massachusetts institute of technology — computer science and artificial intelligence laboratory

Statistical Learning: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization

Sayan Mukherjee, Partha Niyogi,
Tomaso Poggio and Ryan Rifkin

AI Memo 2002-024
CBCL Memo 223

December 2002
Revised July 2003
Second revision December 2003

Abstract

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 leave-one-out stability, called *CVEEE$_{loo}$ stability*. Our main new results are two. We prove that for bounded loss classes CVEEE$_{loo}$ 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 generalization and consistency of ERM*. Thus CVEEE$_{loo}$ stability is a weak form of stability that represents a sufficient condition for generalization for general learning algorithms while subsuming the classical conditions for consistency of ERM. We discuss alternative forms of stability. In particular, we conclude that for ERM a certain form of well-posedness is equivalent to consistency.

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., The Whitaker Foundation, and the SLOAN Foundations.

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[2]. 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 funcion $f \in \mathcal{H}$ 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 expected risk to the minimum risk achievable by functions in $\mathcal{H}$; for ERM generalization is equivalent to consistency [1] and thus for ERM we will often speak of consistency meaning generalization and consistency). 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 later).

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, eg $L : Z^n \rightarrow \mathcal{H}$.

As a major example, let us consider the following, important case for learning due to Cucker and Smale [5]. Assume that the hypothesis space $\mathcal{H}$ is a compact subset of $C(X)$ with X a compact domain in Euclidean space[3]. Compactness ensures[4] the existence of the minimizer of the expected risk for each n and, if the risk functional is convex[5] and regularity conditions on the measure hold, its uniqueness [5, 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

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

[3]Our concept of generalization, ie convergence in probability of the expected error $I[f_S]$ to the empirical error $I_S[f_S]$, corresponds to the uniform estimate of the "defect" of Theorem B in [5] (in their setup); consistency of ERM corresponds to their Theorem C; we do not consider in this paper any result equivalent to their Theorem C*.

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

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

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[6] but not necessary condition for consistency as well [5].

Thus it is natural to ask the question of whether there is a definition of well-posedness, and specifically stability – 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[7] 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 [7] use certain notions of algorithmic stability to prove the consistency of learning algorithms like the k-nearest neighbors classifier. More recently, Kearns and Ron [12] investigated several notions of stability to develop generalization error bounds in terms of the leave one out error. Bousquet and Elisseeff [4] showed that *uniform hypothesis stability* of the learning algorithm may be used to provide exponential bounds on the 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 [14] who introduced a probabilistic notion of change-one 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) Model of Valiant [24].

However, the task of finding a correct characterization of stability that satisfies both (a) and (b) above is subtle and non-trivial. In Kutin and Niyogi (2002) [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 *Cross-validation, error and empirical error Leave-One-Out stability* or, in short, *CVEEE$_{loo}$ stability* –

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

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

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:

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 – CVEEE$_{loo}$ stability – that is sufficient for generalization for any algorithm and for ERM is necessary and sufficient not only for generalization but also for consistency. CVEEE$_{loo}$ stability may, in some ways, be more natural – and perhaps an easier starting point for empirical work[8]– 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 6.1) – 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 CVEEE$_{loo}$ stability in terms of three leave-one-out stability conditions: crossvalidation (CV$_{loo}$) stability, expected error (E$_{loo}$) stability and empirical error (EE$_{loo}$) stability. Of the three leave-one-out stability conditions, CV$_{loo}$ stability is the key one, while the other two are more technical and satisfied by most reasonable algorithms. We prove that CVEEE$_{loo}$ stability is sufficient for generalization for general algorithms. We then prove in painful details the sufficiency of CV$_{loo}$ stability for consistency of ERM and the necessity of it. Finally, we prove that CVEEE$_{loo}$ stability is necessary and sufficient for consistency of ERM. We also discuss alternative definitions of stability. 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 RKHS spaces 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 CV$_{loo}$ 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.

[8] In its distribution-dependent version.

2 Background: learning and ill-posed problems

For notation, definitions and some results, we will assume knowledge of a foundational paper [5] and other review papers [11, 17]. The results of [4, 14] are the starting point for our work. Our interest in stability was motivated by their papers and by our past work in regularization (for reviews see [11, 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* $T(x) = \int_Y y d\mu(y|x)$ mapping X into Y, with $\mu(y|x)$ the conditional probability measure on Y. There is an hypothesis space $\mathcal{H}$ of functions $f : X \to Y$.

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 (in $\mathcal{H}$) that evaluates at a new value x_{new} and (hopefully) predicts the associated value of y:

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

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

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

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 $z \in Z$ V is bounded*, $0 \leq V(f, z) \leq M$.

Given a function f, a loss function V, and a probability distribution μ over X, we define the generalization error that we call here *true 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

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

The basic requirement for any learning algorithm is *generalization*: the empirical error must be a good proxy of the expected error, that is the difference between the two must be "small". Mathematically this means that for the function f_S selected by the algorithm given a training set S

$$\lim_{n\to\infty} |I[f_S] - I_S[f_S]| = 0 \quad \text{in probability.}$$

An algorithm that guarantees good generalization for a given n will predict well, if its empirical error on the training set is small.
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. We will also denote by S, z the training set with the point z added to the n points of S.

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. Similar notation using expectations should be self-explanatory.
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

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

where μ_n is the empirical measure supported on the set $x_1, ..., x_n$. In this notation (see for example [17]) $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$, 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 $f_S^{\varepsilon^E}$ that almost minimizes the empirical risk over all functions $f \in \mathcal{H}$, that is for any given $\varepsilon^E > 0$:*

$$I_S[f_S^{\varepsilon^E}] \leq \inf_{f\in\mathcal{H}} I_S[f] + \varepsilon^E. \tag{1}$$

In the following, we will drop the dependence on ε^E in $f_S^{\varepsilon^E}$. Notice that the term "Empirical Risk Minimization" (see Vapnik [25]) is somewhat misleading: in general the minimum need not exist[9]. In fact, it is precisely for this reason[10] that we use the notion of almost minimizer or ε-minimizer, given in Equation (1) (following others e.g. [1, 17]), 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 $\ell(z) : X \times Y \to \mathbb{R}$ and we define $\mathcal{L} = \{\ell(f) : f \in \mathcal{H}, V\}$. 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 2.2 *A learning map is (universally, weakly) consistent if for any given* $\varepsilon_c > 0$

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left\{ I[f_S] > \inf_{f\in\mathcal{H}} I[f] + \varepsilon_c \right\} = 0. \tag{2}$$

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. In this paper we call consistency what is sometimes defined as weak, universal consistency [6]. 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:

[9] 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.

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

Definition 2.3 *Let $\mathcal{F}$ be a class of functions. $\mathcal{F}$ is a (weak) uniform Glivenko-Cantelli class if*

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left\{ \sup_{f\in\mathcal{F}} |\mathbb{E}_{\mu_n} f - \mathbb{E}_{\mu} f| > \varepsilon \right\} = 0. \tag{3}$$

There may be measurability issues that can be handled by imposing mild conditions on $\mathcal{F}$ (see [8, 9]).

When applied to the loss functions ℓ, the definition implies that for all distributions μ and for each ε_n there exist a $\delta_{\varepsilon_n,n}$ such that

$$\mathbb{P}\left\{ \sup_{\ell\in F} |I[\ell] - I_S[\ell]| > \varepsilon_n \right\} \leq \delta_{\varepsilon_n,n},$$

where the sequences ε_n and $\delta_{\varepsilon_n,n}$ go simultaneously to zero [11]. Later in the proofs we will take the sequence of ε_n^E (in the definition of ε-minimizer) to 0 with a rate faster than $\frac{1}{n}$, therefore faster than the sequence of ε_n (eg the ε_n in the uGC definition).
We are now ready to state the "classical" necessary and sufficient condition for consistency of ERM (from Alon et al., Theorem 4.2, part 3 [1], see also [25, 9]).

Theorem 2.1 *Assuming that the loss functions $\ell \in \mathcal{L}$ are bounded and the collection of functions $\{\ell - \inf_{\mathcal{L}} \ell : \ \ell \in \mathcal{L}\}$ are uniformly bounded*[12] *, a necessary and sufficient condition for consistency of ERM (with respect to all measures) 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 the uGC property of $\mathcal{L}$[13].
Notice that there is a definition of strong uGC classes where, instead of convergence in probability, almost sure convergence is required.

Definition 2.4 *Let $\mathcal{F}$ be a class of functions. $\mathcal{F}$ is a strong uniform Glivenko-Cantelli class if*

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left\{ \sup_{m\geq n} \sup_{f\in\mathcal{F}} |\mathbb{E}_{\mu_m} f - \mathbb{E}_{\mu} f| > \varepsilon \right\} = 0. \tag{4}$$

[11] 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.).

[12] These conditions will be satisfied for bounded loss functions $0 \leq \ell(z) \leq M$

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

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

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 $1 \leq p < \infty$.

For bounded loss functions weak uGC is equivalent to strong uGC (see Theorem 6 in [9]) and weak consistency is equivalent to strong consistency in Theorem 2.1. In the following, we will speak simply of uGC and consistency, meaning – strictly speaking – weak uGC and weak consistency.

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

$$g = Au. \tag{5}$$

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 [10]) 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 (5) the map A^{-1} is continuous on its domain Z if, given any $\varepsilon > 0$, there is a $\delta > 0$ such that for any $z', z'' \in Z$

$$||z' - z''|| \leq \delta$$

with the norm in Z, then

$$||A^{-1}z' - A^{-1}z''|| \leq \varepsilon,$$

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[14] that under certain conditions points to the compactness of $\mathcal{H}$ as sufficient for establishing stability and thus well-posedness[15]. Notice that when the solution of Equation (5) does not exist, the standard approach is to replace it with the following problem, analogous to ERM,

$$\min_{u \in \mathcal{H}} \|Au - g\| \tag{6}$$

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 (6) coincide with the solutions of

$$Au = Pg \tag{7}$$

where P is the projection onto $R(A)$.

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 $S = ((x_1, y_1), \ldots, (x_n, y_n))$ into a function f, that is

$$LS = f_S.$$

In Equation (5) 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 (7) and in which L is linear.

Assume that the x part of the n examples $(x_1, ..., x_n)$ is fixed; then L as an operator on $(y_1, ..., y_n)$ 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 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 [5, 26]. The ERM case corresponds to Equation (6) that is

$$\min_{f \in B_R} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2. \tag{8}$$

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 $\mathcal{H}$ (eg $\|f\|_K \leq R$). Then $\mathcal{H} = \overline{I_K(B_R)}$ is compact – where

[14] Lemma (Tikhonov, [23]) *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.*

[15] In learning, the approach underlying most algorithms such as RBF and SVMs is in fact regularization. These algorithms can therefore be directly motivated in terms of restoring well-posedness of the learning problem.

$I_K : \mathcal{H}_K \hookrightarrow C(X)$ is the inclusion and $C(X)$ is the space of continuous functions with the $\sup$ norm [5]. In this case the minimizer of the generalization error $I[f]$ is well-posed. Minimization of the empirical risk (Equation (8)) is also well-posed: it provides a set of linear equations to compute the coefficients $\mathbf{c}$ of the solution f as

$$K\mathbf{c} = \mathbf{y} \tag{9}$$

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

A particular form of regularization, called Tikhonov regularization, replaces ERM (see Equation (8)) with

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \gamma \|f\|_K^2, \tag{10}$$

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

$$(K + n\gamma I)\mathbf{c} = \mathbf{y}, \tag{11}$$

which for $\gamma = 0$ reduces to Equation (9). In this RKHS case, stability of the empirical risk minimizer provided by Equation (10) 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

$$\frac{\|\Delta f\|}{\|f\|} \leq \|K + n\gamma I\| \left\| (K + n\gamma I)^{-1} \right\| \frac{\|\Delta \mathbf{y}\|}{\|\mathbf{y}\|}, \tag{12}$$

where the condition number $\|K + n\gamma I\| \left\| (K + n\gamma I)^{-1} \right\|$ 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 (8) for any fixed n corresponds to the set of well-posed, linear equations (9), even without the constraint $\|f\|_K^2 \leq R$: 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.

2. Minimization of the functional in Equation (10) 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 (distribution-dependent) definition of stability that takes care of some of the issues above was introduced by [4] with the name of *uniform stability*:

$$\forall S \in Z^n, \forall i \in \{1, \ldots, n\} \quad \sup_{z \in Z} |V(f_S, z) - V(f_{S^i}, z)| \leq \beta, \tag{13}$$

Kutin and Niyogi [14] showed that ERM does not in general exhibit Bousquet and Elisseeff's definition of 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:

$$\mathbb{P}_S \left\{ \sup_{z \in Z} |V(f_S, z) - V(f_{S^i}, z)| \leq \beta \right\} \geq 1 - \delta, \tag{14}$$

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

Interestingly, the results of [4] imply that Tikhonov regularization algorithms are uniformly stable (and of course (β, δ)-hypothesis stable) with $\beta = O(\frac{1}{\gamma n})$. 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 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). In addition, it is interesting to note that the definitions of stability introduced above – and 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 $|\ell_{f_S} - \ell_{f_{S^i}}|$, eg 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. Since we want to deal with spaces without a topology, we propose here the following definition[16] of *leave-one-out cross-validation (in short, CV_{loo}) stability*, which is the key part in the notion of CVEEE_{loo} stability introduced later:

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

$$\forall i \in \{1, \ldots, n\} \quad \mathbb{P}_S \{|V(f_S, z_i) - V(f_{S^i}, z_i)| \leq \beta_{CV}\} \geq 1 - \delta_{CV},$$

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[17]. In turns, *uniform stability implies (β, δ)-hypothesis stability which implies 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 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 to be used to solve it. As we discussed earlier, both 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 [2]).

3 Stability, generalization and consistency of ERM

3.1 Probabilistic preliminaries

The following are easy consequences of our definitions and will be used without further mention throughout the paper:

$$\mathbb{E}_S[I[f_S]] = \mathbb{E}_{S,z}[V(f_S, z)]$$
$$\forall i \in \{1, \ldots, n\} \quad \mathbb{E}_S[I_S[f_S]] = \mathbb{E}_S[V(f_S, z_i)]$$

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

$$\forall i \in \{1,\dots,n\} \quad \mathbb{E}_S[I[f_{S^i}] = \mathbb{E}_S[V(f_{S^i}, z_i)]$$

3.2 Leave-one-out stability properties

This section introduces several definitions of stability – all in the leave-one-out form – and show the equivalence of two of them. The first definition of stability of the learning map L, is *Cross-Validation Leave-One-Out stability* which turns out to be the central one for most of our results. This notion of stability is based upon a variation of a definition of stability introduced in [14].

3.2.1 $\mathbf{CV}_{loo}$ stability

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

$$\forall i \in \{1,\dots,n\} \quad \forall \mu \;\; \mathbb{P}_S \left\{ |V(f_{S^i}, z_i) - V(f_S, z_i)| \leq \beta_{CV}^{(n)} \right\} \geq 1 - \delta_{CV}^{(n)},$$

with $\beta_{CV}^{(n)}$ and $\delta_{CV}^{(n)}$ 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 much weaker than the uniform stability condition [4] and is implied by it.
A definition which turns out to be equivalent was introduced by [4] (see also [12]) under the name of *pointwise hypothesis stability* or *PH stability*, which we give here in its distribution-free version:

Definition 3.2 *The learning map L has distribution-independent, pointwise hypothesis stability if for each n there exists a $\beta_{PH}^{(n)}$*

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

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

We now show that the two definitions of CV_{loo} stability and PH-stability are equivalent (in general, without assuming ERM).

Lemma 3.1 *CV_{loo} stability with β_{loo} and δ_{loo} implies PH stability with $\beta_{PH} = \beta_{loo} + M\delta_{loo}$; PH stability with β_{PH} implies CV_{loo} stability with $(\alpha, \frac{\beta_{PH}}{\alpha})$ for any $\alpha < \beta_{PH}$.*

PROOF:We give the proof for a given distribution. The result extends trivially to the distribution-free ($\forall \mu$) case. From the definition of CV_{loo} stability and the bound on the loss function it follows

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

This proves CV_{loo} stability implies PH stability.

From the definition of PH stability, we have

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

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

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

From this it is immediate that the learning algorithm is $(\alpha, \frac{\beta_{PH}}{\alpha})$ CV_{loo} stable. □

3.2.2 E_{loo} and EE_{loo} stability

We now introduce two rather weak and natural stability conditions which, unlike CV_{loo} stability, are not pointwise and should be satisfied by most reasonable algorithm. The first one – introduced and used by [12] – is leave-one-out stability of the expected error; the second one – which is similar to the *overlap stability* of Kutin and Niyogi – is leave-one-out stability of the empirical error. They are both very reasonable from the point of view of wellposedness of the problem. In particular, for ERM it would be indeed inconsistent if stability of the expected and empirical error were not true. Notice that *uniform stability*, which is the 'normal" definition of continuity for stability (stability in the L_∞ norm), implies both E_{loo} and EE_{loo} stability. In fact (β, δ)-hypothesis stability immediately gives CV_{loo} stability, hypothesis stability, E_{loo} stability, and EE_{loo} stability.

Definition 3.3 *The learning map L is distribution-independent, Error stable – in short E_{loo} stable – if for each n there exists a $\beta_{Er}^{(n)}$ and a $\delta_{Er}^{(n)}$ such that for all $i = 1...n$*

$$\forall \mu \;\; \mathbb{P}_S \left\{ |I[f_{S^i}] - I[f_S]| \leq \beta_{Er}^{(n)} \right\} \geq 1 - \delta_{Er}^{(n)},$$

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

Definition 3.4 *The learning map L is distribution-independent, Empirical error stable – in short EE_{loo} stable – if for each n there exists a $\beta_{EE}^{(n)}$ and a $\delta_{EE}^{(n)}$ such that for all $i = 1...n$*

$$\forall \mu \;\; \mathbb{P}_S \left\{ |I_{S^i}[f_{S^i}] - I_S[f_S]| \leq \beta_{EE}^{(n)} \right\} \geq 1 - \delta_{EE}^{(n)},$$

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

Since the loss function is bounded by M an equivalent definition of EE$_{loo}$ stability is: for each n there exists a $\beta_{EE}^{(n)}$ and a $\delta_{EE}^{(n)}$ such that for all $i = 1...n$

$$\forall \mu \;\; \mathbb{P}_S \left\{ |I_S[f_{S^i}] - I_S[f_S]| \leq \beta_{EE}^{(n)} \right\} \geq 1 - \delta_{EE}^{(n)},$$

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

The $\beta_{EE}^{(n)}$ in the two variants of the definition are within $\frac{M}{n}$ of each other.

3.2.3 CVEEE$_{loo}$ stability

As we will show, the combination of CV$_{loo}$, E$_{loo}$ and EE$_{loo}$ stability is sufficient for generalization for generic, symmetric algorithms and is necessary and sufficient for consistency of ERM. The following definition will be useful

Definition 3.5 *When a learning map L exhibits CV$_{loo}$, E$_{loo}$ and EE$_{loo}$ stability, we will say that it has CVEEE$_{loo}$ stability.*

Notice that uniform stability implies CVEEE$_{loo}$ stability but is not implied by it.

3.3 CVEEE$_{loo}$ stability implies generalization

In this section we prove that CVEEE$_{loo}$ stability is sufficient for generalization for general learning algorithms.
We first prove the following useful Lemma.

Lemma 3.2 *Given the following expectation*

$$\mathbb{E}_x[A_x(B_x - C_x)],$$

with the random variables $0 \leq A_x, B_x, C_x \leq M$ and random varaibles $0 \leq A'_x, B'_x, C'_x \leq M$ where

$$\begin{aligned} \mathbb{P}_x(|A'_x - A_x| > \beta_A) &\leq \delta_A \\ \mathbb{P}_x(|B'_x - B_x| > \beta_B) &\leq \delta_B \\ \mathbb{P}_x(|C'_x - C_x| > \beta_C) &\leq \delta_C, \end{aligned}$$

then

$$\begin{aligned} \mathbb{E}_x[A_x(B_x - C_x)] \;\leq\; & \mathbb{E}_x[A'_x(B'_x - C'_x)] + \\ & 3M\beta_A + 3M^2\delta_A + M\beta_B + M^2\delta_B + M\beta_C + M^2\delta_C. \end{aligned}$$

PROOF:
We first define the following three terms

$$\begin{aligned}\Delta_{A,x} &= A_x - A'_x \\ \Delta_{B,x} &= B_x - B'_x \\ \Delta_{C,x} &= C'_x - C_x.\end{aligned}$$

We now rewrite the expectation and use the fact that the random variables are nonnegative and bounded by M

$$\begin{aligned}\mathbb{E}_x[A_x(B_x - C_x)] &= \mathbb{E}_x[(A'_x + \Delta_{A,x})(B'_x + \Delta_{B,x} - C'_x + \Delta_{C,x})] \\ &\leq \mathbb{E}_x[A'_x(B'_x - C'_x)] + 3M\mathbb{E}_x|\Delta_{A,x}| + M\mathbb{E}_x|\Delta_{B,x}| + 2M\mathbb{E}_x|\Delta_{C,x}|.\end{aligned}$$

By the assumptions given

$$\begin{aligned}|\Delta_{A,x}| \leq \beta_A \text{ with probability } 1 - \delta_A \\ |\Delta_{B,x}| \leq \beta_B \text{ with probability } 1 - \delta_B \\ |\Delta_{C,x}| \leq \beta_C \text{ with probability } 1 - \delta_C.\end{aligned}$$

Sets of x for which $|\Delta_{A,x}| \leq \beta_A$ are called G (the fraction of sets for which this holds is $1 - \delta_A$) while the complement is called G^c

$$\begin{aligned}\mathbb{E}_x|\Delta_{A,x}| &= \mathbb{E}_{x\in G}|\Delta_{A,x}| + \mathbb{E}_{x\in G^c}|\Delta_{A,x}| \\ &\leq (1 - \delta_A)\beta_A + M\delta_A \\ &\leq \beta_A + M\delta_A.\end{aligned}$$

Therefore,

$$\mathbb{E}_x|\Delta_{A,x}| + \mathbb{E}_x|\Delta_{B,x}| + \mathbb{E}_x|\Delta_{C,x}| \leq \beta_A + M\delta_A + \beta_B + M\delta_B + \beta_C + M\delta_C. \ \square$$

We now prove that CVEEE$_{loo}$ stability implies generalization for symmetric algorithms.

Theorem 3.1 *If the symmetric learning map is CVEEE$_{loo}$ stable then with probability* $1 - \delta_{gen}$

$$|I[f_S] - I_S[f_S]| \leq \beta_{gen},$$

where

$$\delta_{gen} = \beta_{gen} = (2M\beta_{CV} + 2M^2\delta_{CV} + 3M\beta_{Er} + 3M^2\delta_{Er} + 5M\beta_{EE} + 5M^2\delta_{EE})^{1/4}.$$

PROOF:
We first state the properties implied by the assumptions. In all cases the probability is over S, z'.
CV stability: with probability $1 - \delta_{CV}$

$$|V(f_S, z') - V(f_{S,z'}, z')| \leq \beta_{CV}.$$

Error stability: with probability $1 - \delta_{Er}$

$$|\mathbb{E}_z V(f_S, z') - \mathbb{E}_z V(f_{S,z'}, z)| \leq \beta_{Er}.$$

Empirical error stability: with probability $1 - \delta_{EE}$

$$\left| \frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) - \frac{1}{n+1} \sum_{z_j \in S,z'} V(f_{S,z'}, z_j) \right| \leq \beta_{EE}.$$

Let us consider

$$\begin{aligned} \mathbb{E}_S(I[f_S] - I_S[f_S])^2 &= \mathbb{E}_S(I[f_S]^2 + I_S[f_S]^2 - 2I[f_S]I_S[f_S]) \\ &= \mathbb{E}_S[I[f_S](I[f_S] - I_S[f_S])] + \mathbb{E}_S[I_S[f_S](I_S[f_S] - I[f_S])]. \end{aligned}$$

We will only upper bound the two terms in the expansion above since a trivial lower bound on the above quantity is zero.

We first bound the first term

$$\begin{aligned} & \mathbb{E}_S \left[\mathbb{E}_z V(f_S, z) \left(\mathbb{E}_{z'} V(f_S, z') - \frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) \right) \right] \\ = \; & \mathbb{E}_{S,z'} \left[\mathbb{E}_z V(f_S, z) \left(V(f_S, z') - \frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) \right) \right]. \end{aligned}$$

Given the stability assumptions and Lemma 3.2

$$\begin{aligned} & \mathbb{E}_{S,z'} \left[\mathbb{E}_z V(f_S, z) \left(V(f_S, z') - \frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) \right) \right] \\ \leq \; & \mathbb{E}_{S,z'} \left[\mathbb{E}_z V(f_{S,z'}, z) \left(V(f_{S,z'}, z') - \frac{1}{n+1} \sum_{z_j \in S,z'} V(f_{S,z'}, z_j) \right) \right] \\ & + 3M\beta_{Er} + 3M^2\delta_{Er} + M\beta_{CV} + M^2\delta_{CV} + M\beta_{EE} + M^2\delta_{EE} \\ = \; & \mathbb{E}_{S,z,z'} \left[V(f_{S,z'}, z) \left(V(f_{S,z'}, z') - \frac{1}{n+1} \sum_{z_j \in S,z'} V(f_{S,z'}, z_j) \right) \right] \\ & + 3M\beta_{Er} + 3M^2\delta_{Er} + M\beta_{CV} + M^2\delta_{CV} + M\beta_{EE} + M^2\delta_{EE} \end{aligned}$$

By symmetry

$$\mathbb{E}_{S,z,z'} \left[V(f_{S,z'}, z) \left(V(f_{S,z'}, z') - \frac{1}{n+1} \sum_{z_j \in S,z'} V(f_{S,z'}, z_j) \right) \right] = 0,$$

since the expectation of the two terms in the inner parentheses is identical: both are measuring the error on a training sample from datasets of $n+1$ samples drawn i.i.d. This bounds the first term as follows

$$\mathbb{E}_S[I[f_S](I[f_S] - I_S[f_S])] \leq M(\beta_{CV} + M\delta_{CV} + 3\beta_{Er} + 3M\delta_{Er} + \beta_{EE} + M\delta_{EE}).$$

We now bound the second term

$$\begin{aligned} & \mathbb{E}_S \left[I_S[f_S] \left(\frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) - \mathbb{E}_z V(f_S, z) \right) \right] \\ = \quad & \mathbb{E}_{S,z} \left[I_S[f_S] \left(\frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) - V(f_S, z) \right) \right]. \end{aligned}$$

Given the stability assumptions and Lemma 3.2

$$\begin{aligned} & \mathbb{E}_{S,z} \left[I_S[f_S] \left(\frac{1}{n} \sum_{z_j \in S} V(f_S, z_j) - V(f_S, z) \right) \right] \\ \leq \quad & \mathbb{E}_{S,z} \left[I_{S,z}[f_{S,z}] \left(\frac{1}{n+1} \sum_{z_j \in S,z} V(f_{S,z}, z_j) - V(f_{S,z}, z) \right) \right] + 4M\beta_{EE} + 4M^2\delta_{EE} + M\beta_{CV} + M^2\delta_{CV} \\ = \quad & \mathbb{E}_{S,z} \left[V(f_{S,z}, z) \left(\frac{1}{n+1} \sum_{z_j \in S,z} V(f_{S,z}, z_j) - V(f_{S,z}, z) \right) \right] + 4M\beta_{EE} + 4M^2\delta_{EE} + M\beta_{CV} + M^2\delta_{CV} \end{aligned}$$

By symmetry

$$\mathbb{E}_{S,z} \left[V(f_{S,z}, z) \left(\frac{1}{n+1} \sum_{z_j \in S,z} V(f_{S,z}, z_j) - V(f_{S,z}, z) \right) \right] = 0.$$

This bounds the second term

$$\mathbb{E}_S[I_S[f_S](I_S[f_S] - I[f_S])] \leq M(\beta_{CV} + M\delta_{CV} + 4\beta_{EE} + 4M\delta_{EE}).$$

Combining the above terms gives

$$\mathbb{E}_S(I[f_S] - I_S[f_S])^2 \leq B,$$

where

$$B = 2M\beta_{CV} + 2M^2\delta_{CV} + 3M\beta_{Er} + 3M^2\delta_{Er} + 5M\beta_{EE} + 5M^2\delta_{EE}.$$

By the Bienaymé-Chebyshev inequality it follows that

$$\mathbb{P}(|I[f_S] - I_S[f_S]| > \delta) \leq \frac{B}{\delta^2}$$

for nonnegative δ, implying that, with probability $1 - \delta$,

$$|I[f_S] \leq I_S[f_S]| + \sqrt{\frac{B}{\delta}}.$$

Setting $\delta = \sqrt{B}$ gives us the result, eg with probability $1 - \delta_{gen}$

$$|I[f_S] - I_S[f_S]| \leq \beta_{gen},$$

where

$$\delta_{gen} = \beta_{gen} = B^{1/4}. \square$$

REMARKS:

1. CV_{loo}, E_{loo} and EE_{loo} stability together are strong enough to imply generalization for general algorithms, but neither condition by itself is sufficient.

2. CV_{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 $f^*(x) = 1$, 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:

   $$f_S(x) = \begin{cases} -1^n & \text{if n is a training point} \\ -1^{n+1} & \text{otherwise.} \end{cases}$$

   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) CV_{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.

3. E_{loo} and EE_{loo} stability by themselves are *not* sufficient for generalization, as the following counterexample shows. Using the same setup as in the previous remark, consider an algorithm which returns 1 at every training point, and -1 otherwise. This algorithm is EE_{loo} and E_{loo} stable (and hypothesis stable), but is not CV_{loo} stable and does not generalize.

4. In [4], Theorem 11, Elisseef and Bousquet claim that PH stability (which is equivalent to our CV_{loo} stability, by Lemma 3.1) is sufficient for generalization. However, there is an error in their proof. The second line of their theorem, translated into our notation, states correctly that

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

   Bousquet and Elisseef use PH stability 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 Elisseef and Bousquet proof can be easily "fixed" by bounding the second term using the more general notion of (non-pointwise) 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 3.1. Hypothesis stability is a strictly stronger notion than error stability and implies it. $Eloo_{err}$ stability (see later) does not imply hypothesis stability but is implied by it[18].

5. Notice that hypothesis stability and CV_{loo} stability imply generalization. Since hypothesis stability implies E_{loo} stability it follows that CV_{loo} stability together with hypothesis stability implies EE_{loo} stability (and generalization).

3.4 Alternative stability conditions

Our main result is in terms of $CVEEE_{loo}$ stability. There are however alternative conditions that together with CV_{loo} stability are also sufficient for generalization and necessary and sufficient for consistency of ERM. One such condition is *Expected-to-Leave-One-Out Error*, in short *$Eloo_{err}$ condition*.

[18] There is an unfortunate confusing proliferation of definitions of stability. The hypothesis stability of Elisseef and Bousquet is essentially equivalent to the L_1 stability of Kutin and Niyogi (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 Elisseef 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 Elisseef's hypothesis stability implies Kutin and Niyogi's weak-$L1$ 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 Elisseef and Bousquet's hypothesis stability).

Definition 3.6 *The learning map L is Eloo$_{err}$ stable in a distribution-independent way, if for each n there exists a $\beta_{EL}^{(n)}$ and a $\delta_{EL}^{(n)}$ such that*

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

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

Thinking of the Eloo$_{err}$ property as a form of stability may seem somewhat of a stretch (though the definition depends on a "perturbation" of the training set from S to S^i). It may be justified however by the fact that the Eloo$_{err}$ property is implied – in the general setting – by a classical leave-one-out notion of stability called *hypothesis stability*[19], which was introduced by DeVroye and Wagner [7] and later used by [12, 4] (and in a stronger change-one form by [14]).

Intuitively, the Eloo$_{err}$ condition seems both weak and strong. It looks weak because the leave-one-out error $\frac{1}{n} \sum_{i=1}^{n} V(f_{S^i}, z_i)$ 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 3.8 *When a learning map L exhibits both CV$_{loo}$ and Eloo$_{err}$ stability, we will say that it has LOO stability.*

3.4.1 LOO stability implies generalization

We now prove that CV$_{loo}$ and Eloo$_{err}$ stability together are sufficient for generalization for general learning algorithms. We will use the following lemma mentioned as Remark 10 in [4], of which we provide a simple proof[20].

Lemma 3.3 *Decomposition of the generalization error*

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

[19] Our definition of hypothesis stability – which is equivalent to leave-one-out stability in the L_1 norm – is:

Definition 3.7 *The learning map L has distribution-independent, leave-one-out hypothesis stability if for each n there exists a $\beta_H^{(n)}$*

$$\forall \mu \;\; \mathbb{E}_S \mathbb{E}_z [|V(f_S, z) - V(f_{S^i}, z)|] \leq \beta_H^{(n)},$$

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

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

PROOF:
By the triangle inequality and inspection

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

We now bound the second term

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

Using the decomposition of the generalization error $I[f_S]-I_S[f_S]$ provided by the lemma it is clear that

Proposition 3.1 *LOO stability implies generalization.*

REMARKS:

1. Eloo_{err} stability by itself is *not* sufficient for generalization, as a previous example showed (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).

2. The converse of Theorem 3.1 is false. Considering the same basic setup as the example 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 possesses none of CV_{loo} or Eloo_{err} (or E_{loo} or EE_{loo}) stability, but it will generalize.

3. CVEEE_{loo} stability implies convergence of $\frac{1}{n}\sum_{j=1}^{n}V(f_{S^j},z_j))$ to I, because we can use the decomposition lemma "in reverse", that is $(I-\frac{1}{n}\sum_{j=1}^{n}V(f_{S^j},z_j))^2 \leq (I-I_S)^2+(I_S-\frac{1}{n}\sum_{j=1}^{n}V(f_{S^j},z_j))^2$ and then use CV_{loo} stability to bound the second term.

We now turn (see section 3.5) to the question of whether CVEEE$_{loo}$ stability (or LOO stability) is general enough to capture the fundamental conditions for consistency of ERM and thus *subsume the "classical" theory*. We will in fact show in the next subsection (3.5) that CV$_{loo}$ stability alone is equivalent to consistency of ERM. To complete the argument, we will also show in subsection 3.5 that E$_{loo}$, EE$_{loo}$ stability (as well as Eloo$_{err}$) are implied by consistency of ERM. Thus CVEEE$_{loo}$ stability (as well as LOO stability) is implied by consistency of ERM.

3.5 CVEEE$_{loo}$ stability is necessary and sufficient for consistency of ERM

We begin showing that CV$_{loo}$ stability is necessary and sufficient for consistency of ERM.

3.5.1 Almost positivity of ERM

We first prove a lemma about the *almost positivity*[21] of $V(f_S, z_i) - V(f_{S^i}, z_i)$, where $|S| = n$, as usual.

Lemma 3.4 (Almost-Positivity) *Under the assumption that ERM finds a ε^E-minimizer,*

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

PROOF:By the definition of almost minimizer (see Equation (1)), we have

$$\frac{1}{n}\sum_{z_j \in S} V(f_{S^i}, z_j) - \frac{1}{n}\sum_{z_j \in S} V(f_S, z_j) \quad \geq \quad -\varepsilon_n^E \tag{15}$$

$$\frac{1}{n}\sum_{z_j \in S^i} V(f_{S^i}, z_j) - \frac{1}{n}\sum_{z_j \in S^i} V(f_S, z_j) \quad \leq \quad \frac{n-1}{n}\varepsilon_{n-1}^E \tag{16}$$

We can rewrite the first inequality as

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

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

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

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

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

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

The following lemma will be key in the proof of our main theorem.

Lemma 3.5 *For any $i \in \{1, 2, \ldots, n\}$, under almost ERM with $\varepsilon_n^E > 0$ chosen such that $\lim_{n\to\infty} n\varepsilon_n^E = 0$, the following (distribution free) bound holds*

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

PROOF:We note that

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

Now we make two observations. First, under the assumption of almost ERM, by Lemma 3.4,

$$\forall i \in \{1, \ldots, n\} \quad V(f_{S^i}, z_i) - V(f_S, z_i) + 2(n-1)\varepsilon_{n-1}^E \geq 0, \tag{17}$$

and therefore

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

Second, by the linearity of expectations,

$$\mathbb{E}_S[V(f_{S^i}, z_i) - V(f_S, z_i)] = \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S], \tag{18}$$

and therefore

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

REMARKS:

1. From *exact positivity* it follows that the leave-one-out error error is greater than or equal to the training error

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

2. CV_{loo} stability implies that the leave-one-out error converges to the training error in probability.

3.5.2 $\mathbf{CV}_{loo}$ stability is necessary and sufficient for consistency of ERM

In the next two theorems we prove first sufficiency and then necessity.

Theorem 3.2 *If the map induced by ERM over a class $\mathcal{H}$ is distribution independent PH stable, and $\mathcal{L}$ is bounded, then ERM over $\mathcal{H}$ is universally consistent.*

PROOF:Given a sample $S = (z_1, \ldots, z_n)$ with n points and a sample $S_{n+1} = (z_1, \ldots, z_{n+1})$ with an additional point then by distribution independent PH stability of ERM, the following holds for all μ:

$$\begin{aligned} \mathbb{E}_{S_{n+1}}[V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})] &\leq \mathbb{E}_{S_{n+1}}[|V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})|] \\ &\leq (\beta_{PH})_{n+1}, \end{aligned} \tag{19}$$

where $(\beta_{PH})_{n+1}$ is associated with S_{n+1} and $|S_{n+1}| = n + 1$.
The following holds for all μ:

$$\mathbb{E}_S I[f_S] - \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] = \mathbb{E}_{S_{n+1}}[V(f_S, z_{n+1}) - V(f_{S_{n+1}}, z_{n+1})]. \tag{20}$$

From Equations (19) and (20), we therefore have

$$\forall \mu \quad \mathbb{E}_S I[f_S] \leq \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] + (\beta_{PH})_{n+1}. \tag{21}$$

Now we will show that

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

Let $\eta_\mu = \inf_{f \in \mathcal{H}} I[f]$ under the distribution μ. Clearly, for all $f \in \mathcal{H}$, we have $I[f] \geq \eta_\mu$ and so $\mathbb{E}_S I[f_S] \geq \eta_\mu$. Therefore, we have (from (21))

$$\forall \mu \quad \eta_\mu \leq \mathbb{E}_S I[f_S] \leq \mathbb{E}_{S_{n+1}} I_{S_{n+1}}[f_{S_{n+1}}] + (\beta_{PH})_{n+1}. \tag{22}$$

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

$$I_{S_{n+1}}[f_{S_{n+1}}] \leq I_{S_{n+1}}[f_{\varepsilon_c, \mu}] + \epsilon^E_{n+1}$$

Taking expectations with respect to S_{n+1} and substituting in eq. (22), we get

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

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

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

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

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

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

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

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

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

This proves distribution independent convergence of $I[f_S]$ to η_μ (consistency), given PH stability. $\square$

Theorem 3.3 *Consistency of ERM (over $\mathcal{H}$) implies PH stability of ERM (over $\mathcal{H}$).*

PROOF:To show PH stability, we need to show that

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

From Lemma 3.5,

$$\forall \mu \; \mathbb{E}_S[|V(f_{S^i}, z_i) - V(f_S, z_i)|] \leq \mathbb{E}_S I[f_{S^i}] - \mathbb{E}_S I_S[f_S] + 4(n-1)\varepsilon^E_{n-1} \quad (23)$$

Given (universal) consistency, Theorem 2.1 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

$$\mathbb{E}_S[I[f_{S^i}] - I_S[f_S]] = \mathbb{E}_S[I[f_{S^i}] - I_S[f_{S^i}]] + \mathbb{E}_S[I_S[f_{S^i}] - I_S[f_S]]. \quad (24)$$

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

$$|I[f_{S^i}] - I_S[f_{S^i}]| \leq \varepsilon_n \quad (25)$$

and therefore

$$\forall \mu \; \mathbb{E}_S[I[f_{S^i}] - I_S[f_{S^i}]] \leq \mathbb{E}_S[|I[f_{S^i}] - I_S[f_{S^i}]|] \leq \varepsilon_n + M\delta_n(\varepsilon_n). \quad (26)$$

From Lemma 3.6, we have

$$\forall \mu \; \mathbb{E}_S[I_S[f_{S^i}] - I_S[f_S]] \leq \frac{M}{n} + \varepsilon^E_{n-1}. \quad (27)$$

Combining Equation (24) with inequalities (26) and (27), we get

$$\forall \mu \; \mathbb{E}_S[I[f_{S^i}] - I_S[f_S]] \leq \varepsilon_n + M\delta_n(\varepsilon_n) + \frac{M}{n} + \varepsilon^E_{n-1}.$$

From inequality (23), we obtain

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

Note that ε^E_n and ε_n may be chosen independently. Also, since we are guaranteed arbitrarily good ε-minimizers, we can choose ε^E_n to be a decreasing sequence such that $\lim_{n\to\infty}(4n-3)\varepsilon^E_n = 0$.
Further, by Lemma 3.7, 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

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

This proves that universal consistency implies PH hypothesis stability. $\square$.

Lemma 3.6 *Under almost ERM,*

$$I_S[f_{S^i}] - I_S[f_S] \leq \frac{M}{n} + \varepsilon^E_{n-1}$$

that is ERM has EE_{loo} stability.

PROOF:

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

Lemma 3.7 *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,

$$\sup_{\mu} \mathbb{P}\left(\sup_{f\in\mathcal{H}} |I[f] - I_S[f]| > \varepsilon\right) \leq \delta_n(\varepsilon)$$

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

For every fixed ε we know that $\lim_{n\to\infty} \delta_n(\frac{1}{k}) = 0$ for every fixed integer k.
Let N_k be such that for all $n \geq N_k$, we have $\delta_n(\frac{1}{k}) < \frac{1}{k}$. Note, that for all $i > j$ $N_i \geq N_j$.
Now choose the following sequence for ε_n. We take $\varepsilon_n = 1$ for all $n < N_2$; $\varepsilon_n = \frac{1}{2}$ for $N_2 \leq n < N_3$ and in general $\varepsilon_n = \frac{1}{k}$ for all $N_k \leq n < N_{k+1}$.
Clearly ε_n is a decreasing sequence converging to 0. Further, for all $N_k \leq n < N_{k+1}$, we have

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

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

REMARKS:

1. Convergence of the empirical error to the expected error follows from either CVEEE$_{loo}$ or LOO stability without assuming ERM (see Theorem 3.1 and Proposition 3.1).

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

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

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

Pseudostability: a remark

It is possible to define a one-sided version of PH stability, called here *pseudoPH stability*:

Definition 3.9 *The learning map L has distribution-independent, pseudo pointwise hypothesis stability if for each n there exists a $\beta_{pPH}^{(n)}$*

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

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

PseudoPH stability is also *necessary and sufficient for universal consistency of ERM*. PseudoPH stability is weaker than PH stability. The proof of its equivalence to consistency of ERM is immediate, following directly from its definition. However, for general (non-ERM) algorithms pseudoPH stability is *not* sufficient in

our approach to ensure convergence in probability of the empirical to the expected risk (eg generalization), when combined with E_{loo} and EE_{loo} (or Eloo_{err}) stability.[22].

Theorems 3.2 and 3.3 can be stated together with the remark on pseudoPH stability to yield

Theorem 3.4 *Either PH or pseudoPH stability of ERM (over $\mathcal{H}$) is necessary and sufficient for consistency of ERM (over $\mathcal{H}$).*

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

A short summary of the argument

The proof just given of necessity and sufficiency of CV_{loo} stability for consistency of ERM has a simple structure, despite the technical details. We summarize it here in the special case of exact minimization of the empirical risk and existence of the minima of the true risk (which we do not assume in the full proof in the previous section) to expose the essence of the proof[23]. In this short summary, we only show that CV_{loo} stability (as well as pseudoPH stability) is necessary and sufficient for consistency of ERM; because ERM on a uGC class is always E_{loo}, EE_{loo} (and Eloo_{err}) stable, the necessity and sufficiency of CVEEE_{loo} (and LOO) stability follows directly.

Theorem 3.5 *Under exact minimization of empirical of the empirical risk and existence of the minima of the true risk, distribution independent (β, δ) CV_{loo} is equivalent to the convergence $I[f_S] \to I[f^*]$ in probability, where $f^* \in \arg\min_{f \in \mathcal{H}} I[f]$.*

PROOF:By the assumption of exact ERM, positivity (instead of almost positivity, see Lemma 3.4) holds, that is

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

Then the following equivalences hold:

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

[22] With pseudoPH stability alone, we are unable to bound the second term in the decomposition of Lemma 3.3.

[23] This short version of the proof could be made shorter by referring to known results on ERM. The argument for almost ERM can be made along similar lines.

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

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

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*,

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

The last step is to show that

$$\lim_{n\to\infty} \mathbb{E}_S I[f_{S^i}] = I[f^*], \tag{28}$$

is equivalent to the statement that $I[f_{S^i}] \to I[f^*]$ *in probability*.
Since $0 \leq I[f] \leq M$ for all f, convergence in probability implies equation (28). The other direction of the statement follows from the fact that

$$I[f_{S^i}] - I[f^*] \geq 0$$

because of the definition of f^*. Therefore,

$$\mathbb{E}_S[I[f_{S^i}] - I[f^*]] \geq 0, \tag{29}$$

which, together with equation (28), implies that *in probability*

$$\lim_{n\to\infty} I[f_{S^i}] = I[f^*]. \tag{30}$$

Finally we note that the convergence in probability of $I[f_{S^i}]$ to $I[f^*]$ is equivalent to consistency. If the draw S_d of the training set has $n+1$ elements, the convergence of $I[f_{S_d^i}]$ to $I[f^*]$ in probability is equivalent to the convergence of $I[f_S]$ to $I[f^*]$ in probability. □

3.5.3 Consistency of ERM implies $\mathbf{E}_{loo}$ and $\mathbf{EE}_{loo}$ (and also $\mathbf{Eloo}_{err}$) stability

ERM is EE_{loo} stable (even when $\mathcal{H}$ is not a uGC class) as shown by Lemma 3.6. Consistency of ERM implies E_{loo} stability as shown by the following Lemma:

Lemma 3.8 *If almost ERM is consistent then*

$$\lim_{n\to\infty} \mathbb{E}_S |I[f_{S^i}] - I[f_S]| = 0$$

The lemma follows immediately from the following condition (because of Jensen inequality)

$$\begin{aligned}\mathbb{E}_S|I[f_S] - I[f_{S^i}]| &\leq \mathbb{E}_S|I[f_S] - I_S[f_S]| + \mathbb{E}_S|I_S[f_S] - I[f_{S^i}]| \\ &\leq \mathbb{E}_S|I[f_S] - I_S[f_S]| + \mathbb{E}_S|I_S[f_{S^i}] - I[f_{S^i}]| + \mathbb{E}_S|I_S[f_{S^i}] - I_S[f_S]|\end{aligned}$$

which implies E_{loo} stability (using consistency to bound the first two terms in the right hand side and EE_{loo} stability to bound the last term). Thus *consistency of ERM implies CVEEE$_{loo}$ stability*.

We now show that consistency of ERM implies Eloo_{err} stability.

Lemma 3.9 *ERM on a uGC class implies*

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

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

PROOF:
By the triangle inequality and inspection

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

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

$$|I_S[f_S] - I[f_S]| \leq \beta_1.$$

Therefore,

$$\mathbb{E}_S\left(I[f_S] - I_S[f_S]\right))^2 \leq M\beta_1 + M^2\delta_1.$$

The following inequality holds for the second term (see proof of Lemma 3.3)

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

Since ERM is on a uGC class (β_2, δ_2) CV_{loo} stability holds, implying

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

Therefore we obtain

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

leading to

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

□

3.5.4 Main result

We are now ready to state the main result of section 3.5.

Theorem 3.6 *Assume that $f_S, f_{S^i} \in \mathcal{H}$ are provided by ERM and $\mathcal{L}$ is bounded. Then distribution independent CVEEE$_{loo}$ stability (as well as LOO stability) is necessary and sufficient for consistency of ERM. Therefore, the following are equivalent*
a) the map induced by ERM is distribution independent CVEEE$_{loo}$ stable
a') the map induced by ERM is distribution independent 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 2.1). We showed that CV$_{loo}$ stability is equivalent to PH stability and that PH stability implies (b). We have also shown in that almost ERM exhibits CVEEE$_{loo}$ stability (and that almost ERM exhibits LOO stability). The theorem follows. □

REMARK:

1. In the classical literature on generalization properties of local classification rules ([7]) 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 3.6. 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*[24]*, or c) $\mathcal{H}$ has a finite number of hypotheses.*

3.5.5 Distribution-dependent stability and consistency

Our main result is given in terms of distribution-free stability and distribution-free consistency. In this distribution-free framework consistency of ERM is equivalent to $\mathcal{L}$ being uGC. Inspection of the proof suggests that it may be possible to reformulate our theorem (see also 3.5) in a distribution dependent way: *for ERM, CV$_{loo}$ stability with respect to a specific distribution is necessary and*

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

sufficient for consistency with respect to the same distribution.[25] Of course, in this distribution-dependent framework $\mathcal{L}$ may not be uGC.

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 theorem 3.2. In general these rates will be very slow, also in the case of ERM.
In this section we outline how CV_{loo} stability can be used to control the expectation and *error stability* can be used to control the variance. The two notions of stability together will be called *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 [14].

Definition 4.1 *The learning map L is strongly stable if*
a. it has $(\beta_{loo}, \delta_{loo})$ CV_{loo} stability
b. it has error stability with a fast rate, eg for each n there exists a $\beta^{(n)}_{error}$ and a $\delta^{(n)}_{error}$ such that

$$\forall i \in \{1, ..., n\} \quad \forall \mu \quad \mathbb{P}_S \left\{ |I[f_S] - I[f_{S^i}]| \leq \beta_{error} \right\} \geq 1 - \delta_{error},$$

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

Our definition of strong stability depends on CV_{loo} stability and on the difference in the expected values of the losses ($I[f_S] - I[f_{S^i}]$).
The following theorem is similar to theorem 6.17 in [14].

Theorem 4.1 *If the learning map is strongly stable then, for any $\varepsilon > 0$,*

$$\mathbb{P}_S \left\{ |I_S[f_S] - I[f_S]| \geq \varepsilon + \beta_{loo} + M\delta_{loo} + \beta_{error} + M\delta_{error} \right\} \leq$$

$$2 \left(\exp \left(\frac{-\varepsilon^2 n}{8(2n\beta_{error} + M)^2} \right) + \frac{n(n+1)2M\delta_{error}}{2n\beta_{error} + M} \right)$$

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

[25] It should be possible to reformulate the definitions of distribution-dependent consistency and CV_{loo} stability appropriately, to avoid the case of trivial consistency (following [25]).

probability of the lack of convergence decreases exponentially as n increases. The proof of the above theorem is in Appendix 6.2 and is based on a version of McDiarmid's inequality.

For the empirical risk to converge to the expected risk in the above bound β_{error} must decrease strictly faster than $O(n^{-1/2})$. 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 [26] with ε^{-1}. Even for hypothesis spaces with VC dimension d the rate of convergence of error stability is not fast enough, with probability $1 - e^{-t}$

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

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 [27];
- 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 Proposition 6 in [5]);
- 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 [26]);
- ERM on VC-subgraph classes that are convex with the square loss.

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 $\mathcal{N}(\mathcal{H}, \varepsilon)$ that are polynomial in ε^{-1} (VC classes fall into this category) or exponential in in ε^{-p} with $p < 1$ (the 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 $O\left(\sqrt{\frac{d}{n}}\right)$ using standard VC or covering number bounds, here d is the metric entropy or shattering dimension of the

[26] 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.).

[27] This case was considered in [14] 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 $(0, e^{-Kn})$ CV stable with respect to the distribution on Z generated by Δ and f_0.*
2. *The VC dimension of $\mathcal{H}$ is finite.*

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 3.1 which gives suboptimal rates that are much slower than $O\left(\sqrt{\frac{d}{n}}\right)$. The following cases fall into the gap between general uGC classes that have slow rates of convergence[28] and those classes that have a fast rate of convergence[29]:

- 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 $\Omega((\frac{1}{n})^b)$: 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 – CVEEE$_{loo}$ 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 a key but independent role in the development of learning theory and learning algorithms respectively, are in fact closely related: well-posedness (defined in terms of CVEEE$_{loo}$ 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 scientific "theories" can be extracted from empirical data (see [25]). 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 CV$_{loo}$ stability fully subsumes the classical *necessary and sufficient* conditions on $\mathcal{H}$ for consistency of ERM.

[28] Obtained using either standard covering number bounds or theorem 3.1.
[29] Obtained using either standard covering number bounds or strong stability.

CVEEE$_{loo}$ (or LOO) stability and its properties may suggest how to develop learning theory beyond the ERM approach. It is a simple observation that CVEEE$_{loo}$ (or 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 CV$_{loo}$ 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 CV$_{loo}$ 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 [10]). 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 CVEEE$_{loo}$ and 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 [3] imply that regularization and SVMs are CVEEE$_{loo}$ (and also LOO) stable; a version of bagging with the number k of regressors increasing with n (with $\frac{k}{n} \to 0$) is CV$_{loo}$ stable and has hypothesis stability (because of [7]) and EE$_{loo}$ stability and is thus CVEEE$_{loo}$ (and LOO) stable; similarly k-NN with $k \to \infty$ and $\frac{k}{n} \to 0$ and kernel rules with the width $h_n \to 0$ and $h_n n \to \infty$ are CVEEE$_{loo}$ (and LOO) stable. Thus all these algorithms satisfy Theorem 3.1 and Proposition and 3.1 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 $O\left(\frac{d}{n}\right)$, 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 $O(n^{-1/2})$. It may be possible to improve these rates using inequalities of the type in [19].

(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 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 [22]). A basic problem in the martingale inequalities is how variance is controlled. Given a random variable $Z = f(X_1, ..., X_n)$ the variance of this random variable is controlled by a term of the form of

$$\text{Var}(Z) \leq \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2\right],$$

where $Z^{(i)} = f(X_1, ..., X_i', ...X_n)$. 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

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

However, for this class of functions we know that

$$\text{Var}(I_S[f_S] - I[f_S]) = \Theta\left(\sqrt{\frac{dlnn}{n}}\right).$$

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 CV$_{loo}$ stability (instead of via consistency).

4. Our results say that for ERM, distribution-independent CV$_{loo}$ 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 CV$_{loo}$ 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 14 . 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 proof is sketched in the Appendix 6.3. 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 CVEEE$_{loo}$ stability is open. Either CVEEE$_{loo}$ or LOO stability answer all the requirements we need: each one is sufficient for generalization in the general setting and subsumes the classical theory for ERM, since it is equivalent to consistency of ERM. It is quite possible, however, that CVEEE$_{loo}$ stability may be equivalent to other, even "simpler" conditions. In particular, we conjecture that CV$_{loo}$ and EE$_{loo}$ stability are sufficient for generalization for general algorithms (without E$_{loo}$ stability). Alternatively, it may be possible to combine CV$_{loo}$ 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, though we conjecture that it does.

The diagram of Figure 1 shows relations between the various properties discussed in this paper. The diagram is by itself a map to a few open questions.

Acknowledgments We thank for many critically helpful comments Ding Zhou, Steve Smale, Richard Dudley, Nicolo' Cesa-Bianchi, Gabor Lugosi, Shahar Mendelson, Andre' Elisseeff and especially Dimitry Panchenko. We are also grateful to Massi (and Theos) for a key, unconscious, ultimately wrong but very useful contribution.

6 Appendices

6.1 Some conditions that are necessary and sufficient for the uGC property

Alon et al. proved a necessary and sufficient conditions for universal (wrt all distributions) and uniform (over all functions in the class) convergence of $|I_S[f] - I[f]|$, in terms of the finiteness for all $\gamma \geq 0$ of a combinatorial quantity called V_γ dimension of $\mathcal{F}$ (which is the set $V(x), f(x), f \in \mathcal{H}$), under some assumptions (such as convexity, continuity, Lipschitz) on V.
Alon's result is based on a necessary and sufficient (distribution independent) condition proved by Vapnik and Dudley et al. which uses the (distribution-independent) metric entropy of $\mathcal{F}$ defined as $H_n(\varepsilon, \mathcal{F}) = \sup_{x_n \in X^n} \log \mathcal{N}(\varepsilon, \mathcal{F}, x_n)$, where $\mathcal{N}(\varepsilon, \mathcal{F}, x_n)$ is the ε-covering of $\mathcal{F}$ with respect to $l^\infty_{x_n}$ ($l^\infty_{x_n}$ is the l^∞ distance on the points x_n).

Theorem (Dudley et al.) *$\mathcal{F}$ is a strong uniform Glivenko-Cantelli class iff* $\lim_{n\to\infty} \frac{H_n(\varepsilon,\mathcal{F})}{n} = 0$ *for all* $\varepsilon > 0$.

Notice (see [16]) that the metric entropy may be defined (and used in the above theorem) with respect to empirical norms other than $l^\infty_{x_n}$.
Thus the following equivalences hold:

$$\begin{aligned} \mathcal{H} \text{ is uGC} \quad &\Leftrightarrow \quad \lim_{n\to\infty} \frac{H_n}{n} = 0, \\ &\Leftrightarrow \quad \text{finite } V_\gamma \ \ \forall \ \ \gamma \geq 0. \end{aligned}$$

Finite VC dimension is the special case of the latter condition when the functions in $\mathcal{H}$ are binary. It is well known that necessary and sufficient condition for uniform convergence in the case of $0-1$ functions is finiteness of the VC dimension (Vapnik). Notice that the V_γ dimension exactly reduces to the VC dimension for $\gamma = 0$.

6.2 Strong stability implies good convergence rates

Proof of Theorem 4.1

This theorem is a variation, using our definition of L-stability, of Theorem 6.17 in [14]. We first give two definitions from [14].

Definition 6.1 *Change-one cross-validation (CV_{co}) stability is defined as*

$$\mathbb{P}_{S,z} \{|V(f_S, z) - V(f_{S^{i,z}}, z)| \leq \beta_{CV}\} \geq 1 - \delta_{CV},$$

where $\beta_{CV} = O(n^{-\alpha})$ *and* $\delta_{CV} = O(n^{-\tau})$ *for* $\tau, \alpha > 0$.

Definition 6.2 *Error stability is defined as*

$$\mathbb{P}_{S,z}\left\{|I[f_S] - I[f_{S^{i,z}}]| \leq \beta'_{Error}\right\} \geq 1 - \delta'_{Error},$$

where $\beta'_{Error} = O(n^{-\alpha})$ *where* $\alpha > 1/2$ *and* $\delta'_{Error} = e^{-\Omega(n)}$.

Notice that both these definitions perturb the data by replacement. In our definition of CV_{loo} stability we use the leave-one-out procedure to perturb the training set.

Kutin and Niyogi 2002 – Theorem 6.17: *If the learning map has* CV_{co} *stability* $(\beta_{CV}, \delta_{CV})$*, and error stability* $(\beta'_{Error}, \delta'_{Error})$ *(where error-stability is defined with respect to a point being replaced)), then, for any* $\varepsilon > 0$*,*

$$\mathbb{P}_S\{|I_S[f_S] - I[f_S]| \geq \varepsilon + \beta_{CV} + M\delta_{CV}\} \leq$$

$$2\left(\exp\left(\frac{-\varepsilon^2 n}{8(n\beta'_{Error} + M)^2}\right) + \frac{n(n+1)M\delta'_{Error}}{n\beta'_{Error} + M}\right)$$

where M *is a bound on the loss.*

The proof of Theorem 6.17 requires two separate steps. It requires first to bound the mean and, second, to bound the variance of the generalization error. The variance is bounded by using error stability; the mean is bounded by using CV stability. Then McDiarmid inequality is used.

Next we will first show (in (a)) that our definition of error stability, with the leave-one-out perturbation, and Kutin's and Niyogi's definition differ by at most a factor 2 (this relates β_Error and δ_Error to β'_Error and δ'_Error). Then (in (b)) we will directly bound the mean of the generalization error.

(a) If we have with probability $1 - \delta_{Error}$

$$|I[f_S] - I[f_{S^i}]| \leq \beta_{Error},$$

then with probability $1 - 2\delta_{Error}$

$$|I[f_S] - I[f_{S^{i,z}}]| \leq 2\beta'_{Error}.$$

This holds because of the following inequalities

$$\begin{aligned} |I[f_S] - I[f_{S^{i,z}}]| &= |I[f_S] - I[f_{S^i}] + I[f_{S^i}] - I[f_{S^{i,z}}]| \\ &\leq |I[f_S] - I[f_{S^i}]| + |I[f_{S^i}] - I[f_{S^{i,z}}]|. \end{aligned}$$

This allows us to replace our definition of error stability with that used in [14].

(b) In the proof of Theorem 6.17 in [14] the terms β_{CV} and δ_{CV} are used to bound the following quantity

$$\mathbb{E}_{S,z}[V(f_S, z) - V(f_{S^{i,z}}, z)] \leq \beta_{CV} + M\delta_{CV},$$

where M is the upper-bound on the error at a point. We need a similar upper bound using CV_{loo} stability instead of CV_{co} stability. We rewrite the left-hand side of the above inequality as

$$\mathbb{E}_{S,z}[V(f_S, z) - V(f_{S^{i,z}}, z)] = \mathbb{E}_S[I[f_S] - I_S[f_S]], \tag{31}$$

which is the expectation with respect to S of the generalization error of f_S. The following inequality holds

$$\begin{aligned} \mathbb{E}_S[I[f_S] - I_S[f_S]] &= \mathbb{E}_S[I[f_{S^i}] - I[f_{S^i}] + I[f_S] - I_S[f_S]] \\ |\mathbb{E}_S[I[f_S] - I_S[f_S]]| &\leq |\mathbb{E}_S[I[f_{S^i}] - I_S[f_S]]| + |E_S[I[f_S] - I[f_{S^i}]]| \\ &\leq |\mathbb{E}_S[I[f_{S^i}] - I_S[f_S]]| + \beta_{Error} + M\delta_{Error}, \end{aligned}$$

the last step following from error stability.
From CV_{loo} stability the following inequality holds

$$|\mathbb{E}_S[I[f_{S^i}] - I_S[f_S]]| \leq \mathbb{E}_S|V(f_{S^i}, z_i) - V(f_S, z_i)| \leq \beta_{LOO} + M\delta_{LOO}.$$

We have now done steps (a) and (b). We restate the KN theorem in terms of our definitions of CV_{loo} stability and error stability:

L-stability implies that the bound in Equation 4.1 holds for all $\varepsilon, \delta > 0$. Then for any f_S obtained by ERM from a dataset S and for any $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}_S\{|I_S[f_S] - I[f_S]| > \varepsilon\} = 0. \square$$

6.3 Compactness and stability

Convexity
In the proofs used here, the convexity of $\mathcal{H}$ and V plays a key role. Given any two functions f_1 and f_2 in a convex $\mathcal{H}$, their *average* $f_A(x) \equiv \frac{1}{2}(f_1(x) + f_2(x))$ is also in $\mathcal{H}$. Furthermore, if V is convex, for any $z \in Z$,

$$\ell(f_A(z)) \leq \frac{1}{2}(\ell(f_1(x)) + \ell(f_2(x))).$$

When V is the square loss, and M is a bound on $(f(x) - y)$, we have the following lemma.

Lemma 6.1
- *If $|\ell(f_1(z)) - \ell(f_2(z))| \geq d_\ell$, then $|f_1(x) - f_2(x)| \geq \sqrt{M} - \sqrt{M - d_\ell}$.*
- *If $|f_1(x) - f_2(x)| = d_f$, then $\ell(f_A(z)) + \frac{d_f^2}{4} = \frac{1}{2}(\ell(f_1(z)) + \ell(f_2(z)))$.*
- *If $|\ell(f_1(z)) - \ell(f_2(z))| \geq d_\ell$, then $\ell(f_A(z)) + \frac{(\sqrt{M} - \sqrt{M - d_\ell})^2}{4} \leq \frac{1}{2}(\ell(f_1(z)) + \ell(f_2(z)))$.*

The first part is obvious from inspection of the loss function. The second part is a simple algebraic exercise. The third part is simply a combination of the first and second parts.
Proof

- **Remark** Notice that under our hypotheses the minimizer of the expected risk exists and is unique when $\mathcal{H}$ is convex and appropriate regularity assumptions on the measure hold [5] (see also [17] p. 5). The minimizer of the empirical risk is not unique in general, but is unique with respect to the empirical measure over the training set.

We first sketch a proof of sufficiency, that is *compactness and convexity of $\mathcal{H}$ implies (β, δ)-hypothesis stability*:

PROOF:

Since $\mathcal{H}$ is compact and convex given appropriate regularity assumptions on the measure there is a unique function f^* that minimizes the expected risk [5]. We first compute the probability that f_S and f^* are close in expected risk:

$$\begin{aligned} |I[f_S] - I[f^*]| &\leq \varepsilon \\ |I[f_{S^i}] - I[f^*]| &\leq \varepsilon. \end{aligned}$$

Lemma 6.2 *With probability* $1 - \delta$

$$|I[f_S] - I[f^*]| \leq \varepsilon,$$

where $\varepsilon = \left(\frac{M}{n}\right)^{1/2-\tau}$ *and* $\delta = \mathcal{N}(\varepsilon, \mathcal{L})\, e^{-O(n^{2\tau})}$, $0 < \tau < 1/2$.

PROOF:
The function f^* has error rate $I[f^*] = \eta$. We define the set of functions f_g as follows

$$f_g = \{f \in \mathcal{F} \text{ for which } |I[f] - I[f^*]| \leq \varepsilon\}.$$

By Chernoff's inequality for a function f_g

$$\mathbb{P}_S \left(I_S[f_g] \geq \eta + \varepsilon/2\right) \leq e^{-\varepsilon^2 n/8M^2}.$$

We define the set of functions f_b as

$$f_b = \{f \in \mathcal{F} \text{ for which } |I[f] - I[f^*]| \geq \varepsilon\}.$$

Also by Chernoff's inequality for a function f_b,

$$\mathbb{P}_S \left(I_S[f_b] \leq \eta + \varepsilon/2\right) \leq e^{-\varepsilon^2 n/8M^2}.$$

Thus for f_S to not be ε close to f^*, at least one of the following events must happen:

$$\begin{aligned} \min_{f \in f_b} I_S[f] &\leq \eta + \varepsilon \\ \max_{f \in f_g} I_S[f] &\geq \eta + \varepsilon \end{aligned}$$

The following uniform convergence bound then holds:

$$\mathbb{P}(|I[f_S] - I[f^*]| \geq \varepsilon) \leq \mathcal{N}(\varepsilon, \mathcal{L})\, e^{-O(\epsilon^2 n)},$$

where $\mathcal{N}(\varepsilon, \mathcal{L})$ is the covering number. Setting $\varepsilon = \left(\frac{M}{n}\right)^{1/2-\tau}$ where $0 < \tau < 1/2$ gives the result stated. □
The above lemma implies that with probability $1 - \delta$ the following inequality holds:

$$|I[f_S] - I[f^*]| \leq \left(\frac{M}{n}\right)^{1/2-\tau},$$

with

$$\delta = \mathcal{N}(\varepsilon, \mathcal{L})\, e^{-O(n^{2\tau})}.$$

We now show (by way of the contrapositive) that under the assumptions of compactness and convexity, a function that is close to the optimal function in generalization error is also close in the sup norm:

$$\sup_x |f_S(x) - f^*(x)| \geq 2\epsilon \;\Rightarrow\; |I[f_S] - I[f^*]| \geq \frac{c\epsilon^2}{2},$$

for some constant c.
We will use Arzelà's theorem [13].
Theorem (Arzelà and Ascoli): A necessary and sufficient condition for a family of continuous functions $f \in \mathcal{F}$ defined on a compact set X to be relatively compact in $C(X)$ is that $\mathcal{F}$ be uniformly bounded and equicontinuous.
Given that

$$\sup_x |f_S(x) - f^*(x)| \geq 2\epsilon,$$

we define the following set:

$$X^\epsilon = \{x : |f_S(x) - f^*(x)| \geq \epsilon\}.$$

By equicontinuity we know that $\mu(X^\epsilon) \geq c$, where the equicontinuity allows us to lower bound the measure of the domain over which the functions differ in the supremum. This constant c will exist for sufficiently regular measures; when it exists, its value will depend on the measure.
If we define a function $f_A = \frac{1}{2}(f_S + f^*)$, the following holds:

$$\begin{aligned} \forall x \in X : I_x[f'] &\leq \frac{1}{2}(I_x[f] + I_x[f^*]) \\ \forall x \in X^\epsilon : I_x[f'] + \frac{\epsilon^2}{4} &\leq \frac{1}{2}(I_x[f] + I_x[f^*]) \end{aligned}$$

where $I_x[f] \equiv \int_y V(f(x), y) d\mu(y)$, the expected risk of f at the point x.
Combining the above inequalities over all points $x \in X$, we see that

$$I[f_A] + \frac{c\epsilon^2}{4} \leq \frac{1}{2}(I[f_S] + I[f^*]).$$

But f^* is the minimizer of $I[f]$, so we also have that

$$I[f^*] + \frac{c\epsilon^2}{4} \leq \frac{1}{2}(I[f_S] + I[f^*]).$$

From the above we have that

$$\begin{aligned} & I[f^*] + \frac{c\epsilon^2}{4} \leq \frac{1}{2}(I[f_S] + I[f^*]) \\ \rightarrow \quad & I[f^*] + \frac{c\epsilon^2}{2} \leq I[f_S] \\ \rightarrow \quad & |I[f] - I[f^*]| \geq \frac{c\epsilon^2}{2}. \end{aligned}$$

From this we can conclude that if the difference in the loss of a function from the optimal function is bounded then so is the difference in the sup norm. Since we know that with probability $1 - \delta$,

$$|I[f_S] - I[f^*]| \leq \beta,$$

where $\beta = \left(\frac{M}{n}\right)^{1/2-\tau}$, then also with probability $1 - \delta$,

$$\sup_x |f_S(x) - f^*(x)| \leq 2\sqrt{\frac{2\beta}{c}},$$

and, applying the same argument to f_{S^i} (we use the uGC constants associated with a training set of size $n - 1$ for both S and S^i), with probability $1 - 2\delta$,

$$\sup_x |f_S(x) - f_{S^i}(x)| \leq 4\sqrt{\frac{2\beta}{c}}.$$

Since $f \in \mathcal{H}$ is bounded and so is Y, the square loss has a Lipschitz property then

$$\sup_x |f_S(x) - f_{S^i}(x)| \leq 4\sqrt{\frac{2\beta}{c}}$$

implies

$$\sup_z |V(f_S, z) - V(f_{S^i}, z)| \leq \frac{K\sqrt{\beta}}{c},$$

which is (β, δ)-hypothesis stability.□

- **Remark.** The proof presented here obviously relies crucially on convexity. It is possible to replace convexity with other assumptions but it seems that compactness alone will not suffice. For example, consider the simple case where X consists of a single point x, $\mathcal{H}$ consists of the two functions $f(x) = 1$ and $f(x) = -1$, and $y(x)$ takes on the values 1 and -1, each with probability $\frac{1}{2}$. The minimizer of the true risk is non-unique here — both functions have identical true risk. However, a simple combinatorial

argument shows f_S and f_{S^i} will only be different functions with probability $O(\sqrt{n})$, so we still have (β, δ)-stability. A possible replacement for the convexity assumption is to relax our definition of (β, δ)-hypothesis stability so that it may not hold on sets of measure zero. Another possibility seems to be the assumption that the target function is continuous.

We now prove necessity, that is *(β, δ)-hypothesis stability implies compactness of $\mathcal{H}$.*

PROOF:

Since (β, δ)-hypothesis stability implies CV_{loo} stability, $\mathcal{H}$ is a uGC class because of Theorem 3.2. Suppose that $(\mathcal{H}, L_\infty)$ is not compact. Then, by Arzelà's theorem, $\mathcal{H}$ is not equicontinuous. In particular, there exists an $\epsilon_\mathcal{H} > 0$, a sequence $\delta_i \to 0$ and a sequence of functions $f_i \subset \mathcal{H}$ satisfying

$$\forall i, \exists x, x' \in X \; s.t. \; |x - x'| \leq \delta_i, |f_i(x) - f_i(x')| \geq \epsilon_\mathcal{H} \tag{32}$$

We first note that each individual f_i, being continuous over a compact domain, is uniformly continuous, and using this fact, it is easy to show that given f_i, there exists a k such that $\forall i' > k$, $|f_i - f_{i'}| \geq \frac{\epsilon_\mathcal{H}}{2}$.
Now, consider $(\mathcal{H}, L_2)$, the set of functions $\mathcal{H}$ equipped with the L_2 norm. This set is totally bounded, because a uGC class is totally bounded in L_2 (Shahar Mendelson, personal communication)[30]. Therefore, any infinite sequence in $(\mathcal{H}, L_2)$ contains a fundamental (Cauchy) subsequence. Using this property, define the sequence g_i to be a subsequence of the f_i referred to in Equation 32 which is Cauchy in $(\mathcal{H}, L_2)$.
The sequence g_i converges in measure in the L_2 metric to some (possibly discontinuous) function f^*, which in turn implies that g_i contains a subsequence h_i which converges almost everywhere to f^* ([13] p. 292, Problem 9).
Next, construct a distribution D that is uniform over the input space X, with "target" $y(x) = f^*(x)$. Consider any finite sample S from D. For i sufficiently large, all $h_{i'}$ for $i' \geq i$ will be ϵ-empirical risk minimizers, and therefore any of them may be returned as either f_S or f_{S^j}.
Consider one such h_i. Let $X^\epsilon_{h_i}$ be those x satisfying $|f^*(x) - h_i(x)| \geq \frac{\epsilon_\mathcal{H}}{16}$. Suppose $\mu(X^\epsilon_{h_i}) \geq 0$. Because the h_i are converging almost everywhere to f^*, there will exist some i' and some x in this set for which $|f^*(x) - h'_i(x)| \leq \frac{\epsilon_\mathcal{H}}{32}$; at this point the losses of h_i and h'_i will differ by at least $\frac{\epsilon^2_\mathcal{H}}{1024}$, showing we do not have (β, δ)-hypothesis stability with $\beta \leq \frac{\epsilon^2_\mathcal{H}}{1024}$ and $\delta \to 0$. The only other possibility is that $\mu(X^\epsilon_{h_i}) = 0$, and $|f^*(x) - h_i(x)| \leq \frac{\epsilon_\mathcal{H}}{16}$ almost everywhere. In this case, because each h_i is continuous over a compact domain and therefore uniformly continuous, find δ such that

$$|x - x'| \leq \delta \to |h_i(x) - h_i(x')| \leq \frac{\epsilon_\mathcal{H}}{8},$$

[30] $(\mathcal{H}, L_2)$ is not (in general) compact, because it is (in general) incomplete.

which in turn implies that

$$|x - x'| \leq \delta, \;\; x, x' \; \in X - X^{\epsilon}_{h_i} \; \rightarrow |f^*(x) - f^*(x')| \leq \frac{\epsilon_{\mathcal{H}}}{4}.$$

Choose an h'_i such that

$$\exists x, x' \; s.t. \; |x - x'| \leq \frac{\delta}{2}, |h'_i(x) - h'_i(x)| \geq \epsilon_{\mathcal{H}}.$$

Because h'_i is uniformly continuous, there will exist sets X^+ and X^-, both of positive measure, such that for any $x \in X^+$ and any $x' \in X^-$,

$$|x - x'| \leq \delta, |h'_i(x) - h'_i(x')| \geq \frac{\epsilon_{\mathcal{H}}}{2}.$$

Since these sets both have positive measure, there must exist a pair x and x' in $X - X^{\epsilon}_{h_i}$ where all of the following hold:

- $|f^*(x) - h_i(x)| \leq \frac{\epsilon_{\mathcal{H}}}{16}$
- $|f^*(x') - h_i(x')| \leq \frac{\epsilon_{\mathcal{H}}}{16}$
- $|f^*(x) - f^*(x')| \leq \frac{\epsilon_{\mathcal{H}}}{4}$
- $|h'_i(x) - h'_i(x')| \geq \frac{\epsilon_{\mathcal{H}}}{2}$

Then at at least one of the two points x and x' referenced in the above (say x), we have $|f^*(x) - h_i(x)| \leq \frac{\epsilon_{\mathcal{H}}}{16}$ but $|f^*(x) - h'_i(x)| \geq \frac{\epsilon_{\mathcal{H}}}{8}$, again showing that we do not have (β, δ)-hypothesis stability.
Assuming the combination of (β, δ)-hypothesis stability, uGC property, and non-compactness led to a contradiction; we therefore conclude that (β, δ)-hypothesis stability combined with uGC implies compactness. □

- **Remark.** Although (β, δ)-hypothesis stability is stated as a condition on the relationship between f_S and f_{S^i}, under the conditions discussed in this paper (namely convexity), we find that the key condition is instead uniqueness of the ERM function f_S for large datasets. In particular, it is possible to show that if $\mathcal{H}$ is not compact, then we can construct a situation where there will be functions with risk arbitrarily close to f^* that are far from f^* in the *sup* norm over a set with positive measure. This leads to a situation where f_S and f_{S^i} are both sets of functions containing elements that differ in the *sup* norm; since any of these functions can be picked as either f_S or f_{S^i}, we will not have (β, δ)-hypothesis stability. In contrast, if we additionally assume compactness and convexity, this cannot happen — that a function cannot simultaneously be far from f^* in the *sup* norm and arbitrarily close to f^* in risk.

The Realizable Setting

We say that the setting is *realizable* when there is some $f_0 \in \mathcal{H}$ which is consistent with the examples. In this case there is a simpler proof of the sufficient part of the conjecture.

Theorem *When the setting is realizable, compactness of $\mathcal{H}$ implies (β, δ)-hypothesis stability).*

Proof

Consider the family of functions $\mathcal{L}$ consisting of $\ell(z) = (f(x) - y)^2$. If $\mathcal{H}$ is bounded and compact then $\mathcal{L} \in C(Z)$ is a compact set of continuous bounded functions with the norm $\|\ell\| = \sup_{z \in Z} |\ell(z)|$. Let $S = (x, y)_1^n$ with $\mathbf{z} = \{z_1, ..., z_n\}$ a set of points in Z. Take a covering of Z defined in terms of a set of n disks $D_i(\mathbf{z}, \nu(\mathbf{z}))$, where $\nu(\mathbf{z})$ is the smallest radius sufficient for the union of the n disks to cover Z. Taking a hint from a proof by Pontil (manuscript in preparation) we write the norm in $C(Z)$ in terms of the empirical norm w. r. t. the set $\mathbf{z}$ for $\ell \in \mathcal{L}$:

$$\sup_{z \in Z} |\ell(z)| = \max_{i=1,...,n} \{ \sup_{z \in D_i(\mathbf{z}, \nu(\mathbf{z}))} |\ell(z)| \} \tag{33}$$

The above equality implies

$$\sup_{z \in Z} |\ell(f_S(z)) - \ell(f_{S^j})(z)| = \max_{i=1,...,n} \{ \sup_{z \in D_i(\mathbf{z}, \nu(\mathbf{z}))} |\ell(f_S(z)) - \ell(f_{S^j}(z))| \} \tag{34}$$

which can be rewritten as

$$\sup_{z \in Z} |\ell(f_S(z)) - \ell(f_{S^j}(z))| = \max_{i=1,...,n} \{ \sup_{z \in D_i(\mathbf{z}, \nu(\mathbf{z}))} |\ell(f_S(z)) - \ell(f_S(z_i)) + \\ + \ell(f_S(z_i)) - (\ell(f_{S^j}(z)) - \ell(f_{S^j}(z_i))) - \ell(f_{S^j}(z_i))| \},$$

leading to the following inequality

$$\sup_{z \in Z} |\ell(f_S(z)) - \ell(f_{S^j}(z))| \leq \max_{i=1,...,n} |\ell(f_S(z_i)) - \ell(f_{S^j}(z_i))| + O(\epsilon(\nu(S))), \tag{35}$$

in which we bound separately the variation of $\ell(f_S)$ and $\ell(f_{S^j})$ within each D_i using the equicontinuity of $\mathcal{L}$. We assume that under regularity conditions (see argument below) on the measure, the radius of the disks ν goes to 0 as $n \to \infty$. Under the realizable setting assumptions all but one of the n terms in the $\max$ operation on the right-hand side of the inequality above disappear (since $\ell(f_S(x)) = 0$ when $x \in S$). (β, δ)-LOO stability (which follows from our main theorem in the paper since $\mathcal{L}$ is uGC because it is compact) can now be used to bound $|\ell(f_S(z_i)) - \ell(f_{S^j}(z_i))|$ for the only non-zero term (the one corresponding to $j = i$, eg the disk centered on the point which is in S but not S^j). Thus

$$\mathbb{P}_S \left\{ \sup_{z \in Z} |\ell(f_S(z)) - \ell(f_{S^j}(z))| \leq \beta_{LOO} + O(\epsilon(\nu(S))) \right\} \geq 1 - \delta \qquad (36)$$

with β_{LOO}, $\epsilon(\nu(S))$ and δ all going to 0 as $n \to \infty$. This is (β, δ)-hypothesis stability. Since $\mathcal{L}$ is compact, it is also uGC and thus fast error stability holds.

To show $\nu(S) \to 0$, consider an arbitrary radius for the disks $a > 0$. Call $\delta_i = P(D_i, a)$ the probability of a point in Z to be covered by the disk D_i when the points x_i for $i = 1, \ldots, n$ are sampled (in general the δ_i are different from each other because the measure μ is in general not uniform in X). Consider $\delta = \min_i \delta_i$. Notice that $\delta = O(a^k)$, where k is the dimensionality of Z. Thus an arbitrary point $x \in X$ will be covered by at least one disk with high probability, $p = 1 - ((1 - \delta)^n)$. It is easy to see that for $n \to \infty$, the probability of a cover $p \to 1$, while simultaneously the disk shrinks ($a \to 0$) slowly enough – with a rate $n^{-\alpha}$ where $\alpha > 0$, $\alpha k < 1$ (because $\lim_{n \to \infty}(1 - n^{\alpha k})^n = 0$ if $\alpha k < 1$).

- **Remark.** Using the assumption of convexity, this proof can be extended to the non-realizable (general) case. Suppose that the difference at the left out training point z_i is less than or equal[31] to β:

  $$\ell(f_{S^i}(z_i)) - \ell(f_S(z_i)) \leq \beta.$$

  Then we will also be able to bound the difference in loss at all the other training points. Suppose we have a training point z_j for which

  $$|\ell(f_{S^i}(z_j)) - \ell(f_S(z_j))| \geq b.$$

  Then, by our convexity lemma, at z_j, the loss of the average function $f_A = \frac{1}{2}(f_S + f_{S^i})$ satisfies

  $$\ell(f_A(z_j)) + c(b) \leq \frac{1}{2}(\ell(f_S(z_j)) + \ell(f_{S^i}(z_j))),$$

  where $c(b) = \frac{(\sqrt{M} - \sqrt{M-b})^2}{4}$. Summing over the set S^i, we have

  $$\begin{aligned} & I_{S^i}[f_A] + \frac{c(b)}{n-1} \leq \frac{1}{2}(I_{S^i}[f_S] + I_{S^i}[f_{S^i}]) \\ \implies \quad & I_{S^i}[f_{S^i}] + \frac{2c(b)}{n-1} \leq I_{S^i}[f_S], \end{aligned}$$

  using the fact that f_{S^i} is the minimizer of $I_{S^i}[\cdot]$. But using the fact that f_S

[31] We don't need an absolute value here, because the loss of f_S at the left out point is always less than the loss of f_{S^i} at that point.

minimizes $I_S[\cdot]$,

$$\begin{aligned}
& I_S[f_S] \leq I_S[f_{S^i}] \\
\implies & \frac{n-1}{n} I_{S^i}[f_S] + \frac{1}{n} V(f_S(x_i), y_i) \leq \frac{n-1}{n} I_{S^i}[f_{S^i}] + \frac{1}{n} V(f_{S^i}(x_i), y_i) \\
\implies & I_{S^i}[f_S] + \frac{1}{n-1} V(f_S(x_i), y_i) \leq I_{S^i}[f_{S^i}] + \frac{1}{n-1} V(f_{S^i}(x_i), y_i) \\
\implies & I_{S^i}[f_S] \leq I_{S^i}[f_{S^i}] + \frac{\beta}{n-1}.
\end{aligned}$$

Combining the above derivations, we obtain

$$\begin{aligned}
& I_{S^i}[f_{S^i}] + \frac{2c(b)}{n-1} \leq I_{S^i}[f_{S^i}] + \frac{\beta}{n-1} \\
\implies & 2c(b) \leq \beta.
\end{aligned}$$

If β is a bound on the difference at the left out point, then at every other training point, the difference in loss is less than b, where

$$\frac{(\sqrt{M} - \sqrt{M-b})^2}{2} \leq \beta.$$

As $\beta \to 0$, $b \to 0$ as well, finishing the proof. Note that we do not provide a rate.

- **Remark.** The assumption of realizability (this setup is also called *proper learning*) is a strong assumption. Throughout this work, we have essentially used compactness (equicontinuity) to enforce a condition that our space will not contain functions that approach the minimizer in risk without also approaching it in the sup norm. If we assume realizability, then we can often get this property without needing compactness. For instance, consider the simple class of functions (from $[0,1]$ to $[0,1]$) where $f_i(x) = \min(ix, 1)$. This is a uGC but non-compact class of functions. If we choose the (non-realizable) target function to be $f^*(x) = 1$, we find that we don't get (β, δ)-hypothesis stability. However, if we require that the target actually be some f_i, we will recover (β, δ)-hypothesis stability; essentially, the slope of the target f_i acts as an equicontinuity constant — those f_j that slope too much more rapidly will not have empirical risk equivalent to f_i for sufficiently large samples. While it is not quite true in general that realizability plus uGC $\to (\beta, \delta)$-hypothesis stability (in the above example, if we add $f(x) = 1$ to $\mathcal{H}$ we lose our (β, δ)-hypothesis stability), we conjecture that a slightly weaker statement holds — if $\mathcal{H}$ is uGC and the target function is in $\mathcal{H}$, then we will get (β, δ)-hypothesis stability over all of Z except possibly a set of measure 0.

Convex hull

We state a simple application of the previous results, together with an obvious property of convex hulls.

Lemma *The convex hull $\mathcal{H}_c$ of $\mathcal{H}$ is compact if and only if $\mathcal{H}$ is compact.*

Theorem *(β, δ)-hypothesis stability of ERM on $\mathcal{H}_c$ for any measure with appropriate regularity conditions is necessary and sufficient for compactness of a uGC $\mathcal{H}$.*

References

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

[2] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889, 1988.

[3] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *Neural Information Processing Systems 14*, Denver, CO, 2000.

[4] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal Machine Learning Research*, 2001.

[5] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49, 2001.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

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

[8] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in advanced mathematics. Cambridge University Press, 1999.

[9] R.M. Dudley, E. Gine, and J. Zinn. Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

[10] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, Holland, 1996.

[11] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

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

[13] A.N. Kolmogorov and S.V. Fomin. *Introductory Real Analysis*. Dover, New York, 1970.

[14] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical report TR-2002-03, University of Chicago, 2002.

[15] S. Kutin and P. Niyogi. Almost everywhere algorithmic stability and generalization error. In *Proceedings of Uncertainty in AI*, Edmonton, Canada, 2002.

[16] S. Mendelson. A few notes on statistical learning theory. Technical report, Preprint, 2002.

[17] S. Mendelson. Geometric parameters in learning theory. Submitted for publication, 2003.

[18] S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. In *Lectures Notes in Statistics: Nonlinear Estimation and Classification, Proceedings from MSRI Workshop, D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick and B. Yu (eds.)*, volume 171, pages 107–124, Springer-Verlag, 2002.

[19] V. De La Pena. A General Class of Exponential Inequalities for Martingales and Ratios. *The Annals of Probability*, 27(1):537–564, 1999.

[20] T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS)*, 50(5):537–544, 2003.

[21] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, Berlin, 1984.

[22] M. Talagrand. A New Look at Independence. *The Annals of Probability*, 24:1–34, 1996.

[23] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, Washington, D.C., 1977.

[24] L.G. Valiant. A theory of learnable. *Proc. of the 1984, STOC*, pages 436–445, 1984.

[25] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[26] D. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.

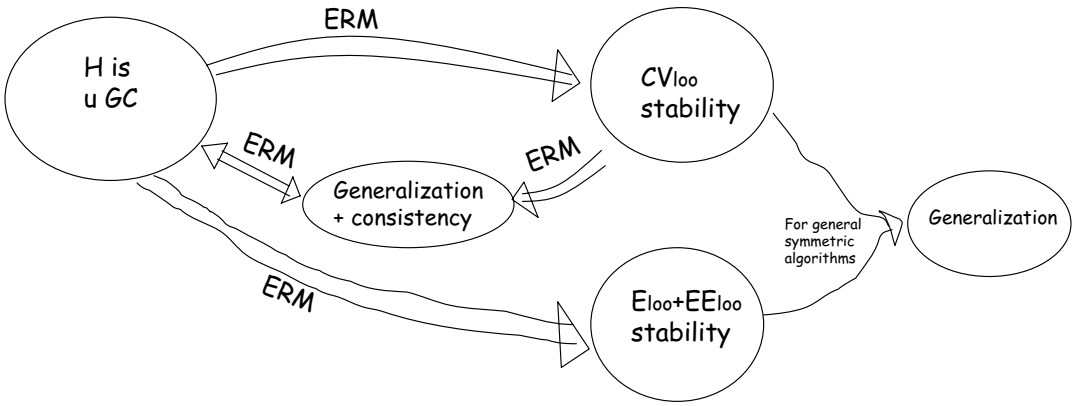

Figure 1: *An overall view of some of the properties discussed in this paper and their relations. Arrows are to be read as implies: for example* $a \Rightarrow b$ *means* a *implies* b*. For ERM the classical result is that generalization and consistency imply that* $\mathcal{H}$ *is uGC and are implied by it. The other relations represent some of the new results of this paper. In the diagram we can substitute the combination of* E_{loo} *and* EE_{loo} *stability with* $Eloo_{err}$ *stability. Notice that for ERM generalization implies consistency. As an example of a non-ERM algorithm, Tikhonov regularization implies uniform hypothesis stability which implies both* $CVEEE_{loo}$ *(ie* CV_{loo}*,* E_{loo} *and* EE_{loo} *stability) and LOO stability (ie* CV_{loo} *and* $Eloo_{err}$*). Note that Ivanov regularization in a RKHS is an example of ERM, whereas Tikhonov regularization is not ERM.*