the collection of concepts that exclude x and the collection of concepts that include x respectively.

This yields the following natural recursive algorithm. For each $x \in \mathcal{U}$ such that $\mathcal{C}[x \to 0], \mathcal{C}[x \to 1] \neq \emptyset$, recursively run the algorithm on $(\mathcal{C}[x \to 0],\mathcal{U}, d-1)$ and $(\mathcal{C}[x \to 1],\mathcal{U}, d-1)$. If both executions return NO for some x, then output NO. Otherwise, output YES. When $d = 0$, there is no need for recursion as we can just check whether $|\mathcal{C}| \leqslant 1$.

Finally, we note that the running time can be easily proved by induction on d. ■