On the Complexity of Learning from Label Proportions

Benjamin Fish and **Lev Reyzin**
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago, Chicago, IL 60022
{bfish3,lreyzin}@uic.edu

Abstract

In the problem of learning with label proportions (also known as the problem of estimating class ratios), the training data is unlabeled, and only the proportions of examples receiving each label are given. The goal is to learn a hypothesis that predicts the proportions of labels on the distribution underlying the sample. This model of learning is useful in a wide variety of settings, including predicting the number of votes for candidates in political elections from polls.

In this paper, we resolve foundational questions regarding the computational complexity of learning in this setting. We formalize a simple version of the setting, and we compare the computational complexity of learning in this model to classical PAC learning. Perhaps surprisingly, we show that what can be learned efficiently in this model is a strict subset of what may be leaned efficiently in PAC, under standard complexity assumptions. We give a characterization in terms of VC dimension, and we show that there are non-trivial problems in this model that can be efficiently learned. We also give an algorithm that demonstrates the feasibility of learning under well-behaved distributions.

1 Introduction

In this paper, we investigate the complexity of the learning problem of estimating the proportion of labels for a given set of instances. For example, this problem appears when predicting the proportion of votes for a given candidate [de Freitas and Kück, 2005]; correctly predicting how each individual votes is not required, only which candidate will win. Variants of this problem also appear in many other domains, including in consumer marketing [Chen *et al.*, 2006], medicine and other health domains [Hernández-González *et al.*, 2013; Wojtusiak *et al.*, 2011], image processing [de Freitas and Kück, 2005], physical processes [Musicant *et al.*, 2007], fraud detection [Rüping, 2010], manufacturing [Stolpe and Morik, 2011], and voting networks [Fish *et al.*, 2016].

In classical PAC learning, we are given labeled data instances from a distribution, and in the idealized case, must find a function that labels all of the data consistent with the observations. In less constrained settings, the goal is to find a function of low error, or at least of error as low as possible on the data presented to the algorithm. There is substantial literature on classical PAC learning outside the scope of this work; see e.g. [Shalev-Shwartz and Ben-David, 2014] for a survey. Once the classifier is found, it is easy to find the proportion of instances with a given label by invoking the classifier on the instances. Algorithms for estimating the proportion of labels with labeled data have been introduced before, for example by Iyer et al. [2014].

However, getting instances with attached labels, as assumed in classical PAC learning, is often difficult. Sometimes this is due to limits on the measurement process [Hernández-González *et al.*, 2013; de Freitas and Kück, 2005; Musicant *et al.*, 2007; Stolpe and Morik, 2011]. At other times, before datasets are released, labels are purposely detached from their instances in order to maintain privacy [Chen *et al.*, 2006; Rüping, 2010; Wojtusiak *et al.*, 2011]. Instead, only the proportion of labels are given for a group of sample instances. For example, in estimating who will win an election, pre-election polls only release the percentage of people planning to vote for a given candidate. Quadrianto et al. [2009] give several other examples where the only data available is of this form.

The goal is then to learn a classifier from a hypothesis class that is able to correctly predict the proportions of labels from a hidden distribution using a training set which consists of a set of instances and the proportions of labels of that set of instances. This is the learning-theoretic problem we formalize and tackle in this paper. The proportion of labels may be inferred by first finding a classifier that predicts the labels for each instance [Patrini *et al.*, 2014; Quadrianto *et al.*, 2009; Rüping, 2010; Yu *et al.*, 2013]. Alternatively, Iyer et al. [2016] propose inferring the proportion of labels directly.

A related setting is Multiple Instance Learning [Dietterich *et al.*, 1997], where the goal is to classify bags of examples with unobserved labels, where the bag is labeled positively by a boolean 'or' function: if any example in the bag is labeled positively, the bag is as well. The goal in Multiple Instance Learning is to label new bags with whether any example in the bag is labeled positively. This is distinct from the problem we tackle in this paper because in our problem we know the

proportion of labels instead.

Yu et al. [2014] introduce a version of a model for learning from label proportions. In their model, each bag of examples comes with the proportions of each label in that bag, and each bag is drawn i.i.d. from a distribution over bags. They give some of the first sample complexity guarantees. Another other approach is where the examples are drawn i.i.d., but the bags may be an arbitrary partition of the examples, as in [Rüping, 2010; Stolpe and Morik, 2011]. Compared to these 'bag' models, our model of learning from label proportions corresponds to the 'one-bag case' with binary labels, where each example is drawn i.i.d. from an arbitrary distribution. However, as we demonstrate, this model is already interesting to study. We formalize this as a PAC-like learning model, which allows us to compare the difficult of learning a hypothesis class in classical PAC learning to learning a hypothesis class in this model.

In particular, we give the following results, including the first computational hardness results for learning label proportions. After formally defining the model in Section 2, we show in Section 3 that under standard complexity assumptions, classes of high VC dimension are not efficiently learnable from label proportions. We also give examples of classes with lower VC dimension that are not efficiently learnable from label proportions, using stronger complexity assumptions. Then, in Section 4 we show that the classes of functions that are learnable from label proportions are a strict subset of the classes that are PAC learnable. Finally, in Section 5 we give some positive results indicating cases where it is possible to PAC learn from label proportions. We also show that n-dimensional half-spaces over the boolean cube are learnable from label proportions under the uniform distribution.

2 Model and sample complexity

For a distribution D over the domain of a function c, call $c(D)$ the resulting distribution over the range of c. For c a function $\{0,1\}^n \to \{0,1\}$, we will call p_c the percentage of positive labels in this distribution, i.e. $c(D)(1)$. For a given sample, we call the percentage labeled positively as $\hat{p_c}$. Where clear, we will abbreviate these as p and $\hat{p}$ respectively.

In this setting, each example x drawn from D has a hidden label $c(x)$, but the learning algorithm does not get to see examples with labels. Instead, the algorithm only gets to see the set of unlabeled examples S and $\hat{p}$, the percentage of S labeled positively by c. The goal is to find a function h in a hypothesis class H such that p_c should be close to p_h with high probability.

Definition 1. *A class of functions H is* PAC learnable from label proportions *if there is an efficient algorithm A such that for every target function c in H, any distribution D over $\{0,1\}^n$, and for any $\epsilon, \delta > 0$, given $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$ examples drawn i.i.d. from D and $\hat{p}$, returns a hypothesis h in H such that*

$$\mathbb{P}[|p_c - p_h| \leq \epsilon] \geq 1 - \delta.$$

We call this form of learning "**PAC learning from label proportions**." In general, we may consider agnostic or improper versions of this PAC model. However, improper learning from the class of all functions here is very easy: We can efficiently learn with a sample complexity that only depends on ϵ and δ:

Observation 1. *The sample complexity for improper PAC learning from label proportions is $O\left(\frac{\ln(1/\delta)}{\epsilon^2}\right)$.*

Proof outline. In improper learning, it is easy to find a function h^* so that not only does $\hat{p}_{h^*} = \hat{p}$, but also $p_{h^*} = \hat{p}$: e.g. h^* may be a randomized function that on any input returns 1 with probability $\hat{p}$ and 0 otherwise. Then $p_{h^*} = \hat{p}$ and a Chernoff bound implies that $\hat{p}$ is close to p. □

For example, if the task is to predict the proportion of votes for a given candidate using only a single poll, *improper* learning in this model is easy simply by virtue of the fact that $\hat{p}$ is an unbiased estimator for p. However, the hypothesis h^* described above will not be a realistic model of voting. So proper learning corresponds to finding a realistic model of voting, one which describes a relationship between examples and labels, that also predicts the proportion of votes correctly. For this reason, for the remainder of this paper, we will only consider *proper* PAC learning from label proportions.

Definition 1 is a distribution-free setting, but when the distribution is known, sample complexity also may be independent of the VC-dimension.

Observation 2. *Let D be a known distribution. Let*

$$\beta = \inf_{\substack{h,h' \in H: \\ h \neq h'}} |p_h - p_{h'}|.$$

Then the sample complexity for PAC learning from label proportions the hypothesis class H is $O\left(\frac{\ln(1/\delta)}{\beta^2}\right)$.

Proof outline. Here, we can use another Chernoff bound to get that with high probability, $\hat{p}$ is within $\beta/2$ of p_c, for c the target hypothesis. But the definition of β implies that there is exactly one value p_{c^*} in $\{p_c : c \in H\}$ such that $\hat{p}$ is closer to p_{c^*} than any other value in $\{p_c : c \in H\}$. Then with high probability $p_c = p_{c^*}$. Thus an algorithm may output any h such that $p_h = p_{c^*}$. □

This analysis of the distribution-free setting only considers sample complexity and not computational complexity. In Section 5, we will give an example where we can efficiently PAC learn from label proportions under the uniform distribution.

We may still wish to bound the sample complexity of PAC learning from label proportions in the setting where the distribution is arbitrary. The same bounds that hold in PAC learning under an arbitrary loss function [Shalev-Shwartz and Ben-David, 2014] (or the absolute value of an arbitrary loss function) also hold here. Namely, we can use the VC dimension of a hypothesis class H to bound generalization error. We denote this quantity by $\mathrm{VC}(H)$. In particular, we have:

Theorem 3 (Occam's razor)**.** *For target function $c \in H$, with probability at least $1 - \delta$, for all $h \in H$,*

$$|p_c - p_h| \leq |\hat{p}_c - \hat{p}_h| + O\left(\frac{1}{\delta}\sqrt{\frac{\log(m/\mathrm{VC}(H))}{m/\mathrm{VC}(H)}}\right).$$

3 Hardness of learning from label proportions

We start by showing it is NP-hard to learn when the VC dimension is sufficiently large. To do this, we start by defining the consistency problem of a hypothesis class. For a hypothesis class C, the *consistency problem* for PAC learning from label proportions is the following: Given a set $X = \{x_i\}$ of points, for each point x_i an integer a_i, and a proportion p, is there a hypothesis $c \in C$ such that $\frac{\sum_{i \in I} a_i}{\sum_i a_i} = p$, where $I = \{i : c(i) = 1\}$? If there is such a c, we will say that c is consistent with X.

Proposition 4. *Let C be a hypothesis class such that $\mathrm{VC}(C) \geq n^\gamma$ for some constant $\gamma > 0$. The consistency problem for C is* NP*-hard.*

Proof. We reduce from SUBSET SUM, which asks, without loss of generality, given a set $S = \{a_1, \ldots, a_m\}$ of positive integers and a positive integer b, if there is a subset $S' \subseteq S$ such that $\sum_{a \in S'} = b$. Let $n = m^{1/\gamma}$. Then $\mathrm{VC}(C) \geq m$, so C shatters some m points, call them $x_1, \ldots, x_m$. We now construct the following instance of the consistency problem:

Define $X = \{x_1, \ldots, x_m\}$, each associated integer to be a_i, and the percent p of positive labels to be $\frac{b}{\sum_i a_i}$.

We now show that X is consistent with a hypothesis $c \in C$ if and only if the given subset sum instance has a solution. If X is consistent with a hypothesis c, then $p = \frac{\sum_{i \in I} a_i}{\sum_i a_i}$. This immediately implies

$$\frac{\sum_{i \in I} a_i}{\sum_i a_i} = \frac{b}{\sum_i a_i},$$

i.e. $b = \sum_{i \in I} a_i$. Then the subset $\{a_i : i \in I\}$ is the solution for the subset sum instance. In the other direction,let $S' \subseteq S$ be the set of integers such that $\sum_{a \in S'} a = b$. Since C shatters $x_1, \ldots, x_m$, there is a hypothesis c that labels positively all and only the points x_i for i such that $a_i \in S'$. That is,

$$\sum_{a \in S'} a = \sum_{i \in I} a_i, \text{ where } I = \{i : c(i) = 1\}.$$

Thus X is consistent, witnessed by this hypothesis:

$$p = \frac{b}{\sum_i a_i} = \frac{\sum_{a \in S'} a}{\sum_i a_i} = \frac{\sum_{i \in I} a_i}{\sum_i a_i}.$$

□

We now show that PAC learning from label proportions is hard whenever the VC-dimension is a fractional power. We reduce from the consistency problem to the learning problem, which is a slightly more involved reduction than in the classical PAC setting.

Theorem 5. *Let C be a hypothesis class such that $\mathrm{VC}(C) \geq n^\gamma$ for some constant $\gamma > 0$. There is no efficient algorithm for PAC learning C from label proportions unless* NP = RP.

Proof. It suffices to reduce from the consistency problem, above. Indeed, using an oracle to an efficient PAC learner for C from label proportions, we merely need to solve the consistency problem with high probability. Given an instance of the consistency problem with input set $X = \{x_i\}$, integers a_i, and proportion p. Define a distribution D that outputs x_i with probability proportional to a_i.

Set

$$\epsilon = \frac{1}{2 \sum_i a_i}.$$

For $\delta > 0$, we will query the oracle with inputs δ, ϵ, and an i.i.d. sample from D of size $m = f(1/\delta, 1/\epsilon)$, where f is the polynomial sample bound for the oracle. Since the sample from D may not be exactly a_i copies of x_i, we do not know $\hat{p}$ to give to the oracle. So instead, we will invoke the oracle $m + 1$ times, setting the input proportion to be each of $0, 1/m, \ldots, 1$, and then check the resulting output hypothesis to see if it is consistent with X. If so, accept, and if no such hypothesis is ever found, reject[1].

Certainly, if we accept, there is a consistent hypothesis by definition: we accept if an oracle outputs a consistent hypothesis. Conversely, if the consistency problem is solvable, then we will accept: Let c be the consistent hypothesis. Since it is consistent, by the definition of D, $p_c = p$. Now consider the invocation of the oracle with the true proportion $\hat{p}$. This invocation will output some hypothesis h that will, except with probability at most δ, satisfy

$$|p_c - p_h| = \left| p - \frac{\sum_{x_i \in S} a_i}{\sum_i a_i} \right| \leq \frac{1}{2 \sum_i a_i},$$

where S is the set of points h labels positively. Since each a_i is an integer, this implies that $p = \frac{\sum_{x_i \in S'} a_i}{\sum_i a_i}$, i.e. that h is consistent with X and therefore we will accept with probability at least $1 - \delta$.

Setting δ to go to 0 in the size of the input of the consistency problem completes the proof. □

A natural question to ask is if there are classes with VC dimension smaller than n^γ that are still hard to learn. We now show that this is the case for parity functions on the first k bits of the input.

Recall in (white-label) noisy PAC learning, each label in the training data is flipped with unknown rate η. We assume the algorithm is given as input some η', where $\eta \leq \eta' < 1/2$ and must only take time polynomial in $\frac{1}{1-2\eta'}$. Noisy PAC learning parity functions under the uniform distribution is presumed to be hard. Blum et al. [2003] give an $2^{O(n/\log n)}$) algorithm, which is the best-current bound.

We now find a specific distribution where PAC learning from label proportions is hard in this sense for parities:

Theorem 6. *For a hypothesis c, Let D_c be the the distribution over $\{0,1\}^n$ that places $\frac{\eta}{2^{n-1}}$ weight on the examples labeled* 0 *and $\frac{1-\eta}{2^{n-1}}$ weight on examples labeled* 1.

PAC learning parities from label proportions under D_c is as least as hard as PAC learning unknown parity c with η white-label noise under the uniform distribution.

[1]The oracle's behavior is undefined if the value input as the proportion of positive labels is not the true value $\hat{p}$. We may assume, however, that the oracle rejects whenever this is the case because the time the oracle takes is polynomially-bounded so we can just wait for that amount of time to see if the oracle returns a hypothesis.

Proof. We use an oracle for PAC learning parities from label proportions under D_c to noisy-PAC learn parities. We get as input η', parameters ϵ and δ, and some m examples x_i, with m to be determined later, with noisy labels $\tilde{\ell}_i$. When $\tilde{\ell}_i = 1$, with probability η, the true label $\ell_i = 0$ and otherwise $\ell_i = 1$. We may assume that the unknown parity c is non-trivial. Then under the uniform distribution over $\{0,1\}^n$, for any such parity function, there are 2^{n-1} points labeled 1 and 2^{n-1} points labeled 0. For any point labeled 0, the probability that it was drawn from the uniform distribution is $\frac{1}{2^{n-1}}$ and the probability that its label was flipped to 1 was η. Then the probability that an example had $\tilde{\ell}_i = 1$ but $\ell_i = 0$ is $\frac{\eta}{2^{n-1}}$ and similarly if $\ell_i = 1$ the probability is $\frac{1-\eta}{2^{n-1}}$. Note that this is exactly the distribution D_c. So if the oracle for PAC learning parities from label proportions is given just the examples where $\tilde{\ell}_i = 1$, the oracle will receive i.i.d. samples from D_c. We will also give to the oracle $\epsilon' = \frac{1/2-\eta'}{2}$ and $\delta' = \delta/3$. The expected proportion of these examples given to the oracle is $1-\eta$, but we do not know the true labels nor do we know η. So instead, we will invoke this oracle $M+1$ times, with the proportion given to the oracle as each of $0, 1/M, \ldots, 1$, where $M = \sum_i \tilde{\ell}_i$, i.e. the number of training examples with noisy label $\tilde{\ell}_i = 1$ [2].

If the oracle returns the correct parity c, then it should agree in expectation with the noisy labels $\tilde{\ell}_i$ on all but η of the examples. For an incorrect parity c', by the orthonormality of the parity functions, the expected disagreement is $1/2$. For h the output of the oracle, if smaller than an $\frac{\eta'+1/2}{2}$ fraction of the noisy labels $\tilde{\ell}_i$ disagree with the corresponding label $h(x_i)$, then we return the hypothesis. Otherwise, we repeat with the next invocation of the oracle.

Let f be the polynomial sample bound for the oracle for PAC learning from label proportions . First, we need to make sure that the oracle receives at least $f(1/\epsilon', 1/\delta')$ examples except with probability at most $\delta/3$. In expectation, $m/2$ of the examples x_i will have $\tilde{\ell}_i = 1$. Using a Chernoff bound, $\mathbb{P}\left[\left|\sum_i \tilde{\ell}_i - m/2\right| > m/4\right] \leq 2e^{-m/8}$. So the oracle will receive at least $\frac{1}{4}m$ examples (and no more than $\frac{3}{4}m$ examples) except with probability no more than $\delta/3$ so long as $m > 8\log(6/\delta)$. This then means that we require $m > 4 \cdot f(1/\epsilon', 1/\delta')$ so that $M \geq f(1/\epsilon', 1/\delta')$.

Now we need to verify that when the proportion given to the oracle is the correct proportion $\hat{p}_c$, the oracle will return c except with probability at most $\delta/3$. The oracle is guaranteed to return a parity h such that except with probability $\delta' = \delta/3$,

$$|p_h - p_c| \leq \epsilon' = \frac{1/2-\eta'}{2}.$$

Using the definition of D_c, $p_c = 1-\eta$. If $h \neq c$, then $p_h = 1/2$ again by orthonormality. But then

$$|p_h - p_c| = |1/2 - \eta| > \frac{1/2-\eta'}{2},$$

so it must be the case that $h = c$. Thus at least one of the invocations of the oracle will return the correct parity.

So it remains to show that we will succeed at returning this parity. If the oracle returns an incorrect parity h, again using a Chernoff bound,

$$\mathbb{P}\left[\left|\frac{\sum_i \mathbb{1}_{h(x_i)\neq\tilde{\ell}_i}}{m} - 1/2\right| \geq \frac{1/2-\eta'}{2}\right] \leq 2e^{-\frac{m(1/2-\eta')^2}{2}} < \frac{1}{M+1}\cdot\frac{\delta}{3}$$

when

$$m = \Omega\left(\frac{\log(\frac{M}{\delta})}{(1/2-\eta')^2}\right) = \Omega\left(\frac{\log\left(\frac{1}{(1/2-\eta')\delta}\right)}{(1/2-\eta')^2}\right)$$

because $M \leq \frac{3}{4}m$, where $\mathbb{1}_A$ is the indicator function that is 1 if A is true and 0 otherwise. This implies that for an incorrect hypothesis, whose expected fraction of disagreements with the noisy labels is $1/2$, the empirical fraction is at least $\frac{\eta'+1/2}{2}$, the threshold we had set. Similarly, for the correct hypothesis, where the expected fraction of disagreements is $\eta < \eta'$, the empirical fraction of disagreements is no more than $\frac{\eta'+1/2}{2}$ except with probability at most $\frac{1}{M+1}\cdot\frac{\delta}{3}$. This means that all of the tests of the hypothesis succeeds except with probability at most $\delta/3$. Then setting

$$m = \Omega\left(\max\left(\left(\frac{\log\left(\frac{1}{(1/2-\eta')\delta}\right)}{(1/2-\eta')^2}\right), 4\cdot f(1/\epsilon', 1/\delta')\right)\right)$$

suffices so that, with the union bound, the total probability of failure is no more than δ, as required. □

Consider parity functions on the first k bits, which have VC dimension equal to k. There is no known algorithm for noisy PAC learning parity functions on the first k bits when $k = \omega(\log n \log\log n)$. It is conjectured that there is no efficient algorithm for PAC-learning noisy parity that runs in time $o(2^{\sqrt{n}})$, which would imply hardness of noisy PAC learning parities on the first k bits for $k = \omega(\log^2 n)$. Calling this the 'parity hardness assumption,' Theorem 6 implies the following:

Corollary 7. *Under the parity hardness assumption, there is no efficient algorithm for PAC learning label proportions of parities on the first k bits for $k = \omega(\log^2 n)$.*

This means there are hypothesis classes with VC dimension $\omega(\log^2 n)$ that aren't PAC learnable from label proportions. However, under a stronger assumption than, say, NP≠RP, we can find a hypothesis class hard to learn with even smaller VC dimension.

Consider the hypothesis class where each hypothesis labels exactly k points positively, say for $k \leq n/2$:

$$H_k = \{h : X \to \{0,1\} : |\{x : h(x) = 1\}| = k\}.$$

We show that in order to find a consistent hypothesis, we will need to solve the k-SUM problem, which we may assume asks when given integers $a_1, \ldots, a_m$, is there a set S of size k whose sum is input b:

[2] As in the proof of Theorem 5, the oracle is undefined when the proportion of positive labels is not the true value $\hat{p}$. And similar to before, we may assume that the oracle returns an arbitrary hypothesis.

Observation 8. *The consistency problem for H_k is as least as hard as k-SUM.*

Proof. Using the notation above, given integers $a_1, \ldots, a_m$, let $X = \{1, \ldots, m\}$, let the associated integer to i be a_i, and let $p = \frac{b}{\sum_i a_i}$. If there is a set S of a_i's whose sum is b, then X is consistent, witnessed by the hypothesis h that labels $h(i) = 1$ if and only if i is in S, and vice versa if there is a consistent hypothesis h for X, then the set S of size k where $h(i) = 1$ will sum to b. □

k-SUM has a well-known $m^{k/2}$ algorithm. On the other hand, this is also a lower bound: Patrascu and Williams [2010] show that k-SUM requires $m^{\Omega(k)}$ time assuming the exponential time hypothesis. Ailon and Chazelle [2005] also give a lower bound of $m^{k/2}$ for linear decision trees. If $k = \log n$, then H_k is hard to lean efficiently in either of these settings. This follows from the same proof as in Theorem 5.

4 Comparing our model to classical PAC

The definition of PAC learning from label proportions makes it harder to learn a class on one hand (by unlinking input from label) but easier on the other hand (by making the loss function easier to satisfy). So it may not be obvious what the relationship with PAC is.

In this section, we show that the hypothesis classes that may be efficiently learned in PAC from label proportions is a subset of the classes that may be efficiently learned in PAC. Theorem 5 then implies it is a strict subset.

Theorem 9. *Suppose $NP \neq RP$. Then if a hypothesis class H is efficiently learnable from label proportions, it is also efficiently (proper) PAC learnable.*

Proof. Let H be learnable from label proportions by some efficient oracle A, and f the polynomial sample size required by this oracle. We now give an efficient algorithm for PAC learning H. Given $\epsilon, \delta > 0$, draw m samples from the unknown distribution D, with m to be determined later. Call the set S of unique inputs $x_1, \ldots, x_m$ and their labels $c(x_1), \ldots, c(x_m)$ for hidden target function c. Let k be the number of positive labels $\sum_j c(x'_j)$. Define a new distribution D' as the following:

$$D'(x) = \left\{ \begin{array}{cc} \frac{m}{km+m-k} & \text{if } x \in S \text{ and } c(x) = 1 \\ \frac{1}{km+m-k} & \text{if } x \in S \text{ and } c(x) = 0 \\ 0 & \text{otherwise} \end{array} \right\}.$$

Let $\epsilon' = 1/(2m^2)$ and $\delta' = \delta$. Draw $m' = f(1/\epsilon', 1/\delta')$ samples x'_j from D' and label each as $c(x'_j)$. We give to the oracle as input ϵ', δ', and the examples x'_j, along with the proportion of positive labels $\hat{p} = \frac{k}{m'}$. Then with probability at least $1 - \delta$ the oracle returns a hypothesis c^* such that

$$|p_{c^*} - p_c| < \frac{1}{2m^2}.$$

The smallest non-zero probability mass in D', however, is

$$\frac{1}{km + m - k} \geq \frac{1}{m^2},$$

minimized when $k = m$. Thus $p_{c^*} = p_c$.

We now show that $c^* = c$ when restricted to the points $x_1, \ldots, x_m$. Suppose there is a point x_i such that $c^*(x_i) \neq c(x_i)$ where $c(x_i) = 1$. Then in order to have $p_{c^*} = p_c$ while $c^*(x_i) = 0$, at least m points labeled 0 by c must be labeled positively by c^*, since D' places (proportional to) m weight on positively labeled points and only unit weight on negative points. This is a contradiction, as there are only m total points. Similarly, if $c(x_i) = 0$ and $c^*(x_i) = 1$, there must be m points labeled 0 by c^* that are labeled 1 by c, but again there are only m distinct points. Thus c and c^* must agree on all m points, i.e. c^* has zero empirical error.

All that remains is to check that the VC dimension of H is sufficiently small so that Occam's razor implies sufficiently small distributional error. But if $\mathrm{VC}(H)$ were, say, super-exponential, (indeed, merely polynomial), then by Corollary 5, it wouldn't be learnable from label proportions, contrary to our assumption that it is. (This only prevents the VC dimension of H from being too large as a function of n. However, a similar proof to Corollary 5 implies that it is hard to learn H if $\mathrm{VC}(H)$ were exponential in, say, the maximum representation size of a hypothesis in H instead.) □

5 Classes PAC learnable from label proportions

Call d the VC dimension of a given hypothesis class H. In Section 3, we showed that if d is a fractional power, H is hard to learn. We also gave examples with d as small as $\log n$ that are hard to learn, under stronger complexity assumptions. On the other hand, as long as labelings in a given hypothesis class are efficiently enumerable, then finite classes H are certainly PAC learnable from label proportions in time $|H|$. Or instead, by enumerating only distinct hypotheses on the sample, assuming that this is efficient, learning can be achieved in m^d time using Sauer's lemma. This immediately implies that all such classes with constant d are learnable from label proportions. We now show that not all classes with $d = \Omega(\log n)$ are hard to learn.

Consider the following variation of H_k which only allows hypotheses whose positive labels are close to each other:

$$H'_k = \{h : \{1, \ldots, 2^n\} \to \{0, 1\} : \max_{h(i)=h(j)=1} |i - j| \leq k\}.$$

There are still exponentially many functions and $VC(H'_k) = k$. H_k was shown to be hard in Section 3. However, for H'_k, this is not the case:

Observation 10. *PAC learning H'_k from label proportions has an $O(2^k m)$ time algorithm.*

Order the m examples in $\{1, \ldots, 2^n\}$, and for each length k subset, of which there are $m - k + 1$ of them, check all 2^k possible labelings. Now when $k = O(\log n)$, this is a polynomial-time algorithm for learning H'_k from label proportions even though the VC dimension is not constant.

In the classical PAC setting, when it is hard to learn under an arbitrary distribution, it is often still valuable to show that learning can still be done in special cases, such as the uniform distribution. We now give an example, namely half-spaces, where it is hard to learn from label proportions under arbitrary distributions (Theorem 5) but easy to learn under the uniform distribution. In other words, when the distribution is well-behaved, learning becomes much easier.

The idea to find a half-space that classifies the given proportion $\hat{p}$ positively is to take a random half-space through the origin, and then move it in the direction of its normal vector, and stop when the half-space classifies the input p proportion of the sample positively. With high probability, this will be possible because no two points in the sample will be projected to the same point on the normal vector.

Proposition 11. *The class of half-spaces in n dimensions is learnable from label proportions under the uniform distribution over $\{0,1\}^n$.*

Proof. Since the VC-dimension of half-spaces is linear in n by Radon's theorem [Mohri *et al.*, 2012], using Theorem 3 it certainly suffices to be able to efficiently find a half-space h such that $\hat{p_h} = p$ with high probability. Consider a hyperplane P of dimension $n-1$ through the origin and v a normal vector defining P.

First, we show that for a randomly chosen vector v, no two points in $\{0,1\}^n$ project more than exponentially close to each other (in terms of n) on v. This allows us to use only a polynomial number of bits to represent each projected point. Consider an arbitrary pair of points x and y in $\{0,1\}^n$ and consider the line ℓ that passes through these two points. If v and ℓ are perpendicular, then x and y will project onto the same point on v. More generally, we can find the maximum obtuse angle between v and ℓ such that the two points so that the points project exponentially close together on v. Any closer, and we will not have enough bits to distinguish between the projection of x and y. Namely, for a pair of points distance d apart, using the Taylor approximation for $\sin(x)$, the difference between $\pi/2$ and this maximum angle is no more than $O\left(\frac{1}{d2^{\omega(n^c)}}\right)$ for constant c. Since the points come from $\{0,1\}^n$, $d \geq 1$, and there are $O(2^{n^2})$ such pairs of points, so the total angle from which a uniformly-random vector v may not be chosen is at most $O\left(\frac{2^{n^2}}{2^{\omega(n^c)}}\right)$, an exponentially small probability. Thus, with high probability, no two points in $\{0,1\}^n$ project to the same point on v, or project more than exponentially close to each other on v.

Given m examples, setting m to be polynomial in n insures with high probability that all examples are distinct, and therefore no two examples project more than exponentially lose to each other on v. Since $\hat{p}_c = i/m$ for some $i \in \{0,1,\ldots,m\}$, we need to find a plane parallel to P such that the corresponding linear threshold function classifies i of the sample points positively. For each pair of consecutive projected points cv and $c'v$ on v for real number c and c', consider the half-space given by the plane defined by the points $p \in \mathbb{R}^n$ satisfying $v\left(p - \left(\frac{c+c'}{2}\right)v\right) = 0$, so that these two points are classified differently by the half-space. Thus one of these half-spaces (or the half-spaces classifying all points positively or negatively) will have $\hat{p}_h = i/m$ since no two points in the sample project onto the same point on v. □

While we have shown that it is strictly harder to PAC learn from label proportions than to PAC learn, introducing noise to the models changes the relationship between these two models. For example, PAC learning parities with unknown η white-label noise is hard under the uniform distribution, as discussed above, but PAC learning parities from label proportions with white-label noise is easy under the uniform distribution. In our model, that means each label is flipped i.i.d. with probability some unknown η, and the proportion of noisy positive labels $\hat{p}^\eta$ is given as input instead, but otherwise the learning requirement remains stays the same.

Observation 12. *The class of parities is learnable from label proportions under the uniform distribution and unknown η white-label noise.*

Proof. Let p_c^η be the proportion of positive labels under η noise and parity c. Note p_c^η is always

$$(1-\eta)p_c + \eta(1-p_c) = p_c(1-2\eta) + \eta,$$

but for any non-trivial parity c, $p_c = 1/2$, so $p_c^\eta = 1/2$. Then Observation 2 implies that we may distinguish efficiently the trivial parity from the non-trivial parities and in the case that $p_c^\eta = 1/2$ we may return any non-trivial parity. □

6 Conclusion

In this paper we formalized a model for learning a hypothesis class by only examples drawn from a distribution and the proportion of them receiving each label, with the goal of finding a hypothesis that matches these statistics on the underlying distribution, and we focused on the binary label setting.

While this task may seem easier than PAC learning, we prove that it is actually no easier, and sometimes harder – namely, we show that everything that is efficiently learnable in our model is also efficiently properly PAC learnable. Moreover, we show that classes of polynomially large VC-dimension are NP-hard to learn in our model, and we give some even stronger lower bounds for specific classes. We give examples where it is possible to efficiently PAC learn from label proportions, which may be surprising given that this is a low-information setting, including half-spaces under the uniform distribution.

These results are for the binary setting and only for the 'one bag' version of the problem. We leave for future work the analysis of the case where there is more than one bag of examples and each bag's proportion of labels is given. For that case, and in other similar settings where the learner is given more information, we expect there to be more positive algorithmic results.

References

[Ailon and Chazelle, 2005] Nir Ailon and Bernard Chazelle. Lower bounds for linear degeneracy testing. *J. ACM*, 52(2):157–171, 2005.

[Blum *et al.*, 2003] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

[Chen *et al.*, 2006] Bee-Chung Chen, Lei Chen, Raghu Ramakrishnan, and David R. Musicant. Learning from aggregate views. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 3, 2006.

[de Freitas and Kück, 2005] Nando de Freitas and Hendrik Kück. Learning about individuals from group statistics. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, UAI '05, Edinburgh, Scotland, July 26-29, 2005*, pages 332–339, 2005.

[Dietterich *et al.*, 1997] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[Fish *et al.*, 2016] Benjamin Fish, Yi Huang, and Lev Reyzin. Recovering social networks by observing votes. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pages 376–384, 2016.

[Hernández-González *et al.*, 2013] Jerónimo Hernández-González, Iñaki Inza, and José Antonio Lozano. Learning bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12):3425–3440, 2013.

[Iyer *et al.*, 2014] Arun Shankar Iyer, J. Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 530–538, 2014.

[Iyer *et al.*, 2016] Arun Shankar Iyer, J. Saketha Nath, and Sunita Sarawagi. Privacy-preserving class ratio estimation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 925–934, 2016.

[Mohri *et al.*, 2012] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

[Musicant *et al.*, 2007] David R. Musicant, Janara M. Christensen, and Jamie F. Olson. Supervised learning by training on aggregate outputs. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA*, pages 252–261, 2007.

[Pătraşcu and Williams, 2010] Mihai Pătraşcu and Ryan Williams. On the possibility of faster SAT algorithms. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 1065–1075, 2010.

[Patrini *et al.*, 2014] Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. (almost) no label no cry. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 190–198, 2014.

[Quadrianto *et al.*, 2009] Novi Quadrianto, Alexander J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

[Rüping, 2010] Stefan Rüping. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 911–918, 2010.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[Stolpe and Morik, 2011] Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML-PKDD 2011, Athens, Greece, September 5-9, 2011*, pages 349–364, 2011.

[Wojtusiak *et al.*, 2011] Janusz Wojtusiak, Katherine Irvin, Aybike Birerdinc, and Ancha V. Baranova. Using published medical results and non-homogenous data in rule learning. In *10th International Conference on Machine Learning and Applications and Workshops, ICMLA 2011, Honolulu, Hawaii, USA, December 18-21, 2011. Volume 2: Special Sessions and Workshop*, pages 84–89, 2011.

[Yu *et al.*, 2013] Felix X. Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$-SVM for learning with label proportions. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 504–512, 2013.

[Yu *et al.*, 2014] Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.