Average-Case Information Complexity of Learning

Ido Nachum IDON@TX.TECHNION.AC.IL
Technion - Israel Institute of Technology

Amir Yehudayoff AMIR.YEHUDAYOFF@GMAIL.COM
Technion - Israel Institute of Technology

Editors: Aurélien Garivier and Satyen Kale

Abstract

How many bits of information are revealed by a learning algorithm for a concept class of VC-dimension d? Previous works have shown that even for $d = 1$ the amount of information may be unbounded (tend to ∞ with the universe size). Can it be that all concepts in the class require leaking a large amount of information? We show that typically concepts do not require leakage. There exists a proper learning algorithm that reveals $O(d)$ bits of information for most concepts in the class. This result is a special case of a more general phenomenon we explore. If there is a low information learner when the algorithm *knows* the underlying distribution on inputs, then there is a learner that reveals little information on an average concept *without knowing* the distribution on inputs.

Keywords: Learning Theory, Information Theory.

1. Introduction

The high-level question that guides this paper is:

when is learning equivalent to compression?

Variants of this question were studied extensively throughout the years in many different contexts. Recently, its importance grew even further due to the growing complexity of learning tasks. In this work, we measure compression using information theory. Our main message is that, in the framework we develop, learning implies compression.

It is well-known that in many contexts, the ability to compress implies learnability. Here is a partial list of examples: sample compression schemes Littlestone and Warmuth (1986); Moran and Yehudayoff (2016), Occam's razor Blumer et al. (1987), minimum description length (Rissanen, 1978; Grünwald, 2007), and differential privacy Dwork et al. (2006, 2015b); Bassily et al. (2016); Rogers et al. (2016); Bassily et al. (2014). We refer the interested reader to Xu and Raginsky (2017); Bassily et al. (2018) for more details.

We use the setting of Xu and Raginsky (2017) and Bassily et al. (2018), where the value of interest is the mutual information $I(S; A(S))$ between the input sample S and the output of the learning algorithm $A(S)$. The above authors suggested that studying this notion may shed additional light on our understanding of the relations between compression and learning.

The rationale is that compression is, in many cases, an information theoretic notion, so it is natural to use information theory to quantify the amount of compression a learning algorithm performs. The quantity $I(S; A(S))$ is a natural information theoretic measure for the amount of compression

the algorithm performs. Additional motivation comes from the connections to privacy, which is about leaking little information while maintaining functionality.

In the information theoretic setting, Xu and Raginsky (2017) and Bassily et al. (2018) showed that for every learning algorithm for which the information $I(S;A(S))$ is much smaller than the sample size m, the true error and the empirical error are typically close. This highlights the following rule of thumb for designing learning algorithms: try to find an algorithm that has small empirical error and at the same time reveals little information for a given input.

What about the other direction? Is it true that *learning* $\Rightarrow$ *compression* in this context? Bassily et al. (2018) answered this question for the class of thresholds and Nachum et al. (2018) extended the result for classes of VC-dimension d (see Section 2 for notations).

Theorem 1 (Bassily et al. (2018); Nachum et al. (2018)) *For every d and every $m \geq 2d^2$, there exists a class $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ of VC-dimension d such that for any proper and consistent (possibly randomized) learning algorithm, there exists a hypothesis $h \in \mathcal{C}$ and a random variable X over $\mathcal{X}$ such that $I(S;A(S)) = \Omega(d \log\log(|\mathcal{X}|/d))$ where $S \sim (X, h(X))^m$.*

The theorem can be interpreted as saying that no, learning does not imply compression in this context. In some cases, for any consistent and proper algorithm, there is always a scenario in which a large amount of information is revealed.

In this work, we shift our attention from a *worst-case* analysis to an *average-case* analysis. In the average-case setting, we show that every prior distribution $\mathcal{P}$ over $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ of VC-dimension d admits an algorithm that *typically* reveals $O(d)$-bits of information on its input (there is an unbounded difference between the worst-case and the average-case).

learning $\Rightarrow$ *compression (on average)* (Theorem 2)

This result is a special case of a more general phenomenon we explore. If there is a low information learner when the algorithm *knows* the underlying distribution on inputs, then there is a learner that reveals little information on an average concept *without knowing* the distribution on inputs (Lemma 4).

The average-case framework is different from the standard worst-case PAC setting. In the standard model, the teacher (or nature) is thought of as being adversarial and is assumed to have perfect knowledge of the learner's strategy.

- From a practical point of view, it is not obvious that such strong assumptions about the environment should be made, since worst-case analysis seems to fail when trying to explain real-life learning algorithms.
- From a biological perspective, to survive, a living organism must perform many tasks (concept class). No human can perform well on all of them (worst case analysis). What matters for survival, is to be able to perform well on most tasks (average case learning).

The average-case framework we study also provides a general mechanism for proving upper bounds on the average sample complexity for classes of functions (not necessarily binary or with 0-1 loss). This framework ("The Information Game") allows the user the freedom to apply his prior knowledge when trying to solve a learning problem. For example, the user can pick only distributions that make sense in his setting (see Discussion).

Related Work

INFORMATION COMPLEXITY IN LEARNING

The mutual information bounds are implicit in the PAC-Bayes bounds (see a survey by McAllester (2013)).

More recent interest comes from applications in adaptive data analysis. In this setting, a user asks a series of queries over some data. Every new query the user decides to ask depends on the answers to the previous queries. Dwork et al. (2015a) used max-information and Feldman and Steinke (2018) used the information theoretic setting and proved generalization bounds for performing adaptive data analysis.

Asadi et al. (2018) applied the information theoretic setting for achieving generalization bounds that depend on the correlations between the functions in the class together with the dependence between the input and the output of the learning algorithm. They mostly investigated Gaussian processes.

AVERAGE-CASE LEARNING

Here is a brief survey of other works that deviate from the worst-case analysis of the PAC learning setting.

Haussler et al. (1994) studied how the sample complexity depends on properties of a prior distribution on the class $\mathcal{C}$ and over the sequence of examples the algorithm receives. Specifically, they studied the probability of an incorrect prediction for an optimal learning algorithm using the Shannon information gain. They also studied stability in the context they investigated.

Wan (2010) can be used as a survey of average-case learning of DNF-formulas. There, the formulas are sampled from the uniform distribution and the distribution over the domain is uniform as well.

Reischuk and Zeugmann (1999) considered the problem of learning monomials. They analyzed the average-case behavior of the Wholist algorithm with respect to the class of binomial distributions.

Finally, we note that many of the lower bounds on the sample complexity of learning algorithms can be casted in the "on average" language. In many cases, the lower bound is proved by choosing an appropriate distribution on the concept class $\mathcal{C}$.

CHANNEL CAPACITY

The information game is also relevant in the following information theoretic scenario. Player two wants to transmit a message through a noisy channel that has several states S and player one wants to prevent that by appropriately choosing S. In the game, player two chooses a distribution on $\mathcal{X}$. Player one chooses a state S that defines the channel; i.e., $p_S(Y|X=x)$ is the distribution on the transmitted data Y conditioned on the input being x. By the minimax theorem this game also has an equilibrium point.

$$\max_X \min_S I(X;Y) = \min_S \max_X I(X;Y).$$

Other variants of this scenario can be found in chapter 7 of El Gamal and Kim (2011).

2. Preliminaries

Here we provide the basic definitions that are needed for this text, and provide references that contain more details and background.

NOTATION

We identify random variables with the distributions they define. The notation $S \sim (X, h(X))^m$ means that S consists of m i.i.d. pairs of the form $(x_i, h(x_i))$ where x_i is distributed as X.

Big O and Ω notations in this text hide absolute constants.

LEARNING THEORY

Part I of Shalev-Shwartz and Ben-David (2014) provides an excellent comprehensive introduction to computational learning theory. Following are some basic definitions.

Let $\mathcal{X}$ and $\mathcal{Y}$ be sets. A set $\mathcal{C} \subseteq \mathcal{Y}^{\mathcal{X}}$ is called a class of hypotheses. $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$ is called the sample space. A realizable sample for $\mathcal{C}$ of size m is

$$S = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in \mathcal{S}^m$$

such that there exists $h \in \mathcal{C}$ satisfying $y_i = h(x_i)$ for all $i \in [m]$.

A learning algorithm A for $\mathcal{C}$ with sample size m is a (possibly randomized) algorithm that takes a realizable sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ for $\mathcal{C}$ as input, and returns a function $h : \mathcal{X} \to \mathcal{Y}$ as output. We say that the learning algorithm is consistent if the output h always satisfies $y_i = h(x_i)$ for all $i \in [m]$. We say the algorithm is proper if it outputs members of $\mathcal{C}$.

The *empirical error* of A with respect to S and a function $h \in \mathcal{C}$ is

$$\text{error}(A; S) = \frac{1}{m} \sum_{i=1}^{m} L_h(x_i, A(S)(x_i)),$$

where $L_h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the loss function. The *true error* of A with respect to a random variable X over $\mathcal{X}$ and a function $h \in \mathcal{C}$ is defined to be

$$\text{error}_h(A; X) = \mathbb{E}_{x \sim X} L_h(x, A(S)(x)).$$

The class $\mathcal{C}$ shatters some finite set $S \subseteq \mathcal{X}$ if the cardinality of $\mathcal{C}|_S = \{h|_S : h \in \mathcal{C}\}$ is $|\mathcal{Y}|^{|S|}$. The VC-dimension of $\mathcal{C}$ denoted $\mathsf{VC}(\mathcal{C})$ is the maximal size of a set $S \subseteq \mathcal{X}$ such that $\mathcal{C}$ shatters S.

INFORMATION THEORY

Let $\mathcal{X}$ be a finite set, and let X be a random variable over $\mathcal{X}$ with probability mass function p such that $p(x) = \Pr(X = x)$. The entropy of X is[1]

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}.$$

The mutual information between two random variables X and Y is

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

See the textbook Cover and Thomas (2006) for additional basic definitions and results from information theory which are used throughout this paper.

1. $\log(x)$ is a shorthand for $\log_2(x)$, and we use the convention that $0 \log \frac{1}{0} = 0$.

AVERAGE COMPLEXITY

Let A be a learning algorithm for $\mathcal{C}$ with sample size m, and let $\mathcal{P}$ be a probability distribution on $\mathcal{C}$. We say that A has at most *average information complexity* of d bits with respect to $\mathcal{P}$, if all random variables X over $\mathcal{X}$ satisfy

$$\mathbb{E}_{h\sim\mathcal{P}} I(S_h; A(S_h)) \leq d.$$

We say that A has error ϵ, confidence $1-\delta$, and at most *average sample complexity* M with respect to $\mathcal{P}$, if for all random variables X over $\mathcal{X}$ and all $m \geq M$,

$$\mathbb{E}_{h\sim\mathcal{P}} \Pr(\text{error}_h(A(S_h); X) > \epsilon) < \delta.$$

3. Information Games

It is helpful to think about the learning framework as a two-player game.

THE INFORMATION GAME

- The two players decide in advance on a class of functions $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ and a sample size m.
- Player one ("Learner") picks a consistent and proper learning algorithm A (possibly randomized).
- Player two ("Nature") picks a function $h \in \mathcal{C}$ and a random variable X over $\mathcal{X}$.
- Learner pays Nature $I(S; A(S))$ coins where $S \sim (X, h(X))^m$.

In the setting of Theorem 1, Nature knows in advance what the learning algorithm A of Learner is. In that case, Nature's optimal strategy leads to a gain of

$$\min_A \max_{(h,X)} I(S, A(S)) = \Omega(d \log\log(|\mathcal{X}|/d)).$$

In other words, when Nature knows what the learner is going to do, Nature's gain can be quite large even in very simple cases.

In Bassily et al. (2018), the other extreme was studied as well. Theorem 13 in Bassily et al. (2018) states that when Learner knows in advance the random variable X of Nature (but not the concept h), the gain of Nature is always much smaller; for all $h \in \mathcal{C}$,

$$\max_X \min_A I(S, A(S)) = O(d \log m).$$

In particular, in this case, Nature's gain does not tend to infinity with the size of the universe.

We see that this information game does not have, in general, a game theoretic equilibrium point. To remedy this, we suggest the following average case information game. We shall see the benefits of considering this game below.

THE AVERAGE INFORMATION GAME

- The two players decide in advance on $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ and m.
- Learner picks a consistent and proper learning algorithm A (possibly randomized).
- Nature picks a random variable X over $\mathcal{X}$.
- Learner pays Nature $\frac{1}{|\mathcal{C}|}\sum_{h\in\mathcal{C}} I(S_h; A(S_h))$ coins where $S_h \sim (X, h(X))^m$.

In the average game, Nature's gain is for an average concept h in the class. Nature can not choose a particular h that would lead to a high payoff. As opposed to the first game, the average information game has an equilibrium point (see the proof of Theorem 2 below):

$$\max_X \min_A \frac{1}{|\mathcal{C}|}\sum_{h\in\mathcal{C}} I(S_h; A(S_h)) = \min_A \max_X \frac{1}{|\mathcal{C}|}\sum_{h\in\mathcal{C}} I(S_h; A(S_h)).$$

By the results mentioned above, if the VC-dimension of $\mathcal{C}$ is d, then Nature's gain in the game is at most $O(d \log m)$, like in the case that Learner knows the underlying distribution. For VC classes, although $I(S; A(S))$ may be extremely large for *all* algorithms under *some* distribution on inputs, the average $\frac{1}{|\mathcal{C}|}\sum_{h\in\mathcal{C}} I(S_h; A(S_h))$ is small for *some* algorithms under *all* distributions on inputs.

An even more general statement holds. If one allows an empirical error of at most ϵ, instead of a consistent algorithm, the dependence on m can be omitted. This is indeed more general as if the empirical error is less than $1/m$ then the algorithm is consistent.

Theorem 2 *For every class $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ of VC-dimension d, every $m \geq 2$, and every $\epsilon > 0$, there is a proper learning algorithm A with empirical error bounded by ϵ such that for all random variables X on $\mathcal{X}$,*

$$\frac{1}{|\mathcal{C}|}\sum_{h\in\mathcal{C}} I(S_h; A(S_h)) = O(d \log(2/\epsilon))$$

where $S_h \sim (X, h(X))^m$.

The above result means that there is a learning algorithm such that for any distribution on inputs, the algorithm reveals little information about its input for at least half of the functions in $\mathcal{C}$, by Markov's inequality. If $d \log m$ is smaller than the entropy of the sample $H(X^m)$, then the algorithm can be thought of as compressing its input.

Remark 3 *The upper bound for the average information complexity that appears on Theorem 2 is tight (up to a multiplicative factor). The average information complexity for the class $\mathcal{C}_d = \{0,1\}^{\mathcal{X}}$ is $\Omega(d)$, where the sample size is $d/4$ and $|\mathcal{X}| = d$.*

Assume by contradiction that the average information complexity of $\mathcal{C}_d$ is $r = o(d)$. By Markov's inequality, atleast half of the functions of $\mathcal{C}_d$ satisfy $I(S_h; A(S_h)) < 2r$. Denote this set of functions $\widehat{\mathcal{C}}_d$. Since $|\widehat{\mathcal{C}}_d| > 2^{d-1}$, by the Sauer-Shelah lemma the VC dimension of $\widehat{\mathcal{C}}_d$ is atleast $d/2$. Let $\widehat{\mathcal{X}} \subset \mathcal{X}$ be the set shattered by $\widehat{\mathcal{C}}_d$. On one hand, the generalization error on $\widehat{\mathcal{C}}_d$ is small by the compression bound (Theorem 8 on (Bassily et al., 2018)). On the other hand, by the no free lunch theorem, the generalization error on $\widehat{\mathcal{C}}_d$ should be large. Hence the average information complexity of $\mathcal{C}_d$ is $\Omega(d)$.

Theorem 2 is a consequence of a more general phenomenon that holds even outside the scope of VC classes. To state it, we need to consider a convex space $\mathcal{D}$ of random variables (or distributions), since the mechanism that underlies its proof is von Neumann's minimax theorem (see Von Neumann (1928); Von Neumann and Morgenstern (1944)).

Lemma 4 *Let $\mathcal{C} \subset \mathcal{Y}^{\mathcal{X}}$ be a class of of hypotheses (not necessarily binary valued) with a loss function that is bounded from above by one. Let $\mathcal{D}$ be a convex set of random variables over the space $\mathcal{X}^m$. Assume that for every $X \in \mathcal{D}$, there exists an algorithm A_X whose output has empirical error $\leq \epsilon$ and $I(S_h; A_X(S_h)) \leq K$ for all $h \in \mathcal{C}$ where $S_h \sim (X, h(X))^m$. Then there exists a learning algorithm A such that for all $X \in \mathcal{D}$, the algorithm outputs a hypothesis with empirical error $\leq \epsilon$ and*

$$\frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} I(S_h; A(S_h)) \leq K.$$

The lemma is proved in Section 4.

Remark 5 *Some natural collections of random variables are not convex. For example, if one starts with a set of i.i.d. random variables over $\mathcal{X}^m$, the relevant convex hull does not consist only of i.i.d. random variables. This point needs to be addressed in the proof of Theorem 2. In the proof of Theorem 2, we apply the lemma with $\mathcal{D}$ being the space of all symmetric distributions on $\mathcal{X}^m$; see Definition 12.*

We call the learning algorithm A that is constructed in the proof of the theorem a *minimax algorithm* for $(\mathcal{C}, \mathcal{D})$ with information K and empirical error ϵ. Such algorithms reveal a small amount of information on most of the hypotheses in $\mathcal{C}$. So, together with the "compression yields generalization" results from Xu and Raginsky (2017) and Bassily et al. (2018) we get that the minimax algorithm has small true error for every $X \in \mathcal{D}$ for most hypotheses in $\mathcal{C}$, as long as $m \gg K$.

Corollary 6 *Let $\mathcal{D}_0$ be a convex set of random variables on $\mathcal{X}$. Let $\mathcal{D}$ be the convex hull of distributions of the form X^m for $X \in \mathcal{D}_0$. Let A be a* minimax algorithm *for $(\mathcal{C}, \mathcal{D})$ with information K and empirical error $\epsilon > 0$. Let $X \in \mathcal{D}_0$. If $m \geq \frac{K}{\epsilon^2 \delta}$, then*

$$\Pr[\mathit{error}_h(A(S); X) > 2\epsilon] < O(\delta) \qquad \textbf{(}\boldsymbol{h}\textbf{ is uniform)}$$

where $S \sim (X, h(X))^m$ and h is uniform in $\mathcal{C}$ and independent of X.

In particular, by Markov's inequality, for at least half of the functions h in $\mathcal{C}$,

$$\Pr[\mathit{error}_h(A(S_h); X) > 2\epsilon] < O(\delta) \qquad \textbf{(}\boldsymbol{h}\textbf{ is fixed)}$$

where $S_h \sim (X, h(X))^m$.

Remark 7 *There is nothing special about the uniform distribution on $\mathcal{C}$. Any other prior distribution $\mathcal{P}$ on $\mathcal{C}$ works just as well. It is important, however, to keep in mind that the algorithm A depends on the choice of the prior $\mathcal{P}$.*

Remark 8 *The convex set of distributions $\mathcal{D}_0$ may be chosen by the algorithm designer. One general choice is to take the space of all distribution on $\mathcal{X}$. Another example is the space of all sub-gaussian probability distributions.*

To complete the proof of Theorem 2, we apply Lemma 4. For the lemma to apply, we need to design an algorithm that reveals little information for VC classes when the distribution of X is known in advance (as mentioned in the remark following Lemma 4 we need to handle even a more general scenario). To do so, we need to extend a result from Bassily et al. (2018). The main ingredient is metric properties of VC classes (see Haussler (1995)). This appears in Section 5.

4. The Minimax Learner

Naturally, the proof requires von Neumann's minimax theorem.

Theorem 9 (Von Neumann (1928); Von Neumann and Morgenstern (1944)) *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^k$ be compact convex sets. Let $f : U \times V \to \mathbb{R}$ be a continuous function that is convex-concave, i.e.,*

- *$f(\cdot, v) : U \to \mathbb{R}$ is convex for every $v \in V$ and*
- *$f(u, \cdot) : V \to \mathbb{R}$ is concave for every $u \in U$.*

Then

$$\min_{u \in U} \max_{v \in V} f(u, v) = \max_{v \in V} \min_{u \in U} f(u, v).$$

Proof [Proof of Lemma 4] We need to verify that the minimax theorem applies. First, as stated in the preliminaries, we deal with a finite space $\mathcal{X}$ so the set of all algorithms (randomized included) with empirical error $\leq \epsilon$ and the set of random variables $\mathcal{D}$ over $\mathcal{X}^m$ can be treated as convex compact sets in high dimensional euclidean space. Specifically, let U be the collection of randomized learning algorithms with empirical error at most ϵ, and let V be the set $\mathcal{D}$ of distributions.

Second, mutual information is a continuous function of both strategies.

Third, the following lemma about mutual information.

Lemma 10 (Theorem 2.7.4 in Cover and Thomas (2006)) *Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $\mathrm{I}(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

We apply the lemma with $p(x)$ being the distribution on S and $p(y|x)$ being the distribution of h conditioned on $S = s$ that the learning algorithm defines. Since a convex combination of convex/concave functions is convex/concave, we see that the map

$$(u, v) \mapsto \frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} I(S_h; A_X(S_h))$$

is convex-concave, where u defines the distribution of $A(S_h)$ conditioned on the value of S_h, and v defines the distribution of S_h.

By assumption,

$$\max_{X \in V} \min_{A_X \in U} \frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} I(S_h; A_X(S_h)) \leq K.$$

By the minimax theorem,

$$\min_{A \in U} \max_{X \in V} \frac{1}{|\mathcal{C}|} \sum_{h \in \mathcal{C}} I(S_h; A(S_h)) \leq K.$$

In other words, there is a randomized algorithm A as needed (points in U are randomized algorithms). ■

Remark 11 *In the proof above, we used the special fact that the mutual information is convex-concave. We are not aware of any other measure of dependence between random variables that satisfy this.*

5. Learning Using Nets

Theorem 13 from Bassily et al. (2018) states the following. For an i.i.d. random variable X over $\mathcal{X}^m$ and $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ with VC-dimension d, there exists a consistent, proper, and deterministic learner that leaks at most $O(d \log(m+1))$-bits of information, where m is the input sample size (for $\mathcal{C}$-realizable samples).

For the minimax theorem to apply, we need to generalize the above statement to work for any convex combination of i.i.d. random variables over $\mathcal{X}^m$. To analyze this collection of random variables, we need to identify some property that we can leverage. We use the fact that such random variables are invariant under permutation of the coordinates.

Definition 12 *A random variable X over $\mathcal{X}^m$ is called symmetric if for any permutation $\sigma : [m] \to [m]$,*

$$\Pr\left(X = (x_1, ..., x_m)\right) = \Pr\left(X = \left(x_{\sigma(1)}, ..., x_{\sigma(m)}\right)\right).$$

The following theorem holds for all symmetric random variables. In this space, we can not assume any kind of independence between the coordinates. This should make the proof more complicated than in Bassily et al. (2018), but in fact it helps to guide the proof and make it quite simple.

Theorem 13 *Let $\epsilon > 0$. For a symmetric random variable X over $\mathcal{X}^m$ and $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$ with VC-dimension d, there exists a proper and deterministic learner A with empirical error $\leq \epsilon$ so that*

$$I(S; A(S)) \leq O(d \log(2/\epsilon))$$

for all $m \geq 2$.

A key component in the proof is Haussler's theorem (see Haussler (1995)) on the size of covers of VC classes. The theorem states that for a given probability distribution μ on $\mathcal{X}$, there are small covers to the metric space whose elements are concepts in $\mathcal{C}$ and the distance between $c_1, c_2 \in \mathcal{C}$ is $\mu(\{x : c_1(x) \neq c_2(x)\})$. The starting point of this theorem is a distribution on $\mathcal{X}$. In the general setting we consider, we start with a non-product distribution on $\mathcal{X}^m$. To apply Haussler's theorem, we need to find the relevant μ (the solution is eventually quite simple).

Proof Since X is symmetric, the marginal distribution is the same on each of the coordinates of $\mathcal{X}^m$ and denote it D. For every integer $j > 0$, pick a minimal ϵ_j-net N_j with respect to the distribution D over $\mathcal{X}$ for $\epsilon_j = \epsilon/2^j$.

The learning algorithm is simple – it outputs the first function that has empirical error of at most ϵ it sees along the sequence of nets. The algorithm stops because $\mathcal{C}$ is finite. It remains to calculate the entropy of its output.

For every $j > 0$ and $h \in \mathcal{C}$, there is a function $f_{j,h}$ in N_j so that

$$D(\{x : h(x) \neq f_{j,h}(x)\}) \leq \epsilon_j.$$

By the linearity of the expectation,

$$\mathop{\mathbb{E}}_{(x_1,\ldots,x_m)} \frac{\sum_{i=1}^{m} 1_{f_{j,h}(x_i) \neq h(x_i)}}{m} \leq \epsilon_j. \qquad \text{(expected empirical error)}$$

So, by Markov's inequality,

$$\Pr(f_{j,h} \text{ has empirical error } > \epsilon) < \frac{\epsilon_j}{\epsilon} = 2^{-j}.$$

In total, for all $j > 0$,

$$\Pr(\exists f \in N_j \text{ with empirical error } \leq \epsilon) \geq \Pr(f_{j,h} \text{ has empirical error } \leq \epsilon) \geq 1 - 2^{-j}.$$

Now take J to be the index of the net where the algorithm stops. For $j \geq 2$ it holds that $P(J = j) \leq 2^{-(j-1)}$. Thus,

$$H(J) \leq O(1),$$

.

By Haussler's theorem (see Haussler (1995)), the size of N_j is at most

$$(4e^2/\epsilon_j)^d = (4e^2 2^j/\epsilon)^d.$$

Therefore,

$$H(A(S)|J) \leq \sum_{j=1}^{\infty} P(J = j) \log |N_j| = O(d \log(2/\epsilon)).$$

Finally,

$$I(S; A(S)) = H(A(S)) \leq H(A(S), J) = H(J) + H(A(S)|J).$$

The first equality follows from A being a deterministic algorithm and the second equality follows from the chain rule.

■

More Generally

The proof of Theorem 13 together with Lemma 4 suggest a general recipe for controlling the average information complexity (and hence the average sample complexity) for pairs of the form $(\mathcal{C}, \mathcal{D})$ (not necessarily binary class or with 0-1 loss).

- For every marginal distribution D over $\mathcal{X}$ from $\mathcal{D}$, find a sequence of small ϵ-nets. This sequence induces an algorithm that leaks little information, for every symmetric random variable $X \in \mathcal{D}$ whose marginal distribution is D (even though it is not necessarily i.i.d.).
- Use the minimax theorem to find an algorithm that leaks little information over all of $\mathcal{D}$.

It will be interesting to see if this setting can be extended to the non-realizable case. It is not immediate to apply the principles seen in the proof of Theorem 2 to this case. In theory, some samples may require large empirical losses (for proper learners). Since the minimax algorithm is a convex combination of those algorithms, it is hard to say what the empirical error of such an algorithm will be, or how far will the empirical error be from the hypothesis in $\mathcal{C}$ with an optimal empirical error.

6. Stability

To describe the minimax algorithm we need to come up with some prior distribution $\mathcal{P}$ on $\mathcal{C}$. In practice, we do not necessarily know the actual prior but we may have some approximation of it. It is natural to ask how does the performance of the minimax algorithm change when our prior $\mathcal{P}$ is wrong, and the true prior is $\mathcal{Q}$.

As an example, if we have a bound $\sup_h \mathcal{Q}(h)/\mathcal{P}(h) \leq C$, then we immediately get

$$\mathbb{E}_{h\sim\mathcal{Q}} I(S_h; A(S_h)) \leq C \cdot \mathbb{E}_{h\sim\mathcal{P}} I(S_h; A(S_h)).$$

As another example, consider the case that the statistical distance $\|\mathcal{P} - \mathcal{Q}\|_1$ is small. If we assume nothing on how $I(S_h; A(S_h))$ distributes, we can get

$$\frac{\mathbb{E}_{h\sim\mathcal{Q}} I(S_h; A(S_h))}{\mathbb{E}_{h\sim\mathcal{P}} I(S_h; A(S_h))} = \Theta(\|\mathcal{P} - \mathcal{Q}\|_1 \log|\mathcal{C}|), \tag{1}$$

which seems too costly to be useful. This can happen when one hypothesis satisfies $I(S_h; A(S_h)) = \Theta(\log|\mathcal{C}|)$, and we move all the allowed weight from one hypothesis with small mutual information to h. If, however, the second moment is bounded, we can get better estimates:

$$\begin{aligned}
&|\mathbb{E}_{h\sim\mathcal{P}} I(S_h; A(S_h)) - \mathbb{E}_{h\sim\mathcal{Q}} I(S_h; A(S_h))| \\
&\qquad \leq \sum_{h\in\mathcal{C}} I(S_h; A(S_h))|\mathcal{P}(h) - \mathcal{Q}(h)| \\
&\qquad = \sum_{h\in\mathcal{C}} \left(I(S_h; A(S_h))\sqrt{\mathcal{P}(h) + \mathcal{Q}(h)} \right) \left(\frac{|\mathcal{P}(h) - \mathcal{Q}(h)|}{\sqrt{\mathcal{P}(h) + \mathcal{Q}(h)}} \right) \\
&\qquad \leq \sqrt{\mathbb{E}_{h\sim\mathcal{P}}[(I(S_h; A(S_h)))^2] + \mathbb{E}_{h\sim\mathcal{Q}}[(I(S_h; A(S_h)))^2]} \cdot \sqrt{\|\mathcal{P} - \mathcal{Q}\|_1}.
\end{aligned}$$

The last inequality is Cauchy-Schwartz. Roughly speaking, this means that if $\mathcal{P}$ is close to $\mathcal{Q}$ then the average information that is leaked is similar, when the map $h \mapsto I(S_h; A(S_h))$ has bounded second moment under both distributions. It is possible to replace the second moment by the p-moment for $p > 1$ using Hölder's inequality.

Remark 14 *We saw that with no assumptions, information cost can grow considerably under small perturbations of $\mathcal{P}$ (see equation 1). The average sample complexity, however, does not. If A has error ϵ, confidence $1 - \delta$, and average sample complexity M with respect to $\mathcal{P}$, it also has error ϵ, confidence $1 - \delta - \|\mathcal{P} - \mathcal{Q}\|_1$, and average sample complexity M with respect to $\mathcal{Q}$.*

7. Discussion

This work leaves the traditional setting of PAC learning and assumes a less hostile environment for learning. We introduce game-theoretic perspectives of the compression learning algorithms perform. In the standard setting, Nature is assumed all powerful and can make the Learner leak quite a lot of information. In the average-case scenario, Nature needs to commit ahead of time on some probability distribution from which the eventual concept is generated. In this case, the minimax theorem allows to lower the amount of information that is leaked.

The average-case framework captures some amount of prior knowledge on the world that the learner can use. It therefore allows to avoid singular or pathological cases.

This work suggests an idea that may be useful in other contexts. Given a class $\mathcal{C} \subset \mathcal{Y}^{\mathcal{X}}$, perform the following four steps.

1. Define a set of reasonable distributions $\mathcal{D}$ over $\mathcal{X}$.
2. Find a collection of ϵ-nets for distributions in $\mathcal{D}$.
3. Look for a distribution over those nets that works well for most distributions in $\mathcal{D}$.
4. Given a sample S, sample a random ϵ-net until finding an hypothesis with small empirical error.

It seems plausible that this will yield acceptable results for samples that come from the real world. All steps above, however, may be quite challenging to implement.

References

Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. *CoRR*, abs/1806.03803, 2018.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464–473. IEEE, 2014.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 2018.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2ed edition, 2006.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506.02629, 2015a. URL http://arxiv.org/abs/1506.02629.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015b.

Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, 2011.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *COLT*, 2018.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217 – 232, 1995.

David Haussler, Michael Kearns, and Robert E. Schapire. Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. *Machine Learning*, 14(1): 83–113, 1994.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, Technical report, University of California, Santa Cruz, 1986.

David A. McAllester. A pac-bayesian tutorial with A dropout bound. *CoRR*, abs/1307.2118, 2013. URL http://arxiv.org/abs/1307.2118.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM (JACM)*, 63(3):21, 2016.

Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. A direct sum result for the information complexity of learning. In *Proceedings of the 2018 Conference on Learning Theory*, 2018.

Rüdiger Reischuk and Thomas Zeugmann. A complete and tight average-case analysis of learning monomials. In *STACS 99*, pages 414–423, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 487–494. IEEE, 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

J Von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

J Von Neumann and O Morgenstern. Theory of games and economic behavior. 1944.

Andrew Wan. Learning, cryptography, and the average case. *Institution Columbia University*, 2010.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30*, pages 2524–2533. 2017.