Agnostic Learning by Refuting*

Pravesh K. Kothari[1] **and Roi Livni**[2]

1 **Princeton University and Institute of Advanced Study, Princeton, NJ, USA**
kothari@cs.princeton.edu

2 **Princeton University, Princeton, NJ, USA**
rlivni@cs.princeton.edu

Abstract

The sample complexity of learning a Boolean-valued function class is precisely characterized by its Rademacher complexity. This has little bearing, however, on the sample complexity of *efficient* agnostic learning.

We introduce *refutation complexity*, a natural computational analog of Rademacher complexity of a Boolean concept class and show that it exactly characterizes the sample complexity of *efficient* agnostic learning. Informally, refutation complexity of a class $\mathcal{C}$ is the minimum number of example-label pairs required to efficiently distinguish between the case that the labels correlate with the evaluation of some member of $\mathcal{C}$ (*structure*) and the case where the labels are i.i.d. Rademacher random variables (*noise*). The easy direction of this relationship was implicitly used in the recent framework for improper PAC learning lower bounds of Daniely and co-authors [6, 8, 10] via connections to the hardness of refuting random constraint satisfaction problems. Our work can be seen as making the relationship between agnostic learning and refutation implicit in their work into an explicit equivalence. In a recent, independent work, Salil Vadhan [25] discovered a similar relationship between refutation and PAC-learning in the realizable (i.e. noiseless) case.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases learning theory, computational learning, Rademacher complexity

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.55

1 Introduction

Statistical complexity characterizes the information theoretic threshold for the amount of data required for any supervised learning task. However, the amount of data required for *efficient* learning, whenever it is possible, can be significantly different from the statistical complexity. For example, algorithms based on polynomial regression ([18, 19, 20]) guarantee efficient (improper, i.e. return a hypothesis not necessarily in the target class) learning while using data that is a polynomial factor larger than the statistical complexity. There is a systematic effort to study the trade-offs between computational and statistical complexity [4, 5] and a growing body of work has provided explicit examples [11, 7, 2] of natural settings where efficient learning provably requires data that is at least a polynomial factor larger than the statistical complexity under some plausible complexity theoretic assumptions.

In the light of the above work, we focus on obtaining a simple and useful characterization of the sample complexity of *efficient* supervised learning. There's a simple and elegant char-

[*] This research was supported by funding from Eric and Wendy Schmidt Fund for Strategic Innovation.

9th Innovations in Theoretical Computer Science Conference (ITCS 2018).
Editor: Anna R. Karlin; Article No. 55; pp. 55:1–55:10
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

acterization of the statistical complexity of learning in terms of the Rademacher complexity [1]. In this note, we give a natural analog of Rademacher complexity that precisely characterizes the amount of data required for *efficient agnostic* (i.e. noisy, see Definition 2) learning.

For a class $\mathcal{C}$ of concepts on $\mathbb{R}^n$, any distribution $\mathcal{D}$ on $\mathbb{R}^n$, the *Rademacher Complexity* of $\mathcal{C}$, $\mathcal{R}_m(\mathcal{C})$ is the following quantity:

$$\mathcal{R}_m(\mathcal{C}) = \mathop{\mathbb{E}}_{\substack{x_i \sim_{i.i.d.} \mathcal{D} \\ 1 \leqslant i \leqslant m}} \left[\mathop{\mathbb{E}}_{\substack{\sigma_i \sim_{i.i.d.} \{\pm 1\} \\ 1 \leqslant i \leqslant m}} \left[\frac{1}{m} \sup_{c \in \mathcal{C}} \sum_{i=1}^{m} \sigma_i c(x_i) \right] \right]. \tag{1.1}$$

Classical results [3] establish that $\mathcal{R}_m(\mathcal{C}) = \varepsilon$ if and only if there's an algorithm to learn $\mathcal{C}$ over $\mathcal{D}$ with error at most ε with $\Theta(m)$ samples, thus characterizing the sample-complexity of ε-error agnostic learning.

In this note, we propose a natural computational analog of Rademacher complexity, called as the *Refutation complexity* and show that it exactly determines the sample complexity of efficient agnostic learning. Given random labeled examples $\{(x_i, y_i)\}_{i \leqslant m}$ where x_is are chosen i.i.d. according to $\mathcal{D}$, we define the problem of *refutation* as the task of distinguishing between the following two cases:

(a) **Structure:** $\{(x_i, y_i)\}_{i \leqslant m}$ are i.i.d. from some distribution $\mathcal{D}'$ with marginal on x_is being $\mathcal{D}$ and $\mathbb{E}_{(x,y)\sim\mathcal{D}'}[c(x)y] = \Omega(1)$. That is, the given example-label pairs come from a distribution that correlates with some $c \in \mathcal{C}$, and

(b) **Noise:** y_is are uniform and independent Rademacher random variables.

We define refutation complexity of $\mathcal{C}$ with respect to the distribution $\mathcal{D}$ at a running time of $T(n)$ as the smallest m for which there's a $T(n)$-time test for distinguishing between structure and noise cases above.

To motivate this definition, observe that we can interpret the statistical complexity (via the connection to Rademacher complexity outlined above) of $\mathcal{C}$ over $\mathcal{D}$ as the smallest m for which no concept in $\mathcal{C}$ correlates with purely random noise (the i.i.d. draws from $\{\pm 1\}$.) Thus, if the Rademacher complexity of $\mathcal{C}$ on $\mathcal{D}$ with m samples is small enough, then, given random labeled examples $\{(x_i, y_i)\}_{i \leqslant m}$, we can (via an inefficient procedure) distinguish between the above two cases by computing the largest correlation of any $c \in \mathcal{C}$ when evaluated at x_is with the y_is. Thus, we can equivalently define statistical complexity as the smallest m for which the above structure vs noise test succeeds. Thus, refutation complexity can be seen as a computational analog of Rademacher complexity.

The main result of this note is the following theorem:

▶ **Theorem 1** (Refutation Complexity = Agnostic Learning Complexity, Informal)**.** *$\mathcal{C}$ has an efficient agnostic learning algorithm over a distribution $\mathcal{D}$ with m samples if and only if the refutation complexity of $\mathcal{C}$ at some polynomial running time is at most $O(m)$.*

1.1 Comparison with [25]

In a recent, independent work, Vadhan [25] used similar arguments to establish a similar equivalence to Theorem 1 between *distribution independent PAC learning* in the realizable case (i.e. when the labels perfectly correlate with some concept in the target class) and a slightly different notion of refutation. In this notion, the refutation algorithm is required to distinguish the case that the sample that realizable (i.e., the labels agree with some concept

[1] The related notion of VC Dimension of $\mathcal{C}$ characterizes the data required to learn $\mathcal{C}$ over *worst-case* distributions.

from the class) from the case that the labels in the sample are i.i.d. Rademacher random variables.

Since agnostic learning is provably different from realizable PAC learning in general, the notions of refutation that characterize the complexity of learning in the two models have to be necessarily different. Another interesting point of difference is that our equivalence is *distribution-specific* and thus slightly more fine-grained in that it allows relating learnability on a given distribution to refutation on the same distribution. In contrast, Vadhan's characterization holds for distribution independent PAC learning. This difference arises entirely due to the the difference in the black-box boosting algorithms one can use in PAC vs agnostic settings[2]: in the PAC learning case, the boosting algorithms modify the distribution of examples over the course of the execution and thus the characterization holds only in a distribution independent setting. In the agnostic setting, there are distribution specific boosting algorithms (such as that of [17, 14]) that work by changing only the distributions of the labels while keeping the distribution of the example points unchanged. It is an interesting direction to investigate notions of refutation that allow *distribution-specific* characterization of PAC learning in realizable case.

It's interesting to note how slight changes to in the formulation of the refutation problem changes the model of learning that it characterizes.

1.2 Discussion

Proper vs Improper Learning and the Framework of [9]

The agnostic learning algorithm we obtain using a refutation algorithm is *improper* - that is, it doesn't necessarily produce a hypothesis from the class $\mathcal{C}$. This is not accidental - it's well known that the flexibility of *improper* learning allows circumventing computational hardness results that afflict *proper* learning. A simple example is the class of 3-term DNF formulas in n variables: unless RP = NP, there's no polynomial time *proper* learning algorithm for this class [21], however, there's a simple poly$(n, 1/\varepsilon)$-time *improper* learning algorithm for it (for a discussion see, [24]). On the flip side, the power of *improper* learning makes the task of proving lower bound against such algorithms harder. The equivalence between refutation and agnostic learning holds for all (and thus, also improper) learning algorithms and thus can serve as a useful handle in understanding the complexity of improper learning.

Indeed this connection and in particular, the implication that learning implies refutation is implicit in the influential work of Daniely and co-authors [6, 10, 8] who showed (in the language of this paper) that a refutation algorithm for the concept classes of halfspaces and DNF formulas can be used to refute certain random constraint satisfaction problems [1, 12]. These works used such a reduction along with standard hardness assumptions for refuting random CSPs to obtain the first hardness results for improper PAC learning for the above classes.

Our equivalence establishes the converse of the connection in these works and makes the connection between refutation and agnostic learning explicit. While a priori, it might appear that refutation (which asks for distinguishing between a pure noise in the labels from a correlated set of labels) is easier than agnostic learning, this work shows that any lower bound on (improper) learning has to necessarily be a lower bound for an associated refutation problem. Thus, to an extent, it shows that the above framework for improper agnostic learning lower bounds is essentially complete.

[2] We thank Salil Vadhan for pointing this out to us.

Connections to Boosting/Property Testing

It is also illuminating to view the equivalence we show as saying that an oracle for refutation is sufficient for agnostic learning. This naturally leads to the question of what kind of oracle access to $\mathcal{C}$ is sufficient for (agnostic) learning. We discuss two natural oracles here: a weak-learning oracle and a property-testing oracle.

Known boosting (see [15, 23], [17, 14]) algorithms imply that a *weak-learning oracle* is sufficient for agnostically learning of $\mathcal{C}$. A weak learning oracle takes random example-label pairs and returns a hypothesis whose correlation with the labels from the input distribution is at least an inverse polynomial fraction of the correlation of the best-fitting hypothesis from $\mathcal{C}$. In learning literature, this is sometimes referred to as a *weak-optimization* oracle for $\mathcal{C}$ - in that, it gives a inverse polynomial (potentially improperly) approximation to the correlation of the best fitting hypothesis from $\mathcal{C}$. It is not hard to see that such an oracle is enough to solve the refutation problem and thus is a potentially stronger access to $\mathcal{C}$ than the refutation algorithm.

Our result implies that an much weaker algorithm is enough to get an agnostic learning algorithm - the refutation oracle doesn't return any hypothesis, it "merely" distinguishes between the case that the labels are completely random and independent of the examples from the case that the labels come from some distribution that correlates with some concept in $\mathcal{C}$.

It is also instructive to compare a refutation oracle (or a "structure" vs "noise" tester) for $\mathcal{C}$ with a "property-tester" for $\mathcal{C}$. An α-approximate property-testing algorithm for $\mathcal{C}$ uses random example-label pairs [3] from some distribution and accepts if the labels achieve a correlation of at least α with $\mathcal{C}$ and rejects if tevery $c \in \mathcal{C}$ has a correlation of at most $\alpha - \varepsilon$ with the labels. We can interpret a property tester, thus, as a variant of the refutation oracle that must treat a distribution on example-label pairs that has a correlation of at most $\alpha - \varepsilon$ with every $c \in \mathcal{C}$ as "noise." In particular, the notion of what is "unstructured/noise" for a property tester is more stringent compared to a refutation algorithm. Indeed, this is not surprising: while testing is known to be no harder than *proper* learning, it can be harder than *improper* learning for some concept classes, once again illustrating the difference between proper and improper learning [16].

Using Refutation to get Learning Algorithms

It will be extremely interesting to understand if the equivalence between refutation and learning allows an application in the direction opposite to the one employed in the work of Daniely and co-authors and get new algorithms for agnostic learning. This is perhaps not too optimistic. The works of Daniely and co-authors establish a natural connection between the refutation problem for a concept class and refuting random CSPs. There are known algorithms for refuting random CSPs (see for e.g. [22, 13, 1]) that use techniques that appear different from the usual tool-kit in agnostic learning (for e.g. the use of semi-definite programming) that might prove useful in obtaining new agnostic learning algorithms by building the required refutation algorithms.

[3] Property testers are usually defined with $\alpha = 1$ and are in general also allowed to use membership queries. We use a definition that is similar in spirit but is more relevant for the comparison here.

1.3 Proof Overview

It is easy to see that efficient learning implies efficient refutation. For the other direction, we give an explicit, efficient algorithm that invokes the refutation algorithm a small number of times to get an agnostic learner for the class $\mathcal{C}$. This algorithm works in two steps - in the first step, it uses a refutation algorithm to come up with a *weak-agnostic* learner: i.e. a hypothesis that achieves a correlation with the labels that is some tiny fraction of the correlation of the best hypothesis from $\mathcal{C}$. In the second step, it combined an off-the-shelf boosting algorithm with the weak learner above to get an agnostic learner with small error.

The key idea in the transformation of a refutation algorithm into a weak-learner is to view the black-box refutation algorithm as a "code" for computing a function by manipulating the example-label pairs that it takes as input. A simple hybrid argument then shows that there's a small list of hypotheses generated by manipulating the inputs to the refutation algorithm that contains a good weak learner. We can find the best weak learner from the list by evaluating the error of each of the hypotheses in the list over a fresh batch of samples from the underlying distribution.

1.4 Preliminaries

We use $\mathcal{U}_m$ to denote the uniform distribution over $\{\pm 1\}^m$ for any $m \in \mathbb{N}$. We define agnostic learning here.

▶ **Definition 2** (Agnostic Learning with respect to a distribution $\mathcal{D}$)**.** Let $\mathcal{C}$ be a class of Boolean concepts $\mathcal{C} \subseteq \{f : \{\pm 1\}^n \to \{\pm 1\}\}$. $\mathcal{C}$ is said to be ε-agnostically learnable in time $T(n, 1/\varepsilon)$ and samples $S(n, 1/\varepsilon)$ if there's an algorithm $\mathcal{A}$ running in time $T(n, 1/\varepsilon)$ that takes $S(n, 1/\varepsilon)$ random labeled examples $\{(x_i, y_i) \mid 1 \leqslant i \leqslant m\}$ where (x_i, y_i)s are i.i.d. from $\mathcal{D}'$, such that the marginal on x_i is $\mathcal{D}$ and outputs with probability at least 3/4, a hypothesis $h : \{\pm 1\}^n \to \{\pm 1\}$ such that $\mathbb{E}_{(x,y)\sim\mathcal{D}'} [\mathbf{1}[h(x) \neq y]] \leqslant \inf_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}'} [\mathbf{1}[c(x) \neq y]] + \varepsilon$.

2 Refutation Complexity

In this section, we define refutation complexity of a class of hypothesis with respect to a distribution $\mathcal{D}$.

▶ **Definition 3** (Refutation Algorithm for Distribution $\mathcal{D}$)**.** Let $\mathcal{C} \subseteq \{f : \mathbb{R}^n \to \{\pm 1\}\}$ be a class of Boolean concepts. Let $\mathcal{D}$ be a distribution on $\mathbb{R}^n$.

A δ-*refutation* algorithm $\mathcal{A}$ for $\mathcal{C}$ on $\mathcal{D}$ with $m = m(n)$ samples is a (possibly randomized) algorithm that takes input an m-tuple of points $\{x_1, x_2, \ldots, x_m\} \subseteq \{\pm 1\}^n$ and an m-tuple of labels $(\sigma_1, \sigma_2, \ldots, \sigma_m) \in \{\pm 1\}^m$ and outputs either noise or structure with the following guarantees:

1. **Completeness:** If $\{(x_i, \sigma_i)\}_{i\leqslant m}$ are i.i.d. from a distribution $\mathcal{D}'$ on $\mathbb{R}^n \otimes \{\pm 1\}$ such that the marginal on $\mathbb{R}^n$ equals $\mathcal{D}$ and $\sup_{c\in\mathcal{C}} \mathbb{E}_{(x,\sigma)\sim\mathcal{D}'}[c(x)\sigma] \geqslant \delta$, then,

$$\mathop{\mathbb{P}}_{\substack{\{(x_i,y_i)\}_{i\leqslant m}\sim_{i.i.d.}\mathcal{D}' \\ \text{internal randomness of } \mathcal{A}}} [\text{ output} = \mathsf{structure}] \geqslant 2/3.$$

2. **Soundness:**

$$\mathop{\mathbb{P}}_{\substack{(\sigma_1,\sigma_2,\ldots,\sigma_m)\sim\mathcal{U}_m \\ x_1,x_2,\ldots,x_m\sim\mathcal{D} \\ \text{internal randomness of } \mathcal{A}}} \mathbb{P}[\text{ output} = \mathsf{noise}] \geqslant 2/3.$$

▶ **Definition 4** (δ-Refutation Complexity). Let $\mathcal{C} \subseteq \{f : \{\pm 1\}^n \to \{\pm 1\}\}$ be a class of Boolean concepts. Let $\mathcal{D}$ be a distribution on $\{\pm 1\}^n$.

The δ-*refutation complexity* of $\mathcal{C}$ on a distribution $\mathcal{D}$ with running time $T(n)$ denoted by $\mathcal{R}_{T(n),\delta}(\mathcal{C})$, is the smallest $m = m(n, \delta)$ such that there exists a δ-refutation algorithm for $\mathcal{C}$ on $\mathcal{D}$ running in time $T(n)$ and m-samples. When $T(n)$ is not stated explicitly, we assume $T(n) = \mathrm{poly}(n)$ for some fixed polynomial in n.

▶ Remark. Observe that the refutation complexity, just as Rademacher complexity is distribution dependent. Further, for $T(n) = \infty$, δ-refutation complexity degenerates into Rademacher complexity. At non-trivially bounded running times (of special interest, of course, is polynomial time algorithms), refutation complexity captures the sample complexity of *efficient* agnostic, improper learning $\mathcal{C}$ over $\mathcal{D}$ as we show next and thus can be much larger than the Rademacher complexity.

3 Learning vs Refutation Complexity

In this section, we establish the equivalence between agnostic learning a class $\mathcal{C}$ over a given distribution $\mathcal{D}$ and the refutation problem with respect to the distribution $\mathcal{D}$ for the concept class $\mathcal{C}$.

We begin by showing the Learning implies Refutation, which is the easy direction.

▶ **Lemma 5** (Learning implies Refutation). *Suppose $\mathcal{C}$ is ε-agnostically learnable in time $T(n, \varepsilon)$ and samples $S(n, \varepsilon)$ over the distribution $\mathcal{D}$. Then, the refutation complexity of $\mathcal{C}$ with respect to the distribution $\mathcal{D}$ at the running time $T(n, \delta/4)$ is at most $2S(n, \delta/4) + 128/\delta^2$.*

Proof. Let $m = S(n, \delta/4) + 64/\delta^2$.

The δ-refutation algorithm gets input $x_1, x_2, \ldots, x_{2m}$ and $\sigma_1, \sigma_2, \ldots, \sigma_{2m}$. It runs the ε-agnostic learner on examples $\{(x_i, \sigma_i)\}_{i=1}^m$ for $\varepsilon = \delta/4$ and obtains a hypothesis h. Let $\mathrm{cor}_h = \frac{1}{m} \sum_{i=m+1}^{2m} \sigma_i \cdot h(x_i)$. If $\mathrm{cor}_h \geqslant \delta/2$, output structure otherwise output noise.

We now analyze the completeness and the soundness properties of this algorithm.

First, suppose $\{(x_i, \sigma_i)\}_{i \leqslant 2m}$ were i.i.d. according to some $\mathcal{D}'$ such that the marginal on $\mathbb{R}^n$ equals $\mathcal{D}$. Let $\mathrm{cor}_f(\mathcal{D}') = \mathbb{E}_{(x,y) \sim \mathcal{D}'}[f(x)y]$. Then, with probability 2/3 over the draw of the sample, the agnostic learner produces a hypothesis h such that $\mathrm{cor}_h \geqslant \mathrm{cor}_h(\mathcal{D}') - \varepsilon \geqslant \mathrm{cor}_c(\mathcal{D}') - 2\varepsilon$ for every $c \in \mathcal{C}$. Thus, if $\mathrm{cor}_c(\mathcal{D}') \geqslant \delta$, then, $\mathrm{cor}_h \geqslant \delta - \varepsilon/2 \geqslant \delta/2$. Thus, in this case, the algorithm above outputs structure as desired.

Now suppose σ_is are i.i.d. Rademacher and independent of x_is. Then, since $\sigma_{m+1}, \ldots, \sim_{2m}$ are independent of $\sigma_1, \ldots, \sigma_m$, $\mathrm{cor}_h \leqslant \frac{4}{\sqrt{m}} < \delta/2$ using that $m > 64/\delta^2$. ◀

▶ **Lemma 6** (Learning by Refutation). *Suppose that the δ-refutation complexity of a class of Boolean concepts $\mathcal{C}$ with respect to a distribution $\mathcal{D}$ at a running time $T(n)$ is $m = \mathcal{R}_{T(n),\delta}(\mathcal{C})$. Then, there's an algorithm that runs in time $T(n)\frac{m^2}{\varepsilon^2}$ and uses $O(\frac{m^3}{\varepsilon^2})$ samples to $(\delta + \varepsilon)$-agnostically learn $\mathcal{C}$ on $\mathcal{D}$.*

The proof is in two steps. In the first step, we show that the refutation algorithm yields a weak agnostic learner for $\mathcal{C}$ with respect to the distribution $\mathcal{D}$. In the second step, we use the distribution specific agnostic boosting algorithm (see [17]) to boost the accuracy of the weak learner to obtain an agnostic learner. We start by defining a weak-agnostic learner :

▶ **Definition 7** (Weak Agnostic Learner). An (γ, α)-weak agnostic learner for a Boolean concept class $\mathcal{C}$ over a distribution $\mathcal{D}$ is an algorithm that takes input random examples from a distribution $\mathcal{D}'$ on example-label pairs (x, y) such that the marginal on x is $\mathcal{D}$ such that with probability at least 3/4 over its random input outputs a (randomized) hypothesis $h : \{\pm 1\}^n \to \{\pm 1\}$ such that $\mathbb{E}_{(x,y)\sim\mathcal{D}'}[y \cdot h(x)] \geqslant \gamma(\sup_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}'}[y \cdot c(x)]) - \alpha$.

▶ **Lemma 8** (Refutation to Weak Agnostic Learner). *Suppose that the δ-refutation complexity of a class of Boolean concepts $\mathcal{C}$ with respect to a distribution $\mathcal{D}$ at a running time $T(n)$ is $m = \mathcal{R}_{T(n),\delta}(\mathcal{C})$. Then, there's an (γ, α)-weak agnostic learner for $\mathcal{C}$ on distribution $\mathcal{D}$ that runs in time $T(n)$ and samples $m(n)$ where $\alpha = \delta \cdot \gamma$, $\gamma = \frac{2}{3m}$.*

We describe a natural class of candidates for a weak learner that come out of running the refutation algorithm on appropriately chosen hybrids of the distribution $\mathcal{D}'$ and $\mathcal{D} \times \mathcal{U}_1$. We begin by defining a class of $2(m+2)$ different functions denoted by $W_{i,b} : \{\pm 1\}^n \to \{0, 1\}$ for $0 \leqslant i \leqslant m+1$ and $b \in \pm 1$ produced by taking these hybrids. Our weak learners will be a simple transformation of this class.

```
Algorithm 1 Hybrid Functions $W_{i,b}$
Input: $x \in \mathbb{R}^n$, $b \in \{\pm 1\}$.
Output: $W_{i,b}(x) = z \in \{\pm 1\}$.
Operation:
  1. Draw $(x_1, \sigma_1), \ldots, (x_{i-1}, \sigma_{i-1})$ i.i.d. from $\mathcal{D} \times \mathcal{U}_1$.
     Draw $(x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2}) \ldots, (x_m, y_m)$ i.i.d. from $\mathcal{D}'$.
  2. Run the $\delta$-refutation algorithm on input
     $(x_1, \sigma_1), (x_2, \sigma_2), \ldots, (x_{i-1}, \sigma_{i-1}), (x, b), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)$.
  3. Let $W_{i,b} = 1$ if the refutation algorithm returns structure and 0 otherwise.
```

We make some simple observations about $W_{i,b}$ that will come handy in the argument below.

Observe that $W_{m+1,b}$ is the function that evaluates to 1 if the output of the refutation algorithm on examples drawn from $\mathcal{D}$ and labels i.i.d Rademacher variables is structure. On the other hand, $W_{0,b}$ is the function obtained when the refutation algorithm is run on example-label pairs from $\mathcal{D}$. Finally, observe that

$$\mathop{\mathbb{E}}_{b\sim\mathcal{U}_1} \mathbb{E}[W_{i,b}(x)] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathbb{E}[W_{i+1,y}(x)] \tag{3.1}$$

Here, the inside expectation is over all the random choices within the procedure for computing $W_{i,b}$s above. We can now present our candidate weak learners.

Candidate Weak Learners

For every $0 \leqslant i \leqslant m+1$, let $h_i(x) = W_{i,+1}(x) - W_{i,-1}(x)$.

Proof of Lemma 8. Our weak learning algorithm is given access to random labeled examples from a distribution $\mathcal{D}'$ on $\mathbb{R}^n \otimes \{\pm 1\}$. The weak learner will draw a sample from $\mathcal{D}'$ of size $O(\log m)$ from $\mathcal{D}'$ and chooses the h_i that has the maximum correlation with the labels. Observe that with $O(\log{(m)})$ samples, the correlations of h_i on $\mathcal{D}'$ will be faithfully preserved with 2/3 probability. Thus, to complete the proof, we only need to argue that one of the h_is is always an (α, γ)-weak learner.

To show this, we must argue that there exists an $0 \leqslant i \leqslant m+1$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}'}[y \cdot h_i(x)] \geqslant \frac{2}{3m} \sup_{c\in\mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}'}[c(x) \cdot y] - \frac{2}{3m}\delta.$$

Observe that the guarantees of the weak learner are trivial if $\sup_{c \in \mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}'}[y \cdot c(x)] < \delta$. Thus assume that $\sup_{c \in \mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}'}[y \cdot c(x)] > \delta$. In this case, we will show that $\mathbb{E}_{(x,y)\sim\mathcal{D}'}[h_i(x)y] \geqslant \frac{2}{3m} \geqslant \frac{2}{3m} \sup_{c \in \mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}'}[c(x) \cdot y] - \frac{2}{3m}\delta$.

Now, observe that over the randomness of both the refutation algorithm and over the draw of i.i.d. sample from $\mathcal{D}'$ of size $m = S(n)$, $\mathbb{E}[W_{0,b}(x)] \geqslant 2/3$ and $\mathbb{E}[W_{m+1,b}(x)] \leqslant 1/3$ for any b. Thus,

$$\sum_{i=0}^{m} \mathbb{E}[W_{i,y}(x) - W_{i+1,y}(x)] \geqslant 1/3,$$

where the expectation is over the randomness in the draw $(x, y) \sim \mathcal{D}'$ and over the randomness in $W_{i,y}$ for $0 \leqslant i \leqslant m+1$.

Thus, there must exist an i such that $\mathbb{E}[W_{i,y}(x) - W_{i+1,y}(x)] > 1/3m$. Observe that by construction

$$\begin{aligned} W_{i,y}(x) = \frac{y+1}{2} \cdot W_{i,1}(x) - \frac{y-1}{2} W_{i,-1}(x) = y \cdot \frac{W_{i,1}(x) - W_{i,-1}(x)}{2} + \frac{1}{2}(W_{i,1}(x) + W_{i,-1}(x)) \\ = \frac{1}{2} y \cdot h_i(x) + \frac{1}{2}(W_{i,1}(x) + W_{i,-1}(x)). \quad (3.2) \end{aligned}$$

Next, observe that by (3.1), $\mathbb{E}[\frac{1}{2}(W_{i,1}(x) + W_{i,-1}(x))] = \mathbb{E}[W_{i+1,y}(x)]$. Taking expectations on both sides of (3.2) and rearranging, we have: $\mathbb{E}[y \cdot h_i(x)] \geqslant \frac{2}{3m}$.

This establishes that for $\gamma = \frac{2}{3m}$ and $\alpha = \delta \cdot \gamma$ our algorithm returns (α, γ)-weak agnostic learner as desired. ◀

We can now use boosting to get a strong agnostic learner for $\mathcal{C}$ over $\mathcal{D}$ by using the weak learning algorithm along with a boosting algorithm. Specifically, we will use the result of Kalai and Kanade [17] (see also [14]) who showed the following agnostic boosting algorithm that takes a (γ, α)-weak learner and outputs a hypothesis whose error is competitive within α with respect to the best fitting hypothesis from the class $\mathcal{C}$.

▶ **Fact 9** (Agnostic Boosting [17]). *Let $\mathcal{C}$ be a class of Boolean concepts. Let $\mathcal{D}$ be a distribution on $\{\pm 1\}^n$ and $\varepsilon > 0$.*

There's an algorithm that takes random labeled examples from a distribution $\mathcal{D}'$ on example-label pairs (x, y) such that the marginal on x is $\mathcal{D}$, invokes a (γ, α)-weak learner for $\mathcal{C}$ $O(\frac{1}{\gamma^2 \varepsilon^2})$ times and outputs a hypothesis $h : \{\pm 1\}^n \to \{\pm 1\}$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}'}[\mathbf{1}[h(x) \neq y]] \leqslant \inf_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}'}[\mathbf{1}[c(x) \neq y]] + \alpha/\gamma + \varepsilon.$$

The algorithm needs $S(n) \cdot O(\frac{1}{\gamma^2 \varepsilon^2})$ samples and runs in time $T(n) \cdot O(\frac{1}{\gamma^2 \varepsilon^2})$ where $S(n)$ and $T(n)$ are the sample complexity and the running time respectively of the (γ, α)-weak agnostic learner.

We get Lemma 6 as an immediate corollary of Fact 9 and Lemma 8.

Acknowledgements. We thank Salil Vadhan for sharing an early version of [25] with us and illuminating follow-up discussions. P.K. thanks Avi Wigderson for many useful comments and suggestions about this work and David Steurer for helpful discussions on related problems.

References

1 Sarah R. Allen, Ryan O'Donnell, and David Witmer. How to refute a random CSP. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 689–708. IEEE Computer Soc., Los Alamitos, CA, 2015.

2 Boaz Barak and Ankur Moitra. Tensor prediction, rademacher complexity and random 3-xor. *CoRR*, abs/1501.06521, 2015.

3 Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL: `http://www.jmlr.org/papers/v3/bartlett02a.html`.

4 Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *CoRR*, abs/1304.0828, 2013.

5 Venkat Chandrasekaran and Michael I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013. `doi:10.1073/pnas.1302293110`.

6 Amit Daniely. Complexity theoretic limitations on learning halfspaces. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 105–117. ACM, 2016. `doi:10.1145/2897518.2897520`.

7 Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. More data speeds up training time in learning halfspaces over sparse vectors. In *NIPS*, pages 145–153, 2013.

8 Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *STOC*, pages 441–448. ACM, 2014.

9 Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 441–448. ACM, 2014.

10 Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In *COLT*, pages 815–830, 2016.

11 Scott E. Decatur, Oded Goldreich, and Dana Ron. Computational sample complexity. *SIAM J. Comput.*, 29(3):854–879, 1999.

12 Uriel Feige. Relations between average case complexity and approximation complexity. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 534–543. ACM, 2002. `doi:10.1145/509907.509985`.

13 Uriel Feige. Refuting smoothed 3cnf formulas. In *FOCS*, pages 407–417. IEEE Computer Society, 2007.

14 Vitaly Feldman. Distribution-specific agnostic boosting. In Andrew Chi-Chih Yao, editor, *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 241–250. Tsinghua University Press, 2010. URL: `http://conference.itcs.tsinghua.edu.cn/ICS2010/content/papers/20.html`.

15 Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

16 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

17 Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 880–888. Curran Associates, Inc., 2009. URL: `http://papers.nips.cc/paper/3676-potential-based-agnostic-boosting`.

18 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`.

19 Daniel M. Kane, Adam R. Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 522–545. JMLR.org, 2013. URL: `http://jmlr.org/proceedings/papers/v30/Kane13.html`.

20 Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 541–550. IEEE Computer Society, 2008. `doi:10.1109/FOCS.2008.64`.

21 Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *J. ACM*, 35(4):965–984, 1988.

22 Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random csps below the spectral threshold. *CoRR*, abs/1605.00058, 2016.

23 Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990. `doi:10.1007/BF00116037`.

24 Shai Shalev-Shwartz, Ohad Shamir, and Eran Tromer. Using more data to speed-up training time. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pages 1019–1027. JMLR.org, 2012. URL: `http://jmlr.csail.mit.edu/proceedings/papers/v22/shalev-shwartz12.html`.

25 Salil Vadhan. On learning vs. refutation. In *Conference on Learning Theory*, pages 1835–1848, 2017.