Almost-everywhere algorithmic stability and generalization error

Samuel Kutin* Partha Niyogi†

March 25, 2002

Abstract

We introduce a new notion of algorithmic stability, which we call training stability. We show that training stability is sufficient for good bounds on generalization error. These bounds hold even when the learner has infinite VC dimension. In the PAC setting, training stability gives necessary and sufficient conditions for exponential convergence, and thus serves as a distribution-dependent analog to VC dimension.

Our proof generalizes an argument of Bousquet and Elisseeff (2001), who show that the more rigid assumption of uniform hypothesis stability implies good bounds on generalization error. We argue that weaker forms of hypothesis stability also give good bounds. We explore the relationships among VC dimension, generalization error, and different notions of stability.

1 Introduction

A major issue in the design of learning algorithms is to ensure that their performance generalizes successfully from training examples to novel examples not encountered before. Correspondingly, a central concern for learning theory is to create frameworks for the accurate analysis of generalization error. To date, the dominant theoretical paradigm for such an analysis has been the framework created by Vapnik and Chervonenkis [Vap98]. Bounds on the generalization error follow as a natural consequence of uniform convergence bounds that are obtained via compactness [CS02], the VC dimension of a space of classifiers, or the closely related V_γ dimension [ABDCBH93].

In this paper, we explore in some detail the viability of an alternative framework for the analysis of generalization error. The object of study is the learning algorithm rather than the space of classifiers. The learning algorithm is a map (effective procedure) from data sets to classifiers and if this map is *stable*, we show that exponential bounds on generalization error may be obtained. Several different notions of algorithmic stability are discussed and their

*Department of Computer Science, University of Chicago, 100 E. 58th Street, Chicago, IL 60637. Email: kutin@cs.uchicago.edu.

†Department of Computer Science, University of Chicago, 100 E. 58th Street, Chicago, IL 60637. Email: niyogi@cs.uchicago.edu.

interrelationships are studied. In particular, we show that the new notion of *training stability* is sufficient for good bounds in general and in the PAC setting [Val84] is both sufficient and necessary for successful generalization. We explore the relationship between VC dimension, generalization error, and various notions of stability. Several examples of learning algorithms are considered. The notion of algorithmic stability potentially allows us to consider a larger class of learning algorithms than Empirical Risk Minimization (ERM) which has been the focus of attention within the VC framework. It provides bounds on the generalization error without the need to prove uniform convergence. Finally, it potentially allows us to deal with learning algorithms that work with hypothesis classes of infinite VC dimension.

Algorithmic stability was first introduced by Devroye and Wagner [DW79]. An algorithm is stable at a training set S if any change of a single point in S yields only a small change in the output hypothesis. (We refer to this notion as "weak hypothesis stability.") Breiman [Bre96b] argues that unstable weak learners benefit from randomization algorithms such as bagging. He finds that, when the weak learner is stable, bagging does not help, and suggests that AdaBoost [FS97] is more effective in this case.

Kearns and Ron [KR99] consider both hypothesis stability and the weaker, related notion of error stability. They prove bounds on the error of leave-one-out estimates of error rates, but their arguments rely on the traditional notion of VC dimension [VC71]. Bousquet and Elisseeff [BE01, BE02] prove that algorithms which are "uniformly hypothesis stable" (i.e., stable at every training set) have low generalization error; their proof does not make any reference to VC dimension.

By avoiding VC theory, stability allows us to focus on a wider class of learning algorithms than empirical risk minimization (ERM). For example, Bousquet and Elisseeff show that regularization networks are stable [BE01]. They use stability to provide generalization error bounds for regularization-based learning algorithms that have been awkward to analyze within the VC framework.

Devroye and Wagner [DW79] were the first to observe a connection between the stability of an algorithm and its generalization error. Bousquet and Elisseeff [BE01] proved that uniform hypothesis stability implies low generalization error. Kearns and Ron [KR99] introduced the notion of error stability, and used it to get bounds on the error of leave-one-out estimates.

We discuss several notions of "almost-everywhere stability," including:

- *strong hypothesis stability*, where an algorithm is stable at most training sets,
- *weak hypothesis stability*, where an algorithm may not be stable at any training set, but changing one point usually leads to a small change in the final hypothesis,
- *training stability*, where changing one point usually leads to a small change in the error on points in the training set, and
- *CV stability*, where changing one point usually leads to a small change in the error on that one point.

Weak hypothesis stability is the definition introduced by Devroye and Wagner [DW79], while strong hypothesis stability lies in between this notion and the uniform stability of Bousquet and Elisseeff. Training stability is strictly weaker than weak hypothesis stability.

Our main result is that training stability gives good concentration bounds on generalization error:

Theorem 1.1 *Suppose a learning algorithm is* $\left(\frac{\lambda}{m}, e^{-\Omega(m)}\right)$*-training-stable. Then, if* $T_{\min} \leq \tau \leq T_{\max}$*, we have*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau) \leq 4 \exp\left(\frac{-\tau^2 m}{1440(\lambda + M)^2}\right)$$

for sufficiently large m*.*

(We use $\operatorname{Err}_D(f_S)$ to represent the true error rate of the algorithm given training set S, and $\operatorname{Err}_S(f_S)$ to represent the training error rate on S. We discuss our notation in detail in Section 2.1.)

Theorem 1.1 is a simplified version of Theorem 6.15. We prove Theorem 1.1 in Section 6.4, where we also state the bounds $T_{\min}$ and $T_{\max}$.

We also prove Theorem 6.13, which gives bounds on generalization error using weak hypothesis stability, and Theorem 6.12, which gives bounds on generalization error using strong hypothesis stability. The stronger assumptions give us tighter concentration bounds. Theorem 6.16 states that, for ERM, CV stability is sufficient for bounds on generalization error

Our proofs follow the argument of Bousquet and Elisseeff [BE01]. They use the method of “independent bounded differences,” developed by McDiarmid [McD89]. We use two extensions of McDiarmid’s Theorem [Kut02], which we state in Section 2.2. These stronger concentration inequalities enable us to prove our more general versions of Bousquet and Elisseeff’s result.

In Section 7, we discuss the implications of CV stability for PAC learnability. In Section 8, we discuss the relationship between finite VC dimension and our various notions of stability. We strengthen a result of Kearns and Ron [KR99]: they show that finite VC dimension implies weak error stability, and we show that finite VC dimension implies strong error stability.

In Section 9, we discuss the stability of various learning algorithms and the trade-offs between stability-based and VC theoretic analyses. Our most involved example of an almost-everywhere stable algorithm appears in a separate paper [KN01]: if a weak learner is uniformly hypothesis stable, then AdaBoost applied to that weak learner is strongly hypothesis stable.

We discuss uniform hypothesis stability in Section 3, and explain why we consider it to be too restrictive. We give various definitions of almost-everywhere stability, and discuss the relationships among these definitions, in Sections 4 and 5. In Section 4.1, we discuss why even weak hypothesis stability is too restrictive. In Section 4.2, we argue that certain weaker notions of stability, error stability and L_1 stability, are insufficient to give good bounds on generalization error.

The proof of Theorem 1.1 appears in Section 6; we also prove a number of variants of Theorem 1.1, where different assumptions about the learning algorithm yield tighter concentration bounds. Finally, in Section 10, we list some open questions and conjectures.

2 Preliminaries

We first describe the setting for our learning algorithms. There is a space $\mathcal{X}$ of points, or *instances*, with some unknown distribution Δ. A target operator takes as input an element $x \in \mathcal{X}$ and outputs some y in a set $\mathcal{Y}$ of labels, according to a conditional distribution function $F(y \mid x)$. For the sake of simplicity, we assume $\mathcal{Y} = \{-1, 1\}$. We let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, the space of *examples*. We write D for the distribution on $\mathcal{Z}$ induced by Δ and F.

One special setting is when there is a *target function* $f \colon \mathcal{X} \to \mathcal{Y}$, and a *noise level* η. In this case, the target operator outputs $f(x)$ with probability $1 - \eta$ and $1 - f(x)$ with probability η. We refer to the case $\eta = 0$ as the *noiseless setting*, and the case $\eta > 0$ as the *noisy setting*.

A *classifier*, also called a *hypothesis*, is a function $h \colon \mathcal{X} \to [-1, 1]$. Informally, a learning algorithm is given a weighted finite subset of $\mathcal{Z}$, and attempts to construct a classifier h such that $h(x)$ is a good approximation to the output of the target operator on input x. We define a learning algorithm formally in Definition 2.5.

We allow our classifiers to output values between -1 and 1; this corresponds to a confidence-rated prediction. However, unless otherwise stated, we assume that classifiers take values in $\{-1, 1\}$. We use $\mathcal{H}$ to denote a set of classifiers.

2.1 Notation

Note 2.1 If D is a distribution on a set X, we use the notation $x \sim D$ to mean that x is chosen from X according to distribution D.

Note 2.2 If we have some probability space Ω, and some property $\Phi : \Omega \to \{\text{true}, \text{false}\}$, then the notation

$$\forall^{\delta} \omega, \qquad \Phi(\omega)$$

means that

$$\Pr_{\omega \in \Omega}(\Phi(\omega)) \geq 1 - \delta.$$

In other words, property Φ holds for all but a δ fraction of Ω. We use this notation only when the underlying space Ω is understood.

Definition 2.3 The *cost* of a classifier h on a point $z \in \mathcal{Z}$ is a measure of the error h makes on z. We denote this cost by $c(h, z)$, and we require $0 \leq c(h, z) \leq M$ for some constant M.

Definition 2.4 The error rate of a classifier $h \colon \mathcal{X} \to [-1, 1]$, with respect to a distribution D, is

$$\mathrm{Err}_D(h) = \mathbf{E}_{z \sim D}(c(h, z)).$$

The error rate of a collection $\mathcal{H}$ of classifiers is

$$\mathrm{Err}_D(\mathcal{H}) = \min_{h \in \mathcal{H}} \{\mathrm{Err}_D(h)\}.$$

Note that we use this notation even when D has finite support. For $S \in \mathcal{Z}^m$, we use the notation Err_S to mean Err_p where p is the uniform distribution on S.

We leave off the subscript D (or S) when the underlying distribution is clear from context.

Definition 2.5 A *learning algorithm* $\mathcal{A}$ is a process which takes as input a finite training set $S \in \bigcup_m \mathcal{Z}^m$ and outputs a function $f_S \colon \mathcal{X} \to [-1, 1]$.

A *symmetric* learning algorithm depends only on the instances given to it, not on the order in which they are presented. For simplicity, we assume that all of our learning algorithms are symmetric.

Note 2.6 For $S \in \mathcal{Z}^m$, $S = (z_1, \ldots, z_m)$, we let S^i denote $S \setminus z_i$, the training set with the ith instance removed. So, for each i, $S^i \in \mathcal{Z}^{m-1}$. For $u \in \mathcal{Z}$, we let $S^{i,u}$ denote $S^i \cup \{u\}$, the training set with the ith instance replaced by u. So $S^{i,u} \in \mathcal{Z}^m$.

Definition 2.7 The *training error* of a learning algorithm on an input S is the error rate of f_S on the set S, or $\mathrm{Err}_S(f_S)$. The *true error* of a learning algorithm on an input S is the error rate of f_S on a randomly chosen example, or $\mathrm{Err}_D(f_S)$. The *generalization error* of a learning algorithm on an input S is the difference between the observed error rate and the true error rate, or

$$|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)|.$$

We use $\mathrm{gen}(S)$ to denote the (signed) generalization error of $\mathcal{A}$ on S:

$$\mathrm{gen}(S) = \mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S).$$

We write $\mu = \mathbf{E}_S(\mathrm{gen}(S))$. Hence μ is a function of m.

Our focus is on bounding the probability that the generalization error of a learning algorithm $\mathcal{A}$ is large. Note that this is not sufficient to imply that $\mathcal{A}$ is a good algorithm. However, it is easier to determine empirically whether an algorithm has good training error. Also, in general, generalization error bounds are more elusive than training error bounds.

Remark 2.8 There are other ways of analyzing true error. Some authors [Niy98, NG99, CS02] write true error as the sum of *approximation error*, which is the error rate $\mathrm{Err}_D(h_*)$ of the optimal classifier $h_* \in \mathcal{H}$, and *estimation error*, which is $\mathrm{Err}_D(f_S) - \mathrm{Err}_D(h_*)$. Approximation error measures how well $\mathcal{H}$ can fit the data, and estimation error measures the gap between the function obtained from S and the best possible function. Since our focus is on the learning algorithm $\mathcal{A}$, rather than on the representative power of the space $\mathcal{H}$, we do not use this decomposition.

Others (see, for example, Geman, et al. [GBD92]) write true error as the sum of *bias* and *variance*. This decomposition applies when we use the squared cost function $c(h, (x, y)) = (h(x) - y)^2$. Our focus in this paper is on the classification setting, where we generally use an absolute-value cost function.

2.2 Extensions of McDiarmid's Inequality

McDiarmid's method of independent bounded differences [McD89, McD98] gives concentration bounds on multivariate functions in terms of the maximum effect that changing one coordinate of the input can have on the output:

Definition 2.9 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω. We say that X is *uniformly difference-bounded by* c if the following holds: for any k, if $\omega, \omega' \in \Omega$ differ only in the kth coordinate, then

$$|X(\omega) - X(\omega')| \leq c. \tag{1}$$

Theorem 2.10 (McDiarmid [McD89]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is uniformly difference-bounded by c. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(-\frac{2\tau^2}{mc^2}\right)$$

Here, we state two generalizations [Kut02] of McDiarmid's Theorem which apply when Inequality (1) holds with high probability: changing one coordinate usually leads to a small change in the output, but not always. We need these results to extend Bousquet and Elisseeff's argument that uniform hypothesis stability gives good bounds on generalization error [BE01] to weaker notions of stability. We first describe the weaker conditions which we allow our random variables to satisfy, and we then state the theorems.

Definition 2.11 ([Kut02]) Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω. We say that X is *strongly difference-bounded by* (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \Pr(\omega \in B)$. If $\omega, \omega' \in \Omega$ differ only in the kth coordinate, and $\omega \notin B$, then

$$|X(\omega) - X(\omega')| \leq c.$$

Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b.$$

Definition 2.12 ([Kut02]) Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω. We say that X is *weakly difference-bounded by* (b, c, δ) if the following holds: for any k,

$$\forall^\delta (\omega, \upsilon) \in \Omega \times \Omega_k, \qquad |X(\omega) - X(\omega')| \leq c,$$

where $\omega'_k = \upsilon$ and $\omega'_i = \omega_i$ for $i \neq k$. In words, if we choose $\omega \in \Omega$, and $\upsilon \in \Omega_k$, and we construct ω' by replacing the kth entry of ω with υ, then the inequality holds for all but a δ fraction of the choices. Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b.$$

Theorem 2.13 ([Kut02]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by (b, c, δ). Assume $b \geq c > 0$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{8mc^2}\right) + \frac{mb\delta}{c}\right).$$

Theorem 2.14 ([Kut02]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by (b, c, δ). Assume $b \geq c > 0$, and assume $\delta \leq (c/b)^6$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2 \left(\exp \left(\frac{-\tau^2}{10mc^2 \left(1 + \frac{2\tau}{15mc}\right)} + \frac{mb\delta^{1/2}}{c} \exp\left(\frac{\tau b}{4mc^2}\right) \right) + m\delta^{1/2} \right).$$

For the proof of Theorem 1.1, we use the following simplified version of Theorem 2.14:

Theorem 2.15 ([Kut02]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by $(b, \frac{\lambda}{m}, \exp(-Km))$. Let $\mu = \mathbf{E}(X)$. If $0 < \tau \leq T(b, \lambda, K)$, and $m \geq N(b, \lambda, K, \tau)$, then*

$$\Pr(|X - \mu| \geq \tau) \leq 4 \exp\left(-\frac{\tau^2 m}{40\lambda^2}\right).$$

The bounds T and N are:

$$T(b, \lambda, K) = \min\left\{\frac{15\lambda}{2}, 4\lambda\sqrt{K}, \frac{\lambda^2 K}{b}\right\}$$
$$N(b, \lambda, K, \tau) = \max\left\{\frac{b}{\lambda}, \lambda\sqrt{40}, 3\left(\frac{24}{K} + 3\right)\ln\left(\frac{24}{K} + 3\right), \frac{1}{\tau}\right\}.$$

The above theorems do not give good bounds in the case $c = 0$. In Section 9.5, we instead use the following result:

Lemma 2.16 ([Kut02]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω.*

1. *If X is weakly difference-bounded by $(b, 0, \delta)$, then, for any $\omega, \chi \in \Omega$,*

$$|X(\omega) - X(\chi)| \leq mb.$$

2. *If X is weakly difference-bounded by $(b, 0, \delta)$, then there is some $\chi \in \Omega$ such that*

$$\Pr(X \neq X(\chi)) \leq m\delta.$$

3. *If X is strongly difference-bounded by $(b, 0, \delta)$, then there is some $\chi \in \Omega$ such that*

$$\Pr(X \neq X(\chi)) \leq \frac{m\delta}{2}.$$

3 Uniform stability

Definition 3.1 (Devroye and Wagner [DW79]) We say that a learning algorithm $\mathcal{A}$ has *leave-one-out stability* β if the following holds:

$$\forall S \in \mathcal{Z}^m, \quad \forall i, \quad \forall z \in \mathcal{Z}, \qquad |c(f_S, z) - c(f_{S^i}, z)| \leq \beta.$$

Definition 3.2 (Bousquet and Elisseeff [BE01]) We say that a learning algorithm $\mathcal{A}$ has *change-one stability* β if the following holds:

$$\forall S \in \mathcal{Z}^m, \quad \forall i, \quad \forall u, z \in \mathcal{Z}, \qquad |c(f_S, z) - c(f_{S^{i,u}}, z)| \leq \beta.$$

We also say that $\mathcal{A}$ has *uniform hypothesis stability* β, or is *uniformly* β*-hypothesis-stable*. (We use this terminology when we wish to contrast this notion with that of Definition 4.1 or 4.4.)

Note that, in both of these definitions, we view β as a function of m. We are most interested in the case where $\beta = \lambda/m$ for a constant λ.

Observation 3.3 *For any function β of m, if $\mathcal{A}$ has leave-one-out stability β, then $\mathcal{A}$ has change-one stability 2β.*

Remark 3.4 We include both change-one stability and leave-one-out stability for the sake of completeness, but the distinction between them is minor. Observation 3.3 states that leave-one-out stability implies change-one stability. The converse is not true (for example, $\mathcal{A}$ might have radically different behavior when the size of the training set is even or odd).[1] However, for any constant λ, change-one stability λ/m implies leave-one-out stability λ/m [KN01].

Devroye and Wagner [DW79] use stability to get bounds on the error of leave-one-out estimates. Bousquet and Elisseeff [BE01] were the first to obtain bounds on generalization error:

Theorem 3.5 (Bousquet and Elisseeff [BE01]) *If $\mathcal{A}$ has uniform hypothesis stability β, then, for all $\tau > 0$,*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \beta) \leq 2 \exp\left(\frac{-\tau^2 m}{2(m\beta + M)^2}\right).$$

Theorem 3.5 gives good bounds on generalization error when $\beta = O(1/m)$.

[1] Kearns and Ron [KR99, Theorem 12] make use of such an algorithm, where the behavior depends on whether the size of the training set is even or odd, to show that the error of the leave-one-out estimate depends on VC dimension.

3.1 Why uniform stability is too restrictive

Our goal in this paper is to extend Bousquet and Elisseeff's Theorem 3.5 to weaker notions of stability. In this section, we motivate this analysis by presenting some simple learning algorithms which are not uniformly hypothesis-stable.

We first show that uniform hypothesis stability is not directly applicable in the classification setting where f_S outputs a binary value. Bousquet and Elisseeff [BE01] discuss the setting in which the binary value is obtained by thresholding a real-valued quantity, in which case uniform hypothesis stability can be applied.

Definition 3.6 A learning algorithm $\mathcal{A}$ is a ± 1-*algorithm* if, for all $S \in \mathcal{Z}^m$, and for all $x \in \mathcal{X}$, $f_S(x) \in \{-1, 1\}$.

Theorem 3.7 *If $\mathcal{A}$ is a ± 1-algorithm, and $\mathcal{A}$ is uniformly β-hypothesis-stable for some $\beta < 1$, then $\mathcal{A}$ is the constant algorithm.*

Proof: For any $S \in \mathcal{Z}^m$ and $z \in \mathcal{Z}$, we must have $c(f_S, z) = 0$ or 1; f_S is either right or wrong on the example z. So, for any $\beta < 1$, the condition

$$\forall S \in \mathcal{Z}^m, \quad \forall i, \quad \forall u, z \in \mathcal{Z}, \qquad |c(f_S, z) - c(f_{S^{i,u}}, z)| \leq \beta$$

implies that

$$\forall S \in \mathcal{Z}^m, \quad \forall i, \quad \forall u \in \mathcal{Z}, \qquad f_S = f_{S^{i,u}},$$

which in turn implies that f_S is constant for every S. ■

We next show that uniform stability cannot apply in the case when $\mathcal{H}$ is finite.

Theorem 3.8 *Let $\mathcal{H}$ be a finite collection of classifiers, and let $\mathcal{A}$ be a learning algorithm which performs ERM over $\mathcal{H}$. Then, assuming $\mathcal{A}$ is not constant, $\mathcal{A}$ is not uniformly β-hypothesis-stable for any $\beta = o(1)$.*

Proof: Write $\mathcal{H} = \{h_1, \ldots, h_n\}$. For any i, define the region $R_{i,m} \subset \mathcal{Z}^m$ by $R_{i,m} = \{S \in \mathcal{Z}^m \mid f_S = h_i\}$. Since $\mathcal{A}$ is not constant, we assume that (for infinitely many m) some $R_{i,m}$ and $R_{j,m}$ are nonempty, so there must be some points on the boundary between regions: for some S, k, and u, $S \in R_{i,m}$, but $S^{k,u} \in R_{j,m}$. We conclude that $\mathcal{A}$ is not uniformly β-hypothesis-stable for any $\beta < \min_{i,j}\{d(i,j)\}$, where $d(i,j) = \sup_z\{|c(h_i, z) - c(h_j, z)|\}$. ■

In Sections 4 and 5, we introduce new notions of stability which allow us to handle these examples.

4 Strong and weak stability

In Section 3, we discuss uniform notions of hypothesis stability. Definition 3.2 requires that, for every training set S, and every S' which differs from S in only one point, the hypotheses f_S and $f_{S'}$ are close on all of $\mathcal{X}$. As we observe in Section 3.1, this definition is too restrictive.

We now discuss some more relaxed notions of stability. In Section 6, we prove that weak hypothesis stability can also give good bounds on generalization error. However, we show in Section 4.1 that even weak hypothesis stability is too restrictive. In Section 5, we introduce training stability, our most general notion of stability which is sufficient for good generalization error bounds.

In this section, we also discuss notions of L_1 stability and error stability. We use error stability, in combination with other properties, in the proofs of Section 6. However, we argue in Section 4.2 that L_1 and error stability, without any additional assumptions, are not sufficient for good bounds on generalization error.

The parameters β and δ in the definitions below are functions of m. We will chiefly be interested in the case where $\beta = O(1/m)$ and $\delta = \exp(-\Omega(m))$. In some natural examples, we have $\beta = 0$.

Definition 4.1 A learning algorithm $\mathcal{A}$ is *β-hypothesis-stable at S* if

$$\forall i \in \{1, \ldots, m\}, \quad \forall u \in \mathcal{Z}, \qquad \max_{z \in \mathcal{Z}} \{|c(f_S, z) - c(f_{S^{i,u}}, z)|\} \leq \beta.$$

We say that $\mathcal{A}$ is *strongly (β, δ)-hypothesis-stable*, or has *strong hypothesis stability* (β, δ), if

$$\forall^\delta S, \qquad \mathcal{A} \text{ is } \beta\text{-hypothesis-stable at } S,$$

where S is chosen according to the distribution D^m.

Definition 4.2 A learning algorithm $\mathcal{A}$ is *β-L_1-stable at S* if

$$\forall i \in \{1, \ldots, m\}, \quad \forall u \in \mathcal{Z}, \qquad \mathbf{E}_{z \in Z}\left(|c(f_S, z) - c(f_{S^{i,u}}, z)|\right) \leq \beta.$$

We say that $\mathcal{A}$ is *strongly (β, δ)-L_1-stable*, or has *strong L_1 stability* (β, δ), if

$$\forall^\delta S, \qquad \mathcal{A} \text{ is } \beta\text{-}L_1\text{-stable at } S,$$

where S is chosen according to the distribution D^m. We say that $\mathcal{A}$ is *uniformly β-L_1-stable*, or has *uniform L_1 stability β*, if $\mathcal{A}$ is β-L_1-stable at every training set S.

Definition 4.3 A learning algorithm $\mathcal{A}$ is *β-error-stable at S* if

$$\forall i \in \{1, \ldots, m\}, \quad \forall u \in \mathcal{Z}, \qquad |\operatorname{Err}_D(f_S) - \operatorname{Err}_D(f_{S^{i,u}})| \leq \beta.$$

We say that $\mathcal{A}$ is *strongly (β, δ)-error-stable*, or has *strong error stability* (β, δ), if

$$\forall^\delta S, \qquad \mathcal{A} \text{ is } \beta\text{-error-stable at } S,$$

where S is chosen according to the distribution D^m. We say that $\mathcal{A}$ is *uniformly β-error-stable*, or has *uniform error stability β*, if $\mathcal{A}$ is β-error-stable at every training set S.

Definition 4.4 (Devroye and Wagner [DW79]) A learning algorithm $\mathcal{A}$ is *weakly (β, δ)-hypothesis-stable*, or has *weak hypothesis stability* (β, δ), if, for any $i \in \{1, \ldots, m\}$,

$$\forall^\delta S, u, \qquad \max_{z \in \mathcal{Z}} \{|c(f_S, z) - c(f_{S^{i,u}}, z)|\} \leq \beta,$$

where $S \sim D^m$ and $u \sim D$.

Definition 4.5 A learning algorithm $\mathcal{A}$ is *weakly (β, δ)-L_1-stable*, or has *weak L_1 stability* (β, δ), if, for any $i \in \{1, \ldots, m\}$,

$$\forall^\delta S, u, \qquad \mathbf{E}_{z \in Z} \left(|c(f_S, z) - c(f_{S^{i,u}}, z)| \right) \leq \beta,$$

where $S \sim D^m$ and $u \sim D$.

Definition 4.6 (Kearns and Ron [KR99]) A learning algorithm $\mathcal{A}$ is *weakly (β, δ)-error-stable*, or has *weak error stability* (β, δ), if, for any $i \in \{1, \ldots, m\}$,

$$\forall^\delta S, u, \qquad |\operatorname{Err}_D(f_S) - \operatorname{Err}_D(f_{S^{i,u}})| \leq \beta,$$

where $S \sim D^m$ and $u \sim D$.

Remark 4.7 We use the term L_1 stability because

$$\mathbf{E}_{z \in Z} \left(|c(f_S, z) - c(f_{S^{i,u}}, z)| \right)$$

is the L_1 distance between the functions e_{f_S} and $e_{f_{S^{i,u}}}$ on $\mathcal{Z}$. We could also define L_p stability for any $p < \infty$. Note that hypothesis stability can be viewed as L_∞ stability in this sense.

Remark 4.8 For the notions of weak stability, we could instead include i inside the $\forall^\delta$, choosing i according to any distribution on $\{1, \ldots, m\}$. Since we require learning algorithms to be symmetric (see Definition 2.5), there is no difference.

Remark 4.9 The notion of weak hypothesis stability was first formulated by Devroye and Wagner [DW79], who refer to the concept simply as stability. Kearns and Ron [KR99] use the term hypothesis stability for what we now call weak hypothesis stability. Kearns and Ron also introduce error stability, which is what we call weak error stability. Both Devroye and Wagner and Kearns and Ron phrase their definitions in terms of leave-one-out stability, but, for consistency, we use change-one stability (see Remark 3.4).

Note 4.10 The definition of strong error stability in effect says that the random variable $\operatorname{Err}_D(f_S)$ (i.e., the true error rate) is strongly difference bounded in the sense of Definition 2.11. (The only difference is that strong error stability does not provide an absolute bound on $|\operatorname{Err}_D(f_S) - \operatorname{Err}_D(f_{S'})|$, but this quantity is clearly at most M.) Similarly, weak error stability says that $\operatorname{Err}_D(f_S)$ is weakly difference-bounded (see Definition 2.12).

We discuss this relationship explicitly in Section 6.3.

Observation 4.11 *The following all hold:*

- *Strong (β, δ) hypothesis stability implies weak (β, δ) hypothesis stability.*
- *Strong (β, δ) L_1 stability implies weak (β, δ) L_1 stability.*
- *Strong (β, δ) error stability implies weak (β, δ) error stability.*
- *Strong (β, δ) hypothesis stability implies strong (β, δ) L_1 stability.*

$$\begin{tikzcd}
\text{Strong hypothesis} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{Strong } L_1 \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{Strong error} \arrow[d, Rightarrow] \\
\text{Weak hypothesis} \arrow[r, Rightarrow] & \text{Weak } L_1 \arrow[r, Rightarrow] & \text{Weak error}
\end{tikzcd}$$

Figure 1: Implications among six notions of almost-everywhere stability

- *Strong (β, δ) L_1 stability implies strong (β, δ) error stability.*
- *Weak (β, δ) hypothesis stability implies weak (β, δ) L_1 stability.*
- *Weak (β, δ) L_1 stability implies weak (β, δ) error stability.*

We summarize the implications of Observation 4.11 in Figure 1.

Remark 4.12 There are no other implications between these concepts: for example, we can describe a learning algorithm which is strongly error stable, and weakly hypothesis stable, but not strongly L_1 stable. Examples refuting some other implications can be found in Section 9.

Remark 4.13 If $\mathcal{A}$ is absolutely β-hypothesis-stable, then it is strongly $(\beta, 0)$-hypothesis-stable. The converse is false, since there may be bad training sets which occur with probability zero. However, if $\mathcal{X}$ is discrete, then weak hypothesis stability $(\beta, 0)$ implies uniform hypothesis stability β.

4.1 Learning a one-dimensional threshold

In Section 3.1, we give two examples illustrating why uniform hypothesis stability is too restrictive. We can analyze both of these examples within the framework of strong hypothesis stability: it is possible for a nonconstant ± 1-algorithm to be strongly hypothesis stable, and we show in Section 9.3 that, under certain conditions, ERM over a finite space $\mathcal{H}$ is strongly hypothesis stable.

We now give an example illustrating why strong hypothesis stability is too restrictive. We also show that it is possible for ERM over a space of VC dimension 1 not to be weakly hypothesis stable.

Example 4.14 Let $\mathcal{X} = [0, 1]$. We consider the class $\mathcal{H}$ of threshold functions $g_\theta(x) = \text{sign}(x - \theta)$. (We assume for purposes of discussion that $g_\theta(\theta) = 1$).

Let $\mathcal{A}$ be an algorithm performing ERM over this space $\mathcal{H}$. If the labels on the training set S are generated by some θ, then we will output some θ_S close to θ. Specifically, we output some $\theta_S \in (a, b]$, where a is the maximal value for which $(a, -1) \in S$ and b is the minimal value for which $(b, 1) \in S$.

Note that $\text{VCdim}(\mathcal{H}) = 1$.

Suppose we run this algorithm on a training set S, obtaining a threshold θ_S. Regardless of how we compute θ_S from S, the probability that $\theta = \theta_S$ is 0. If $\theta \neq \theta_S$, we can construct an S' by removing some point in S and replacing it with a point lying between θ and θ_S. So, $f_{S'}$ will differ from f_S somewhere, which implies that $\max_z |c(f_S, z) - c(f_{S'}, z)| = 1$. Hence, $\mathcal{A}$ is not β-hypothesis-stable at any set S for any $\beta < 1$. We conclude that $\mathcal{A}$ is not strongly (β, δ)-hypothesis-stable for any $\beta < 1$, $\delta < 1$.

This argument does not apply to weak hypothesis stability: as $m \to \infty$, our threshold θ_S approaches θ. So the probability that a randomly chosen point would lie between θ and θ_S is very small.

However, we did not specify how $\mathcal{A}$ determines θ_S from θ. For example, we could let ξ be the average of all points in S, and output $\theta_S = a + (b - a)\xi$. In this case, any small change to S will yield a small change to θ_S, which is enough to force $\max_z |c(f_S, z) - c(f_{S'}, z)|$ to be 1.

Even weak hypothesis stability is therefore too restrictive a definition to handle ERM over spaces of finite VC dimension. We feel that weak hypothesis stability is not the "right" framework for discussing generalization error.

In Section 5, we will introduce the notion of training stability. We will see that the algorithm of this section is training stable, regardless of how $\mathcal{A}$ selects the threshold θ_S consistent with S.

Remark 4.15 We discuss Example 4.14 in more detail in Section 9.4.

4.2 Overtraining

We showed in Section 9.4 that ERM over a space of VC dimension 1 is not necessarily weakly hypothesis stable. However, by Proposition 9.9, the algorithm of Example 9.8 is strongly L_1-stable. We will show in Section 8 that ERM over any space of finite VC dimension is necessarily strongly L_1-stable.

It is reasonable to ask whether strong L_1 stability, or even some form of error stability, might be sufficient to prove bounds on generalization error. We now show that even uniform L_1 stability does not give good bounds on generalization error.

Example 4.16 Let $\mathcal{X} = [0, 1]$. We have some function $d : \mathbb{N} \to (0, 1]$, where $d(m) = o(1/m)$. Given a training set $S = \{(x_i, y_i)\}$, we define f_S to be the following function from $\mathcal{X}$ to $\{-1, 1\}$:

1. Given $x \in \mathcal{X}$, let $j = \arg\min_{1 \le i \le m} |x - x_i|$. So x_j is the nearest neighbor to x in $\mathcal{X}$.

2. If $|x - x_j| < d(m)$, return y_j. Otherwise, return 1.

This defines a learning algorithm $\mathcal{A}\colon S \mapsto f_S$.

If we take $d(m) = 0$, then $\mathcal{A}$ returns the correct labels on the training set and 1 elsewhere. For $d(m) > 0$, we use a nearest-neighbor approximation near the training points.

Theorem 4.17 *The learning algorithm $\mathcal{A}$ of Example 4.16 has uniform L_1 stability $4d(m)$.*

Proof: Given any two training sets S and S' which differ in one element, f_S and $f_{S'}$ differ on a region of size at most $4d(m)$. So

$$\mathbf{E}_z|c(f_S,z) - c(f_{S'},z)| \leq 4d(m).$$

We conclude that $\mathcal{A}$ has uniform L_1 stability $4d(m)$. ■

We now consider the generalization error of $\mathcal{A}$. Let η be the error rate (with respect to D) of the constant hypothesis 1; we can assume for purposes of discussion that $\eta = 1/2$.

Since the measure of $\{x : f_S(x) = -1\}$ is at most $2md(m)$, the generalization error is at least

$$\text{gen}(S) = \text{Err}_D(f_S) \geq \eta - 2md(m).$$

Since $d(m) = o(1/m)$, we have $\text{gen}(S) \to \eta$ as $m \to \infty$.

We therefore conclude:

Theorem 4.18 *Strong error stability, or even uniform L_1 stability, is not sufficient to prove bounds on generalization error.*

Remark 4.19 Kearns and Ron [KR99, Theorem 14] give an example of an algorithm which has weak error stability, but where leave-one-out estimates have high error. Their algorithm in fact has strong error stability, and is another example where strong error stability does not imply good bounds on generalization error. However, their algorithm is not weakly L_1 stable.

Remark 4.20 Example 4.16 can also be used to show that, for any $p < \infty$, strong L_p stability (see Remark 4.7) is insufficient to imply good bounds on generalization error.

5 Training stability and CV stability

We observe in Section 4.2 that error stability, and even L_1 stability, are not sufficient to prove good bounds on generalization error. Error stability implies that $\text{Err}_D(f_S)$ is concentrated about its mean. As we will see in Section 6, there are two other parts to our proof of good generalization error bounds:

- The training error $\text{Err}_S(f_S)$ is concentrated about its mean.
- The mean generalization error μ is small.

Remark 5.1 The algorithm of Example 4.16 satisfies the first of these conditions, but not the second. As $m \to \infty$, the mean μ approaches the error rate $\eta = \text{Err}_D(1)$.

In this section, we introduce new notions of stability which are chosen precisely to complete the proof described above. We first introduce *CV stability*, which will give us a bound on μ. We then introduce *overlap stability*, which we will use to show that training error is concentrated about its mean.

We show in Section 5.1 that CV stability implies weak error stability. Hence, we conclude Theorem 1.1: the combination of CV stability and overlap stability, which we call *training stability*, is sufficient for good bounds on generalization error.[2]

Definition 5.2 A learning algorithm $\mathcal{A}$ is (β,δ)*-cross-validation-stable*, or (β,δ)*-CV-stable*, if, for any $i \in \{1,\ldots,m\}$,

$$\forall^\delta S, u \qquad |c(f_S, u) - c(f_{S^{i,u}}, u)| \leq \beta.$$

We also say that $\mathcal{A}$ has *cross-validation stability* or *CV stability* (β,δ).

Definition 5.3 A learning algorithm $\mathcal{A}$ is (β,δ)*-overlap-stable*, or has *overlap stability* (β,δ), if, for any $i \in \{1,\ldots,m\}$,

$$\forall^\delta S, u, \qquad |\operatorname{Err}_{S^i}(f_S) - \operatorname{Err}_{S^i}(f_{S^{i,u}})| \leq \beta.$$

We call this notion overlap stability because it says that, for most training sets S, S' differing in only one coordinate, f_S and $f_{S'}$ have similar performance on $S \cap S'$. (Recall that S^i is S with the ith example removed, so $S^i = S \cap S'$.)

We now combine these two notions into one definition. Since both apply to the performance of S and S' on the training set $S \cup S'$, we call this joint notion training stability:

Definition 5.4 A learning algorithm $\mathcal{A}$ is (β,δ)*-training-stable*, or has *training stability* (β,δ), if

1. $\mathcal{A}$ has CV stability (β,δ).
2. $\mathcal{A}$ has overlap stability (β,δ).

Remark 5.5 In the definition of CV stability, the roles of S and $S^{i,u}$ are interchangeable, so we could have stated the definition as:

$$\forall^\delta S, u \qquad |c(f_S, z_i) - c(f_{S^{i,u}}, z_i)| \leq \beta,$$

where z_i is the ith element of S. If we use this formulation, then training stability refers only to the behavior of f_S and $f_{S^{i,u}}$ on the training set S.

Observation 5.6 *Weak hypothesis stability (β,δ) implies training stability (β,δ).*

Remark 5.7 The formulation of training stability is similar to the weak notions of stability defined in Section 4. We could also define a strong or uniform notion of training stability. However, for some common algorithms, for any z, $c(f_S, z) - c(f_{S^{i,u}}, z)$ is maximized at $u = z$, so strong CV stability would not be significantly different from strong hypothesis stability. It is in the weak sense that training stability is useful.

Remark 5.8 The algorithm of Example 4.16 always has training error 0, so it has overlap stability $(0, 0)$. We could thus augment Theorem 4.18 as follows: uniform L_1 stability, together with overlap stability, is not sufficient to prove bounds on generalization error.

[2]By Corollary 6.8, any algorithm performing ERM has training error concentrated about its mean. Hence, for ERM, CV stability alone is sufficient for good bounds on generalization error. This is the statement of Theorem 6.16.

5.1 CV stability and weak error stability

As we remark in Observation 5.6, weak hypothesis stability implies training stability (and, therefore, CV stability). We now show that CV stability implies weak L_1 stability, and hence weak error stability.

Theorem 5.9 *Let $\mathcal{A}$ be a (β, δ)-CV-stable learning algorithm. Then, for any function $\alpha(m) > 0$, $\mathcal{A}$ has weak L_1 stability $(2\beta + 2M\alpha, 2\delta/\alpha)$.*

Proof: Fix some i. By the definition of CV stability,

$$\forall^\delta S, z \qquad |c(f_S, z) - c(f_{S^{i,z}}, z)| \leq \beta.$$

We say that S is "good" if

$$\forall^\alpha z \qquad |c(f_S, z) - c(f_{S^{i,z}}, z)| \leq \beta.$$

We have

$$\begin{aligned}
\delta &\geq \Pr_{S,z}(|c(f_S, z) - c(f_{S^{i,z}}, z)| > \beta) \\
&= \Pr_{S,z}(|c(f_S, z) - c(f_{S^{i,z}}, z)| > \beta \mid S \text{ is bad}) \Pr_S(S \text{ is bad}) \\
&\geq \alpha \Pr_S(S \text{ is bad}),
\end{aligned}$$

so $\Pr_S(S \text{ is bad}) \leq \delta/\alpha$.

Similarly, $\Pr_{S,u}(S^{i,u} \text{ is bad}) \leq \delta/\alpha$. So, for all but a $2\delta/\alpha$ fraction of choices of S, u, both S and $S^{i,u}$ are good.

Now, suppose that S and $S^{i,u}$ are good. Then

$$\forall^\alpha z \qquad |c(f_S, z) - c(f_{S^{i,z}}, z)| \leq \beta. \tag{2}$$

$$\forall^\alpha z \qquad |c(f_{S^{i,u}}, z) - c(f_{S^{i,z}}, z)| \leq \beta. \tag{3}$$

(We use the fact that $(S^{i,u})^{i,z} = S^{i,z}$.)

So, all but a 2α fraction of z satisfy the inequalities of (2) and (3). Hence, by the triangle inequality,

$$\forall^{2\alpha} z \qquad |c(f_S, z) - c(f_{S^{i,u}}, z)| \leq 2\beta,$$

and therefore, whenever S and $S^{i,u}$ are both good,

$$\mathbf{E}_z(|c(f_S, z) - c(f_{S^{i,u}}, z)|) \leq 2\beta + 2\alpha M.$$

Since S and $S^{i,u}$ are both good with probability at least $1 - 2\delta/\alpha$, this implies weak L_1 stability $(2\beta + 2\alpha M, 2\delta/\alpha)$. ■

Corollary 5.10 *Let $\mathcal{A}$ be a (β, δ)-CV-stable learning algorithm. Then, for any function $\alpha(m) > 0$, $\mathcal{A}$ has weak error stability $(2\beta + 2\alpha M, 2\delta/\alpha)$.*

In Figure 2, we give an expanded version of Figure 1 which includes all of our notions of stability.

$$
\begin{tikzcd}
\text{Uniform hypothesis} \arrow[rr, Rightarrow] \arrow[d, Rightarrow] & & \text{Uniform } L_1 \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{Uniform error} \arrow[d, Rightarrow] \\
\text{Strong hypothesis} \arrow[rr, Rightarrow] \arrow[d, Rightarrow] & & \text{Strong } L_1 \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{Strong error} \arrow[d, Rightarrow] \\
\text{Weak hypothesis} \arrow[rr, Rightarrow] \arrow[d, Rightarrow] & & \text{Weak } L_1 \arrow[r, Rightarrow] & \text{Weak error} \\
\text{Training} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{CV} \arrow[ur, Rightarrow] & & \\
\text{Overlap} & & &
\end{tikzcd}
$$

Figure 2: Implications among twelve notions of almost-everywhere stability

6 Stability and generalization error

In this section, we prove our stronger versions of Theorem 3.5 [BE01]: that various notions of almost-everywhere stability imply good bounds on generalization error. We prove our main result, Theorem 1.1, in Section 6.4.

We prove a number of results, using different stability assumptions to draw slightly different conclusions. However, each of the proofs is essentially the same, and proceeds in three parts. Recall that $\mathrm{gen}(S) = \mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)$, and that $\mu = \mathbf{E}(\mathrm{gen}(S))$.

1. If $\mathcal{A}$ has CV stability (β, δ), then $|\mu| \leq \beta + \delta M$.
2. Under certain stability assumptions, $\mathrm{Err}_S(f_S)$ is weakly (or strongly) difference-bounded.
3. Weak (or strong) error stability implies that $\mathrm{Err}_D(f_S)$ is weakly (or strongly) difference-bounded.

We then make use of the following observation, which is immediate from Definitions 2.11 and 2.12.

Observation 6.1 *Let X, Y be random variables on $\Omega = \prod_{k=1}^{m} \Omega_k$. If X is strongly (or weakly) difference-bounded by (b_1, c_1, δ_1), and Y is strongly (or weakly) difference-bounded by (b_2, c_2, δ_2), then $X + Y$ is strongly (or weakly) difference-bounded by $(b_1 + b_2, c_1 + c_2, \delta_1 + \delta_2)$.*

We now conclude that $\mathrm{gen}(S)$ is strongly or weakly difference-bounded, so we apply Theorem 2.13 or 2.14 to obtain concentration bounds.

In Section 6.1, we show that CV stability implies that μ is small. In Section 6.2, we show that, under certain conditions, $\mathrm{Err}_S(f_S)$ is difference-bounded. In Section 6.3, we show that, under certain conditions, $\mathrm{Err}_D(f_S)$ is difference-bounded.

Finally, in Section 6.4, we explore how results from Sections 6.1, 6.2, and 6.3 can be combined to yield theorems relating stability to generalization error.

6.1 Mean generalization error is small

We now prove that CV stability implies that μ is small.

Lemma 6.2 (Bousquet and Elisseeff [BE01]) *For any learning algorithm $\mathcal{A}$, for any $i \in \{1, \ldots, m\}$,*

$$\mu = \mathbf{E}_{S,z \sim D^{m+1}} \left(c(f_S, z) - c(f_{S^{i,z}}, z) \right).$$

Proof: Fix i. Writing $S = (z_1, \ldots, z_m)$, we note that

$$\mathbf{E}_{S \sim D^m}(\mathrm{Err}_S(f_S)) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{E}_{S \sim D^m}(c(f_S, z_j)) = \mathbf{E}_{S \sim D^m}(c(f_S, z_i))$$

because $\mathcal{A}$ is symmetric. Since $S^{i,z}$ is simply S with the ith entry replaced by z, we also have

$$\mathbf{E}_{S \sim D^m}(\mathrm{Err}_S(f_S)) = \mathbf{E}_{S,z \sim D^{m+1}}(\mathrm{Err}_{S^{i,z}}(f_{S^{i,z}})) = \mathbf{E}_{S,z \sim D^{m+1}}(c(f_{S^{i,z}}, z)). \tag{4}$$

By definition,

$$\mathbf{E}_{S\sim D^m}(\mathrm{Err}_D(f_S)) = \mathbf{E}_{S,z\sim D^{m+1}}(c(f_S,z)). \tag{5}$$

So, subtracting (4) from (5), we get that

$$\mu = \mathbf{E}_{S\sim D^m}(\mathrm{gen}(S)) = \mathbf{E}_{S,z\sim D^{m+1}}\left(c(f_S,z) - c(f_{S^{i,z}},z)\right)$$

■

Theorem 6.3 *CV stability (β,δ) implies that $|\mu| \leq \beta + \delta M$.*

Proof: Fix some i. Let $\psi(S,u) = c(f_S,u) - c(f_{S^{i,u}},u)$. By Lemma 6.2,

$$\mu = \mathbf{E}_{S,u\sim D^{m+1}}(\psi(S,u)),$$

so

$$|\mu| \leq \mathbf{E}_{S,u\sim D^{m+1}}(|\psi(S,u)|).$$

By assumption, $\mathcal{A}$ is (β,δ)-CV-stable. So, if we choose $S \sim D^m$ and $u \sim D$, then, with probability at least $1-\delta$, S and u satisfy

$$|\psi(S,u)| = |c(f_S,u) - c(f_{S^{i,u}},u)| \leq \beta.$$

For any S and u, we know $|\psi(S,u)| \leq M$. We conclude that

$$\mathbf{E}_{S,u\sim D^{m+1}}(|\psi(S,u)|) \leq (1-\delta)\beta + \delta M \leq \beta + \delta M.$$

■

6.2 Training error is difference-bounded

We now consider the random variable $\mathrm{Err}_S(f_S)$. We show that $\mathrm{Err}_S(f_S)$ is strongly difference-bounded under an assumption of strong hypothesis stability, and weakly difference-bounded under an assumption of overlap stability. (We introduced overlap stability precisely because it is the relevant quantity in this argument.) We also show that, if $\mathcal{A}$ performs ERM, then $\mathrm{Err}_S(f_S)$ is uniformly difference-bounded.

We first prove a poor uniform bound on $\mathrm{Err}_S(f_S)$.

Observation 6.4 $\mathrm{Err}_S(f_S)$ *is uniformly difference-bounded by M.*

Proof: For any S, $0 \leq \mathrm{Err}_S(f_S) \leq M$. So, for any $S \in \mathcal{Z}^m$, any i, and any $z \in \mathcal{Z}$,

$$|\,\mathrm{Err}_S(f_S) - \mathrm{Err}_{S^{i,z}}(f_{S^{i,z}})| \leq M.$$

■

We next consider the consequences of strong hypothesis stability.

Lemma 6.5 *If $\mathcal{A}$ is strongly (β,δ)-hypothesis-stable, then* $\mathrm{Err}_S(f_S)$ *is strongly difference-bounded by $(M, \beta + \frac{M}{m}, \delta)$.*

Proof: Choose some S, and suppose that $\mathcal{A}$ is β-hypothesis-stable at S. Choose any $i \in \{1, \ldots, m\}$ and $u \in \mathcal{Z}$, and let S' denote $S^{i,u}$.

Writing $S = (z_1, \ldots, z_m)$, we have

$$|\operatorname{Err}_S(f_S) - \operatorname{Err}_{S'}(f_{S'})| \leq \frac{1}{m}|c(f_S, z_i) - c(f_{S'}, u)| + \frac{1}{m}\sum_{j \neq i}|c(f_S, z_j) - c(f_{S'}, z_j)| \leq \frac{M}{m} + \beta.$$

Since $\mathcal{A}$ is (β, δ) hypothesis stable, the above bound holds for all i, z with probability at least $1 - \delta$. Together with Observation 6.4, this completes the proof. ■

Next, we show that $\operatorname{Err}_S(f_S)$ is weakly difference-bounded using overlap stability.

Lemma 6.6 *If $\mathcal{A}$ is (β, δ)-overlap-stable, then $\operatorname{Err}_S(f_S)$ is weakly difference-bounded by $(M, \beta + \frac{M}{m}, \delta)$.*

Proof: Fix some $i \in \{1, \ldots, m\}$. Choose some $S \sim D^m$ and $u \sim D$; let $S' = S^{i,u}$, and let $T = S^i$ (i.e., T is the overlap between S and S'.

Writing $S = (z_1, \ldots, z_m)$, we have

$$|\operatorname{Err}_S(f_S) - \operatorname{Err}_{S'}(f_{S'})| \leq \frac{1}{m}|c(f_S, z_i) - c(f_{S'}, u)| + \frac{1}{m}\left|\sum_{j \neq i}(c(f_S, z_j) - c(f_{S'}, z_j))\right| \leq \frac{M}{m} + \frac{m-1}{m}|\operatorname{Err}_{S^i}(f_S) - \operatorname{Err}_{S^i}(f_{S'})|.$$

By overlap stability, all but a δ fraction of S, u are good: i.e.,

$$|\operatorname{Err}_{S^i}(f_S) - \operatorname{Err}_{S^i}(f_{S'})| \leq \beta$$

and hence

$$|\operatorname{Err}_S(f_S) - \operatorname{Err}_{S'}(f_{S'})| \leq \frac{M}{m} + \beta.$$

Together with Observation 6.4, this completes the proof. ■

Finally, we show that, if an algorithm performs ERM, then $\operatorname{Err}_S(f_S)$ is uniformly difference-bounded by $\frac{M}{m}$. This bound is better than either of the above bounds obtained via stability.

Lemma 6.7 *Let $\mathcal{H}$ be a space of classifiers. For any $S \in \mathcal{Z}^m$, $i \in \{1, \ldots, m\}$, and $z \in \mathcal{Z}$,*

$$|\operatorname{Err}_S(\mathcal{H}) - \operatorname{Err}_{S^{i,z}}(\mathcal{H})| \leq \frac{M}{m}.$$

Proof: Let S' denote $S^{i,z}$. Write $S = (z_1, \ldots, z_m)$ and $S' = (z'_1, \ldots, z'_m)$, where $z_j = z'_j$ for all $j \neq i$. Since S and S' are simply two sets which differ in the ith coordinate, we may assume without loss of generality that $\operatorname{Err}_S(\mathcal{H}) \leq \operatorname{Err}_{S'}(\mathcal{H})$.

Let h_* be a classifier so that $\operatorname{Err}_S(\mathcal{H}) = \operatorname{Err}_S(h_*)$. Then

$$\operatorname{Err}_{S'}(h_*) = \frac{1}{m}\sum_{j=1}^{m} c(h_*, z'_j) = \frac{1}{m}\left(\sum_{j \neq i} c(h_*, z_j)\right) + \frac{1}{m}c(h_*, z'_i) \leq \operatorname{Err}_S(h_*) + \frac{M}{m}.$$

Hence,

$$\mathrm{Err}_S(\mathcal{H}) \le \mathrm{Err}_{S'}(\mathcal{H}) \le \mathrm{Err}_{S'}(h_*) \le \mathrm{Err}_S(h_*) + \frac{M}{m} = \mathrm{Err}_S(\mathcal{H}) + \frac{M}{m}.$$

This proves that

$$|\,\mathrm{Err}_S(\mathcal{H}) - \mathrm{Err}_{S'}(\mathcal{H})| \le \frac{M}{m}.$$

■

Corollary 6.8 *If $\mathcal{A}$ performs ERM, then* $\mathrm{Err}_S(f_S)$ *is uniformly difference-bounded by* $\frac{M}{m}$.

Proof: For any $S \in \mathcal{Z}^m$, any $i \in \{1, \ldots, m\}$, and any $z \in \mathcal{Z}$,

$$|\,\mathrm{Err}_S(f_S) - \mathrm{Err}_{S^{i,z}}(f_{S^{i,z}})| \le \frac{M}{m}.$$

■

6.3 True error is difference-bounded

We now use stability to show that $\mathrm{Err}_D(f_S)$ is difference-bounded. First, we note an analog of Observation 6.4.

Observation 6.9 $\mathrm{Err}_D(f_S)$ *is uniformly difference-bounded by M.*

Proof: For any S, $0 \le \mathrm{Err}_D(f_S) \le M$. So, for any $S \in \mathcal{Z}^m$, any i, and any $z \in \mathcal{Z}$,

$$|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_D(f_{S^{i,z}})| \le M.$$

■

We now note that the definition of error stability is precisely what we need to prove that $\mathrm{Err}_D(f_S)$ is weakly or strongly difference-bounded.

Observation 6.10 *If $\mathcal{A}$ is strongly (β, δ)-error-stable, then* $\mathrm{Err}_D(f_S)$ *is strongly difference-bounded by (M, β, δ).*

Observation 6.11 *If $\mathcal{A}$ is weakly (β, δ)-error-stable, then* $\mathrm{Err}_D(f_S)$ *is weakly difference-bounded by (M, β, δ).*

6.4 Putting it all together

We can construct various theorems from the components in the preceding sections. We combine the following:

1. CV stability for Theorem 6.3.

2. Overlap stability for Lemma 6.6 or ERM for Corollary 6.8. Note that we can prove stronger bounds by assuming strong hypothesis stability and using Lemma 6.5.

3. Weak error stability for Observation 6.11. If we assume strong error stability, we can instead use Observation 6.10. Note that weak error stability follows from CV stability by Corollary 5.10.

We now list some ways of combining the above ingredients into theorems about generalization error. This list is not intended to be exhaustive.

Theorem 6.12 *If $\mathcal{A}$ is strongly (β, δ)-hypothesis-stable, then, for any $\tau > 0$,*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \beta + M\delta) \leq 2\left(\exp\left(\frac{-\tau^2 m}{8(2m\beta + M)^2}\right) + \frac{4m^2 M\delta}{2m\beta + M}\right).$$

Proof: By Lemma 6.5, $\operatorname{Err}_S(f_S)$ is strongly difference-bounded by $(M, \beta + \frac{M}{m}, \delta)$. Strong hypothesis stability implies strong error stability, so, by Observation 6.10, $\operatorname{Err}_D(f_S)$ is strongly difference-bounded by (M, β, δ). We conclude, using Observation 6.1, that $\operatorname{gen}(S)$ is strongly difference-bounded by $(2M, 2\beta + \frac{M}{m}, 2\delta)$.

Strong hypothesis stability implies CV stability. Hence, by Theorem 6.3, $\mu \leq \beta + \delta M$. The result now follows from Theorem 2.13. ■

Theorem 6.13 *If $\mathcal{A}$ is weakly (β, δ)-hypothesis-stable, then, for any $\tau > 0$,*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \beta + M\delta) \leq 2\sqrt{2}m\delta^{1/2} + 2\exp\left(\frac{-\tau^2 m}{10(2m\beta + M)^2\left(1 + \frac{2\tau}{30m\beta + 15M}\right)} + \frac{2\sqrt{2}m^2 M\delta^{1/2}}{2m\beta + M}\exp\left(\frac{\tau m M}{2(2m\beta + M)^2}\right)\right).$$

Proof: Weak hypothesis stability implies overlap stability, so, by Lemma 6.6, $\operatorname{Err}_S(f_S)$ is weakly difference-bounded by $(M, \beta + \frac{M}{m}, \delta)$. Weak hypothesis stability also implies weak error stability, so, by Observation 6.11, $\operatorname{Err}_D(f_S)$ is weakly difference-bounded by (M, β, δ). We conclude, using Observation 6.1, that $\operatorname{gen}(S)$ is weakly difference-bounded by $(2M, 2\beta + \frac{M}{m}, 2\delta)$.

Weak hypothesis stability implies CV stability. Hence, by Theorem 6.3, $\mu \leq \beta + \delta M$. The result now follows from Theorem 2.14. ■

Remark 6.14 We can slightly improve the constants in Theorems 6.12 and 6.13 by observing that the same S and S' are "bad" for $\operatorname{Err}_S(f_S)$ and $\operatorname{Err}_D(f_S)$. Hence, strong hypothesis stability implies that $\operatorname{gen}(S)$ is strongly difference-bounded by $(2M, 2\beta + \frac{M}{m}, \delta)$, and similarly for weak hypothesis stability.

Theorem 6.15 *If $\mathcal{A}$ is (β, δ)-training-stable, then, for any $\tau > 0$,*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \beta + M\delta) \leq 2m(2m+1)^{1/2}\delta^{1/2} + 2\exp\left(\frac{-\tau^2 m}{10(3m\beta + 3M)^2\left(1 + \frac{2\tau}{45m\beta + 45M}\right)} + \frac{2m^2 M(2m+1)^{1/2}\delta^{1/2}}{3m\beta + 3M}\exp\left(\frac{\tau m M}{2(3m\beta + 3M)^2}\right)\right).$$

Proof: Training stability implies overlap stability, so, by Lemma 6.6, $\mathrm{Err}_S(f_S)$ is weakly difference-bounded by $(M, \beta + \frac{M}{m}, \delta)$.

Also, training stability (β, δ) implies CV stability (β, δ). We now apply Corollary 5.10 with $\alpha = \frac{1}{m}$; we conclude that $\mathcal{A}$ has weak error stability $(2\beta + \frac{2M}{m}, 2m\delta)$. Hence, by Observation 6.11, $\mathrm{Err}_D(f_S)$ is weakly difference-bounded by $(M, 2\beta + \frac{2M}{m}, 2m\delta)$.

We conclude, using Observation 6.1, that $\mathrm{gen}(S)$ is weakly difference-bounded by $(2M, 3\beta + \frac{3M}{m}, (2m+1)\delta)$.

By Theorem 6.3, $\mu \leq \beta + \delta M$. The result now follows from Theorem 2.14. ■

Proof of Theorem 1.1: The proof of Theorem 1.1 is almost the same as the proof of Theorem 6.15. We assume training stability $\left(\frac{\lambda}{m}, \delta\right)$ where $\delta = \exp(-Km)$, and we conclude that $\mu \leq \frac{\lambda}{m} + \delta M$ and that $\mathrm{gen}(S)$ is weakly difference-bounded by $\left(2M, \frac{3\lambda+3M}{m}, (2m+1)\delta\right)$. Choose some $K' < K$; then, for some N, for $m \geq N$, we know $(2m+1)\delta \leq \exp(-K'm)$.

We now apply Theorem 2.15. For $\tau \leq T_2(2M, 3\lambda + 3M, K')$, and for $m \geq M_2(2M, 3\lambda + 3M, K', \tau)$ (and $m \geq N$),

$$\Pr\left(|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)| > \tau + \frac{\lambda}{m} + Me^{-K'm}\right) \leq 4\exp\left(\frac{-\tau^2 m}{360(\lambda + M)^2}\right).$$

If we now assume $\tau \geq \frac{\lambda}{m} + Me^{-K'm}$, we get

$$\Pr(|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)| > 2\tau) \leq 4\exp\left(\frac{-\tau^2 m}{360(\lambda + M)^2}\right).$$

Finally, we replace 2τ with τ, achieving the desired result.

Note that

$$T_{\min} = 2\left(\frac{\lambda}{m} + Me^{-K'm}\right),$$
$$T_{\max} = 2T(2M, 3\lambda + 3M, K'),$$

where T is the function of Theorem 2.15. ■

Theorem 6.16 *If $\mathcal{A}$ performs empirical risk minimization, and has CV stability (β, δ), then, for any $\tau > 0$,*

$$\Pr(|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)| > \tau + \beta + M\delta) \leq \sqrt{8m^3\delta} +$$
$$2\exp\left(\frac{-\tau^2 m}{10(2m\beta + 3M)^2\left(1 + \frac{2\tau}{15(2m\beta+3M)}\right)} + \frac{m(m+1)M\sqrt{2m\delta}}{2m\beta + 3M}\exp\left(\frac{\tau(m+1)M}{4(2m\beta + 3M)^2}\right)\right).$$

Proof: By Corollary 6.8, $\mathrm{Err}_S(f_S)$ is uniformly difference-bounded by $\frac{M}{m}$. By Corollary 5.10 with $\alpha = \frac{1}{m}$, $\mathcal{A}$ is weakly $(2\beta + \frac{2M}{m}, 2m\delta)$-error-stable; so, by Observation 6.11, $\mathrm{Err}_D(f_S)$ is weakly difference-bounded by $(M, 2\beta + \frac{2M}{m}, 2m\delta)$.

We conclude that $\mathrm{gen}(S)$ is weakly difference-bounded by $(M + \frac{M}{m}, 2\beta + \frac{3M}{m}, 2m\delta)$. The result now follows from Theorem 6.3 and Theorem 2.14. ■

Each of the above theorems has a simple hypothesis. We could also combine the results of the preceding sections in more complicated ways. The next theorem serves as an example of this approach; we also use it in Section 9.2.

Theorem 6.17 *If $\mathcal{A}$ performs ERM, is $(\beta_{CV}, \delta_{CV})$-CV-stable, and is strongly (β, δ)-error-stable, then, for any $\tau > 0$,*

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \beta_{CV} + M\delta_{CV}) \leq 2\left(\exp\left(\frac{-\tau^2 m}{8(m\beta + M)^2}\right) + \frac{m(m+1)M\delta}{m\beta + M}\right).$$

Proof: By Corollary 6.8, $\operatorname{Err}_S(f_S)$ is uniformly difference-bounded by $\frac{M}{m}$. By Observation 6.10, $\operatorname{Err}_D(f_S)$ is strongly difference-bounded by (M, β, δ). We conclude that $\operatorname{gen}(S)$ is strongly difference-bounded by $(M + \frac{M}{m}, \beta + \frac{M}{m}, \delta)$.

By Theorem 6.3, $\mu \leq \beta_{\mathrm{CV}} + \delta_{\mathrm{CV}} M$. The result now follows from Theorem 2.13. ■

7 CV stability and learnability

In this section, we discuss the relationship between CV stability and learnability.

Theorem 6.3 states that, if $\mathcal{A}$ is CV stable, then the average generalization error μ approaches 0 as $m \to \infty$. This implies that CV stable algorithms are good in the following sense: the expected error rate of the output classifier approaches the optimal error rate.

Corollary 7.1 *Let $\mathcal{H}$ be a space of classifiers, and let $\mathcal{A}$ be a learning algorithm performing ERM over $\mathcal{H}$. If $\mathcal{A}$ has CV stability (β, δ), where $\beta, \delta \to 0$ as $m \to \infty$, then $\mathbf{E}_S(\operatorname{Err}_D(f_S)) \to \operatorname{Err}_D(\mathcal{H})$ as $m \to \infty$.*

Note that this statement does not refer to the VC dimension of $\mathcal{H}$.

Proof: By Theorem 6.3,

$$\lim_{m\to\infty} (\mathbf{E}_S(\operatorname{Err}_D(f_S)) - \mathbf{E}_S(\operatorname{Err}_S(f_S))) = 0. \tag{6}$$

Let $h_* \in \mathcal{H}$ be an optimal classifier; i.e., $\operatorname{Err}_D(h_*) = \operatorname{Err}_D(\mathcal{H})$. Clearly

$$\operatorname{Err}_D(h_*) = \mathbf{E}_S(\operatorname{Err}_S(h_*)). \tag{7}$$

By our choice of h_*, $\operatorname{Err}_D(h_*) \leq \operatorname{Err}_D(f_S)$ for every S. And, by the definition of empirical risk minimization, $\operatorname{Err}_S(f_S) \leq \operatorname{Err}_S(h_*)$ for every S. Combining these inequalities with Equations (6) and (7) yields the desired result. ■

In the PAC setting, we can say something even stronger:

Theorem 7.2 *Let $\mathcal{H}$ be a space of ± 1-classifiers, and let $\mathcal{A}$ be a learning algorithm performing ERM over $\mathcal{H}$. Suppose that our examples are generated to be consistent with some $h_0 \in \mathcal{H}$.*

Then $\mathcal{A}$ has CV stability $(0, \delta)$, with $\delta = \exp(-\Omega(m))$, if and only if $\operatorname{Err}_D(f_S) \to 0$ exponentially in m.

Remark 7.3 In the setting of Theorem 7.2, ERM always has overlap stability $(0, 0)$. So CV stability is equivalent to training stability.

Proof: First, we note that, for ± 1-classifiers, any $\beta < 1$ is equivalent, so we assume $\beta = 0$. We also assume $M = 1$.

Since $\mathrm{Err}_D(h_0) = 0$, we have $\mathrm{Err}_S(h_0) = 0$ for all S. By the definition of ERM, we therefore have $\mathrm{Err}_S(f_S) = 0$ for all S, which implies $\mathrm{gen}(S) = \mathrm{Err}_D(f_S)$. We also have, for all S and z, $c(\mathrm{Err}_{S^{i,z}}, z) = 0$. Hence,

$$\begin{array}{lll} \text{CV stability } (0,\delta) & \iff & \forall^\delta S, z \quad c(f_S, z) - c(f_{S^{i,z}}, z) = 0 \\ & \iff & \forall^\delta S, z \quad c(f_S, z) = 0 \\ & \iff & \mathbf{E}_{S,z}(c(f_S, z)) \leq \delta \\ & \iff & \mathbf{E}_S(\mathbf{E}_z(c(f_S, z))) \leq \delta \\ & \iff & \mathbf{E}_S(\mathrm{Err}_D(f_S)) \leq \delta. \end{array}$$

So, $\mathbf{E}_S(\mathrm{Err}_D(f_S)) \to 0$ exponentially in m if and only if $\mathcal{A}$ is $(0,\delta)$-CV-stable where $\delta \to 0$ exponentially in m.

By Theorem 1.1, CV stability $(0, \exp(-\Omega(m)))$ implies exponential concentration bounds on $\mathrm{gen}(S)$. So, if we assume exponential CV stability, we get $\mathrm{Err}_D(f_S) \to 0$ exponentially in m. ■

CV stability thus gives necessary and sufficient conditions for distribution-dependent PAC learning. The bounds of Theorem 7.2 are analogous to obtained using annealed entropy (see, for example, Vapnik [Vap98]); however, VC theory does not give necessary and sufficient distribution-dependent conditions for fast convergence. We conjecture that "CV theory" is to the distribution-dependent setting what VC theory is to the distribution-free setting.

Corollary 7.4 *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 $\mathcal{X}$, and any $h_0 \in \mathcal{H}$, ERM over $\mathcal{H}$ is $(0, e^{-Km})$-CV-stable (or, equivalently, $(0, e^{-Km})$-training-stable) with respect to the distribution on $\mathcal{Z}$ generated by Δ and h_0.*

2. $\mathrm{VCdim}(\mathcal{H}) < \infty$.

Proof: By Theorem 7.2, statement 1 is equivalent to saying that $\mathrm{Err}_D(f_S) \to 0$ exponentially with m for every distribution on $\mathcal{Z}$. By VC theory, this is equivalent to statement 2. ■

8 VC dimension and stability

Finite VC dimension [Vap98] is a necessary and sufficient condition for good distribution-free bounds on generalization error. Our goal is to use stability to get distribution-dependent bounds on generalization error. It is reasonable to hope that distribution-dependent information may help us to bound generalization error even for learning algorithms with infinite VC dimension: Bousquet and Elisseeff [BE01] show that regularization has uniform hypothesis stability $O(1/m)$, and we discuss the stability of finite language learning in Section 9.5.

It may be that stability is a strictly more general concept than VC dimension; i.e., finite VC dimension may imply training stability.

In this section, we show that ERM over a class of finite VC dimension is strongly error stable, and even strongly L_1 stable. In Section 9.4, we discuss an example where ERM over a class of finite VC dimension is not weakly hypothesis stable. It remains open whether ERM over a class of finite VC dimension is necessarily training stable.

We begin with Vapnik's uniform convergence theorem [Vap82]. We follow the notation of Kearns and Ron [KR99].

Theorem 8.1 (Vapnik [Vap82, Theorem 6.7]) *Let $\mathcal{H}$ be a hypothesis class with VC dimension $d < m$. Then, for every $m > 4$, and for any $\delta > 0$,*

$$\forall^\delta S, \quad \forall h \in \mathcal{H}, \qquad |\operatorname{Err}_D(h) - \operatorname{Err}_S(h)| < M \ \operatorname{VC}(d, m, \delta),$$

where

$$\operatorname{VC}(d, m, \delta) = 2\sqrt{\frac{d(\ln \frac{2m}{d} + 1) + \ln \frac{9}{\delta}}{m}}.$$

Kearns and Ron [KR99] show that finite VC dimension implies error stability in their sense, which we call *leave-one-out error stability.* (They make the simplifying assumption $M = 1$.)

Proposition 8.2 (Kearns and Ron [KR99, Lemma 5]) *ERM over a hypothesis class of VC dimension d has leave-one-out error stability* $(2\operatorname{VC}(d, m-1, \delta/2), \delta)$ *for any* $\delta > 0$.

Since their definition of error stability is slightly different from ours, we now state and prove our version of this statement:

Proposition 8.3 *ERM over a hypothesis class of VC dimension d has weak error stability* $(2M\operatorname{VC}(d, m, \delta/2), \delta)$ *for any* $\delta > 0$.

Proof: Fix some $\delta > 0$. Suppose we choose S, S' randomly differing in only one coordinate. Then, by Theorem 8.1, each of the following holds with probability at least $1 - \frac{\delta}{2}$:

$$\forall h \in \mathcal{H}, \qquad |\operatorname{Err}_D(h) - \operatorname{Err}_S(h)| \le M \operatorname{VC}\left(d, m, \frac{\delta}{2}\right). \tag{8}$$

$$\forall h \in \mathcal{H}, \qquad |\operatorname{Err}_D(h) - \operatorname{Err}_{S'}(h)| \le M \operatorname{VC}\left(d, m, \frac{\delta}{2}\right). \tag{9}$$

So, with probability at least $1 - \delta$, both (8) and (9) hold.

Assuming (8), and using the fact that $\operatorname{Err}_S(f_S) \le \operatorname{Err}_S(f_{S'})$, we get

$$\begin{aligned}
\operatorname{Err}_D(f_S) &\le \operatorname{Err}_S(f_S) + M \operatorname{VC}\left(d, m, \frac{\delta}{2}\right) \\
&\le \operatorname{Err}_S(f_{S'}) + M \operatorname{VC}\left(d, m, \frac{\delta}{2}\right) \qquad (10) \\
&\le \operatorname{Err}_D(f_{S'}) + 2M \operatorname{VC}\left(d, m, \frac{\delta}{2}\right).
\end{aligned}$$

Similarly, if (9) holds, we get

$$\mathrm{Err}_D(f_{S'}) \le \mathrm{Err}_D(f_S) + 2M\,\mathrm{VC}\left(d, m, \frac{\delta}{2}\right). \tag{11}$$

Combining Inequalities (10) and (11), we conclude that

$$\forall^\delta S, S' \qquad |\,\mathrm{Err}_D(f_S) - \mathrm{Err}_D(f_{S'})| \le 2M\,\mathrm{VC}\left(d, m, \frac{\delta}{2}\right).$$

■

With only a bit more work, we can prove the following:

Theorem 8.4 *ERM over a hypothesis class of VC dimension d has strong error stability $\left(2M\,\mathrm{VC}(d, m, \delta) + \frac{2M}{m}, \delta\right)$ for any $\delta > 0$.*

Proof: Fix some δ. By Theorem 8.1, if we choose $S \sim D^m$, then, with probability at least $1 - \delta$,

$$\forall h \in \mathcal{H} \qquad |\,\mathrm{Err}_D(h) - \mathrm{Err}_S(h)| \le M\,\mathrm{VC}(d, m, \delta). \tag{12}$$

Assume S satisfies (12). Write $S = (z_1, \ldots, z_m)$. Let $S' = S^{i,z}$ for some i, z, and write $S' = (z'_1, \ldots, z'_m)$.

We first note that, for any $h \in \mathcal{H}$,

$$\begin{aligned} |\,\mathrm{Err}_S(h) - \mathrm{Err}_{S'}(h)| &= \left| \frac{1}{m}\sum_{i=1}^m c(h, z_i) - \frac{1}{m}\sum_{i=1}^m c(h, z'_i) \right| \\ &= \left| \frac{1}{m}(c(h, z_i) - c(h, z)) \right| \le \frac{M}{m}. \end{aligned} \tag{13}$$

So, we have

$$\begin{aligned} |\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)| &\le M\,\mathrm{VC}(d, m, \delta) && \text{by (12)} \\ |\,\mathrm{Err}_S(f_S) - \mathrm{Err}_{S'}(f_{S'})| &\le \frac{M}{m} && \text{by Corollary 6.8} \\ |\,\mathrm{Err}_{S'}(f_{S'}) - \mathrm{Err}_S(f_{S'})| &\le \frac{M}{m} && \text{by (13)} \\ |\,\mathrm{Err}_S(f_{S'}) - \mathrm{Err}_D(f_{S'})| &\le M\,\mathrm{VC}(d, m, \delta) && \text{by (12)} \end{aligned}$$

Combining these four inequalities yields

$$|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_D(f_{S'})| \le 2M\,\mathrm{VC}(d, m, \delta) + \frac{2M}{m}.$$

Since this holds for any $S' = S^{i,z}$, the theorem is proved. ■

It may seem that there is a trade-off between Proposition 8.3 and Theorem 8.4; we prove a stronger notion of stability in Theorem 8.4, but with a different parameter β. In fact, Theorem 8.4 is a strict improvement over Proposition 8.3:

Proposition 8.5 *For any $m > 4$, $d < m$, and $\delta > 0$,*

$$2M\,\mathrm{VC}(d,m,\delta) + \frac{2M}{m} < 2M\,\mathrm{VC}\left(d,m,\frac{\delta}{2}\right).$$

Proof: Let $A = \mathrm{VC}(d,m,\delta)$, and $B = \mathrm{VC}(d,m,\frac{\delta}{2})$. ¿From the definition of the VC function (see Theorem 8.1), we have

$$\left(\frac{A}{2}\right)^2 + \frac{\ln 2}{m} = \left(\frac{B}{2}\right)^2,$$

or, equivalently,

$$(B-A)(B+A) = B^2 - A^2 = \frac{4\ln 2}{m}.$$

Since $0 < A < B < 1$, we have $B + A < 2$, and hence

$$B - A > \frac{2\ln 2}{m}.$$

So,

$$2M\,\mathrm{VC}\left(d,m,\frac{\delta}{2}\right) - 2M\,\mathrm{VC}(d,m,\delta) > \frac{4M\ln 2}{m} > \frac{2M}{m},$$

which is what we wanted to prove. ■

Remark 8.6 It would be nice if we could combine Theorem 8.4, Corollary 6.8, and Observation 6.10 to draw conclusions about the generalization error of ERM over a space of finite VC dimension. Such an argument would indicate that the statement of strong error stability captures the essence of finite VC dimension, and would be necessary for the claim that stability subsumes VC theory.

However, no matter what value we choose for δ, the value of β from Theorem 8.4 is at least $\Theta((\ln m/m)^{1/2})$. We need strong error stability with $\beta = O(1/m)$ to apply our extended version of McDiarmid's Theorem. So we cannot use the results of this section to get good generalization error bounds on learners of finite VC dimension.

We can also draw a conclusion about the L_1 stability of ERM over a space with finite VC dimension:

Theorem 8.7 *Let $\mathcal{H}$ be a space of $\{-1,1\}$-classifiers with $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Let $\eta = \mathrm{Err}_D(\mathcal{H})$. Then, for any $\delta > 0$, ERM over $\mathcal{H}$ has strong L_1 stability $\left(2\eta + 4M\,\mathrm{VC}(d,m,\delta) + \frac{2M}{m}, \delta\right)$.*

Proof: Fix some δ. By Theorem 8.1, with probability at least $1-\delta$, a randomly-generated training set S satisfies (12) from Theorem 8.4.

Let h_* be the optimal classifier $h_* = \arg\min_{h\in\mathcal{H}} \mathrm{Err}_D(h)$. Assume S satisfies (12). Then

$$\begin{aligned}\mathrm{Err}_S(f_S) &\le \mathrm{Err}_S(h_*)\\ &\le \mathrm{Err}_D(h_*) + M\,\mathrm{VC}(d,m,\delta)\\ &\le \eta + M\,\mathrm{VC}(d,m,\delta).\end{aligned}$$

By (12),

$$|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| \leq M\,\mathrm{VC}(d,m,\delta),$$

and hence

$$\operatorname{Err}_D(f_S) \leq \eta + 2M\,\mathrm{VC}(d,m,\delta). \tag{14}$$

By the argument in the proof of Theorem 8.4, Inequalities (12) and (13), together with Corollary 6.8, imply

$$|\operatorname{Err}_D(f_{S'}) - \operatorname{Err}_S(f_S)| \leq M\,\mathrm{VC}(d,m,\delta) + \frac{2M}{m},$$

and hence

$$\operatorname{Err}_D(f_{S'}) \leq \eta + 2M\,\mathrm{VC}(d,m,\delta) + \frac{2M}{m}. \tag{15}$$

For $\{-1,1\}$-classifiers, $\operatorname{Err}_D(h)$ is a measure of the probability that h is wrong. So, for any classifiers h_1, h_2,

$$\Pr_z(c(h_1,z) \neq c(h_2,z)) \leq \operatorname{Err}_D(h_1) + \operatorname{Err}_D(h_2).$$

Applying this inequality to f_S and $f_{S'}$, and using (14) and (15), we get

$$\mathbf{E}_z(|c(f_S,z) - c(f_{S'},z)|) \leq 2\eta + 4M\,\mathrm{VC}(d,m,\delta) + \frac{4M}{m}. \tag{16}$$

Let β denote the right-hand-side of Inequality (16). We conclude that $\mathcal{A}$ is β-L_1-stable at S.

The probability that $\mathcal{A}$ is β-L_1-stable at S is thus at least $1-\delta$. ■

Theorem 8.7 is most useful when $\eta = 0$; i.e., when the examples in $\mathcal{Z}$ are generated by a rule in $\mathcal{H}$.

9 Examples of stable and unstable learners

Bousquet and Elisseeff [BE01] prove that regularization is uniformly β-hypothesis-stable for $\beta = O(1/m)$. Corollary 7.4 states that, in the PAC setting, ERM over a class of finite VC dimension is training stable. In this section, we discuss the stability of several other learners.

9.1 Stability-preserving algorithms

Our most involved example of a stable algorithm is AdaBoost [FS97]. AdaBoost is a widely-used sequential reweighting scheme for converting a weak learner to a strong learner. It can be shown [KN01] that, if the weak learner is uniformly hypothesis stable, the strong learner is weakly hypothesis stable. The proof is technical, and beyond the scope of this paper.

Bagging [Bre96a] is another procedure for converting a weak learner to a strong learner: in each round, the weak learner is called on m points sampled with replacement from the original input. Evgeniou, et al. [EPE01] argue that bagging increases uniform hypothesis stability (see Definition 3.2), but they use fewer than m points each time they resample the input, and their bounds do not apply in the case $\beta = O(1/m)$ which we consider.

Poggio, et al. [PRM02] prove that, if the weak learner is uniformly hypothesis stable, then the bagged learner is strongly hypothesis stable. Their argument has the same limitation as that of Evgeniou, et al.: they use fewer than m points each time they resample the input.

Freund, at al. [FMS01] discuss a variant of bagging, which outputs a linear combination of a set of hypotheses with the weight of a hypothesis determined by its error rate. They show that this algorithm is stable.

9.2 $|\mathcal{H}| = 2$

Example 9.1 Consider the learning algorithm $\mathcal{A}$ which performs ERM over the class $\mathcal{H}$ containing only two functions: the constants 1 and -1. $\mathcal{A}$ simply returns the label which occurs most frequently in the training set. (If the training set is perfectly evenly divided between 1-examples and -1-examples, we output 1.)

Let p be $\Pr_{(x,y)\sim D}(y = 1)$; for purposes of discussion, we assume $p \geq 1/2$. The question is: for which p is $\mathcal{A}$ stable?

Theorem 9.2 *Let $\mathcal{A}$ be the learning algorithm of Example 9.1.*

1. *If $p \geq 1/2$, then $\mathcal{A}$ is uniformly $(2p-1)$-error-stable. $\mathcal{A}$ is weakly $(0, \delta)$-L_1-stable and $(0, \delta)$-training-stable for $\delta(m) \sim 1/\sqrt{2\pi m}$.*

2. *If $p > 1/2$, then $\mathcal{A}$ is strongly $(0, \delta)$-hypothesis-stable, where*

$$\delta = \exp\left(-\frac{1}{8}\left(2 - \frac{1}{p}\right)^2 m + O(1)\right).$$

3. *If $p = 1/2$, then $\mathcal{A}$ is not weakly (β, δ)-L_1-stable, (β, δ)-CV-stable, or (β, δ)-overlap-stable for any $\beta < 1$ and any $\delta = o(m^{-1/2})$.*

Proof: We begin with Parts 1 and 3.

Proof of Parts 1 and 3: For every S, f_S is 1 or -1, so $\mathrm{Err}_D(f_S)$ is p or $1-p$. Hence, $|\mathrm{Err}_D(f_S) - \mathrm{Err}_D(f'_S)|$ is always at most $2p-1$. This proves that $\mathcal{A}$ is uniformly $(1-2p)$-error-stable.

For any S and S', either $f_S = f_{S'}$ everywhere or $f_S \neq f_{S'}$ everywhere. So, for any $\beta < 1$, $\mathcal{A}$ is weakly (β, δ)-L_1-stable, (β, δ)-CV-stable, and (β, δ)-overlap-stable if and only if

$$\Pr_{S,S'}(f_S \neq f_{S'}) \leq \delta,$$

where the probability is taken over all S, S' differing in only one coordinate.

Suppose we choose S, S' by first choosing $T \sim D^{m+1}$, and then selecting two points to be removed. If $m+1$ is even, then $f_S \neq f'_S$ if and only if:

1. T contains exactly $(m+1)/2$ each of 1-examples and -1-examples, and

2. S and S' are each formed by removing examples with different labels (i.e., $y_i \neq y'_i$). Since T contains the same number of 1-examples and -1 examples, the probability that this occurs (given Condition 1) is exactly $(m+1)/2m$.

So, when m is odd,

$$\begin{aligned}\Pr_{S,S'}(f_S \neq f_{S'}) &= \binom{m+1}{(m+1)/2} p^{(m+1)/2}(1-p)^{(m+1)/2}\frac{m+1}{2m} \\ &\leq \binom{m+1}{(m+1)/2}\frac{1}{2^{m+1}}\frac{m+1}{2m} \sim \frac{1}{\sqrt{2\pi m}}\end{aligned}$$

by Stirling's formula. When m is even, we get the same asymptotic behavior.

This proves Part 1.

When $p = 1/2$, we have, for odd m,

$$\Pr_{S,S'}(f_S \neq f_{S'}) = \binom{m+1}{(m+1)/2}\frac{1}{2^{m+1}}\frac{m+1}{2m} \sim \frac{1}{\sqrt{2\pi m}},$$

and similarly for even m. So, in this case, we cannot have (β, δ) stability (in the weak L_1, CV, or overlap sense) for any $\beta < 1$ and any $\delta = o(m^{-1/2})$. This proves Part 3. □

Proof of Part 2: We now consider the case $p > 1/2$.

We choose some training set $S \sim D^m$. Let $\xi_i(S) = 1$ if the ith example in S is labeled -1, and $\xi_i(S) = 1 - 1/p$ if the ith example is labeled 1. Then $\mathbf{E}_S(\xi_i) = 0$. We let $\zeta = \sum_i \xi_i$. By Chernoff's inequality, for any α,

$$\Pr(\zeta \geq \alpha) < e^{-\alpha^2/2m}.$$

If S contains at least $(m/2)+2$ examples labeled 1, then $\mathcal{A}$ is 0-stable at S; the output is 1, and remains 1 if any coordinate of S is changed. Therefore,

$$\begin{aligned}\delta &= \Pr(\mathcal{A} \text{ is not 0-stable at } S) \\ &\leq \Pr(S \text{ contains at most } (m/2)+1 \text{ examples labeled } 1) \\ &\leq \Pr\left(\zeta \geq \left(\frac{m}{2}-1\right) + \left(\frac{m}{2}+1\right)\left(1-\frac{1}{p}\right)\right) = \Pr\left(\zeta \geq \frac{m}{2}\left(2-\frac{1}{p}\right)+\frac{1}{p}\right) \\ &\leq \exp\left(-\frac{1}{2m}\left(\frac{m}{2}\left(2-\frac{1}{p}\right)+\frac{1}{p}\right)^2\right) = \exp\left(-\frac{1}{8}\left(2-\frac{1}{p}\right)^2 m + O(1)\right).\end{aligned}$$

□

This concludes the theorem. ■

Remark 9.3 When $p = 1/2$, we can combine Part 1 of Theorem 9.2 with Theorem 6.17: $\beta = \delta = \beta_{\mathrm{CV}} = 0$, and $\delta_{\mathrm{CV}} \sim 1/\sqrt{2\pi m}$. So, for any $\tau > 0$, for sufficiently large m,

$$\Pr\left(|\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)| > \tau + \frac{C}{\sqrt{m}}\right) \leq 2\left(\exp\left(\frac{-\tau^2 m}{8}\right)\right) \quad (17)$$

for any $C > 1/\sqrt{2\pi}$. Note that the right-hand side of (17) is a constant when $\tau = \Theta(1/\sqrt{m})$, so it is reasonable to assume $\tau > C/\sqrt{m}$. Writing $\epsilon = 2\tau$, we get the following: For any $\epsilon > 2C/\sqrt{m}$,

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \epsilon) \leq 2\left(\exp\left(\frac{-\epsilon^2 m}{32}\right)\right).$$

This is exactly the sort of bound we hope to get. It may be possible to use such an argument for p close to $1/2$, and use Part 2 of Theorem 9.2 with Theorem 6.12 for p significantly larger than $1/2$, to derive distribution-free bounds on the generalization error of the learning algorithm of Example 9.1.

Remark 9.4 A similar example, and the corresponding $\Omega(m^{-1/2})$ bound on hypothesis stability, also appears in Kearns and Ron [KR99, Theorem 11], who cite Devroye, et al. [DGL96, Chapter 24].

9.3 $|\mathcal{H}| < \infty$

We now attempt to generalize Theorem 9.1 to the case where $\mathcal{H}$ contains more than two classifiers, but is still finite.

Theorem 9.5 *Let $\mathcal{H}$ be a finite collection of classifiers, and let $\mathcal{A}$ be a learning algorithm which performs ERM over $\mathcal{H}$. Suppose that there exists a unique classifier $h_* \in \mathcal{H}$ which minimizes true error (i.e., there is a unique h_* for which $\operatorname{Err}_D(h_*) = \operatorname{Err}_D(\mathcal{H})$).*

Then $\mathcal{A}$ is strongly $(0, \delta)$-hypothesis-stable for $\delta = \exp(-\Omega(m))$.

Proof: We argue that, for sufficiently large m, f_S is almost always equal to the optimal classifier h_*.

Let $\eta = \operatorname{Err}_D(h_*)$. Let ϵ be the minimum possible gap between the error rate of h_* and the error rate of another classifier $h \in \mathcal{H}$:

$$\epsilon = \min_{h \neq h_*}\{\operatorname{Err}_D(h) - \operatorname{Err}_D(h_*)\}.$$

Suppose we are given a training set S of size m. We can use Chernoff's inequality to bound the probability that the error rate of h_* is at least $\eta + \epsilon/2$:

$$\Pr_S\left(\operatorname{Err}_S(h_*) \geq \eta + \frac{\epsilon}{2}\right) \leq \exp(-\epsilon^2 m/8).$$

Similarly, since every $h \neq h_*$ has $\operatorname{Err}_D(h) \geq \eta + \epsilon$,

$$\forall h \neq h_*, \qquad \Pr_S\left(\operatorname{Err}_S(h) \leq \eta + \epsilon - \frac{\epsilon}{2}\right) \leq \exp(-\epsilon^2 m/8).$$

In order for f_S to be anything other than h_*, we must have $\operatorname{Err}_S(f_S) \leq \operatorname{Err}_S(h_*)$, which implies that either $\operatorname{Err}_S(f_S) \leq \eta + \epsilon/2$ or $\operatorname{Err}_S(h_*) \geq \eta + \epsilon/2$. So,

$$\Pr_S(f_S \neq h_*) \leq |\mathcal{H}| \exp(-\epsilon^2 m/8).$$

■

Remark 9.6 In the case where the labels are actually generated according to some $h_0 \in \mathcal{H}$, Theorem 9.5 follows from a lemma of Blumer, et al. [BEHW87], who state that the probability that there is a hypothesis h with error rate $\mathrm{Err}_D(h) = \epsilon$ which makes no errors on a training set of size m is less than $|\mathcal{H}|(1-\epsilon)^m$.

9.4 Hyperplanes and maximum margin

If we wish to argue that some notion of stability is the "right" framework for discussing generalization error, we should be able to prove directly that ERM over a class of finite VC dimension is stable. In this section, we focus on ERM over spaces of hyperplanes in $\mathbb{R}^k$.

Remark 9.7 By Corollary 7.4, in the PAC setting, ERM over a space of finite VC dimension is training stable. In particular, all of the algorithms discussed in this section are training stable.

However, in the proof of Corollary 7.4, we simply argue that any good learner is training stable, and then cite VC theory. Where possible, we prefer a direct proof of stability.

We begin with the one-dimensional case first introduced in Section 4.1.

Example 9.8 Let $\mathcal{X} = [0,1]$. We consider the class $\mathcal{H}$ of Example 4.14: $\mathcal{H}$ contains all threshold functions $g_\theta(x) = \mathrm{sign}(x-\theta)$. (We assume for purposes of discussion that $g_\theta(\theta) = 1$).

Suppose we are given a training set S of examples generated consistently with some g_θ. Let a be the maximal value for which $(a,-1) \in S$, and b the minimal value for which $(b,1) \in S$. Then $\theta \in (a,b]$, and S is consistent with any $g_{\hat{\theta}}$ for $\hat{\theta} \in (a,b]$.

A learning algorithm which performs ERM over $\mathcal{H}$ can return any $\hat{\theta} \in (a,b]$. We consider three schemes for choosing $\hat{\theta}$:

1. Return $\hat{\theta} = b$. This classifier is guaranteed to have one-sided error.
2. Return $\hat{\theta} = (a+b)/2$. This is the maximum margin classifier.
3. Determine $\hat{\theta} \in (a,b]$ using some complicated function of the entire data set.

The question is: what is the stability of these three schemes?

Proposition 9.9 *All three schemes are strongly L_1-stable. None of the schemes is strongly (β,δ)-hypothesis-stable for $\beta < 1$, $\delta < 1$.*

Proof: Let S be a training set. If we remove $(b,1)$ from S, and replace it with some other point, we will change the final hypothesis. This proves that none of the schemes is strongly (β,δ)-hypothesis-stable for any $\beta < 1$, $\delta < 1$.

If we choose any point $z \in S$, and construct S' by replacing z with some z', then f_S and $f_{S'}$ will agree, except possibly on $(a,b]$, unless we remove $(a,-1)$ or $(b,1)$. Even in this case, the sets will agree on $(a',b]$ or $(a,b']$, where $(a',-1)$ and $(b',1)$ are the points in S closest to $(a,-1)$ and $(b,1)$. By standard VC theoretic arguments [Vap98], the sizes of these intervals are small for almost all S, so we will get strong L_1 stability for reasonable values of β and δ. See Theorem 8.7 for a more detailed, general version of this argument. ■

We have almost categorized the stability of Example 9.8 according to the six definitions of stability in Section 4. The remaining notion is weak hypothesis stability. Since each scheme only outputs ± 1-classifiers, we may take $\beta = 0$.

Proposition 9.10 *Scheme 1 of Example 9.8 is weakly $(0, \frac{2}{m+1})$-hypothesis-stable. Scheme 2 is weakly $(0, \frac{4}{m+1})$-hypothesis-stable. Scheme 3 is not weakly $(0, \delta)$-hypothesis-stable for any $\delta < 1$.*

Proof: We begin with Scheme 3. Since the threshold $\hat{\theta}$ depends on the entire data set, any small change will move the threshold slightly, yielding a different finite hypothesis.

Next, consider Scheme 1. Choosing S and $S^{i,z}$ is equivalent to first choosing $m+1$ points independently from $\mathcal{Z}$, and then deciding which one will be called z and which will be called z_i. So, choose these $m+1$ points, and let $(b, 1)$ be the minimal point labeled 1. Clearly, if $(b, 1) \in S$ and $(b, 1) \in S^{i,z}$, we have $f_S = f_{S^{i,z}}$. So the hypotheses are different only when $(b, 1) = z$ or $(b, 1) = z_i$. This happens with probability $2/(m+1)$.

Finally, we consider Scheme 2, the maximum margin approach. Again, we first choose $m+1$ points, and then decide which will be z and which will be z_i. Let $(b, 1)$ and $(a, -1)$ be the minimal 1-example and maximal -1-example. We have $f_S \neq f_{S^{i,z}}$ only when z or z_i is $(a, -1)$ or $(b, 1)$. The probability that this happens is

$$\frac{2}{m+1} + \frac{2}{m+1} - \frac{2}{m(m+1)} = \frac{4m-2}{m(m+1)} \leq \frac{4}{m+1}.$$

■

We conclude the following:

Corollary 9.11 *Empirical risk minimization over a class of finite VC dimension is not necessarily weakly hypothesis stable.*

However, the above analysis leaves open the following possibility:

Conjecture 9.12 *Given any class of finite VC dimension d, there exists an algorithm which performs ERM over that class which is weakly $(0, \delta)$ hypothesis stable for $\delta = O(d/m)$.*

Note that such a δ is not sufficiently small for us to get good bounds using Theorem 6.13.

Remark 9.13 Proposition 9.10 implies that Scheme 1 is "more stable" than Scheme 2 (maximum margin). However, Scheme 2 would seem to be a better approach. This is a sign that the stability analysis misses some important aspects of learning.

If we generalize to axis-parallel hyperplanes in higher dimensions, the same argument applies: there will be one or two extremal points which determine the exact location of the hyperplane, and we will have weak $(0, O(1/m))$ hypothesis stability. If we consider rectangular regions (i.e., regions of the form $\prod_{i=1}^{k} I_i$, where $I_i \subset \mathbb{R}$ is an interval), the same argument applies, with $2k$ extremal points.

Now, suppose our space $\mathcal{H}$ is the space of hyperplanes in k-dimensional space.

Example 9.14 Suppose we are given a training set $\{(\mathbf{x}_i, y_i)\}$, with $\mathbf{x}_i \in \mathbb{R}^k$. The *maximum margin* hyperplane separating the data is the hyperplane $\mathbf{w} \cdot \mathbf{x} + b$ which minimizes $\mathbf{w} \cdot \mathbf{w}$, subject to

$$\forall i, \qquad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1.$$

The *maximum margin algorithm* returns the maximum margin hyperplane. Note that, if there is a separating hyperplane, this algorithm finds it; in this setting, we can say that the maximum margin algorithm performs ERM.

We consider the action of the maximum margin algorithm on a training set S. There will be some subset of *support vectors* in S for which $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$. These are the points in S closest to the hyperplane.

Suppose we start with $m+1$ points, and let T be the set of support points. We now choose z and z_i from among our $m+1$ points, forming S and S'. We will have $f_S = f_{S'}$ unless z or z_i lies inside T. We conclude that:

Theorem 9.15 *The maximum margin algorithm of Example 9.14 is weakly* $(0, \delta)$*-hypothesis-stable, where*

$$\delta = \frac{2\mathbf{E}_{S \sim D^{m+1}}(|T|)}{m+1},$$

and $|T|$ *is the number of support points among a training set of size* $m+1$.

Remark 9.16 Theorem 9.15 is reminiscent of a result of Vapnik [Vap95, Theorem 5.2], who argues that the expected error rate of support vector machines is at most $\frac{\mathbf{E}(|T|)}{m-1}$. Evgeniou, et al. [EPP00] also give a VC analysis of support vector machines.

Remark 9.17 Under certain conditions on D, the following argument applies: the system of equations $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$ has $k+1$ unknowns (b and the coordinates of w). So, with probability 1, there will only be $k+1$ support points. (In fact, if there are more, than we should get the same hyperplane even if one of them is removed.)

9.5 Learning finite languages

We now consider the problem of language learning. There is some finite alphabet Σ, and our space of instances X is Σ^*, the set of words over Σ (with some distribution). There is a finite target language $L \subset \Sigma^*$, and an instance x is labeled 1 if $x \in L$ and 0 if $x \notin L$.

Example 9.18 Consider the learning algorithm $\mathcal{A}$ which, given S, returns the language L_S, where $L_S = \{x : (x, 1) \in S\}$. We ignore negative examples, and return the smallest language containing all positive examples.

It is clear that the space of languages returned by $\mathcal{A}$ has infinite VC dimension. However, $\mathcal{A}$ has good generalization error. To see this, we define the following random variables:

$$n_0(S) = |\{x \in L : (x, 1) \notin S\}|$$
$$n_1(S) = |\{x \in L : (x, 1) \text{ occurs exactly once in } S\}|$$

Observation 9.19 *$L_S = L$ exactly when $n_0(S) = 0$. The algorithm $\mathcal{A}$ of Example 9.18 is 0-hypothesis-stable at S exactly when $n_0(S) = n_1(S) = 0$.*

Proposition 9.20 *Let L be a finite language. Let $p = \min_{x \in L} \Pr(x)$, where the probability is with respect to the distribution on Σ^*. Then*

$$\Pr(n_0(S) \neq 0) \leq |L|(1-p)^m. \tag{18}$$

$$\Pr(n_0(S) \neq 0 \text{ or } n_1(S) \neq 0) \leq |L| \left(\frac{1 + (m-1)p}{1-p} \right) (1-p)^m. \tag{19}$$

Proof: We begin with Inequality (18). The probability that a particular word x does not appear in S is $(1 - \Pr(x))^m \leq (1-p)^m$. So,

$$\Pr(n_0(S) \neq 0) \leq \sum_{x \in L} \exp(-\Pr(x)m) \leq |L|(1-p)^m.$$

Similarly, the probability that x occurs at most once in S is

$$\begin{aligned}
(1 - \Pr(x))^m + m \Pr(x)(1 - \Pr(x))^{m-1} &\leq (1-p)^m + mp(1-p)^{m-1} \\
&= (1-p)^m \left(1 + \frac{mp}{1-p} \right) \\
&= (1-p)^m \left(\frac{1 + (m-1)p}{1-p} \right).
\end{aligned}$$

Summing the above over all $x \in L$ yields Inequality (19). ■

Observation 9.19 and Proposition 9.20 combine to give us the following theorem:

Theorem 9.21 *If the target language L is finite, then the language-learning algorithm of Example 9.18 has strong hypothesis stability $(0, \delta)$, where*

$$\delta = (1-p)^{m - O(\log m)}$$

for $p = \min_{x \in L} \Pr(x)$.

Remark 9.22 Suppose we use Theorem 9.21 to draw conclusions about the behavior of the language-learning algorithm. By Part 3 of Lemma 2.16, an algorithm which is strongly $(0, \delta)$-hypothesis-stable outputs the same answer (in this case, the correct hypothesis L) with probability at least $1 - m\delta/2$. So,

$$\Pr(L_S \neq L) \leq \frac{m}{2}|L| \left(\frac{1 + (m-1)p}{1-p} \right) (1-p)^m = (1-p)^{m - O(\log m)}.$$

In contrast, if we simply use Observation 9.19 and Inequality (18) of Proposition 9.20, we get

$$\Pr(L_S \neq L) = \Pr(n_0(S) \neq 0) \leq |L|(1-p)^m = (1-p)^{m - O(1)}.$$

In fact, letting $x_0 = \arg\min_{x \in L} \Pr(x)$,

$$\Pr(L_S \neq L) \geq \Pr(x_0 \notin S) = (1-p)^m.$$

So the concentration bounds on finite language learning obtained through Lemma 2.16 are log-asymptotically equal to the best possible bounds.

If, however, we apply Theorem 6.12 directly, we get that, for any $\tau > 0$,

$$\Pr(|\operatorname{Err}_D(f_S) - \operatorname{Err}_S(f_S)| > \tau + \delta) \leq 2\left(\exp\left(-\tau^2 m/8\right) + 2m^2\delta\right).$$

For the language-learning algorithm, $\operatorname{Err}_S(f_S) = 0$, and $L_S \neq L$ implies that $\operatorname{Err}_D(f_S) \geq p$. So, assuming $\delta < p/2$, we take $\tau = p/2$, yielding

$$\Pr(L_S \neq L) \leq 2\left(\exp(-p^2 m/32) + 2m^2\delta\right).$$

Here, the $\exp(-p^2 m/32)$ term dominates. We get an exponential concentration bound, but it is not log-asymptotically optimal.

9.5.1 The noisy setting

We now consider finite language learning in the noisy setting: each instance x is labeled incorrectly with probability $\eta < 1/2$ and correctly with probability $1-\eta$. It seems reasonable to ask whether the language-learning algorithm is stable in this context. In this section, we argue that, in the noisy setting, the algorithm of Example 9.18 is not strongly (β, δ)-hypothesis-stable for any $\beta < 1$ and $\delta = \exp(-\Omega(m))$.

The first question is how the algorithm should be modified to deal with the possibility of incorrect labels. One option is: for each x which occurs somewhere in the training set, take a majority vote of the labels associated with x.

The problem is that, assuming the distribution on Σ^* has infinite support, there must be some $x \notin L$ which does not appear in S. Changing some element of S to $(x, 1)$ then changes the output of f_S on x.

Remark 9.23 The above argument may feel like cheating: after all, we do not allow changing an element to $(x, 1)$ in the noiseless case. However, assuming the distribution on Σ^* has infinite support, it is likely (for sufficiently large m) that a given training set S contains some instance $x \notin L$ exactly once. If this happens, there is a probability of at least η that such an x will be incorrectly labeled by f_S.

There are two natural ways to modify the language-learning algorithm for the noisy setting:

1. If x occurs at least $g(m)$ times in S, and $(x, 1)$ occurs more often than $(x, 0)$, then $f_S(x) = 1$. Otherwise, $f_S(x) = 0$.

2. If $(x, 1)$ occurs more often than $(x, 0)$ in S, and the margin of victory is at least $g(m)$, then $f_S(x) = 1$. Otherwise, $f_S(x) = 0$.

For either of these two definitions, we have the same problem: if $g(m) = o(m)$, then it is likely that some $x \notin L$ will show up between $g(m)$ and $2g(m)$ times in S. The probability that x is mislabeled each time is at least $\eta^{2g(m)}$. So, the probability that the algorithm misclassifies some $x \notin L$ is $\exp(-o(m))$.[3]

If we take $g(m) = \Theta(m)$, then the above problem does not occur; the algorithm is stable. However, if $g(m) = \alpha m$ and some $x \in L$ occurs with probability less than α, then that x will be misclassified with overwhelming probability. So, if we try too hard to stabilize the algorithm, it no longer learns the language.

10 Open questions

10.1 The "right" definition of stability

Question 10.1 We argue in Section 4.2 that strong error stability, or even uniform L_1 stability, does not imply good bounds on generalization error. Theorem 1.1 states that training stability does imply good bounds on generalization error. We can get even better bounds with hypothesis stability, but, as we show in Section 9, hypothesis stability is too restrictive to be the "right" definition of stability.

Is training stability the best notion of stability? Is there some other notion of stability which yields similar bounds? (This could be something weaker than training stability, or it could be something stronger which gives better bounds and is more widely applicable than weak hypothesis stability.)

Vapnik [Vap01] suggests *functional stability*, in which we bound the difference $|\phi(f_S) - \phi(f_{S'})|$ for a set of functionals ϕ. Error stability is (roughly) the case where we use a single function $\phi = \mathrm{Err}_S$.)

Question 10.2 It may be that a version of Theorem 2.13 or Theorem 2.14 could be proved which uses more complete information about the quantity $|X(\omega) - X(\omega')|$, not just a bound which holds with high probability. If this could be achieved, the stability of an algorithm would depend not on two parameters β, δ, but on a single term, such as $\mathbf{E}_{S,i,u}(\max_z\{|c(f_S, z) - c(f_{S^{i,u}}, z)|\})$. We may even be able to describe a counterpart to the VC dimension—a quantity d where we get good convergence bounds if d is finite, and not if d is infinite. In this case, we can hope that any algorithm with finite VC dimension will be finite in this new sense as well; we might then be able to use stability to give an alternate explanation of the power of learners with finite VC dimension.

Question 10.3 Theorem 6.13 gives good bounds on generalization error for weakly (β, δ)-hypothesis-stable algorithms when $\beta = O(1/m)$ and $\delta = \exp(-\Omega(m))$. In Section 9.4, we see that ERM over classes of hyperplanes is weakly $(0, \delta)$ hypothesis stable with $\delta = \Theta(1/m)$.

Can we use stability to get good bounds on generalization error when $\delta = \Theta(1/m)$? This might require a stronger version of Theorem 2.14, or it might require a new definition

[3] And, since the probability of such an event is concentrated on the margin of victory being as small as possible, most training sets where x is misclassified will be unstable.

of stability, in between weak and strong hypothesis stability, which would be satisfied by hyperplanes and which would give good generalization error bounds.

Question 10.4 What is the right definition of stability for probabilistic algorithms? In particular, it would nice to determine if a randomized bootstrapping algorithm (e.g., bagging [Bre96a]) preserves (or even enhances) stability. It seems difficult to adapt our notions of strong (β, δ) stability to the probabilistic case, but the weak (β, δ) stability definitions (including CV stability) carry over naturally.

Question 10.5 Boosting algorithms work by reducing training error at the expense of generalization error. If the weak learner has poor performance on the training set but generalizes well, then the ensemble strikes a better balance between training error and generalization error.

Can we, instead, boost stability? We would start with an unstable weak learner that does well on its training set but generalizes poorly. By designing a stable ensemble, we might be able to construct a learner which generalizes well without sacrificing too much training error. As with the usual boosting algorithms, we would strike a better balance, and have a lower true error rate.

10.2 Leave-one-out error estimates

The goal of an analysis of learning algorithms is to bound the true error rate $\mathrm{Err}_D(f_S)$. We have focused on the generalization error $|\,\mathrm{Err}_D(f_S) - \mathrm{Err}_S(f_S)$.

Others have focused on *leave-one-out estimates* of the true error rate: if $S = (z_1, \ldots, z_m)$, the leave-one-out estimate of the error rate of f_S is

$$\frac{1}{m} \sum_{i=1}^{m} c(f_{S^i}, z_i),$$

where S^i is the set S with z_i removed. (These are also called *cross-validation error estimates*.) Several authors [DW79, KR99, BE02, EPE01] have observed connections between algorithmic stability and the accuracy of leave-one-out error estimates.

What can we prove about leave-one-out error using almost-everywhere stability? In particular:

Question 10.6 Bousquet and Elisseeff [BE02] use uniform hypothesis stability to get bounds on the error of leave-one-out error estimates. Can we use our extended versions of McDiarmid's inequality (see Section 2.2) to derive similar bounds from an assumption of training stability, or of weak or strong hypothesis stability?

Question 10.7 Kearns and Ron [KR99] show that, in the case of finite VC dimension, weak error stability gives good bounds on the error of leave-one-out error estimates. Can we remove this VC theoretic assumption? Does it help to assume some form of L_1 stability?

10.3 Stability, VC dimension, and examples

Question 10.8 We show (Theorem 7.2) that, in the PAC setting, CV stability gives necessary and sufficient conditions for good generalization error. Can we prove a similar statement in a more general setting?

Question 10.9 By Corollary 7.4, distribution-free CV stability is equivalent to finite VC dimension in the PAC setting. Our proof relies on the VC theoretic argument that finite VC dimension is equivalent to PAC learnability. Can we give a direct proof of Corollary 7.4?

In one direction, this seems straightforward: if $\mathrm{VCdim}(\mathcal{H})$ is infinite, then $\mathcal{H}$ shatters large sets of points. By putting all of the weight of our distribution on such a set, we prevent the algorithm from being CV stable.

Can we formalize this argument without reproving the necessity of finite VC dimension for PAC learnability? Can we give a direct proof of the converse, that ERM over a class of finite VC dimension is CV stable with respect to any distribution?

Also: can we extend Corollary 7.4 beyond the PAC setting, to $\{-1, 1\}$-classifiers in general?

Question 10.10 Finite VC dimension is necessary and sufficient for distribution-free bounds on generalization error. More generally, annealed entropy (see, for example, Vapnik [Vap98]) can be used to obtain sufficient conditions for distribution-dependent fast uniform convergence.

What is the relationship between training stability and annealed entropy? Is ERM over a space with low annealed entropy (i.e., $\lim_{m\to\infty} \mathcal{H}^{\mathcal{H}}_{\mathrm{ann}}(m)/m = 0$) necessarily training stable? Can we prove a direct connection between training stability and annealed entropy, without connecting each to convergence bounds?

We believe that some version of Part 3 of Theorem 9.2 also applies in the setting of Section 9.3. That is,

Conjecture 10.11 *If there are two or more classifiers in a finite space $\mathcal{H}$ of equal accuracy, then there exists a distribution D for which ERM is not $(0, \delta)$-training-stable for any $\delta = o(m^{-1/2})$.*

Question 10.12 Can we use stability to get distribution-free generalization error bounds in the case where $|\mathcal{H}| = 2$? (See Remark 9.3 for one possible approach.) Can we extend such an argument to the setting of Section 9.3, where $|\mathcal{H}| < \infty$?

Question 10.13 In Section 9.5, we show that learning a finite language is strongly hypothesis stable in the noiseless setting. In Section 9.5.1, we show that a straightforward extension of the noiseless algorithm must either be unstable or fail to learn the language.

Is there another approach to learning a finite language which yields a good, stable learner in the noiseless setting? Which sense of stability should we use?

Question 10.14 In Section 9.4, we show that ERM over a class of finite VC dimension is not necessarily weakly hypothesis stable. We conjecture (see Conjecture 9.12) that, for any class of finite VC dimension, there exists an algorithm performing ERM which is weakly $(0, O(d/m))$-hypothesis-stable.

We show that ERM over a class of finite VC dimension is strongly error stable (see Theorem 8.4).

Is empirical risk minimization over a class of finite VC dimension necessarily weakly L_1 stable? Strongly L_1 stable?

(For a discussion of VC dimension and CV stability, see Question 10.9.)

Question 10.15 Can we prove that ERM over a class of finite VC dimension is weakly hypothesis stable under some additional continuity assumptions? Can we give a margin-based argument for weak hypothesis stability, using V_γ dimension [ABDCBH93]? Rifkin, et al. [RMP] show that, under certain conditions, strong hypothesis stability is equivalent to finite V_γ dimension.

Question 10.16 Under what conditions, and in what senses, is Occam learning [BEHW87] stable?

References

[ABDCBH93] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proc. of the 34th IEEE FOCS*, pages 292–301, 1993.

[BE01] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems 13: Proc. NIPS'2000*, 2001.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2002. To appear.

[BEHW87] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

[Bre96a] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bre96b] L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996.

[CS02] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49, 2002.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Number 31 in Applications of Mathematics. Springer, 1996.

[DW79] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory*, 25(5):601–604, 1979.

[EPE01] T. Evgeniou, M. Pontil, and A. Elisseeff. Leave one out error, stability, and generalization of voting combination of classifiers. Technical Report 2001-21-TM, INSEAD, 2001.

[EPP00] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[FMS01] Y. Freund, Y. Mansour, and R. Schapire. Why averaging classifiers can protect against overfitting. In *Workshop on Artificial Intelligence and Statistics*, 2001.

[FS97] Y. Freund and R. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[GBD92] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[KN01] S. Kutin and P. Niyogi. The interaction of stability and weakness in AdaBoost. Technical Report TR-2001-30, Department of Computer Science, The University of Chicago, 2001.

[KR99] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427–1453, 1999.

[Kut02] S. Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical report, Department of Computer Science, The University of Chicago, 2002. In preparation.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[McD98] C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, Berlin, 1998.

[NG99] P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics*, 10(1):51–80, 1999.

[Niy98] P. Niyogi. *The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar.* Kluwer Academic Publishers, 1998.

[PRM02] T. Poggio, R. Rifkin, and S. Mukherjee. Bagging regularizes. Technical Report 214/AI Memo #2002-003, MIT CBCL, 2002.

[RMP] R. Rifkin, S. Mukherjee, and T. Poggio. Stability, uniform convergence, and regularization. In preparation.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.

[Vap95] V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.

[Vap98] V. Vapnik. *Statistical learning theory*. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[Vap01] V. Vapnik. Personal communication, January 2001.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.