# Learnability and Stability in the General Learning Setting

**Shai Shalev-Shwartz**
TTI-Chicago
shai@tti-c.org

**Ohad Shamir**
The Hebrew University
ohadsh@cs.huji.ac.il

**Nathan Srebro**
TTI-Chicago
nati@uchicago.edu

**Karthik Sridharan**
TTI-Chicago
karthik@tti-c.org

## Abstract

We establish that stability is necessary and sufficient for learning, even in the General Learning Setting where uniform convergence conditions are not necessary for learning, and where learning might only be possible with a non-ERM learning rule. This goes beyond previous work on the relationship between stability and learnability, which focused on supervised classification and regression, where learnability is equivalent to uniform convergence and it is enough to consider the ERM.

## 1 Introduction

We consider the General Setting of Learning [10] where we would like to minimize a population risk functional (stochastic objective)

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

where the distribution $\mathcal{D}$ of $Z$ is unknown, based on i.i.d. sample $z_1,\dots,z_m$ drawn from $\mathcal{D}$ (and full knowledge of the function $f$). This General Setting subsumes supervised classification and regression, certain unsupervised learning problems, density estimation and stochastic optimization. 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.

For supervised classification and regression problems, it is well known that a problem is *learnable* (see precise definition in Section 2) if and only if the empirical risks

$$F_S(\mathbf{h}) = \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{h}, z_i) \tag{2}$$

converge uniformly to their expectations [1]. If uniform convergence holds, then the empirical risk minimizer (ERM) is *consistent*, i.e. the population risk of the ERM converges to the optimal population risk, and the problem is learnable using the ERM. That is, learnability is equivalent to learnability by ERM, and so we can focus our attention solely on empirical risk minimizers.

Stability has also been suggested as an explicit alternate 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 has been established that various forms of stability of the ERM are sufficient for learnability. Mukherjee *et al* [7] argue that since uniform convergence also implies stability of the ERM, and is necessary for (distribution independent) learning in the supervised classification and regression setting, then stability of the ERM is necessary and sufficient for learnability *in the supervised classification and regression setting*. It is important to emphasize that this characterization of stability as necessary for learnability goes through uniform convergence, i.e. the chain of implications is:

$$\begin{tikzcd}\boxed{\text{ERM Stable}} \arrow[r, Rightarrow] & \boxed{\text{Learnable with ERM}} \arrow[r, Leftrightarrow] & \boxed{\text{Uniform Convergence}} \arrow[ll, Rightarrow, bend left]\end{tikzcd}$$

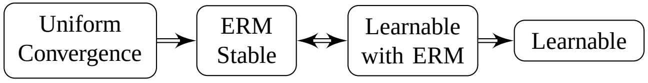
However, the equivalence between (distribution independent) consistency of empirical risk minimization and uniform convergence is specific to supervised classification and regression. The results of Alon *et al* [1] establishing this equivalence do *not* always hold in the General Learning Setting. In particular, we recently showed that in strongly convex stochastic optimization problems, the ERM is stable and thus consistent, even though the empirical risks do *not* converge to their expectations uniformly (Example 7.1, taken from [9]). Since the other implications in the chain above still hold in the general learning setting (e.g., uniform convergence implies stability and stability implies learnability by ERM), this example demonstrates that stability is a strictly more general sufficient condition for learnability.

A natural question is then whether, in the General Setting, stability is also necessary for learning. Here we establish that indeed, even in the general learning setting, (distribution independent) stability of ERM is necessary and sufficient for (distribution independent) consistency of the ERM. The situation is therefore as follows:

$$\begin{tikzcd}\boxed{\text{Uniform Convergence}} \arrow[r, Rightarrow] & \boxed{\text{ERM Stable}} \arrow[r, Leftrightarrow] & \boxed{\text{Learnable with ERM}}\end{tikzcd}$$

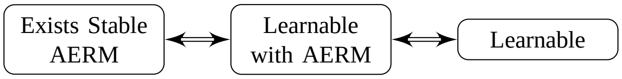
We emphasize that, unlike the arguments of Mukherjee *et al* [7], the proof of necessity does *not* go through uniform convergence, allowing us to deal also with settings beyond supervised classification and regression.

The discussion above concerns only stability and learnability of the ERM. In supervised classification and regression there is no need to go beyond the ERM, since learnability is equivalent to learnability by empirical risk minimization. But as we recently showed, there are learning problems in the General Setting which are learnable using some alternate learning rule, but in which ERM is neither stable nor consistent (Example 7.2, taken from [9]). Stability of ERM is therefore a sufficient, but not necessary, condition for learnability:



This prompts us to study the stability properties of non-ERM learning rules.

We establish that, even in the General Setting, any consistent and generalizing learning rule (i.e. where in addition to consistency, the empirical risk is also a good estimate of the population risk) must be asymptotically empirical risk minimizing (AERM, see precise definition in Section 2). We thus focus on such rules and show that also for an AERM, stability is sufficient for consistency and generalization. The converse is a bit weaker for AERMs, though. We show that a strict notion of stability, which is necessary for ERM consistency, is not necessary for AERM consistency, and instead suggest a weaker notion (averaging out fluctuations across random training sets) that is necessary and sufficient for AERM consistency. Noting that any consistent learning rule can be transformed to a consistent and generalizing learning rule, we obtain a sharp characterization of learnability in terms of stability—learnability is equivalent to the existence of a stable AERM:



## 2 The General Learning Setting

A "learning problem" is specified by a hypothesis domain $\mathcal{H}$, an instance domain $\mathcal{Z}$ and an objective function (e.g. "loss functional" or "cost function") $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. Throughout this paper we assume the function is bounded by some constant $B$, i.e. $|f(\mathbf{h}, z)| \leq B$ for all $\mathbf{h} \in \mathcal{H}$ and $z \in \mathcal{Z}$.

A "learning rule" is a mapping $\mathbf{A} : \cup_m \mathcal{Z}^m \to \mathcal{H}$ from sequences of instances in $\mathcal{Z}$ to hypotheses. We refer to sequences $S = \{z_1, \dots, z_m\}$ 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 is said to be *symmetric*. We will generally consider samples $S \sim \mathcal{D}^m$ of $m$ i.i.d. draws from $\mathcal{D}$.

A possible approach to learning is to minimize the empirical risk $F_S(\mathbf{h})$. We say that a rule $\mathbf{A}$ is an **ERM (Empirical Risk Minimizer)** if it minimizes the empirical risk

$$F_S(\mathbf{A}(S)) = F_S(\hat{\mathbf{h}}) = \min_{\mathbf{h} \in \mathcal{H}} F_S(\mathbf{h}). \tag{3}$$

where we use $F_S(\hat{\mathbf{h}}) = \min_{\mathbf{h} \in \mathcal{H}} F_S(\mathbf{h})$ to refer to the minimum empirical risk. But since there might be several hypotheses minimizing the empirical risk, $\hat{\mathbf{h}}$ 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 (Asymptotical Empirical Risk Minimizer)** with rate $\epsilon_{\mathrm{erm}}(m)$ under distribution $\mathcal{D}$ if:

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right] \leq \epsilon_{\mathrm{erm}}(m) \tag{4}$$

Here and whenever talking about a "rate" $\epsilon(m)$, we require it be monotone decreasing with $\epsilon(m) \overset{m \to \infty}{\longrightarrow} 0$. A learning rule is **universally an AERM** with rate $\epsilon_{\mathrm{erm}}(m)$, if it is an AERM with rate $\epsilon_{\mathrm{erm}}(m)$ under all distributions $\mathcal{D}$ over $\mathcal{Z}$.

Returning to our goal of minimizing the expected risk, we say a rule $\mathbf{A}$ is **consistent** with rate $\epsilon_{\mathrm{cons}}(m)$ under distribution $\mathcal{D}$ if for all $m$,

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[F(\mathbf{A}(S)) - F(\mathbf{h}^*)\right] \leq \epsilon_{\mathrm{cons}}(m). \tag{5}$$

where we denote $F(\mathbf{h}^*) = \inf_{\mathbf{h} \in \mathcal{H}} F(\mathbf{h})$. A rule is **universally consistent** with rate $\epsilon_{\mathrm{cons}}(m)$ if it is consistent with rate $\epsilon_{\mathrm{cons}}(m)$ under all distributions $\mathcal{D}$ over $\mathcal{Z}$. A problem is said to be **learnable** if there exists some universally consistent learning rule for the problem. 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 [4] to Vapnik's General Setting of Learning as studied by Haussler [3] and others.

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

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[|F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))|\right] \leq \epsilon_{\mathrm{gen}}(m). \tag{6}$$

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

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

## 3 Stability

We define a sequence of progressively weaker notions of stability, all based on leave-one-out validation. For a sample $S$ of size $m$, let $S^{\backslash i} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_m\}$ be a sample of $m-1$ points obtained by deleting the $i$-th observation of $S$. All our measures of stability concern the effect deleting $z_i$ has on $f(\mathbf{h}, z_i)$, where $\mathbf{h}$ is the hypotheses returned by the learning rule. That is, all measures consider the magnitude of $f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i)$.

**Definition 1.** *A rule* $\mathbf{A}$ *is* **uniform-LOO stable** *with rate* $\epsilon_{\mathrm{stable}}(m)$ *if for all samples* $S$ *of* $m$ *points and for all* $i$*:*

$$\left| f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i) \right| \leq \epsilon_{\mathrm{stable}}(m).$$

**Definition 2.** *A rule* $\mathbf{A}$ *is* **all-i-LOO stable** *with rate* $\epsilon_{\mathrm{stable}}(m)$ *under distributions* $\mathcal{D}$ *if for all* $i$*:*

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\left| f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i) \right|\right] \leq \epsilon_{\mathrm{stable}}(m).$$

**Definition 3.** *A rule* $\mathbf{A}$ *is* **LOO stable** *with rate* $\epsilon_{\mathrm{stable}}(m)$ *under distributions* $\mathcal{D}$ *if*

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{S \sim \mathcal{D}^m}\left[\left| f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i) \right|\right] \leq \epsilon_{\mathrm{stable}}(m).$$

For symmetric learning rules, Definitions 2 and 3 are equivalent. Example 7.5 shows that the symmetry assumption is necessary, and the two definitions are not equivalent for non-symmetric learning rules.

Our weakest notion of stability, which we show is still enough to ensure learnability, is:

**Definition 4.** *A rule* $\mathbf{A}$ *is* **on-average-LOO stable** *with rate* $\epsilon_{\text{stable}}(m)$ *under distributions* $\mathcal{D}$ *if*

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

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

**Claim 3.1.** *Uniform-LOO stability with rate* $\epsilon_{\text{stable}}(m)$ *implies all-i-LOO stability with rate* $\epsilon_{\text{stable}}(m)$*, which implies LOO stability with rate* $\epsilon_{\text{stable}}(m)$*, which implies on-average-LOO stability with rate* $\epsilon_{\text{stable}}(m)$*.*

### Relationship to Other Notions of Stability

Many different notions of stability, some under multiple names, have been suggested in the literature.

In particular, our notion of all-i-LOO stability has been studied by several authors under different names: pointwise-hypothesis stability [2], $\text{CV}_{\text{loo}}$ stability [7], and cross-validation-(deletion) stability [8]. All are equivalent, though the rate is sometimes defined differently. Other authors define stability with respect to replacing, rather then deleting, an observation. E.g. “CV stability” [5] and “cross-validation-(replacement)” [8] are analogous to all-i-LOO stability and “average stability” [8] is analogous to average-LOO stability for symmetric learning rules. In general the deletion and replacement variants of stability are incomparable—in Appendix A we briefly discuss how the results in this paper change if replacement stability is used.

A much stronger notion is *uniform stability* [2], which is strictly stronger than any of our notions, and is sufficient for tight generalization bounds. However, this notion is far from necessary for learnability ([5] and Example 7.3 below).

In the context of symmetric learning rules, all-i-LOO stability and LOO stability are equivalent. In order to treat non-symmetric rules more easily, we prefer working with LOO stability.

For an elaborate discussion of the relationships between different notions of stability, see [5].

## 4 Main Results

We first establish that existence of a stable AERM is sufficient for learning:

**Theorem 4.1.** *If a rule is an AERM with rate* $\epsilon_{\text{erm}}(m)$ *and stable (under any of our definitions) with rate* $\epsilon_{\text{stable}}(m)$ *under* $\mathcal{D}$*, then it is consistent and generalizes under* $\mathcal{D}$ *with rates*

$$\epsilon_{\text{cons}}(m)\leq 3\epsilon_{\text{erm}}(m)+\epsilon_{\text{stable}}(m+1)+\tfrac{2B}{m+1}$$
$$\epsilon_{\text{gen}}(m)\leq 4\epsilon_{\text{erm}}(m)+\epsilon_{\text{stable}}(m+1)+\tfrac{6B}{\sqrt{m}}$$

**Corollary 4.2.** *If a rule is universally an AERM and stable then it is universally consistent and generalizing.*

Seeking a converse to the above, we first note that it is not possible to obtain a converse for each distribution $\mathcal{D}$ separately, i.e. to Theorem 4.1. In Example 7.6, we show a specific learning problem and distribution $\mathcal{D}$ in which the ERM (in fact, any AERM) is consistent, but not stable, even under our weakest notion of stability.

However, we are able to obtain a converse to Corollary 4.2. That is, establish that a *universally* consistent ERM, or even AERM, must also be stable. For exact ERMs we have:

**Theorem 4.3.** *For an ERM the following are equivalent:*

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

Recall that for a symmetric rule, LOO stability and all-i-LOO stability are equivalent, and so consistency or generalization of a symmetric ERM (the typical case) also imply all-i-LOO stability.

Theorem 4.3 only guarantees LOO stability as a necessary condition for consistency. Example 7.3 (adapted from [5]) establishes that we cannot strengthen the condition to uniform-LOO stability, or any stronger definition: there exists a learning problem for which the ERM is universally consistent, but not uniform-LOO stable.

For AERMs, we obtain a weaker converse, ensuring only on-average-LOO stability:

**Theorem 4.4.** *For an AERM, the following are equivalent:*

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

On-average-LOO stability is strictly weaker then LOO stability, but this is the best that can be assured. In Example 7.4 we present a learning problem and an AERM that is universally consistent, but is not LOO stable.

The exact rate conversions of Theorems 4.3 and 4.4 are specified in the corresponding proofs (Section 6), and are all polynomial. In particular, an $\epsilon_{\text{cons}}$-universal consistent $\epsilon_{\text{erm}}$-AERM is on-average-LOO stable with rate

$$\epsilon_{\text{stable}}(m)\leq 3\epsilon_{\text{erm}}(m-1)+3\epsilon_{\text{cons}}((m-1)^{1/4})+\tfrac{6B}{\sqrt{m-1}}.$$

The above results apply only to AERMs, for which we also see that universal consistency and generalization are equivalent. Next we show that if in fact we seek universal consistency and generalization, then we must consider only AERMs:

**Theorem 4.5.** *If a rule* $\mathbf{A}$ *is universally consistent with rate* $\epsilon_{\text{cons}}(m)$ *and generalizing with rate* $\epsilon_{\text{gen}}(m)$*, then it is universally an AERM with rate*

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

Combining theorems 4.4 and 4.5, we get that the existence of a universally on-average-LOO stable AERM is a necessary (and sufficient) condition for existence of some universally consistent and generalizing rule. As we show in Example 7.7, there might still be a universally consistent learning rule (hence the problem is learnable by our definition) that is *not* stable even by our weakest definition (and is not an AERM nor generalizing). Nevertheless, any universally consistent learning rule can be transformed into a universally consistent and generalizing learning rule (Lemma 6.11). Thus by Theorems 4.5 and 4.4 this rule must also be a stable AERM, establishing:

**Theorem 4.6.** *A learning problem is learnable if and only if there exists a universally on-average-LOO stable AERM.*

In particular, if there exists a $\epsilon_{\text{cons}}$-universally consistent rule, then there exists a rule that is $\epsilon_{\text{stable}}$-on-average-LOO stable and $\epsilon_{\text{erm}}$-AERM where:

$$\begin{aligned}\epsilon_{\text{erm}}(m) &= 3\epsilon_{\text{cons}}(m^{1/4}) + \tfrac{7B}{\sqrt{m}} \ , \\ \epsilon_{\text{stable}}(m) &= 6\epsilon_{\text{cons}}((m-1)^{1/4}) + \tfrac{19B}{\sqrt{m-1}}\end{aligned} \tag{7}$$

# 5 Comparison with Prior Work

## 5.1 Theorem 4.1 and Corollary 4.2

The consistency implication in Theorem 4.1 (specifically, that all-LOO stability of an AERM implies consistency) was established by Mukherjee *et al* [7, Theorem 3.15].

As for the generalization guarantee, Rakhlin *et al* [8] prove that for ERM, average-LOO stability is equivalent to generalization. For more general learning rules, [2] attempted to show that all-i-LOO stability implies generalization. However, Mukherjee *et al* [7] (in remark 3, pg. 173) provide a simple counterexample and note that the proof of [2] is wrong, and in fact all-i-LOO stability alone is not enough to ensure generalization. To correct this, Mukherjee *et al* [7] introduced an additional condition, referred to as Eloo$_{\text{err}}$ stability, which together with all-i-LOO stability ensures generalization. For AERMs, they use arguments specific to supervised learning, arguing that universal consistency implies uniform convergence, and establish generalization only via this route. And so, Mukherjee *et al* obtain a version of Corollary 4.2 that is specific to supervised learning.

In summary, comparing Theorem 4.1 to previous work, our results extend the generalization guarantee also to AERMs in the general learning setting.

Rakhlin *et al* [8] also show that the replacement (rather then deletion) version stability implies generalization for any rule (even non-AERM), and hence consistency for AERMs. Recall that the deletion and replacement version are not equivalent. We are not aware of strong converses for the replacement variant.

## 5.2 Converse Results

Mukherjee *et al* [7] argue that all-i-LOO stability of the ERM (in fact, of any AERM) is also necessary for ERM universal consistency and thus learnability. However, their arguments are specific to supervised learning, and establish stability only via uniform convergence of $F_S(\mathbf{h})$ to $F(\mathbf{h})$, as discussed in the introduction. As we now know, in the general learning setting, ERM consistency is not equivalent to this uniform convergence, and furthermore, there might be a non-ERM universally consistent rule even though the ERM is not universally consistent. Therefore, our results here apply to the general learning setting and do not use uniform convergence arguments.

For an ERM, Rakhlin *et al* [8] show that generalization is equivalent to on-average-LOO stability, for any distribution and without resorting to uniform convergence arguments. This provides a partial converse to Theorem 4.1. However, our results extend to AERM's as well and more importantly, provide a converse to AERM consistency rather than just generalization. This distinction between consistency and generalization is important, as there are situations with consistent but not stable AERM's (Example 7.6), or even universally consistent learning rules which are not stable, generalizing nor AERM's (Example 7.7).

Another converse result that does not use uniform convergence arguments, but is specific only to the *realizable* binary learning setting was given by Kutin and Niyogi [5]. They show that in this setting, for any distribution $\mathcal{D}$, all-i-LOO stability of the ERM under $\mathcal{D}$ is necessary for ERM consistency under $\mathcal{D}$. This is a much stronger form of converse as it applies to any specific distribution separately, rather then requiring universal consistency. However, not only is it specific to supervised learning, but further requires the distribution be realizable (i.e. zero error is achievable). As we show in Example 7.6, a distribution-specific converse is not possible in the general setting.

All the papers cited above focus on symmetric learning rules where all-i-LOO stability is equivalent to LOO stability. We prefer not to limit our attention to symmetric rules, and instead use LOO stability.

# 6 Detailed Results and Proofs

We first establish that for AERMs, on-average-LOO stability and generalization are equivalent, and that for ERMs the equivalence extends also to LOO stability. This extends the work of Rakhlin *et al* [8] from ERMs to AERMs, and with somewhat better rate conversions.

## 6.1 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 $\mathbf{A}$ **on-average generalizes** with rate $\epsilon_{\text{oag}}(m)$ under distribution $\mathcal{D}$ if for all $m$,

$$\left|\mathbb{E}_{S\sim\mathcal{D}^m}\left[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right]\right| \leq \epsilon_{\text{oag}}(m). \tag{8}$$

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 on-average stability, establishing the equivalence between generalization and on-average stability (for AERMs).

**Lemma 6.1 (For AERMs: on-average generalization ⇔ on-average stability).** *Let* $\mathbf{A}$ *be AERM with rate* $\epsilon_{\text{erm}}(m)$ *under* $\mathcal{D}$*. If* $\mathbf{A}$ *is on-average generalizing with rate* $\epsilon_{\text{oag}}(m)$ *then it is on-average LOO stable with rate* $\epsilon_{\text{oag}}(m-1) + 2\epsilon_{\text{erm}}(m-1) + 2B/m$*. If* $\mathbf{A}$ *is on-average LOO stable with rate* $\epsilon_{\text{stable}}(m)$ *then it is on-average generalizing with rate* $\epsilon_{\text{stable}}(m+1) + 2\epsilon_{\text{erm}}(m) + 2B/m$*.*

*Proof.* For the ERMs of $S$ and $S^{\setminus i}$ we have $\left|F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_S(\hat{\mathbf{h}}_S)\right| \leq \frac{2B}{m}$, and so since $A$ is AERM:

$$\mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F_{S^{\setminus i}}(\mathbf{A}(S^{\setminus i}))\right|\right] \leq 2\epsilon_{\text{erm}}(m-1) + \tfrac{2B}{m} \tag{9}$$

**generalization ⇒ stability** Applying (8) to $S^{\setminus i}$ and combining with (9) we have $\left|\mathbb{E}\left[F(\mathbf{A}(S^{\setminus i})) - F_S(\mathbf{A}(S))\right]\right| \le \epsilon_{\mathrm{oag}}(m-1) + 2\epsilon_{\mathrm{erm}}(m-1) + 2B/m$, which does not actually depend on $i$, hence:

$$\begin{aligned}
&\epsilon_{\mathrm{oag}}(m-1) + 2\epsilon_{\mathrm{erm}}(m-1) + 2B/m \\
&\ge \left|\mathbb{E}\left[F(\mathbf{A}(S^{\setminus i})) - F_S(\mathbf{A}(S))\right]\right| \qquad (10)\\
&= \left|\mathbb{E}_i\left[\mathbb{E}\left[F(\mathbf{A}(S^{\setminus i})) - F_S(\mathbf{A}(S))\right]\right]\right| \\
&= \left|\mathbb{E}_{S^{\setminus i}, z_i}\left[\mathbb{E}_i\left[f(\mathbf{A}(S^{\setminus i}), z_i)\right]\right] - \mathbb{E}_S\left[f(\mathbf{A}(S), z_i)\right]\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}_i\left[f(\mathbf{A}(S^{\setminus i}), z_i) - f(\mathbf{A}(S), z_i)\right]\right]\right| \qquad (11)
\end{aligned}$$

which establishes on-average stability.

**stability ⇒ generalization** Bounding (11) by $\epsilon_{\mathrm{stable}}(m)$ and working back we get that (10) is also bounded by $\epsilon_{\mathrm{stable}}(m)$. Combined with (9) we get $\left|\mathbb{E}\left[F(\mathbf{A}(S^{\setminus i})) - F_{S^{\setminus i}}(\mathbf{A}(S^{\setminus i}))\right]\right| \le \epsilon_{\mathrm{stable}}(m) + 2\epsilon_{\mathrm{oag}}(m-1) + 2B/m$ which establishes on-average generalization. □

**Lemma 6.2 (AERM + on-average generalization ⇒ generalization).** *If $\mathbf{A}$ is an AERM with rate $\epsilon_{\mathrm{erm}}(m)$ and on-average generalizes with rate $\epsilon_{\mathrm{oag}}(m)$ under $\mathcal{D}$, then $\mathbf{A}$ generalizes with rate $\epsilon_{\mathrm{oag}}(m) + 2\epsilon_{\mathrm{erm}}(m) + \frac{2B}{\sqrt{m}}$ under $\mathcal{D}$.*

*Proof.* Using respective optimalities of $\hat{\mathbf{h}}_S$ and $\mathbf{h}^\star$ we can bound:

$$\begin{aligned}
&F_S(\mathbf{A}(S)) - F(\mathbf{A}(S)) \\
&= F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S) + F_S(\hat{\mathbf{h}}_S) - F_S(\mathbf{h}^\star) \\
&\quad + F_S(\mathbf{h}^\star) - F(\mathbf{h}^\star) + F(\mathbf{h}^\star) - F(\mathbf{A}(S)) \\
&\le F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S) + F_S(\mathbf{h}^\star) - F(\mathbf{h}^\star) = Y \qquad (12)
\end{aligned}$$

Where the final equality defines a new random variable $Y$. By Lemma 6.3 and the AERM guarantee we have $\mathbb{E}\left[|Y|\right] \le \epsilon_{\mathrm{erm}}(m) + B/\sqrt{m})$. From Lemma 6.4 we can conclude that

$$\begin{aligned}
&\mathbb{E}\left[|F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))|\right] \\
&\le |\mathbb{E}\left[F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))\right]| + 2\mathbb{E}\left[|Y|\right] \\
&\le \epsilon_{\mathrm{oag}}(m) + 2\epsilon_{\mathrm{erm}}(m) + \tfrac{2B}{\sqrt{m}}.
\end{aligned}$$

□

**Utility Lemma 6.3.** *For i.i.d. $X_i$, $|X_i| \le B$ and $X = \frac{1}{m}\sum_{i=1}^m X_i$ we have $\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right] \le B/\sqrt{m}$.*

*Proof.* $\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right] \le \sqrt{\mathrm{Var}[X]} = \sqrt{\mathrm{Var}[X_i]/m} \le B/\sqrt{m}$. □

**Utility Lemma 6.4.** *Let $X, Y$ be random variables s.t. $X \le Y$ almost surely. Then $\mathbb{E}\left[|X|\right] \le |\mathbb{E}\left[X\right]| + 2\mathbb{E}\left[|Y|\right]$.*

*Proof.* Denote $a_+ = \max(0, a)$ and observe that $X \le Y$ implies $X_+ \le Y_+$ (this holds when both have the same sign, and when $X \le 0 \le Y$, while $Y < 0 < X$ is not possible). We therefor have $\mathbb{E}\left[X_+\right] \le \mathbb{E}\left[Y_+\right] \le \mathbb{E}\left[|Y|\right]$. Also note that $|X| = 2X_+ - X$. We can now calculate: $\mathbb{E}\left[|X|\right] = \mathbb{E}\left[2X_+ - X\right] = 2\mathbb{E}\left[X_+\right] - \mathbb{E}\left[X\right] \le 2\mathbb{E}\left[|Y|\right] + |\mathbb{E}\left[X\right]|$. □

For exact ERM, we get a stronger equivalence:

**Lemma 6.5 (ERM+on-average-LOO⇒LOO stable).** *If an exact ERM $\mathbf{A}$ is on-average-LOO stable with rate $\epsilon_{\mathrm{stable}}(m)$ under $\mathcal{D}$, then it is also LOO stable under $\mathcal{D}$ with the same rate.*

*Proof.* By optimality of $\hat{\mathbf{h}}_S = A(S)$:

$$\begin{aligned}
f(\hat{\mathbf{h}}_{S^{\setminus i}}, z_i) - f(\hat{\mathbf{h}}_S, z_i) &= F_S(\hat{\mathbf{h}}_{S^{\setminus i}}) - F_S(\hat{\mathbf{h}}_S) \\
&\quad + F_{S^{\setminus i}}(\hat{\mathbf{h}}_S) - F_{S^{\setminus i}}(\hat{\mathbf{h}}_{S^{\setminus i}}) \ge 0. \qquad (13)
\end{aligned}$$

Then using on-average-LOO stability:

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left|f(\hat{\mathbf{h}}_{S^{\setminus i}}, z_i) - f(\hat{\mathbf{h}}_S, z_i)\right|\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left(f(\hat{\mathbf{h}}_{S^{\setminus i}}, z_i) - f(\hat{\mathbf{h}}_S, z_i)\right)\right] \le \epsilon_{\mathrm{stable}}(m)
\end{aligned}$$

□

Lemma 6.5 can be extended also to AERMs with rate $o(\frac{1}{n})$. However, for AERMs with a slower rate, or at least with rate $\Omega(\frac{1}{\sqrt{n}})$, Example 7.4 establishes that this stronger converse is not possible.

We have now **established the stability↔generalization parts of Theorems 4.1, 4.3 and 4.4** (in fact, even a slightly stronger converse than in Theorems 4.3 and 4.4, as it does not require universality).

### 6.2 A Sufficient Condition for Consistency

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

**Lemma 6.6 (AERM+generalization⇒consistency).** *If $\mathbf{A}$ is AERM with rate $\epsilon_{\mathrm{erm}}(m)$ and it on-average generalizes with rate $\epsilon_{\mathrm{oag}}(m)$ under $\mathcal{D}$ then it is consistent with rate $\epsilon_{\mathrm{oag}}(m) + \epsilon_{\mathrm{erm}}(m)$ under $\mathcal{D}$.*

*Proof.*

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{A}(S)) - F(\mathbf{h}^\star)\right] = \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{h}^\star)\right] \\
&= \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right] + \mathbb{E}\left[F_S(\mathbf{A}(S)) - F_S(\mathbf{h}^\star)\right] \\
&\le \mathbb{E}\left[F(\mathbf{A}(S)) - F_S(\mathbf{A}(S))\right] + \mathbb{E}\left[F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right] \\
&\le \epsilon_{\mathrm{gen}}(m) + \epsilon_{\mathrm{erm}}(m)
\end{aligned}$$

□

Combined with the results of Section 6.1, this completes the **proof of Theorem 4.1** and the **stability→consistency and generalization→consistency parts of Theorems 4.3 and 4.4**.

### 6.3 Converse Direction

Lemma 6.1 already provides a converse result, establishing that stability is necessary for generalization. However, in order to establish that stability is also necessary for universal consistency we must prove that universal consistency of an AERM implies universal generalization. Note that consistency under a specific distribution for an AERM does *not* imply generalization nor stability (Example 7.6). We must instead rely on universal consistency. The main tool we use is the following lemma:

**Lemma 6.7** (**Main Converse Lemma).** *If a problem is learnable, i.e. there exists a universally consistent rule* $\mathbf{A}$ *with rate* $\epsilon_{\text{cons}}(m)$*, then under any distribution,*

$$\mathbb{E}\left[\left|F_S(\hat{\mathbf{h}}_S) - F(\mathbf{h}^\star)\right|\right] \le \epsilon_{\text{emp}}(m) \qquad \textit{where}$$

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

*for any sequence* $m'$ *is such that* $m' \to \infty$ *and* $m' = o(\sqrt{m})$*.*

*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 $S' = \{z_{I_i}\}_{i=1}^{m'}$, i.e. 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"):

$$\Pr\left(I \text{ has duplicates}\right) \le \frac{\sum_{i=1}^{m'}(i-1)}{m} \le \frac{m'^2}{2m} \tag{14}$$

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

$$\mathbb{E}\left[\left|F(\mathbf{A}(S')) - F(\mathbf{h}^\star)\right| \mid \text{no dups}\right] \le \epsilon_{\text{cons}}(m') \tag{15}$$

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

$$\mathbb{E}_{S'}\left[\left|F_S(A(S')) - F_S(\hat{\mathbf{h}}_S)\right|\right] \le \epsilon_{\text{cons}}(m') \tag{16}$$

Conditioned on having no duplications in $I$, $S \setminus S'$ (i.e. those samples in $S$ not chosen by $I$) is independent of $S'$, and $|S \setminus S'| = m - m'$, and so by Lemma 6.3:

$$\mathbb{E}_S\left[\left|F(\mathbf{A}(S')) - F_{S\setminus S'}(\mathbf{A}(S'))\right|\right] \le \frac{B}{\sqrt{m-m'}} \tag{17}$$

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

$$\left|F_S(A(S')) - F_{S\setminus S'}(A(S'))\right| \le \frac{2Bm'}{m} \tag{18}$$

Combining (15),(16),(17) and (18), accounting for a maximal discrepancy of $B$ when we do have duplicates, and assuming $2 \le m' \le m/2$, we get the desired bound. □

In the supervised learning setting, Lemma 6.7 is just an immediate consequence of learnability being equivalent to consistency and generalization of the ERM. However, the Lemma applies also in the General Setting, where universal consistency might be achieved only by a non-ERM. The Lemma states that if a problem is learnable, even though the ERM might not be consistent (as in, e.g. Example 7.2), the empirical error achieved by the ERM is in fact an asymptotically unbiased estimator of $F(\mathbf{h}^\star)$.

Equipped with Lemma 6.7, we are now ready to show that universal consistency of an AERM implies 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 6.7 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 6.8 (learnable+AERM+consistent⇒generalizing).** *If Lemma 6.7 holds with rate* $\epsilon_{\text{emp}}(m)$*, and* $\mathbf{A}$ *is an* $\epsilon_{\text{erm}}$*-AERM and* $\epsilon_{\text{cons}}$*-consistent under* $\mathcal{D}$*, then it is generalizing under* $\mathcal{D}$ *with rate* $\epsilon_{\text{emp}}(m) + \epsilon_{\text{erm}}(m) + \epsilon_{\text{cons}}(m)$*.*

*Proof.*

$$\begin{aligned} \mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))\right|\right] &\le \mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right|\right] \\ + \mathbb{E}\left[\left|F(\mathbf{h}^\star) - F(\mathbf{A}(S))\right|\right] &+ \mathbb{E}\left[\left|F_S(\hat{\mathbf{h}}_S) - F(\mathbf{h}^\star)\right|\right] \\ &\le \epsilon_{\text{erm}}(m) + \epsilon_{\text{cons}}(m) + \epsilon_{\text{emp}}(m) \quad \square \end{aligned}$$

**Lemma 6.9 (learnable+consistent+generalizing⇒AERM).** *If Lemma 6.7 holds with rate* $\epsilon_{\text{emp}}(m)$*, and* $\mathbf{A}$ *is* $\epsilon_{\text{cons}}$*-consistent and* $\epsilon_{\text{gen}}$*-generalizing under* $\mathcal{D}$*, then it is AERM under* $\mathcal{D}$ *with rate* $\epsilon_{\text{emp}}(m) + \epsilon_{\text{gen}}(m) + \epsilon_{\text{cons}}(m)$*.*

*Proof.*

$$\begin{aligned} \mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F_S(\hat{\mathbf{h}}_S)\right|\right] &\le \mathbb{E}\left[\left|F_S(\mathbf{A}(S)) - F(\mathbf{A}(S))\right|\right] \\ + \mathbb{E}\left[\left|F(\mathbf{A}(S)) - F(\mathbf{h}^\star)\right|\right] &+ \mathbb{E}\left[\left|F(\mathbf{h}^\star) - F_S(\hat{\mathbf{h}}_S)\right|\right] \\ &\le \epsilon_{\text{gen}}(m) + \epsilon_{\text{cons}}(m) + \epsilon_{\text{emp}}(m) \quad \square \end{aligned}$$

Lemma 6.8 establishes that universal consistency of an AERM implies universal generalization, and thus **completes the proof of Theorems 4.3 and 4.4**. Lemma 6.9 **establishes Theorem 4.5**. To get the rates in 4, we use $m' = m^{1/4}$ in Lemma 6.7.

Lemmas 6.6, 6.8 and 6.9 together establish an interesting relationship:

**Corollary 6.10.** *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:

- The ERM in Example 7.2 is neither consistent nor generalizing, despite the problem being learnable.

- Example 7.7 demonstrates 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 and regression 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 6.10 can also provide a certificate of non-learnability. E.g. for the problem in Example 7.6 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.

### 6.4 Existence of a Stable Rule

Theorems 4.5 and 4.4, which we just completed proving, already establish that for AERMs, universal consistency is equivalent to universal on-average-LOO stability. Existence of a universally on-average-LOO 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 on-average-LOO stable by Theorem 4.4.

We actually show how to transform a consistent rule to a consistent and generalizing rule. If this rule is universally consistent, then by Lemma 6.9 we can then conclude it must an AERM, and by 6.1 that it must be on-average-LOO stable.

**Lemma 6.11.** *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* $\epsilon_{\text{cons}}$*-consistent under* $\mathcal{D}$ *then* $\mathbf{A}'$ *is* $\epsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor)$ *consistent under* $\mathcal{D}$*.*

*Proof.* For a sample $S$ of size $m$, let $S'$ be a sub-sample consisting of the first $\lfloor\sqrt{m}\rfloor$ observation in $S$. Define $\mathbf{A}'(S) = \mathbf{A}(S')$. That is, $\mathbf{A}'$ applies $A$ to only $\lfloor\sqrt{m}\rfloor$ of the observation in $S$.

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

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

The first term can be bounded by $2B/\lfloor\sqrt{m}\rfloor$. As for the second term, $S \setminus S'$ is statistically independent of $S'$ and so we can use Lemma 6.3 to bound its expected magnitude to obtain:

$$\mathbb{E}\left[|F_S(\mathbf{A}(S')) - F(\mathbf{A}(S'))|\right] \le \frac{2B}{\lfloor\sqrt{m}\rfloor} + (1 - \frac{1}{\lfloor\sqrt{m}\rfloor})\frac{B}{\sqrt{m - \lfloor\sqrt{m}\rfloor}} \le \frac{3B}{\sqrt{m}} \qquad (19)$$

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

$$\mathbb{E}\left[F(\mathbf{A}'(S)) - \inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})\right] \le \mathbb{E}\left[F(\mathbf{A}(S')) - \inf_{\mathbf{h}\in\mathcal{H}} F(\mathbf{h})\right] \le \epsilon_{\text{cons}}(\lfloor\sqrt{m}\rfloor) \quad \square$$

**Proof of Converse in Theorem 4.6** If there exists a universally consistent rule with rate $\epsilon_{\text{cons}}(m)$, by Lemma 6.11 there exists $\mathbf{A}'$ which is universally consistent and generalizing. Choosing $m' = m^{1/4}$ in Lemma 6.7 and applying Lemmas 6.9 and 6.1 we get the rates specified in (7). □

**Remark** We can strengthen the above theorem to show existence of an on-average-LOO stable, always AERM (ie. a rule which for every sample approximately minimizes $F_S(\mathbf{h})$). The new learning rule for this purpose chooses the hypothesis returned by the original rule whenever empirical risk is small and chooses an ERM otherwise. The proof is completed via Markov inequality to bound the probability that we don't choose the hypothesis returned by the original learning rule.

## 7 Examples

Our first example (taken from [9]) shows that uniform convergence is *not* necessary for ERM consistency. I.e. universal ERM consistency holds without uniform convergence. Of course, this can also happen in "trivial" settings where there is one hypothesis $\mathbf{h}_0$ which dominates all other hypothesis (i.e. $f(\mathbf{h}_0, z) < f(\mathbf{h}, z)$ for all $z$ and all $\mathbf{h} \neq \mathbf{h}_0$) [10]. However, the example below demonstrates a non-trivial situation with ERM universal consistency but no uniform convergence: there is no dominating hypothesis, and finding the optimal hypothesis does require learning. In particular, unlike "trivial" problems with a dominating hypothesis, in the example below there is not even local uniform convergence. I.e. there is no uniform convergence even among hypotheses that are close to being population optimal.

**Example 7.1.** *There exists a learning problem for which any ERM is universally consistent, but the empirical risks do not converge uniformly to their expectations.*

*Proof.* Consider a convex stochastic optimization problem given by:

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

where $\mathbf{w}$ is the hypothesis, $\mathbf{w}, \mathbf{x}$ are elements in a unit ball around the origin of a Hilbert space with a countably infinite orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$, and $\alpha$ is an infinite binary sequence. $\alpha[i]$ is the i-th coordinate of $\alpha$, $\mathbf{w}[i] := \langle \mathbf{w}, \mathbf{e}_i \rangle$, and $\mathbf{x}[i]$ is defined similarly. In our other submission [9], we show that the ERM is stable, hence consistent. However, when $x = 0$ a.s. and $\alpha$ is i.i.d. uniform, there is no uniform convergence, not even locally. To see why, note that for a random sample $S$ of any finite size, with probability one there exists an "excluded" basis vector $e_j$ such that $\alpha_i[j] = 0$ for all $(\mathbf{x}_i, \alpha_i) \in S$. For any $t > 0$, we have $F(t\mathbf{e}_j) - F_S(t\mathbf{e}_j) \geq t^2$, regardless of the sample size. Setting $t = 1$ establishes $\sup_{\mathbf{w}} |F(\mathbf{w}) - F_S(\mathbf{w})| \geq 1$ even as $m \to \infty$, and so there is no uniform convergence. Choosing $t$ arbitrarily small, we see that even when $F(t\mathbf{e}_j)$ is close to optimal, the deviations $|F(\mathbf{w}) - F_S(\mathbf{w})|$ still do not converge to zero as $m \to \infty$. □

Perhaps more surprisingly, the next example (also taken from [9]) shows that in the general setting, learnability might require using a non-ERM.

**Example 7.2.** *There exists a learning problem with a universally consistent learning rule, but for which no ERM is universally consistent.*

*Proof.* Consider the same hypothesis space and sample space as before, with:

$$f(\mathbf{w}, z) = \frac{\|\alpha * (\mathbf{w} - \mathbf{x})\|}{2} + \frac{\epsilon}{2}\sum_{i=1}^{\infty} 2^{-i}(w_i - 1)^2 ,$$

where $\epsilon = 0.01$. When $\mathbf{x} = 0$ a.s. and $\alpha$ is i.i.d. uniform, then the ERM must have $\|\hat{\mathbf{w}}\| = 1$. To see why, note that for an excluded $\mathbf{e}_j$ (which exists a.s.) increasing $\mathbf{w}[j]$ towards one decreases the objective. But since $\|\hat{\mathbf{w}}\| = 1$, we have $F(\hat{\mathbf{w}}) \geq 1/2$, while $\inf_{\mathbf{w}} F(\mathbf{w}) \leq F(0) = \epsilon$, and so $F(\hat{\mathbf{w}}) \not\to \inf_{\mathbf{w}} F(\mathbf{w})$.

On the other hand, $\mathbf{A}(S) = \arg\min F_S(\mathbf{w}) + \frac{20}{\sqrt{m}} \|\mathbf{w}\|^2$ is a uniformly-LOO stable AERM and hence by Theorem 4.1 universally consistent. □

In the next three examples, we show that in a certain sense, Theorem 4.3 and Theorem 4.4 cannot be improved with stronger stability notions. Viewed differently, they also constitute separation results between our various stability notions, and show which are strictly stronger than the other. Example 7.4 also demonstrates the gap between supervised learning and a general learning setting, by presenting a learning problem and an AERM that is universally consistent, but not LOO stable.

**Example 7.3.** *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 [5]. Consider the hypothesis space $\{0, 1\}$, the instance space $\{0, 1\}$, and the objective function $f(h, z) = |h - z|$.

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

*Proof.* Let the instance space, hypothesis space and objective function be as in Example 7.3. 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 $\mathrm{Parity}(S) = (z_1 + \ldots z_m) \bmod 2$.

This learning rule is an AERM, with $\epsilon_{\mathrm{erm}}(m) = \sqrt{2\log(4)/m}$. 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 $\epsilon_{\mathrm{gen}}(m) = \sqrt{\log(4/\delta)/2m}$), and by Theorem 4.4, this implies that the learning rule is also universally consistent and 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 $i$,

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

This implies that

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left|\Big(f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i)\Big)\right|\right] \tag{20}$$

$$\geq \frac{1}{2}\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\{z_i=1\}}\middle|\sqrt{\frac{\log(4)}{2m}} \geq \left|\sum_{i=1}^{m}\frac{\mathbb{1}_{\{z_i=1\}}}{m} - \frac{1}{2}\right|\right]$$

$$\geq \frac{1}{2}\left(\frac{1}{2} - \sqrt{\frac{\log(4)}{2m}}\right) \longrightarrow \frac{1}{4} \,,$$

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 average-LOO stability cannot be replaced even by weaker stability notions than LOO stability. For instance, a natural stability notion intermediate between average-LOO stability and LOO stability is

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\Big(f(\mathbf{A}(S^{\backslash i}); z_i) - f(\mathbf{A}(S); z_i)\Big)\right|\right], \tag{21}$$

where the absolute value is now over the entire sum, but inside the expectation. In the example used in the proof, (21) is still lower bounded by (20), which does not converge to zero with the sample size.

**Example 7.5.** *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 $\epsilon_{\mathrm{erm}}(m) = 1/m$. To see universal consistency, let $\Pr(z_1) = p$. 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 $\mathbb{E}_S\left[F(\mathbf{A}(S))\right] \leq p(1 - p)^{m-2}$, which can be easily verified to be at most $1/(m - 1)$, so the learning rule is consistent with rate $\epsilon_{\mathrm{cons}}(m) \leq 1/(m - 1)$. 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 $\epsilon_{\mathrm{stable}}(m) \leq 2/m$ (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, $|f(\mathbf{A}(S^{\backslash 1}), z_1) - f(\mathbf{A}(S), z_1)| = 1$ with probability $1$, so it obviously cannot be all-i-LOO-stable with respect to $i = 1$. □

Next we show that for *specific* distributions, even ERM consistency does not imply even our weakest notion of stability.

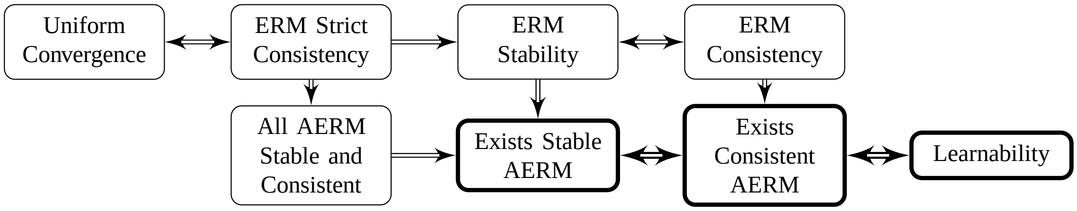


Figure 1: Implications of various properties of learning problems. Consistency refers to univeral consistency and stability refers to univeral on-average-LOO stability.

**Example 7.6.** *There exists a learning problem and a distribution on the instance space, such that the ERM (or any AERM) is consistent but is not average-LOO stable.*

*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 $f(h,z) = \mathbb{1}_{\{z \notin h\}})$. Consider any continuous distribution on the instance space. Since the underlying distribution $\mathcal{D}$ is continuous, we have $F(h) = 1$ 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 $F_S(\hat{h}_S) = 0$, so any AERM cannot generalize, or even on-average-generalize (by Lemma 6.2), hence cannot be average-LOO stable (by Lemma 6.1). □

Finally, the following example shows that while learnability is equivalent to the existence of stable and consistent AERM's (Theorem 4.4 and Theorem 4.6), there might still exist other learning rules, which are neither of the above.

**Example 7.7.** *There exists a learning problem with a universally consistent learning rule, which is not average-LOO 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 $f(h,z) = \mathbb{1}_{\{z \in h\}}$. 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.

This learning rule is not an AERM, nor does it necessarily generalize or is average-LOO stable. Consider any continuous distribution on $[0,1]$. The learning rule always returns a countable 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, $f(\mathbf{A}(S), z_i) = 0$ while $f(\mathbf{A}(S^{\setminus i}0, z_i) = 1$ with probability 1, so it is not average-LOO 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 , $F(S) = 0 = \inf_h F(h)$. 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 [6, Theorem 3], we have that for any $\delta$, 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 event 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 least $1 - O(\log(m/\delta)/\sqrt{m})$, 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 $\delta$, it is easy to see that the learning rule is universally consistent. □

## 8 Discussion

In the familiar setting of supervised classification or regression, the question of learnability is reduced to that of uniform convergence of empirical risks to their expectation, and in turn to finiteness of the fat-shattering dimension [1]. Furthermore, due to the equivalence of learnability and uniform convergence, there is no need to look beyond the ERM.

We recently showed [9] that the situation in the General Learning Setting is substantially more complex. Universal ERM consistency might *not* be equivalent to uniform convergence, and furthermore, learnability might be possible only with a non-ERM. We are therefore in need of a new understanding of the question of learnability that applies more broadly then just to supervised classification and regression.

In studying learnability in the General Setting, Vapnik [10] focuses solely on empirical risk minimization, which we now know is not sufficient for understanding learnability (e.g. Example 7.2). 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 now know that even in rather non-trivial problems (e.g. Example 7.1 taken from [9]), where the ERM is consistent and generalizes, strict consistency does not hold. Furthermore, Example 7.1 also demonstrates that ERM stability guarantees ERM consistency, but *not* strict consistency, perhaps giving another indication that strict consistency might be too strict (this and other relationships are depicted in Figure 1).

In Examples 7.1 and 7.2 we see that stability is a strictly more general sufficient condition for learnability. This makes stability an appealing candidate for understanding learnability in the more general setting.

Indeed, we show that stability is not only sufficient, but is also necessary for learning, even in the General Learning Setting. A previous such characterization was based on uniform convergence and thus applied only to supervised classification and regression [7]. Extending the characterization beyond these settings is particularly interesting, since for supervised classification and regression the question of learnability is already essentially solved. Extending the characterization, without relying on uniform convergence, 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 1).

In studying the question of learnability and its relation to stability, we encounter several differences between this more general setting, and settings such as supervised classification and regression 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 and regression, 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 and regression, 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 7.7 demonstrates. However, Theorem 4.5 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 6.10, which concerns 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 and regression, if a problem is learnable then generalization always holds (for any rule), and so universal consistency and AERM imply each other.
- In supervised classification and regression, 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 6.10 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 7.6).
- In the General Setting, universal consistency of an AERM only guarantees on-average-LOO stability, but not LOO stability as in the supervised classification setting [7]. As we show in Example 7.4, this is a real difference and not merely a deficiency of our proofs.

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.

Throughout the paper we ignored the issue of getting high-confidence concentration guarantees. 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 and using Markov's inequality we can translate a rate of the form $\mathbb{E}\left[|\cdots|\right] \leq \epsilon(m)$ to a "low confidence" guarantee $\Pr(|\cdots| > \epsilon(m)/\delta) \leq \delta$. Can we also obtain exponential concentration results of the form $\Pr(|\cdots| > \epsilon(m)\text{polylog}(1/\delta)) \leq \delta$ ? It is possible to construct examples in the General Setting in which convergence in expectation of the stability does *not* imply exponential concentration of consistency and generalization. Is it possible to show that exponential concentration of stability is equivalent to exponential concentration of consistency and generalization?

We showed that existence of an average-LOO stable AERM is necessary and sufficient for learnability (Theorem 4.6). Although specific AERMs might be universally consistent and generalizing without being LOO stable (Example 7.4), it might still be possible to show that for a learnable problem, there always exists some LOO stable AERM. This would tighten our converse result and establish existence of a LOO stable AERM as equivalent to learnability.

Even existence of a LOO stable AERM 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.

## References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[2] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

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

[4] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *Proc. of COLT 5*, 1992.

[5] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proc. of UAI 18*, 2002.

[6] D.A. McAllester and R.E. Schapire. On the convergence rate of good-turing estimators. In *Proc. of COLT 13*, 2000.

[7] S. Mukherjee, P. Niyogi, T. Poggio, and R. M. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25(1-3):161–193, 2006.

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

[9] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of COLT 22*, 2009.

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

# A Replacement Stability

We used the Leave-One-Out version of stabilities throughout the paper, however many of the results hold when we use the replacement versions instead. Here we briefly survey the differences in the main results as they apply to replacement-based stabilities.

Let $S^{(i)}$ denote the sample $S$ with $z_i$ replaced by some other $z'_i$ drawn from the same unknown distribution $\mathcal{D}$.

**Definition 5.** *A rule* $\mathbf{A}$ *is* **uniform-RO stable** *with rate* $\epsilon_{\text{stable}}(m)$ *if for all samples* $S$ *of* $m$ *points and* $\forall z', z'_1, ..., z'_m \in \mathbb{Z}$ *:*

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

**Definition 6.** *A rule* $\mathbf{A}$ *is* **on-average-RO stable** *with rate* $\epsilon_{\text{stable}}(m)$ *under distributions* $\mathcal{D}$ *if*

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

With the above definitions replacing uniform-LOO stability and on-average-LOO stability respectively, all theorems in Section 4 other than Theorem 4.3 hold (i.e. Theorem 4.1, Corollary 4.2, Theorem 4.4 and Theorem 4.6).

We do not know how to obtain a replacement-variant of Theorem 4.3—even for a consistent ERM, we can only guarantee on-average-RO stability (as in Theorem 4.4), but we do not know if this is enough to ensure RO stability.

However, although for ERMs we can only obtain a weaker converse, we can guarantee the existence of an AERM that is not only on-average-RO stable but actually uniform-RO stable. That is, we get a much stronger variant of Theorem 4.6:

**Theorem A.1.** *A learning problem is learnable if and only if there exists an uniform-RO stable AERM.*

*Proof.* Clearly if there exists any rule $\mathbf{A}$ that is uniform-RO stable and AERM then the problem is learnable, since the learning rule $\mathbf{A}$ is in fact universally consistent by theorem 4.1. On the other hand if there exists a rule $\mathbf{A}$ that is universally consistent, then consider the rule $\mathbf{A}'$ as in the construction of Lemma 6.11. As shown in the lemma this rule is consistent. Now note that $\mathbf{A}'$ only uses the first $\sqrt{m}$ samples of $S$. Hence for $i > \sqrt{m}$ we have $\mathbf{A}'(S^{(i)}) = \mathbf{A}'(S)$ and so:

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

We thus showed that this rule is consistent, generalizes, and is $\frac{2B}{\sqrt{m}}$-uniformly RO stable. □