Journal of Machine Learning Research 11 (2010) 2635-2670 Submitted 4/10; Revised 9/10; Published 10/10

Learnability, Stability and Uniform Convergence

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Givat Ram, Jerusalem 91904, Israel

Ohad Shamir OHADSH@MICROSOFT.COM
Microsoft Research
One Memorial Drive
Cambridge, MA 02142, USA

Nathan Srebro NATI@TTIC.EDU
Karthik Sridharan KARTHIK@TTIC.EDU
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave.
Chicago, IL 60637, USA

Editor: Nicolò Cesa-Bianchi

Abstract

The problem of characterizing 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. Moreover, we show that the conditions for learnability in the general setting are significantly more complex than in supervised classification and regression.

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

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)

$$F(\mathbf{h}) = \mathbb{E}_{Z\sim\mathcal{D}}[f(\mathbf{h};Z)] \tag{1}$$

over some hypothesis class $\mathcal{H}$, where the distribution $\mathcal{D}$ of Z is unknown, based on i.i.d. sample $z_1,\ldots,z_m$ drawn from $\mathcal{D}$ (and full knowledge of f and $\mathcal{H}$). This General Setting subsumes supervised classification and regression, certain unsupervised learning problems, density estimation and more. For example, in supervised 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. See Section 2 for formal definitions and further examples.

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 $z_1,\ldots,z_m$.

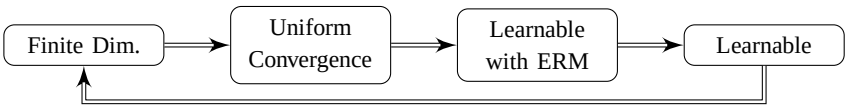
For supervised classification and regression problems, it is well known that a problem is learnable if and only if the empirical risks

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

for all $\mathbf{h} \in \mathcal{H}$ converge uniformly to the population risk (Blumer et al., 1989; Alon et al., 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.

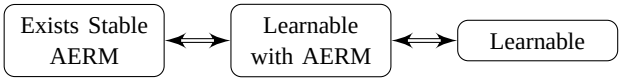
The equivalence of uniform convergence and learnability was formally established only in the supervised classification and regression setting. In the more general setting, the "rightward" implications in the diagram above still hold: finite fat-shattering dimensions, uniform convergence, as well as ERM 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 learnable without uniform convergence (see Section 3.1). 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 Section 4.1 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, in Section 4.2 we show a modified example which *is* learnable using empirical risk minimization, 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 in Section 3.1, 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" and (AERM - see precise definition in Section 2). 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. In Section 6, 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 Section 7, 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).

2. 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 $\mathcal{H}$, an instance set $\mathcal{Z}$ (with a sigma-algebra), and an objective function (e.g., "loss" or "cost") $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. Throughout this paper we assume the function is bounded by some constant B, that is $|f(\mathbf{h};\mathbf{z})| \leq B$ for all $\mathbf{h} \in \mathcal{H}$ and $z \in \mathcal{Z}$.

Given a distribution $\mathcal{D}$ on $\mathcal{Z}$, the quality of each hypothesis $\mathbf{h} \in \mathcal{H}$ is measured by its *risk* $F(\mathbf{h})$, which is defined as $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{h};\mathbf{z})]$. While $\mathcal{H}$, $\mathcal{Z}$ and $f(\mathbf{h};\mathbf{z})$ are known to the learner, we assume that $\mathcal{D}$ is unknown. Ideally, we would like to pick $\mathbf{h} \in \mathcal{H}$ whose risk is as close as possible to $\inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})$. Since the underlying distribution $\mathcal{D}$ 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 $\mathbf{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ from sequences of instances in $\mathcal{Z}$ to hypotheses. We refer to sequences $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ 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 $S \sim \mathcal{D}^m$ of m i.i.d. draws from $\mathcal{D}$.

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 $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, let $\mathcal{H}$ be a set of functions $\mathbf{h} : \mathcal{X} \mapsto \{0,1\}$, and let $f(\mathbf{h};(\mathbf{x},y)) = \mathbb{1}_{[\mathbf{h}(\mathbf{x})\neq y]}$. 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 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are bounded subsets of $\mathbb{R}^n$ and $\mathbb{R}$ respectively, let $\mathcal{H}$ be a set of bounded functions $\mathbf{h} : \mathcal{X}^n \mapsto \mathbb{R}$, 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 $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, where $\mathcal{X}$ is a bounded subset of an RKHS, let $\mathcal{H}$ 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 $\mathcal{Z} = \mathbb{R}^n$, let $\mathcal{H}$ be all subsets of $\mathbb{R}^n$ of size k, and let $f(\mathbf{h};\mathbf{z}) = \min_{\mathbf{c}\in\mathbf{h}} \|\mathbf{c} - \mathbf{z}\|^2$. 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 $\mathbf{h}$.

- **Density Estimation:** Let $\mathcal{Z}$ be a subset of $\mathbb{R}^n$, let $\mathcal{H}$ be a set of bounded probability densities on $\mathcal{Z}$, and let $f(\mathbf{h};\mathbf{z}) = -\log(\mathbf{h}(\mathbf{z}))$. Here, $f(\cdot)$ is simply the negative log-likelihood of an instance $\mathbf{z}$ according to the hypothesis density $\mathbf{h}$. 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 $\mathbf{z} \in \mathcal{Z}$.

- **Stochastic Convex Optimization in Hilbert Spaces:** Let $\mathcal{Z}$ be an arbitrary measurable set, let $\mathcal{H}$ 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 $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{h};\mathbf{z})]$, where the distribution over $\mathcal{Z}$ 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 $\mathbf{h} \in \mathcal{H}$ 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 $\mathbf{A}$ is consistent with rate $\varepsilon_{\text{cons}}(m)$ under distribution $\mathcal{D}$ if for all m,

$$\mathbb{E}_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S)) - F^*] \leq \varepsilon_{\text{cons}}(m), \tag{2}$$

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

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

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

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

A learning rule $\mathbf{A}$ *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

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

$\mathcal{Z}$,$\mathbf{z}$	Instance domain and a specific instance.
$\mathcal{H}$,$\mathbf{h}$	Hypothesis class and a specific hypothesis.
$f(\mathbf{h},\mathbf{z})$	Loss of hypothesis $\mathbf{h}$ on instance $\mathbf{z}$
B	$\sup_{\mathbf{h},\mathbf{z}} \lvert f(\mathbf{h};\mathbf{z})\rvert$
$\mathcal{D}$	Underlying distribution on instance domain $\mathcal{Z}$
S	Empirical sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$, sampled i.i.d. from $\mathcal{D}$
m	Size of empirical sample S
$\mathbf{A}(S)$	Learning rule $\mathbf{A}$ applied on empirical sample S
$\varepsilon_{\text{cons}}(m)$	Rate of consistency for a learning rule
$F(\mathbf{h})$	Risk of hypothesis $\mathbf{h}$, $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{h};\mathbf{z})]$
F^*	$\inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})$
$F_S(\mathbf{h})$	Empirical risk of hypothesis $\mathbf{h}$ on sample S, $\frac{1}{m}\sum_{\mathbf{z}\in S} f(\mathbf{h};\mathbf{z})$
$\hat{\mathbf{h}}_S$	An ERM hypothesis, $F_S(\hat{\mathbf{h}}_S) = \inf_{\mathbf{h}\in\mathcal{H}} F_S(\mathbf{h})$
$\varepsilon_{\text{erm}}(m)$	Rate of AERM for a learning rule
$\varepsilon_{\text{stable}}(m)$	Rate of stability for a learning rule
$\varepsilon_{\text{gen}}(m)$	Rate of generalization for a learning rule

Table 1: Table of Notation

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

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

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

We say that a rule $\mathbf{A}$ is an *AERM (Asymptotic Empirical Risk Minimizer)* with rate $\varepsilon_{\text{erm}}(m)$ under distribution $\mathcal{D}$ if:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right] \leq \varepsilon_{\text{erm}}(m)$$

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

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

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\lvert F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\rvert\right] \leq \varepsilon_{\text{gen}}(m).$$

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

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 $\mathbb{E}[\lvert X\rvert] \leq \varepsilon(m)$ to a "low confidence" guarantee $\mathbb{P}[\lvert X\rvert > \varepsilon(m)/\delta] \leq \delta$. See Section 7 for a further discussion on this issue.

3. 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.

3.1 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:

$$\sup_{\mathcal{D}} \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{\mathbf{h}\in\mathcal{H}} |F(\mathbf{h}) - F_S(\mathbf{h})|\right] \stackrel{m\to\infty}{\longrightarrow} 0.$$

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 $\mathcal{Z} = \mathcal{X}\times\{0,1\}$, each hypothesis is a mapping from $\mathcal{X}$ to $\{0,1\}$, and $f(\mathbf{h};(\mathbf{x},y)) = \mathbb{1}_{\{\mathbf{h}(\mathbf{x})\neq y\}}$), 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 $\mathcal{H}$), and add to $\mathcal{H}$ a single hypothesis $\tilde{\mathbf{h}}$ such that $f(\tilde{\mathbf{h}},\mathbf{z}) < \inf_{\mathbf{h}\in\mathcal{H}} f(\mathbf{h},\mathbf{z})$ for all $\mathbf{z}\in\mathcal{Z}$ (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 $\mathcal{H}$ - 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 $f(h;(\mathbf{x},y)) = \mathbb{1}_{\{\mathbf{h}(\mathbf{x})\neq y\}}$, since on any (x,y) we will have hypotheses with $f(h;(\mathbf{x},y)) = f(\tilde{h};(\mathbf{x},y))$ and thus if $\mathcal{H}$ 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

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

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

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

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 characterizes learnability, at least using the ERM learning rule. However, as we will see later on, the situation is in fact not that simple.

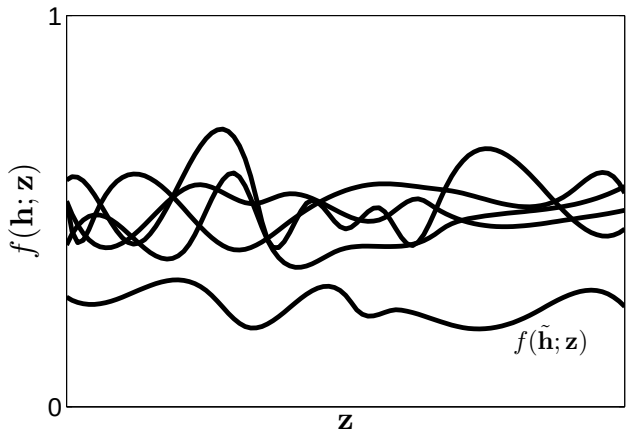


Figure 1: An example of a "trivial" learning situation. Each line represents some $\mathbf{h} \in \mathcal{H}$, and shows the value of $f(\mathbf{h},\mathbf{z})$ for all $\mathbf{z} \in \mathcal{Z}$. The hypothesis $\tilde{\mathbf{h}}$ dominates any other hypothesis (e.g., $f(\tilde{\mathbf{h}};\mathbf{z})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

3.2 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 can be traced back at least to the work 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

1. In a more general mathematical context, stability has been around for much longer. The necessity of stability for so-called inverse problems to be well posed was first recognized by Hadamard (1902). The idea of regularization (that is, introducing stability into ill-posed inverse problems) became widely known through the works of Tikhonov (1943) and Phillips (1962). We return to the notion of regularization later on.

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. For further discussion of this issue, see Appendix A.

4. Gaps Between Learnability, Uniform Convergence and ERM

In this section, we study a special case of the General Learning Setting, where there is a real gap between learnability and uniform convergence, in the sense that there are non-trivial problems where no uniform convergence holds (not even in a local sense), but they are still learnable. Moreover, some of these problems are learnable with an ERM (again, without any uniform convergence), and some are not learnable with an ERM, but rather with a different mechanism. We also discuss why this peculiar behavior does not formally contradict Vapnik's results on the equivalence of strict consistency of the ERM and uniform convergence, as well as the important role that regularization seems to play here, but in a different way than in standard theory.

4.1 Learnability without Uniform Convergence : Stochastic Convex Optimization

A stochastic convex optimization problem is a special case of the General Learning Setting discussed above, with the added constraints that the objective function $f(\mathbf{h};\mathbf{z})$ is Lipschitz-continuous and convex in $\mathbf{h}$ for every $\mathbf{z}$, and that $\mathcal{H}$ is closed, convex and bounded. We will focus here on problems where $\mathcal{H}$ is a subset of a Hilbert space. A special case is the familiar linear prediction setting, where $\mathbf{z} = (\mathbf{x}, y)$ is an instance-label pair, each hypothesis $\mathbf{h}$ belongs to a subset $\mathcal{H}$ of a Hilbert space, and $f(\mathbf{h};\mathbf{x},y) = \ell(\langle \mathbf{h}, \phi(\mathbf{x})\rangle, y)$ for some feature mapping ϕ and a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, which is convex w.r.t. its first argument.

The situation in which the stochastic dependence on $\mathbf{h}$ is linear, as in the preceding example, is fairly well understood. When the domain $\mathcal{H}$ and the mapping ϕ are bounded, we have uniform convergence, in the sense that $|F(\mathbf{h}) - F_S(\mathbf{h})|$ is uniformly bounded over all $\mathbf{h} \in \mathcal{H}$ (see Sridharan et al., 2008). This uniform convergence of $F_S(\mathbf{h})$ to $F(\mathbf{h})$ justifies choosing the empirical minimizer $\hat{\mathbf{h}}_S = \arg\min_{\mathbf{h}} F_S(\mathbf{h})$, and guarantees that the expected value of $F(\hat{\mathbf{h}}_S)$ converges to the optimal value $F^* = \inf_{\mathbf{h}} F(\mathbf{h})$.

Even if the dependence on $\mathbf{h}$ is not linear, it is still possible to establish uniform convergence (using covering number arguments) provided that $\mathcal{H}$ is finite dimensional. Unfortunately, when we turn to infinite dimensional hypothesis spaces, uniform convergence might not hold and the problem might not be learnable with empirical minimization. Surprisingly, it turns out that this does not imply that the problem is unlearnable. We will show that using a regularization mechanism, it is possible to devise a learning algorithm for any stochastic convex optimization problem, even when uniform convergence does not hold. This mechanism is fundamentally related to the idea of stability, and will be a good starting point for our more general treatment of stability and learnability in the next section of the paper.

We now turn to discuss our first concrete example. Consider the convex stochastic optimization problem given by

$$f^{(3)}(\mathbf{h};(\mathbf{x},\alpha)) \;=\; \|\alpha * (\mathbf{h}-\mathbf{x})\| \;=\; \sqrt{\sum_i \alpha^2[i](\mathbf{h}[i]-\mathbf{x}[i])^2}\,, \tag{3}$$

where for now we let $\mathcal{H}$ to be the d-dimensional unit sphere $\mathcal{H} = \left\{\mathbf{h} \in \mathbb{R}^d : \|\mathbf{h}\| \leq 1\right\}$, we let $\mathbf{z} = (\mathbf{x}, \alpha)$ with $\alpha \in [0,1]^d$ and $\mathbf{x} \in \mathcal{H}$, and we define $u * v$ to be an element-wise product. We will first consider a sequence of problems, where $d = 2^m$ for any sample size m, and establish that we cannot expect a convergence rate which is independent of the dimensionality d. We then formalize this example in infinite dimensions.

One can think of the problem in Equation (3) as that of finding the "center" of an unknown distribution over $\mathbf{x} \in \mathbb{R}^d$, where we also have stochastic per-coordinate "confidence" measures $\alpha[i]$. We will actually focus on the case where some coordinates are missing, namely that $\alpha[i] = 0$.

Consider the following distribution over $(\mathbf{x}, \alpha)$: $\mathbf{x} = 0$ with probability one, and α is uniform over $\{0,1\}^d$. That is, $\alpha[i]$ are i.i.d. uniform Bernoulli. For a random sample $(\mathbf{x}_1, \alpha_1), \ldots, (\mathbf{x}_m, \alpha_m)$ if $d > 2^m$ then we have that with probability greater than $1 - e^{-1} > 0.63$, there exists a coordinate $j \in 1 \ldots d$ such that all confidence vectors α_i in the sample are zero on the coordinate j, that is $\alpha_i[j] = 0$ for all $i = 1..m$. Let $\mathbf{e}_j \in \mathcal{H}$ be the standard basis vector corresponding to this coordinate. Then

$$F_S^{(3)}(\mathbf{e}_j) \;=\; \frac{1}{m}\sum_{i=1}^{m} \left\|\alpha_i * (\mathbf{e}_j - 0)\right\| \;=\; \frac{1}{m}\sum_{i=1}^{m} |\alpha_i[j]| \;=\; 0,$$

where $F_S^{(3)}(\cdot)$ denotes the empirical risk w.r.t. the function $f^{(3)}(\cdot)$. On the other hand, letting $F^{(3)}(\cdot)$ denote the actual risk w.r.t. $f^{(3)}(\cdot)$, we have

$$F^{(3)}(\mathbf{e}_j) \;=\; \mathbb{E}_{\mathbf{x},\alpha}\left[\left\|\alpha * (\mathbf{e}_j - 0)\right\|\right] \;=\; \mathbb{E}_{\mathbf{x},\alpha}\left[|\alpha[j]|\right] \;=\; 1/2.$$

Therefore, for any m, we can construct a convex Lipschitz-continuous objective in a high enough dimension such that with probability at least 0.63 over the sample, $\sup_{\mathbf{h}} \left|F^{(3)}(\mathbf{h}) - F_S^{(3)}(\mathbf{h})\right| \geq 1/2$. Furthermore, since $f(\cdot;\cdot)$ is non-negative, we have that $\mathbf{e}_j$ is an empirical minimizer, but its expected value $F^{(3)}(\mathbf{e}_j) = 1/2$ is far from the optimal expected value $\min_{\mathbf{h}} F^{(3)}(\mathbf{h}) = F^{(3)}(0) = 0$.

To formalize the example in a sample-size independent way, take $\mathcal{H}$ to be the unit sphere 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 $\mathbf{v}[j] = \langle \mathbf{v}, \mathbf{e}_j \rangle$ w.r.t this basis. The confidences α are now a mapping of each coordinate to $[0,1]$. That is, an infinite sequence of reals in $[0,1]$. The element-wise product operation $\alpha * \mathbf{v}$ is defined with respect to this basis and the objective function $f^{(3)}(\cdot)$ of Equation (3) is well defined in this infinite-dimensional space.

We again take a distribution over $z = (\mathbf{x}, \alpha)$ where $\mathbf{x} = 0$ and α is an infinite i.i.d. sequence of uniform Bernoulli random variables (that is, a Bernoulli process with each α_i uniform over $\{0,1\}$ and independent of all other α_j). Now, for any finite sample there is almost surely a coordinate j with $\alpha_i[j] = 0$ for all i, and so we a.s. have an empirical minimizer $F_S^{(3)}(\mathbf{e}_j) = 0$ with $F^{(3)}(\mathbf{e}_j) = 1/2 > 0 = F^{(3)}(0)$.

As a result, we see that the empirical values $F_S^{(3)}(\mathbf{h})$ do not converge uniformly to their expectations, and empirical minimization is not guaranteed to solve the problem. Moreover, it is possible to construct a sharper counterexample, in which the *unique* empirical minimizer $\hat{\mathbf{h}}_S$ is far from having optimal expected value. To do so, we augment $f^{(3)}(\cdot)$ by a small term which ensures its empirical minimizer is unique, and far from the origin. Consider:

$$f^{(4)}(\mathbf{h}; (\mathbf{x}, \alpha)) = f^{(3)}(\mathbf{h}; (\mathbf{x}, \alpha)) + \varepsilon \sum_i 2^{-i} (\mathbf{h}[i] - 1)^2 \tag{4}$$

where $\varepsilon = 0.01$. The objective is still convex and $(1+\varepsilon)$-Lipschitz. Furthermore, since the additional term is strictly convex, we have that $f^{(4)}(\mathbf{h}; \mathbf{z})$ is strictly convex w.r.t. $\mathbf{h}$ and so the empirical minimizer is unique.

Consider the same distribution over z: $\mathbf{x} = 0$ while $\alpha[i]$ are i.i.d. uniform zero or one. The empirical minimizer is the minimizer of $F_S^{(4)}(\mathbf{h})$ subject to the constraints $\|\mathbf{h}\| \leq 1$. Identifying the solution to this constrained optimization problem is tricky, but fortunately not necessary. It is enough to show that the optimum of the *unconstrained* optimization problem $h^*_{\text{UC}} = \arg\min F_S^{(4)}(\mathbf{h})$ (without constraining $\mathbf{h} \in \mathcal{H}$) has norm $\|\mathbf{h}^*_{\text{UC}}\| \geq 1$. Notice that in the unconstrained problem, whenever $\alpha_i[j] = 0$ for all $i = 1..n$, only the second term of $f^{(4)}$ depends on $\mathbf{h}[j]$ and we have $\mathbf{h}^*_{\text{UC}}[j] = 1$. Since this happens a.s. for some coordinate j,

we can conclude that the solution to the constrained optimization problem lies on the boundary of $\mathcal{H}$, that is $\left\|\hat{\mathbf{h}}_S\right\| = 1$. But for such a solution we have

$$F^{(4)}(\hat{\mathbf{h}}_S) \geq \mathbb{E}_\alpha\left[\sqrt{\sum_i \alpha[i]\hat{\mathbf{h}}_S^2[i]}\right] \geq \mathbb{E}_\alpha\left[\sum_i \alpha[i]\hat{\mathbf{h}}_S^2[i]\right] = \sum_i \hat{\mathbf{h}}_S^2[i]\mathbb{E}_\alpha\left[\alpha[i]\right] = \frac{1}{2}\left\|\hat{\mathbf{h}}_S\right\|^2 = \frac{1}{2},$$

while $F^* \leq F(0) = \varepsilon$.

In conclusion, no matter how big the sample size is, the unique empirical minimizer $\hat{\mathbf{h}}_S$ of the stochastic convex optimization problem in Equation (4) is a.s. much worse than the population optimum, $F(\hat{\mathbf{h}}_S) \geq \frac{1}{2} > \varepsilon \geq F^*$, and certainly does not converge to it.

4.2 Learnability via Stability

At this point, we have seen an example in the stochastic convex optimization framework where uniform convergence does not hold, and the ERM algorithm fails. Surprisingly, we will now show that such problems are in fact learnable using an alternative mechanism which has nothing to do with uniform convergence.

Given a stochastic convex optimization problem with an objective function $f(\mathbf{h};\mathbf{z})$, consider a *regularized* version of it: instead of minimizing the expected risk $\mathbb{E}_\mathbf{z}\left[f(\mathbf{h};\mathbf{z})\right]$ over $\mathbf{h} \in \mathcal{H}$, we will try to minimize

$$\mathbb{E}_\mathbf{z}\left[f(\mathbf{h};\mathbf{z}) + \frac{\lambda}{2}\left\|\mathbf{h}\right\|^2\right]$$

for some $\lambda > 0$. Notice that this is simply a stochastic convex optimization problem w.r.t. the objective function $f(\mathbf{h};\mathbf{z}) + \frac{\lambda}{2}\left\|\mathbf{h}\right\|^2$. We will show that this regularized problem *is* learnable using the ERM algorithm (namely, by attempting to minimize $\frac{1}{m}\sum_i f(\mathbf{h};\mathbf{z}_i) + \frac{\lambda}{2}\left\|\mathbf{h}\right\|^2$), by showing that the ERM algorithm is *stable*. By taking $\lambda \to 0$ at an appropriate rate as the sample size increases, we are able to solve the original stochastic problem optimization problem, w.r.t. $f(\mathbf{h};\mathbf{z})$.

The key characteristic of the regularized objective function we need is that it is λ-*strongly convex*. Formally, we say that a real function $g(\cdot)$ over a domain $\mathcal{H}$ in a Hilbert space is λ-strongly convex (where $\lambda \geq 0$), if the function $g(\cdot) - \frac{\lambda}{2}\left\|\cdot\right\|^2$ is convex. In this case, it is easy to verify that if $\mathbf{h}$ minimizes g then

$$\forall \mathbf{h}',\ g(\mathbf{h}') - g(\mathbf{h}) \geq \tfrac{\lambda}{2}\left\|\mathbf{h}' - \mathbf{h}\right\|^2 .$$

When $\lambda = 0$, strong convexity corresponds to standard convexity. In particular, it is immediate from the definition that $f(\mathbf{h};\mathbf{z}) + \frac{\lambda}{2}\left\|\mathbf{h}\right\|^2$ is λ-strongly convex w.r.t. $\mathbf{h}$ (assuming $f(\mathbf{h};\mathbf{z})$ is convex).

The arguments above are formalized in the following two theorems:

Theorem 2 *Consider a stochastic convex optimization problem such that $f(\mathbf{h};\mathbf{z})$ is λ-strongly convex and L-Lipschitz with respect to $\mathbf{h} \in \mathcal{H}$. Let $\mathbf{z}_1,\ldots,\mathbf{z}_m$ be an i.i.d. sample and let $\hat{\mathbf{h}}_S$ be the empirical minimizer. Then, with probability at least $1-\delta$ over the sample we have*

$$F(\hat{\mathbf{h}}_S) - F^* \ \leq\ \frac{4L^2}{\delta\lambda m} .$$

Theorem 3 *Let $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be such that $\mathcal{H}$ is bounded by B and $f(\mathbf{h},\mathbf{z})$ is convex and L-Lipschitz with respect to $\mathbf{h}$. Let $\mathbf{z}_1,\ldots,\mathbf{z}_m$ be an i.i.d. sample and let $\hat{\mathbf{h}}_\lambda$ be the minimizer of*

$$\hat{\mathbf{h}}_\lambda = \min_{\mathbf{h}\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m f(\mathbf{h},\mathbf{z}_i) + \tfrac{\lambda}{2}\left\|\mathbf{h}\right\|^2\right) \tag{5}$$

with $\lambda = \sqrt{\frac{16L^2}{\delta B^2 m}}$. Then, with probability at least $1-\delta$ we have

$$F(\hat{\mathbf{h}}_\lambda) - F^* \ \leq\ 4\sqrt{\frac{L^2B^2}{\delta m}}\left(1 + \frac{8}{\delta m}\right) .$$

Proof [Proof of Theorem 2] To prove the theorem, we use a stability argument. Denote

$$F_S^{(i)}(\mathbf{h}) = \frac{1}{m}\left(f(\mathbf{h},\mathbf{z}_i') + \sum_{j\neq i} f(\mathbf{h},\mathbf{z}_j)\right).$$

the empirical average with $\mathbf{z}_i$ replaced by an independently and identically drawn $\mathbf{z}_i'$, and consider its minimizer:

$$\hat{\mathbf{h}}_S^{(i)} = \arg\min_{\mathbf{h}\in\mathcal{H}} F_S^{(i)}(\mathbf{h}).$$

We first use strong convexity and Lipschitz-continuity to establish that empirical minimization is stable in the following sense:

$$\forall \mathbf{z}\in\mathbb{Z},\quad \left|f(\hat{\mathbf{h}}_S,\mathbf{z}) - f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z})\right| \leq \tfrac{4L^2}{\lambda m}\,. \tag{6}$$

We have that

$$\begin{aligned}
&F_S(\hat{\mathbf{h}}_S^{(i)}) - F_S(\hat{\mathbf{h}}_S)\\
&= \frac{f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_i) - f(\hat{\mathbf{h}}_S,\mathbf{z}_i)}{m} + \frac{\sum_{j\neq i}\left(f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_j) - f(\hat{\mathbf{h}}_S,\mathbf{z}_j)\right)}{m}\\
&= \frac{f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_i) - f(\hat{\mathbf{h}}_S,\mathbf{z}_i)}{m} + \frac{f(\hat{\mathbf{h}}_S,\mathbf{z}_i') - f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_i')}{m}\\
&\quad + \left(F_S^{(i)}(\hat{\mathbf{h}}_S^{(i)}) - F_S^{(i)}(\hat{\mathbf{h}}_S)\right)\\
&\leq \frac{|f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_i) - f(\hat{\mathbf{h}}_S,\mathbf{z}_i)|}{m} + \frac{|f(\hat{\mathbf{h}}_S,\mathbf{z}_i') - f(\hat{\mathbf{h}}_S^{(i)},\mathbf{z}_i')|}{m}\\
&\leq \frac{2L}{m}\left\|\hat{\mathbf{h}}_S^{(i)} - \hat{\mathbf{h}}_S\right\|
\end{aligned} \tag{7}$$

where the first inequality follows from the fact that $\hat{\mathbf{h}}_S^{(i)}$ is the minimizer of $F_S^{(i)}(\mathbf{h})$ and for the second inequality we use Lipschitz continuity. But from strong convexity of $F_S(\mathbf{h})$ and the fact that $\hat{\mathbf{h}}_S$ minimizes $F_S(\mathbf{h})$ we also have that

$$F_S(\hat{\mathbf{h}}_S^{(i)}) \geq F_S(\hat{\mathbf{h}}_S) + \tfrac{\lambda}{2}\left\|\hat{\mathbf{h}}_S^{(i)} - \hat{\mathbf{h}}_S\right\|^2. \tag{8}$$

Combining Equation (8) with Equation (7) we get $\left\|\hat{\mathbf{h}}_S^{(i)} - \hat{\mathbf{h}}_S\right\| \leq 4L/(\lambda m)$ and combining this with Lipschitz continuity of f we obtain that Equation (6) holds. Later on in this paper, we show that a stable ERM is sufficient for learnability. More formally, Equation (6) implies that the ERM is uniform-RO stability (Definition 4) with rate $\varepsilon_{\text{stable}}(m) = 4L^2/(\lambda m)$ and therefore Theorem 8 implies that the ERM is consistent with rate $\leq \varepsilon_{\text{stable}}(m)$, namely

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[F(\hat{\mathbf{h}}_S) - F^*\right] \leq \tfrac{4L^2}{\lambda m}\,.$$

Since the random variable in the expectation is non-negative, the theorem follows by Markov's inequality. ■

We now turn to the proof of Theorem 3.

Proof [Proof of Theorem 3] Let $r(\mathbf{h};\mathbf{z}) = \frac{\lambda}{2}\|\mathbf{h}\|^2 + f(\mathbf{h};\mathbf{z})$ and let $R(\mathbf{h}) = \mathbb{E}_{\mathbf{z}}[r(\mathbf{h},\mathbf{z})]$. Note that $\hat{\mathbf{h}}_\lambda$ is the empirical minimizer for the stochastic optimization problem defined by $r(\mathbf{h};\mathbf{z})$.

We apply Theorem 2 to $r(\mathbf{h};\mathbf{z})$, to this end note that since f is L-Lipschitz and $\forall \mathbf{h}\in\mathcal{H},\ \|\mathbf{h}\|\leq B$ we see that r is in fact $L+\lambda B$-Lipschitz. Applying Theorem 2, we see that

$$\tfrac{\lambda}{2}\left\|\hat{\mathbf{h}}_\lambda\right\|^2 + F(\hat{\mathbf{h}}_\lambda) = R(\hat{\mathbf{h}}_\lambda) \leq \inf_{\mathbf{h}} R(\mathbf{h}) + \frac{4(L+\lambda B)^2}{\delta\lambda m}$$

Now note that $\inf_{\mathbf{h}} R(\mathbf{h}) \leq \inf_{\mathbf{h}} F(\mathbf{h}) + \frac{\lambda}{2}B^2 = F^* + \frac{\lambda}{2}B^2$, and so we get that

$$\begin{aligned} F(\hat{\mathbf{h}}_\lambda) &\leq F^* + \frac{\lambda}{2}B^2 + \frac{4(L+\lambda B)^2}{\delta\lambda m} \\ &\leq F^* + \frac{\lambda}{2}B^2 + \frac{8L^2}{\delta\lambda m} + \frac{8\lambda B^2}{\delta m} \end{aligned}$$

Plugging in the value of λ given in the theorem statement we see that

$$F(\hat{\mathbf{h}}_\lambda) \leq F^* + 4\sqrt{\frac{L^2B^2}{\delta m}} + \frac{32}{\delta m}\sqrt{\frac{L^2B^2}{\delta m}}$$

This gives us the required bound. ■

From the above theorem, we see that regularization is essential for convex stochastic optimization. It is important to note that even for the strongly convex optimization problem in Theorem 2, where the ERM algorithm does work, it is not due to uniform convergence. To see this, consider augmenting the objective function $f^{(3)}(\cdot)$ from Equation (3) with a strongly convex term:

$$f^{(9)}(\mathbf{h};\mathbf{x},\alpha) = f^{(3)}(\mathbf{h};\mathbf{x},\alpha) + \frac{\lambda}{2}\|\mathbf{h}\|^2. \tag{9}$$

The modified objective $f^{(9)}(\cdot;\cdot)$ is λ-strongly convex and $(1+\lambda)$-Lipschitz over $\mathcal{H} = \{\mathbf{h} : \|\mathbf{h}\| \leq 1\}$ and thus satisfies the conditions of Theorem 2. Now, consider the same distribution over $z = (\mathbf{x},\alpha)$ used earlier: $\mathbf{x} = 0$ and α is an i.i.d. sequence of uniform zero/one Bernoulli variables. Recall that almost surely we have a coordinate j that is never "observed", namely such that $\forall_i \alpha_i[j] = 0$. Consider a vector $t\mathbf{e}_j$ of magnitude $0 < t \leq 1$ in the direction of this coordinate. We have that $F_S^{(9)}(t\mathbf{e}_j) = \frac{\lambda}{2}t^2$ (where $F_S^{(9)}(\cdot)$ is the empirical risk w.r.t. $f^{(9)}(\cdot)$) but $F^{(9)}(t\mathbf{e}_j) = \frac{1}{2}t + \frac{\lambda}{2}t^2$. Hence, letting $F^{(9)}(\cdot)$ denote the risk w.r.t. $f^{(9)}(\cdot)$, we have that $F^{(9)}(t\mathbf{e}_j) - F_S^{(9)}(t\mathbf{e}_j) = t/2$. In particular, we can set $t = 1$ and establish $\sup_{\mathbf{h}\in\mathcal{H}}(F^{(9)}(\mathbf{h}) - F_S^{(9)}(\mathbf{h})) \geq \frac{1}{2}$ regardless of the sample size.

We see then that the empirical averages $F_S^{(9)}(\mathbf{h})$ do *not* converge uniformly to their expectations. Moreover, the example above shows that there is no uniform convergence even in a local sense, namely over all hypotheses whose risk is close enough to F^*, or those close enough to the minimizer of $f^{(9)}(\mathbf{h};\mathbf{x},\alpha)$.

Finally, we note that the learning algorithm we have discussed here is mainly for pedagogical reasons. A different generic algorithm for stochastic convex optimization is already known in the literature, by combining Zinkevich's algorithm (Zinkevich, 2003) for online convex optimization, with an online-to-batch conversion (e.g., Cesa-Bianchi et al., 2004). While different than our algorithm, Shalev-Shwartz (2007) showed that Zinkevich's online learning algorithm can be viewed as approximate coordinate ascent optimization of the dual of the regularized problem Equation (5). Thus, this algorithm still uses the same mechanisms of regularization and stability. Also, we note that the algorithm also enjoys bounds which depend only logarithmically on $1/\delta$, while the bounds we have obtained above depend linearly on $1/\delta$. However, we suspect that the dependence on δ in Theorem 2 can be improved to $\log(1/\delta)$. For instance, such bounds has been obtained whenever the objective function is a generalized linear function of h (Sridharan et al., 2008).

4.3 How to Interpret Regularization: Uniform Convergence vs Stability

The technique of regularizing the objective function by adding a "bias" term is old and well known. In particular, adding $\|\mathbf{h}\|^2$ is the so-called Tikhonov Regularization technique, which has been known for more than half a century (see Tikhonov, 1943). However, the role of regularization in our case is very different than in familiar settings such as ℓ_2 regularization in SVMs and ℓ_1 regularization in LASSO. In those settings regularization serves to constrain our domain to a low-complexity domain (e.g., low-norm predictors), where

we rely on uniform convergence. In fact, almost all learning guarantees that we are aware of can be expressed in terms of some sort of uniform convergence.

In our case, constraining the norm of $\mathbf{h}$ does *not* ensure uniform convergence. Consider the example $f^{(3)}(\cdot)$ we have seen earlier. Even over a restricted domain $\mathcal{H}_r = \{\mathbf{h} : \|\mathbf{h}\| \leq r\}$, for arbitrarily small $r > 0$, the empirical averages $F_S(\mathbf{h})$ do *not* uniformly converge to $F(\mathbf{h})$. Furthermore, consider replacing the regularization term $\lambda \|\mathbf{h}\|^2$ with a constraint on the norm of $\|\mathbf{h}\|$, namely, solving the problem

$$\tilde{\mathbf{h}}_r = \arg\min_{\|\mathbf{h}\| \leq r} F_S(\mathbf{h})$$

We cannot solve the stochastic optimization problem by setting r in a distribution-independent way (i.e., without knowing the solution...). To see this, note that when $\mathbf{x} = 0$ a.s. we must have $r \to 0$ to ensure $F(\tilde{\mathbf{h}}_r) \to F^*$. However, if $\mathbf{x} = \mathbf{e}_1$ a.s., we must set $r \to 1$. No constraint will work for all distributions over $\mathbb{Z} = (\mathcal{X}, \alpha)$! This sharply contrasts with traditional uses of regularization, where learning guarantees are typically stated in terms of a constraint on the norm rather than in terms of a parameter such as λ, and adding a regularization term of the form $\frac{\lambda}{2}\|\mathbf{h}\|^2$ is viewed as a proxy for bounding the norm $\|\mathbf{h}\|$.

4.4 Contradiction to Vapnik?

In Section 3.1, we discussed how Vapnik showed that uniform convergence is in fact necessary for learnability with the ERM. At first glance, this might seem confusing in light of the examples presented above, where we have problems learnable with the ERM without uniform convergence whatsoever.

The solution for this apparent paradox is that our examples are not "strictly consistent" in Vapnik's sense. Recall that in order to exclude "trivial" cases, Vapnik defined strict consistency of empirical minimization as (in our notation):

$$\forall c \in \mathbb{R}, \quad \inf_{\mathbf{h}:F(\mathbf{h})\geq c} F_S(\mathbf{h}) \longrightarrow \inf_{\mathbf{h}:F(\mathbf{h})\geq c} F(\mathbf{h}) \ , \tag{10}$$

where the convergence is in probability. This condition indeed ensures that $F(\hat{\mathbf{h}}_S) \xrightarrow{P} F^*$. Vapnik's Key Theorem on Learning Theory (Vapnik, 1998, Theorem 3.1) then states that *strict* consistency of empirical minimization is equivalent to one-sided[2] uniform convergence. In the example presented above, even though Theorem 2 establishes $F^{(9)}(\hat{\mathbf{h}}_S) \xrightarrow{P} F^*$, the consistency isn't "strict" by the definition above. To see this, for any $c > 0$, consider the vector $t\mathbf{e}_j$ (where $\forall_i \alpha_i[j] = 0$) with $t = 2c$. We have $F^{(9)}(t\mathbf{e}_j) = \frac{1}{2}t + \frac{\lambda}{2}t^2 > c$ but $F_S^{(9)}(t\mathbf{e}_j) = \frac{\lambda}{2}t^2 = 2\lambda c^2$. Focusing on $\lambda = \frac{1}{2}$ we get:

$$\inf_{F^{(9)}(\mathbf{h})\geq c} F_S^{(9)}(\mathbf{h}) \leq c^2$$

almost surely for any sample size m, violating the strict consistency requirement Equation (10).

We emphasize that stochastic convex optimization is far from "trivial" in that there is no dominating hypothesis that will always be selected. Although for convenience of analysis we took $\mathbf{x} = 0$, one should think of situations in which $\mathbf{x}$ is stochastic with an unknown distribution. This shows that uniform convergence is a sufficient, but not at all necessary, condition for consistency of empirical minimization in non-trivial settings.

5. Learnability in the General Learning Setting: the role of Stability

In the previous section, we have shown that in the General Learning Setting, it is possible for problems to be learnable without uniform convergence, in sharp contrast to previously considered settings. The key underlying mechanism which allowed us to learn is stability. In this section, we study the connection between learnability and stability in greater depth, and show that stability can in fact *characterize* learnability. Also, we will see how various "common knowledge facts", which we usually take for granted and are based on a

2. "One-sided" meaning requiring only $\sup(F(\mathbf{h}) - F_S(\mathbf{h})) \longrightarrow 0$, rather then $\sup|F(\mathbf{h}) - F_S(\mathbf{h})| \longrightarrow 0$.

"uniform convergence equivalent to learnability" assumption, do not hold in the General Learning Setting, and things can be much more delicate.

We will refer to settings where learnability is equivalent to uniform convergence as "supervised classification" settings. While supervised classification does not encompass all settings where this equivalence holds, most equivalence results refer to it either explicitly or implicitly (by reduction to a classification problem).

5.1 Stability : Definitions

We start by giving the exact definition of the stability notions that we will use. As discussed earlier, there are many possible stability measures, some of which can be used to obtain results of a similar flavor to the ones below. The definition we use seems to be the most convenient for the goal of characterizing learnability in the General Learning Setting. In Appendix A, we provide a few illustrating examples to the subtle differences that can arise from slight variations in the stability measure.

Our two stability notions are based on replacing one of the training sample instances. For a sample S of size m, let $S^{(i)} = \{\mathbf{z}_1, ..., \mathbf{z}_{i-1}, \mathbf{z}'_i, \mathbf{z}_{i+1}, ..., \mathbf{z}_m\}$ be a sample obtained by replacing the i-th observation of S with some different instance $\mathbf{z}'_i$. When not discussed explicitly, the nature of how $\mathbf{z}'_i$ is obtained should be obvious from context.

Definition 4 *A rule* $\mathbf{A}$ *is* **uniform-RO stable**[3] *with rate* $\varepsilon_{\text{stable}}(m)$*, if for all possible* $S^{(i)}$ *and any* $\mathbf{z}' \in \mathcal{Z}$*,*

$$\frac{1}{m}\sum_{i=1}^{m}\left|f(\mathbf{A}(S^{(i)});\mathbf{z}') - f(\mathbf{A}(S);\mathbf{z}')\right| \leq \varepsilon_{\text{stable}}(m).$$

Definition 5 *A rule* $\mathbf{A}$ *is* **average-RO stable** *with rate* $\varepsilon_{\text{stable}}(m)$ *under distributions* $\mathcal{D}$ *if*

$$\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m,(\mathbf{z}'_1,...,\mathbf{z}'_m)\sim\mathcal{D}^m}\left[f(\mathbf{A}(S^{(i)});\mathbf{z}'_i) - f(\mathbf{A}(S);\mathbf{z}'_i)\right]\right| \leq \varepsilon_{\text{stable}}(m).$$

Note that this definition corresponds to assuming that the expected empirical risk of the learning rule converges to the expected risk - see Lemma 11.

We say that a rule is *universally* stable with rate $\varepsilon_{\text{stable}}(m)$, if the stability property holds with rate $\varepsilon_{\text{stable}}(m)$ for all distributions.

Claim 6 *Uniform-RO stability with rate* $\varepsilon_{\text{stable}}(m)$ *implies average-RO stability with rate* $\varepsilon_{\text{stable}}(m)$*.*

5.2 Characterizing Learnability : Main Results

Our overall goal is to characterize learnable problems (namely, problems for which there exists a universally consistent learning rule, as in Equation (2)). That means finding some condition which is both *necessary* and *sufficient* for learnability. In the uniform convergence setting, such a condition is the stability of the ERM (under any of several possible stability measures, including both variants of RO-stability defined above). This is still sufficient for learnability in the General Learning Setting, but far from being necessary, as we have seen in Section 4.

The most important result in this section is a condition which is necessary and sufficient for learnability in the General Learning Setting:

Theorem 7 *A learning problem is learnable if and only if there exists a uniform-RO stable, universally AERM learning rule.*

In particular, if there exists a $\varepsilon_{\text{cons}}(m)$*-universally consistent rule, then there exists a rule that is* $\varepsilon_{\text{stable}}(m)$*-uniform-RO stable and universally* $\varepsilon_{\text{erm}}(m)$*-AERM where:*

$$\varepsilon_{\text{erm}}(m) = 3\varepsilon_{\text{cons}}(m^{1/4}) + \tfrac{8B}{\sqrt{m}} \ ,$$
$$\varepsilon_{\text{stable}}(m) = \tfrac{2B}{\sqrt{m}}.$$

3. RO is short for "replace-one".

In the opposite direction, if a learning rule is $\varepsilon_{\text{stable}}(m)$-uniform-RO stable and universally $\varepsilon_{\text{erm}}(m)$-AERM, then it is universally consistent with rate

$$\varepsilon_{\text{cons}}(m) \leq \varepsilon_{\text{stable}}(m) + \varepsilon_{\text{erm}}(m)$$

Thus, while we have seen in Section 4 that the ERM rule might fail for learning problems which are in fact learnable, there is always an AERM rule which will work. In other words, when designing learning rules, we might need to look beyond empirical risk minimization, but not beyond AERM learning rules. On the downside, we must choose our AERM carefully, since not any AERM will work. This contrasts with supervised classification, where any AERM will work if the problem is learnable at all.

How do we go about proving this assertion? The easier part is showing sufficiency. Namely, that a stable AERM must be consistent (and generalizing). In fact, this holds both separately for any particular distribution $\mathcal{D}$s, and uniformly over all distributions:

Theorem 8 *If a rule is an AERM with rate $\varepsilon_{\text{erm}}(m)$ and average-RO stable (or uniform-RO stable) with rate $\varepsilon_{\text{stable}}(m)$ under $\mathcal{D}$, then it is consistent and generalizes under $\mathcal{D}$ with rates*

$$\varepsilon_{\text{cons}}(m) \leq \varepsilon_{\text{stable}}(m) + \varepsilon_{\text{erm}}(m)$$
$$\varepsilon_{\text{gen}}(m) \leq \varepsilon_{\text{stable}}(m) + 2\varepsilon_{\text{erm}}(m) + \tfrac{2B}{\sqrt{m}}$$

The second part of Theorem 7 follows as a direct corollary. We note that close variants of Theorem 8 has already appeared in previous literature (e.g., Mukherjee et al., 2006 and Rakhlin et al., 2005).

The harder part is showing that a uniform-RO stable AERM is *necessary* for learnability. This is done in several steps.

First, we show that consistent AERMs have to be average-RO stable:

Theorem 9 *For an AERM, the following are equivalent:*
- *Universal average-RO stability.*
- *Universal consistency.*
- *Universal generalization.*

The exact conversion rate of Theorem 9 is specified in the corresponding proof (Section 5.3), and are all polynomial. In particular, an $\varepsilon_{\text{cons}}$-universal consistent ε_{erm}-AERM is average-RO stable with rate

$$\varepsilon_{\text{stable}}(m) \leq \varepsilon_{\text{erm}}(m) + 3\varepsilon_{\text{cons}}(m^{1/4}) + \tfrac{4B}{\sqrt{m}}.$$

Next, we show that if we seek universally consistent and generalizing learning rules, then we must consider only AERMs:

Theorem 10 *If a rule $\mathbf{A}$ is universally consistent with rate $\varepsilon_{\text{cons}}(m)$ and generalizing with rate $\varepsilon_{\text{gen}}(m)$, then it is universally an AERM with rate*

$$\varepsilon_{\text{erm}}(m) \leq \varepsilon_{\text{gen}}(m) + 3\varepsilon_{\text{cons}}(m^{1/4}) + \frac{4B}{\sqrt{m}}$$

Now, recall that learnability is defined as the existence of some universally consistent learning rule. Such a rule might not be generalizing, stable or even an AERM (see example 2 below). However, it turns out that if a universally consistent learning rule exist, then there is *another* learning rule for the same problem, which is generalizing (Lemma 20). Thus, by Theorems 9-10, this rule must also be average-RO stable AERM. In fact, by another application of Lemma 20, such an AERM must also be uniform-RO stable, leading to Theorem 7.

5.3 Detailed Results and Proofs

We first establish that for AERMs, average-RO stability and generalization are equivalent.

5.3.1 EQUIVALENCE OF STABILITY AND GENERALIZATION

It will be convenient to work with a weaker version of generalization as an intermediate step: We say a rule **A on-average generalizes** with rate $\varepsilon_{\text{oag}}(m)$ under distribution $\mathcal{D}$ if for all m,

$$|\mathbb{E}_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))]| \leq \varepsilon_{\text{oag}}(m). \tag{11}$$

It is straightforward to see that generalization implies on-average generalization with the same rate. We show that for AERMs, the converse is also true, and also that on-average generalization is equivalent to average-RO stability. This establishes the equivalence between generalization and average-RO stability (for AERMs).

Lemma 11 (on-average generalization $\Leftrightarrow$ average-RO stability) *If* $\mathbf{A}$ *is on-average generalizing with rate* $\varepsilon_{\text{oag}}(m)$ *then it is average-RO stable with rate* $\varepsilon_{\text{oag}}(m)$*. If* $\mathbf{A}$ *is average-RO stable with rate* $\varepsilon_{\text{stable}}(m)$ *then it is on-average generalizing with rate* $\varepsilon_{\text{stable}}(m)$*.*

Proof For any i, $\mathbf{z}_i$ and $\mathbf{z}'_i$ are both drawn i.i.d. from $\mathcal{D}$, we have that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[f(\mathbf{A}(S);\mathbf{z}_i)] = \mathbb{E}_{S\sim\mathcal{D}^m,\mathbf{z}'_i\sim\mathcal{D}}\left[f(\mathbf{A}(S^{(i)});\mathbf{z}'_i)\right].$$

Hence,

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[F_S(\mathbf{A}(S))] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i=1}^m f(\mathbf{A}(S);\mathbf{z}_i)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{S\sim\mathcal{D}^m}[f(\mathbf{A}(S);\mathbf{z}_i)] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{S\sim\mathcal{D}^m,\mathbf{z}'_i\sim\mathcal{D}}\left[f(\mathbf{A}(S^{(i)});\mathbf{z}'_i)\right]
\end{aligned}$$

Also note that $F(\mathbf{A}(S)) = \mathbb{E}_{\mathbf{z}'_i\sim\mathcal{D}}[f(\mathbf{A}(S);\mathbf{z}'_i)] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathbf{z}'_i\sim\mathcal{D}}[f(\mathbf{A}(S);\mathbf{z}'_i)]$. Hence we can conclude that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{S\sim\mathcal{D}^m,(\mathbf{z}'_1,\ldots,\mathbf{z}'_m)\sim\mathcal{D}^m}\left[f(\mathbf{A}(S);\mathbf{z}'_i) - f(\mathbf{A}(S^{(i)});\mathbf{z}'_i)\right]$$

Hence we have the required result. ■

For the next result, we will need the following two short utility lemmas.

Utility Lemma 12 *For i.i.d.* X_i*,* $|X_i| \leq B$ *and* $X = \frac{1}{m}\sum_{i=1}^m X_i$ *we have* $\mathbb{E}[|X - \mathbb{E}[X]|] \leq B/\sqrt{m}$*.*

Proof $\mathbb{E}[|X - \mathbb{E}[X]|] \leq \sqrt{\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right]} \leq \sqrt{\text{Var}[X]} = \sqrt{\text{Var}[X_i]/m} \leq B/\sqrt{m}$. ■

Utility Lemma 13 *Let* X, Y *be random variables s.t.* $X \leq Y$ *almost surely. Then* $\mathbb{E}[|X|] \leq |\mathbb{E}[X]| + 2\mathbb{E}[|Y|]$*.*

Proof

$$\mathbb{E}[|X|] = \mathbb{E}[|(Y-X) - Y|] \leq \mathbb{E}[Y - X] + \mathbb{E}[|Y|] \leq |\mathbb{E}[X]| + 2|\mathbb{E}[Y]|.$$

■

Lemma 14 (AERM + on-average generalization ⇒ generalization) *If* $\mathbf{A}$ *is an AERM with rate* $\varepsilon_{\text{erm}}(m)$ *and on-average generalizes with rate* $\varepsilon_{\text{oag}}(m)$ *under* $\mathcal{D}$*, then* $\mathbf{A}$ *generalizes with rate* $\varepsilon_{\text{oag}}(m)+2\varepsilon_{\text{erm}}(m)+\frac{2B}{\sqrt{m}}$ *under* $\mathcal{D}$*.*

Proof Recall that $F^* = \inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})$. For an arbitrarily small $\nu > 0$, let $\mathbf{h}_\nu$ be a fixed hypothesis such that $F(\mathbf{h}_\nu) \leq F^* + \nu$. Using respective optimalities of $\hat{\mathbf{h}}_S$ and F^* we can bound:

$$\begin{aligned}
&F_S(\mathbf{A}(S)) - F(\mathbf{A}(S)) \\
&= F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S) + F_S(\hat{\mathbf{h}}_S) - F_S(\mathbf{h}_\nu) + F_S(\mathbf{h}_\nu) - F(\mathbf{h}_\nu) + F(\mathbf{h}_\nu) - F(\mathbf{A}(S)) \\
&\leq F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S) + F_S(\mathbf{h}_\nu) - F(\mathbf{h}_\nu) + \nu = Y_\nu
\end{aligned}$$

Where the final equality defines a new random variable Y_ν. By Lemma 12 and the AERM guarantee we have $\mathbb{E}[|Y_\nu|] \leq \varepsilon_{\text{erm}}(m) + B/\sqrt{m} + \nu$. From Lemma 13 we can conclude that

$$\mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))\right|\right] \leq \left|\mathbb{E}\left[F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))\right]\right| + 2\mathbb{E}\left[|Y_\nu|\right] \leq \varepsilon_{\text{oag}}(m) + 2\varepsilon_{\text{erm}}(m) + \tfrac{2B}{\sqrt{m}} + \nu.$$

Notice that the l.h.s. is a fixed quantity which does not depend on ν. Therefore, we can take ν in the r.h.s. to zero, and the result follows. ■

Combining Lemma 11 and Lemma 14, we have now **established the stability↔generalization parts of Theorem 8 and Theorem 9** (in fact, even a slightly stronger converse than in Theorem 9, as it does not require universality).

5.3.2 A SUFFICIENT CONDITION FOR CONSISTENCY

It is fairly straightforward to see that generalization (or even on-average generalization) of an AERM implies its consistency:

Lemma 15 (AERM+generalization⇒consistency) *If* $\mathbf{A}$ *is AERM with rate* $\varepsilon_{\text{erm}}(m)$ *and it on-average generalizes with rate* $\varepsilon_{\text{oag}}(m)$ *under* $\mathcal{D}$ *then it is consistent with rate* $\varepsilon_{\text{oag}}(m)+\varepsilon_{\text{erm}}(m)$ *under* $\mathcal{D}$*.*

Proof For any $\nu > 0$, let $\mathbf{h}_\nu$ be a hypothesis such that $F(\mathbf{h}_\nu) \leq F^* + \nu$. We have

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{A}(S)) - F^*\right] = \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{h}_\nu) + \nu\right] \\
&= \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right] + \mathbb{E}\left[F_S(\mathbf{A}(S)) - F_S(\mathbf{h}_\nu)\right] + \nu \\
&\leq \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right] + \mathbb{E}\left[F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right] + \nu \\
&\leq \varepsilon_{\text{oag}}(m) + \varepsilon_{\text{erm}}(m) + \nu.
\end{aligned}$$

Since this upper bound holds for any ν, we can take ν to zero, and the result follows. ■

Combined with the results of Lemma 11, this completes the **proof of Theorem 8** and the **stability → consistency and generalization → consistency parts of Theorem 9**.

5.3.3 CONVERSE DIRECTION

Lemma 11 already provides a converse result, establishing that stability is necessary for generalization. However, as it will turn out, in order to establish that stability is also necessary for *universal consistency*, we must prove that universal consistency of an AERM implies *universal* generalization. The assumption of *universal* consistency for the AERM is crucial here: mere consistency of an AERM with respect to a specific distribution does *not* imply generalization nor stability with respect to that distribution. The following example briefly illustrates this point.

Example 1 *There exists a learning problem and a distribution on the instance space, such that the ERM (or any AERM) is consistent with rate* $\varepsilon_{\text{cons}}(m) = 0$*, but does not generalize and is not average-RO stable (namely,* $\varepsilon_{\text{gen}}(m), \varepsilon_{\text{stable}}(m) = \Omega(1)$*).*

Proof Let the instance space be $[0,1]$, the hypothesis space consist of all finite subsets of $[0,1]$, and define the objective function as $f(\mathbf{h},z) = \mathbb{1}_{\{z \notin \mathbf{h}\}}$). Consider any continuous distribution on the instance space. Since the underlying distribution $\mathcal{D}$ is continuous, we have $F(\mathbf{h}) = 1$ for any hypothesis h. Therefore, any learning rule (including any AERM) will be consistent with $F(\mathbf{A}(S)) = 1$. On the other hand, the ERM here always achieves $F_S(\hat{\mathbf{h}}_S) = 0$, so any AERM cannot generalize, or even on-average-generalize (by Lemma 14), hence cannot be average-RO stable (by Lemma 11). ■

The main tool we use to prove our desired converse result is the following lemma. It is here that we crucially use the universal consistency assumption (i.e., consistency with respect to *any* distribution). Intuitively, it states that if a problem is learnable at all, then although the ERM rule might fail, its empirical risk is a consistent estimator of the minimal achievable risk.

Lemma 16 (Main Converse Lemma) *If a problem is learnable, namely there exists a universally consistent rule* $\mathbf{A}$ *with rate* $\varepsilon_{\text{cons}}(m)$*, then under any distribution,*

$$\mathbb{E}\left[\left|F_S(\hat{\mathbf{h}}_S) - F^*\right|\right] \leq \varepsilon_{\text{emp}}(m) \qquad \textit{where} \tag{12}$$

$$\varepsilon_{\text{emp}}(m) = 2\varepsilon_{\text{cons}}(m') + \tfrac{2B}{\sqrt{m}} + \tfrac{2Bm'^2}{m}$$

for any m' *such that* $2 \leq m' \leq m/2$*.*

Proof Let $I = \{I_1, \ldots, I_{m'}\}$ be a random sample of m' indexes in the range $1..m$ where each I_i is independently uniformly distributed, and I is independent of S. Let $S' = \{z_{I_i}\}_{i=1}^{m'}$, that is, a sample of size m' drawn from the uniform distribution over samples in S (with replacements). We first bound the probability that I has no repeated indexes ("duplicates"):

$$\mathbb{P}\left[I \text{ has duplicates}\right] \leq \frac{\sum_{i=1}^{m'}(i-1)}{m} \leq \frac{m'^2}{2m} \tag{13}$$

Conditioned on not having duplicates in I, the sample S' is actually distributed according to $\mathcal{D}^{m'}$, that is, can be viewed as a sample from the original distribution. We therefore have by universal consistency:

$$\mathbb{E}\left[\left|F(\mathbf{A}(S')) - F^*\right| \,\middle|\, \text{no dups}\right] \leq \varepsilon_{\text{cons}}(m') \tag{14}$$

But viewed as a sample drawn from the uniform distribution over instances in S, we also have:

$$\mathbb{E}_{S'}\left[\left|F_S(\mathbf{A}(S')) - F_S(\hat{\mathbf{h}}_S)\right|\right] \leq \varepsilon_{\text{cons}}(m') \tag{15}$$

Conditioned on having no duplications in I, the set of those samples in S not chosen by I (i.e., $S \setminus S'$) is independent of S', and $|S \setminus S'| = m - m'$, and so by Lemma 12:

$$\mathbb{E}_S\left[\left|F(\mathbf{A}(S')) - F_{S\setminus S'}(\mathbf{A}(S'))\right|\right] \leq \frac{B}{\sqrt{m-m'}} \tag{16}$$

Finally, if there are no duplicates, then for any hypothesis, and in particular for $\mathbf{A}(S')$ we have:

$$\left|F_S(\mathbf{A}(S')) - F_{S\setminus S'}(\mathbf{A}(S'))\right| \leq \frac{2Bm'}{m} \tag{17}$$

Combining Equation (14),Equation (15),Equation (16) and Equation (17), accounting for a maximal discrepancy of B when we do have duplicates, and assuming $2 \leq m' \leq m/2$, we get the desired bound. ■

Equipped with Lemma 16, we are now ready to show that universal consistency of an AERM implies universal generalization and that any universally consistent and generalizing rule must be an AERM. What we show is actually a bit stronger: that if a problem is learnable, and so Lemma 16 holds, then for any distribution $\mathcal{D}$ separately, consistency of an AERM under $\mathcal{D}$ implies generalization under $\mathcal{D}$ and also any consistent and generalizing rule under $\mathcal{D}$ must be an AERM.

Lemma 17 (learnable+AERM+consistent⇒generalizing) *If Equation (12) in Lemma 16 holds with rate $\varepsilon_{\text{emp}}(m)$, and $\mathbf{A}$ is an ε_{erm}-AERM and $\varepsilon_{\text{cons}}$-consistent under $\mathcal{D}$, then it is generalizing under $\mathcal{D}$ with rate $\varepsilon_{\text{emp}}(m)+\varepsilon_{\text{erm}}(m)+\varepsilon_{\text{cons}}(m)$.*

Proof

$$\begin{aligned}\mathbb{E}\left[\left|F_S(\mathbf{A}(S))-F(\mathbf{A}(S))\right|\right] &\leq \mathbb{E}\left[\left|F_S(\mathbf{A}(S))-F_S(\hat{\mathbf{h}}_S)\right|\right]+\mathbb{E}\left[\left|F^*-F(\mathbf{A}(S))\right|\right]+\mathbb{E}\left[\left|F_S(\hat{\mathbf{h}}_S)-F^*\right|\right] \\ &\leq \varepsilon_{\text{erm}}(m)+\varepsilon_{\text{cons}}(m)+\varepsilon_{\text{emp}}(m)\ .\end{aligned}$$

■

Lemma 18 (learnable+consistent+generalizing⇒AERM) *If Equation (12) in Lemma 16 holds with rate $\varepsilon_{\text{emp}}(m)$, and $\mathbf{A}$ is $\varepsilon_{\text{cons}}$-consistent and ε_{gen}-generalizing under $\mathcal{D}$, then it is AERM under $\mathcal{D}$ with rate $\varepsilon_{\text{emp}}(m)+\varepsilon_{\text{gen}}(m)+\varepsilon_{\text{cons}}(m)$.*

Proof

$$\begin{aligned}\mathbb{E}\left[\left|F_S(\mathbf{A}(S))-F_S(\hat{\mathbf{h}}_S)\right|\right] &\leq \mathbb{E}\left[\left|F_S(\mathbf{A}(S))-F(\mathbf{A}(S))\right|\right]+\mathbb{E}\left[\left|F(\mathbf{A}(S))-F^*\right|\right]+\mathbb{E}\left[\left|F^*-F_S(\hat{\mathbf{h}}_S)\right|\right] \\ &\leq \varepsilon_{\text{gen}}(m)+\varepsilon_{\text{cons}}(m)+\varepsilon_{\text{emp}}(m)\ .\end{aligned}$$

■

Lemma 17 establishes that universal consistency of an AERM implies universal generalization, and thus **completes the proof of Theorem 9**. Lemma 18 **establishes Theorem 10**. To get the rates in Section 5.2, we use $m' = m^{1/4}$ in Lemma 16.

Lemma 15, Lemma 17 and Lemma 18 together establish an interesting relationship:

Corollary 19 *For a (universally) learnable problem, for any distribution $\mathcal{D}$ and learning rule $\mathbf{A}$, any two of the following imply the third :*

- $\mathbf{A}$ *is an AERM under $\mathcal{D}$.*
- $\mathbf{A}$ *is consistent under $\mathcal{D}$.*
- $\mathbf{A}$ *generalizes under $\mathcal{D}$.*

Note, however, that any one property by itself is possible, even universally:

- In Section 4.1, we have discussed an example where the ERM learning rule is neither consistent nor generalizing, despite the problem being learnable.
- In the next subsection (Example 2) we demonstrate a universally consistent learning rule which is neither generalizing nor an AERM.
- A rule returning a fixed hypothesis always generalizes, but of course need not be consistent nor an AERM.

In contrast, for learnable supervised classification problems, it is not possible for a learning rule to be just universally consistent, without being an AERM and without generalization. Nor is it possible for a learning rule to be a universal AERM for a learnable problem, without being generalizing and consistent.

Corollary 19 can also provide a *certificate* of non-learnability. In other words, for the problem in Example 1 we show a specific distribution for which there is a consistent AERM that does not generalize. We can conclude that there is *no* universally consistent learning rule for the problem, otherwise the corollary is violated.

5.3.4 EXISTENCE OF A STABLE RULE

Theorem 9 and Theorem 10, which we just completed proving, already establish that for AERMs, universal consistency is equivalent to universal average-RO stability. Existence of a universally average-RO stable AERM is thus sufficient for learnability. In order to prove that it is also necessary, it is enough to show that existence of a universally consistent learning rule implies existence of a universally consistent AERM. This AERM must then be average-RO stable by Theorem 9.

We actually show how to transform a consistent rule to a consistent and generalizing rule (Lemma 20 below). If this rule is universally consistent, then by Lemma 18 we can then conclude it must be an AERM, and by Lemma 11 it must be average-RO stable.

Lemma 20 *For any rule* $\mathbf{A}$ *there exists a rule* $\mathbf{A}'$*, such that:*

- $\mathbf{A}'$ *universally generalizes with rate* $\frac{3B}{\sqrt{m}}$.
- *For any* $\mathcal{D}$*, if* $\mathbf{A}$ *is* $\varepsilon_{\text{cons}}$*-consistent under* $\mathcal{D}$ *then* $\mathbf{A}'$ *is* $\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$ *consistent under* $\mathcal{D}$*.*
- $\mathbf{A}'$ *is uniformly-RO-stable with rate* $\frac{2B}{\sqrt{m}}$.

Proof For a sample S of size m, let S' be a sub-sample consisting of some $\lfloor\sqrt{m}\rfloor$ observation in S. To simplify the presentation, assume that $\lfloor\sqrt{m}\rfloor$ is an integer. Define $\mathbf{A}'(S) = \mathbf{A}(S')$. That is, $\mathbf{A}'$ applies A to only $\sqrt{m}$ of the observation in S.

$\mathbf{A}'$ *generalizes:* We can decompose:

$$F_S(\mathbf{A}(S')) - F(\mathbf{A}(S')) = \tfrac{1}{\sqrt{m}}(F_{S'}(\mathbf{A}(S')) - F(\mathbf{A}(S'))) + (1-\tfrac{1}{\sqrt{m}})(F_{S\setminus S'}(\mathbf{A}(S')) - F(\mathbf{A}(S')))$$

The first term can be bounded by $2B/\sqrt{m}$. As for the second term, $S \setminus S'$ is statistically independent of S' and so we can use Lemma 12 to bound its expected magnitude to obtain:

$$\mathbb{E}\left[\left|F_S(\mathbf{A}(S')) - F(\mathbf{A}(S'))\right|\right] \leq \tfrac{2B}{\sqrt{m}} + (1-\tfrac{1}{\sqrt{m}})\tfrac{B}{\sqrt{m-\sqrt{m}}} \leq \tfrac{3B}{\sqrt{m}}$$

$\mathbf{A}'$ *is consistent:* If $\mathbf{A}$ is consistent, then:

$$\mathbb{E}\left[F(\mathbf{A}'(S)) - \inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})\right] = \mathbb{E}\left[F(\mathbf{A}(S')) - \inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})\right] \leq \varepsilon_{\text{cons}}(\sqrt{m})$$

$\mathbf{A}'$ *is uniformly-RO-stable:* Since $\mathbf{A}'$ only uses the first $\sqrt{m}$ samples of S, for any $i > \sqrt{m}$ we have $\mathbf{A}'(S^{(i)}) = \mathbf{A}'(S)$ and so:

$$\frac{1}{m}\sum_{i=1}^{m}\left|f(\mathbf{A}'(S^{(i)});\mathbf{z}') - f(\mathbf{A}'(S);\mathbf{z}')\right| = \frac{1}{m}\sum_{i=1}^{\sqrt{m}}\left|f(\mathbf{A}'(S^{(i)});\mathbf{z}') - f(\mathbf{A}'(S);\mathbf{z}')\right| \leq \frac{2B}{\sqrt{m}}$$

■

Proof of Converse in Theorem 7 If there exists a universally consistent rule with rate $\varepsilon_{\text{cons}}(m)$, by Lemma 20 there exists $\mathbf{A}'$ which is $\varepsilon_{\text{cons}}(\sqrt{m})$- universally consistent, $\frac{2B}{\sqrt{m}}$-generalizing and $\frac{2B}{\sqrt{m}}$-uniformly-RO-stable.

Further by Lemma 18 and Lemma 16 (with $m' = m^{1/4}$), we can conclude that A' is ε_{erm}-universally AERM where,

$$\varepsilon_{\text{erm}}(m) \leq 3\varepsilon_{\text{cons}}(m^{1/4}) + \frac{8B}{\sqrt{m}}.$$

Hence we get the specified rate for the converse direction. To see that if there exists a rule that is a universal AERM and stable it is consistent, we simply use Lemma 15.

As a final note, the following example shows that while learnability is equivalent to the existence of stable and consistent AERM's (Theorem 7 and Theorem 9), there might still exist other learning rules, which are neither stable, nor generalize, nor AERM's. In this sense, our results characterize learnability, but do not characterize all learning rules which "work".

Example 2 *There exists a learning problem with a universally consistent learning rule, which is not average-RO stable, generalizing nor an AERM.*

Proof Let the instance space be $[0,1]$. Let the hypothesis space consist of all finite subsets of $[0,1]$, and the objective function be the indicator function $f(\mathbf{h},z) = \mathbb{1}_{[z\in\mathbf{h}]}$. Consider the following learning rule: given a sample $S \subseteq [0,1]$, the learning rule checks if there are any two identical instances in the sample. If so, the learning rule returns the empty set $\emptyset$. Otherwise, it returns the sample.

Consider any continuous distribution on $[0,1]$. In that case, the probability of having two identical instances is 0. Therefore, the learning rule always returns a countable non-empty set $\mathbf{A}(S)$, with $F_S(\mathbf{A}(S)) = 1$, while $F_S(\emptyset) = 0$ (so it is not an AERM) and $F(\mathbf{A}(S)) = 0$ (so it does not generalize). Also, $f(\mathbf{A}(S), z_i) = 1$ while $f(\mathbf{A}(S^{(i)}), z_i) = 0$ with probability 1, so it is not average-RO stable either.

However, the learning rule is universally consistent. If the underlying distribution is continuous on $[0,1]$, then the returned hypothesis is S, which is countable hence , $F(S) = 0 = \inf_{\mathbf{h}} F(\mathbf{h})$. For discrete distributions, let M_1 denote the proportion of instances in the sample which appear exactly once, and let M_0 be the probability mass of instances which did not appear in the sample. Using (McAllester and Schapire, 2000, Theorem 3), we have that for any δ, it holds with probability at least $1-\delta$ over a sample of size m that

$$|M_0 - M_1| \leq O\left(\frac{\log(m/\delta)}{\sqrt{m}}\right),$$

uniformly for any discrete distribution. If this occurs, then either $M_1 < 1$, or $M_0 \geq 1 - O(\log(m/\delta)/\sqrt{m})$. But in the first event, we get duplicate instances in the sample, so the returned hypothesis is the optimal $\emptyset$, and in the second case, the returned hypothesis is the sample, which has a total probability mass of at most $O(\log(m/\delta)/\sqrt{m})$, and therefore $F(\mathbf{A}(S)) \leq O(\log(m/\delta)/\sqrt{m})$. As a result, regardless of the underlying distribution, with probability of at least $1-\delta$ over the sample,

$$F(\mathbf{A}(S)) \leq O\left(\frac{\log(m/\delta)}{\sqrt{m}}\right).$$

Since the r.h.s. converges to 0 with m for any δ, it is easy to see that the learning rule is universally consistent. ■

6. Randomization, Convexification, and a Generic Learning Algorithm

The strongest result we were able to obtain for characterizing learnability so far is Theorem 7, which stated that a problem is learnable if and only if there exists a universally uniform-RO stable AERM. In fact, this result was obtained under the assumption that the learning rule $\mathbf{A}$ is deterministic: given a fixed sample S, $\mathbf{A}$ returns a single specific hypothesis $\mathbf{h}$. However, we might relax this assumption and also consider *randomized* learning rules: given any fixed S, $\mathbf{A}(S)$ returns a distribution over the hypothesis class $\mathcal{H}$.

With this relaxation, we will see that we can obtain a stronger version of Theorem 7, and even provide a generic learning algorithm (at least for computationally unbounded learners) which successfully learns any learnable problem.

6.1 Stronger Results with Randomized Learning Rules

For simplicity, we will override the notations $f(\mathbf{A}(S),\mathbf{z})$, $F(\mathbf{A}(S))$ and $F_S(\mathbf{A}(S))$ to mean $\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[f(\mathbf{h},\mathbf{z})]$, $\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[F(\mathbf{h})]$ and $\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[F_S(\mathbf{h})]$. In other words, $\mathbf{A}$ returns a distribution over $\mathcal{H}$ and $f(\mathbf{A}(S),\mathbf{z})$ for some fixed $S,\mathbf{z}$ is the expected loss of a random hypothesis picked according to that distribution, with respect to $\mathbf{z}$. Similarly, $F(\mathbf{A}(S))$ for some fixed S is the expected generalization error, and $F_S(\mathbf{A}(S))$ is the expected empirical risk on the fixed sample S. With this slight abuse of notation, all our previous definitions hold. For instance, we still define a learning rule $\mathbf{A}$ to be consistent with rate $\varepsilon_{\text{cons}}(m)$ if $\mathbb{E}_{S\sim\mathcal{D}^m}[F(\mathbf{A}(S))-F^*]\leq\varepsilon_{\text{cons}}(m)$, only now we actually mean

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[F(\mathbf{h})-F^*]\right]\leq\varepsilon_{\text{cons}}(m).$$

The definitions for AERM, generalization etc. also hold with this subtle change in meaning.

An alternative way to view randomization is as a method to *linearize* the learning problem. In other words, randomization implicitly replaces the arbitrary hypothesis class $\mathcal{H}$ by the space of probability distributions over $\mathcal{H}$,

$$\mathcal{M}=\left\{\alpha:\mathcal{H}\to[0,1]\ \text{ s.t. }\int\alpha[\mathbf{h}]=1\right\},$$

and replaces the arbitrary function $f(\mathbf{h};\mathbf{z})$ by a *linear* function in its first argument

$$f(\alpha;\mathbf{z})=\mathbb{E}_{\mathbf{h}\sim\alpha}[f(\mathbf{h},\mathbf{z})]=\int f(\mathbf{h};\mathbf{z})\alpha[\mathbf{h}]\ .$$

Linearity of the loss and convexity of $\mathcal{M}$ are the key mechanism which allows us to obtain our stronger results. Moreover, if the learning problem is already convex (i.e., f is convex and $\mathcal{H}$ is covex), we can achieve the same results using a deterministic learning rule, as the following claim demonstrates:

Claim 21 *Assume that the hypothesis class $\mathcal{H}$ is convex subset of a vector space, such that $\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[\mathbf{h}]$ is a well-defined element of $\mathcal{H}$ for any S. Moreover, assume that $f(\mathbf{h};\mathbf{z})$ is convex in $\mathbf{h}$. Then from any (possibly randomized) learning rule $\mathbf{A}$, it is possible to construct a deterministic learning rule $\mathbf{A}'$, such that $f(\mathbf{A}'(S),\mathbf{z})\leq f(\mathbf{A}(S),\mathbf{z})$ for any $S,\mathbf{z}$. As a result, it also holds that $F_S(\mathbf{A}'(S))\leq F_S(\mathbf{A}(S))$ and $F(\mathbf{A}'(S))\leq F(\mathbf{A}(S))$.*

Proof Given a sample S, define $\mathbf{A}'(S;\mathbf{z})$ as the single hypothesis $\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[\mathbf{h}]$. The proof of the theorem is immediate by Jensen's inequality: since $f()$ is convex in its first argument,

$$f(\mathbf{A}'(S);\mathbf{z})=f(\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[\mathbf{h}],\mathbf{z})\leq\mathbb{E}_{\mathbf{h}\sim\mathbf{A}(S)}[f(\mathbf{h},\mathbf{z})],$$

where the r.h.s. is in fact $f(\mathbf{A}(S),\mathbf{z})$ by the abuse of notation we have defined previously. ■

Although linearization is the real mechanism at play here, we find it more convenient to display our results and proofs in the language of randomized learning rules.

Allowing randomization allows us to obtain results with respect to the following very strong notion of stability:[4]

Definition 22 *A rule $\mathbf{A}$ is* **strongly-uniform-RO stable** *with rate $\varepsilon_{\text{stable}}(m)$ if for all samples S of m points, for all i, and any $\mathbf{z}',\mathbf{z}'_i\in\mathcal{Z}$, it holds that*

$$\left|f(\mathbf{A}(S^{(i)});\mathbf{z}')-f(\mathbf{A}(S);\mathbf{z}')\right|\leq\varepsilon_{\text{stable}}(m).$$

The strengthening of Theorem 7 that we will prove here is the following:

4. This definition of stability is very similar to the so-called "uniform stability", discussed in Bousquet and Elisseeff (2002), although Bousquet and Elisseeff (2002) consider deterministic learning rules. See Appendix A for more details.

Theorem 23 *A learning problem is learnable if and only if there exists a (possibly randomized) learning rule which is an always AERM and strongly-uniform-RO stable.*

Compared to Theorem 7, we have replaced universal AERM by the stronger notion of an always AERM, and uniform-RO stability by strongly-uniform-RO stability. This makes the result strong enough to formulate a generic learning algorithm, as we will see later on.

The theorem is an immediate consequence of Theorem 7 and the following lemma:

Lemma 24 *For any deterministic learning rule* $\mathbf{A}$*, there exists a randomized learning rule* $\mathbf{A}'$ *such that:*

- *For any* $\mathcal{D}$*, if* $\mathbf{A}$ *is* $\varepsilon_{\text{cons}}$*-consistent under* $\mathcal{D}$ *then* $\mathbf{A}'$ *is* $\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$ *consistent under* $\mathcal{D}$*.*
- $\mathbf{A}'$ *universally generalizes with rate* $4B/\sqrt{m}$*.*
- *If* $\mathbf{A}$ *is uniform-RO stable with rate* $\varepsilon_{\text{stable}}(m)$*, then* $\mathbf{A}'$ *is strongly-uniform-RO stable with rate* $\varepsilon_{\text{stable}}(\lfloor\sqrt{m}\rfloor)$*.*
- *If* $\mathbf{A}$ *is universally* $\varepsilon_{\text{cons}}$*-consistent, then* $\mathbf{A}'$ *is an always AERM with rate* $2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$*.*

Moreover, $\mathbf{A}'$ *is a symmetric learning rule (it does not depend on the order of elements in the sample on which it is applied).*

Proof Consider the learning rule A' which given a sample S, returns a uniform distribution over $A(S')$, where S' ranges over all subsets of S of size $\lfloor\sqrt{m}\rfloor$.

The fact that $\mathbf{A}'$ is symmetric is trivial. We now prove the other assertions in the lemma.

$\mathbf{A}'$ *is consistent:* First note that $F(\mathbf{A}'(S)) = \mathbb{E}_{S'}\left[F(\mathbf{A}(S'))\right]$, and so:

$$\mathbb{E}_S\left[\left|F(\mathbf{A}'(S)) - F^*\right|\right] \;\le\; \mathbb{E}_{S,S'}\left[\left|F(\mathbf{A}(S')) - F^*\right|\right] \;=\; \mathbb{E}_{[S']}\left[\mathbb{E}_{S|[S']}\left[\left|F(\mathbf{A}(S')) - F^*\right|\right]\right]$$

where $[S']$ designates a choice of indices for S'. This decomposition of the random choice of S' (e.g., first deciding on the indices and only then sampling S) allows us think of $[S']$ and S as statistically independent. Given a fixed choice of indices $[S']$, S' is simply an i.i.d. sample of size $\lfloor\sqrt{m}\rfloor$. Therefore, if $\mathbf{A}$ is consistent, $|F(\mathbf{A}(S')) - F^*| \le \varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$, this holds for any possible fixed $[S']$, and therefore

$$\mathbb{E}_{[S']}\left[\mathbb{E}_{S|[S']}\left[\left|F(\mathbf{A}(S')) - F^*\right|\right]\right] \;=\; \mathbb{E}_{[S']}\left[\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)\right] \;\le\; \varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor).$$

$\mathbf{A}'$ *generalizes:* For convenience, let $b(S,S') = |F_S(\mathbf{A}(S')) - F(\mathbf{A}(S'))|$. Using similar arguments and notation as above:

$$\begin{aligned}
&\mathbb{E}_S\left[\left|F_S(\mathbf{A}'(S)) - F(\mathbf{A}'(S))\right|\right] \\
&\le \mathbb{E}_{[S']}\left[\mathbb{E}_{S|[S']}\left[b(S,S')\right]\right] \\
&\le \mathbb{E}_{[S']}\left[\mathbb{E}_{S|[S']}\left[\frac{\lfloor\sqrt{m}\rfloor}{m} b(S',S')\right] \;+\; \mathbb{E}_{S|[S']}\left[\left(1 - \frac{\lfloor\sqrt{m}\rfloor}{m}\right) b(S\setminus S', S')\right]\right] \\
&\le \mathbb{E}_{[S']}\left[\frac{\lfloor\sqrt{m}\rfloor}{m} 2B + \left(1 - \frac{\lfloor\sqrt{m}\rfloor}{m}\right) \frac{B}{\sqrt{m - \lfloor\sqrt{m}\rfloor} + 1}\right],
\end{aligned}$$

where the last line follows from Lemma 12 and the fact that $b(S,S') \le 2B$ for any S, S'. It is not hard to show that the expression above is at most $4B/\sqrt{m}$, assuming $m \ge 1$.

$\mathbf{A}'$ *is strongly-uniform-RO stable:* For any sample S, any i and replacement instance $\mathbf{z}_i$, and any instance $\mathbf{z}'$, we have that

$$\left|f(\mathbf{A}'(S^{(i)}); \mathbf{z}') - f(\mathbf{A}'(S); \mathbf{z}')\right| \;\le\; \mathbb{E}_{S'}\left[\left|f(\mathbf{A}(S'^{(i)}); \mathbf{z}') - f(\mathbf{A}(S'); \mathbf{z}')\right|\right],$$

where we take $S'^{(i)}$ in the expectation to mean S' if $i \notin [S']$. Notice that if $i \notin [S']$, then $f(\mathbf{A}(S'^{(i)});\mathbf{z}_i) - f(\mathbf{A}(S');\mathbf{z}_i)$ is trivially 0. Thus, we can upper bound the expression above by

$$\mathbb{E}_{S'}\left[\left|f(\mathbf{A}(S'^{i});\mathbf{z}') - f(\mathbf{A}(S');\mathbf{z}')\right| \;\middle|\; i \in [S']\right].$$

Since S' is chosen uniformly over all $\lfloor\sqrt{m}\rfloor$-subsets of S, all permutations of $[S']$ are equally happen to occur, and therefore the above is equal to

$$\mathbb{E}_{S'}\left[\frac{1}{\lfloor\sqrt{m}\rfloor}\sum_{j\in S'}\left|f(\mathbf{A}(S'^{(j)});\mathbf{z}') - f(\mathbf{A}(S');\mathbf{z}')\right|\right] \;\leq\; \mathbb{E}_{S'}\left[\varepsilon_{\text{stable}}(\lfloor\sqrt{m}\rfloor)\right] \;=\; \varepsilon_{\text{stable}}(\lfloor\sqrt{m}\rfloor).$$

$\mathbf{A}'$ *is an always AERM:* For any fixed sample S, we note that

$$\begin{aligned}|F_S(\mathbf{A}'(S)) - F_S(\hat{\mathbf{h}}_S)| \;&=\; \mathbb{E}_{S'}\left[F_S(\mathbf{A}(S')) - F_S(\hat{\mathbf{h}}_S)\right]\\ &= \mathbb{E}_{S'\sim\mathcal{U}(S)^{\lfloor\sqrt{m}\rfloor}}\left[F_S(\mathbf{A}(S')) - F_S(\hat{\mathbf{h}}_S) \mid \text{no dups}\right],\end{aligned}$$

where $\mathcal{U}(S)^{\lfloor\sqrt{m}\rfloor}$ signifies the distribution of i.i.d. samples of size $\lfloor\sqrt{m}\rfloor$, picked uniformly at random (with replacement) from $\lfloor\sqrt{m}\rfloor$, and 'no dups' signifies the event that no element in S was picked twice. By the law of total expectation, this is at most

$$\frac{\mathbb{E}_{S'\sim\mathcal{U}(S)^{\lfloor\sqrt{m}\rfloor}}\left[F_S(\mathbf{A}(S')) - F_S(\hat{\mathbf{h}}_S)\right]}{\mathbb{P}\left[\text{no dups}\right]}.$$

Since the learning rule $\mathbf{A}$ is universally consistent, it is in particular consistent with respect to the distribution $\mathcal{U}(S)$, and therefore the expectation in the expression above is at most $\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$. As to $\mathbb{P}\left[\text{no dups}\right]$, an analysis identical to the one performed in the proof of Lemma 16 (see Equation (13)) implies that it is at least $1-(\lfloor\sqrt{m}\rfloor)^2/m \geq 1/2$. Overall, we get that $F_S(\mathbf{A}'(S)) - F_S(\hat{\mathbf{h}}_S) \leq 2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$, so in particular

$$\frac{\mathbb{E}_{S'\sim\mathcal{U}(S)^{\lfloor\sqrt{m}\rfloor}}\left[F_S(\mathbf{A}(S')) - F_S(\hat{\mathbf{h}}_S)\right]}{\mathbb{P}\left[\text{no dups}\right]} \leq 2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor),$$

from which the claim follows.

■

6.2 A Generic Learning Algorithm

Recall that a symmetric learning rule $\mathbf{A}$ is such that $\mathbf{A}(S) = \mathbf{A}(S')$ whenever S, S' are identical samples up to permutation. When we deal with randomized learning rules, we assume that the distribution of $\mathbf{A}(S)$ is identical to the distribution of $\mathbf{A}(S')$. Also, let $\mathcal{H}$ denote the set of all distributions on $\mathcal{H}$. An element $\bar{\mathbf{h}} \in \bar{H}$ will be thought of as a possible outcome of a randomized learning rule.

Consider the following learning rule: given a sample size m, find a minimizer over all symmetric[5] functions $\mathbf{A} : \mathcal{Z}^m \to \mathcal{H}$ of

$$\sup_{S\in\mathcal{Z}^m}\left(F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right) \;+\; \sup_{S\in\mathcal{Z}^m,\mathbf{z}'}\left|f(\mathbf{A}(S);\mathbf{z}') - f(\mathbf{A}(S^{(i)});\mathbf{z}')\right|, \tag{18}$$

with i being an arbitrary fixed element in $\{1,\ldots,m\}$. Once such a function $\mathbf{A}$ is found, return $\mathbf{A}_m(S)$.

5. The algorithm would still work, with slight modifications, if we minimize over all functions - symmetric or not. However, the search space would be larger.

Theorem 25 *If a learning problem is learnable (namely, there exist a universally consistent learning rule with rate $\varepsilon_{\text{cons}}(m)$), the learning algorithm described above is universally consistent with rate*

$$4\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{8B}{\sqrt{m}}.$$

Proof By Lemma 24, if a learning problem is learnable, there exists a (possibly randomized) symmetric learning rule $\mathbf{A}'$, which is an always AERM and strongly-uniform-RO stable. More specifically, we have that

$$\sup_{S\in\mathcal{Z}^m}\left(F_S(\mathbf{A}'(S))-F_S(\hat{\mathbf{h}}_S)\right)\leq 2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor),$$

as well as

$$\sup_{S\in\mathcal{Z}^m,\mathbf{z}'}\left|f(\mathbf{A}'(S);\mathbf{z}')-f(\mathbf{A}'(S^{(i)});\mathbf{z}')\right|\leq\frac{4B}{\sqrt{m}}.$$

In particular, there exists some symmetric $\mathbf{A}:Z^m\to\bar{\mathcal{H}}$, for which the expression in Equation (18) is at most

$$2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{4B}{\sqrt{m}}.$$

Therefore, by definition, the $\mathbf{A}$ found satisfies

$$\sup_{S\in\mathcal{Z}^m}\left(F_S(\mathbf{A}_m(S))-F_S(\hat{\mathbf{h}}_S)\right)\ \leq\ 2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{4B}{\sqrt{m}},\tag{19}$$

as well as

$$\sup_{S\in\mathcal{Z}^m}\left|f(\mathbf{A}_m(S);\mathbf{z}')-f(\mathbf{A}_m(S^{(i)});\mathbf{z}')\right|\ \leq\ 2\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{4B}{\sqrt{m}}.\tag{20}$$

In Theorem 9, we have seen that a universally average-RO stable AERM learning rule has to be universally consistent. The inequalities above essentially say that $\mathbf{A}$ is in fact both strongly-uniform-RO stable (and in particular, universally average-RO stable) and an AERM, and thus is a universally consistent learning rule. Formally speaking, this is not entirely accurate, because $\mathbf{A}$ is defined only with respect to samples of size m, and hence is not formally a learning rule which can be applied to samples of any size. However, the analysis we have done earlier in fact carries through also for learning rules $\mathbf{A}$ which are defined just on a specific sample size m. In particular, the analysis of Lemma 11 and Lemma 15 hold verbatim for $\mathbf{A}$ (with trivial modifications due to the fact that $\mathbf{A}$ is randomized), and together imply that since Equation (19) and Equation (20) hold, then

$$\mathbb{E}\left[F(\mathbf{A}(S))-F^*\right]\leq 4\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{8B}{\sqrt{m}}.$$

Therefore, our learning algorithm is consistent with rate $4\varepsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)+\frac{8B}{\sqrt{m}}$. ■

The main drawback of the algorithm we described is that it is completely infeasible: in practice, we cannot hope to efficiently perform minimization of Equation (18) over all functions from $\mathcal{Z}^m$ to $\bar{\mathcal{H}}$. Nevertheless, we believe it is conceptually important for three reasons: First, it hints that generic methods to develop learning algorithms might be possible in the General Learning Setting (similar to the more specific supervised classification setting); Second, it shows that stability might play a crucial role in the way such methods will work; And third, that stability might act in a similar manner to regularization. Indeed, Equation (18) can be seen as a "regularized ERM" in the space of learning rules (i.e., functions from samples to hypotheses): if we take just the first term in Equation (18), $\sup_{S\in\mathcal{Z}^m}\left(F_S(\mathbf{A}(S))-F_S(\hat{\mathbf{h}}_S)\right)$, then its minimizer is trivially the ERM learning rule. If we take just the second term in Equation (18), $\sup_{S\in\mathcal{Z}^m,\mathbf{z}}\left|f(\mathbf{A}(S);\mathbf{z}')-f(\mathbf{A}(S^{(i)});\mathbf{z}')\right|$, then its minimizers are trivial learning rules which return the same hypothesis irrespective of the training sample. Minimizing a sum of both terms forces us to choose a learning rule which is an "almost"-ERM but also stable - a learning rule which must exist if the problem is learnable at all, as Theorem 23 proves.

In any case, using these results and intuitions to design a generic, *practical* method to learn in the General Learning Setting - remains a very interesting open problem.

7. 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 $1-\delta$ over the sample, and the bounds depend linearly on $1/\delta$. However, in supervised classification, if we have learnability at all, then we have learnability at rates which are logarithmic in $1/\delta$. Can such results be attained in the General Learning Setting?

Fortunately, there is a generic method already known in the literature ("Boosting the Confidence", see Schapire, 1989) which allows us to convert any learning algorithm with linear dependence on δ to an algorithm with logarithmic dependence on $1/\delta$, at a certain price in terms of the sample complexity. This technique is reviewed below.

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 $1/\delta$. 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 $1/\delta$.

Theorem 26 *Let* $\mathbf{A}$ *be a universally consistent learning rule with rate* $\varepsilon_{\text{cons}}(m)$*, namely that*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[F(\mathbf{A}(S))-F^*\right]\leq\varepsilon_{\text{cons}}(m). \tag{21}$$

Then there exists another universally consistent learning rule $\mathbf{A}'$ *such that with probability at least* $1-\delta$ *over a sample* S *of size* m*,*

$$F(\mathbf{A}'(S))-F^*\leq e\varepsilon_{\text{cons}}\left(\frac{m}{\log(2/\delta)+1}\right)+2B\sqrt{\frac{\log(2/\delta)+\log(\log(2/\delta))}{2m}}$$

Proof Applying Markov's inequality on Equation (21), we have with probability at least $1-1/e$ over a sample S of size m that

$$F(\mathbf{A}(S))-F^*\leq e\varepsilon_{\text{cons}}(m). \tag{22}$$

Now, define the learning rule $\mathbf{A}'$ as follows: given a sample of size m, split it randomly into $a+1$ parts $S_1,\ldots,S_{a+1}$ of size $m/(a+1)$ each (where a is a constant to be determined later). Apply $\mathbf{A}$ separately on $S_1,\ldots,S_a$, to create a hypotheses $\mathbf{A}(S_1),\ldots,\mathbf{A}(S_a)$. Now, return the hypothesis $\mathbf{A}(S_t)$ which minimizes $F_{S_{a+1}}(\mathbf{A}(S_t))$ (namely, the hypothesis with lowest empirical risk on S_{a+1}), where ties are broken arbitrarily. By Equation (22), we have for any S_t separately that with probability at least $1-1/e$,

$$F(\mathbf{A}(S_t))-F^*\leq e\varepsilon_{\text{cons}}\left(\frac{m}{a+1}\right).$$

Since $F(\mathbf{A}(S_1)),\ldots,F(\mathbf{A}(S_a))$ are independent random variables, we have that with probability at least $1-(1/e)^a$, there exists at least one S_t such that

$$F(\mathbf{A}(S_t))-F^*\leq e\varepsilon_{\text{cons}}\left(\frac{m}{a+1}\right).$$

Assume w.l.o.g. that this holds for S_1. Using Hoeffding's inequality and a union bound, it also holds with probability at least $1-\delta_1$ over S that

$$F_{S_{a+1}}(\mathbf{A}(S_1))-F(\mathbf{A}(S_1))\leq B\sqrt{\frac{\log(2a/\delta_1)}{2m}},$$

and also

$$F(\mathbf{A}(S_t))-F_{S_{a+1}}(\mathbf{A}(S_t))\leq B\sqrt{\frac{\log(2a/\delta_1)}{2m}}$$

simultaneously for every $t = 2,\ldots,a$. If this happens, it means that we will pick a hypothesis whose risk is at most $2B\sqrt{\log(2a/\delta_1)/2m}$ larger than $F(\mathbf{A}(S_1))$. Overall, we have that with probability at least $1-\delta_1-(1/e)^a$,

$$F(\mathbf{A}'(S)) - F^* \leq e\varepsilon_{\text{cons}}\left(\frac{m}{a+1}\right) + 2B\sqrt{\frac{\log(2a/\delta_1)}{2m}}\ .$$

Picking $a = \log(2/\delta)$ and $\delta_1 = \delta/2$, we get that with probability at least $1-\delta$,

$$F(\mathbf{A}'(S)) - F^* \leq e\varepsilon_{\text{cons}}\left(\frac{m}{\log(2/\delta)+1}\right) + 2B\sqrt{\frac{\log(4/\delta)+\log(\log(4/\delta))}{2m}}$$

as required. ■

After we have seen how to convert a low-confidence learning rule (linear in δ) to a high-confidence learning rule (logarithmic in δ), we show that such conversions might actually be necessary, in sharp contrast to supervised classification.

Example 3 *There exists a learning problem where any ERM algorithm is universally consistent and average-RO stable with rates $\Theta(1/\sqrt{m})$, but for any ERM algorithm,*

$$\mathbb{P}\left[F(\hat{\mathbf{h}}_S) - F^* = 1\right] = \Theta\left(\frac{1}{\sqrt{m}}\right). \tag{23}$$

The $\Theta(\cdot)$ notation hides only absolute constants.

This example implies that no high-confidence bound is possible, at least without foregoing polynomial dependence on m. To see this, note that a high-confidence result corresponds to $\mathbb{P}\left[F(\hat{\mathbf{h}}_S) - F^*\right] > \varepsilon)$ decreasing exponentially in m for any fixed $\varepsilon > 0$, while in the case above we only have convergence at the rate of $1/\sqrt{m}$.

Proof Consider the instance space $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z} = [0,1]\times\{-1,+1\}\times\{-1,+1\}$, with any joint distribution such that $p(y,z|x)$ for any x is uniform on $\{-1,+1\}^2$, and the marginal distribution on $\mathcal{X}$ is continuous.

Consider the hypothesis class $\mathcal{H} = \mathcal{G}\cup\mathcal{B}$, where $\mathcal{G}$ consists of the constant function 1 and the constant function -1 over $[0,1]$, and $\mathcal{B}$ consists of all functions $\mathbf{h} : [0,1] \mapsto \{-1,0,+1\}$, such that each $\mathbf{h}(\cdot)$ equals 0 on all but a non-empty finite subset of $[0,1]$, and is uniformly either $+1$ or -1 on this finite subset.

Finally, define the objective function as

$$f(\mathbf{h},(x,y,z)) = \left(\mathbf{1}(\mathbf{h}\in\mathcal{G})y + \frac{\mathbf{1}(\mathbf{h}\in\mathcal{B})z}{2|\mathbf{h}|}\right)\mathbf{h}(x) + \mathbf{1}(\mathbf{h}(x)=0),$$

where $|\mathbf{h}| = |\{x\in[0,1] : \mathbf{h}(x)\neq 0\}|$ (namely, the number of points in $[0,1]$ on which the function $h(\cdot)$ is not zero). For $\mathbf{h}\in\mathcal{G}$, where the number of such points is infinite, we take $|\mathbf{h}| = \infty$.

First, notice that for any $\mathbf{h}\in\mathcal{G}$, $F(\mathbf{h}) = 0$, and for any $\mathbf{h}\in B$, $F(\mathbf{h}) = 1$. Thus, we can think of $\mathcal{G}$ as the set of "good" hypotheses, and $\mathcal{B}$ as the set of 'bad" hypotheses. Our goal is to show that any ERM will pick a hypothesis from $\mathcal{B}$ with probability $\Theta(1/\sqrt{m})$.

We need to do a bit of case-by-case analysis. Let $(x_1,y_1,z_1),\ldots,(x_m,y_m,z_m)$ be the sample. If $\sum_{i=1}^m y_i \neq 0$, then using hypotheses in G, it is possible to achieve an empirical risk of

$$-\left|\sum_{i=1}^m y_i\right| \leq -1,$$

while using hypotheses in B, it is only possible to achieve an empirical risk of

$$\frac{\sum_{i=1}^m z_i\mathbf{h}(x_i)}{2|\mathbf{h}|} + \sum_{i=1}^m \mathbf{1}(\mathbf{h}(x_i)=0) \geq -\frac{1}{2}.$$

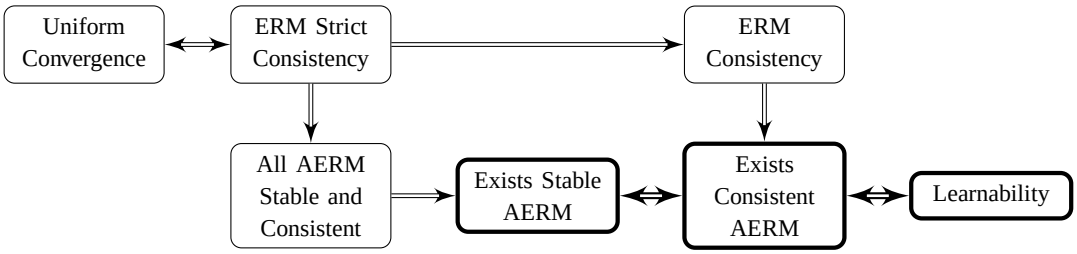


Figure 2: Implications of various properties of learning problems. Consistency refers to univeral consistency and stability refers to universal uniform-RO stability.

Thus, with probability $1-\Theta(1/\sqrt{m})$ (the probability that $\sum_{i=1}^{m} y_i \neq 0$ in the sample), any ERM algorithm will pick $\hat{\mathbf{h}}_S \in G$.

If $\sum_{i=1}^{m} y_i = 0$, then any $\mathbf{h} \in G$ achieves an empirical objective value of 0. On the other hand, unless $\sum_{i=1}^{m} z_i = 0$, we can choose some $\mathbf{h} \in B$, which is non-zero on all points in the sample, and achieves an empirical risk smaller than 0. The probability that $\sum_{i=1}^{m} y_i = 0$ and $\sum_{i=1}^{m} z_i \neq 0$ is $\Theta(1/\sqrt{m})(1-\Theta(1/\sqrt{m}))$, or $\Theta(1/\sqrt{m})$.

So we have that any ERM picks $\hat{\mathbf{h}}_S \in G$ with probability $1-\Theta(1/\sqrt{m})$, and some $\hat{\mathbf{h}}_S \in B$ with probability $\Theta(1/\sqrt{m})$, from which the consistency rate and Equation (23) in the theorem statement follows. Finally, note that replacing a single instance in the training set will lead to the ERM picking a different hypothesis, only if $\sum_{i=1}^{m} y_i = 0$ before or after the replacement. The probability for getting a training set where this happens is $O(1/\sqrt{m})$, and from this it is easy to see that the ERM is average-RO stable with rate $O(1/\sqrt{m})$. ■

8. 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 minimization, 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 non-trivial 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. This also allows us to frame stability as the core condition guaranteeing learnability, with uniform convergence only a sufficient, but not necessary, condition for stability (see Figure 2).

In studying the question of learnability and its relation to stability, we encountered several differences between this more general setting, and settings such as supervised classification where learnability is equivalent to uniform convergence. We summarize some of these distinctions:

- Perhaps the most important distinction is that in the General Setting learnability might be possible only with a non-ERM. In this paper we establish that if a problem is learnable, although it might not be learnable with an ERM, it must be learnable with some AERM. And so, in the General Setting we must look beyond empirical risk minimization, but not beyond asymptotic empirical risk minimization.

- In supervised classification, if one AERM is universally consistent then all AERMs are universally consistent. In the General Setting we must choose the AERM carefully.

- In supervised classification, a universally consistent rule must also generalize and be AERM. In the General Setting, a universally consistent rule need not generalize nor be an AERM, as example 2 demonstrates. However, Theorem 10 establishes that, even in the General Setting, if a rule is universally consistent *and* generalizing then it must be an AERM. This gives us another reason to not look beyond asymptotic empirical risk minimization, even in the General Setting.

  The above distinctions can also be seen through Corollary 19, which is concerned with the relationship between AERM, consistency and generalization in learnable problems. In the General Setting, any two conditions imply the other, but it is possible for any one condition to exist without the others. In supervised classification, if a problem is learnable then generalization always holds (for any rule), and so universal consistency and AERM imply each other.
- In supervised classification, ERM inconsistency for some distribution is enough to establish non-learnability. Establishing non-learnability in the General Setting is trickier, since one must consider all AERMs. We show how Corollary 19 can provide a *certificate* for non-learnability, in the form of a rule that is consistent and an AERM for some specific distribution, but does not generalize (Example 1).

- In supervised classification, any learnable problem is learnable with an ERM, *and* the ERM "works" with high-confidence (namely, $F(\hat{\mathbf{h}}_S) - F^*$ can be bounded with probability $1-\delta$ by an expression with logarithmic dependence on $1/\delta$). In Section 7 we have seen that in the General Learning Setting, even if the ERM is universally consistent, high-confidence bounds for the ERM might be impossible to obtain.

We have begun exploring the issue of learnability in the General Setting, and uncovered important relationships between learnability and stability. But many problems are left open, some of which are listed below.

First, is it possible to come up with well-known machine learning applications, where learnability is achievable despite uniform convergence failing to hold?

In Section 6.2, we have managed to obtain a completely generic learning algorithm: an algorithm which in principle allows us to learn any learnable problem. However, the algorithm suffers from the severe drawback that in general, it requires unbounded computational power. Can we derive an efficient algorithm, or characterize classes of learning problems where our algorithm, or some other generic learning algorithm using the notion of stability, can be executed efficiently? For instance, can we always learn using a regularized ERM learning rule?

On a related vein, it would be interesting to develop learning algorithms (perhaps for specific settings rather than generic learning problems) which directly use stability in order to learn. Convex regularization is one such mechanism, as discussed in Section 4. Are there other mechanisms, which use the notion of stability in a different way?

Another issue is that even the existence of uniform-RO stable AERM (or strongly-uniform-RO stable, always-AERM allowing for convexity/randomization) is not as elegant and simple as having finite VC dimension or fat-shattering dimension. It would be very interesting to derive equivalent but more "combinatorial" conditions for learnability.

Yet another open question: We showed that existence of an uniform-RO stable AERM is necessary and sufficient for learnability (Theorem 7). However, it is possible that learnability is an equivalent to the existence of an AERM with a stronger notion of stability, without resorting to convexity/randomization as we

have done in Section 6.2. This might perhaps lead to generic learning algorithms which perform minimization over a search space more feasible than the one our algorithm (in Section 6.2) uses.

Finally, we do not know whether it is enough to consider symmetric learning rules: that is, learning rules which do not depend on the order of the instances in the training sample. Intuitively, this should be true, since the instances were sampled i.i.d. Can our characterization of learnability (e.g., existence of a uniform-RO stable AERM learning rule) be strengthened to existence of symmetric uniform-RO stable AERM learning rule, without allowing convexity/randomization?

Acknowledgments

We would like to thank Leon Bottou, Vladimir Vapnik and Tong Zhang for helpful discussions. We would also like to thank the anonymous reviewers for their detailed and helpful comments.

Appendix A. Alternative Notions of Stability

In this appendix, we discuss how our definition of stability compares to previous definitions in the literature, as well as demonstrate how subtleties involved in the precise choice of the definition can have a significant effect on the results which can be obtained.

A.1 Previous Definitions in the Literature

The existing literature on stability in learning, briefly surveyed in Section 3.2, uses many different stability measures. All of them measure 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, they differ in how "output", "amount of change to the output", and "small changes to the sample" are defined. In Section 5, we used three stability measures. Roughly speaking, one measure (average-RO stability) is the expected change in the objective value on a particular instance, after that instance is replaced with a different instance. The second measure and third measure (uniform-RO stability and strongly-uniform-RO stability respectively) basically deal with the maximal possible change in the objective value with respect to a particular instance, by replacing a single instance in the training set. However, instead of measuring the objective value on a specific instance, we could have measured the change in the risk of the returned hypothesis, or any other distance measure between hypotheses. Instead of replacing an instance, we could have talked about adding or removing one instance from the sample, either in expectation or in some arbitrary manner. Such variations are common in the literature.

To relate our stability definitions to the ones in the literature, we note that our definitions of uniform-RO stability and strongly-uniform-RO stability are somewhat similar to uniform stability (Bousquet and Elisseeff, 2002), which in our notation is defined as $\sup_{S,\mathbf{z}} \max_i |f(\mathbf{A}(S;\mathbf{z})) - f(\mathbf{A}(S^{\setminus i});\mathbf{z})|$, where $S^{\setminus i}$ is the training sample S with instance $\mathbf{z}_i$ removed. Compared to uniform-RO stability, here we measure maximal change over any particular instance, rather than average change over all instances in the training sample. Also, we deal with removing an instance rather than replacing it. Strongly-uniform-RO stability is more similar, with the only formal difference being removal vs. replacement of an instance. However, the results for uniform stability mostly assume deterministic learning rules, while in this paper we have used strongly-uniform-RO stability solely in the context of randomized learning rules. For deterministic learning rules, the differences outlined above are sufficient to make uniform stability a strictly stronger requirement than uniform-RO stability, since it is easy to come up with learning problems and (non-symmetric) learning rules which are uniform-RO stable but not uniformly stable. Moreover, we show in this paper that uniform-RO stable AERM's characterize learnability, while it is well known that uniformly stable AERM's are not necessary for learnability (see Kutin and Niyogi, 2002). For the same reason, our notion of strongly-uniform-RO stability is apparently too strong to characterize learnability when we deal with deterministic learning rules, as opposed to randomized learning rules.

Our definition of average-RO stable is similar to "average stability" defined in Rakhlin et al. (2005), which in our notation is defined as $\mathbb{E}_{S\sim\mathcal{D}^m,\mathbf{z}'_1}\left[f(\mathbf{A}(S^{(i)});\mathbf{z}_1)-f(\mathbf{A}(S);\mathbf{z}_1)\right]$. Compared to average-RO stability, the main difference is that the change in the objective value is measured with respect to $\mathbf{z}_1$ rather than an average over $\mathbf{z}_i$ for all i, and stems from the assumption there that the learning algorithm is symmetric. Notice that in this paper we do not make such an assumption.

For an elaborate study on other stability notions and their relationships, see Kutin and Niyogi (2002).

Unfortunately, many of the stability notions in the literature are incomparable, and even slight changes in the definition radically affect their behavior. We go into this in much more detail in the following subsections.

A.2 LOO Stability vs. RO Stability

The stability definitions we have used in this paper are all based on the idea of replacing one instance in the training sample by another instance (e.g., "RO" or "replace-one" stability). An alternative set of definitions can be obtained based on *removing* one instance in the training sample (e.g., "LOO" or "leave-one-out" stability). In fact, these were the definitions used in our preliminary paper (Shalev-Shwartz et al., 2009b). Despite seeming like a small change, it turns out there is a considerable discrepancy in terms of the obtainable results, compared to RO stability. In this subsection, we wish to discuss these discrepancies, as well as show how small changes to the stability definition can materially affect its strength.

Specifically, we consider the following four LOO stability measures, each slightly weaker than the previous one. The first and last are similar to our notion of uniform-RO stability and average-RO stability respectively. However, we emphasize that RO stability and LOO stability are in general incomparable notions, as we shall see later on. Also, we note that some of these definitions appeared in previous literature. For instance, the notion of "all-i-LOO" below has been studied by several authors under different names (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Rakhlin et al., 2005). The notation $S^{\setminus i}$ below refer to a training sample S with instance $\mathbf{z}_i$ removed.

Definition 27 *A rule* $\mathbf{A}$ *is* **uniform-LOO stable** *with rate* $\varepsilon_{\text{stable}}(m)$ *if for all samples* S *of* m *points and for all* i*:*

$$\left|f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right|\leq\varepsilon_{\text{stable}}(m).$$

Definition 28 *A rule* $\mathbf{A}$ *is* **all-i-LOO stable** *with rate* $\varepsilon_{\text{stable}}(m)$ *under distribution* $\mathcal{D}$ *if for all* i*:*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\left|f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right|\right]\leq\varepsilon_{\text{stable}}(m).$$

Definition 29 *A rule* $\mathbf{A}$ *is* **LOO stable** *with rate* $\varepsilon_{\text{stable}}(m)$ *under distribution* $\mathcal{D}$ *if*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\left|f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right|\right]\leq\varepsilon_{\text{stable}}(m).$$

Definition 30 *A rule* $\mathbf{A}$ *is* **on-average-LOO stable** *with rate* $\varepsilon_{\text{stable}}(m)$ *under distribution* $\mathcal{D}$ *if*

$$\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right]\right|\leq\varepsilon_{\text{stable}}(m).$$

While some of the definitions above might look rather similar, we show below that each one is strictly stronger than the other. Example 6 is interesting in its own right, since it presents a learning problem and an AERM that is universally consistent, but not LOO stable. While this is possible in the General Learning Setting, in supervised classification every such AERM has to be LOO stable (this is essentially proven in Mukherjee et al., 2006).

Example 4 *There exists a learning problem with a universally consistent and all-i-LOO stable learning rule, but there is no universally consistent and uniform LOO stable learning rule.*

Proof This example is taken from Kutin and Niyogi (2002). Consider the hypothesis space $\{0,1\}$, the instance space $\{0,1\}$, and the objective function $f(\mathbf{h},z) = |h-z|$.

It is straightforward to verify that an ERM is a universally consistent learning rule. It is also universally all-i-LOO stable, because removing an instance can change the hypothesis only if the original sample had an equal number of 0's and 1's (plus or minus one), which happens with probability at most $O(1/\sqrt{m})$ where m is the sample size. However, it is not hard to see that the only uniform LOO stable learning rule, at least for large enough sample sizes, is a constant rule which always returns the same hypothesis h regardless of the sample. Such a learning rule is obviously not universally consistent. ■

Example 5 *There exists a learning problem with a universally consistent and LOO-stable AERM, which is not symmetric and is not all-i-LOO stable.*

Proof Let the instance space be $[0,1]$, the hypothesis space $[0,1] \cup 2$, and the objective function $f(h,z) = \mathbb{1}_{\{h=z\}}$. Consider the following learning rule $\mathbf{A}$: given a sample, check if the value z_1 appears more than once in the sample. If no, return z_1, otherwise return 2.

Since $F_S(2) = 0$, and z_1 returns only if this value constitutes $1/m$ of the sample, the rule above is an AERM with rate $\varepsilon_{\text{erm}}(m) = 1/m$. To see universal consistency, let $\mathbb{P}[z_1] = p$. With probability $(1-p)^{m-2}$, $z_1 \notin \{z_2,\ldots,z_m\}$, and the returned hypothesis is z_1, with $F(z_1) = p$. Otherwise, the returned hypothesis is 2, with $F(2) = 0$. Hence $\mathbb{E}_S[F(\mathbf{A}(S))] \leq p(1-p)^{m-2}$, which can be easily verified to be at most $1/(m-1)$, so the learning rule is consistent with rate $\varepsilon_{\text{cons}}(m) \leq 1/(m-1)$. To see LOO-stability, notice that our learning hypothesis can change by deleting z_i, $i > 1$, only if z_i is the only instance in $z_2,\ldots,z_m$ equal to z_1. So $\varepsilon_{\text{stable}}(m) \leq 2/m$ (in fact, LOO-stability holds even without the expectation). However, this learning rule is not all-i-LOO-stable. For instance, for any continuous distribution, $|f(\mathbf{A}(S^{\setminus 1}),z_1) - f(\mathbf{A}(S),z_1)| = 1$ with probability 1, so it obviously cannot be all-i-LOO-stable with respect to $i = 1$. ■

Example 6 *There exists a learning problem with a universally consistent (and on-average-LOO stable) AERM, which is not LOO stable.*

Proof Let the instance space, hypothesis space and objective function be as in Example 4. Consider the following learning rule, based on a sample $S = (z_1,\ldots,z_m)$: if $\sum_i \mathbb{1}_{\{z_i=1\}}/m > 1/2 + \sqrt{\log(4)/2m}$, return 1. If $\sum_i \mathbb{1}_{\{z_i=1\}}/m < 1/2 - \sqrt{\log(4)/2m}$, return 0. Otherwise, return $\text{Parity}(S) = (z_1 + \ldots z_m) \bmod 2$.

This learning rule is an AERM, with $\varepsilon_{\text{erm}}(m) = \sqrt{2\log(4)/m}$. Since we have only two hypotheses, we have uniform convergence of $F_S(\cdot)$ to $F(\cdot)$ for any hypothesis. Therefore, our learning rule universally generalizes (with rate $\varepsilon_{\text{gen}}(m) = \sqrt{\log(4/\delta)/2m}$), and by Theorem 9, this implies that the learning rule is also universally consistent and on-average-LOO stable.

However, the learning rule is not LOO stable. Consider the uniform distribution on the instance space. By Hoeffding's inequality, $|\sum_i \mathbb{1}_{\{z_i=1\}}/m - 1/2| \leq \sqrt{\log(4)/2m}$ with probability at least $1/2$ for any sample size m. In that case, the returned hypothesis is the parity function (even when we remove an instance from the sample, assuming $m \geq 3$). When this happens, it is not hard to see that for any i,

$$f(\mathbf{A}(S),z_i) - f(\mathbf{A}(S^{\setminus i}),z_i) = \mathbb{1}_{\{z_i=1\}}(-1)^{\text{Parity(S)}}.$$

This implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left|\left(f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right)\right|\right] \qquad (24)\\
&\geq \frac{1}{2}\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\{z_i=1\}}\left|\sqrt{\frac{\log(4)}{2m}}\geq\left|\sum_{i=1}^{m}\frac{\mathbb{1}_{\{z_i=1\}}}{m}-\frac{1}{2}\right|\right]\right.\\
&\geq \frac{1}{2}\left(\frac{1}{2}-\sqrt{\frac{\log(4)}{2m}}\right) \;\longrightarrow\; \frac{1}{4}\ ,
\end{aligned}$$

which does not converge to zero with the sample size m. Therefore, the learning rule is not LOO stable. ■

Note that the proof implies that on-average-LOO stability cannot be replaced even by something between on-average-LOO stability and LOO stability. For instance, a natural candidate would be

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\left(f(\mathbf{A}(S^{\setminus i});\mathbf{z}_i)-f(\mathbf{A}(S);\mathbf{z}_i)\right)\right|\right], \qquad (25)$$

where the absolute value is now over the entire sum, but inside the expectation. In the example used in the proof, Equation (25) is still lower bounded by Equation (24), which does not converge to zero with the sample size.

After showing that the hierarchy of definitions above is indeed strict, we turn to the question of what can be characterized in terms of LOO stability. In Shalev-Shwartz et al. (2009b), we show a version of Theorem 7, which asserts that a problem is learnable if and only if there is an on-average-LOO stable AERM. However, on-average-LOO stability is qualitatively much weaker than the notion of uniform-RO stability used in Theorem 7 (see Definition 4). Rather, we would expect to prove a version of the theorem with the notion of uniform-LOO stability or at least LOO stability, which are more analogous to uniform-RO stability. However, the proof of Theorem 7 does not work for these stability definitions (technically, this is because the proof relies on the sample size remaining constant, which is true for replacement stability, but not when we remove an instance as in LOO stability). We do not know if one can prove a version of Theorem 7 with an LOO stability notion stronger than on-average-LOO stability.

On the plus side, LOO stability allows us to prove the following interesting result, specific to ERM learning rules.

Theorem 31 *For an ERM the following are equivalent:*

- *Universal LOO stability.*
- *Universal consistency.*
- *Universal generalization.*

In particular, the theorem implies that LOO stability is a necessary property for consistent ERM learning rules. This parallels Theorem 9, which dealt with AERM's in general, and used RO stability. As before, we do not know how to obtain something akin to Theorem 9 with RO stability.

Proof Lemma 15 and Lemma 17 from Section 5.3.3 already tell us that for ERM's, universal consistency is equivalent to universal generalization. Moreover, Lemma 14 implies that for ERM's, generalization is equivalent to on-average generalization (see Equation (11) for the exact definition). Thus, is left to prove that for ERM's, generalization implies LOO stability, and LOO stability implies on-average generalization. stability.

First, suppose the ERM learning rule is generalizing with rate $\varepsilon_{\text{gen}}(m)$. Note that $f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i) - f(\hat{\mathbf{h}}_S; z_i)$ is always nonnegative. Therefore the LOO stability of the ERM can be upper bounded as follows:

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left|f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i) - f(\hat{\mathbf{h}}_S; z_i)\right|\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i) - f(\hat{\mathbf{h}}_S; z_i)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[F(\hat{\mathbf{h}}_{S^{\setminus i}})\right] - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} f(\hat{\mathbf{h}}_S; z_i)\right] \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}}) + \varepsilon_{\text{gen}}(m-1)\right] - \mathbb{E}\left[F_S(\hat{\mathbf{h}}_S)\right] \\
&= \varepsilon_{\text{gen}}(m-1) + \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_S(\hat{\mathbf{h}}_S)\right] \\
&\leq \varepsilon_{\text{gen}}(m-1).
\end{aligned}$$

For the opposite direction, suppose the ERM learning rule is LOO stable with rate $\varepsilon_{\text{stable}}(m)$. Notice that we can get any sample of size $m-1$ by picking a sample S of size m and discarding any instance i. Therefore, the on-average generalization rate of the ERM for samples of size $m-1$ is equal to the following:

$$\begin{aligned}
&\left|\mathbb{E}\left[F(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}})\right]\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[F(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}})\right]\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i)\right] - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}})\right]\right|
\end{aligned}$$

Now, note that for the ERM's of S and $S^{\setminus i}$ we have $\left|F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_S(\hat{\mathbf{h}}_S)\right| \leq \frac{2B}{m}$. Therefore, we can upper bound the above by

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i)\right] - \mathbb{E}\left[F_S(\hat{\mathbf{h}}_S)\right]\right| + \frac{2B}{m} \\
&= \left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[f(\hat{\mathbf{h}}_{S^{\setminus i}}; z_i) - f(\hat{\mathbf{h}}_{S,}; z_i)\right]\right| \\
&\leq \varepsilon_{\text{stable}}(m)
\end{aligned}$$

using the assumption that the learning rule is $\varepsilon_{\text{stable}}(m)$-stable. ■

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.

A. 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.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

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

J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

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.

D. McAllester and R. Schapire. On the convergence rate of Good-Turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 1–6, 2000.

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.

D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM*, 9(1):84–97, 1962.

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. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

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.

K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 22*, pages 1545–1552, 2008.

A. N. Tikhonov. On the stability of inverse problems. *Dolk. Akad. Nauk SSSR*, 39(5):195–198, 1943.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI(2):264–280, 1971.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, 2003.