Investigations Governing With Uniform Convergence On The Basis Of Stability And Learnability

Krishna Kumar Sen[1]*, Dr. Ram Narayan Yadava[2], Dr. Satish Agnihotri[3]
***Corresponding Author: -** Krishna Kumar Sen
Address: [1, 2, 3] Madhyanchal Professional University (M.P)

Abstract

The key issue pertaining with characterization of learnability is the most basic question of statistical learning theory. A fundamental and long-standing answer, at least for the case of supervised classification and regression, is that learnability is equivalent to uniform convergence of the empirical risk to the population risk, and that if a problem is learnable, it is learnable via empirical risk minimization. In this paper, we consider the General Learning Setting (introduced by Vapnik), which includes most statistical learning problems as special cases. We show that in this setting, there are non-trivial learning problems where uniform convergence does not hold, empirical risk minimization fails, and yet they are learnable using alternative mechanisms. Instead of uniform convergence, we identify stability as the key necessary and sufficient condition for learnability. More- over, we show that the conditions for learnability in the general setting are significantly more complex than in supervised classification and regression.

Keywords: learnability, statistical learning theory, stability, uniform convergence, stochastic convex optimization.

DOI Number: 10.14704/nq.2022.20.9.NQ44580 **Neuro Quantology 2022; 20(9):5028-5040**

1. Introduction

We consider the General Setting of Learning introduced by Vapnik (1995) where we would like to minimize a population risk functional (stochastic objective)

$$F(\mathbf{h}) = \mathrm{E}_{Z\sim D}\,[\mathrm{f}(\mathbf{h};Z)] \quad (1)$$

Over some hypothesis class H, where the distribution D of Z is unknown, based on i.i.d. sample $z_1, \ldots, z_m$ drawn from D (and full knowledge of f and H). This General Setting subsumes supervised classification and regression, certain unsupervised learning problems, density estimation and more. For example, in super- vised learning $z = (\mathbf{x},y)$ is an instance-label pair, $\mathbf{h}$ is a predictor, and $f(h;(\mathbf{x},y)) = \text{loss}\,(h(\mathbf{x}),y)$ is the loss functional.

In the context of this general setting, we are concerned with the question of statistical "learnability". That is, when can Equation (1) be minimized to within arbitrary precision based only on a finite sample $z_1,\ldots,z_m$, as $m \to \infty$? We are not concerned here with computational aspects of this problem, that is, whether this approximate minimization can be carried out efficiently, but only whether it is statistically possible to do so based only on the sample $z_1, \ldots, z_m$.

For supervised classification and regression problems, it is well known that a problem is learnable if and only if the empirical risks

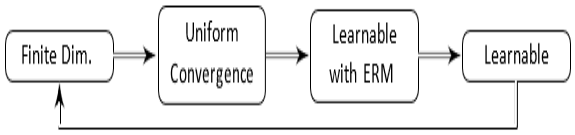
$$F_S(\mathbf{h}) = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{h}, z_i)$$

for all $\mathbf{h}$ H converge uniformly to the population risk (Blumer et al., 1989; Alone tal., 1997). If uniform convergence holds, then the empirical risk minimizer (ERM) is *consistent*, that is, the population risk of the ERM converges to the optimal population risk, and the problem is learnable using the ERM. We therefore have:

- A necessary and sufficient condition for learnability, namely uniform convergence of the empirical risks. Furthermore, this can be shown to be equivalent to a combinatorial condition: having finite VC-dimension in the case of classification, and having finite fat-shattering dimensions in the case of regression.
- A complete understanding of *how* to learn:

since learnability is equivalent to learnability by ERM, we can focus our attention solely on empirical risk minimizers.

The situation, for supervised classification and regression, can be depicted as follows:

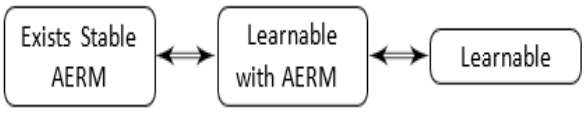


Other than uniform convergence, certain notions of stability have also been suggested as an explicit condition for learnability. Intuitively, stability notions focus on particular algorithms, or learning rules, and measure their sensitivity to perturbations in the training set. In particular, it is known that stability of the ERM is *sufficient* for learnability. In Mukherjee et al. (2006), it is argued that stability is also a *necessary* for learnability. However, that argument relied on the assumption that uniform convergence is equivalent to learnability. Therefore, stability was shown to characterize learnability only in situations where uniform convergence characterizes learnability anyway.

The equivalence of uniform convergence and learnability was formally established only in the supervised classification and regression setting. In the more general setting, the "rightward" implications in the diagram above still hold: finite fat-shattering dimensions, uniform convergence, as well as ERM stability, are indeed sufficient conditions for learnability using the ERM. As for the reverse implication, Vapnik showed that a notion of "non-trivial" or "strict" learnability with the ERM is indeed equivalent to uniform convergence of the empirical risks. This notion was meant to exclude certain "trivial" learning problems, which are learn- able without uniform convergence. Even in such problems, learnability is still possible by empirical risk minimization. Thus, it would seem that in the General Learning Setting, as in supervised classification and regression, a problem is learnable if and only if it is learnable by empirical risk minimization.

In this paper we show that the situation in the General Learning Setting is actually much more complex. In particular, in which we show an example of a learning problem in the General Learning Setting, which is learnable (using an online algorithm and an online-to-batch conversion), but which is *not* learnable using empirical risk minimization. To the best of our knowledge this is the first example shown of this type. Furthermore, iwe show a modified example which *is* learnable using empirical risk mini- mization, but for which the empirical risks of the hypotheses do *not* converge uniformly to their expectations, not even locally for hypotheses very close to the true hypothesis. We argue that unlike the examples discussed. This example is far from being "trivial". We use this example to discuss how Vapnik's notion of "strict" learnability with the ERM is too strict, and precludes cases which are far from trivial and in which learnability with empirical risk minimization is *not* equivalent to uniform convergence.

Having shown that learnability does not imply learnability with the ERM, and learnability with the ERM does not imply uniform convergence (unlike supervised classification and regression), we proceed in Section 5 to characterize learnability in the General Learning Setting, unveiling stability as a key notion.
In particular, we show that for learnable problems, even when they are not learnable with ERM, they are always learnable with some learning rule which is "asymptotically ERM". Moreover, such an AERM must be stable (under a suitable notion of stability). Namely, we have the following characterization of learnability in the General Learning Setting:



Note that this characterization holds even for learnable problems with no uniform convergence. In this sense, stability emerges as a strictly more powerful notion than uniform convergence for characterizing learnability. Other than this, we also discuss several related results, which above all imply that the conditions for learnability in the General Learning Setting are substantially different and more complex than in supervised classification and regression.

Our results point not to a specific learning rule

(such as an ERM), but rather to a class of learning rules (AERM learning rules) as possible candidates for learning. Also we explore how our results can be strengthened if we allow randomized learning rules. In particular, randomization allows us to pinpoint not a general class of learning rules, but rather a specific (though highly impractical) learning rule, which learns if and only if the problem is learnable.

Throughout most of the paper we discuss learning rates (as a function of the sample size), but do not pay much attention to the confidence at which the learning rule succeeds (i.e., the dependence of the sample size on the allowed probability of failure). This issue is addressed, and again we show that in the General Learning Setting, things can behave rather differently than in supervised classification and regression.

In summary, this paper opens a door to the complexity of learnability in the General Learning Setting, and provides some understanding of the situation, including highlighting the important role of stability. Many gaps in our understanding remain, and we hope that future progress will close some of these gaps, as well as connect the theoretical insights gained to machine learning as used in practice.

This paper is partially based on the results obtained in Shalev-Shwartz et al. (2009a) and Shalev-Shwartz et al. (2009b).

The General Learning Setting: Formal Definition and Notation

In this paper we focus on the General Learning Setting, which was introduced by Vapnik (1995) as a unifying framework for the problem of statistical learning from empirical data.

The General Learning Setting deals with *learning problems*. Formally, a learning problem is specified by a hypothesis class H, an instance set Z (with a sigma-algebra), and an objective function (e.g., "loss" or "Cost") f: $H \times Z \to$ R. Throughout this paper we assume the function is bounded by some constant B, that is $|f(\mathbf{h}; \mathbf{z})| \le B$ for all $\mathbf{h} \in H$ and $z \in Z$.

Given a distribution D on Z, the quality of each hypothesis $\mathbf{h} \in H$ is measured by its *risk* $F(\mathbf{h})$, which is defined as $E_{\mathbf{z}\sim D}[f(\mathbf{h}; \mathbf{z})]$. While H, Z and $f(\mathbf{h}; \mathbf{z})$ are known to the learner, we assume that D is unknown. Ideally, we would like to pick $\mathbf{h}$ H whose risk is as close as possible to $\inf_{\mathbf{h}\ H} F(\mathbf{h})$. Since the underlying distribution D is unknown, we cannot do this directly, but instead need to rely on a finite empirical *training sample* $S = \{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$. On this sample, we apply a *learning rule* to pick a hypothesis. Formally, a learning rule is a mapping $\mathbf{A}$: $\cup^{=} Z^m \to H$ from sequences of instances in Z to hypotheses. We refer to sequences S = z1,…,$\mathbf{z}_m$ as "sample sets", but it is important to remember that the order and multiplicity of instances may be significant. A learning rule that does not depend on the order of the instances in the training sample is said to be *symmetric*. We will generally consider samples $S\ D^m$ of m i.i.d. draws from D.

This framework is sufficiently general to include a large portion of the statistical learning and optimization problems we are aware of, such as:

- **Binary Classification:** Let $Z = X \times \{0, 1\}$, let H be a set of functions $\mathbf{h}$: $X \to \{0, 1\}$, and let $f(\mathbf{h}; (\mathbf{x}, y)) = 1_{\{\mathbf{h}(\mathbf{x}) \neq y\}}$. Here, $f(\cdot)$ is simply the 0 – 1 loss function, measuring whether the binary hypothesis $\mathbf{h}(\cdot)$ misclassified the example $(\mathbf{x}, y)$.

- **Regression:** Let $Z = X \times Y$ where X and Y are bounded subsets of R and R respectively, let H be a set of bounded functions $\mathbf{h}$: X^n R, and let $f(\mathbf{h}; (\mathbf{x}, y)) = (\mathbf{h}(\mathbf{x}) - y)^2$. Here, $f(\cdot)$ is simply the squared loss function.

Large Margin Classification in a Reproducing Kernel Hilbert Space (RKHS): Let $Z = X \times \{0, 1\}$, where X is a bounded subset of an RKHS, let H be another bounded subset of the RKHS, and let $f(\mathbf{h}; (\mathbf{x}, y)) = \max(0, 1 - y\langle \mathbf{x}, \mathbf{h}\rangle)$. Here, $f(\cdot)$ is the well-known hinge loss function, and our goal is to perform margin-based linear classification in the RKHS.

K-Means Clustering in Euclidean Space: Let Z = R, let H be all of size k, and let subsets of R $f(\mathbf{h}; \mathbf{z}) = \min_{\mathbf{c}\in\mathbf{h}} \|\mathbf{c} - \mathbf{z}\|$. Here, each $\mathbf{h}$ represents a set of k centroids, and $f(\cdot)$ measures the Euclidean distance squared between an instance $\mathbf{z}$ and its nearest centroid,

according to the hypothesis **h**.

- **Density Estimation:** Let Z be a subset of R, let H be a set of bounded probability densities on Z, and let $f(\mathbf{h};\mathbf{z}) = -\log(\mathbf{h}(\mathbf{z}))$. Here, $f(\cdot)$ is simply the negative log-likelihood of an instance **z** according to the hypothesis density **h**. Note that to ensure boundedness of $f(\cdot)$, we need to assume that $\mathbf{h}(\mathbf{z})$ is lower bounded by a positive constant for all $\mathbf{z} \in Z$.

- **Stochastic Convex Optimization in Hilbert Spaces:** Let Z *be* an arbitrary measurable set, let H be a closed, convex and bounded subset of a Hilbert space, and let $f(\mathbf{h}; \mathbf{z})$ be Lipschitz-continuous and convex w.r.t. its first argument. Here, we want to approximately minimize the objective function $E_{\mathbf{z}\sim D}[f(\mathbf{h};\mathbf{z})]$, where the distribution over Z is unknown, based on an empirical sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$.

Our overall goal in this setting is to pick a hypothesis **h** H with approximately minimal possible risk, based on a finite sample. Generally, we expect the approximation to get better with the sample size. Learning rules which allow us to choose such hypotheses are said to be *consistent*. Formally, we say a rule A is consistent with rate $\varepsilon_{cons}(m)$ under distribution D if for all m,

$$E_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S)) - F^*] \leq \varepsilon_{cons}(m), \qquad (2)$$

Where we denote $F^* = \inf_{\mathbf{h}\in H} F(\mathbf{h})$ (here and whenever talking about a "rate" $\varepsilon(m)$, we require it to be monotone decreasing with $\varepsilon_{cons}(m) \xrightarrow{m\to\infty} 0$).

D is unknown, we cannot choose a learning rule based on D stronger requirement, namely that the rule is consistent with rate $\varepsilon_{cons}(m)$ under *all* distributions D over Z. This leads to the following central definition:

Definition 1 A learning problem is learnable, if there exist a learning rule A and a monotonically decreasing sequence $\varepsilon cons(m)$, *such that* $\varepsilon cons(m) \xrightarrow{m\to} \varepsilon 0$, *and*

$$\forall \mathcal{D},\ E_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S)) - F^*] \leq \varepsilon_{cons}(m).$$

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

This definition of learnability, requiring a uniform rate for all distributions, is the relevant notion for studying learnability of a hypothesis class. It is a direct generalization of agnostic PAC-learnability (Kearns et al., 1992) to Vapnik"s General Setting of Learning as studied by Haussler (1992) and others.

A possible approach to learning is to minimize the *empirical risk* $F_S(\mathbf{h})$ over a sample S, defined as

$$F_S(\mathbf{h}) = \frac{1}{m}\sum_{\mathbf{z}\in S} f(\mathbf{h};\mathbf{z}).$$

Z,z	Instance domain and a specific instance.	
$\mathcal{H}$,h	Hypothesis class and a specific hypothesis.	
$f(h, z)$	Loss of hypothesis h on instance z	
B	$\sup_{h,z} \lvert f(h;z)\rvert$	
$\mathcal{D}$	Underlying distribution on instance domain Z	
S	Empirical sample $z_1, \ldots, z_m$ sampled i.i.d. from $\mathcal{D}$	
m	Size of empirical sample S	
$A(S)$	Learning rule A applied on empirical sample S	
$\varepsilon_{cons}(m)$	Rate of consistency for a learning rule	
$F(h)$	Risk of hypothesis h, $E_{z\sim\mathcal{D}}[f(h;z)]$	
F^*	$\inf_{h\in\mathcal{H}} F(h)$	
$F_S(h)$	Empirical risk of hypothesis h on sample S, $\frac{1}{m}\sum_{z\in S} f(h;z)$	
$\hat{h}_S$	An ERM hypothesis, $F_S(\hat{h}_S) = \inf_{h\in\mathcal{H}} F_S(h)$	
$\varepsilon_{erm}(m)$	Rate of AERM for a learning rule	
$\varepsilon stable(m)$	Rate of stability for a learning rule	
$\varepsilon_{gen}(m)$	Rate of generalization for a learning rule	

Table 1: Table of Notation

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

$$F_S(\mathbf{A}(S)) = F_S(\hat{\mathbf{h}}_S) = \inf_{\mathbf{h}\in H} F_S(\mathbf{h}).$$

where we use $F_S(\hat{\mathbf{h}}_S) = \inf_{\mathbf{h}\ H} F_S(\mathbf{h})$ to refer risk. But since there might be several hypotheses minimizing the empirical risk, $\hat{\mathbf{h}}_S$ *does* not refer to to the minimal empirical a specific hypotheses and there might be many
rules which are all ERM.

We say that a rule **A** is an *AERM (Asymptotic Empirical Risk Minimizer)* with rate $\varepsilon_{erm}(m)$

under distri- bution D if:

$$E_{S\sim D^m}\Big[F_S(\mathbf{A}(S)) - F_S(\mathbf{h}_S)\Big] \le \varepsilon_{erm}(m)$$

A learning rule is *universally an AERM* with rate $\varepsilon_{erm}(m)$, if it is an AERM with rate $\varepsilon_{erm}(m)$ under all distributions D over Z. A learning rule is an *always AERM* with rate $\varepsilon_{erm}(m)$, if for *any* sample S of size m, it holds that $F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S) \le \varepsilon_{erm}(m)$.

We say a rule $\mathbf{A}$ *generalizes* with rate $\varepsilon_{gen}(m)$ under distribution D if for all m,

$$E_{S\sim D^m}\left[\left|F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right|\right] \le \varepsilon_{gen}(m).$$

A rule *universally generalizes* with rate $\varepsilon_{gen}(m)$ if it generalizes with rate $\varepsilon_{gen}(m)$ under all distributions D over Z.

We note that other authors sometimes define "consistent", and thus also "learnable" as a combination of our notions of "consistent" and "generalizing".

In the above definitions, we choose to use convergence in expectation, and defined the rates as rates on the expectation. Since the objective f is bounded, convergence in expectation is equivalent to convergence in Probability. Furthermore, using Markov's inequality we can translate a rate of the form $E[|X|] \le \varepsilon(m)$ to a "low confidence" guarantee $P[|X| \le \varepsilon(m)/\delta] \le \delta$.

Background: Characterization of Learnability

Before presenting our results, we begin with a review of the known connections between learnability, stability, and uniform convergence, highlighting the issues which will be of importance later on.

Learnability and Uniform Convergence

As discussed in the introduction, a central notion for characterizing learnability is uniform convergence. Formally, we say that uniform convergence holds for a learning problem, if the empirical risks of hypotheses in the hypothesis class converges to their population risk uniformly, with a distribution-independent rate:

$$\sup_{D} E_{S\sim D^m} \sup_{\mathbf{h}\in\mathcal{H}} |F(\mathbf{h}) - F_S(\mathbf{h})| \xrightarrow{m\to\infty} 0.$$

It is straightforward to show that if uniform convergence holds, then a problem can be learned with the ERM learning rule.

For binary classification problems (where $Z = X \times \{0, 1\}$, each hypothesis is a mapping from X to $\{0, 1\}$, and $f(\mathbf{h};(\mathbf{x}, y)) = \mathbf{1}_{\mathbf{h}(\mathbf{x})\neq y}$), Vapnik and Chervonenkis (1971) showed that the finiteness of a simple combinatorial measure known as the VC-dimension implies uniform convergence. Furthermore, it can be shown that binary classification problems with infinite VC-dimension are not learnable in a distribution-independent sense. This establishes the condition of having finite VC-dimension, and thus also uniform convergence, as a necessary and sufficient condition for learnability.

Such a characterization can also be extended to regression, such as regression with squared loss, where $\mathbf{h}$ is now a real-valued function, and $f(\mathbf{h};(\mathbf{x}, y)) = (\mathbf{h}(\mathbf{x}) - y)^2$. The property of having finite fat-shattering dimension at all finite scales now replaces the property of having finite VC-dimension, but the basic equivalence still holds: a problem is learnable if and only if uniform convergence holds (Alon et al., 1997, see also Anthony and Bartlet, 1999, Chapter 19). These results are usually based on clever reductions to binary classification. However, the General Learning Setting that we consider is much more general than classification and regression, and includes setting where a reduction to binary classification is impossible.

To justify the necessity of uniform convergence even in the General Learning Setting, Vapnik attempted to show that in this setting, learnability with the ERM learning rule is equivalent to uniform convergence (Vapnik, 1998). Vapnik noted that this result does not hold, due to "trivial" situations. In particular, consider the case where we take an arbitrary learning problem (with hypothesis class H), and add to H a single Hypothesis $\tilde{\mathbf{h}}$ such that $f(\tilde{\mathbf{h}},\mathbf{z}) < \inf_{\mathbf{h}\in H} f(\mathbf{h},\mathbf{z})$ for all $\mathbf{z} \in Z$ (see figure 1 below). This learning problem is now trivially learnable, with the ERM learning rule which always picks $\tilde{\mathbf{h}}$.

Note that no assumptions Whatsoever are made on H - in particular, it can be arbitrarily complex, with no uniform convergence or any other particular property. Note also that such a phenomenon is not possible in the binary

classification setting,

whatsoever are made on H - in particular, it can be arbitrarily complex, with no uniform convergence or any other particular property. Note also that such a phenomenon is not possible in the binary classification setting, where $f(h;(\mathbf{x},y)) = \mathbf{1}_{\{h(x) \neq y\}}$, since on any (x,y) we will have hypotheses with $f(h;(\mathbf{x},y)) = f(\tilde{h};(\mathbf{x},y))$ and thus if H is very complex (has infinite VC dimension) then on every training set there will be many hypotheses with zero empirical error.

To exclude such "trivial" cases, Vapnik introduced a stronger notion of consistency, termed as "strict consistency", which in our notation is defined as

$$\forall c \in \mathbb{R}. \quad \inf_{\mathbf{h}:F(\mathbf{h}) \geq c} F_S(\mathbf{h}) \xrightarrow{m \to \infty} \inf_{\mathbf{h}:F(\mathbf{h}) \geq c} F(\mathbf{h}) .$$

where the convergence is in probability. The intuition is that we require the empirical risk of the ERM to converge to the lowest possible risk, even after discarding all the "good" hypotheses whose risk is smaller than some threshold. Vapnik then showed that such strict consistency of the ERM is in fact equivalent to (one-sided) uniform convergence, of the form

$$\sup_{\mathbf{h} \in \mathcal{H}} (F(\mathbf{h}) - F_S(\mathbf{h})) \xrightarrow{m \to \infty} 0$$

in probability. Note that this equivalence holds for every distribution separately, and does not rely on universal consistency of the ERM.

These results seem to imply that up to "trivial" situations, a uniform convergence property indeed char- acterizes learnability, at least using the ERM learning rule. However, as we will see later on, the situation is in fact not that simple.

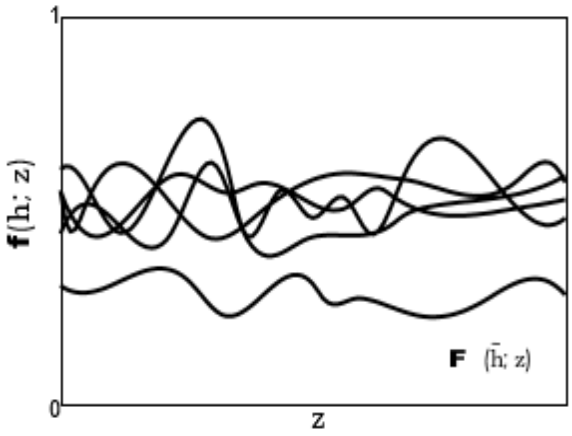


Figure 1: An example of a "trivial" learning situation. Each line represents some $\mathbf{h}$ H , and shows the value of $f(\mathbf{h},\mathbf{z})$ for all $\mathbf{z}$ Z. The hypothesis $\tilde{\mathbf{h}}$ dominates any other hypothesis (e.g., $f(\tilde{\mathbf{h}};\mathbf{z}) < f(\mathbf{h};\mathbf{z})$ uniformly for all $\mathbf{z}$), and thus the problem is learnable without uniform convergence or any other
property of H .

Uniform Convergence

1. Mathematical Definitions

A power series,

$$f(x) = \sum_{n=0} a_n x^n ,$$

is an example of a sum over a series of functions

$$f(x) = \sum_{n=0} g_n(x), \qquad (1)$$

where $g_n(x) = a_n x^n$. It is useful to consider the more general case. Let us considera sum of the form given in eq. (1) and ask whether the sum is convergent. If we consider each x separately, then we can determine whether the sum converges atthe point x. Suppose that the sum is determined to converge for all points $x \in A$, where A is some interval on the real axis. Typical intervals are: the *open* interval $a < x < b$, which we will denote by (a, b) and the *closed* interval $a \leq x \leq b$, which we will denote by $[a, b]$. Of course, we could also consider half-open, half-closed intervals, such as $a < x \leq b$, denoted by $(a, b]$ and $a \leq x < b$, denoted by $[a, b)$. In this notation, the parenthesis indicates that the endpoint is not included in the interval,

whereas the square bracket indicates that the endpoint is included in the Interval.

If $f(x)$ converges for all $x \in A$, we say that the sum given by eq. (1) is *point wise convergent* over the interval $x \in A$. In this case, A is called the *interval of convergence*. A classic example is the infinite geometric series,

$$\frac{1}{1-x} = \sum_{n=0} x^n, \qquad |x| < 1. \tag{2}$$

The above sum converges point wise to the function $(1 - x)^{-1}$ inside the open interval $-1 < x < 1$. Using the standard procedures, it is easy to see that the sum diverges for all $|x| \geq 1$.

Let us return to the general case of

$$f(x) = \sum_{n=0} g_n(x), \qquad x \in A,$$

where A is the interval of convergence. Suppose that the $g_n(x)$ are continuous functions. Does this imply that $f(x)$ is continuous? The great mathematician Augustin Louis Cauchy got the answer wrong. In 1821, he claimed to prove that all infinite sums of continuous functions are continuous. It took over 30 years before the error was properly corrected.

It is simple to provide a counterexample to Cauchy's claim. Consider the series:

$$f(x) = \sum_{n=1} g_n(x) = \sum_{n=1} \frac{x^2}{(1+nx^2)[1+(n-1)x^2]}. \tag{3}$$

Although this series looks complicated, we can simplify it using partial fractions. The following is an algebraic identity:

$$g_n(x) \equiv \frac{x^2}{(1+nx^2)[1+(n-1)x^2]} = \frac{1}{1+(n-1)x^2} - \frac{1}{1+nx^2}. \tag{4}$$

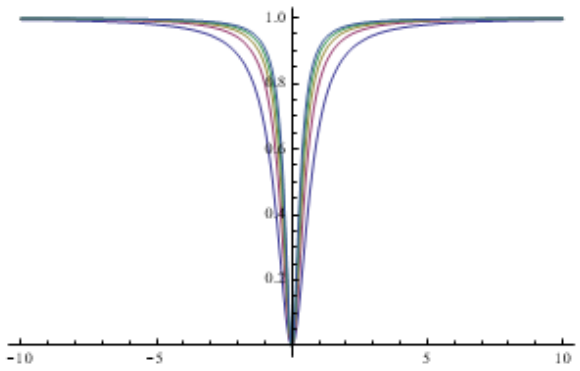
The sum over the first N terms is a telescoping series, which yields

$$\sum_{n=1} \frac{x^2}{(1+nx^2)[1+(n-1)x^2]} = 1 - \frac{1}{1+x^2} + \frac{1}{1+x^2} - \frac{1}{1+2x^2} + \cdots + \frac{1}{1+(N-1)x^2} - \frac{1}{1+Nx^2} = 1 - \frac{1}{1+Nx^2} = \frac{Nx^2}{1+Nx^2}. \tag{5}$$

Thus,

$$f(x) = \lim_{N\to\infty} \frac{Nx^2}{1+Nx^2} = \begin{cases} 1, & \text{for } x \neq 0, \\ 0, & \text{for } x = 0. \end{cases}$$

That is, $f(x)$ is discontinuous at $x = 0$. Below, I have plotted $Nx^2/(1 + Nx^2)$ for $N = 2, 4, 6, 8$ and 10. As N increases, the curves begin to approach the function $f(x)$, which is equal to 1 for all $x \neq 0$. Of course, $f(0) = 0$, so in the limit of $N \to \infty$, the functions exhibits a discontinuity.



$$f_N(x) \equiv \sum_{n=1} \frac{x^2}{(1+nx^2)[1+(n-1)x^2]} = \frac{Nx^2}{1+Nx^2}.$$

Then for $x \neq 0$,

$$|f(x) - f_N(x)| = \frac{1}{1+Nx^2} < \varrho, \tag{6}$$

Which implies that

$$N > \frac{1-\varrho}{\varrho x^2}. \tag{7}$$

That is, given an ϱ no matter how small, one can always find an N such that $|f(x) - f_N(x)| < \varrho$. This implies that for $x \neq 0$, $\lim_{N\to\infty} f_N(x) = f(x)$. Note that as $x \to 0$,

the required value of N grows arbitrarily large.This means that the *rate* of convergence is getting slower and slower the closer x is to the origin. This is the source of the ultimate discontinuity at $x = 0$.

These observations motivate the following definition.

Definition: Given $f_n(x) = \sum_{n=1} f_n(x)$ and $\lim f_N(x) = f(x)$, where $x \in A$, the sum is *uniformly convergent* in the interval $x \in A$ if given a positive error bound ϱ, there always exists a response N such that for any $n \geq N$ we have $|f(x) - f_n(x)| < \varrho$. The *same* N must work for all $x \in A$.

The concept of uniform convergence is illustrated in the above figure, taken from ref. 2. We see that the approximation f_n to f lies completely within an ϱ-band over the entire interval.

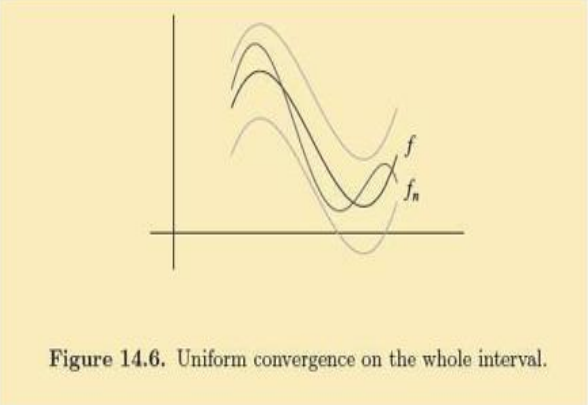


Figure 14.6. Uniform convergence on the whole interval.

By comparing the definition of uniform convergence with eqs. (6) and (7), it is clear that the sum given in eq. (3) is point wise convergent but is *not* uniformly convergent over any interval that contains the point $x = 0$. Indeed, uniform convergence is a more stringent requirement than point wise convergence. However, the advantage of uniform convergence is that the properties of the functions $g_n(x)$ (such as continuity) are preserved by the infinite sum.

2 Theorems about uniformly converging series

We quote a few key theorems without proofs (for details, see e.g., refs. 1 and 2 given at the end of these notes).

Theorem 1: If $f(x) = \sum_{n=1} g_n(x)$ converges uniformly over the interval $x \in A$,and if the $g_n(x)$ are continuous at every point $x \in A$, then $f(x)$ is also continuous at every point $x \in A$.

Since $f(x)$ defined in eq. (3) is discontinuous at $x = 0$, it follows that $f(x)$ is not uniformly convergent over any interval that includes the point $x = 0$. However, the converse to Theorem 1 is false. In particular, there are many examples of non-uniformly convergent sums that are continuous over the interval of convergence.

Theorem 2: If $f(x) = \sum_{n=1} g_n(x)$ converges uniformly over the open interval $x \in A$, and if the $g_n(x)$ are differentiable at every point in A, then $f(x)$ is also differentiable at every point $x \in A$, and $f'(x) = \sum_{n=1} g_n(x)$

That is, it is legal to interchange the order of summation and differentiation. While this interchange is clearly valid for finite sums, it requires proof if the sum is over an infinite number of terms. For non-uniformly convergent sums, interchanging the order of an infinite summation and differentiation may fail. As an example, consider

$$f(x) = \sin x \sum_{n=1} g_n(x), \qquad (8)$$

Where $g_n(x)$ is defined in eq. (4). Using eq. (5), it follows that:

$$f_N(x) \equiv \sin x \sum_{n=1} g_n(x) = \frac{Nx^2 \sin x}{1 + Nx^2}$$

Thus,

$$f(x) = \lim_{N \to \infty} f_N(x) = \sin x,$$

Which holds for all real values of x including the point $x = 0$. Because of the x^2 in the numerator of eq. (4), it is clear that $g'(0) = 0$ for all n. Thus,

$$\sum_{n=1} g_n'(0) = 0.$$

However, since $f(x) = \sin x$, we have $f'(x) = \cos x$, and $f'(0) = 1$. Thus,

$$f'(0) \quad \sum_{n=1} g_n \quad (0), \tag{9}$$

The reason for this odd behavior is that $f(x)$ is not uniformly convergent at $x = 0$.

Theorem 3: If $f(x) = \sum_{n=1} g_n(x)$ converges uniformly over the closed interval $x \in [a, b]$, and if the $g_n(x)$ are integrable over the interval $[a, b]$, then $f(x)$ is also integrable over the interval $[a, b]$, and

$$\int_a^b f(x)dx = \sum_{n=1} \int_a^b g_n(x)dx.$$

That is, it is mathematically permissible to interchange the order of summation and integration. While this interchange is clearly valid for finite sums, it requires proof if the sum is over an infinite number of terms. For non-uniformly convergent sums, interchanging the order of an infinite summation and integration may fail. As an example, consider

$$f(x) = \sum_{n=1} g_n(x), \qquad -1 < x \leq 1, \tag{10}$$

Where

$$g_n(x) \equiv x^{n-1}(1-x)^2 \qquad (n-1)^2 \tag{11}$$

We recognize that the sum of the first N terms of $f(x)$ is a telescoping sum, whichis easily computed:

$$f_N(x) = \sum_{n=1} g_n(x) = N^2 x^N (1-x).$$

12)

Thus, $f(x) = \lim_{N\to\infty} f_N(x) = 0$ for $|x| < 1$ and $f(1) = 0$ due to the factor of $1 - x$ in the $g_n(x)$. Hence,

$$\int_0^1 f(x)\,dx = 0. \tag{13}$$

However, notice that

$$\sum_{n=1} \int_0^1 g_n(x)\,dx = \int_0^1 dx \sum_{n=1} g_n(x) = \int_0^1 f_N(x)\,dx = N^2 \int_0^1 x^N(1-x)\,dx$$

$$= N^2 \left[\frac{1}{N+1} - \frac{1}{N+2}\right] = \frac{N^2}{(N+1)(N+2)}, \tag{14}$$

Where we have used the fact that it is always permissible to interchange the order of integration and summation if the latter is a sum over a finite number of terms.Taking the limit of eq. (14) as $N \to \infty$ yields

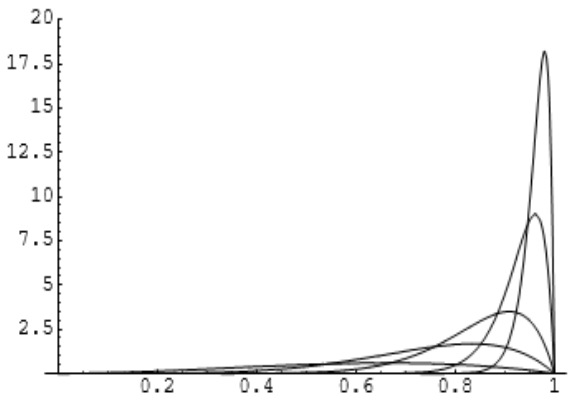
$$\sum_{n=1} \int_0^1 g_n(x)\,dx = 1. \tag{15}$$

Comparing eq. (13) with eq. (15), we conclude that

$$\int_0^1 dx \sum_{n=1} g_n(x) \neq \sum_{n=1} \int_0^1 g_n(x)dx. \tag{16}$$

Again, this behavior can be attributed to the fact that the sum is not uniformly convergent at $x = 1$.

To understand what is happening, consider a plot of $f_N(x)$ [eq. (12)] for various values of N. Below, I have plotted $f_N(x)$ for $N = 2$, 5, 10, 25 and 50. As N increases, the hump in the graph of $f_N(x)$ gets higher and narrower and is pushed further to the right.



As $N \to \infty$, $f_N(x) \to 0$ but in a very non-uniform way, since the area under thegraph of $f_N(x)$ approaches 1 in the same limit!

3 A test for uniform convergence

The Weierstrass *M*-test: If $f(x) = \Sigma_{\infty}^{n=1} g_n(x)$, and if $|g_n(x)| \le M_n$ for every n greater than orequal to some some fixed integer N, for all $x \in A$, and if $\Sigma_{n=1}^{\infty} M_n$ converges, then $\Sigma_{n=1}^{\infty} g_n(x)$ converges uniformly to $f(x)$ over the intervalx $\in A$. Moreover, at every point $x \in A$, the convergence is absolute.

This is the analog of the comparison test for numerical series. The converse of the Weierstrass *M* -test is false. Namely, it is possible for a uniformly convergent series to fail the Weierstrass *M*-test. An example of this is

$$\sum_{n=1} \frac{(-1)^{n+1}}{x^2+n}. \qquad (17)$$

This sum is only conditionally convergent. Thus, it cannot pass the Weierstrass *M* -test, since any series that satisfies the Weierstrass *M* -test must be absolutely convergent. Nevertheless, one can prove that eq. (17) is also uniformly convergent.

4. The uniform convergence of power series

Finally, we can now understand why power series are so nice. As a consequence of the Weierstrass *M*-test, we have this important result.

Theoram 4: *If* $f(x) = \Sigma_{n=0}^{\infty} a_n x^n$ is a power series with a finite radius convergence $R > 0$, then for any $0 < r < R$, this power series converges uniformly and absolutely over the *closed* interval $[-r, r]$.* If the radius of of convergence is infinite, then the power series converges uniformly over the closed interval $[-r, r]$ for any finite positive value of r.

Proof: For any $|x| \le r$, we have $|a_n x^n| \le |a_n| r^n$. But, $\Sigma_{n=0}^{\infty} |a_n| r^n$ converges absolutely (as a consequence of the ratio test). Hence the Weierstrass *M*–test applies.

As a result, $f(x)$ is continuous over any closed interval $[r, r]$ (for $r < R$), and one can differentiate and integrate power series by differentiating or integrating each term in the series. Returning to the example of the infinite geometric series, we conclude that eq. (2) is uniformly convergent over any closed interval $[-r, r]$, where $0 < r < 1$. However, it would be *incorrect* to claim that eq. (2) is uniformly Convergent over the open interval $(-1, 1)$. The problem here is that the rate of convergence is infinitely slow as $x \to \pm 1$. This is not surprising, since eq. (2) diverges at $x = \pm 1$.

For the general power series $f(x) = \Sigma_{n=0}^{\infty} a_n x^n$ with radius of convergence R, the series converges for $|x| < R$ and diverges for $|x| > R$. However the convergence properties at $x = \pm R$ must be checked separately. The following theoremis relevant.

Theoram 5: Given a power series $F(x) = \Sigma_{n=0}^{\infty} a_n x^n$ with a finite radius of convergence $R > 0$ and an interval of convergence A. Then,

a. If $A = [-R, R]$, then the series converges uniformly (but not necessarily absolutely) over the closed interval $[-R, R]$.
b. If $A = (-R, R]$, then the series converges uniformly (but not necessarilyabsolutely) over any closed interval $[a, R]$ for all $-R < a < R$.
c. If $A = [-R, R)$, then the series converges uniformly (but not necessarilyabsolutely) over any closed interval $[-R, b]$ for all $-R < b < R$.
d. If $A = (-R, R)$, then the series converges uniformly and absolutely over anyclosed interval $[a, b]$ for all $-R < a < b < R$.

The infinite geometric series [eq. (2)] is an example of case 4 of Theorem 5.
An example of case 2 is

$$\ln(1+x) = \sum_{n=1} (-1)^{n+1} \frac{x^n}{n}, \qquad -1 < x \le 1.$$

This sum is conditionally convergent at $x = 1$ and divergent at $x = -1$. Thus, we conclude that the series converges uniformly on any closed interval $a \le x \le 1$ where $-1 < a < 1$. In particular, the sum is uniformly convergent at $x = 1$.

Our final example is Euler's dilogarithm,

which is defined by its power series

$$\mathrm{Li}_2(x) = \sum_{n=1} \frac{x^n}{n^2}, \qquad |x| \le 1.$$

The radius of convergence of this power series is 1, and the series converges uniformly and absolutely at $x = \pm 1$. In fact, one can show that:

$$\mathrm{Li}_2(1) = \frac{\pi^2}{6}, \qquad \mathrm{Li}_2(-1) = -\frac{\pi^2}{12}.$$

This is an example of case 1 of Theorem 5.

Learnability and Stability

Instead of focusing on the hypothesis class, and ensuring uniform convergence of the empirical risks of hypothesis in this class, an alternative approach is to directly control the variance of the learning rule. Here, it is not the complexity of the hypothesis class which matters, but rather the way that the learning rule explores this hypothesis class. This alternative approach leads to the notion of stability in learning. It is important to note that stability is a property of a learning rule, not of the hypothesis class.

In the context of modern learning theory, [1] the use of stability of Rogers and Wagner (1978), which noted that the sensitivity of a learning algorithm with regard to small changes in the sample controls the variance of the leave-one-out estimate. The authors used this observation to obtain generalization bounds (w.r.t. the leave-one-out estimate) for the *k*-nearest neighbor algorithm. It is interesting to note that a uniform convergence approach for analyzing this algorithm simply cannot work, because the "hypothesis class" in this case has unbounded complexity. These results were later extended to other "local" learning algorithms (see Devroye et al., 1996 and references therein). In addition, practical methods have been developed to introduce stability into learning algorithms, in particular the Bagging technique introduced by Breiman (1996).

Over the last decade, stability was studied as a generic condition for learnability. Kearns and Ron (1999) showed that an algorithm operating on a hypothesis class with finite VC dimension is also stable (under a certain definition of stability). Bousquet and Elisseeff (2002) introduced a strong notion of stability (denoted as *uniform stability*) and showed that it is a sufficient condition for learnability, satisfied by popular learning algorithms such as regularized linear classifiers and regressors in Hilbert spaces (including several variants of SVM). Kutin and Niyogi (2002) introduced several weaker variants of stability, and showed how they are sufficient to obtain generalization bounds for algorithms stable in their sense.

The papers above mainly considered stability as a *sufficient* condition for learnability. A more recent line of work (Rakhlin et al., 2005; Mukherjee et al., 2006) studied stability as a *necessary* condition for learnability. However, the line of argument is specific to settings where uniform convergence holds and is necessary for learning. With this assumption, it is possible to show that the ERM algorithm is stable, and thus stability is also a necessary condition for learning. However, as we will see later on in our paper, uniform convergence is in fact not necessary for learning in the General Learning Setting, and stability plays there a key role which has nothing to do with uniform convergence.

Finally, it is important to note that the results cited above make use of many different definitions of stability, which unfortunately are not always comparable. All of them measure stability as the amount of change in the algorithm's output as a function of small changes to the sample on which the algorithm is run. However, "amount of change to the output" and "small changes to the sample" can be defined in many different ways. "Amount of change to the output" can mean change in risk, change in loss with respect to particular examples, or supremum of change in loss over all examples. "Small changes to the sample" usually mean either deleting one example or replacing it with another one (and even here, one can talk about removing/replacing one instance at random, or in some arbitrary manner). Finally, this measure of change can be measured with respect to any arbitrary sample, in expectation over samples drawn from the underlying distribution; or in high probability over samples. High Confidence Learnability

So far, we have presented all our results in terms of expectation: namely, the rate at which the expected risk converges to the lowest possible risk. By Markov's inequality, we can always convert these bounds to bounds which hold with probability 1 δ over the sample, and the bounds depend linearly on 1/δ. However, in supervised classification, if we have learnability at all, then we have learnability at rates which are logarithmic in 1/δ .

Moreover, we show that such conversions can in fact be necessary: we give a learning problem which is learnable with an ERM algorithm, and the ERM is stable, but the dependence on the confidence parameter δ cannot be better than linear. This shows that both learnability and stability (under our definitions) of the ERM learning rule are not sufficient to ensure logarithmic dependence on 1/δ. Also, this gives a nice illustration to the fundamental differences between the General Learning Setting and supervised classification, where in the latter case learnability implies logarithmic dependence on 1/δ.

Discussion and Conclusions

In the familiar setting of supervised classification problems, the question of learnability is reduced to that of uniform convergence of empirical risks to their expectations. Therefore, for the purposes of establishing learnability, there is no need to look beyond the ERM.

In this paper, we showed that in the General Learning Setting, which includes more general problems, this equivalence does not hold, and the situation is substantially more complex. ERM might work without any uniform convergence, and learnability might be possible only with a non-ERM algorithm. We are therefore in need of a new understanding of the question of learnability, that applies more broadly than just to supervised classification.

In studying learnability in the General Setting, Vapnik (1995) focuses solely on empirical risk minimiza/- tion, which we have seen to be insufficient for understanding learnability. Furthermore, for empirical risk minimization, Vapnik establishes uniform convergence as a necessary and sufficient condition not for ERM consistency, but rather for *strict* consistency of the ERM. We have seen that even in rather nontrivial problems, where the ERM is consistent and generalizes, strict consistency does not hold. This perhaps gives an indication that strict consistency might be too strict.

On the other hand, we have seen that stability is both a sufficient and necessary condition for learning, even in the General Learning Setting where uniform convergence fails to characterize learnability. A previous stability-based characterization (Mukherjee et al., 2006) relied on uniform convergence and thus applied only to restricted setting. Extending the characterization beyond these settings is particularly interesting, since for supervised classification the question of learnability is already essentially solved.

In studying the question of learnability and its relation to stability, we encountered several differences be- tween this more general setting, and settings such as supervised classification where learnability is equivalent to uniform convergence.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM, 44(4):615–631, 1997.

M. Anthony and P. Bartlet. Neural Network Learning: Theoretical Foundations. Cambridge University Press, 1999.

Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. Journal of the Association for Computing Machinery, 36(4):929–965, October 1989.

O. Bousquet and A. Elisseeff. Stability and generalization. Journal of Machine Learning Research, 2:499– 526, 2002. ISSN 1533-7928.

L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, 1996.

L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Springer, 1996.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and Computation, 100(1):78–150, 1992.

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation.
Neural Computation, 11(6):1427–1453, 1999.

M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, pages 341–352, 1992.

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, pages 275–282, 2002.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. Advances in Computational Mathematics, 25(1-3):161–193, 2006.

S. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. Analysis and Applications, 3 (4):397–419, 2005.

W. Rogers and T. Wagner. A finite sample distribution-free performance bound for local discrimination rules.

Annals of Statistics, 6(3):506–514, 1978.

R.E. Schapire. The strength of weak learnability. In 30th Annual Symposium on Foundations of Computer Science, pages 28–33, October 1989.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In Proceedings of the 22nd Annual Conference on Computational Learning Theory, 2009a.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability and stability in the general learning setting. In Proceedings of the 22nd Annual Conference on Computational Learning Theory, 2009b.

David Bressoud, A Radical Approach to Real Analysis (Mathematical Associ- ation of America, Washington, DC, 1994).

Brian S. Thomson, Judith B. Bruckner and Andrew M. Bruckner, Elementary Real Analysis (Prentice-Hall, Inc., Upper Saddle River, NJ, 2001).