On statistical learning via the lens of compression

Ofir David
Department of Mathematics
Technion - Israel Institute of Technology
ofirdav@tx.technion.ac.il

Shay Moran
Department of Computer Science
Technion - Israel Institute of Technology
shaymrn@cs.technion.ac.il

Amir Yehudayoff
Department of Mathematics
Technion - Israel Institute of Technology
amir.yehudayoff@gmail.com

Abstract

This work continues the study of the relationship between sample compression schemes and statistical learning, which has been mostly investigated within the framework of binary classification. The central theme of this work is establishing equivalences between learnability and compressibility, and utilizing these equivalences in the study of statistical learning theory. We begin with the setting of multiclass categorization (zero/one loss). We prove that in this case learnability is equivalent to compression of logarithmic sample size, and that uniform convergence implies compression of constant size. We then consider Vapnik's general learning setting: we show that in order to extend the compressibility-learnability equivalence to this case, it is necessary to consider an approximate variant of compression. Finally, we provide some applications of the compressibility-learnability equivalences.

1 Introduction

This work studies statistical learning theory using the point of view of compression. The main theme in this work is *establishing equivalences between learnability and compressibility, and making an effective use of these equivalences to study statistical learning theory.*

In a nutshell, the usefulness of these equivalences stems from that compressibility is a combinatorial notion, while learnability is a statistical notion. These equivalences, therefore, translate statistical statements to combinatorial ones and vice versa. This translation helps to reveal properties that are otherwise difficult to find, and highlights useful guidelines for designing learning algorithms.

We first consider the setting of *multiclass categorization*, which is used to model supervised learning problems using the zero/one loss function, and then move to *Vapnik's general learning setting* [23], which models many supervised and unsupervised learning problems.

Zero/one loss function (Section 3) This is the setting in which sample compression schemes were defined by Littlestone and Warmuth [16], as an abstraction of a common property of many learning algorithms. For more background on sample compression schemes, see e.g. [16, 8, 9, 22].

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

We use an agnostic version of sample compression schemes, and show that learnability is equivalent to some sort of compression. More formally, that any learning algorithm can be transformed to a compression algorithm, compressing a sample of size m to a sub-sample of size roughly $\log(m)$, and that such a compression algorithm implies learning. This statement is based on arguments that appear in [16, 10, 11]. We conclude this part by describing some applications:

(i) Equivalence between PAC and agnostic PAC learning from a statistical perspective (i.e. in terms of sample complexity). For binary-labelled classes, this equivalence follows from basic arguments in Vapnik-Chervonenkis (VC) theory, but these arguments do not seem to extend when the number of labels is large.

(ii) A dichotomy for sample compression - if a non-trivial compression exists (e.g. compressing a sample of size m to a sub-sample of size $m^{0.99}$), then a compression to logarithmic size exists (i.e. to a sub-sample of size roughly $\log m$). This dichotomy is analogous to the known dichotomy concerning the growth function of binary-labelled classes: the growth function is either polynomial (when the VC dimension is finite), or exponential (when the VC dimension is infinite).

(iii) Compression to constant size versus uniform convergence - every class with the uniform convergence property has a compression of constant size. The proof has two parts. The first part, which is based on arguments from [18], shows that finite graph dimension (a generalization of VC dimension for multiclass categorization [19]) implies compression of constant size. The second part, which uses ideas from [1, 24, 7], shows that the uniform convergence rate is captured by the graph dimension. In this part we improve upon the previously known bounds.

(iv) Compactness for learning - if finite sub-classes of a given class are learnable, then the class is learnable as well. Again, for binary-labelled classes, such compactness easily follows from known properties of VC dimension. For general multi-labeled classes we derive this statement using a corresponding compactness property for sample compression schemes, based on the work by [2].

General learning setting (Section 4). We continue with investigating general loss functions. This part begins with a simple example in the context of linear regression, showing that for general loss functions, learning is not equivalent to compression. We then consider an approximate variant of compression schemes, which was used by [13, 12] in the context of classification, and observe that learnability is equivalent to possessing an approximate compression scheme, whose size is roughly the statistical sample complexity. This is in contrast to (standard) sample compression schemes, for which the existence of such an equivalence (under the zero/one loss) is a long standing open problem, even in the case of binary classification [25]. We conclude the paper by showing that - unlike for zero/one loss functions - for general loss functions, PAC learnability and agnostic PAC learnability are *not* equivalent. In fact, this is derived for a loss function that takes just three values. The proof of this non-equivalence uses Ramsey theory for hypergraphs. The combinatorial nature of compression schemes allows to clearly identify the place where Ramsey theory is helpful. More generally, the study of statistical learning theory via the lens of compression may shed light on additional useful connections with different fields of mathematics.

We begin our investigation by breaking the definition of sample compression schemes into two parts. The first part (which may seem useless at first sight) is about *selection schemes*. These are learning algorithms whose output hypothesis depends on a selected small sub-sample of the input sample. The second part of the definition is the sample-consistency guarantee; so, sample compression schemes are selection schemes whose output hypothesis is consistent with the input sample. We then show that selection schemes of small size do not overfit in that their empirical risk is close to their true risk. Roughly speaking, this shows that for selection schemes there are no surprises: “what you see is what you get”.

2 Preliminaries

The definitions we use are based on the textbook [22].

Learnability and uniform convergence

A learning problem is specified by a set $\mathcal{H}$ of hypotheses, a domain $\mathcal{Z}$ of examples, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$. To ease the presentation, we shall only discuss loss functions that are bounded from above by 1, although the results presented here can be extended to more general loss functions. A sample S is a finite sequence $S = (z_1, \ldots, z_m) \in \mathcal{Z}^m$. A *learning algorithm* is a mapping that gets as an input a sample and outputs an hypothesis h.

In the context of supervised learning, hypotheses are functions from a domain $\mathcal{X}$ to a label set $\mathcal{Y}$, and the examples domain is the cartesian product $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. In this context, the loss $\ell(h, (x, y))$ depends only on $h(x)$ and y, and therefore in this case we it is modelled as a function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

Given a distribution $\mathcal{D}$ on $\mathcal{Z}$, the *risk* of an hypothesis $h : \mathcal{X} \to \mathcal{Y}$ is its expected loss: $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} \left[\ell(h, z)\right]$. Given a sample $S = (z_1, \ldots, z_m)$, the *empirical risk* of an hypothesis h is $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z)$.

An *hypothesis class* $\mathcal{H}$ is a set of hypotheses. A distribution $\mathcal{D}$ is *realizable* by $\mathcal{H}$ if there exists $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$. A sample S is *realizable* by $\mathcal{H}$ if there exists $h \in \mathcal{H}$ such that $L_S(h) = 0$.

A hypothesis class $\mathcal{H}$ has *the uniform convergence property*[1] if there exists a rate function $d : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta > 0$ and distribution $\mathcal{D}$ over $\mathcal{Z}$, if S is a sample of $m \geq d(\epsilon, \delta)$ i.i.d. pairs generated by $\mathcal{D}$, then with probability at least $1-\delta$ we have: $\forall h \in \mathcal{H}$ $|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon$.

The class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a learner A and a rate function $d : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta > 0$ and distribution $\mathcal{D}$ over $\mathcal{Z}$, if S is a sample of $m \geq d(\epsilon, \delta)$ i.i.d. pairs generated by $\mathcal{D}$, then with probability at least $1 - \delta$ we have $L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$. The class $\mathcal{H}$ is *PAC learnable* if this condition holds for every realizable distribution $\mathcal{D}$. The parameter ϵ is referred to as the *error* parameter and δ as the *confidence* parameter.

Note that the uniform convergence property implies agnostic PAC learnability with the same rate via any learning algorithm which outputs $h \in \mathcal{H}$ that minimizes the empirical risk, and that agnostic PAC learnability implies PAC learnability with the same rate.

Selection and compression schemes

The variants of sample compression schemes that are discussed in this paper, are based on the following object, which we term *selection scheme*. We stress here that unlike sample compression schemes, selection schemes are not associated with any hypothesis class.

A selection scheme is a pair (κ, ρ) of maps for which the following holds:

- κ is called the selection map. It gets as an input a sample S and outputs a pair (S', b) where S' is a sub-sample[2] of S and b is a finite binary string, which we think of as side information.
- ρ is called the reconstruction map. It gets as an input a pair (S', b) of the same type as the output of κ and outputs an hypothesis h.

The size of (κ, ρ) on a given input sample S is defined to be $|S'| + |b|$ where $\kappa(S) = (S', b)$. For an input size m, we denote by $k(m)$ the maximum size of the selection scheme on all inputs S of size at most m. The function $k(m)$ is called the *size* of the selection scheme. If $k(m)$ is uniformly bounded by a constant, which does not depend on m, then we say that the selection scheme has a constant size; otherwise, we say that it has a variable size.

The definition of selection schemes is very similar to that of sample compression schemes. The difference is that sample compression schemes are defined relative to a fixed hypothesis class with respect to which they are required to have "correct" reconstructions whereas selection schemes do not provide any correctness guarantee. The distinction between the 'selection' part and the 'correctness' part is helpful for our presentation, and also provides some more insight into these notions.

A selection scheme (κ, ρ) is a *sample compression scheme* for $\mathcal{H}$ if for every sample S that is realizable by $\mathcal{H}$, $L_S\left(\rho\left(\kappa\left(S\right)\right)\right) = 0$. A selection scheme (κ, ρ) is an *agnostic sample compression scheme* for $\mathcal{H}$ if for every sample S, $L_S\left(\rho\left(\kappa\left(S\right)\right)\right) \leq \inf_{h \in \mathcal{H}} L_S(h)$.

In the following sections, we will see different manifestations of the statement "compression $\Rightarrow$ learning". An essential part of these statements boils down to a basic property of selection schemes,

[1] We omit the dependence on the loss function ℓ from this and similar definitions, since ℓ is clear from the context.

[2] That is, if $S = (z_1, \ldots, z_m)$ then S' is of the form $(z_{i_1}, \ldots, z_{i_\ell})$ for $1 \leq i_1 < \ldots < i_\ell \leq m$.

that as long as $k(m)$ is sufficiently smaller than m, a selection scheme based learner does not overfit its training data (the proof appears in the full version of this paper).

Theorem 2.1 ([22, Theorem 30.2])**.** *Let (κ, ρ) be a selection scheme of size $k = k(m)$, and let $A(S) = \rho(\kappa(S))$. Then, for every distribution $\mathcal{D}$ on $\mathcal{Z}$, integer m such that $k \le m/2$, and $\delta > 0$, we have*

$$\Pr_{S\sim\mathcal{D}^m}\left[\left|L_{\mathcal{D}}\left(A\left(S\right)\right) - L_S\left(A\left(S\right)\right)\right| \ge \sqrt{\epsilon \cdot L_S\left(A\left(S\right)\right)} + \epsilon\right] \le \delta,$$

where $\epsilon = 50\frac{k\log(m/k)+\log(1/\delta)}{m}$.

3 Zero/one loss functions

In this section we consider the zero/one loss function, which models categorization problems. We study the relationships between uniform convergence, learnability, and sample compression schemes under this loss. Subsection 3.1 establishes equivalence between learnability and compressibility of a sublinear size. In Subsection 3.2 we use this equivalence to study the relationships between the properties of uniform convergence, PAC, and agnostic learnability. In Subsection 3.2.1 we show that agnostic learnability is equivalent to PAC learnability, In Subsection 3.2.2 we observe a dichotomy concerning the size of sample compression schemes, and use it to establish a compactness property of learnability. Finally, in Subsection 3.2.3 we study an extension of the Littlestone-Floyd-Warmuth conjecture concerning an equivalence between learnability and sample compression schemes of fixed size.

3.1 Learning is equivalent to sublinear compressing

The following theorem shows that if $\mathcal{H}$ has a sample compression scheme of size $k = o(m)$, then it is learnable. Its proof appears in the full version of this paper.

Theorem 3.1 (Compressing implies learning [16])**.** *Let (κ, ρ) be a selection scheme of size k, let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{D}$ be a distribution on $\mathcal{Z}$.*

1. *If (κ, ρ) is a sample compression scheme for $\mathcal{H}$, and m is such that $k(m) \le m/2$, then*

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\rho\left(\kappa\left(S\right)\right)\right) > 50\frac{k\log\frac{m}{k} + k + \log\frac{1}{\delta}}{m}\right) < \delta.$$

2. *If (κ, ρ) is an agnostic sample compression scheme for $\mathcal{H}$, and m is such that $k(m) \le m/2$, then*

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}\left(\rho\left(\kappa\left(S\right)\right)\right) > \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 100\sqrt{\frac{k\log\frac{m}{k} + k + \log\frac{1}{\delta}}{m}}\right) < \delta.$$

The following theorem shows that learning implies compression. We present its proof in the full version of this paper.

Theorem 3.2 (Learning implies compressing)**.** *Let $\mathcal{H}$ be an hypothesis class.*

1. *If $\mathcal{H}$ is agnostic PAC learnable with learning rate $d(\epsilon, \delta)$, then it is PAC learnable with the same learning rate.*
2. *If $\mathcal{H}$ is PAC learnable with learning rate $d(\epsilon, \delta)$, then it has a sample compression scheme of size $k(m) = O(d_0\log(m)\log\log(m) + d_0\log(m)\log(d_0))$, where $d_0 = d(1/3, 1/3)$.*
3. *If $\mathcal{H}$ has a sample compression scheme of size $k(m)$, then it has an agnostic sample compression scheme of the same size.*

Remark. *The third part in Theorem 3.2 does not hold when the loss function is general. In Section 4 we show that even if the loss function takes three possible values, then there are instances where a class has a sample compression scheme but not an agnostic sample compression scheme.*

3.2 Applications

3.2.1 Agnostic and PAC learnability are equivalent

Theorems 3.1 and 3.2 imply that if $\mathcal{H}$ is PAC learnable, then it is agnostic PAC learnable. Indeed, a summary of the implications between learnability and compression given by Theorems 3.1 and 3.2 gives:

- An agnostic learner with rate $d\left(\epsilon, \delta\right)$ implies a PAC learner with rate $d\left(\epsilon, \delta\right)$.
- A PAC learner with rate $d\left(\epsilon, \delta\right)$ implies a sample compression scheme of size $k\left(m\right) = O\left(d_0 \cdot \log\left(m\right) \log\left(d_0 \cdot \log\left(m\right)\right)\right)$ where $d_0 = d(1/3, 1/3)$.
- A sample compression scheme of size $k\left(m\right)$ implies an agnostic sample compression scheme of size $k\left(m\right)$.
- An agnostic sample compression scheme of size $k\left(m\right)$ implies an agnostic learner with error $\epsilon\left(d, \delta\right) = 100\sqrt{\frac{k(d)\log\frac{d}{k(d)} + k(d) + \log\frac{1}{\delta}}{d}}$.

Thus, for multiclass categorization problems, agnostic learnability and PAC learnability are equivalent. When the size of the label set $\mathcal{Y}$ is $O(1)$, this equivalence follows from previous works that studied extensions of the VC dimension to multiclass categorization problems [24, 3, 19, 1]. These works show that PAC learnability and agnostic PAC learnability are equivalent to the uniform convergence property, and therefore any ERM algorithm learns the class. Recently, [7] separated PAC learnability and uniform convergence for large label sets by exhibiting PAC learnable hypothesis classes that do not satisfy the uniform convergence property. In contrast, this shows that the equivalence between PAC and agnostic learnability remains valid even when $\mathcal{Y}$ is large.

3.2.2 A dichotomy and compactness

Let $\mathcal{H}$ be an hypothesis class. Assume that $\mathcal{H}$ has a sample compression scheme of size, say, $m/500$ for some large m. Therefore, by Theorem 3.1, $\mathcal{H}$ is weakly PAC learnable with confidence $2/3$, error $1/3$, and $O(1)$ examples. Now, Theorem 3.2 implies that $\mathcal{H}$ has a sample compression scheme of size $k(m) \leq O(\log(m)\log\log(m))$. In other words, the following dichotomy holds: every hypothesis class $\mathcal{H}$ either has a sample compression scheme of size $k(m) = O(\log(m)\log\log(m))$, or any sample compression scheme for it has size $\Omega(m)$.

This dichotomy implies the following compactness property for learnability under the zero/one loss.

Theorem 3.3. *Let $d \in \mathbb{N}$, and let $\mathcal{H}$ be an hypothesis class such that each finite subclass of $\mathcal{H}$ is learnable with error $1/3$, confidence $2/3$ and d examples. Then $\mathcal{H}$ is learnable with error $1/3$, confidence $2/3$ and $O(d\log^2(d)\log\log(d))$ examples.*

When $\mathcal{Y} = \{0, 1\}$, the theorem follows by the observing that if every subclass of $\mathcal{H}$ has VC dimension at most d, then the VC dimension of $\mathcal{H}$ is at most d. We are not aware of a similar argument that applies for a general label set. A related challenge, which was posed by [6], is to find a "combinatorial" parameter, which captures multiclass learnability like the VC dimension captures it in the binary-labeled case.

A proof of Theorem 3.3 appears in the full version of this paper. It uses an analogous[3] compactness property for sample compression schemes proven by [2].

3.2.3 Uniform convergence versus compression to constant size

Since the introduction of sample compression schemes by [16], they were mostly studied in the context of binary-labeled hypothesis classes (the case $\mathcal{Y} = \{0, 1\}$). In this context, a significant number of works were dedicated to studying the relationship between VC dimension and the minimal size of a compression scheme (e.g. [8, 14, 9, 2, 15, 4, 21, 20, 17]). Recently, [18] proved that any class of VC dimension d has a compression scheme of size exponential in the VC dimension. Establishing whether a compression scheme of size linear (or even polynomial) in the VC dimension remains open [9, 25].

[3] Ben-David and Litman proved a compactness result for sample compression schemes when $\mathcal{Y} = \{0, 1\}$, but their argument generalizes for a general $\mathcal{Y}$.

This question has a natural extension to multiclass categorization: Does every hypothesis class $\mathcal{H}$ have a sample compression scheme of size $O(d)$, where $d = d_{PAC}(1/3, 1/3)$ is the minimal sample complexity of a weak learner for $\mathcal{H}$? In fact, in the case of multiclass categorization it is open whether there is a sample compression scheme of size depending only on d.

We show here that the arguments from [18] generalize to uniform convergence.

Theorem 3.4. *Let $\mathcal{H}$ be an hypothesis class with uniform convergence rate $d^{UC}(\epsilon, \delta)$. Then $\mathcal{H}$ has a sample compression scheme of size $\exp(d)$, where $d = d^{UC}(1/3, 1/3)$.*

The proof of this theorem uses the notion of the graph dimension, which was defined by [19].

Theorem 3.4 is proved using the following two ingredients. First, the construction in [18] yields a sample compression scheme of size $\exp(\dim_G(\mathcal{H}))$. Second, the graph dimension determines the uniform convergence rate, similarly to that the VC dimension does it in the binary-labeled case.

Theorem 3.5. *Let $\mathcal{H}$ be an hypothesis class, let $d = \dim_G(\mathcal{H})$, and let $d^{UC}(\epsilon, \delta)$ denote the uniform convergence rate of $\mathcal{H}$. Then, there exist constants C_1, C_2 such that*

$$C_1 \cdot \frac{d + \log(1/\delta) - C_1}{\epsilon^2} \le d^{UC}(\epsilon, \delta) \le C_2 \cdot \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}.$$

Parts of this result are well-known and appear in the literature: The upper bound follows from Theorem 5 of [7], and the core idea of the argument dates back to the articles of [1] and of [24]. A lower bound with a worse dependence on ϵ follows from Theorem 9 of [7]. A proof of Theorem 3.5 appears in the full version of this paper.

4 General loss functions

We have seen that in the case of the zero/one loss function, an existence of a sublinear sample compression scheme is equivalent to learnability. It is natural to ask whether this phenomenon extends to other loss functions. The direction "compression $\Longrightarrow$ learning" remains valid for general loss functions. In contrast, as will be discussed in this section, the other direction fails for general loss functions.

However, a natural adaptation of sample compression schemes, which we term *approximate sample compression schemes*, allows the extension of the equivalence to arbitrary loss functions. Approximate compression schemes were previously studied in the context of classification (e.g. [13, 12]). In Subsection 4.1 we argue that in general sample compression schemes are not equivalent to learnability; specifically, there is no agnostic sample compression scheme for linear regression. In Subsection 4.2 we define approximate sample compression schemes and establish their equivalence with learnability.

Finally, in Subsection 4.3 we use this equivalence to demonstrate classes that are PAC learnable but not agnostic PAC learnable. This manifests a difference with the zero/one loss under which agnostic and PAC learning are equivalent (see 3.2.1). It is worth noting that the loss function we use to break the equivalence takes only three values (compared to the two values of the zero/one loss function).

4.1 No agnostic compression for linear regression

We next show that in the setup of linear regression, which is known to be agnostic PAC learnable, there is no agnostic sample compression scheme. For convenience, we shall restrict the discussion to zero-dimensional linear regression. In this setup[4], the sample consists of m examples $S = (z_1, z_2, \ldots, z_m) \in [0, 1]^m$, and the loss function is defined by $\ell(h, z) = (h - z)^2$. The goal is to find $h \in \mathbb{R}$ which minimizes $L_S(h)$. The empirical risk minimizer (ERM) is exactly the average $h^* = \frac{1}{m} \sum_i z_i$, and for every $h \neq h^*$ we have $L_S(h) > L_S(h^*)$. Thus, an agnostic sample compression scheme in this setup should compress S to a subsequence and a binary string of side information, from which the average of S can be reconstructed. We prove that there is no such compression.

Theorem 4.1. *There is no agnostic sample compression scheme for zero-dimensional linear regression with size $k(m) \le m/2$.*

[4] One may think of X as a singleton.

The proof appears in the full version of this paper. The idea is to restrict our attention to sets $\Omega \subseteq [0,1]$ for which every subset of Ω has a distinct average. It follows that any sample compression scheme for samples from Ω must perform a compression that is information theoretically impossible.

4.2 Approximate sample compression schemes

The previous example suggests the question of whether one can generalize the definition of compression to fit problems where the loss function is not zero/one. Taking cues from PAC and agnostic PAC learning, we consider the following definition. We say that the selection scheme (κ, ρ) is an ϵ-*approximate* sample compression scheme for $\mathcal{H}$ if for every sample S that is realizable by $\mathcal{H}$, $L_S(\rho(\kappa(S))) \leq \epsilon$. It is called an ϵ-*approximate* agnostic sample compression scheme for $\mathcal{H}$ if for every sample S, $L_S(\rho(\kappa(S))) \leq \inf_{h \in \mathcal{H}} L_S(h) + \epsilon$.

Let us start by revisiting the case of zero-dimensional linear regression. Even though it does not have an agnostic compression scheme of sublinear size, it does have an ϵ-approximate agnostic sample compression scheme of size $k = O(\log(1/\epsilon)/\epsilon)$ which we now describe.

Given a sample $S = (z_1, \ldots, z_m) \in [0,1]$, the average $h^* = \sum_{i=1}^m z_i/m$ is the ERM of S. Let

$$L^* = L(h^*) = \sum_{i=1}^m z_i^2/m - \left(\sum_{i=1}^m z_i/m\right)^2.$$

It is enough to show that there exists a sub-sample $S' = (z_{i_1}, \ldots, z_{i_\ell})$ of size $\ell = \lceil 1/\epsilon \rceil$ such that $L_S\left(\sum_{j=1}^{\ell} z_{i_j}/\ell\right) \leq L^* + \epsilon$. It turns out that picking S' at random suffices. Let $Z_1, \ldots, Z_\ell$ be independent random variables that are uniformly distributed over S and let $H = \frac{1}{\ell}\sum_{i=1}^{\ell} Z_i$ be their average. Thus, $\mathbb{E}[H] = h^*$ and $\mathbb{E}\big[L_S(H)\big] = L^* + \mathrm{Var}[H] \leq L^* + \epsilon$. In particular, this means that there exists some sub-sample of size ℓ whose average has loss at most $L^* + \epsilon$. Encoding such a sub-sample requires $O(\log(1/\epsilon)/\epsilon)$ additional bits of side information.

We now establish the equivalence between approximate compression and learning (the proof is similar to the proof of Theorem 3.1).

Theorem 4.2 (Approximate compressing implies learning)**.** *Let (κ, ρ) be a selection scheme of size k, let $\mathcal{H}$ be an hypothesis class, and let $\mathcal{D}$ be a distribution on $\mathcal{Z}$.*

1. *If (κ, ρ) is an ϵ-approximate sample compression scheme for $\mathcal{H}$, and m is such that $k(m) \leq m/2$, then*

$$\Pr_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(\rho(\kappa(S))) > \epsilon + 100\sqrt{\frac{k \log \frac{m}{k} + \log \frac{1}{\delta}}{m}}\right) < \delta.$$

2. *If (κ, ρ) is an ϵ-approximate agnostic sample compression scheme for $\mathcal{H}$, and m is such that $k(m) \leq m/2$, then*

$$\Pr_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(\rho(\kappa(S))) > \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon + 100\sqrt{\frac{k \log \frac{m}{k} + \log \frac{1}{\delta}}{m}}\right) < \delta.$$

The following Theorem shows that every learnable class has an approximate sample compression scheme. The proof of this theorem is straightforward - in contrast with the proof of the analog statement in the case of zero/one loss functions and compression schemes without error.

Theorem 4.3 (Learning implies approximate compressing)**.** *Let $\mathcal{H}$ be an hypothesis class.*

1. *If $\mathcal{H}$ is PAC learnable with rate $d(\epsilon, \delta)$, then it has an ϵ-approximate sample compression scheme of size $k \leq O(d \log(d))$ with $d = \min_{\delta < 1} d(\epsilon, \delta)$.*

2. *If $\mathcal{H}$ is agnostic PAC learnable with rate $d(\epsilon, \delta)$, then it has an ϵ-approximate agnostic sample compression scheme of size $k \leq O(d \log(d))$ with $d = \min_{\delta < 1} d(\epsilon, \delta)$.*

The proof appears in the full version of this paper.

4.3 A separation between PAC and agnostic learnability

Here we establish a separation between PAC and agnostic PAC learning under loss functions which take more than two values:

Theorem 4.4. *There exist hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and loss function $l : \mathcal{Y} \times \mathcal{Y} \to \{0, \frac{1}{2}, 1\}$ such that $\mathcal{H}$ is PAC learnable and not agnostic PAC learnable.*

The main challenge in proving this theorem is showing that $\mathcal{H}$ is not agnostic PAC learnable. We do this by showing that $\mathcal{H}$ does not have an approximate sample compression scheme. The crux of the argument is an application of Ramsey theory; the combinatorial nature of compression allows to identify the place where Ramsey theory is helpful. The proof appears in the full version of this paper.

5 Discussion and further research

The compressibility-learnability equivalence is a fundamental link in statistical learning theory. From a theoretical perspective this link can serve as a guideline for proving both negative/impossibility results, and positive/possibility results.

From the perspective of positive results, just recently, [5] relied on this paper in showing that every learnable problem is learnable with robust generalization guarantees. Another important example appears in the work of boosting weak learners [11] (see Chapter 4.2). These works follow a similar approach, that may be useful in other scenarios: (i) transform the given learner to a sample compression scheme, and (ii) utilize properties of compression schemes to derive the desired result. The same approach is also used in this paper in Section 3.2.1, where it is shown that PAC learning implies agnostic PAC learning under 0/1 loss; we first transform the PAC learner to a realizable compression scheme, and then use the realizable compression scheme to get an agnostic compression scheme that is also an agnostic learner. We note that we are not aware of a proof that directly transforms the PAC learner to an agnostic learner without using compression.

From the perspective of impossibility/hardness results, this link implies that to show that a problem is not learnable, it suffices to show that it is not compressible. In Section 4.3, we follow this approach when showing that PAC and agnostic PAC learnability are not equivalent for general loss functions.

This link may also have a practical impact, since it offers a thumb rule for algorithm designers; if a problem is learnable then it can be learned by a compression algorithm, whose design boils down to an intuitive principle "find a small insightful subset of the input data." For example, in geometrical problems, this insightful subset often appears on the boundary of the data points (see e.g. [12]).

References

[1] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of {0,...,n}-valued functions. *J. Comput. Syst. Sci.*, 50(1):74–86, 1995. 2, 5, 6

[2] Shai Ben-David and Ami Litman. Combinatorial Variability of Vapnik-Chervonenkis Classes with Applications to Sample Compression Schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998. 2, 5

[3] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. 5

[4] A. Chernikov and P. Simon. Externally definable sets and dependent pairs. *Israel Journal of Mathematics*, 194(1):409–425, 2013. 5

[5] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 772–814, 2016. 8

[6] A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *COLT*, volume 35, pages 287–316, 2014. 5

[7] Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015. 2, 5, 6

[8] S. Floyd. Space-Bounded Learning and the Vapnik-Chervonenkis Dimension. In *COLT*, pages 349–364, 1989. 1, 5

[9] Sally Floyd and Manfred K. Warmuth. Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. *Machine Learning*, 21(3):269–304, 1995. 1, 5

[10] Yoav Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995. 2

[11] Yoav Freund and Robert E. Schapire. *Boosting: Foundations and Algorithms*. Adaptive computation and machine learning. MIT Press, 2012. 2, 8

[12] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics. *CoRR*, abs/1502.06208, 2015. 2, 6, 8

[13] Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-Bayesian Compression Bounds on the Prediction Error of Learning Algorithms for Classification. *Machine Learning*, 59(1-2):55–76, 2005. 2, 6

[14] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth. Learning integer lattices. *SIAM J. Comput.*, 21(2):240–266, 1992. 5

[15] Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. 5

[16] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. *Unpublished*, 1986. 1, 2, 4, 5

[17] Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *COLT*, pages 77–92, 2013. 5

[18] Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, June 2016. 2, 5, 6

[19] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989. 2, 5, 6

[20] B. I. P. Rubinstein and J. H. Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. 5

[21] Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. H. Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *J. Comput. Syst. Sci.*, 75(1):37–59, 2009. 5

[22] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. 1, 2, 4

[23] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. 1

[24] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971. 2, 5, 6

[25] Manfred K. Warmuth. Compressing to VC dimension many points. In *COLT/Kernel*, pages 743–744, 2003. 2, 5