Advances in Computational Mathematics (2006) 25: 161–193

Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization

Sayan Mukherjee [a,b], Partha Niyogi [c], Tomaso Poggio [a,⋆] and Ryan Rifkin [a,d]

[a] *Center for Biological and Computational Learning, Artificial Intelligence Laboratory,*
and McGovern Institute, USA
E-mail: tp@ai.mit.edu; rif@mit.edu

[b] *MIT/Whitehead Institute, Center for Genome Research, Massachusetts Institute of Technology,*
Cambridge, MA 02139, USA
E-mail: sayan@mit.edu

[c] *Department of Computer Science and Statistics, University of Chicago, Chicago, IL 60637, USA*
E-mail: niyogi@cs.uchicago.edu

[d] *Honda Research Institute, Boston, MA 02111, USA*

Received 15 October 2003; accepted 1 April 2004
Communicated by Y. Xu

Dedicated to Charles A. Micchelli on his 60th birthday

Solutions of learning problems by Empirical Risk Minimization (ERM) – and almost-ERM when the minimizer does not exist – need to be *consistent*, so that they may be predictive. They also need to be well-posed in the sense of being *stable*, so that they might be used robustly. We propose a statistical form of stability, defined as *leave-one-out* (LOO) *stability*. We prove that for bounded loss classes LOO stability is (a) *sufficient for generalization*, that is convergence in probability of the empirical error to the expected error, for any algorithm satisfying it and, (b) *necessary and sufficient for consistency of ERM*. Thus LOO stability is a weak form of stability that represents a sufficient condition for generalization for symmetric learning algorithms while subsuming the classical conditions for consistency of ERM. In particular, we conclude that a certain form of well-posedness and consistency are equivalent for ERM.

Keywords: stability, inverse problems, generalization, consistency, empirical risk minimization, uniform Glivenko–Cantelli.

Mathematics subject classifications (2000): 68T05, 68T10, 68Q32, 62M20.

[⋆] Corresponding author.

1. Introduction

In learning from a set of examples, the key property of a learning algorithm is *generalization*: the empirical error must converge to the expected error when the number of examples n increases.[⋆] An algorithm that guarantees good generalization for a given n will predict well, if its empirical error on the training set is small. Empirical risk minimization (ERM) on a class of functions $\mathcal{H}$, called the *hypothesis space*, represents perhaps the most natural class of learning algorithms: the algorithm selects a function $f \in \mathcal{H}$ that minimizes the empirical error – as measured on the training set.

Classical learning theory was developed around the study of ERM. One of its main achievements is a complete characterization of the necessary and sufficient conditions for generalization of ERM, and for its *consistency* (consistency requires convergence of the empirical risk to the expected risk for the minimizer of the empirical risk *together* with convergence of the expected risk to the minimum risk achievable by functions in $\mathcal{H}$). It turns out that consistency of ERM is equivalent to a precise property of the hypothesis space: $\mathcal{H}$ has to be a *uniform Glivenko–Cantelli* (uGC) class of functions (see definition 2.4).

Less attention has been given to another requirement on the ERM solution of the learning problem, which has played an important role in the development of several learning algorithms but not in learning theory proper. In general, empirical risk minimization is ill-posed (for any fixed number of training examples n). Any approach of practical interest needs to ensure well-posedness, which usually means existence, uniqueness and stability of the solution. The critical condition is stability of the solution; in this paper we refer to well-posedness, meaning, in particular, stability. In our case, stability refers to continuous dependence on the n training data. Stability is equivalent to some notion of continuity of the learning map (induced by ERM) that maps training sets into the space of solutions, e.g., $L : \bigcup_{n \geqslant 1} Z^n \to \mathcal{H}$.

As a major example, let us consider the following, important case for learning developed in [6]. Assume that the hypothesis space $\mathcal{H}$ is a compact subset of $C(X)$ with X a compact domain in Euclidean space. Compactness[⋆⋆] ensures[‡] the existence of the minimizer of the expected risk for each n and, if the risk functional is convex[‡‡] and regularity conditions on the measure hold, its uniqueness [6, 21]. Compactness guarantees continuity of the learning operator L, measured in the sup norm in $\mathcal{H}$ (see section 2.4.3). However, compactness is not necessary for well-posedness of ERM (it is well known, at least since Tikhonov, that compactness is sufficient but not necessary for well-posedness of a large class of inverse problems involving linear operators). Interestingly, compactness is a sufficient[§] but not necessary condition for consistency as well [6].

[⋆] The precise notion of generalization defined here roughly agrees with the informal use of the term in learning theory.

[⋆⋆] With the sup norm as the distance metric.

[‡] Together with continuity and boundedness of the loss function V.

[‡‡] For convex loss function $V(f, z)$.

[§] Compactness of $\mathcal{H}$ implies the uGC property of $\mathcal{H}$ since it implies *finite covering numbers*.

Thus it is natural to ask the question of whether there is a definition of well-posedness, and specifically stability, of ERM – if any – that is sufficient to guarantee generalization for any algorithm. Since some of the key achievements of learning theory revolve around the conditions equivalent to consistency of ERM, it is also natural to ask whether the same notion of stability could subsume the classical theory of ERM. In other words, is it possible that some specific form of well-posedness is sufficient for generalization and necessary and sufficient for generalization and consistency[⋆] of ERM? *Such a result would be surprising* because, a priori, there is no reason why there should be a connection between well-posedness and generalization – or even consistency (in the case of ERM): they are both important requirements for learning algorithms but they seem quite different and independent of each other.

In this paper, we define a notion of stability that guarantees generalization and in the case of ERM is in fact equivalent to consistency.

There have been many different notions of stability that have been suggested in the past. The earliest relevant notion may be traced to Tikhonov where stability is described in terms of continuity of the learning map L. In learning theory, Devroye and Wagner [9] use certain notions of algorithmic stability to prove the consistency of learning algorithms like the k-nearest neighbors classifier. More recently, Kearns and Ron [14] investigated several notions of stability to develop generalization error bounds in terms of the leave one out error. Bousquet and Elisseeff [5] showed that *uniform hypothesis stability* of the learning algorithm may be used to provide exponential bounds on generalization error without recourse to notions such as the VC dimension.

These various notions of algorithmic stability are all seen to be sufficient for (a) the generalization capability (convergence of the empirical to the expected risk) of learning algorithms. However, until recently, it was unclear whether there is a notion of stability that (b) is also both necessary and sufficient for consistency of ERM. The first partial result in this direction was provided by Kutin and Niyogi [15] who introduced a probabilistic notion of stability called Cross Validation or CV stability. This was shown to be necessary and sufficient for consistency of ERM in the Probably Approximately Correct (PAC) framework of Valiant [25].

However, the task of finding a correct characterization of stability that satisfies both (a) and (b) above is subtle and nontrivial. In [15] at least ten different notions were examined. An answer for the general setting, however, was not found.

In this paper we give a new definition of stability – which we call *Leave-one-out stability* or, in short, *LOO stability* – of the learning map L. This definition answers the open questions mentioned above.

Thus, our somewhat surprising new result is that this notion of stability is sufficient for generalization and is both necessary and sufficient for consistency of ERM. Consistency of ERM is in turn equivalent to $\mathcal{H}$ being a uGC class. To us the result seems interesting for at least three reasons:

[⋆] In the case of ERM it is well known that generalization is equivalent to consistency.

1. it proves the very close relation between two different, and apparently independent, motivations to the solution of the learning problem: consistency and well-posedness;
2. it provides a condition – LOO stability – that is sufficient for generalization for any algorithm and for ERM is necessary and sufficient not only for generalization but also for consistency. LOO stability may, in some ways, be more natural – and perhaps an easier starting point for empirical work[⋆] – than classical conditions such as complexity measures of the hypothesis space $\mathcal{H}$, for example, finiteness of V_γ or VC dimension;
3. it provides a necessary and sufficient condition for consistency of ERM that – unlike all classical conditions (see appendix A.1) – is a condition on the mapping induced by ERM and not directly on the hypothesis space $\mathcal{H}$.

The plan of the paper is as follows. We first give some background and definitions for the learning problem, ERM, consistency and well-posedness. In section 3, which is the core of the paper, we define LOO stability in terms of two conditions: $\mathrm{CV}_{\mathrm{loo}}$ stability and $\mathrm{Eloo}_{\mathrm{err}}$ stability. We prove that LOO stability is sufficient for generalization for general algorithms. We then prove that LOO stability is necessary and sufficient for consistency of ERM. After the main results of the paper we outline in section 4 stronger stability conditions that imply faster rates of convergence and are guaranteed only for "small" uGC classes. Examples are hypothesis spaces with finite VC dimension when the target is in the hypothesis space and balls in Sobolev spaces or Reproducing Kernel Hilbert Spaces (RKHS) with a sufficiently high modulus of smoothness. We then discuss a few remarks and open problems: they include stability conditions and associated concentration inequalities that are equivalent to uGC classes of intermediate complexity – between the general uGC classes characterized by LOO stability (with arbitrary rate) and the small classes mentioned above; they also include the extension of our approach to non-ERM approaches to the learning problem.

2. Background: learning and ill-posed problems

For notation, definitions and some results, we will assume knowledge of a foundational paper [6] and other review papers [13, 16]. The results of [5, 15] are the starting point for our work. Our interest in stability was motivated by the above papers and by our past work in regularization (for reviews see [13, 20]).

2.1. *The supervised learning problem*

There is an unknown probability distribution $\mu(x, y)$ on the product space $Z = X \times Y$. We assume X to be a compact domain in Euclidean space and Y to be a closed subset of $\mathbb{R}^k$. The measure μ defines an unknown *true function* $T(x) = \int_Y y \, \mathrm{d}\mu(y|x)$ mapping X into Y, with $\mu(y|x)$ the conditional probability measure on Y.

[⋆] In its distribution-dependent version.

We are given a training set S consisting of n samples (thus $|S| = n$) drawn i.i.d. from the probability distribution on Z^n:

$$S = (x_i, y_i)_{i=1}^n = (z_i)_{i=1}^n.$$

The basic goal of supervised learning is to use the training set S to "learn" a function f_S that evaluates at a new value x_{new} and (hopefully) predicts the associated value of y:

$$y_{\text{pred}} = f_S(x_{\text{new}}).$$

If y is real-valued, we have regression. If y takes values from $\{-1, 1\}$, we have binary pattern classification. In this paper we consider only symmetric learning algorithms, for which the function output does not depend on the ordering in the training set.

In order to measure goodness of our function, we need a loss function V. We denote by $V(f, z)$ (where $z = (x, y)$) the price we pay when the prediction for a given x is $f(x)$ and the true value is y. An example of a loss function is the square loss which can be written as

$$V(f, z) = \bigl(f(x) - y\bigr)^2.$$

In this paper, *we assume that the loss function V is the square loss*, though most results can be extended to many other "good" loss functions. Throughout the paper we also *require that for any $f \in \mathcal{H}$ and $z \in Z$ the loss is bounded,* $0 \leqslant V(f, z) \leqslant M$.

Given a function f, a loss function V, and a probability distribution μ over X, we define the *expected error* of f as:

$$I[f] = \mathbb{E}_z V(f, z)$$

which is also the expected loss on a new example drawn at random from the distribution. In the case of square loss

$$I[f] = \mathbb{E}_z V(f, z) = \int_{X,Y} \bigl(f(x) - y\bigr)^2 \,\mathrm{d}\mu(x, y) = \mathbb{E}_\mu |f - y|^2.$$

In the following we denote by S^i the training set with the point z_i removed and $S^{i,z}$ the training set with the point z_i replaced with z. For empirical risk minimization, the functions f_S, f_{S^i}, and $f_{S^{i,z}}$ are almost minimizers (see definition 2.1) of $I_S[f]$, $I_{S^i}[f]$, and $I_{S^{i,z}}[f]$, respectively. As we will see later, this definition of perturbation of the training set is a natural one in the context of the learning problem: it is natural to require that the prediction should be asymptotically *robust* against deleting a point in the training set.

2.2. *Empirical risk minimization*

For generalization, that is for correctly predicting new data, we would like to select a function f for which $I[f]$ is small, but in general we do not know μ and cannot compute $I[f]$.

In the following, we will use the notation $\mathbb{P}_S$ and $\mathbb{E}_S$ to denote respectively the probability and the expectation with respect to a random draw of the training set S of size $|S| = n$, drawn i.i.d, from the probability distribution on Z^n.

Given a function f and a training set S consisting of n data points, we can measure the *empirical error (or risk) of f* as:

$$I_S[f] = \frac{1}{n}\sum_{i=1}^{n} V(f, z_i).$$

When the loss function is the square loss

$$I_S[f] = \frac{1}{n}\sum_{i=1}^{n}\bigl(f(x_i) - y_i\bigr)^2 = \mathbb{E}_{\mu_n}(f - y)^2.$$

where μ_n is the empirical measure supported on the set $x_1, \ldots, x_n$. In this notation (see, for example, [16]) $\mu_n = (1/n)\sum_{i=1}^{n}\delta_{x_i}$, where δ_{x_i} is the point evaluation functional on the set x_i.

Definition 2.1. Given a training set S and a function space $\mathcal{H}$, we define almost-ERM (Empirical Risk Minimization) to be a *symmetric* procedure that selects a function $f_S^{\varepsilon^{\mathrm{E}}}$ that *almost minimizes* the empirical risk over all functions $f \in \mathcal{H}$, that is for any given $\varepsilon^{\mathrm{E}} > 0$:

$$I_S\bigl[f_S^{\varepsilon^{\mathrm{E}}}\bigr] \leqslant \inf_{f\in\mathcal{H}} I_S[f] + \varepsilon^{\mathrm{E}}. \tag{2.1}$$

Definition 2.2. An algorithm is defined as symmetric if over training sets S

$$\mathbb{E}_S V(f_S, z) = \mathbb{E}_{S,\pi} V(f_{S(\pi)}, z)$$

for any z and $S(\pi) = \{z_{\pi(1)}, \ldots, z_{\pi(n)}\}$ for every permutation π from $\{1, \ldots, n\}$ onto itself.

In the following, we will drop the dependence on ε^{E} in $f_S^{\varepsilon^{\mathrm{E}}}$. Notice that the term "Empirical Risk Minimization" (see [26]) is somewhat misleading: in general, the minimum need not exist.⋆ In fact, it is precisely for this reason⋆⋆ that we use the notion of almost minimizer or ε-minimizer, given in equation (2.1) (following others, e.g., [1, 16]), since the infimum of the empirical risk always exists. In this paper, we use the term ERM to refer to *almost-ERM*, unless we say otherwise.

We will use the following notation for the *loss class* $\mathcal{L}$ of functions induced by V and $\mathcal{H}$. For every $f \in \mathcal{H}$, let $\ell(z) = V(f, z)$, where z corresponds to x, y. Thus

⋆ When $\mathcal{H}$ is the space of indicator functions, minimizers of the empirical risk exist, because either a point x_i is classified as an error or not.

⋆⋆ It is worth emphasizing that ε-minimization is *not* assumed to take care of algorithmic complexity issues (or related numerical precision issues) that are outside the scope of this paper.

$\ell(z) : X \times Y \to \mathbb{R}$ and we define $\mathcal{L} = \{\ell(f)\colon\ f \in \mathcal{H}, V\}$. The use of the notation ℓ emphasizes that the loss function ℓ is a new function of z induced by f (with the measure μ on $X \times Y$).

2.3. Consistency of ERM and uGC classes

The key problem of learning theory was posed by Vapnik as the problem of statistical consistency of ERM and of the necessary and sufficient conditions to guarantee it. In other words, how can we guarantee that the empirical minimizer of $I_S[f]$ – the distance in the empirical norm between f and y – will yield a small $I[f]$? It is well known (see [1]) that convergence of the empirical error to the expected error guarantees for ERM its consistency.

Our definition of consistency is:

Definition 2.3. A learning map is (universally, weakly) consistent if for any given $\varepsilon_c > 0$

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\Big\{ I[f_S] > \inf_{f\in\mathcal{H}} I[f] + \varepsilon_c \Big\} = 0.$$

Universal consistency means that the above definition holds with respect to the set of all measures on Z. Consistency can be defined with respect to a specific measure on Z. Weak consistency requires only convergence in probability, strong consistency requires almost sure convergence. For bounded loss functions weak consistency and strong consistency are equivalent [11]. In this paper we call consistency what is sometimes defined as weak, universal consistency [7].

The work of Vapnik and Dudley showed that consistency of ERM can be ensured by restricting sufficiently the hypothesis space $\mathcal{H}$ to ensure that a function that is close to a target T for an empirical measure will also be close with respect to the original measure. The key condition for consistency of ERM can be formalized in terms of *uniform convergence in probability* of the functions $\ell(z)$ induced by $\mathcal{H}$ and V. Function classes for which there is uniform convergence in probability are called uniform Glivenko–Cantelli classes of functions:

Definition 2.4. Let $\mathcal{F}$ be a class of functions. $\mathcal{F}$ is a (weak) uniform Glivenko–Cantelli class if

$$\forall\varepsilon > 0 \quad \lim_{n\to\infty} \sup_{\mu} \mathbb{P}\Big\{ \sup_{f\in\mathcal{F}} |\mathbb{E}_{\mu_n} f - \mathbb{E}_{\mu} f| > \varepsilon \Big\} = 0.$$

Measurability issues can be handled by imposing mild conditions on $\mathcal{F}$ (see [10, 11]).

When applied to the loss functions ℓ, the definition implies that for all distributions μ there exist ε_n and δ_n such that

$$\mathbb{P}\Big\{\sup_{f\in\mathcal{F}}\big|I[f]-I_S[f]\big|>\varepsilon_n\Big\}\leqslant\delta_{\varepsilon_n,n},$$

where the sequences ε_n and $\delta_{\varepsilon_n,n}$ go simultaneously to zero.[⋆] Later in the proofs we will take the sequence of $\varepsilon_n^{\mathrm{E}}$ (in the definition of ε-minimizer) to 0 with a rate faster than $1/n$, therefore faster than the sequence of ε_n (e.g., the ε_n in the uGC definition).

We are now ready to state the "classical" necessary and sufficient condition for consistency of ERM (from [1, theorem 4.2, part 3], see also [11, 26]).

Theorem 2.5. Assuming that the loss functions $\ell\in\mathcal{L}$ are bounded and the collection of functions $\{\ell-\inf_{\mathcal{L}}\ell\colon\ \ell\in\mathcal{L}\}$ are uniformly bounded,[⋆⋆] a necessary and sufficient condition for universal consistency of ERM is that $\mathcal{L}$ is uGC.

We observe that for many "good" loss functions V – in particular, the square loss – with ℓ bounded, the uGC property of $\mathcal{H}$ is equivalent to the uGC property of $\mathcal{L}$.[‡]

Notice that there is a definition of strong uGC classes where, instead of convergence in probability, almost sure convergence is required.

Definition 2.6. Let $\mathcal{F}$ be a class of functions. $\mathcal{F}$ is a strong uniform Glivenko–Cantelli class if

$$\forall\varepsilon>0\quad\lim_{n\to\infty}\sup_{\mu}\mathbb{P}\Big\{\sup_{m\geqslant n}\sup_{f\in\mathcal{F}}|\mathbb{E}_{\mu_m}f-\mathbb{E}_{\mu}f|>\varepsilon\Big\}=0.$$

For bounded loss functions weak uGC is equivalent to strong uGC (see [11, theorem 6]) and weak consistency is equivalent to strong consistency in theorem 2.5. In the following, we will speak simply of uGC and consistency, meaning – strictly speaking – weak uGC and weak consistency.

[⋆] This fact follows from the metrization of the convergence of random variables in probability by the Ky Fan metric and its analogue for convergence in outer probability. The rate can be slow, in general (Dudley, Pers. com.).

[⋆⋆] These conditions will be satisfied for bounded loss functions $0\leqslant\ell(z)\leqslant M$.

[‡] Assume that the loss class has the following Lipschitz property for all $x\in X$, $y\in Y$, and $f_1, f_2\in\mathcal{H}$:

$$c_1|V(f_1(x),y)-V(f_2(x),y)|\leqslant|f_1(x)-f_2(x)|\leqslant c_2|V(f_1(x),y)-V(f_2(x),y)|,$$

where $0<c_1<c_2$ are Lipschitz constants that upper and lower-bound the functional difference. Then $\mathcal{L}$ is uGC iff $\mathcal{H}$ is uGC because there are Lipschitz constants that upper and lower bound the difference between two functions ensuring that the cardinality of $\mathcal{H}$ and $\mathcal{L}$ at a scale ε differ by at most a constant. Bounded L_p losses have this property for $1\leqslant p<\infty$.

2.4. *Inverse and well-posed problems*

2.4.1. The classical case

Hadamard introduced the definition of ill-posedness. Ill-posed problems are often inverse problems.

As an example, assume g is an element of Z and u is a function in $\mathcal{H}$, with Z and $\mathcal{H}$ metric spaces. Then given the operator A, consider the equation

$$g = Au. \tag{2.2}$$

The direct problem is to compute g given u; the inverse problem is to compute u given the data g. The inverse problem of finding u is well-posed when

- the solution exists,
- is unique, and
- is *stable*, that is depends continuously on the initial data g. In the example above this means that A^{-1} has to be continuous. Thus stability has to be defined in terms of the relevant norms.

Ill-posed problems (see [12]) fail to satisfy one or more of these criteria. In the literature the term ill-posed is often used for problems that are *not stable*, which is the key condition. In equation (2.2) the map A^{-1} is continuous on its domain Z if, given any $\varepsilon > 0$, there is a $\delta > 0$ such that for any $z', z'' \in Z$

$$\|z' - z''\| \leqslant \delta$$

with the norm in Z, then

$$\left\|A^{-1}z' - A^{-1}z''\right\| \leqslant \varepsilon,$$

with the norm in $\mathcal{H}$.

The basic idea of regularization for solving ill-posed problems is to restore existence, uniqueness and stability of the solution by an appropriate choice of $\mathcal{H}$ (the hypothesis space in the learning framework). Usually, existence can be ensured by redefining the problem and uniqueness can often be restored in simple ways (for instance, in the learning problem we choose randomly one of the several equivalent *almost minimizers*). However, stability of the solution is usually much more difficult to guarantee. The regularization approach has its origin in a topological lemma[⋆] that under certain conditions points to the compactness of $\mathcal{H}$ as sufficient for establishing stability and thus well-posedness.[⋆⋆]

[⋆] Lemma (Tikhonov [24]). If operator A maps a compact set $\mathcal{H} \subset H$ onto $Z \subset Q$, H and Q metric spaces, and A is continuous and one-to-one, then the inverse mapping is also continuous.

[⋆⋆] In learning, the approach underlying most algorithms such as Radial Basis Functions (RBFs) and Support Vector Machines (SVMs) is in fact regularization. These algorithms can therefore be directly motivated in terms of restoring well-posedness of the learning problem.

Notice that when the solution of equation (2.2) does not exist, the standard approach is to replace it with the following problem, analogous to ERM,

$$\min_{u \in \mathcal{H}} \|Au - g\|, \tag{2.3}$$

where the norm is in Z. Assuming for example that Z and $\mathcal{H}$ are Hilbert spaces and A is linear and continuous, the solutions of equation (2.3) coincide with the solutions of

$$Au = Pg, \tag{2.4}$$

where P is the projection onto $R(A) = \{Au \mid u \in \mathcal{H}\}$.

2.4.2. Classical framework: regularization of the learning problem

For the learning problem it is clear, but often neglected, that ERM is, in general, *ill-posed* for any given S_n. ERM defines a map L which maps any discrete data $S = ((x_1, y_1), \dots, (x_n, y_n))$ into a function f, that is

$$LS = f_S.$$

In equation (2.2) L corresponds to A^{-1} and g to the discrete data S. In general, the operator L induced by ERM cannot be expected to be linear. In the rest of this subsection, we consider a simple, "classical" case that corresponds to equation (2.4) and in which L is linear.

Assume that the x part of the n examples $(x_1, \dots, x_n)$ is fixed; then L as an operator on $(y_1, \dots, y_n)$ can be defined in terms of a set of evaluation functionals F_i on $\mathcal{H}$, that is $y_i = F_i(u)$. If $\mathcal{H}$ is a Hilbert space and in it the evaluation functionals F_i are *linear and bounded*, then $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) and the F_i can be written as $F_i(u) = (u, K_{x_i})_K$ where K is the kernel associated with the RKHS and we use the inner product in the RKHS. For simplicity we assume that K is positive definite and sufficiently smooth [6, 28]. The ERM case corresponds to equation (2.3) that is

$$\min_{f \in B_R} \frac{1}{n} \sum_{i=1}^{n} \bigl(f(\mathbf{x}_i) - y_i\bigr)^2. \tag{2.5}$$

Compactness is ensured by enforcing the solution f – which has the form $f(\mathbf{x}) = \sum_{1=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x})$ since it belongs to the RKHS – to be in the ball B_R of radius R in $\mathcal{H}$ (e.g., $\|f\|_K \leqslant R$). Then $\mathcal{H} = \overline{I_K(B_R)}$ is compact – where $I_K : \mathcal{H}_K \hookrightarrow C(X)$ is the inclusion and $C(X)$ is the space of continuous functions with the sup norm [6]. In this case the minimizer of the generalization error $I[f]$ is well-posed. Minimization of the empirical risk (equation (2.5)) is also well-posed: it provides a set of linear equations to compute the coefficients $\mathbf{c}$ of the solution f as

$$K\mathbf{c} = \mathbf{y} \tag{2.6}$$

where $\mathbf{y} = (y_1, \dots, y_n)$ and $(K)_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$.

A particular form of regularization, called Tikhonov regularization, replaces ERM (see equation (2.5)) with

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\bigl(f(\mathbf{x}_i)-y_i\bigr)^2+\gamma\|f\|_K^2, \tag{2.7}$$

which gives the following set of equations for $\mathbf{c}$ (with $\gamma \geqslant 0$)

$$(K+n\gamma I)\mathbf{c}=\mathbf{y}, \tag{2.8}$$

which for $\gamma = 0$ reduces to equation (2.6). In this RKHS case, stability of the empirical risk minimizer provided by equation (2.7) can be characterized using the classical notion of *condition number* of the problem. The change in the solution f due to a variation in the data $\mathbf{y}$ can be bounded as

$$\frac{\|\Delta f\|}{\|f\|}\leqslant \|K+n\gamma I\|\bigl\|(K+n\gamma I)^{-1}\bigr\|\frac{\|\Delta\mathbf{y}\|}{\|\mathbf{y}\|}, \tag{2.9}$$

where the condition number $\|K+n\gamma I\|\|(K+n\gamma I)^{-1}\|$ is controlled by $n\gamma$. A large value of $n\gamma$ gives condition numbers close to 1, whereas ill-conditioning may result if $\gamma = 0$ and the ratio of the largest to the smallest eigenvalue of K is large.

Remarks.

1. Equation (2.5) for any fixed n corresponds to the set of well-posed, linear equations (2.6), even without the constraint $\|f\|_K^2 \leqslant R$: if K is symmetric and positive definite and the x_i are distinct then K^{-1} exists and $\|f\|_K^2$ is automatically bounded (with a bound that increases with n). For any fixed n, the condition number is finite but typically increases with n by equation (2.9).
2. Minimization of the functional in equation (2.7) with $\gamma > 0$ implicitly enforces the solution to be in a ball in the RKHS, whose radius can be bounded “a priori” before the data set S is known (see [18]).

2.4.3. Stability of learning: a more general case

The approach to defining stability described above for the RKHS case cannot be used directly in the more general setup of the supervised learning problem introduced in section 2.1. In particular, the training set S_n is drawn i.i.d. from the probability distribution on Z, the x_i are not fixed and we may not even have a norm in $\mathcal{H}$ (in the case of RKHS the norm in $\mathcal{H}$ bounds the sup norm).

The probabilistic case for $\mathcal{H}$ with the sup *norm.* A definition of stability that takes care of some of the issues above was introduced by [5] with the name of *uniform stability*:

$$\forall S\in Z^n,\ \forall i\in\{1,\dots,n\}\quad \sup_{z\in Z}\bigl|V(f_S,z)-V(f_{S^i},z)\bigr|\leqslant\beta. \tag{2.10}$$

Kutin and Niyogi [15] showed that ERM does not, in general, exhibit uniform stability. Therefore they extended it in a probabilistic sense with the name of (β,δ)

hypothesis stability, which is a natural stability criterion for hypothesis spaces equipped with the sup norm. We give here a slightly different version:

$$\mathbb{P}_S\Big\{\sup_{z\in Z}\big|V(f_S,z)-V(f_{S^i},z)\big|\leqslant\beta\Big\}\geqslant 1-\delta, \tag{2.11}$$

where β and δ go to zero with $n\to\infty$.

Interestingly, the results of [5] imply that Tikhonov regularization algorithms are uniformly stable (and of course (β,δ) hypothesis stable) with $\beta=\mathrm{O}(1/\gamma n)$. Thus, this definition of stability recovers the key parameters for good conditioning number of the regularization algorithms. As discussed later, we conjecture that in the case of ERM, (β,δ) hypothesis stability is related to the compactness of $\mathcal{H}$ with respect to the sup norm in $C(X)$.

A more general definition of stability. The definitions of stability introduced in the past are not general enough to be equivalent to the classical necessary and sufficient conditions on $\mathcal{H}$ for consistency of ERM.[⋆] The key ingredient in our definitions of stability given above is some measure on $|\ell_{f_S}-\ell_{f_{S^i}}|$, e.g., a measure of the difference between the error made by the predictor obtained by using ERM on the training set S vs. the error of the predictor obtained from a slightly perturbed training set S^i. We propose here the following definition[⋆⋆] of *leave-one-out cross-validation* ($\mathrm{CV_{loo}}$) *stability*, which is the key part in the notion of LOO stability introduced later:

$$\forall i\in\{1,\dots,n\}\quad \mathbb{P}_S\big\{\big|V(f_S,z_i)-V(f_{S^i},z_i)\big|\leqslant\beta_{\mathrm{CV}}\big\}\geqslant 1-\delta_{\mathrm{CV}}.$$

Here we measure the difference between the errors at a point z_i which is in the training set of one of the predictors but not in the training set of the other. Notice that the definitions of stability we discussed here are progressively weaker: a good condition number (for increasing n) implies good uniform stability.[‡] In turns, *uniform stability implies* (β,δ) *hypothesis stability which implies* $\mathrm{CV_{loo}}$ *stability*. For the case of supervised learning all the definitions capture the basic idea of stability of a well-posed problem: the function "learned" from a training set should, with high probability, change little in its pointwise predictions for a small change in the training set, such as deletion of one of the examples.

Remarks.

1. In the learning problem, *uniqueness* of the solution of ERM is always meant in terms of uniqueness of ℓ and therefore uniqueness of the equivalence class induced in $\mathcal{H}$ by the loss function V. In other words, multiple $f\in\mathcal{H}$ may provide the same ℓ. Even

[⋆] In addition, the above definitions of stability are not appropriate for hypothesis spaces for which the sup norm is not meaningful, at least in the context of the learning problem (for instance, for hypothesis spaces of indicator functions).

[⋆⋆] The definition is given here in its distribution-dependent form.

[‡] Note that $n\gamma$ which controls the quality of the condition number in regularization also controls the rate of uniform stability.

in this sense, ERM on a uGC class is not guaranteed to provide a unique "almost minimizer". Uniqueness of an almost minimizer therefore is a rather weak concept since uniqueness is valid *modulo the equivalence classes* induced by the loss function *and* by ε-minimization.

2. Stability of *algorithms* is almost always violated, even in good and useful algorithms (Smale, Pers. comm.). In this paper, we are not concerned about stability of algorithms but *stability of problems*. Our notions of stability of the map L are in the same spirit as the condition number of a linear problem, which is independent of the algorithm used to solve it. As we discussed earlier, both $\mathrm{CV_{loo}}$ stability and uniform stability can be regarded as extensions of the notion of condition number (for a discussion in the context of inverse ill-posed problems see [3]).

3. $\mathrm{CV_{loo}}$ and $\mathrm{ELoo_{err}}$ stability, generalization and consistency of ERM

3.1. Probabilistic preliminaries

The following are consequences of the linearity of expectations and the symmetry of the learning algorithm. They will be used throughout the paper:

$$\mathbb{E}_S\big[I[f_S]\big] = \mathbb{E}_S\big[\mathbb{E}_z V(f_S, z)\big] = \mathbb{E}_{S,z}\big[V(f_S, z)\big],$$

for all $i \in \{1, \dots, n\}$

$$\mathbb{E}_S\big[I_S[f_S]\big] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} V(f_S, z_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\big[\mathbb{E}_S V(f_S, z_i)\big] = \mathbb{E}_S\big[V(f_S, z_i)\big],$$

$$\mathbb{E}_S\big[I[f_{S^i}]\big] = \mathbb{E}_S\mathbb{E}_{z^i} V(f_{S^i}, z_i) = \mathbb{E}_S\big[V(f_{S^i}, z_i)\big].$$

3.2. Forms of stability

This section introduces several definitions of stability and shows the equivalence of two of them. The first definition of stability of the learning map L, is *Cross-Validation leave-one-out* ($\mathrm{CV_{loo}}$) *stability*. This notion of stability is a variation of a definition of stability introduced in [15].

Definition 3.1. The learning map L is distribution-independent, $(\beta_{\mathrm{CV}}^{(n)}, \delta_{\mathrm{CV}}^{(n)})$ $\mathrm{CV_{loo}}$ *stable* if for each n there exists a $\beta_{\mathrm{CV}}^{(n)}$ and a $\delta_{\mathrm{CV}}^{(n)}$ such that

$$\forall i \in \{1, \dots, n\},\ \forall \mu \quad \mathbb{P}_S\big\{\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big| \leqslant \beta_{\mathrm{CV}}^{(n)}\big\} \geqslant 1 - \delta_{\mathrm{CV}}^{(n)},$$

with $\beta_{\mathrm{CV}}^{(n)}$ and $\delta_{\mathrm{CV}}^{(n)}$ going to zero for $n \to \infty$.

Notice that our definition of the stability of L depends on the pointwise value of $|V(f_S, z_i) - V(f_{S^i}, z_i)|$. This definition is weaker than the uniform stability condition stated in [5] and is implied by it.

A definition which turns out to be equivalent was introduced in [5] (see also [14]) under the name of *pointwise hypothesis* (*PH*) *stability*.

Definition 3.2. The learning map L is distribution-independent, (*PH*) *stable* if for each n there exists a $\beta_{PH}^{(n)}$

$$\forall i \in \{1,\dots,n\},\ \forall\mu \quad \mathbb{E}_S\big[\big|V(f_S,z_i) - V(f_{S^i},z_i)\big|\big] \leqslant \beta_{PH}^{(n)},$$

with $\beta_{PH}^{(n)}$ going to zero for $n \to \infty$.

We now show that the two definitions of $\mathrm{CV}_{\mathrm{loo}}$ stability and *PH* stability are equivalent.

Lemma 3.3. $\mathrm{CV}_{\mathrm{loo}}$ stability with β_{loo} and δ_{loo} implies *PH* stability with $\beta_{PH} = \beta_{\mathrm{loo}} + M\delta_{\mathrm{loo}}$ and *PH* stability with β_{PH} implies $\mathrm{CV}_{\mathrm{loo}}$ stability with $(\alpha, \beta_{PH}/\alpha)$ for any $\alpha < \beta_{PH}$.

Proof. The following proof holds for any distribution μ and therefore it is distribution independent. From the definition of $\mathrm{CV}_{\mathrm{loo}}$ stability and the bound on the loss function it follows that

$$\forall i \in \{1,\dots,n\} \quad \mathbb{E}_S\big[\big|V(f_S,z_i) - V(f_{S^i},z_i)\big|\big] \leqslant \beta_{\mathrm{loo}} + M\delta_{\mathrm{loo}}.$$

This proves the first statement.

From the definition of *PH* stability, we have

$$\mathbb{E}_S\big[\big|V(f_{S^i},z_i) - V(f_S,z_i)\big|\big] \leqslant \beta_{PH}.$$

Since $|V(f_{S^i},z_i) - V(f_S,z_i)| \geqslant 0$, by Markov's inequality, we have

$$\mathbb{P}\big[\big|V(f_{S^i},z_i) - V(f_S,z_i)\big| > \alpha\big] \leqslant \frac{\mathbb{E}_S[|V(f_{S^i},z_i) - V(f_S,z_i)|]}{\alpha} \leqslant \frac{\beta_{PH}}{\alpha}.$$

This proves the second statement. □

We now introduce a condition that we call *Expected-to-leave-one-out error* ($\mathrm{Eloo}_{\mathrm{err}}$) *stability*.

Definition 3.4. The learning map L is distribution-independent, $\mathrm{Eloo}_{\mathrm{err}}$ *stable* if for each n there exists a $\beta_{\mathrm{EL}}^{(n)}$ and a $\delta_{\mathrm{EL}}^{(n)}$ such that

$$\forall i \in \{1,\dots,n\},\ \forall\mu \quad \mathbb{P}_S\left\{\left|I[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i},z_i)\right| \leqslant \beta_{\mathrm{EL}}^{(n)}\right\} \geqslant 1 - \delta_{\mathrm{EL}}^{(n)},$$

with $\beta_{\mathrm{EL}}^{(n)}$ and $\delta_{\mathrm{EL}}^{(n)}$ going to zero for $n \to \infty$.

Our use of the term stability for the $\mathrm{Eloo}_{\mathrm{err}}$ property may seem somewhat of a stretch (though the definition depends on a “perturbation” of the training set from S to S^i). It is justified however by the fact that the $\mathrm{Eloo}_{\mathrm{err}}$ property is implied – in the general setting – by a classical leave-one-out notion of stability called *hypothesis stability*, which was introduced by Devroye and Wagner [9] and later used in [5, 14] (and in a stronger change-one form in [15]). Our definition of hypothesis stability is equivalent to leave-one-out stability in the L_1 norm.

Definition 3.5. The learning map L is distribution-independent, leave-one-out *hypothesis stable* if for each n there exists a $\beta_H^{(n)}$

$$\forall\mu \quad \mathbb{E}_{S,z}\left[\left|V(f_S,z)-V(f_{S^i},z)\right|\right]\leqslant\beta_H^{(n)},$$

with $\beta_H^{(n)}$ going to zero for $n\to\infty$.

Intuitively, the $\mathrm{Eloo}_{\mathrm{err}}$ condition may seem both strong and weak. In particular, it looks weak because the leave-one-out error $I_{\mathrm{loo}}=n^{-1}\sum_{i=1}^n V(f_{S^i},z_i)$ seems a good empirical proxy for the expected error $\mathbb{E}_z V(f_S,z)$ and it is in fact routinely used in this way for evaluating empirically the expected error of learning algorithms.

Definition 3.6. A learning map L is LOO *stable* if it exhibits both $\mathrm{CV}_{\mathrm{loo}}$ and $\mathrm{Eloo}_{\mathrm{err}}$ stability.

3.3. LOO stability implies generalization

We now prove that $\mathrm{CV}_{\mathrm{loo}}$ and $\mathrm{Eloo}_{\mathrm{err}}$ stability together are sufficient for generalization of symmetric learning algorithms. The following lemma was mentioned as remark 10 in [5].[⋆]

Lemma 3.7. The generalization error can be decomposed as follows

$$\mathbb{E}_S\big(I[f_S]-I_S[f_S]\big)^2\leqslant 2\mathbb{E}_S\left(I[f_S]-\frac{1}{n}\sum_{i=1}^n V(f_{S^i},z_i)\right)^2+2M\mathbb{E}_S\left|V(f_S,z_i)-V(f_{S^i},z_i)\right|.$$

Proof. By the triangle inequality and inspection

$$\begin{aligned}\mathbb{E}_S\big(I[f_S]-I_S[f_S]\big)^2\leqslant{}&2\mathbb{E}_S\left(I[f_S]-\frac{1}{n}\sum_{j=1}^n V(f_{S^j},z_j)\right)^2\\&+2\mathbb{E}_S\left(I_S[f_S]-\frac{1}{n}\sum_{j=1}^n V(f_{S^j},z_j)\right)^2.\end{aligned}$$

[⋆] Bousquet and Elisseeff attribute the result to Devroye and Wagner.

We now bound the second term

$$\begin{aligned}
&\mathbb{E}_S\left(I_S[f_S] - \frac{1}{n}\sum_{j=1}^{n} V(f_{S^j}, z_j)\right)^2 \\
&\quad = \mathbb{E}_S\left(\frac{1}{n}\sum_{j=1}^{n} V(f_S, z_j) - \frac{1}{n}\sum_{j=1}^{n} V(f_{S^j}, z_j)\right)^2 \\
&\quad = \mathbb{E}_S\frac{1}{n}\left|\sum_{j=1}^{n}\bigl[V(f_S, z_j) - V(f_{S^j}, z_j)\bigr]\right|^2 \leqslant M\mathbb{E}_S\frac{1}{n}\left|\sum_{j=1}^{n}\bigl[V(f_S, z_j) - V(f_{S^j}, z_j)\bigr]\right| \\
&\quad \leqslant M\mathbb{E}_S\frac{1}{n}\sum_{j=1}^{n}\bigl|V(f_S, z_j) - V(f_{S^j}, z_j)\bigr| = M\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_S\bigl|V(f_S, z_j) - V(f_{S^j}, z_j)\bigr| \\
&\quad = M\mathbb{E}_S\bigl|V(f_S, z_i) - V(f_{S^i}, z_i)\bigr|. \qquad \square
\end{aligned}$$

The following proposition follows directly from the above decomposition of the generalization error.

Proposition 3.8. LOO stability implies generalization.

Remarks.

1. Other stability conditions can be derived that ensure generalization for symmetric algorithms [17, 19].
2. $\mathrm{CV}_{\mathrm{loo}}$ and $\mathrm{Eloo}_{\mathrm{err}}$ stability together are strong enough to imply generalization for symmetric algorithms, but neither condition by itself is sufficient.
3. $\mathrm{CV}_{\mathrm{loo}}$ stability by itself is *not* sufficient for generalization, as the following counterexample shows. Let X be uniform on $[0, 1]$. Let $Y \in \{-1, 1\}$. Let the "target function" be $t(x) = 1$, and the loss-function be the $\{0, 1\}$-loss.
   Given a training set of size n, our (non-ERM) algorithm ignores the y values and produces the following function:
   $$f_S(x) = \begin{cases} (-1)^n & \text{if } x \text{ is a training point,} \\ (-1)^{n+1} & \text{otherwise.} \end{cases}$$
   Now consider what happens when we remove a single training point to obtain f_{S^i}. Clearly,
   $$f_{S^i}(x) = \begin{cases} f_S(x) & \text{if } x = x_i, \\ -f_S(x) & \text{otherwise.} \end{cases}$$
   In other words, when we remove a training point, the value of the output function switches at every point except that training point. The value at the training point removed does not change at all, so the algorithm is (β_C, δ_C) $\mathrm{CV}_{\mathrm{loo}}$ stable with $\beta_C =$

$\delta_C = 0$. However, this algorithm does not generalize at all; for every training set, depending on the size of the set, either the training error is 0 and the testing error is 1, or vice versa.

4. Eloo_{err} stability by itself is *not* sufficient for generalization, as the following example shows. Using the same setup in the previous remark, consider an algorithm which returns 1 for every training point, and -1 for every test point. This algorithm is Eloo_{err} stable (as well as hypothesis stable), but does not generalize.

5. The converse of proposition 3.8 is false. Using the same setup as in the previous remark, consider an algorithm that, given a training set of size n, yields the constant function $f(x) = (-1)^n$. This algorithm is neither CV_{loo} or Eloo_{err} stable, but it will generalize.

6. In [5, theorem 11], it is claimed that *PH* stability (which is equivalent to CV_{loo} stability, by lemma 3.3) is sufficient for generalization. However, there is an error in this proof. The second line of the theorem, translated into our notation, states correctly that

$$\begin{aligned}\mathbb{E}_{S,z}\big[\big|V(f_S, z_i) - V(f_{S^{i,z}}, z_i)\big|\big] \leqslant{}& \mathbb{E}_S\big[\big|V(f_S, z_i) - V(f_{S^i}, z_i)\big|\big] \\ &+ \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_{S^{i,z}}, z_i)\big|\big].\end{aligned}$$

PH stability is used to bound both terms in the expansion. While the first term can be bounded using *PH* stability, the second term involves the difference in performance on z_i between functions generated from two different test sets, neither of which contain z_i; this cannot be bounded using *PH* stability. The proof can be easily "fixed" by bounding the second term using the more general notion of hypothesis stability; this would then prove that the combination of CV_{loo} stability and hypothesis stability are sufficient for generalization, which also follows directly from proposition 3.8. Hypothesis stability is a stronger notion than Eloo_{err} stability since hypothesis stability implies Eloo_{err} stability but Eloo_{err} stability does not imply hypothesis stability.[⋆]

We now ask whether Eloo_{err} and CV_{loo} stability together are enough to capture the fundamental conditions for consistency of ERM and thus *subsume the "classical" theory.* We will in fact show in the next section 3.4.2 that CV_{loo} stability alone is equivalent

[⋆] There is an unfortunate confusing proliferation of definitions of stability. The hypothesis stability of Elisseeff and Bousquet [5] is essentially equivalent to the L_1 stability of Kutin and Niyogi [15] (modulo probabilistic versus non-probabilistic and change-one versus leave-one-out differences); similarly, what Kutin and Niyogi call (β, δ) hypothesis stability is a probabilistic version of the (very strong) uniform stability of Elisseeff and Bousquet. It is problematic that many versions of stability exist in both change-one and leave-one-out forms. If a given form of stability measures error at a point that is not in either training set, the change-one form implies the leave-one-out form (for example, Bousquet and Elisseeff's hypothesis stability implies Kutin and Niyogi's weak-L_1 stability), but if the point at which we measure is added to the training set, this does not hold (for example, our CV_{loo} stability does not imply the change-one CV stability of Kutin and Niyogi; in fact, Kutin and Niyogi's CV stability is roughly equivalent to the combination of our CV_{loo} stability and Elisseeff and Bousquet's hypothesis stability).

to consistency of ERM. To complete the argument, we will also show in subsection 3.4.3 that Eloo_{err} stability is implied by consistency of ERM.

3.4. LOO stability is necessary and sufficient for consistency of ERM

The main result of this section is the following theorem.

Theorem 3.9. Assume that $f_S, f_{S^i} \in \mathcal{H}$ are provided by ERM and the loss is bounded. Then LOO stability is necessary and sufficient for consistency of ERM. Therefore, the following are equivalent

(a) the map induced by almost ERM is LOO stable,

(b) almost ERM is universally consistent,

(c) $\mathcal{L}$ is uGC.

Proof. The equivalence of (b) and (c) is well known (see theorem 2.5). The the equivalence of (a) and (b) is a result of the following theorem and lemma which are proven in sections 3.4.2 and 3.4.3, respectively.

Theorem 3.10. CV_{loo} stability is necessary and sufficient for consistency of ERM on a function class $\mathcal{H}$.

Lemma 3.11. ERM on a uGC class implies Eloo_{err} stability

$$\mathbb{E}_S\left(I[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant \beta_n,$$

where $\lim_{n\to\infty} \beta_n = 0$.

As a result of the above theorem and lemma (b) implies (a). By proposition 3.8 (a) implies (b). The equivalence of (a) and (b) follows. □

Remark. If we make specific assumptions on the loss function V (see footnote[12]), then theorem 3.9 can be stated in terms of $\mathcal{H}$ being uGC.

3.4.1. Almost positivity of ERM

The two lemmas in this section will be important in proving theorem 3.10.

We first prove a lemma about the *almost positivity*[⋆] of $V(f_S, z_i) - V(f_{S^i}, z_i)$.

Lemma 3.12. Under the assumption that ERM finds a ε^{E}-minimizer,

$$\forall i \in \{1, \ldots, n\} \quad V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon^{\text{E}} \geqslant 0.$$

[⋆] Shahar Mendelson's comments prompted us to define the notion of *almost positivity*.

Proof. By the definition of almost minimizer (see equation (2.1)), we have

$$\frac{1}{n}\sum_{z_j\in S} V(f_{S^i}, z_j) - \frac{1}{n}\sum_{z_j\in S} V(f_S, z_j) \geqslant -\varepsilon_n^{\mathrm{E}},$$
$$\frac{1}{n}\sum_{z_j\in S^i} V(f_{S^i}, z_j) - \frac{1}{n}\sum_{z_j\in S^i} V(f_S, z_j) \leqslant \frac{n-1}{n}\varepsilon_{n-1}^{\mathrm{E}}.$$

We can rewrite the first inequality as

$$\left[\frac{1}{n}\sum_{z_j\in S^i} V(f_{S^i}, z_j) - \frac{1}{n}\sum_{z_j\in S^i} V(f_S, z_j)\right] + \frac{1}{n}V(f_{S^i}, z_i) - \frac{1}{n}V(f_S, z_i) \geqslant -\varepsilon_n^{\mathrm{E}}.$$

The term in the bracket is less than or equal to $((n-1)/n)\varepsilon_{n-1}^{\mathrm{E}}$ (because of the second inequality) and thus

$$V(f_{S^i}, z_i) - V(f_S, z_i) \geqslant -n\varepsilon_n^{\mathrm{E}} - (n-1)\varepsilon_{n-1}^{\mathrm{E}}.$$

Because the sequence of $n\varepsilon_n^N$ is a decreasing sequence of positive terms, we obtain

$$V(f_{S^i}, z_i) - V(f_S, z_i) \geqslant -2(n-1)\varepsilon_{n-1}^{\mathrm{E}}. \qquad \square$$

Lemma 3.13. Under almost ERM with $\varepsilon_n^{\mathrm{E}} > 0$ chosen such that $\lim_{n\to\infty} n\varepsilon_n^{\mathrm{E}} = 0$, the following bound holds

$$\forall i \in \{1,\ldots,n\} \quad \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] \leqslant \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S] + 4(n-1)\varepsilon_{n-1}^{\mathrm{E}}.$$

Proof. We note that

$$\begin{aligned}
&\mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big]\\
&\quad = \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}} - 2(n-1)\varepsilon_{n-1}^{\mathrm{E}}\big|\big]\\
&\quad \leqslant \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}}\big|\big] + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}}.
\end{aligned}$$

Now we make two observations. By lemma 3.12,

$$\forall i \in \{1,\ldots,n\} \quad V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}} \geqslant 0,$$

and therefore

$$\mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}}\big|\big] = \mathbb{E}_S\big[V(f_{S^i}, z_i) - V(f_S, z_i)\big] + 2(n-1)\varepsilon_{n-1}^{\mathrm{E}}.$$

Second, by the linearity of expectations,

$$\mathbb{E}_S\big[V(f_{S^i}, z_i) - V(f_S, z_i)\big] = \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S],$$

and therefore

$$\mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] \leqslant \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S] + 4(n-1)\varepsilon_{n-1}^{\mathrm{E}}. \qquad \square$$

Remark. In the case when ERM finds a minima (exact minimization), *positivity* holds

$$\forall i \in \{1, \dots, n\} \quad V(f_{S^i}, z_i) - V(f_S, z_i) \geqslant 0,$$

and the leave-one-out error error is greater than or equal to the training error

$$\frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i) \geqslant I_S[f_S].$$

3.4.2. CV$_{\text{loo}}$ *stability is necessary and sufficient for consistency of ERM*

Before proving theorem 3.10 in detail for the general case of almost ERM we prove the statement for the case of exact ERM.

Exact ERM. The following short proof could be made much shorter by referring to known results on ERM. The argument for almost ERM can be made along similar lines with a few additional, annoying ε terms.

Theorem 3.14. Under exact minimization of the empirical risk and the existence of the minima of the true risk, $I[f^*]$ where $f^* \in \arg\min_{f\in\mathcal{H}} I[f]$, then (β, δ) CV$_{\text{loo}}$ stability is equivalent to universal consistency of ERM.

Proof. By the assumption of exact ERM positivity holds

$$V(f_{S^i}, z_i) - V(f_S, z_i) \geqslant 0.$$

Then the following equivalences hold:

$$\begin{aligned}
(\beta,\delta)\ \text{CV}_{\text{loo}}\ \text{stability} \quad &\Leftrightarrow \quad \lim\nolimits_{n\to\infty} \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] = 0 \\
&\Leftrightarrow \quad \lim\nolimits_{n\to\infty} \mathbb{E}_S\big[V(f_{S^i}, z_i) - V(f_S, z_i)\big] = 0 \\
&\Leftrightarrow \quad \lim\nolimits_{n\to\infty} \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S] = 0 \\
&\Leftrightarrow \quad \lim\nolimits_{n\to\infty} \mathbb{E}_S I[f_{S^i}] = \lim\nolimits_{n\to\infty} \mathbb{E}_S I_S[f_S].
\end{aligned}$$

Now, $I[f^*] \leqslant I[f_{S^i}]$ and $I_S[f_S] \leqslant I_S[f^*]$. Therefore,

$$I[f^*] \leqslant \lim_{n\to\infty} \mathbb{E}_S I[f_{S^i}] = \lim_{n\to\infty} \mathbb{E}_S I_S[f_S] \leqslant \lim_{n\to\infty} \mathbb{E}_S I_S[f^*] = I[f^*],$$

resulting in

$$\lim_{n\to\infty} \mathbb{E}_S I[f_{S^i}] = \lim_{n\to\infty} \mathbb{E}_S I[f^*] = I[f^*],$$

which implies that in probability,

$$\lim_{n\to\infty} I[f_{S^i}] = I[f^*].$$

Finally, we note that the convergence in probability of $I[f_{S^i}]$ to $I[f^*]$ is equivalent to the convergence of $I[f_S]$ to $I[f^*]$ in probability which is universal consistency. □

Almost ERM. The next two theorems combined prove theorem 3.10. We prove first sufficiency and then necessity.

Theorem 3.15. If the map induced by ERM over a class $\mathcal{H}$ is distribution-independent $\mathrm{CV}_{\mathrm{loo}}$ stable, and the loss is bounded by M, then ERM over $\mathcal{H}$ is universally consistent.

Proof. Given a sample $S = (z_1, \dots, z_n)$ with n points and a sample $S_{n+1} = (z_1, \dots, z_{n+1})$ then by $\mathrm{CV}_{\mathrm{loo}}$ stability of ERM, the following holds for all μ:

$$\begin{aligned}\mathbb{E}_{S_{n+1}}\bigl[V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})\bigr] &\leqslant \mathbb{E}_{S_{n+1}}\bigl[\bigl|V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})\bigr|\bigr] \\ &\leqslant \beta_{PH}^{(n+1)}, \end{aligned} \tag{3.1}$$

where $\beta_{PH}^{(n+1)} = \beta_{\mathrm{CV}_{\mathrm{loo}}}^{(n+1)} + M\delta_{\mathrm{CV}_{\mathrm{loo}}}^{(n+1)}$.

The following holds for all μ:

$$\mathbb{E}_S I[f_S] - \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] = \mathbb{E}_{S_{n+1}}\bigl[V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})\bigr]. \tag{3.2}$$

From equations (3.1) and (3.2), we therefore have

$$\forall\mu \quad \mathbb{E}_S I[f_S] \leqslant \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] + \beta_{PH}^{(n+1)}. \tag{3.3}$$

Now we will show that

$$\lim_{n\to\infty} \sup_\mu \Bigl(\mathbb{E}_S I[f_S] - \inf_{f\in\mathcal{H}} I[f]\Bigr) = 0.$$

Let $\eta_\mu = \inf_{f\in\mathcal{H}} I[f]$ under the distribution μ. Clearly, for all $f \in \mathcal{H}$, we have $I[f] \geqslant \eta_\mu$ and so $\mathbb{E}_S I[f_S] \geqslant \eta_\mu$. Therefore, we have (from (3.3))

$$\forall\mu \quad \eta_\mu \leqslant \mathbb{E}_S I[f_S] \leqslant \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] + \beta_{PH}^{(n+1)}. \tag{3.4}$$

For every $\varepsilon_c > 0$, there exists $f_{\varepsilon_c,\mu} \in \mathcal{H}$ such that $I[f_{\varepsilon_c,\mu}] < \eta_\mu + \varepsilon_c$. By the almost ERM property, we also have

$$I_{S_{n+1}}[f_{S_{n+1}}] \leqslant I_{S_{n+1}}[f_{\varepsilon_c,\mu}] + \varepsilon_{n+1}^{\mathrm{E}}.$$

Taking expectations with respect to S_{n+1} and substituting in equation (3.4), we get

$$\forall\mu \quad \eta_\mu \leqslant \mathbb{E}_S I[f_S] \leqslant \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{\varepsilon_c,\mu}] + \varepsilon_{n+1}^{\mathrm{E}} + \beta_{PH}^{(n+1)}.$$

Now we make the following observations. First, $\lim_{n\to\infty} \varepsilon_{n+1}^{\mathrm{E}} = 0$. Second, $\lim_{n\to\infty} \beta_{PH}^{(n)} = 0$. Finally, by considering the fixed function $f_{\varepsilon_c,\mu}$, we get

$$\forall\mu \quad \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{\varepsilon_c,\mu}] = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}_{S_{n+1}} V(f_{\varepsilon_c,\mu}, z_i) = I[f_{\varepsilon_c,\mu}] \leqslant \eta_\mu + \varepsilon_c.$$

Therefore, for every fixed $\varepsilon_c > 0$, for n sufficiently large,

$$\forall\mu \quad \eta_\mu \leqslant \mathbb{E}_S I[f_S] \leqslant \eta_\mu + \varepsilon_c$$

from which we conclude, for every fixed $\varepsilon_c > 0$,

$$0 \leqslant \liminf_{n\to\infty} \sup_\mu \big(\mathbb{E}_S I[f_S] - \eta_\mu\big) \leqslant \limsup_{n\to\infty} \sup_\mu \big(\mathbb{E}_S I[f_S] - \eta_\mu\big) \leqslant \varepsilon_c.$$

From this it follows that $\lim_{n\to\infty} \sup\limits_\mu (\mathbb{E}_S I[f_S] - \eta_\mu) = 0$. Consider the random variable $X_S = I[f_S] - \eta_\mu$. Clearly, $X_S \geqslant 0$. Also, $\lim_{n\to\infty} \sup\limits_\mu \mathbb{E}_S X_S = 0$. Therefore, we have (from Markov's inequality applied to X_S):

For every $\alpha > 0$,

$$\lim_{n\to\infty} \sup_\mu \mathbb{P}\big[I[f_S] > \eta_\mu + \alpha\big] = \lim_{n\to\infty} \sup \mathbb{P}[X_S > \alpha] \leqslant \lim_{n\to\infty} \sup \frac{\mathbb{E}_S[X_S]}{\alpha} = 0.$$

This proves distribution independent convergence of $I[f_S]$ to η_μ (consistency), given $\mathrm{CV_{loo}}$ stability. □

Theorem 3.16. Consistency of ERM implies $\mathrm{CV_{loo}}$ stability of ERM when the loss is bounded.

Proof. $\mathrm{CV_{loo}}$ stability and *PH* are equivalent when the loss is bounded by lemma 3.3. To show *PH* stability, we need to show that

$$\lim_{n\to\infty} \sup_\mu \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] = 0.$$

From lemma 3.13,

$$\forall\mu \quad \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] \leqslant \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S] + 4(n-1)\varepsilon^{\mathrm{E}}_{n-1}. \tag{3.5}$$

Given (universal) consistency, theorem 2.5 implies that $\mathcal{L}$ is a uGC class. Because $\mathcal{L}$ is uGC, $I[f_{S^i}]$ is close to $I_S[f_{S^i}]$. Because we are performing ERM, $I_S[f_{S^i}]$ is close to $I_S[f_S]$. Combining these results, $I[f_{S^i}] - I_S[f_S]$ is small.

We start with the equality

$$\mathbb{E}_S\big[I[f_{S^i}] - I_S[f_S]\big] = \mathbb{E}_S\big[I[f_{S^i}] - I_S[f_{S^i}]\big] + \mathbb{E}_S\big[I_S[f_{S^i}] - I_S[f_S]\big]. \tag{3.6}$$

Since $\mathcal{L}$ is uGC, with probability at least $1 - \delta_n(\varepsilon_n)$,

$$\big|I[f_{S^i}] - I_S[f_{S^i}]\big| \leqslant \varepsilon_n$$

and therefore

$$\forall\mu \quad \mathbb{E}_S\big[I[f_{S^i}] - I_S[f_{S^i}]\big] \leqslant \mathbb{E}_S\big[\big|I[f_{S^i}] - I_S[f_{S^i}]\big|\big] \leqslant \varepsilon_n + M\delta_n(\varepsilon_n). \tag{3.7}$$

From lemma 3.17, we have

$$\forall\mu \quad \mathbb{E}_S\big[I_S[f_{S^i}] - I_S[f_S]\big] \leqslant \frac{M}{n} + \varepsilon^{\mathrm{E}}_{n-1}. \tag{3.8}$$

Combining equation (3.6) with inequalities (3.7) and (3.8), we get

$$\forall\mu \quad \mathbb{E}_S\big[I[f_{S^i}] - I_S[f_S]\big] \leqslant \varepsilon_n + M\delta_n(\varepsilon_n) + \frac{M}{n} + \varepsilon^{\mathrm{E}}_{n-1}.$$

From inequality (3.5), we obtain

$$\forall\mu \quad \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] \leqslant \varepsilon_n + M\delta_n(\varepsilon_n) + \frac{M}{n} + \varepsilon^{\mathrm{E}}_{n-1} + 4(n-1)\varepsilon^{\mathrm{E}}_{n-1}.$$

Note that $\varepsilon^{\mathrm{E}}_n$ and ε_n may be chosen independently. Also, since we are guaranteed arbitrarily good ε-minimizers, we can choose $\varepsilon^{\mathrm{E}}_n$ to be a decreasing sequence such that $\lim_{n\to\infty}(4n-3)\varepsilon^{\mathrm{E}}_n = 0$.

Further, by lemma 3.18, it is possible to choose a sequence ε_n such that $\varepsilon_n \to 0$ and $\delta_n(\varepsilon_n) \to 0$. These observations taken together prove that

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{E}_S\big[\big|V(f_{S^i}, z_i) - V(f_S, z_i)\big|\big] = 0.$$

This proves that universal consistency implies *PH* hypothesis stability. □

Lemma 3.17. Under almost ERM,

$$I_S[f_{S^i}] - I_S[f_S] \leqslant \frac{M}{n} + \varepsilon^{\mathrm{E}}_{n-1}.$$

Proof.

$$\begin{aligned} I_S[f_{S^i}] &= \frac{(n-1)I_{S^i}[f_{S^i}] + V(f_{S^i}, z_i)}{n} \\ &\leqslant \frac{(n-1)(I_{S^i}[f_S] + \varepsilon^{\mathrm{E}}_{n-1}) + V(f_{S^i}, z_i)}{n} \quad \text{(by almost ERM)} \\ &= \frac{(n-1)I_{S^i}[f_S] + V(f_S, z_i) - V(f_S, z_i) + V(f_{S^i}, z_i)}{n} + \frac{n-1}{n}\varepsilon^{\mathrm{E}}_{n-1} \\ &\leqslant I_S[f_S] + \frac{M}{n} + \varepsilon^{\mathrm{E}}_{n-1} \quad \text{since } 0 \leqslant V \leqslant M. \end{aligned}$$

□

Lemma 3.18. If $\mathcal{L}$ is uGC, there exists a sequence $\varepsilon_n > 0$ such that:

(1) $\lim_{n\to\infty} \varepsilon_n = 0$,

(2) $\lim_{n\to\infty} \delta_n(\varepsilon_n) = 0$.

Proof. Because $\mathcal{L}$ is uGC,

$$\sup_{\mu} \mathbb{P}\Big(\sup_{f\in\mathcal{H}} \big|I[f] - I_S[f]\big| > \varepsilon\Big) \leqslant \delta_n(\varepsilon),$$

where $\lim_{n\to\infty} \delta_n(\varepsilon) = 0$.

For every fixed ε we know that $\lim_{n\to\infty}\delta_n(1/k) = 0$ for every fixed integer k.

Let N_k be such that for all $n \geqslant N_k$, we have $\delta_n(1/k) < 1/k$. Note, that for all $i > j$, $N_i \geqslant N_j$.

Now choose the following sequence for ε_n. We take $\varepsilon_n = 1$ for all $n < N_2$; $\varepsilon_n = \frac{1}{2}$ for $N_2 \leqslant n < N_3$ and in general $\varepsilon_n = 1/k$ for all $N_k \leqslant n < N_{k+1}$.

Clearly ε_n is a decreasing sequence converging to 0. Further, for all $N_k \leqslant n < N_{k+1}$, we have

$$\delta_n(\varepsilon_n) = \delta_n\left(\frac{1}{k}\right) \leqslant \frac{1}{k}.$$

Clearly $\delta_n(\varepsilon_n)$ also converges to 0. □

Remarks.

1. $\mathrm{CV}_{\mathrm{loo}}$ stability implies that the leave-one-out error converges to the training error in probability.

2. In general the bounds above are not exponential in δ. However, since for ERM $\mathrm{CV}_{\mathrm{loo}}$ stability implies that $\mathcal{L}$ is uGC, the standard uniform bound holds, which for any given ε is exponential in δ

$$\sup_{\mu}\mathbb{P}\left\{\sup_{f\in\mathcal{H}}\left|I[f] - I_S[f]\right| > \varepsilon\right\} \leqslant C\mathcal{N}\left(\frac{\varepsilon(n)}{8}, \mathcal{H}\right)\mathrm{e}^{-n\varepsilon^2/(8M^2)}.$$

   Notice that the covering number can grow arbitrarily fast in $1/\varepsilon$ resulting in an arbitrarily slow rate of convergence between $I_S[f]$ and $I[f]$.

3. It is possible to define a one-sided version of *PH* stability, called here *pseudo-PH stability*.

Definition 3.19. The learning map L is distribution-independent, *pseudo-pointwise hypothesis stable* if for each n there exists a $\beta_{pPH}^{(n)}$

$$\forall i \in \{1, \ldots, n\}, \forall\mu \quad \mathbb{E}_S\big[V(f_{S^i}, z_i) - V(f_S, z_i)\big] \leqslant \beta_{pPH}^{(n)},$$

with $\beta_{pPH}^{(n)}$ going to zero for $n \to \infty$.

Pseudo-stability is also *necessary and sufficient for universal consistency of ERM*. Pseudo-stability is weaker than *PH* stability. The proof of its equivalence to consistency of ERM is immediate from its definition. However, for non-ERM algorithms pseudo-*PH* stability is *not* sufficient in our approach to ensure convergence in probability of the empirical to the expected risk (e.g., generalization), when combined with $\mathrm{Eloo}_{\mathrm{err}}$ stability.[⋆]

[⋆] With pseudo-*PH* stability alone, we are unable to bound the second term in the decomposition of lemma 3.7.

3.4.3. *Consistency of ERM implies Eloo*$_{\mathrm{err}}$ *stability*

We now show that consistency of ERM implies Eloo$_{\mathrm{err}}$ stability.

Lemma 3.20. ERM on a uGC class implies

$$\mathbb{E}_S\left(I[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant \beta_n,$$

where $\lim_{n\to\infty} \beta_n = 0$.

Proof. By the triangle inequality and inspection

$$\mathbb{E}_S\left(I[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant 2\mathbb{E}_S\bigl(I[f_S] - I_S[f_S]\bigr)^2 + 2\mathbb{E}_S\left(I_S[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2.$$

We first bound the first term. Since we have are performing ERM on a uGC class we have with probability $1 - \delta_1$

$$\bigl|I_S[f_S] - I[f_S]\bigr| \leqslant \beta_1.$$

Therefore,

$$\mathbb{E}_S\bigl(I[f_S] - I_S[f_S]\bigr)^2 \leqslant M\beta_1 + M^2\delta_1.$$

The following inequality holds for the second term (see proof of lemma 3.7)

$$\mathbb{E}_S\left(I_S[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant M\mathbb{E}_S\bigl|V(f_S, z_i) - V(f_{S^i}, z_i)\bigr|.$$

Since ERM is on a uGC class (β_2, δ_2) CV$_{\mathrm{loo}}$ stability holds, implying

$$M\mathbb{E}_S\bigl|V(f_S, z_i) - V(f_{S^i}, z_i)\bigr| \leqslant M\beta_2 + M^2\delta_2.$$

Therefore we obtain

$$\mathbb{E}_S\left(I_S[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant M\beta_2 + M^2\delta_2$$

leading to

$$\mathbb{E}_S\left(I[f_S] - \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}, z_i)\right)^2 \leqslant 2M\beta_1 + 2M^2\delta_1 + 2M\beta_2 + 2M^2\delta_2. \qquad \square$$

Remark. In the classical literature on generalization properties of local classification rules [9] hypothesis stability was proven (and used) to imply Eloo_{err} stability. It is thus natural to ask whether we could replace Eloo_{err} stability with hypothesis stability in theorem 3.9. Unfortunately, we have been unable to either prove that ERM on a uGC class has hypothesis stability or provide a counterexample. The question remains therefore open. It is known that *ERM on a uGC class has hypothesis stability when either* (a) $\mathcal{H}$ *is convex, or* (b) *the setting is realizable*,[⋆] *or* (c) $\mathcal{H}$ *is finite-dimensional.*

4. Stability conditions, convergence rates and size of uGC classes

The previous section concludes the main body of the paper. This section consists of a few "side" observations. It is possible to provide rates of convergence of the empirical risk to the expected risk as a function of CV_{loo} stability using proposition 3.8. In general, these rates will be very slow, even in the case of ERM.

CV_{loo} and Eloo_{err} stability together ensure the convergence of $I[f_S]$ to $I[f_S]$ for non-ERM algorithms by controlling simultaneously both the expectation and variance of the quantity $V(f_{S^i}, z_i) - V(f_S, z_i)$[⋆⋆] and the difference between expected and leave-one-out error.

In this section we outline how CV_{loo} stability can be used to control the expectation and another notion of stability, *error stability*, can be used to control the variance. The two notions of stability together will be called here *strong stability* when the rate of convergence of error stability is fast enough. Strong stability yields faster rates of convergence of the empirical error to the expected error. In this section we define strong stability and list several "small" hypothesis spaces for which ERM is strongly stable.

The following definition of the continuity of the learning map L is based upon a variation of two definitions of stability first introduced in [15].

Definition 4.1. The learning map L is *strongly stable* if

(a) it is $(\beta_{\text{loo}}, \delta_{\text{loo}})$ CV_{loo} stable,

(b) it is error stable with a fast rate, e.g., for each n there exists a $\beta_{\text{error}}^{(n)}$ and a $\delta_{\text{error}}^{(n)}$ such that

$$\forall i \in \{1, \ldots, n\},\ \forall \mu \quad \mathbb{P}_S\bigl\{\bigl|I[f_S] - I[f_{S^i}]\bigr| \leqslant \beta_{\text{error}}^{(n)}\bigr\} \geqslant 1 - \delta_{\text{error}}^{(n)},$$

where $\beta_{\text{error}}^{(n)} = \mathrm{O}(n^{-\alpha})$ where $\alpha > 1/2$ and $\delta_{\text{error}}^{(n)} = \mathrm{e}^{-\Omega(n)}$.

The following theorem is similar to theorem 6.17 in [15].

[⋆] We say that the setting is realizable when there is some $f_0 \in \mathcal{H}$ which is consistent with the examples.

[⋆⋆] For ERM – because of almost positivity – it is sufficient that the expectation of the above quantity is controlled to ensure convergence.

Theorem 4.2. If the learning map is strongly stable then, for any $\varepsilon > 0$,

$$\mathbb{P}_S\left\{\left|I_S[f_S] - I[f_S]\right| \geqslant \varepsilon + \beta_{\text{loo}} + M\delta_{\text{loo}} + \beta_{\text{error}} + M\delta_{\text{error}}\right\} \leqslant 2\left(\exp\left(\frac{-\varepsilon^2 n}{8(2n\beta_{\text{error}} + M)^2}\right) + \frac{n(n+1)2M\delta_{\text{error}}}{2n\beta_{\text{error}} + M}\right),$$

where M is a bound on the loss.

The above bound states that with high probability the empirical risk converges to the expected risk at the rate of the slower of the two rates β_{loo} and β_{error}. The probability of the lack of convergence decreases exponentially as n increases. The proof of the above theorem is in [17, appendix 6.2.2]. For the empirical risk to converge to the expected risk in the above bound β_{error} must decrease strictly faster than $\mathrm{O}(n^{-1/2})$. For ERM the rate of convergence of β_{error} is the same rate as the convergence of the empirical error to the expected error.

Error stability with a fast rate of convergence is a strong requirement. In general, for a uGC class the rate of convergence of error stability can be arbitrarily slow because the covering number associated with the function class can grow arbitrarily fast[⋆] with ε^{-1}. Even for hypothesis spaces with VC dimension d the rate of convergence of error stability is not fast enough, with probability $1 - \mathrm{e}^{-t}$

$$I[f_S] - I[f_{S^i}] \leqslant \mathrm{O}\left(\sqrt{\frac{d \log n}{n}} + \sqrt{\frac{t}{n}}\right).$$

Fast rates for error stability can be achieved for ERM with certain hypothesis spaces and settings:

- ERM on VC classes of indicator functions in the realizable setting;[⋆⋆]
- ERM with square loss function on balls in Sobolev spaces $H^s(X)$, with compact $X \subset \mathbb{R}^d$, if $s > d$ (this is due to [6, proposition 6]);
- ERM with square loss function on balls or in RKHS spaces with a kernel K which is C^{2s} with $s > d$ (this is can be inferred from [28]);
- ERM on VC-subgraph classes that are convex with the square loss.

[⋆] Take a compact set K of continuous functions in the sup norm, so that $N(\varepsilon, K)$ is finite for all $\varepsilon > 0$. The set is uniform Glivenko–Cantelli. $N(\varepsilon, K)$ can go to infinity arbitrarily fast as $\varepsilon \to 0$ in the sup norm (Dudley, pers. com.).

[⋆⋆] This case was considered in [15, theorem 7.4].

Theorem. Let $\mathcal{H}$ be a space of ± 1-classifiers. The following are equivalent

1. There is a constant K such that for any distribution Δ on Z and any $f_0 \in \mathcal{H}$, ERM over $\mathcal{H}$ is $(0, \mathrm{e}^{-Kn})$ CV stable with respect to the distribution on Z generated by Δ and f_0.
2. The VC dimension of $\mathcal{H}$ is finite.

A requirement for fast rates of error stability is that the class of functions $\mathcal{H}$ is "small": hypothesis spaces with with empirical covering numbers $\mathcal{N}(\varepsilon, \mathcal{H})$ that are polynomial in ε^{-1} (VC classes fall into this category) or exponential in ε^{-p} with $p < 1$ (Sobolev spaces and RKHS spaces fall into this category). Simply having a "small" function class is not enough for fast rates: added requirements such as either the realizable setting or assumptions on the convexity of $\mathcal{H}$ and square loss are needed.

There are many situations where convergence of the empirical risk to the expected risk can have rates of the order of $\mathrm{O}(\sqrt{d/n})$ using standard VC or covering number bounds, here d is the metric entropy or shattering dimension of the class $\mathcal{H}$. For these cases we do not have stability based bounds that allow us to prove rates of convergence of the empirical error to the expected error faster than the polynomial bound in theorem 3.8 which gives suboptimal rates that are much slower than $\mathrm{O}(\sqrt{d/n})$. The following cases fall into the gap between general uGC classes that have slow rates of convergence[⋆] and those classes that have a fast rate of convergence:[⋆⋆]

- ERM on convex hulls of VC classes.
- ERM on balls in Sobolev spaces $H^s(X)$ if $2s > d$, which is the condition that ensures that functions in the space are defined pointwise – a necessary requirement for learning. In this case the standard union bounds give rates of convergence $\Omega((1/n)^b)$: for the general case $b = 1/4$ and for the convex case $b = 1/3$.
- ERM on VC classes of indicator functions in the non-realizable setting.

5. Discussion

The results of this paper are interesting from two quite different points of view. From the point of view (A) of the foundations of learning theory, they provide a condition – LOO stability – that extends the classical conditions beyond ERM and subsumes them in the case of ERM. From the point of view (B) of inverse problems, our results show that the conditions of well-posedness of the algorithm (specifically stability), and the condition of predictivity (specifically generalization) that played key but independent roles in the development of learning theory and learning algorithms respectively, are in fact closely related: well-posedness (defined in terms of LOO stability) implies predictivity and it is equivalent to it for ERM algorithms.

A. Learning techniques start from the basic and old problem of fitting a multivariate function to measurement data. The characteristic feature central to the learning framework is that the fitting should be *predictive*, in the same way that cleverly fitting data from an experiment in physics can uncover the underlying physical law, which should then be usable in a predictive way. In this sense, the same generalization results of learning theory also characterize the conditions under which predictive and therefore

[⋆] Obtained using either standard covering number bounds or proposition 3.8.

[⋆⋆] Obtained using either standard covering number bounds or strong stability.

scientific "theories" can be extracted from empirical data (see [26]). It is surprising that a form of stability turns out to play such a key role in learning theory. It is somewhat intuitive that stable solutions are predictive but it is surprising that our specific definition of $\mathrm{CV}_{\mathrm{loo}}$ stability fully subsumes the classical *necessary and sufficient* conditions on $\mathcal{H}$ for consistency of ERM.

LOO stability and its properties may suggest how to develop learning theory beyond the ERM approach. It is a simple observation that LOO stability can provide generalization bounds for algorithms other than ERM. For some of them a "VC-style" analysis in terms of complexity of the hypothesis space can still be used; for others, such as k-nearest neighbor, such an analysis is impossible because the hypothesis space has unbounded complexity or is not even defined, whereas $\mathrm{CV}_{\mathrm{loo}}$ stability can still be used.

B. Well-posedness and, in particular, stability are at the core of the study of inverse problems and of the techniques for solving them. The notion of $\mathrm{CV}_{\mathrm{loo}}$ stability may be a tool to bridge learning theory and the broad research area of the study of inverse problems in applied math and engineering (for a review see [12]). As we mentioned in the introduction, while predictivity is at the core of classical learning theory, another motivation drove the development of several of the best existing algorithms (such as regularization algorithms of which SVMs are a special case): well-posedness and, in particular, stability of the solution. These two requirements – consistency and stability – have been treated so far as "de facto" separate and in fact there was no a priori reason to believe that they are related (see [20]). Our new result shows that these two apparently different motivations are closely related and actually completely equivalent for ERM.

Some additional remarks and open questions are:

1. It would be interesting to analyze LOO stability properties – and thereby estimate bounds on rate of generalization – of several non-ERM algorithms. Several observations can be already inferred from existing results. For instance, the results of [4] imply that regularization and SVMs are LOO stable; a version of bagging with the number k of regressors increasing with n (with $(k/n) \to 0$) is $\mathrm{CV}_{\mathrm{loo}}$ stable and has hypothesis stability (because of [9]) and thus LOO stable; similarly k-NN with $k \to \infty$ and $(k/n) \to 0$ and kernel rules with the width $h_n \to 0$ and $h_n n \to \infty$ are LOO stable. Thus all these algorithms satisfy proposition 3.8 and have the generalization property, that is $I_S[f_S]$ converges to $I[f_S]$ (and some are also universally consistent).

2. The rate of convergence of the empirical error to the expected error for the empirical minimizer for certain hypothesis spaces differ, depending on whether we use the stability approaches or measures of the complexity of the hypothesis space, for example VC dimension or covering numbers. This discrepancy is illustrated by the following two gaps.

   (a) The hypothesis spaces in section 4 that have a fast rate of error stability have a rate of convergence of the empirical error of the minimizer to the expected error at a rate of $\mathrm{O}(d/n)$, where d is the VC dimension or metric entropy. This rate is obtained using VC-type bounds. The strong stability approach, which uses a variation of McDiarmid's inequality, gives a rate of convergence of $\mathrm{O}(n^{-1/2})$. It may be possible to improve these rates using inequalities of the type in [8].

(b) For the hypothesis spaces described at the end of section 4 standard martingale inequalities cannot be used to prove convergence of the empirical error to the expected error for the empirical minimizer.

It is known that martingale inequalities do not seem to yield results of optimal order in many situations (see [23]). A basic problem in the martingale inequalities is how variance is controlled. Given a random variable $Z = f(X_1, \dots, X_n)$ the variance of this random variable is controlled by a term of the form of

$$\mathrm{Var}(Z) \leqslant \mathbb{E}\left[\sum_{i=1}^{n}\bigl(Z - Z^{(i)}\bigr)^2\right],$$

where $Z^{(i)} = f(X_1, \dots, X_i', \dots, X_n)$. If we set $Z = I_S[f_S] - I[f_S]$ then for a function class with VC dimension d the upper bound on the variance is a constant since

$$\mathbb{E}\bigl[\bigl(Z - Z^{(i)}\bigr)^2\bigr] = K\frac{d}{n}.$$

However, for this class of functions we know that

$$\mathrm{Var}\bigl(I_S[f_S] - I[f_S]\bigr) = \Theta\left(\sqrt{\frac{d \log n}{n}}\right).$$

It is an open question if some other concentration inequality can be used to recover optimal rates.

3. We have a direct proof of the following statement for ERM: *If $\mathcal{H}$ has infinite VC dimension, then $\forall n$, $(\beta_{PH})_n > \frac{1}{8}$.* This shows that distribution-free β_{PH} does not converge to zero if $\mathcal{H}$ has infinite VC dimension and therefore provides a direct link between VC and $\mathrm{CV}_{\mathrm{loo}}$ stability (instead of via consistency).

4. Our results say that for ERM, distribution-independent $\mathrm{CV}_{\mathrm{loo}}$ stability is equivalent to the uGC property of $\mathcal{L}$. What can we say about compactness? Compactness is a stronger constraint on $\mathcal{L}$ than uGC (since compact spaces are uGC but not vice versa). Notice that the compactness case is fundamentally different because a compact $\mathcal{H}$ is a metric space, whereas in our main theorem we work with spaces irrespectively of their topology. The specific question we ask is whether there exists a stability condition that is related to compactness – as $\mathrm{CV}_{\mathrm{loo}}$ stability is related to the uGC property. Bousquet and Elisseeff showed that Tikhonov regularization (which enforces compactness but is NOT empirical risk minimization) gives uniform stability (with fast rate). Kutin and Niyogi showed that Bousquet and Elisseeff's uniform stability is unreasonably strong for ERM and introduced the weaker notion of (β, δ)-hypothesis stability in equation (2.11). It should also be noted (observation by Steve Smale)

that both these definitions of stability effectively require a hypothesis space with the sup norm topology. The following theorems illustrate some relations. For these theorems we assume that the hypothesis space $\mathcal{H}$ is a bounded subset of $C(X)$ where X is a closed, compact subset $X \in \mathbb{R}^k$ and Y is a closed subset $Y \in \mathbb{R}$.

Theorem 5.1. Given (β, δ)-hypothesis stability for ERM with the square loss, the hypothesis space $\mathcal{H}$ is compact.

Theorem 5.2. If $\mathcal{H}$ is compact and convex then ERM with the square loss is (β, δ)-hypothesis stable under regularity conditions of the underlying measure.

The proofs of the above theorems and the regularity condition required are in [17]. The theorems are not symmetric, since the second requires convexity and constraints on the measure. Thus they do not answer in a satisfactory way the question we posed about compactness and stability. In fact it can be argued on general grounds that compactness is not an appropriate property to consider in connection with hypothesis stability (Mendelson, pers. com.).

Finally, the search for "simpler" conditions than LOO stability is open. LOO stability answers all the requirements we need: it is sufficient for generalization in the general setting and subsumes the classical theory for ERM, since it is equivalent to consistency of ERM. There are other "simple" stability conditions equivalent to LOO stability [17, 19]. A prime candidate would be to replace Eloo_{err} stability with a "strong" condition such as hypothesis stability. We know that hypothesis stability implies Eloo_{err} stability; we do not know whether or not ERM on a uGC class implies hypothesis stability. Alternatively, it may be possible to replace Eloo_{err} stability with a "weak" condition such as error stability, which is implied by ERM on a uGC class. The open question here would be to show that the new condition together with CV_{loo} stability is sufficient for generalization in the general setting.

Appendix A

A.1. Necessary and sufficient conditions for a class of functions to be uniform Glivenko–Cantelli

In this section we state two conditions each of which is necessary and sufficient for a function class to be a uGC class. The first condition is on the metric entropy of the class and the second is on a combinatorial quantity of the class. For technical reasons some of the conditions are stated not for the class $\mathcal{H}$ but the class $\widetilde{\mathcal{H}} = \{f - \inf f\colon f \in \mathcal{H}\}$. For a uniformly bounded function class this difference is of no consequence.

The ε-covering number of the class $\widetilde{\mathcal{H}}$ is $\mathcal{N}(\varepsilon, \widetilde{\mathcal{H}}, l^p_{x_n})$ where $p \in [1, \infty)$ and $l^p_{x_n}$ is the empirical l^p distance on points x_n. The metric entropy of the class is

$$H_{n,p}(\varepsilon, \widetilde{\mathcal{H}}) = \sup_{x_n \in X^n} \log \mathcal{N}(\varepsilon, \widetilde{\mathcal{H}}, l^p_{x_n}).$$

A class of functions γ-shatters a set $A = \{x_i, \dots, x_n\}$ if for some point $x \in A$ there exists a level $\alpha(x) \in \mathbb{R}$ such that, given a subset B of A, we can find a function in $\mathcal{H}$ such that $f(x) \leqslant \alpha(x)$ for all $x \in B$ and $f(x) \geqslant \alpha(x) + \gamma$ for $x \in A \backslash B$. The V_γ or γ-shattering dimension of $\mathcal{H}$ is the cardinality of the smallest set that cannot be γ-shattered in all possible ways.

If the function class $\mathcal{H}$ is uniformly bounded then the following statements are equivalent assuming certain measurability conditions ($\widetilde{\mathcal{H}}$ is image admissible [11]):

1. $\widetilde{\mathcal{H}}$ is a weak (convergence in probability) uniform Glivenko–Cantelli class,
2. $\widetilde{\mathcal{H}}$ is a strong (almost sure convergence) uniform Glivenko–Cantelli class,
3. $\lim_{n\to\infty}(H_{n,p}(\varepsilon, \widetilde{\mathcal{H}})/n) = 0$ for all $\varepsilon > 0$,
4. $V_\gamma(\mathcal{H})$ is finite for all $\gamma \geqslant 0$.

The first three statements are from [11] and the last from [1].

Uniform Glivenko–Cantelli classes were first characterized for classes $\mathcal{H}$ of indicator functions. In this case, $\mathcal{H}$ is a uGC class if and only if it has finite VC($\mathcal{H}$) dimension. The "if" part of the statement was proven by Vapnik and Červonenkis [27] and the "only if" part was proven by Assouad and Dudley [2]. For uniformly bounded real-valued functions Dudley et al. [11] first stated necessary and sufficient conditions for uGC classes based upon an asymptotic condition of the metric entropy. Alon et al. [1] stated necessary and sufficient conditions for uGC classes based upon finiteness of $V_\gamma(\mathcal{H})$. This same property was used by Talagrand [22] but for problems in convex geometry. $V_\gamma(\mathcal{H})$ reduces to VC($\mathcal{H}$) when we set the scale parameter $\gamma = 0$.

Acknowledgements

We thank for many critically helpful comments Ding Zhou, Steve Smale, Richard Dudley, Nicolo' Cesa-Bianchi, Gabor Lugosi, Shahar Mendelson and especially Dimitry Panchenko. We are also grateful to Massi (and Theos).

References

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

[2] P. Assouad and R.M. Dudley, Minimax nonparametric estimation over classes of sets, unpublished manuscript (1990).

[3] M. Bertero, T. Poggio and V. Torre, Ill-posed problems in early vision, Proc. IEEE 76 (1988) 869–889.

[4] O. Bousquet and A. Elisseeff, Algorithmic stability and generalization performance, in: *Neural Information Processing Systems*, Vol. 14, Denver, CO (2000).

[5] O. Bousquet and A. Elisseeff, Stability and generalization, J. Mach. Learning Res. (2001).

[6] F. Cucker and S. Smale, On the mathematical foundations of learning, Bulletin AMS 39 (2001) 1–49.

[7] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, Vol. 31 (Springer, New York, 1996).

[8] V. De La Pena, A general class of exponential inequalities for martingales and ratios, Ann. Probab. 27(1) (1999) 537–564.

[9] L. Devroye and T. Wagner, Distribution-free performance bounds for potential function rules, IEEE Trans. Inform. Theory 25(5) (1979) 601–604.

[10] R.M. Dudley, *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics (Cambridge Univ. Press, Cambridge, 1999).

[11] R.M. Dudley, E. Giné, and J. Zinn, Uniform and universal Glivenko–Cantelli classes, J. Theoret. Probab. 4 (1991) 485–510.

[12] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems* (Kluwer Academic, Dordrecht, 1996).

[13] T. Engniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.

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

[15] S. Kutin and P. Niyogi, Almost-everywhere algorithmic stability and generalization error, Technical Report TR-2002-03, University of Chicago (2002).

[16] S. Mendelson, Geometric parameters in learning theory (2003) submitted for publication.

[17] S. Mukherjee, P. Niyogi, T. Poggio and R. Rifkin, Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization, AI Memo 2002-024, Massachusetts Institute of Technology (2002).

[18] S. Mukherjee, R. Rifkin and T. Poggio, Regression and classification with regularization, in: *Nonlinear Estimation and Classification, Proc. of MSRI Workshop*, eds. D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick and B. Yu, *Lectures Notes in Statistics*, Vol. 171 (Springer, New York, 2002) pp. 107–124.

[19] T. Poggio, R. Rifkin, S. Mukherjee and P. Niyogi, General conditions for predictivity in learning theory, Nature 343 (February 2004) 644–647.

[20] T. Poggio and S. Smale, The mathematics of learning: Dealing with data, Notices Amer. Math. Soc. 50(5) (2003) 537–544.

[21] D. Pollard, *Convergence of Stochastic Processes* (Springer, Berlin, 1984).

[22] M. Talagrand, Type, infratype and the Elton–Pajor theorem, Invent. Math. 107 (1992) 41–59.

[23] M. Talagrand, A new look at independence, Ann. Probab. 24 (1996) 1–34.

[24] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-posed Problems* (W.H. Winston, Washington, 1977).

[25] L.G. Valiant, A theory of learnable, in: *Proc. of the 1984 STOC* (1984) pp. 436–445.

[26] V.N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[27] V.N. Vapnik and A.Y. Chervonenkis, On the uniform convergence of relative frequences of events to their probabilities, Theory Probab. Appl. 17(2) (1971) 264–280.

[28] D. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.