Improper Learning Equals Refutation — Theory of Machine Learning

Rachit Nimavat

May 30, 2018

We study the equivalence between the notions of *efficient improper learning* and *efficient refutation* in the distribution specific setting. We generalize the approach of Kothari and Livni [1] to distribution specific PAC-learning, and obtain a result of Vadhan [2] as corollary; albeit with weaker, but nevertheless polynomial bound on running time.

1 Introduction

Sample complexity is the information theoretical threshold on the number of samples required to learn. There is a vast body of literature showing that VC-dimension is the correct measure of sample complexity for learning, and Rademacher complexity is the correct measure if we are looking at fine-grained distribution (and even, data) dependent sample complexity. However, the number of samples required to *'efficiently'* learn can be significantly different from sample complexity, depending on the notion of efficiency involved. An extreme (and incredibly surprising) instructive example of this is a result of Ran Raz [3] that shows that learning parity under uniform distribution requires exponential samples if allowed less than $n^2/25$ bits of space, while if allowed more than n^2 bits, can be done in linear number of samples by simple Gaussian elimination.

In this note, we relate distribution-dependent time efficiency of learning algorithms to 'refutation-complexity' of a class of hypothesis. This is an extension of the work of [1], and in flavor similar to [2]. Here, we do not make any effort to optimize time efficiency, and leave it at some polynomial running time to make our exposition simpler.

2 Preliminaries

Let $\mathcal{C} \subseteq \{f : \{\pm 1\}^n \mapsto \{\pm 1\}\}$ be a class of boolean concepts and $\mathcal{D}$ be a distribution over $\{\pm 1\}^n$ for some $n \in \mathbb{N}$. In the rest of the note, we fix $\mathcal{C}$, $\mathcal{D}$ and the underlying parameter n. Whenever we say *efficient* algorithms, refer to (potentially randomized) algorithms running in time $\text{poly}(n)$. The hypothesis that we consider are (potentially randomized) functions $h : \{\pm 1\}^n \mapsto \{\pm 1\}$. We refer by $\mathcal{U}$ the uniform distribution on $\{\pm 1\}$. An important theme is the notion of *extension* of $\mathcal{D}$ to incorporate labels of data-points distributed according to $\mathcal{D}$, that we define next.

Definition 1 *Let $\mathcal{D}'$ be an arbitrary distribution supported $\{\pm 1\}^n \times \{\pm 1\}$. We say that $\mathcal{D}'$ is an* extension *of $\mathcal{D}$ iff its marginal on first n coordinates is $\mathcal{D}$.*

Definition 2 *Given an extension $\mathcal{D}'$ of $\mathcal{D}$ and a hypothesis $h : \{\pm 1\}^n \mapsto \{\pm 1\}$, its error on $\mathcal{D}'$ is given by $\mathsf{err}_{\mathcal{D}'}(h) := \mathbf{E}_{(x,y)\sim\mathcal{D}'}\left[\mathbf{1}_{h(x)\neq y}\right]$ and its correlation by $\mathsf{cor}_{\mathcal{D}'}(h) := \mathbf{E}_{(x,y)\sim\mathcal{D}'}\left[h(x)\cdot y\right] = 1-2\cdot\mathsf{err}_{\mathcal{D}'}(h)$. The optimal error and correlation are denoted by $\mathsf{err}^*_{\mathcal{D}'} = \inf_{c\in\mathcal{C}} \mathsf{err}_{\mathcal{D}'}(c)$ and $\mathsf{cor}^*_{\mathcal{D}'} = \sup_{c\in\mathcal{C}} \mathsf{cor}_{\mathcal{D}'}(c)$.*

Next, we define the notions of distribution-specific agnostic learning and refutability.

Definition 3 *We say that $\mathcal{C}$ is ϵ-agnostically learnable if there is an efficient algorithm $\mathcal{L}$ with access to $\text{poly}(n)$ i.i.d labelled samples from* each *extension $\mathcal{D}'$ of $\mathcal{D}$, outputs with probability at least $3/4$, a hypothesis h such that $\mathsf{err}_{\mathcal{D}'}(h) \leq \mathsf{err}^*_{\mathcal{D}'} + \epsilon$.*

Definition 4 *We say that $\mathcal{C}$ is δ-refutable if there is an efficient algorithm $\mathcal{R}$ that with access to* $\mathrm{poly}(n)$ *i.i.d labelled samples from* each *extension $\mathcal{D}'$ of $\mathcal{D}$, outputs either* STRUCTURE *or* NOISE *with following guarantees:*

- *If $\mathsf{cor}^*_{\mathcal{D}'} \geq \delta$, then* $\mathbf{Pr}\left[\mathsf{STRUCTURE}\right] \geq 2/3$*; and*
- *If $\mathcal{D}' = \mathcal{D} \times \mathcal{U}$, (which implies, $\mathsf{cor}^*_{\mathcal{D}'} = 0$), then* $\mathbf{Pr}\left[\mathsf{NOISE}\right] \geq 2/3$.

Notice that the definition of refutability is computational analogue of Rademacher Complexity, which exactly characterizes sample complexity of agnostic learning. Our definition can be also contrasted with that of refuting random-CSPs, where the clauses come from some distribution, but the 'labels' are fixed. For instance, there is a known algorithm to efficiently refute random-3SAT formulas, given more than $n^{3/2}$ clauses. Our refutation is more general, in a sense that we allow adversarial labelings, that are allowed to not satisfy all the 'constraints'. Finally, our definition can also be contrasted with the notion of property testing, where NOISE is defined as $\mathsf{cor}^*_{\mathcal{D}'} = 0$, which is a subset of NOISE that we consider. It has been known that this property-testing analogue of distinguisher, at least for even weaker definitions of noise, can be *harder* than improper learning for some concept classes [6].

3 Main Result

Our result is the following theorem:

Theorem 3.1 *There is a $\epsilon' = 1/\mathrm{poly}(n)$ depending only on $\mathcal{C}$ and $\mathcal{D}$ such that $\mathcal{C}$ is ϵ-agnostically learnable on $\mathcal{D}$ iff $\mathcal{C}$ is $(2\epsilon + \epsilon')$-refutable.*

The theorem follows from the following two lemmata by appropriately setting the parameter ζ'. Their proofs are present in Sections 3.1 and 3.2 respectively.

Lemma 3.2 *If $\mathcal{C}$ is ϵ-agnostically learnable on $\mathcal{D}$, then $\mathcal{C}$ is $(2\epsilon + \zeta')$-refutable on $\mathcal{D}$ for any choice of $\zeta' > 1/\mathrm{poly}(n)$.*

Lemma 3.3 *If $\mathcal{C}$ is δ-refutable on $\mathcal{D}$, then $\mathcal{C}$ is $(\delta/2 - \zeta')$-agnostically learnable on $\mathcal{D}$ for some small enough $\zeta' = 1/\mathrm{poly}(n)$.*

3.1 From Learning to Refutation — Proof of Lemma 3.2

Fix an extension $\mathcal{D}'$ of $\mathcal{D}$. Our goal is to come up with a $2\epsilon + \zeta'$-refutation algorithm $\mathcal{R}$ for $\mathcal{C}$ on $\mathcal{D}'$ using ϵ-agnostic learning algorithm $\mathcal{L}$. We let $\zeta := \zeta'/2$. Let m be the number of samples required to agnostically learn using $\mathcal{L}$. If $m < 6/\zeta^2$, we set $m = 6/\zeta^2$.

The refutation algorithm $\mathcal{R}$ works as follows. First, it requests m i.i.d samples from $\mathcal{D}'$. It runs $\mathcal{L}$ on these samples and obtains a hypothesis h. It then requests m fresh i.i.d samples $\{(x_i, y_i) : i \in [m]\}$ and computes the empirical correlation $c = \frac{1}{m} \cdot \sum_{i=1}^{m} h(x_i) \cdot y_i$. If the computed empirical correlation $c \geq \zeta$, it outputs STRUCTURE; and NOISE otherwise. Claims 3.4 and 3.5 now complete the proof of the lemma 3.2.

Claim 3.4 *If $\mathcal{D}' = \mathcal{D} \times \mathcal{U}$, then $c < \zeta$ with probability at least* $2/3$.

Proof: Notice that for the latter half of the algorithm, h is a fixed hypothesis. Thus, since each y_i is a uniform random variable in $\{\pm 1\}$, $h(x_i) \cdot y_i$ is also a uniform random variable in $\{\pm 1\}$. Notice

that if we flip the value of $h(x_i) \cdot y_i$ for any i, the value of c changes by at most $2/m$. Thus, blindly applying McDiarmid's inequality and using the fact that $m \geq 6/\zeta^2$, we obtain

$$\mathbf{Pr}\left[c \geq \zeta\right] \leq \exp\left(\frac{-2 \cdot \zeta^2}{m \cdot (2/m)^2}\right) = \exp(-\zeta^2 m/2) < 1/3.$$

□

Claim 3.5 *If $\mathsf{cor}^*_{\mathcal{D}'} \geq 2\epsilon + \zeta'$, then $c \geq \zeta$ with probability at least $2/3$.*

Proof: We know that $\mathsf{err}^*_{\mathcal{D}'} = (1 - \mathsf{cor}^*_{\mathcal{D}'})/2 \leq 1/2 - \epsilon - \zeta$. Thus, with probability at least $3/4$, we get a hypothesis h from $\mathcal{L}$ such that $\mathsf{err}_{\mathcal{D}'}(h) \leq \mathsf{err}^*_{\mathcal{D}'} + \epsilon \leq 1/2 - \zeta$. From bounded-range Chernoff bound and using the fact that $m \geq 6/\zeta^2$, we obtain

$$\mathbf{Pr}\left[c < \zeta\right] \leq \exp\left(\frac{-2 \cdot (\zeta m)^2}{m \cdot 4}\right) = \exp(-\zeta^2 m/2) < 1/12.$$

The claim now follows by taking union bound over both the above-mentioned bad events. □

3.2 From Refutation to Learning — Proof of Lemma 3.3

We will come up with a weak-learning algorithm $\mathcal{L}'$ and then using an off-the-shelf boosting algorithm, we will obtain $\mathcal{L}$ with claimed properties. We begin by defining *weak agnostic learning* algorithm, that can recover at least polynomial fraction of the edge over random guessing of the best hypothesis in $\mathcal{C}$.

Definition 5 *An efficient learning algorithm $\mathcal{L}'$ is called α-weak agnostic learner if given access to* $\mathrm{poly}(n)$ *i.i.d labelled samples from* each *extension $\mathcal{D}'$ of $\mathcal{D}$, outputs with probability at least $2/3$ a (potentially randomized) hypothesis $h : \{\pm 1\}^n \mapsto [-1, 1]$ such that $\mathsf{err}_{\mathcal{D}'}(h) \leq 1/2 - 1/\mathrm{poly}(n)$ whenever $\mathsf{err}^*_{\mathcal{D}'} \leq 1/2 - \alpha$.*

Theorem 3.6 (Agnostic Boosting [5]) *There is an efficient algorithm, that given access to a α-weak learner $\mathcal{L}'$ outputs a hypothesis h such that $\mathsf{err}_{\mathcal{D}'}(h) \leq \mathsf{err}^*_{\mathcal{D}'} + \alpha + \epsilon$ for any choice of $\epsilon > 1/\mathrm{poly}(n)$.*

Now we will present our $(\delta/2 - 2\tau')$-weak-learning algorithm $\mathcal{L}'$ for predicting the label y^{**} of challenge example x^{**} coming from an extension $\mathcal{D}'$ of $\mathcal{D}$. First, we will describe an algorithm to obtain a bunch of hypothesis, which will then be tested on a fresh sample S and then we will pick the best hypothesis and label x^{**} accordingly. Let $m = \mathrm{poly}(n)$ be the number of samples requested by $\mathcal{R}$, and we set $\tau' = 1/m^4$. If $\mathsf{err}^*_{\mathcal{D}'} > 1/2 - \delta/2 + 2\tau'$, then we have nothing to show. Thus, we assume from now on that $\mathsf{err}^*_{\mathcal{D}'} \leq 1/2 - \delta/2 + 2\tau'$ and hence, $\mathsf{cor}^*_{\mathcal{D}'} \geq \delta - 4\tau'$. Notice that if we we look at only $O(m^3)$ i.i.d examples from $\mathcal{D}'$, then with high probability, it is information theoretically impossible to distinguish $\mathcal{D}'$ from a distribution $\mathcal{D}''$ with $\mathsf{cor}^*_{\mathcal{D}''} \geq \delta$. Thus, we can assume from now on that $\mathsf{cor}^*_{\mathcal{D}'} \geq \delta$.

As a first step, we define a class of $2(m+2)$ hybrid functions obtained by running appropriately chosen hybrids of the distributions $\mathcal{D}'$ and $\mathcal{D} \times \mathcal{U}$. [1] For each $i \in \{0, \ldots, m+1\}$ and $b \in \{+1, -1\}$, we get the function $W_{i,b} : \{\pm 1\}^n \mapsto \{0, 1\}$ labelling x^* as follows. We draw $i - 1$ fresh i.i.d examples (x_j, y_j) from $\mathcal{D}'$. We let $x_i = x^*$. Finally, we draw $m - i$ fresh unlabeled examples x_j from $\mathcal{D}'$ and label them uniformly randomly from $\{\pm 1\}$. We run $\mathcal{R}$ on $\{(x_1, y_1), \ldots, (x_i, b), \ldots, (x_m, y_m)\}$, and let the output be 1 if $\mathcal{R}$ outputs $\mathsf{STRUCTURE}$ and 0 otherwise. Now, for each $i \in \{0, \ldots, m+1\}$, we get our weak learners $h_i(x^*) = W_{i,1}(x^*) - W_{i,-1}(x^*)$. Notice that $h_i : \{\pm 1\}^n \mapsto \{-1, 0, 1\}$.

Now, using an idea very similar to the beautiful hybrid argument of Yao [7], we get the following claim, whose proof is deferred to end of this section.

[1] The term 'hybrid of a pair of distributions' refers to a new distribution obtained by picking a few samples from the first distribution and remaining from the second one. It is the cornerstone of Yao's [7] hybrid argument which is used extensively in cryptography and pseudorandomness.

Claim 3.7 *There is a* $i \in \{0, \ldots, m+1\}$ *such that* $\mathsf{err}_{\mathcal{D}'}(h_i) \leq 1/2 - 1/3m$.

Finally, we draw $O(m)$ fresh i.i.d samples S from $\mathcal{D}'$ and for each $i \in \{0, \ldots, m+1\}$, we obtain a hypothesis h_i classifying each example in $S \cup x^{**}$. Notice that in this process we use $O(m^3)$ examples of $\mathcal{D}'$, and hence, our sneaky pretention that $\mathsf{cor}^*_{\mathcal{D}'} \geq \delta$ is warranted whp. We then pick the hypothesis h_{i^*} which empirically performs best on S. The learning algorithm $\mathcal{L}'$ labels x^{**} as $h_{i^*}(x^{**})$. From the deviation bounds $\mathsf{err}_{\mathcal{D}'}(h_{i^*}) \leq 1/2 - 1/4m$ whp, as a random sample of size $O(m)$ will faithfully preserve error bounds whp. Thus, the following claim follows.

Claim 3.8 $\mathcal{L}'$ *is a* $\delta/2 - 2\tau'$*-weak agnostic learning algorithm.*

Now invoking Theorem 3.6, Lemma 3.3 follows after setting $\epsilon = \tau'$. The rest of the section is devoted to prove Claim 3.7.

Proof of Claim 3.7. Let $(x^{**}, y^{**}) = (x, y)$ be the challenge example. From the guarantees of our refutation algorithm $\mathcal{R}$, $\mathbf{E}\left[W_{0,b}(x)\right] \leq 1/3$ and $\mathbf{E}\left[W_{m+1,b}(x)\right] \geq 2/3$ for each $b \in \{\pm 1\}$. [2]

Using Yao's hybrid argument idea[7], we can write the difference of these expectations as:

$$\frac{1}{3} \leq \mathbf{E}\left[W_{m+1,y}(x)\right] - \mathbf{E}\left[W_{0,y}(x)\right] = \sum_{i=0}^{m} \mathbf{E}\left[W_{i+1,y}(x) - W_{i,y}(x)\right]$$

Thus, there is $i^* \in \{1, \ldots, m+1\}$ such that $\mathbf{E}\left[W_{i^*,y}(x) - W_{i^*-1,y}(x)\right] \geq 1/3m$. But from definition of the functions h_i, we have:

$$W_{i^*,y}(x) = \frac{y}{2} h_{i^*}(x) + \frac{1}{2}(W_{i^*,1}(x) + W_{i^*,-1}(x))$$

But from our construction of functions $W_{i,b}(x)$, we know that:

$$\mathbf{E}\left[\frac{1}{2} W_{i^*,1}(x) + \frac{1}{2} W_{i^*,-1}(x)\right] = \mathbf{E}\left[W_{i^*-1,y}(x)\right]$$

Combining the previous two equations, we obtain $\mathbf{E}\left[y \cdot h_{i^*}(x)\right] = \mathsf{cor}_{\mathcal{D}'}(h_{i^*}) \geq 2/3m$. In other words, $\mathsf{err}_{\mathcal{D}'}(h_{i^*}) \leq 1/2 - 1/3m$ as claimed. □

4 Learning in Realizable Scenario

We improved the analysis of [1] in the near-realizable and realizable case; while resolving one of their open questions. In [1], there was a slack in the relation between refutability and learning. They showed that ϵ-agnostic learning implies 4ϵ-refutation; and δ-refutation implies $(\delta + \zeta)$-agnostic learning for any $\zeta \geq 1/\operatorname{poly}(n)$. Their result has an intrinsic limitation that it cannot be used to show equivalence between learning and refutation in the realizable case. Indeed, to comment about possibility of learning, they required a strong $(1/2 - \zeta)$-refutation algorithm, which might be hard to obtain. Say, for instance, we can use Gaussian elimination to distinguish random assignments from completely satisfiable set of equations; but it may not always be possible to distinguish random case from the case where 0.6 fraction of equations are satisfiable. They left it as open question.

This problem, in realizable case, was also studied in distribution-independent setting by Salil Vadhan [2], where he gave equivalence between PAC-learning and a slightly different notion of refutation. His

[2] This expectation, and all other future expectations in this proof will be over the challenge example (x, y), all other draws form $\mathcal{D}'$ and internal randomness of our algorithm.

notion of refutation was to differentiate between NOISE and STRUCTURE for $\mathcal{C}$, but under *worst-case* distribution $\mathcal{D}$ of data. We also note that if we lift our distribution specific restriction and allow worst case distributions, then we can use any boosting algorithm to lift weak PAC-learning to standard PAC-learning, and in particular, we obtain the result of [2]. We remark with 'regret' [3] that [2] has many interesting applications, for instance, generalization the work of Daniely and Shalev-Schwartz on hardness of PAC-learning and agnostic learning half-spaces [4], as well as some other connections to cryptographic hardness assumptions, that we did not find enough space to cover.

First, we formally define the problem of distribution-dependent (weak) PAC-learnability and then show its equivalence to 1-refutability in Corollary 4.1.

Definition 6 *We say that $\mathcal{C}$ is weakly* PAC*-learnable over $\mathcal{D}$ iff there is an efficient algorithm $\mathcal{L}_{\mathsf{PAC}}$ that for* each *extension $\mathcal{D}'$ such that $\mathsf{err}^*_{\mathcal{D}'} = 0$, outputs with probability $3/4$ a hypothesis h such that $\mathsf{err}_{\mathcal{D}'}(h) \leq 1/2 - 1/\operatorname{poly}(n)$.*

Corollary 4.1 *$\mathcal{C}$ is weakly* PAC*-learnable on $\mathcal{D}$ iff $\mathcal{C}$ is 1-refutable on $\mathcal{D}$.*

We note that we cannot improve our guarantee to match standard PAC-learning guarantee of error $1/3$ because of a fundamental bottleneck in Theorem 3.6. We leave it as an open question to find a correct notion of *refutation* so as to establish an equivalence between PAC-learning (where the error probability is $1/3$) and refutability in the distribution specific setting.

5 Conclusion

We showed a slight improvement over analysis of [1] to obtain equivalence between agnostic learning and refutability, which can even be extended to *weak* PAC-learning and refutability. This partially solves an open question posed in [1]. We also generalized the result of [2] to distribution specific setting. Unfortunately, we could not find any analogue of *proper-boosting* [4] in the literature. Apart from being interesting in its own right, it may have immediate applications to obtain analogous equivalence between a notion of refutability and agnostic (or, even PAC) proper learning.

References

[1] Pravesh Kothari and Roi Livni. *Agnostic Learning by Refuting*, ITCS, 2018.

[2] Salil Vadhan. *On Learning versus Refutation*, COLT, 2017.

[3] Ran Raz. *Fast Learning Requires Good Memory: A Time-Space Lower Bound for Parity Learning*, FOCS, 2016.

[4] Amit Daniely and Shai Shalev-Schwartz. *Complexity Theoretic Limitations on Learning DNFs*, COLT, 2016.

[5] Vitaly Feldman. *Distribution-Specific Agnostic Boosting*, ICS, 2010.

[6] Oded Goldreich, Shafi Goldwasser and Dana Ron. *Property Testing and Its Connection to Learning and Approximation*, FOCS, 1995.

[7] Andrew Yao. *Theory and Applications of Trapdoor Functions*, FOCS, 1982.

[3] When compared to *experts* like Avrim Blum, who can magically fit entire proofs on a single slide!

[4] Where we want the boosted hypothesis to lie in the class of weak-hypothesis, and at the same time be comparable with the best hypothesis present in the same class.