Journal of Artificial Intelligence Research 69 (2020) 1333–1349 Submitted 03/2020; published 12/2020

On the Complexity of Learning a Class Ratio from Unlabeled Data

Benjamin Fish BENJAMIN.FISH@MILA.QUEBEC
Mila - Quebec AI Institute
6666 St-Urbain, #200, Montreal, QC, H2S 3H1

Lev Reyzin LREYZIN@UIC.EDU
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
851 S Morgan St, Chicago, IL 60607

Abstract

In the problem of learning a class ratio from unlabeled data, which we call CR learning, the training data is unlabeled, and only the ratios, or 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 applicable to a wide variety of settings, including predicting the number of votes for candidates in political elections from polls.

In this paper, we formally define this class and resolve foundational questions regarding the computational complexity of CR learning and characterize its relationship to PAC learning. Among our results, we show, perhaps surprisingly, that for finite VC classes what can be efficiently CR learned is a strict subset of what can be learned efficiently in PAC, under standard complexity assumptions. We also show that there exist classes of functions whose CR learnability is independent of ZFC, the standard set theoretic axioms. This implies that CR learning cannot be easily characterized (like PAC by VC dimension).

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 & 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 & Kück, 2005), physical processes (Musicant et al., 2007), fraud detection (Rüping, 2010), manufacturing (Stolpe & Morik, 2011), and voting networks (Fish et al., 2016). This problem may also arise when attempting to correct for differences between training and testing distributions (Du Plessis & Sugiyama, 2014; Saerens et al., 2002).

In classical Probably Approximately Correct (PAC) learning (Valiant, 1984), 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

©2020 AI Access Foundation. All rights reserved.

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. Alternatives that directly estimate the proportion of labels with labeled data have been introduced by, for example, Iyer et al. (2014).

However, getting instances with attached labels for training, 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 & Kück, 2005; Musicant et al., 2007; Stolpe & Morik, 2011). At other times, before data sets 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. In the case of binary labels, we will refer to this proportion as the *class ratio*. For example, in estimating who will win an election, training data consisting of 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.

Our goal is to use a training set which consists of a set of instances and the proportions of labels of that set of instances to learn a classifier from a hypothesis class that is able to correctly predict the class ratio, i.e. the proportions of labels, from a hidden distribution. This is the learning-theoretic problem we formalize and tackle in this paper. With only the proportion of labels given as training data, 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.

Yu, Choromanski, Kumar, Jebara, and Chang (2014) introduce a version of a model for a related problem, 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. Their goal is to learn a classifier with low error. They give some of the first sample complexity guarantees. Another approach is where the examples are drawn i.i.d., but the bags may be an arbitrary partition of the examples, as is done by Rüping (2010) and Stolpe and Morik (2011). Compared to these "bag" models, our model of learning corresponds to the "one-bag" case with binary labels, where each example is drawn i.i.d. from an arbitrary distribution (and the goal is to learn a classifier that correctly predicts the class ratio). However, as our results demonstrate, we believe this model is already interesting to study. We formalize this as a PAC-like learning model, which allows us to compare the difficult problem of learning a hypothesis class in classical PAC learning to learning a hypothesis class in this model, which we call *CR learnability*.

In particular, we give the following results, including the first computational hardness results for learning a class ratio. After formally defining the model in Section 2, we start in Section 3 by pointing out that CR learning satisfies a natural uniform convergence property, meaning that any class with finite VC dimension is learnable in this model (see, e.g., Shalev-Shwartz & Ben-David, 2014, for more on VC dimension). However, CR learnability is not characterized by finite VC dimension: We give a simple class of functions with infinite VC dimension that is learnable in this model. In addition, we give a class of functions whose learnability under label proportions is independent of the Zermelo-Fraenkel set theory with

the axiom of choice (ZFC), implying that learnability in this model does not admit any simple characterization like VC dimension.

In Section 4, we compare efficient PAC to efficient CR learnability, where the learning algorithm must take only polynomial time in the size of its input. We show that for classes with finite VC dimension, if it is efficiently CR learnable, it is also efficiently properly PAC learnable. We then go on to show that CR learning can be harder than PAC learning in Section 4.1: classes that are efficiently PAC learnable like parities and monotone disjunctions are hard to learn in this setting. Finally, in Section 5 we give some positive results indicating cases where it is possible to CR learn. We also show that n-dimensional half-spaces over the boolean cube are CR learnable under the uniform distribution.

2. Model and Sample Complexity

For c a function $c : X \to \{0, 1\}$ and D a distribution over the domain of c, we will call p_c the percentage of positive labels in this distribution, i.e. the class ratio $p_c = \mathbb{P}_{x\sim D}[c(x) = 1]$. For a given sample S, we call the percentage labeled positively as the class ratio $\hat{p}_c = \frac{1}{|S|}\sum_{x\in S} c(x)$. 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 $\hat{p}_c$, 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.

We now define CR learnability (where "CR" stands for "class ratio").

Definition 1 (CR Learnability)**.** *A class of functions H is* **CR learnable** *if there is an algorithm A such that for every target function c in H, any distribution D over X, 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 $\hat{p}_c$, returns a hypothesis h in H such that*

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

If in addition, A is an efficient algorithm (i.e. running in time polynomial in the size of its input), then we call such a class **efficiently CR learnable**.

In general, we may consider agnostic or improper versions of this learning model. Improper learning is where the returned hypothesis need not belong to the same class as the class from which the target functions come, but instead need only belong to some larger class instead. If this larger class is just the class of all functions $h : X \to \{0, 1\}$, CR learning is very easy. We can efficiently improperly CR learn with a sample complexity that only depends on ϵ and δ:

Observation 2. *The sample complexity of improper CR learning 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 Hoeffding 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. When the hypothesis class represents all plausible models of voting, proper learning corresponds to finding a plausible 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* CR learning.

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

Observation 3. *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 CR learning the hypothesis class H is $O\left(\frac{\ln(1/\delta)}{\beta^2}\right)$.

Proof outline. Here, we can use another Hoeffding 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 $p_h = p_{c^*}$. □

In Section 5, we return to distribution-specific learning.

3. Comparing PAC to CR Learnability

We start by considering sample complexity in the distribution-free setting. Is CR learning harder or easier than PAC learning? We first give a uniform convergence result that implies that if a class has finite VC dimension, then it is CR learnable. Despite this result, unlike the PAC setting, VC dimension does *not* characterize CR learnability. Perhaps surprisingly, we show that there are classes with infinite VC dimension that are CR learnable and those whose CR learnability are independent from ZFC (for more, see below).

First, we prove a uniform convergence bound. Following the proof of the equivalent bounds in PAC learning under an arbitrary loss function (see Shalev-Shwartz & Ben-David, 2014), we can show the same bounds also hold here. Namely, we can use the VC dimension of a hypothesis class H to bound generalization error. In particular, we have:

Theorem 4 (Uniform convergence)**.** *For target function $c \in H$ with finite VC dimension d, with probability at least $1 - \delta$, for all $h \in H$,*

$$|p_h - p_c| \leq |\hat{p}_h - \hat{p}_c| + \sqrt{\frac{8d\log(em/d)}{m}} + \sqrt{\frac{2\log(4/\delta)}{m}}.$$

Proof. Consider the empirical loss $|\hat{p}_h - \hat{p}_c| = \left|\frac{1}{m}\sum_{i=1}^{m} h(x_i) - c(x_i)\right|$ for sample $S = \{x_1, \ldots, x_m\}$. Unfortunately its expectation $\mathbb{E}_{S\sim D^m}[|\hat{p}_h - \hat{p}_c|]$ *does not* equal the distributional loss $|p_h - p_c| = |\mathbb{E}_{x\sim D}[h(x)] - \mathbb{E}_{x\sim D}[c(x)]|$, so instead, we consider convergence of the *signed* loss instead: $\ell(h, x) := h(x) - c(x)$, $\mathcal{L}_D(h) := \mathbb{E}_{x\sim D}[\ell(h, x)]$, and

$\mathcal{L}_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(h, x_i)$. This is a loss function with $|\ell(h,x)| \leq 1$, so we can bound generalization error by the empirical Rademacher complexity (see, e.g., Shalev-Shwartz and Ben-David (2014)): For $R(\cdot)$ the Rademacher complexity, with probability of at least $1-\delta$, for all $h \in H$,

$$\mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 2\mathbb{E}_{z_i \sim D}[R(\{(\ell(h,z_1),\ldots,\ell(h,z_m)) \mid h \in H\})] + \sqrt{\frac{2\ln(2/\delta)}{m}}.$$

The Sauer-Shelah lemma implies that $|\{(h(z_1),\ldots,h(z_m))|h \in H\}| \leq \left(\frac{em}{d}\right)^d$, so we also have $|\{(\ell(h,z_1),\ldots,\ell(h,z_m)) \mid h \in H\}| \leq \left(\frac{em}{d}\right)^d$. Again since $|\ell(h,x)| \leq 1$, Massart's lemma then implies that $R(\{(\ell(h,z_1),\ldots,\ell(h,z_m)) \mid h \in H\}) \leq \sqrt{\frac{2d\log(em/d)}{m}}$. Thus with probability of at least $1-\delta/2$, for all $h \in H$,

$$\mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 2\sqrt{\frac{2d\log(em/d)}{m}} + \sqrt{\frac{2\ln(4/\delta)}{m}}.$$

Repeating this argument for $-\ell(h,z)$, and then applying the union bound, we conclude

$$|\mathcal{L}_D(h) - \mathcal{L}_S(h)| \leq 2\sqrt{\frac{2d\log(em/d)}{m}} + \sqrt{\frac{2\ln(4/\delta)}{m}}.$$

But if the signed losses are bounded to each other, then so must the unsigned losses be bounded to each other:

$$||p_h - p_c| - |\hat{p}_h - \hat{p}_c|| \leq 2\sqrt{\frac{2d\log(em/d)}{m}} + \sqrt{\frac{2\ln(4/\delta)}{m}}.$$

□

This immediately implies that PAC learnable classes form a subset of CR learnable classes. In the remainder of this section, we show first that it is a strict subset. We then show that we can find a hypothesis class whose learnability is independent of ZFC, the common axiomatic system of mathematics. Following the work of Ben-David et al. (2019) who showed the first independence result for a machine learning problem, Reyzin (2019) suggested that independence results may hold in other models of machine learning, such as this one.

To demonstrate this, we make use of a recent result by Ben-David et al. (2019), who defined EMX-learnability and gave a class of functions whose learnability is independent of ZFC. First, we give the definition of EMX learnability.

Definition 5 (Ben-David et al., 2019)**.** *An (ϵ,δ)-EMX learner for a class of functions C is an algorithm that, for some integer $m = m(\epsilon,\delta)$, produces a hypothesis $c_S \in C$ for which*

$$\mathbb{P}_{S\sim D^m}[\mathbb{E}_D[c_S] \leq \sup_{c\in C} \mathbb{E}_D[c] - \epsilon] \leq \delta$$

for every distribution that is finitely-supported over the σ-algebra over all subsets of the input space.

This definition asks the learner to produce a (proper) function having as high distributional weight as possible. Ben-David et al. (2019) give the following surprising result.

Theorem 6 (Ben-David et al., 2019)**.** *EMX learnability of finite subsets of* $[0,1]$ *over finitely supported distributions (over the unit interval) is independent of ZFC.*

Now, we demonstrate that in some cases, EMX learning suffices to be able to CR learn:

Theorem 7. *Consider a space Z equipped with a total order. For any interval of Z (induced by the total order), suppose finite subsets of that interval are EMX learnable over finitely supported distributions. Then finite subsets of Z are CR learnable over finitely supported distributions.*

Proof. To CR learn finite subsets of Z, we first show that even though the finite subset may be arbitrarily large, it is contained in a small number of intervals of Z. We then show how an EMX learner can achieve any class ratio on Z (within a factor of ϵ) using these intervals.

First, define *heavy* points as those whose distributional mass is at least $\epsilon/4$ – note there can be no more than $4/\epsilon$ such points. Let the remaining points be called *light*. Let W_L be the distributional weight on the light points. There must be a subset of the heavy points whose weight adds up to within W_L of the target proportion p. Between these heavy points there are $4/\epsilon + 1$ intervals of Z containing light points, so there are also at most $4/\epsilon + 1$ sub-intervals that combined with the subset of the heavy points reach within $\epsilon/4$ of the target proportion: each additional light point included in a sub-interval can add only at most $\epsilon/4$ to the total mass included.

Hence, there exists a union of at most $4/\epsilon + 1$ intervals containing light points, that together with a subset of at most $4/\epsilon$ heavy points (around which we can also add intervals around the single points), whose weight adds up to the right proportion (within $\epsilon/4$). Since the VC dimension d of a union of $k = 8/\epsilon + 1$ intervals scales as $2k + 1$, Theorem 4 tells us that $\tilde{O}(d/\epsilon^2) = \tilde{O}(1/\epsilon^3)$ samples suffice to find such intervals via empirical risk minimization on the sample to within $\epsilon/4$, meaning we can find k such intervals with proportion within $\epsilon/2$ of the true proportion.

Finally, for each interval, we use an EMX learner to find a finite subset within that interval with mass approximating the total mass of the interval, within sufficiently small error; e.g. within $\epsilon^2/18$ suffices. Then the union of these finite subsets will be the finite subset of Z with mass within ϵ of the true proportion (except with probability δ). Since there are at most $8/\epsilon + 1$ such intervals to approximate, the union bound implies that the total error contributed in this part is also bounded by $\epsilon/2$. This is sufficient to achieve CR learnability. □

When Z is the unit interval over the reals, this means that EMX learning finite subsets is equivalent to CR learning finite subsets, and because EMX learning finite subsets is independent of ZFC, so is CR learning:

Corollary 8. *CR learnability of finite subsets of* $[0,1]$ *over finitely supported distributions is independent of ZFC.*

Proof. Using Theorem 6, it suffices to show that finite subsets of $[0,1]$ over finitely supported distributions over the unit interval are CR learnable if and only if they are EMX learnable. The "if" direction follows from Theorem 7.

Now for the other direction, since we know there exists a finite subset of the unit interval that contains the entire distributional mass, i.e. has true proportion 1, we can feed a CR learner $\hat{p} = 1$, along with the sample. If this learner can successfully CR learn finite subsets of $[0, 1]$, this learns a finite subset that maximizes the proportion, as desired.

Finally, we note that while the polynomial sample complexity requirement of CR learning is not present in EMX, this doesn't cause a problem. This is because we can always define a representation for which the finite sample complexity bound is polynomial. While this is not normally sensible, it is sufficient to establish the existence of a class for which the result holds. □

And when $Z = \mathbb{N}$, Theorem 7 implies that finite subsets of $\mathbb{N}$ are CR learnable. This is because intervals of $\mathbb{N}$ are either isomorphic to itself, or are finite numbers of points. Finite subsets of either of these are EMX learnable (Ben-David et al., 2019). But actually, it's worth noting something stronger, that finite subsets of $\mathbb{N}$ are *efficiently* learnable. The class of finite subsets of $\mathbb{N}$ has infinite VC dimension, implying it is a class that is efficiently CR learnable but is not PAC learnable.

Corollary 9. *Finite subsets of* $\mathbb{N}$ *are efficiently CR learnable over finitely supported distributions.*

Examining the proof of Theorem 7, it suffices to show that CR learning unions of intervals and then EMX learning each of those intervals can be done efficiently. But note that in the case of $\mathbb{N}$, first finding a set of intervals amongst a size m set of sample points using Theorem 4 is equivalent to choosing a subset of those points whose mass is within $\epsilon/4$ of the empirical proportion by empirical risk minimization. Given that subset of points, EMX learning for each interval is just selecting all of the subset, so it suffices to perform ERM on the original set of m sample points. This can be done efficiently in time polynomial in $1/\epsilon$ using a dynamic program for subset sum.

4. Comparing Efficient PAC to Efficient CR Learning

In this section, we consider how requiring efficiency of the learning algorithm changes the relationship between CR and PAC. In the previous section we showed that it is easier to CR learn than it is to PAC learn, in the sense that PAC is a strict subset of CR. In this section, we show that this relationship does not hold between efficient CR and efficient PAC. From Corollary 9, we already know that there are hypothesis classes with infinite VC dimension that are efficiently CR learnable but not efficiently PAC learnable, but here we show that there are classes that are hard to CR learn even though they are efficiently PAC learnable. Moreover, the only classes that are efficiently CR learnable that are not efficiently PAC learnable have infinite VC dimension:

Theorem 10. *Suppose* NP $\neq$ RP*. Then if a hypothesis class H with finite VC dimension is efficiently CR learnable, it is also efficiently (properly) PAC learnable.*

Proof. Let H be CR learnable 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 we need only a polynomial sample size to use uniform convergence (Theorem 4). This only requires $\mathrm{VC}(H) = \mathrm{poly}(1/\epsilon, 1/\delta, n, \mathrm{size}(c))$, where $\mathrm{VC}(H)$ denotes the VC dimension of H. If H is finite, recall that $\mathrm{VC}(H) \leq \log |H|$. But since $\mathrm{size}(c) \geq \log |H|$, we certainly have $\mathrm{VC}(H) = \mathrm{poly}(\mathrm{size}(c))$. If H is infinite, then the size of the input and classifiers are unbounded. Then the assumption that $\mathrm{VC}(H) < \infty$ implies that it is also a constant, and therefore the bound given by Theorem 4 is indeed a polynomial. □

4.1 Hardness of CR Learning

We now give three examples of hypothesis classes that are hard to CR learn: parities, monotone disjunctions, and monotone conjunctions. Parities are functions $h_a : \{0,1\}^n \to \{0,1\}$ of the form $h_a(x) = a \cdot x \bmod 2$, i.e. the parity of the bits of a subset of x determined by the bit mask a. Monotone disjuctions are disjunctions on n variables without negations of the literals and likewise monotone conjunctions are conjunctions without negations. Each of these has linear VC dimension and is efficiently PAC learnable (Shalev-Shwartz & Ben-David, 2014).

We start by showing that parities are hard to CR learn, even when the subset of the bits is always amongst the first k bits of the input for k poly-logarithmic in n. This class has VC dimension linear in k. To do this, 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 CR learning is hard in this sense for parities:

Theorem 11. *For a hypothesis c, let D_c be 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.*

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

Proof. We use an oracle for CR learning parities from 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 CR learning parities 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 $\tilde{\ell}_i = 1$ [1].

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 CR learning. 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 Hoeffding bound,

$$\mathbb{P}\left[\left|\sum_i \tilde{\ell}_i - m/2\right| > m/4\right] \leq 2e^{-m/8}.$$

1. The oracle is undefined when the proportion of positive labels is not the true value $\hat{p}$. We may assume that the oracle returns an arbitrary hypothesis in this case.

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 Hoeffding bound,

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

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 11 implies the following:

Corollary 12. *Under the parity hardness assumption, there is no algorithm for efficiently CR learning parities on the first k bits for $k = \omega(\log^2 n)$.*

The above result relies on the hardness of PAC learning parities, rather than NP-hardness. We can also establish problems that are NP-hard to learn. To do this, we start by defining the consistency problem of a hypothesis class. For a hypothesis class C, the *consistency problem* for CR learning is the following: Given a multi-set X of points and an integer k, where each unique x_i in X appears some a_i times in X, is there a hypothesis $c \in C$ such that the size of the multi-set of points that are labeled 1 is exactly k? That is, is there a hypothesis c such that $\sum_{x_i : c(x_i)=1} a_i = k$? If there is such a c, we will say that c is consistent with X.

We reduce from the consistency problem to the learning problem, which is a slightly more involved reduction than in the classical PAC setting.

Theorem 13. *Suppose that the consistency problem for a hypothesis class C is NP-hard. There is no efficient algorithm for CR learning C unless* NP = RP.

Proof. It suffices to reduce from the consistency problem to the learning problem. Indeed, using an oracle to an efficient CR learner for C, we merely need to solve the consistency problem with high probability. Given an instance of the consistency problem with input multi-set $X = \{x_i\}$ and integer k, define a distribution D that outputs each unique x_i with probability proportional to a_i.

Set

$$\epsilon = \frac{1}{2|X|}.$$

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[2].

Certainly, if we accept, there is a consistent hypothesis by definition: we accept if an oracle outputs a consistent hypothesis. Conversely, if there is a consistent hypothesis c, then the oracle will accept: Let c be the consistent hypothesis. Since it is consistent, by the definition of D, $p_c = k/|X|$. 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

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

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

2. 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.

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

Now, we can show that there are classes with linear VC dimension that are NP-hard to learn. In particular, we start with monotone disjunctions. Recall a monotone disjunction is a disjunction over n variables without negations, i.e. $\bigvee_{j \in J} x^j$ for a subset $J \subset [n]$, where we use x^j to refer to the jth bit of a vector x. We reduce the consistency problem for monotone disjunctions from a decision problem we will refer to as Exact Partial Set Cover, which asks, given a universe U, a collection of subsets $S \subset 2^U$ and an integer k, is there a subfamily $S' \subset S$ such that $|\cup_{s \in S'} s| = k$.

Lemma 14. Exact Partial Set Cover *is NP-hard.*

Proof. The reduction is from the well-known NP-hard problem Exact Cover by 3-Sets, X3C, which asks, given a universe U of exactly $3t$ elements, and a collection S of 3-element subsets of U, if there is a sub-collection $S' \subset S$ of size t such that $\cup_{s \in S'} s = U$. For each 3-element subset $s_i \in S$, define auxiliary elements $z_{i,1}, \ldots, z_{i,\ell}$, with ℓ to be determined shortly. Given an instance of X3C, we construct an instance of Exact Partial Set Cover. The universe we define as $U \cup \{z_{i,j}\}$ and the collection of subsets we define as the collection $\{s_i \cup \{z_{i,1}, \ldots, z_{i,\ell}\}\}$. Finally, let $k = |U| + \ell t$. This is a polynomial-time reduction as long as ℓ is only polynomially large.

If there is an exact cover by 3-sets, then the corresponding collection of subsets will have union exactly $|U| + \ell t$: The non-auxiliary elements of this union must cover U, and since it is a size t collection, there are ℓt auxiliary elements in the union. Conversely, if there is an exact partial set cover of size b whose union is size $k = |U| + \ell t$, this collection must be exactly size t: If $b > t$, then the size of the union, which is at least ℓb, must be strictly larger than k as long as $\ell > |U|$. And if $b < t$, then the size of the union is no more than $|U| + \ell b$, strictly smaller than k. Thus the corresponding collection of subsets must be an exact cover, as it is size t and its union is size $|U|$. □

Corollary 15. *CR learning monotone disjunctions is NP-hard.*

Proof. Using Theorem 13, it suffices to reduce the consistency problem from Exact Partial Set Cover. Given an instance (U, S, k) of this problem, we construct an instance of the consistency problem $m = |U|$ examples x_i, each with $|S|$ bits. Without loss of generality, we assume that $U = \{1, \ldots, m\}$. We define the jth bit of x_i to be 1 if the jth set s_j of the collection includes i, and otherwise we set the bit to be 0. We set the number of points to be labeled 1 as k.

The disjunction $d(x) = x^{j_1} \vee \ldots \vee x^{j_\ell}$ will label exactly k of these m examples positively if and only if the union of the collection $S' = \{s_{j_1}, \ldots, s_{j_\ell}\}$ is size exactly k: If $d(x_i) = 1$, then $x_i^j = 1$ for at least one $j \in \{j_1, \ldots, j_\ell\}$, i.e. i is in S'. And if $d(x_i) = 0$, then $x_i^j = 0$ for all $j \in \{j_1, \ldots, j_\ell\}$, i.e. i is not in any of the sets in S'. Thus, the number of examples that d labels positively is the size of the union of S'. □

Using a similar construction, it can be shown that learning monotone conjunctions is also NP-hard:

Corollary 16. *CR learning monotone conjunctions is NP-hard.*

5. Efficiently CR Learnable Classes

In Section 4.1, we gave a number of examples of classes that are hard to CR learn. We now turn to examples of classes that can be efficiently CR learned. Certainly, as long as labelings in a given hypothesis class are efficiently enumerable, then finite classes H are certainly CR learnable 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 CR learnable. But there are also examples of classes with finite but larger VC dimension that can be efficiently CR learned.

Consider the following hypothesis class 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 $VC(H_k) = k$. For k sufficiently small, this class is efficiently learnable:

Observation 17. *There is an $O(2^k m)$ time algorithm for CR learning H_k.*

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 CR learning H_k 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 easy to learn under the uniform distribution.

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 18. *The class of half-spaces in n dimensions is CR learnable 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 4 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.

Our goal will be to find such a vector v such that 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 while still being able to find a hyperplane that separates every pair of consecutive projected points.

It suffices to choose v uniformly at random. To show this, 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 angle between v and ℓ so that the distance as a function of n, let's call this $d'(n)$, between the two projected points is not too small. In particular, suppose $d'(n) = o(1/2^{n^c})$ for all constants $c > 0$. A distance of $o(1/2^{n^c})$ would require $\omega(n^c)$ bits to distinguish the two points. For x and y distance d apart, this maximum angle between v and ℓ is $\sin^{-1}(d'(n)/d) = O\left(d'(n)/d\right)$ using the Taylor approximation for $\sin^{-1}(x)$.

There are $O(2^{n^2})$ pairs of points since the points come from $\{0,1\}^n$. Then the total sum of angles that would result in any pair of projected points being too close is $O\left(\frac{2^{n^2}d'(n)}{d}\right)$, which goes to 0 exponentially quickly in n because d is at least a constant and $d'(n) = o(1/2^{n^2})$. 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 ensures with high probability that all examples are distinct, and therefore no two examples project more than exponentially close 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. □

Introducing noise to CR learning and PAC learning changes the relationship between the two models. For example, PAC learning parities with unknown η white-label noise is hard under the uniform distribution, as discussed above, but CR learning parities 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 19. *The class of parities is CR learnable 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 3 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.

We give some initial results into a learning theory for this task, including that in the case of finite VC dimension, classes that are efficiently CR learnable are automatically also efficiently properly PAC learnable. On the other hand, we give an independence result that implies that CR learning will not admit a nice VC-like characterization. We also give examples where it is possible to efficiently CR learn, which may be surprising given that this is a low-information setting, including half-spaces under the uniform distribution.

These results imply that the classes that are PAC learnable form a strict subset of those classes that are CR learnable, but that the only classes that are CR learnable but not PAC learnable have infinite VC dimension (e.g. finite subsets of $\mathbb{N}$). Moreover, there are classes that are efficiently PAC learnable but not efficiently CR learnable (e.g. monotone disjunctions) and vice versa (e.g. finite subsets of $\mathbb{N}$), but again in the latter case these must have infinite VC dimension.

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.

Acknowledgments

We thank Avrim Blum for pointing out that in the version of this paper that was published in IJCAI 2017 (Fish & Reyzin, 2017) the proof and statement of Theorem 5 of that version are incorrect; in addition to including new results and renaming the model from "LLP" to "CR" learning (one of many helpful suggestions from the JAIR reviewers), we've corrected this error herein. This work was supported in part by NSF awards CCF-1848966 and CCF-1934915.

References

Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., & Yehudayoff, A. (2019). Learnability can be undecidable. *Nature Machine Intelligence*, *1*(1), 44.

Blum, A., Kalai, A., & Wasserman, H. (2003). Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, *50*(4), 506–519.

Chen, B., Chen, L., Ramakrishnan, R., & Musicant, D. R. (2006). Learning from aggregate views. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, p. 3.

de Freitas, N., & Kück, H. (2005). 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*, pp. 332–339.

Du Plessis, M. C., & Sugiyama, M. (2014). Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, *50*, 110–119.

Fish, B., Huang, Y., & Reyzin, L. (2016). Recovering social networks by observing votes. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pp. 376–384.

Fish, B., & Reyzin, L. (2017). On the complexity of learning from label proportions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 1675–1681.

Hernández-González, J., Inza, I., & Lozano, J. A. (2013). Learning bayesian network classifiers from label proportions. *Pattern Recognition*, *46*(12), 3425–3440.

Iyer, A. S., Nath, J. S., & Sarawagi, S. (2014). 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*, pp. 530–538.

Iyer, A. S., Nath, J. S., & Sarawagi, S. (2016). 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*, pp. 925–934.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press.

Musicant, D. R., Christensen, J. M., & Olson, J. F. (2007). 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*, pp. 252–261.

Patrini, G., Nock, R., Caetano, T., & Rivera, P. (2014). (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*, pp. 190–198.

Quadrianto, N., Smola, A. J., Caetano, T. S., & Le, Q. V. (2009). Estimating labels from label proportions. *Journal of Machine Learning Research*, *10*, 2349–2374.

Reyzin, L. (2019). Unprovability comes to machine learning. *Nature*.

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

Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, *14*(1), 21–41.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Stolpe, M., & Morik, K. (2011). 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*, pp. 349–364.

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

Wojtusiak, J., Irvin, K., Birerdinc, A., & Baranova, A. V. (2011). 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*, pp. 84–89.

Yu, F. X., Choromanski, K., Kumar, S., Jebara, T., & Chang, S.-F. (2014). On learning from label proportions. *arXiv preprint arXiv:1402.5902*.

Yu, F. X., Liu, D., Kumar, S., Jebara, T., & Chang, S. (2013). $\propto$-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*, pp. 504–512.