World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:17, No:3, 2023

A Family of Distributions on Learnable Problems without Uniform Convergence

César Garza

Abstract—In supervised binary classification and regression problems, it is well-known that learnability is equivalent to uniform convergence of the hypothesis class, and if a problem is learnable, it is learnable by empirical risk minimization. For the general learning setting of unsupervised learning tasks, there are non-trivial learning problems where uniform convergence does not hold. We present here the task of learning centers of mass with an extra feature that "activates" some of the coordinates over the unit ball in a Hilbert space. We show that the learning problem is learnable under a stable RLM rule. We introduce a family of distributions over the domain space with some mild restrictions for which the sample complexity of uniform convergence for these problems must grow logarithmically with the dimension of the Hilbert space. If we take this dimension to infinity, we obtain a learnable problem for which the uniform convergence property fails for a vast family of distributions.

Keywords—Statistical learning theory, learnability, uniform convergence, stability, regularized loss minimization

I. Introduction

INTUITIVELY, a problem is *learnable* if there exists a training algorithm producing a learning rule such that, with *high probability* on a randomly selected training set, the generalization error is small. This is defined rigorously in Definition 1. For supervised tasks such as binary classification, regression, or multiclass prediction, different conditions have been shown to be equivalent to learnability. Learning with the empirical risk minimization rule (ERM) in the supervised case is equivalent to uniform convergence of the empirical risk to the true risk with a rate that is independent of the distribution over the instance set. For binary classification, Vapnik and Chervonenkis [1] showed that finiteness of a combinatorial condition known as the VC-dimension is a necessary and sufficient condition for learnability under the ERM rule. For some regression problems, finite fat-shattering dimension characterizes learnability [2] and the Natarajan dimension characterizes learnability of some multiclass learning problems [3].

For the general learning setting, there is no equivalence between learnability and uniform convergence, as Shalev-Shwartz et al. showed in [4]. Instead, the key notion is stability, as defined in Section III. Examples of learnable problems without uniform convergence can shed more light into how Vapnik's notion of "strict" learnability fails in the framework of unsupervised learning.

In this paper we present two unsupervised tasks for the center of mass over the unit ball in some Hilbert spaces. The second problem is a modification of the first that makes it a strictly convex, bounded, smooth problem. We show that these tasks are learnable by exhibiting stability using smoothness of the corresponding loss functions, with a coefficient that is independent of the dimension d of the Hilbert space. We also show that these tasks possess distributions $\mathcal{D}$ concentrated in a small ball around the origin for which the uniform convergence property does not hold in the infinite dimensional case.

In Section II, we present the formal definitions of learning under the ERM rule in the supervised case and discuss the equivalent notions of uniform convergence and finiteness of VC-dimension in the binary classification case. In Section III we introduce the generalized concept of learning as defined in [4]. After defining the equivalent concept of stability, we present a theorem from [4] that shows that for convex-smooth-bounded problems, the Regularized Loss Minimization rule (RLM) with Tikhonov regularization leads to a stable learning algorithm. Finally, in Section IV, we introduce our learning problem where the uniform convergence property fails. This can be described as the task of finding the "center of mass" of a distribution over the unit ball of $\mathbb{R}^d$, where an extra parameter $\boldsymbol{\alpha}$ indicates which of the coordinates are marked as "active" or "inactive". We show that this problem is learnable using the RLM rule with a sample complexity that does not depend on d. Then we choose a probability distribution for the instance space such that if $m < \log_2(d)$, there is a high probability that a sample of i.i.d. labeled points of size m has a high estimation error. We say that such samples are not "ϵ-representative". The main theorems of this paper are Theorems 5 and 6 where we show that for distributions $\mathcal{D}$ on the domain $\mathcal{Z}$ concentrated in a ball of radius $1/4$ and yielding a uniform Bernoulli distribution on the parameter $\boldsymbol{\alpha}$, a.s. the ERM rule does not converge to a minimizer of the true population risk as the sample size $m \to \infty$, not even for strictly convex bounded smooth problems where the ERM minimizer is unique. We hope that the distributions presented here are only the starting point for a rich variety of stable problems in unsupervised settings that lack the uniform convergence property.

II. The Supervised Learning Setting

In the supervised learning setting, we have an instant space $\mathcal{X}$, a label set $\mathcal{Y}$, and a hypothesis class $\mathcal{H}$. The domain $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ has a sigma-algebra structure and we have a "loss" function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ that is measurable for all $h \in \mathcal{H}$. We also assume the loss function is bounded over $\mathcal{H} \times \mathcal{Z}$.

Given a probability distribution $\mathcal{D}$ over $\mathcal{Z}$, the *risk* or *true error* of a hypothesis $h \in \mathcal{H}$ denoted as $L_{\mathcal{D}}(h)$ is defined as the expected value of the loss function over $\mathcal{Z}$; that is,

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h,(\mathbf{x},y))]$$

C. Garza is with the Department of Mathematics & Statistics, University of Houston Downtown, Houston, TX, 77002 USA (e-mail: garzace@uhd.edu).

While $\mathcal{H}, \mathcal{Z}$ and the loss function ℓ are known to the learner, we assume that $\mathcal{D}$ is unknown. It is thus not possible to simply choose $h \in \mathcal{H}$ that minimizes $L_{\mathcal{D}}(h)$. Instead, we consider *training samples* $S \sim \mathcal{D}^m$ of m i.i.d. draws from $\mathcal{Z}$. Each sample S is a sequence of the form $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, where $\mathbf{x}_i \in \mathcal{X}$ and y_i is the corresponding label. For any $m \in \mathbb{N}$, we will use the notation $[m]$ to denote $\{1, \ldots, m\}$. Our overall goal in this setting is to have a learning algorithm A that picks a hypothesis $A(S) \in \mathcal{H}$ based on the training sample S with approximately minimal possible risk. Generally, we expect the approximation to get better with the sample size. Before we give the formal definition of learnability, we present some examples of supervised statistical learning tasks:

- **Binary Classification:** Let $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{H}$ be the set of functions $h : \mathcal{X} \to \{0, 1\}$. The loss function is the indicator function $\ell(h, (\mathbf{x}, y)) = \mathbb{1}_{h(\mathbf{x}) \neq y}$. This is also known as the $0 - 1$ loss function, which measures if h labeled the example $(\mathbf{x}, y)$ properly or not.

- **Linear Regression:** Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^n$ and let $\mathcal{Y}$ be a bounded subset of $\mathbb{R}$. Let $\mathcal{H}$ be a set of bounded functions $h : \mathcal{X} \to \mathbb{R}$, and let ℓ be the square loss function: $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$.

- **Ranking:** We can consider ranking problems for classification or information retrieval purposes. The training data are a list of items and we assign a partial order to the items in the list. If $\mathcal{X}$ is the set of instances, let $\mathcal{X}^* = \bigcup_{n=1}^{\infty} \mathcal{X}^n$ be the set of all sequences of instances from $\mathcal{X}$ of arbitrary length. Here $\mathcal{Z} = \bigcup_{r=1}^{\infty} (\mathcal{X}^r \times \mathbb{R}^r)$. The hypothesis class $\mathcal{H}$ is the set of ranking hypotheses h that receive a sequence of instances $\overline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_r) \in \mathcal{X}^*$ and return a vector $\mathbf{y} \in \mathbb{R}^r$. By sorting the elements of $\mathbf{y}$ in increasing order, we obtain a permutation of $[r]$. There are many possible ways to define a loss function for ranking. If we denote by $\pi(\mathbf{y})$ the permutation of $[r]$ induced by the vector $\mathbf{y} \in \mathbb{R}^r$, then one example is the $0 - 1$ loss function $\ell(h, (\overline{\mathbf{x}}, \mathbf{y})) = \mathbb{1}_{[\pi(h(\overline{\mathbf{x}})) \neq \pi(\mathbf{y})]}$. Better examples of loss functions for ranking are the Kendall-Tau loss or the Normalized Discounted Cumulative Gain loss; see [5] for more details.

Ideally, we wish to pick in this setting a hypothesis $h \in \mathcal{H}$ that minimizes the true risk $L_{\mathcal{D}}(h)$, but since $\mathcal{D}$ is unknown to the learner, this is not feasible. We wish to obtain a learning rule A such that, upon receiving a training sample S of size m, A outputs a hypothesis $A(S)$ and the expected value of the difference between the true risk of $A(S)$ and the minimal risk is small, with this value approaching 0 as the sample size $m \to \infty$. That is,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)] \leq \epsilon(m)$$

where $\mathcal{D}^m$ is the probability over m-tuples in $\mathcal{Z}$ induced by applying $\mathcal{D}$ to pick each element of the tuple independently of the other members of the tuple. We also require the rate $\epsilon(m)$ to be monotonically decreasing with $\epsilon(m) \xrightarrow{m \to \infty} 0$.

Since $\mathcal{D}$ is unknown, we ask for learnability that the above inequality is consistent over *all* distributions $\mathcal{D}$ on $\mathcal{Z}$. This leads us to the formal definition of learnability of supervised tasks.

Definition 1. A learning problem is *learnable* if there exist a learning rule A and a monotonically decreasing sequence $\epsilon_{\text{const}}(m)$, such that $\epsilon_{\text{const}}(m) \xrightarrow{m \to \infty} 0$ and for all distributions $\mathcal{D}$ on $\mathcal{Z}$,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)] \leq \epsilon_{\text{const}}(m). \tag{1}$$

A learning rule A for which this holds is denoted as a *universally consistent* learning rule.

This is a direct generalization of agnostic PAC-learnability as seen in [2]. Note that instead of asking for an inequality similar to (1) that holds with probability $1 - \delta$ over all samples S, we ask for a uniform rate over the expected value of the difference of errors for all distributions on $\mathcal{Z}$.

A. Equivalent Forms of Learnability

The learner does not have access to the distribution $\mathcal{D}$ of the domain. Nevertheless, the learner can compute an *empirical error* or *empirical risk* based on the training sample S. This is denoted by $L_S(h)$ and it is defined as the error a hypothesis h incurs over the training sample. If $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, then

$$L_S(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h, (\mathbf{x}_i, y_i))$$

We say that a rule A is an ERM (*Empirical Risk Minimizer*) if it minimizes the empirical risk

$$A(S) \in \operatorname*{arg\,min}_{h \in \mathcal{H}} L_S(h).$$

Here argmin denotes the collection of hypotheses in $\mathcal{H}$ for which the value of $L_S(h)$ over $\mathcal{H}$ is minimal.

We say that a problem is learnable under the ERM rule if the ERM rule described above satisfies (1) for all distributions $\mathcal{D}$ over $\mathcal{Z}$.

A simple idea that is related to learnability is to have a hypothesis class $\mathcal{H}$ for which the empirical risk of any hypothesis $h \in \mathcal{H}$ is a good approximation of its true risk. This is formalized in the definition of *uniform convergence* of a learning problem.

Definition 2. A learning problem with domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and hypothesis class $\mathcal{H}$ is said to have the *uniform convergence property* if

$$\sup_{\mathcal{D}} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \right] \xrightarrow{m \to \infty} 0$$

More intuitively, given any $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that for any distribution $\mathcal{D}$ on $\mathcal{Z}$ and any hypothesis $h \in \mathcal{H}$, the mean value of $|L_{\mathcal{D}}(h) - L_S(h)|$ is less than ϵ.

The uniform convergence property says that the empirical risks of hypotheses in the hypothesis class converge to their population risk uniformly, with a distribution-independent rate. We offer a third combinatorial concept that is used in binary classification problems only. Let $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$, where each hypothesis $h \in \mathcal{H}$ is a mapping $h : \mathcal{X} \to \{0,1\}$ and ℓ is the $0-1$ loss function $\ell(h,(\mathbf{x},y)) = \mathbb{1}_{h(\mathbf{x})\neq y}$.

Definition 3. Let C be a finite subset of $\mathcal{X}$. We say that a hypothesis class $\mathcal{H}$ *shatters* C if any function from C to $\{0,1\}$ can be obtained as a restriction of an element $h \in \mathcal{H}$ to C.

Vapnik and Chervonenkis defined in [1] a simple combinatorial measure that implies uniform convergence.

Definition 4. Let $\mathcal{H}$ be a hypothesis class. The VC-dimension of $\mathcal{H}$, denoted VCdim$(\mathcal{H})$, is the maximal cardinal D such that a set of cardinality D in $\mathcal{X}$ is shattered by $\mathcal{H}$.

For binary classification problems we have a chain of equivalences explained in the next theorem.

Theorem 1 (The Fundamental Theorem of Statistical Learning, see [5] Theorem 6.7)**.** *Let $\mathcal{X}$ be the set of instances and let $\mathcal{H}$ be a hypothesis class of binary functions on $\mathcal{X}$. Then, under the $0-1$ loss function, the following are equivalent:*

1) *$\mathcal{H}$ has a finite VC-dimension.*
2) *$\mathcal{H}$ has the uniform convergence property.*
3) *Any ERM rule is a successful learner for $\mathcal{H}$.*
4) *$\mathcal{H}$ is learnable according to Definition 1.*

The situation is depicted in Fig. 1.

In the case of regression problems, a similar characterization holds. This time a hypothesis h is a real-valued function $h : \mathcal{X} \to \mathbb{R}$ and the loss function is the squared-loss function $\ell(h,(\mathbf{x},y)) = (h(\mathbf{x}) - y)^2$. The VC dimension is replaced by the fat-shattering dimension, but the basic equivalence still holds: a problem is learnable if and only if uniform convergence holds if and only if the uniform convergence property is present (see [6]).

Remark 1. In our definitions of learnability, uniform convergence, and stability in the next section, we have used convergence in expectation, and defined the rates as rates on the expectation. Since the loss function ℓ is bounded, by the dominated convergence theorem, convergence in expectation is equivalent to convergence in probability. Furthermore, using Markov's inequality we can translate a rate of the form $\mathbb{E}[|X|] \leq \epsilon(m)$ to a "low confidence" guarantee $\mathbb{P}[|X| > \epsilon(m)/\delta] \leq \delta$. Thus "learnability" can be replaced with agnostic PAC learnability as defined in [5] in Theorem 1. For simplicity, we will not discuss in this paper the computational aspects of learnability, although for the tasks presented here there are well-known efficient algorithms such as SGD that solve the problem.

III. General Learning Framework

We now consider the general learning setting, where the domain $\mathcal{Z}$ is an arbitrary measurable space. There is still a hypothesis class $\mathcal{H}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ that is measurable on $\mathcal{Z}$ and bounded by some constant B. That is, $\ell(h,\mathbf{z}) \leq B$ for all $h \in \mathcal{H}$ and $\mathbf{z} \in \mathcal{Z}$.

Some examples of general learning tasks that do not fit in the supervised setting are:

- **K-means clustering:** Let $\mathcal{Z} = \mathbb{R}^n$, let $\mathcal{H}$ be all subsets of $\mathbb{R}^n$ with k elements, and let $\ell(h,\mathbf{z}) = \min_{\mathbf{c}\in h} \|\mathbf{c}-\mathbf{z}\|^2$. Here, each h represents a set of k centroids, and ℓ measures the square of the Euclidean distance between an instance $\mathbf{z}$ and its nearest centroid, according to the hypothesis h.
- **Stochastic Convex Optimization in Hilbert Spaces:** Let $\mathcal{Z}$ be any measurable set, let $\mathcal{H}$ be a closed, convex and bounded subset of a Hilbert space, and let $\ell(h,\mathbf{z})$ be Lipschitz and and convex with respect to its first argument. The task is to minimize the true risk function $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})]$, where the distribution $\mathcal{D}$ over Z is unknown, based on a training sample $S = (\mathbf{z}_1,\ldots,\mathbf{z}_m)$.

The definitions of learnability and uniform convergence in the general case are exactly as in Definitions 1 and 2 respectively, with the only difference being a more general domain space $\mathcal{Z}$.

In the next section we will prove that in the general framework learnability is no longer equivalent to uniform convergence. We will define an equivalent notion of learnability that is no longer concerned about the complexity of the hypothesis class. Instead, we wish to control the variance of the learning rule. Intuitively, an algorithm is considered stable if a slight change of its input does not change its output much. To be more precise, given the training set S and an additional example $\mathbf{z}'$ from $\mathcal{Z}$, let $S^{(i)}$ be the training set obtained by replacing the i'th example of S with $\mathbf{z}'$. That is,

$$S^{(i)} = (\mathbf{z}_1,\ldots,\mathbf{z}_{i-1},\mathbf{z}',\mathbf{z}_{i+1},\ldots,\mathbf{z}_m)$$

By "a small change of the input" we mean that we feed the learner A the sample $S^{(i)}$ instead of S. We observe that only one training sample is replaced. We then compare the loss of the hypothesis $A(S)$ on the element $\mathbf{z}_i$ to the loss of $A(S^{(i)})$ on the same element $\mathbf{z}_i$. We say that A is a *stable* algorithm if changing a single example in the training set does not lead to a significant change. Formally,

Definition 5. Let $\epsilon_{\text{st}}(m)$ be a monotonically decreasing function with $\epsilon_{\text{st}}(m) \xrightarrow{m\to\infty} 0$ and let $U(m)$ be the uniform distribution over $[m]$. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon_{\text{st}}(m)$ if for every distribution $\mathcal{D}$ over $\mathcal{Z}$

$$\mathop{\mathbb{E}}_{(S,\mathbf{z}')\sim\mathcal{D}^{m+1},i\sim U(m)}\left[\ell(A(S^{(i)}),\mathbf{z}_i) - \ell(A(S),\mathbf{z}_i)\right] \leq \epsilon_{\text{st}}(m)$$

For simplicity, we will call a learning algorithm that is on-average-replace-one-stable just *universally stable* or simply *stable*.

For supervised learning tasks like binary classification or regression, by the Fundamental Theorem of Statistical Learning, if a problem is learnable then it is learnable under any ERM rule. This is no longer true in the general setting. In this case, the correct approach is to choose a rule that is "asymptotically" ERM or AERM for short. The precise definition is as follows.

Definition 6. A rule A is *universally* an AERM rule with rate $\epsilon_{\text{erm}}(m) \xrightarrow{m\to\infty} 0$ if

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_S(A(S)) - \min_{h\in\mathcal{H}} L_S(h)] \leq \epsilon_{\text{erm}}(m)$$

for all distributions $\mathcal{D}$ over $\mathcal{Z}$.

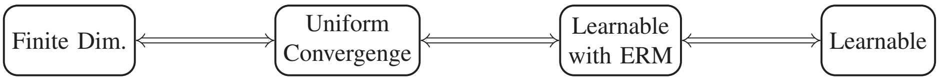

Fig. 1 The Fundamental Theorem of Statistical Learning

In [4], Shalev-Shwarz et al. proved that general learnability is equivalent to having a stable universal AERM rule.

Theorem 2 ([4], Theorem 7)**.** *A learning problem is learnable if and only if there exists a stable universally AERM learning rule.*

The theorem even relates the three distinct convergence rates $\epsilon_{\text{const}}(m), \epsilon_{\text{erm}}(m)$, and $\epsilon_{\text{st}}(m)$, although we shall not be concerned about this. It should also be mentioned that, contrary to binary classification, not any AERM rule is enough for learnability; the AERM rule must also be stable. Fig. 2 illustrates the correspondences in the general learning framework.

Note that the uniform convergence property no longer appears as an equivalent condition for learnability. It can be shown (see [4]) that uniform convergence is a sufficient condition for a stable universally AERM rule and hence learnability, but it is by no means a necessary condition.

In the next section, we consider two types of learning problems that are learnable by a stable universally AERM rule, but do not possess the uniform convergence property.

IV. A Learnable Problem without Uniform Convergence

Let $\mathcal{B}$ be the unit ball in $\mathbb{R}^d$, let $\mathcal{H} = \mathcal{B}$, and let $\mathcal{Z} = \mathcal{B} \times [0,1]^d$. We define a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ as:

$$\ell(\mathbf{h},(\mathbf{x},\boldsymbol{\alpha})) = \sum_{i=1}^{d} \alpha_i (x_i - h_i)^2 = \left\|\sqrt{\boldsymbol{\alpha}} * (\mathbf{x}-\mathbf{h})\right\|^2 \qquad (2)$$

where $\sqrt{\boldsymbol{\alpha}}$ is the element-wise square root and $\mathbf{u} * \mathbf{v}$ denotes an element-wise product. This is an unsupervised learning task where we try to find the "center of mass" of the distribution over $\mathcal{B}$ and the vector $\boldsymbol{\alpha}$ represents a vector of stochastic per-coordinate "confidence" weights α_i for each coordinate in $\mathbb{R}^d$.

We will prove that this problem is learnable using smoothness properties of the loss function. First we define formally the concept of smoothness that we will use in this paper.

Definition 7. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz. That is, for all $\mathbf{v}, \mathbf{w}$ in $\mathbb{R}^d$ we have $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|$.

Now we can show that our loss function is β-smooth in its first argument for a constant β that does not depend on the dimension d.

Lemma 1. *The loss function $\ell(\cdot,(\mathbf{x},\boldsymbol{\alpha}))$ in* (2) *is* 2*-smooth for all $d \in \mathbb{N}$.*

Proof 1. *Fix $(\mathbf{x},\boldsymbol{\alpha})$ in $\mathcal{Z}$. Then for $\mathbf{v}, \mathbf{w}$ in $\mathcal{H}$,*

$$\begin{aligned}\|\nabla\ell(\mathbf{v}) - \nabla\ell(\mathbf{w})\| &= 2\|\langle \alpha_1(v_1 - w_1), \ldots, \alpha_d(v_d - w_d)\rangle\| \\ &\leq 2\|\mathbf{v} - \mathbf{w}\|\end{aligned}$$

where the last inequality follows since each α_i satisfies $0 \leq \alpha_i \leq 1$. □

We have thus a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ where the following holds:

- $\mathcal{H}$ is a convex bounded subset of $\mathbb{R}^d$.
- For all $\mathbf{z} \in \mathcal{H}$, the loss function $\ell(\cdot, \mathbf{z})$ is a convex, nonnegative, 2-smooth function such that $\ell(\mathbf{0}, \mathbf{z}) = \sum_{i=1}^d \alpha_i x_i^2 \leq \|\mathbf{x}\|^2 \leq 1$.

This is known as a Convex-Smooth-Bounded Learning problem (see [5, Definition 12.13]). Instead of working directly with the loss function ℓ, we use a "regularized" version of it; namely, we use an ERM rule for the regularized loss function

$$\ell(\mathbf{h}, \mathbf{z}) + \frac{\lambda}{2}\|\mathbf{h}\|^2$$

for some parameter $\lambda > 0$ to be chosen later. This is also known as the "Regularized Loss Minimization" (RLM) rule. The extra function $\lambda\|\mathbf{h}\|^2/2$ is known as Tikhonov regularization. If A is an RLM learner for some parameter $\lambda > 0$, then upon receiving a sample $S \sim \mathcal{D}^m$, the algorithm returns a hypothesis

$$A(S) \in \underset{\mathbf{h} \in \mathcal{H}}{\arg\min} \left(\ell(\mathbf{h}, \mathbf{z}) + \frac{\lambda}{2}\|\mathbf{h}\|^2 \right)$$

Theorem 3 says that the RLM rule is a successful learner for Convex-Smooth-Bounded Learning problems with a suitable boundedness condition on the loss function.

Theorem 3 ([5], Corollary 13.11)**.** *Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be a convex-smooth-bounded learning problem with parameters β, B, where $\|\mathbf{h}\| \leq B$ for all $\mathbf{h} \in \mathcal{H}$. We assume in addition that $\ell(\mathbf{0}, \mathbf{z}) \leq 1$ for all $\mathbf{z} \in \mathcal{Z}$. For any $\epsilon \in (0,1)$, let $m \geq \dfrac{150\beta B^2}{\epsilon^2}$ and set $\lambda = \epsilon/(3B^2)$. Let A be an RLM learner with parameter λ. Then, for every distribution $\mathcal{D}$ of $\mathcal{Z}$,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - \min_{\mathbf{h} \in \mathcal{H}} L_{\mathcal{H}}(\mathbf{h})\right] \leq \epsilon$$

For our center of mass problem, $\beta = 2$ and $B = 1$. Thus the RLM rule is stable and the problem is learnable under Definition 1 for any $d \in \mathbb{N}$.

Now we take $\mathcal{H}$ to be the unit sphere $\mathcal{B}$ of an infinite-dimensional Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$, where for $\mathbf{v} \in \mathcal{H}$, we refer to its coordinates $\mathbf{v}_j = \langle \mathbf{v}, \mathbf{e}_j \rangle$. The weights $\boldsymbol{\alpha}$ are now a mapping of each coordinate to $[0,1]$. That is, $\boldsymbol{\alpha}$ is an infinite sequence of reals in $[0,1]$. The loss function in (2) is defined with respect to this orthonormal basis and is still well-defined in this Hilbert space. Since $\beta = 2$ was independent of the dimension d, the infinite-dimensional problem $(\mathcal{H}, \mathcal{Z}, \ell)$ is still a convex-smooth-bounded learning problem and we thus obtain

Theorem 4. *Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be the infinite-dimensional problem where $\mathcal{H} = \mathcal{B}$, $\mathcal{Z}$ is formed of pairs $(\mathbf{x}, \boldsymbol{\alpha})$ where $\mathbf{x} \in \mathcal{B}$ and $\boldsymbol{\alpha}$ is a sequence of numbers in $[0,1]$, and $\ell(\mathbf{h}, (\mathbf{x}, \boldsymbol{\alpha}))$ is as*

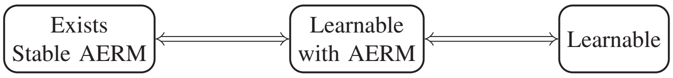

Fig. 2 The General Learning Framework

in (2)*. Then* $(\mathcal{H}, \mathcal{Z}, \ell)$ *is a convex-smooth bounded problem learnable under a stable RLM rule.*

Next we present a family of distributions on $\mathcal{Z}$ for which $\sup_{\mathbf{h}\in\mathcal{H}} |L_{\mathcal{D}}(\mathbf{h}) - L_S(\mathbf{h})|$ does not converge in mean to 0 as $m \to \infty$, showing that the uniform convergence property fails for this problem. This is an extension of the work in [4], where only one such distribution was shown. The main goal of this paper is to show how easy it is to find distributions where the true risk of a hypothesis is considerably bigger than the empirical risk, even when the true risk converges in mean to the minimal risk achievable by elements in the hypothesis class.

We start with the finite-dimensional case. Let d be a positive integer and consider the learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ defined above where $\mathcal{H}$ is the unit ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $\mathcal{D}$ be a distribution only over $\mathcal{B} \times \{0,1\}^d$ satisfying the following two conditions:

1) $\underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{P}}(\|\mathbf{x}\| > 1/4) = 0$
2) $\forall i \in [d] : \underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{P}}(\boldsymbol{\alpha}_i = 1) = \dfrac{1}{2}$

The two conditions state that the distribution is 0 away from the ball of radius $1/4$ in $\mathbb{R}^d$ and that the marginal distribution on $\{0,1\}^d$ is a sum of independent, uniform Bernoulli random variables. Here are two examples of such distributions:

Let C be a finite subset of $\mathcal{B}$ such that $\|\mathbf{x}\| \le 1/4$ for all $\mathbf{x} \in C$. Let $\mathcal{D}_1$ be the uniform distribution on $C \times \{0,1\}^d$. More generally, let $C = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ be a denumerable collection in $\mathbb{R}^d$ such that $\|\mathbf{x}_i\| \le 1/4$ for all i. Assign to each $(\mathbf{x}_i, \boldsymbol{\alpha})$ a probability of 2^{-i-d}. This yields a distribution $\mathcal{D}_2$ satisfying the two conditions above.

As a generalization of the previous example, let μ be any probability measure on $\mathcal{B}_{1/4}$, the ball of radius $1/4$ in $\mathbb{R}^d$ centered at $\mathbf{0}$ and let U be the uniform distribution on $\{0,1\}^d$. Then for $\mathcal{D}_3 = \mu \times U$ extended to 0 over $\mathcal{B} \times \{0,1\}^d$ the two conditions hold.

We will show that the rate of uniform convergence for the problem $(\mathcal{H}, \mathcal{Z}, \ell)$ grows with d. First we define a notion of "representative" samples with respect to a distribution $\mathcal{D}$.

Definition 8. Let $\epsilon > 0$. A training set S is called ϵ-representative (with respect to domain $\mathcal{Z}$, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall \mathbf{h} \in \mathcal{H}, \qquad |L_{\mathcal{D}}(\mathbf{h}) - L_S(\mathbf{h})| \le \epsilon$$

Lemma 2. *Let* $\mathcal{D}$ *be a distribution over* $\mathcal{B} \times \{0,1\}^d$ *satisfying* (1) *and* (2)*. We assume* $2^m < d$*. Then with probability of at least* $1 - e^{-1}$*, a sample* S *of size* m *is not* $\frac{1}{5}$*-representative w.r.t.* $(\mathcal{H}, \mathcal{Z}, \ell, \mathcal{D})$*.*

Proof 2. *Let* $S = ((\mathbf{x}^{(1)}, \boldsymbol{\alpha}^{(1)}), \ldots, (\mathbf{x}^{(m)}, \boldsymbol{\alpha}^{(m)}))$ *be a sample of* m *i.i.d. draws from* $\mathcal{Z}$ *with distribution* $\mathcal{D}$*. We will show that with probability at least* $1 - e^{-1} > 0.63$*, there exists a coordinate* $j \in [d]$ *such that* $\boldsymbol{\alpha}_j^{(i)} = 0$ *for all* $i \in [m]$*. Indeed, the probability that this occurs is given by*

$$\mathbb{P}\left(\bigcup_{j\in[d]} \bigcap_{i\in[m]} \{\boldsymbol{\alpha}_j^{(i)} = 0\}\right) = 1 - \mathbb{P}\left(\bigcap_{j\in[d]} \bigcup_{i\in[m]} \{\boldsymbol{\alpha}_j^{(i)} = 1\}\right)$$

By our choice of $\mathcal{D}$*, the* $\boldsymbol{\alpha}_j^{(i)}$ *are independent uniform Bernoulli random variables. Hence*

$$\begin{aligned} 1 - \mathbb{P}&\left(\bigcap_{j\in[d]} \bigcup_{i\in[m]} \{\boldsymbol{\alpha}_j^{(i)} = 1\}\right) \\ &= 1 - \prod_{j\in[d]} \left(1 - \mathbb{P}\left(\boldsymbol{\alpha}_j^{(1)} + \ldots + \boldsymbol{\alpha}_j^{(m)} = 0\right)\right) \\ &= 1 - (1 - 2^{-m})^d \\ &\ge 1 - (e^{-2^{-m}})^d \\ &= 1 - e^{-d2^{-m}} \\ &\ge 1 - e^{-1} \end{aligned} \tag{3}$$

Now we show that a sample S *for which* $\boldsymbol{\alpha}_j^{(i)} = 0$ *for some coordinate* $j \in [d]$ *and all* $i \in [m]$ *cannot be* $\frac{1}{5}$ *representative with respect to this distribution* $\mathcal{D}$*. Let* $\mathbf{e}_j$ *be the standard unit vector along coordinate* j *in* $\mathbb{R}^d$*. Then* $\mathbf{e}_j \in \mathcal{H}$ *and*

$$L_S(\mathbf{e}_j) = \frac{1}{m} \sum_{i\in[m]} \ell(\mathbf{e}_j, (\mathbf{x}^{(i)}, \boldsymbol{\alpha}^{(i)}))$$

since $\boldsymbol{\alpha}_j^{(i)} = 0$ *for all* i *and* $\mathbf{e}_j$ *has only one nonzero coordinate,*

$$\begin{aligned} &= \frac{1}{m} \sum_{i\in[m]} \sum_{k\in[d]\setminus\{j\}} \boldsymbol{\alpha}_k^{(i)} (\mathbf{x}_k^{(i)})^2 \\ &\le \frac{1}{m} \sum_{i\in[m]} \left\|\mathbf{x}^{(i)}\right\|^2 \\ &\le \frac{1}{16} \end{aligned}$$

On the other hand, by the law of total expectation,

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{e}_j) &= \underset{(\mathbf{x},\boldsymbol{\alpha})\sim\mathcal{D}}{\mathbb{E}}[\ell(\mathbf{e}_j, (\mathbf{x}, \boldsymbol{\alpha}))] \\ &= \underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[\ell(\mathbf{e}_j, (\mathbf{x}, \boldsymbol{\alpha}))|\boldsymbol{\alpha}_j = 1]\, \mathbb{P}(\boldsymbol{\alpha}_j = 1) \\ &+ \underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[\ell(\mathbf{e}_j, (\mathbf{x}, \boldsymbol{\alpha}))|\boldsymbol{\alpha}_j = 0]\, \mathbb{P}(\boldsymbol{\alpha}_j = 0) \\ &\ge \frac{1}{2} \underset{(\mathbf{x},\boldsymbol{\alpha})}{\mathbb{E}}[(\mathbf{x}_j - 1)^2 + \ldots] \\ &\ge \frac{1}{2}\left(\frac{3}{4}\right)^2 = \frac{9}{32} \end{aligned}$$

We conclude that with probability of at least $1 - e^{-1}$ *we obtain a sample* S *of size* m *for which* $|L_{\mathcal{D}}(\mathbf{e}_j) - L_S(\mathbf{e}_j)| \ge \frac{9}{32} - \frac{1}{16} > \frac{1}{5}$ *for some* $j \in [d]$*. Such a sample is not* $\frac{1}{5}$*-representative.* □

Open Science Index, Cognitive and Language Sciences Vol:17, No:3, 2023 publications.waset.org/10012987.pdf

Lemma 2 shows that the sample complexity m for uniform convergence in the $(\mathcal{H}, \mathcal{Z}, \ell)$ finite-dimensional problem is $\Omega(\log(d))$.

In the infinite dimensional case where $\mathcal{H} = \mathcal{B}$ is the unit sphere in a Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \dots$ and ℓ has the coordinate-free form in (2), we consider distributions $\mathcal{D}$ of the following form:

1) $\mathcal{D}$ is nonzero only on $\mathcal{B}_{1/4} \times [0,1]$, where we have identified sequences $\boldsymbol{\alpha}$ in $\{0,1\}$ with real numbers in the interval $[0,1]$.
2) The marginal distribution on $[0,1]$ is the uniform distribution. If we regard $\boldsymbol{\alpha} \in [0,1]$ as a sequence where each $\boldsymbol{\alpha}_j$ is 0 or 1, this means that all $\boldsymbol{\alpha}_j$ are independent uniform Bernoulli random variables.

Theorem 5. *Let $(\mathcal{H}, \mathcal{Z}, \ell)$ be the learnable problem as in Theorem 4. Then the problem does not have the uniform convergence property.*

Proof 5. *Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$ with the two properties defined above. The estimates in* (3) *carry out the same in the infinite dimensional case. If we take the limit as $d \to \infty$, we obtain that a.s. if we take a training sample S of size m, there is a coordinate j such that $\boldsymbol{\alpha}_j^{(i)} = 0$ for all $i \in [m]$. By the same computations as in Lemma 2, we obtain that for such distributions $\mathcal{D}$ and for all m,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\sup_{\mathbf{h} \in \mathcal{H}} |L_{\mathcal{D}}(\mathbf{h}) - L_S(\mathbf{h})| \right] \geq \frac{1}{5}$$

Therefore $(\mathcal{H}, \mathcal{Z}, \ell)$ does not have the uniform convergence property. □

For the learning problem presented here, we were able to show that there exists some hypothesis $\mathbf{h} \in \mathcal{H}$ for which the true risk does not converge to the empirical risk as $m \to \infty$. We can sharpen this example by exhibiting a problem for which the empirical risk minimizer $\mathbf{h}^*$ also exhibits this problem.

We consider the problem $(\mathcal{H}, \mathcal{Z}, \ell')$ where $\mathcal{H}$ and $\mathcal{Z}$ are as in Theorem 4, but the loss function is now

$$\ell'(\mathbf{h}, (\mathbf{x}, \boldsymbol{\alpha})) = \left\| \sqrt{\boldsymbol{\alpha}} * (\mathbf{x} - \mathbf{h}) \right\|^2 + \eta \sum_{j=1}^{\infty} b_j (\mathbf{h}_j - 1)^2 \quad (4)$$

where $\eta = 0.01$ and $\{b_j : j \in \mathbb{N}\}$ is any set of positive numbers such that $\sum b_j = 1$.

The new loss function ℓ' is $(2 + 2\eta)$-smooth and since the additional term is strictly convex, ℓ' is strictly convex. Therefore, for any training sample S of any size,

$$\mathbf{h}^* \in \arg\min_{\mathbf{h} \in \mathcal{H}} L'_S(\mathbf{h})$$

is unique, where $L'_S(\mathbf{h})$ is the empirical risk of the hypothesis $\mathbf{h}$ with respect to the new loss function ℓ'. We assume that $\mathcal{D}$ is a product measure on $\mathcal{B}_{1/4} \times [0,1]$ satisfying conditions (1) and (2). In particular, $\mathbf{x}$ and $\boldsymbol{\alpha}$ are independent random variables.

Lemma 3. *Let S be a i.i.d. sample of size m of $\mathcal{Z}$ according to $\mathcal{D}^m$. If $\mathbf{h}^* \in \mathcal{H}$ is the unique empirical risk minimizer of ℓ', then a.s. $\|\mathbf{h}^*\| = 1$.*

Proof 3. *First consider the unconstrained optimization problem of finding*

$$\mathbf{h}^*_{UC} \in \arg\min_{\mathbf{h}} L'_S(\mathbf{h})$$

For any training sample S of size m, a.s. there exists a coordinate j such that $\boldsymbol{\alpha}_j^{(i)} = 0$ for all $i \in [m]$. Thus only the second term in (4) *depends on $\mathbf{h}_j$. Since $\mathbf{h}^*_{UC}$ is the unique minimizer of L'_S, we obtain $\mathbf{h}^*_{UC,j} = 1$. Consequently, $\|\mathbf{h}^*_{UC}\| \geq 1$. It follows that in the constrained case where $\mathbf{h} \in \mathcal{H}$, we must have $\|\mathbf{h}^*\| = 1$.* □

We will denote by $L'_{\mathcal{D}}(\mathbf{h})$ the true risk of hypothesis $\mathbf{h}$ under the loss function (4).

Theorem 6. *Let S be a i.i.d. sample of size m of $\mathcal{Z}$ according to $\mathcal{D}^m$. Let $\mathbf{h}^* \in \mathcal{H}$ be the unique empirical risk minimizer of ℓ', and let $L^* = \min_{\mathbf{h} \in \mathcal{H}} L'_{\mathcal{D}}(\mathbf{h})$. Then a.s.*

$$|L'_{\mathcal{D}}(\mathbf{h}^*) - L^*| \geq \frac{1}{5}$$

Proof 6. *By Lemma 3, $\|\mathbf{h}^*\| = 1$. We write*

$$\begin{aligned} L'_{\mathcal{D}}(\mathbf{h}^*) &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})} [\ell'(\mathbf{h}^*, (\mathbf{x}, \boldsymbol{\alpha}))] \\ &\geq \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})} \left[\sum_{k=1}^{\infty} \boldsymbol{\alpha}_k (\mathbf{x}_k - \mathbf{h}^*_k)^2 \right] \end{aligned}$$

since $\mathbf{x}$ and $\boldsymbol{\alpha}$ are independent with our choice of $\mathcal{D}$:

$$\begin{aligned} &= \sum_{k=1}^{\infty} \mathop{\mathbb{E}}_{\boldsymbol{\alpha}} [\boldsymbol{\alpha}_k] \mathop{\mathbb{E}}_{\mathbf{x}} (\mathbf{x}_k - \mathbf{h}^*_k)^2 \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x}} [\|\mathbf{x} - \mathbf{h}^*\|^2] \\ &\geq \frac{9}{32} \end{aligned}$$

On the other hand,

$$\begin{aligned} L^* &= \min_{\mathbf{h} \in \mathcal{H}} L'_{\mathcal{D}}(\mathbf{h}) \\ &\leq L'_{\mathcal{D}}(\mathbf{0}) \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})} [\ell'(\mathbf{0}, (\mathbf{x}, \boldsymbol{\alpha}))] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})} [\left\| \sqrt{\boldsymbol{\alpha}} * \mathbf{x} \right\|^2] + \eta \\ &\leq \mathop{\mathbb{E}}_{(\mathbf{x}, \boldsymbol{\alpha})} [\|\mathbf{x}\|^2] + \eta \\ &\leq \frac{1}{16} + \eta \end{aligned}$$

Since $\eta = .01$, $|L'_{\mathcal{D}}(\mathbf{h}^*) - L^*| \geq \frac{9}{32} - \frac{1}{16} - .01 > \frac{1}{5}$. □

Theorem 6 says that for any product measure $\mathcal{D} = \mu \times U_{[0,1]}$ on $\mathcal{B}_{1/4} \times [0,1]$ where $U_{[0,1]}$ is the uniform distribution on $[0,1]$, then a.s. the unique empirical risk minimizer $\mathbf{h}^*$ of $(\mathcal{H}, \mathcal{Z}, \ell')$ performs much worse than the population optimum L^* and therefore does not converge to it as $m \to \infty$. Thus this problem is not learnable under the ERM rule, although we already showed that is learnable under a RLM rule.

V. Conclusion

Theorems 5 and 6 are generalizations of Example 4.1 in [4]. In that paper, the authors only presented a single distribution $\mathcal{D}$ concentrated on $\mathbf{0} \in \mathcal{B}$ where the uniform convergence property fails. We have presented a very rich family of distributions over $\mathcal{B} \times [0, 1]$ where the gap between the empirical risk and the true risk is bounded away from 0 a.s. for any sample size m. The only restrictions on $\mathcal{D}$ that we have imposed are a concentration of $\mathcal{D}$ in a smaller ball of radius $1/4$, and a corresponding distribution of independent, uniform Bernouilli random variables on the $\boldsymbol{\alpha}$ variable. We can even relax some conditions on the problem we have studied here. For example, we can ask that the hypothesis class $\mathcal{H}$ is a bounded convex set only. This family of distributions for the weighted center of mass problem also shows that "pathogenic" distributions on $\mathcal{Z}$ where the uniform convergence property fails are much more common than originally thought.

References

[1] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*, pp. 11–30, Springer, Cham, 2015. Reprint of Theor. Probability Appl. **1**6 (1971), 264–280.

[2] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, pp. 115–141, 1994.

[3] B. K. Natarajan, "On learning sets and functions," *Machine Learning*, vol. 4, pp. 67–97, 2004.

[4] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *J. Mach. Learn. Res.*, vol. 11, pp. 2635–2670, 2010.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

[6] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *J. ACM*, vol. 44, no. 4, pp. 615–631, 1997.

Dr. César Garza is an assistant professor of Mathematics & Statistics at the University of Houston-Downtown. His research focus is on the area of convex learning problems. He has supervised student research projects in PAC learnability. He has also published papers on the construction of hyperkähler metrics through Riemann-Hilbert problems and he is the author of a monograph in the area of low-dimensional topology about hyperbolic knots with toroidal surgeries.