Analysis and Applications, Vol. 3, No. 4 (2005) 397–417

World Scientific
www.worldscientific.com

STABILITY RESULTS IN LEARNING THEORY

ALEXANDER RAKHLIN
Center for Biological and Computational Learning
Massachusetts Institute of Technology
45 Carleton St. E25-201, Cambridge, MA 02139, USA
rakhlin@mit.edu

SAYAN MUKHERJEE
Institute of Genome Sciences and Policy
Institute for Statistics and Decision Sciences
Duke University, Durham, NC 27708, USA
sayan@stat.duke.edu

TOMASO POGGIO
Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
tp@ai.mit.edu

Received 2 July 2005
Revised 22 August 2005

The problem of proving generalization bounds for the performance of learning algorithms can be formulated as a problem of bounding the bias and variance of estimators of the expected error. We show how various *stability assumptions* can be employed for this purpose. We provide a necessary and sufficient stability condition for bounding the bias and variance for the Empirical Risk Minimization algorithm, and various sufficient conditions for bounding bias and variance of estimators for general algorithms. We discuss settings in which it is possible to obtain exponential bounds, and we prove an extension of the bounded-difference inequality for "almost always" stable algorithms.

Keywords: Stability; generalization; estimators; empirical risk minimization.

Mathematics Subject Classification 2000: 22E46, 53C35, 57S20

1. Introduction

One of the central problems of Statistical Learning Theory is to quantify the generalization ability of learning algorithms within a probabilistic framework. The standard setting for the problem is the following. Let $\mathcal{F}$ be a class of real-valued functions on a space $\mathcal{X}$, mapping $\mathcal{X}$ into $\mathcal{Y} \subset \mathbb{R}$. Denote by μ an unknown probability measure on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. An algorithm $\mathcal{A}$ is a mapping $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{F}$, $n \in \mathbb{Z}^+$. In plain words, a learning algorithm observes n input-output pairs and produces (learns) a function

which describes well the underlying input-output process. Throughout this article, we will focus on *symmetric* algorithms, i.e. $\mathcal{A}(z_1, \ldots, z_n) = \mathcal{A}(\pi(z_1, \ldots, z_n))$ for any permutation $\pi \in S_n$, the symmetric group.

Let $\mathcal{A}(z_1, \ldots, z_n; x)$ denote the evaluation of the function $\mathcal{A}(z_1, \ldots, z_n)$ at a point x. To measure the quality of $\mathcal{A}(z_1, \ldots, z_n)$, a loss function $\ell : \mathcal{F} \times \mathcal{Z} \mapsto [0, M]$ is introduced, such that $\ell(\mathcal{A}(z_1, \ldots, z_n); z)$ is a measure of how well $\mathcal{A}(z_1, \ldots, z_n)$ predicts y at point x, where $(x, y) = z \in \mathcal{Z}$. The function ℓ is often taken to be the square loss.

If the algorithm $\mathcal{A}$ is clear from the context, we will write $\ell(z_1, \ldots, z_n; z)$ instead of $\ell(\mathcal{A}(z_1, \ldots, z_n); z)$. The functions $\ell(f; \cdot)$ are called the *loss functions* and the class $\mathcal{L}(\mathcal{F}) = \{\ell(f; \cdot) : f \in \mathcal{F}\}$ is called the *loss class*.

The main quantity of interest is the *expected error* of the function $\mathcal{A}(z_1, \ldots, z_n)$,

$$I_{\exp}(z_1, \ldots, z_n) := \mathbb{E}_{z \sim \mu}[\ell(z_1, \ldots, z_n; z)] = \int_{\mathcal{Z}} \ell(z_1, \ldots, z_n; z)\, d\mu(z).$$

This quantity measures the accuracy of $\mathcal{A}(z_1, \ldots, z_n)$ on the unseen data z drawn from μ. Unfortunately, the measure μ is unknown and this quantity cannot be computed. The key assumption made in the Statistical Learning Theory is that the observed sample $z_1, \ldots, z_n$ is independent and identically distributed (i.i.d.) with the generating distribution μ. The problem thus is to estimate $I_{\exp}(z_1, \ldots, z_n)$ based on the finite sample $z_1, \ldots, z_n$.

Although the expected error is unknown, several important quantities *can* be computed from the sample. The first one is the *empirical error* (or *resubstitution estimate*),

$$I_{\text{emp}}(z_1, \ldots, z_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(z_1, \ldots, z_n; z_i).$$

The second one is the *leave-one-out error* (or *deleted estimate*),[a]

$$I_{\text{loo}}(z_1, \ldots, z_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n; z_i).$$

These quantities are employed to estimate the expected error, and the Statistical Learning Theory is concerned with providing bounds on the deviations of these estimates from the expected error. Denote these deviations

$$\Psi_n(z_1, \ldots, z_n) := I_{\exp}(z_1, \ldots, z_n) - I_{\text{emp}}(z_1, \ldots, z_n),$$
$$\Phi_n(z_1, \ldots, z_n) := I_{\exp}(z_1, \ldots, z_n) - I_{\text{loo}}(z_1, \ldots, z_n).$$

If one can show that Ψ_n (or Φ_n) is "small", then the empirical error (resp. leave-one-out error) is a good proxy for the expected error. In particular, we are interested in

[a]It is understood that the first term in the sum is $\ell(z_2, \ldots, z_n; z_1)$ and the last term is $\ell(z_1, \ldots, z_{n-1}; z_n)$.

the rate of the convergence of Ψ_n and Φ_n to zero as n increases. Such statements, of course, have to be made in probability.

Let us first focus on the random variable $\Psi_n(z_1, \ldots, z_n)$. Recall that the Central Limit Theorem (CLT) guarantees that the average of n i.i.d. random variables converges to their mean (under the assumption of finiteness of second moment). Unfortunately, the random variables $\ell(z_1, \ldots, z_n; z_1), \ldots, \ell(z_1, \ldots, z_n; z_n)$ are dependent, and the CLT is not applicable. In fact, the interdependence of these random variables makes the resubstitution estimate *positively biased*, as the next example shows.

Example 1.1. Let $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, $\mu(x) = U[0,1]$, $\mu(y|x) = \delta_{y=1}$, $\ell(y, y') = |y - y'|$, and $\mathcal{A}$ is defined as $\mathcal{A}(z_1, \ldots, z_n; x) = 1$ if $x \in \{x_1, \ldots, x_n\}$ and 0 otherwise. In other words, the algorithm observes n data points $(x, 1)$, where x is distributed uniformly on $[0,1]$, and generates a hypothesis which fits exactly the observed data, but outputs 0 for unseen points x. The empirical error of $\mathcal{A}$ is 0, while the expected error is 1, i.e. $\Psi_n(z_1, \ldots, z_n) = 1$ for any $z_1, \ldots, z_n$.

The algorithm in Example 1.1 is the *Empirical Risk Minimization* (*ERM*) algorithm

$$\mathcal{A}(z_1, \ldots, z_n) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f; z_i).$$

In the above example, the function class $\mathcal{F} = \bigcup_{n \geq 1} \{f_{\mathbf{x}} : \mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n\}$ where $f_{\mathbf{x}}(x) = 1$ if $x = x_i$ for some $1 \leq i \leq n$ and $f_{\mathbf{x}}(x) = 0$ otherwise.

Though an exact minimizer of empirical risk might not exist, an almost-minimizer always exists. The results of this paper hold for almost-minimizers, but, for the sake of clarity, we consider exact minimization.

Minimizing the empirical error is a natural idea, as long as guarantees on smallness of $\Psi_n(z_1, \ldots, z_n)$ can be made. Note that no such guarantee can be made in Example 1.1. Intuitively, this is due to the fact that the algorithm can fit *any* data, i.e. the space of functions $\mathcal{L}(\mathcal{F})$ is too large. Indeed, convergence of empirical errors to the expected errors is completely characterized by the "size" of $\mathcal{L}(\mathcal{F})$. Such a characterization disregards the algorithm $\mathcal{A}$, and only focuses on the loss function ℓ and the class $\mathcal{F}$, from which the functions are chosen. The class $\mathcal{L}(\mathcal{F})$ is called *uniform Glivenko–Cantelli* if for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup_{\mu} \mathbb{P}\left(\sup_{\ell \in \mathcal{L}(\mathcal{F})} \left| \mathbb{E}\ell - \frac{1}{n} \sum_{i=1}^{n} \ell(z_i) \right| \geq \varepsilon \right) = 0,$$

where $z_1, \ldots, z_n$ are i.i.d. random variables distributed according to μ.

Non-asymptotic results of the form

$$\mathbb{P}\left(\sup_{\ell \in \mathcal{L}(\mathcal{F})} \left| \mathbb{E}\ell - \frac{1}{n} \sum_{i=1}^{n} \ell(z_i) \right| \geq \varepsilon \right) \leq \delta(\varepsilon, n, \mathcal{L}(\mathcal{F}))$$

give *uniform* (over class $\mathcal{L}(\mathcal{F})$) rates of convergence of empirical means to the expected means. Since the guarantee is given for *all* functions in the class, the

defect $\Psi_n(z_1, \ldots, z_n)$ is bounded by ε with probability $1 - \delta(\varepsilon, n, \mathcal{L}(\mathcal{F}))$, no matter what the algorithm is.

Albeit interesting from the theoretical point of view, the uniform bounds are in general loose, as they are "worst case" over all functions in the class. As an extreme example, consider the algorithm that always ouputs the same function (the constant algorithm)

$$\mathcal{A}(z_1, \ldots, z_n) = f_0, \quad \forall\, (z_1, \ldots, z_n) \in \mathcal{Z}^n.$$

The bound on $\Psi_n(z_1, \ldots, z_n)$ follows from the CLT and an analysis based upon the complexity of a class $\mathcal{F}$ does not make sense. Recent advances in the Statistical Learning Theory shift the focus from uniform bounds to non-uniform bounds of the form

$$\mathbb{P}(|\Psi_n(z_1, \ldots, z_n)| > \varepsilon) < \delta(\varepsilon, n, \mathcal{A}), \tag{1.1}$$

or

$$\mathbb{P}(|\Psi_n(z_1, \ldots, z_n)| > \varepsilon(\delta, n, \mathcal{A}, z_1, \ldots, z_n)) < \delta,$$

where in the last bound ε depends on the sample. In this article, we will focus on the bounds of type (1.1). **The goal is to derive bounds on Ψ_n (or Φ_n) such that $\lim_{n\to\infty} \delta(\varepsilon, n, \mathcal{A}) = 0$ for any fixed $\varepsilon > 0$.** If the rate of decrease of $\delta(\varepsilon, n, \mathcal{A})$ is not important, we will write $|\Psi_n| \xrightarrow{P} 0$ and $|\Phi_n| \xrightarrow{P} 0$.

Notice that Ψ_n and Φ_n are bounded random variables, as the loss function $\ell \subset [0, M]$. By Markov's inequality,

$$\forall\, \varepsilon \geq 0, \quad \mathbb{P}(|\Psi_n| \geq \varepsilon) \leq \frac{\mathbb{E}|\Psi_n|}{\varepsilon}$$

and also,

$$\forall\, \varepsilon' \geq 0, \quad \mathbb{E}|\Psi_n| \leq M\mathbb{P}(|\Psi_n| \geq \varepsilon') + \varepsilon'.$$

Therefore, showing $|\Psi_n| \xrightarrow{P} 0$ is *equivalent* to showing $\mathbb{E}|\Psi_n| \to 0$. The latter is equivalent to $\mathbb{E}\Psi_n^2 \to 0$ since $|\Psi_n| \leq M$. Further, notice that $\mathbb{E}\Psi_n^2 = \mathrm{var}(\Psi_n) + (\mathbb{E}\Psi_n)^2$. We will call $\mathbb{E}\Psi_n$ the *bias*, $\mathrm{var}(\Psi_n)$ the *variance*, and $\mathbb{E}\Psi_n^2$ the *second moment* of Ψ_n. The same derivations and terminology hold for Φ_n.

We have shown that studying conditions for convergence in probability of the estimators to zero is equivalent to studying their mean and variance (*or the second moment alone*).

In this paper, we consider various *stability* conditions which allow one to bound bias and variance or the second moment, and thus imply convergence of Ψ_n and Φ_n to zero in probability. Though the reader should expect a number of definitions of stability, the common flavor of these notions is the comparison of the "behavior" of the algorithm $\mathcal{A}$ on similar samples. We hope that the present work sheds light on the important stability aspects of algorithms, suggesting principles for designing predictive learning systems.

We now sketch the organization of this paper. In Sec. 2, we motivate the use of stability and give some historical background. In Sec. 3, we show how bias (Sec. 3.1) and variance (Sec. 3.2) can be bounded by various stability quantities. Sometimes it is mathematically more convenient to bound the second moment instead of bias and variance, and this is done in Sec. 4. In particular, Sec. 4.1 deals with the second moment $\mathbb{E}\Phi_n^2$ in the spirit of [4], while in Secs. 4.3 and 4.2, we bound $\mathbb{E}\Psi_n^2$ in the spirit of [10] and [2], respectively. The goal of Secs. 4.1 and 4.2 is to re-derive some known results in a simple manner that allows one to compare the proofs side by side. The results of these sections hold for general algorithms. Furthermore, for specific algorithms the results can be improved, i.e. simpler quantities might govern the convergence of the estimators to zero. To illustrate this, in Sec. 4.4 we prove that for the Empirical Risk Minimization algorithm, a bound on the bias $\mathbb{E}\Psi_n$ implies a bound on the second moment $\mathbb{E}\Psi_n^2$. We therefore provide a simple necessary and sufficient condition for consistency of ERM. If rates of convergence are of importance, rather than using Markov's inequality, one can make use of more sophisticated concentration inequalities with a cost of requiring more stringent stability conditions. In Sec. 5, we discuss the most rigid stability, Uniform Stability, and provide exponential bounds in the spirit of [2]. In Sec. 5.2, we consider less rigid notions of stability and prove exponential inequalities based on powerful moment inequalities of [1]. Finally, Sec. 6 summarizes the paper and discusses further directions and open questions.

2. Historical Remarks and Motivation

Devroye, Rogers and Wagner (see, e.g., [4]) were the first, to our knowledge, to observe that sensitivity of the algorithms with regard to small changes in the sample is related to the behavior of the leave-one-out estimate. The authors were able to obtain results for the k-Nearest-Neighbor algorithm, where VC theory fails because of large class of potential hypotheses. These results were further extended for k-local algorithms and for potential learning rules. Kearns and Ron [6] later discovered a connection between finite VC-dimension and stability. Bousquet and Elisseeff [2] showed that a large class of learning algorithms, based on *Tikhonov Regularization*, is stable in a very strong sense, which allowed the authors to obtain exponential bounds without much work. Kutin and Niyogi [8] introduced a number of notions of stability and showed implications between them. The authors emphasized the importance of "almost-everywhere" stability and proved valuable extensions of McDiarmid's exponential inequality [7]. Mukherjee *et al.* [10] proved that a combination of three stability notions is sufficient to bound the difference between the empirical estimate and the expected error, while for Empirical Risk Minimization these notions are necessary and sufficient. The latter result showed an alternative to VC theory condition for consistency of Empirical Risk Minimization. In this paper, we prove, in a unified framework, some of the important results mentioned above, as well as show new ways of incorporating stability notions in the Learning Theory.

We now give some intuition for using algorithmic stability. First, note that without any assumptions on the algorithm, nothing can be said about the mean and the variance of Ψ_n. One can easily come up with settings when the mean is converging to zero, but not the variance, or vice versa (e.g., Example 1.1), or both quantities diverge from zero.

The assumptions of this paper that allow us to bound the mean and the variance of Ψ_n and Φ_n are loosely termed as *stability assumptions*. Recall that if the algorithm is a constant algorithm, Ψ_n is bounded by the Central Limit Theorem. Of course, this is an extreme and the most "stable" case. It turns out that the "constancy" assumption on the algorithm can be relaxed while still achieving tight bounds. A central notion here is that of *Uniform Stability* [2]:

Definition 2.1. Uniform Stability $\beta_\infty(n)$ of an algorithm $\mathcal{A}$ is

$$\beta_\infty(n) := \sup_{z_1,\ldots,z_n,z\in\mathcal{Z},x\in\mathcal{X}} |\mathcal{A}(z_1,\ldots,z_n;x) - \mathcal{A}(z,z_2,\ldots,z_n;x)|.$$

Intuitively, if $\beta_\infty(n) \to 0$, the algorithm resembles more and more the constant algorithm when considered on similar samples (although it can produce distant functions on different samples). It can be shown that some well-known algorithms possess Uniform Stability with a certain rate on $\beta_\infty(n)$ (see [2] and Sec. 5.1).

In the following sections, we will show how the bias and variance (or second moment) can be upper-bounded or decomposed in terms of quantities over "similar" samples. The advantage of this approach is that it allows one to check "stability" for a specific algorithm and derive generalization bounds without much further work. For instance, it is easy to show that k-Nearest Neighbors algorithm is L_1-stable and a generalization bound follows immediately (see Sec. 4.1).

3. Bounding Bias and Variance

3.1. *Decomposing the bias*

The bias of the resubstitution estimate and the deleted estimate can be written as quantities over similar samples:

$$\begin{aligned}\mathbb{E}\Psi_n &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_z\ell(z_1,\ldots,z_n;z) - \ell(z_1,\ldots,z_n;z_i)\right]\right]\\ &= \mathbb{E}[\ell(z_1,\ldots,z_n;z) - \ell(z_1,\ldots,z_n;z_1)]\\ &= \mathbb{E}[\ell(z,z_2,\ldots,z_n;z_1) - \ell(z_1,\ldots,z_n;z_1)].\end{aligned}$$

The first equality above follows because $\mathbb{E}\ell(z_1,\ldots,z_n;z_k) = \mathbb{E}\ell(z_1,\ldots,z_n;z_m)$ for any k,m. The second equality holds by noticing that $\mathbb{E}\ell(z_1,\ldots,z_n;z) = \mathbb{E}\ell(z,z_2,\ldots,z_n;z_1)$ because the roles of z and z_1 can be switched. We will employ this trick many times in the later proofs, and for convenience, we shall denote this "renaming" process by $z \leftrightarrow z_1$.

Let us inspect the quantity $\mathbb{E}[\ell(z,z_2,\ldots,z_n;z_1) - \ell(z_1,\ldots,z_n;z_1)]$. It is the average difference between the loss at a point z_1 when it is not present in the

learning sample (out-of-sample) and the loss at z_1 when it is present in the n-tuple (in-sample). Hence, the bias $\mathbb{E}\Psi_n$ will decrease if and only if the average behavior on in-sample and out-of-sample points is becoming more and more similar. This is a stability property and we will give a name to it:

Definition 3.1. Average Stability $\beta_{\text{bias}}(n)$ of an algorithm $\mathcal{A}$ is

$$\beta_{\text{bias}}(n) := \mathbb{E}[\ell(z, z_2, \ldots, z_n; z_1) - \ell(z_1, \ldots, z_n; z_1)].$$

We now turn to the deleted estimate. The bias $\mathbb{E}\Phi_n$ can be written as

$$\begin{aligned}\mathbb{E}\Phi_n &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_z\ell(z_1, \ldots, z_n; z) - \ell(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n; z_i))\right] \\ &= \mathbb{E}[\ell(z_1, \ldots, z_n; z) - \ell(z_2, \ldots, z_n; z_1)] \\ &= \mathbb{E}[I_{\text{exp}}(z_1, \ldots, z_n) - I_{\text{exp}}(z_2, \ldots, z_n)].\end{aligned}$$

We will not give a name to this quantity, as it will not be used explicitly later. One can see that the bias of the deleted estimate should be small for reasonable algorithms. Unfortunately, the variance of the deleted estimate is large in general (see, e.g., [5, p. 415]). The opposite is believed to be true for the resubstitution estimate. We refer the reader to [5, Chaps. 23, 24 and 31] for more information. Surprisingly, we will show in Sec. 4.4 that for Empirical Risk Minimization algorithms, if one shows that the bias of the resubstitution estimate decreases, one also obtains that the variance decreases.

3.2. *Bounding the variance*

Having shown a decomposition of the bias of Ψ_n and Φ_n in terms of stability conditions, we now show a simple way to bound the variance in terms of quantities over "similar" samples.

Theorem 3.2 (Efron–Stein). *Let $\xi : \mathcal{Z}^n \mapsto \mathbb{R}$ be a measurable function of n variables and define $\Gamma = \xi(z_1, \ldots, z_n)$ and $\Gamma_i' = \xi(z_1, \ldots, z_i', \ldots, z_n)$, where $z_1, \ldots, z_n, z_1', \ldots, z_n'$ are i.i.d. random variables. Then*

$$\text{var}(\Gamma) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\big[(\Gamma - \Gamma_i')^2\big]. \tag{3.1}$$

A "removal" version of the above is the following:

Theorem 3.3 (Efron–Stein). *Let $\xi : \mathcal{Z}^n \mapsto \mathbb{R}$ be a measurable function of n variables and $\xi' : \mathcal{Z}^{n-1} \mapsto \mathbb{R}$ of $n-1$ variables. Define $\Gamma = \xi(z_1, \ldots, z_n)$ and $\Gamma_i = \xi'(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, where $z_1, \ldots, z_n$ are i.i.d. random variables. Then*

$$\text{var}(\Gamma) \leq \sum_{i=1}^{n}\mathbb{E}\big[(\Gamma - \Gamma_i)^2\big]. \tag{3.2}$$

The idea of the proofs of the above result is based on the fact that $\mathrm{var}(\Gamma) \leq \mathbb{E}(\Gamma - c)^2$ for any constant c, and so

$$\mathrm{var}_i(\Gamma) = \mathbb{E}_{z_i}(\Gamma - \mathbb{E}_{z_i}\Gamma)^2 \leq \mathbb{E}_{z_i}(\Gamma - \Gamma_i)^2.$$

Thus, we artificially introduce a quantity over a "similar" sample to upper-bound the variance. If the increments $\Gamma - \Gamma_i$ and $\Gamma - \Gamma_i'$ are small, the variance is small. When applied to the function $\Psi_n(z_1, \ldots, z_n)$, this translates exactly into controlling the behavior of $\mathcal{A}$ on similar samples:

$$\begin{aligned}\mathrm{var}(\Psi_n) &\leq n\mathbb{E}(\Psi_n(z_1, \ldots, z_n) - \Psi_n(z_2, \ldots, z_n))^2 \\ &\leq 2n\mathbb{E}(I_{\mathrm{exp}}(z_1, \ldots, z_n) - I_{\mathrm{exp}}(z_2, \ldots, z_n))^2 \\ &\quad + 2n\mathbb{E}(I_{\mathrm{emp}}(z_2, \ldots, z_n) - I_{\mathrm{emp}}(z_1, \ldots, z_n))^2.\end{aligned}$$

Here we used the fact that the algorithm is invariant under permutation of coordinates, and therefore all the terms in the sum of (3.2) are equal. This symmetry will be exploited to a great extent in the later sections. Note that similar results can be obtained using the "replacement" version of Efron–Stein's bound.

The meaning of the above bound is that if the mean *square* of the difference between expected errors of functions, learned from samples differing in one point, is decreasing faster than n^{-1}, and if the same holds for the empirical errors, then the variance of the resubstitution estimate is decreasing. Let us give names to the above quantities.

Definition 3.4. Empirical-Error (Removal) Stability of an algorithm $\mathcal{A}$ is

$$\beta^2_{\mathrm{emp}}(n) := \mathbb{E}|I_{\mathrm{emp}}(z_1, \ldots, z_n) - I_{\mathrm{emp}}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)|^2.$$

Definition 3.5. Expected-Error (Removal) Stability of an algorithm $\mathcal{A}$ is

$$\beta^2_{\mathrm{exp}}(n) := \mathbb{E}|I_{\mathrm{exp}}(z_1, \ldots, z_n) - I_{\mathrm{exp}}(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)|^2.$$

With the above definitions, the following theorem follows:

Theorem 3.6.

$$\mathrm{var}(\Psi_n) \leq 2n\big(\beta^2_{\mathrm{exp}}(n) + \beta^2_{\mathrm{emp}}(n)\big).$$

The following example shows that the ERM algorithm is always Empirical-Error Stable with $\beta_{\mathrm{emp}}(n) \leq M(n-1)^{-1}$. We deduce that $\Psi_n \xrightarrow{P} 0$ for ERM whenever $\beta_{\mathrm{exp}} = o(n^{-1/2})$. As we will show in Sec. 4.4, the decay of the Average Stability, $\beta_{\mathrm{bias}}(n) = o(1)$, is both necessary and sufficient for $\Psi_n \xrightarrow{P} 0$ for ERM.

Example 3.7. For an Empirical Risk Minimization algorithm, $\beta_{\text{emp}}(n) \leq \frac{M}{n-1}$:

$$\begin{aligned}
&I_{\text{emp}}(z_2, \ldots, z_n) - I_{\text{emp}}(z_1, \ldots, z_n) \\
&\quad \leq \frac{1}{n-1} \sum_{i=2}^{n} \ell(z_2, \ldots, z_n; z_i) - \frac{1}{n-1} \sum_{i=1}^{n} \ell(z_1, \ldots, z_n; z_i) + \frac{M}{n-1} \\
&\quad \leq \frac{1}{n-1} \sum_{i=2}^{n} \ell(z_2, \ldots, z_n; z_i) - \frac{1}{n-1} \sum_{i=2}^{n} \ell(z_1, \ldots, z_n; z_i) \\
&\qquad + \frac{M}{n-1} - \frac{1}{n-1} \ell(z_1, \ldots, z_n; z_1) \leq \frac{M}{n-1}
\end{aligned}$$

and the other direction is proved similarly.

We will show in the following sections that a direct study of the second moment leads to better bounds. For the bound on the variance in Theorem 3.6 to decrease, β_{exp} and β_{emp} have to be $o(n^{-1/2})$. With an additional assumption, we will be able to remove the factor n by upper-bounding the second moment and by exploiting the structure of the random variables Φ_n and Ψ_n.

4. Bounding the 2nd Moment

Instead of bounding the mean and variance of the estimators, we can bound the second moment. The reason for doing so is for mathematical convenience and is due to the following straightforward bounds on the second moment:

$$\begin{aligned}
\mathbb{E}\Psi_n^2 &= \mathbb{E}[\mathbb{E}_z \ell(z_1, \ldots, z_n; z)]^2 - \mathbb{E}\left[\mathbb{E}_z \ell(z_1, \ldots, z_n; z) \frac{1}{n} \sum_{i=1}^{n} \ell(z_1, \ldots, z_n; z_i)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_1, \ldots, z_n; z_i)\right]^2 - \mathbb{E}\left[\mathbb{E}_z \ell(z_1, \ldots, z_n; z) \frac{1}{n} \sum_{i=1}^{n} \ell(z_1, \ldots, z_n; z_i)\right] \\
&\leq \mathbb{E}[\mathbb{E}_z \ell(z_1, \ldots, z_n; z) \mathbb{E}_{z'} \ell(z_1, \ldots, z_n; z') - \mathbb{E}_z \ell(z_1, \ldots, z_n; z) \ell(z_1, \ldots, z_n; z_1)] \\
&\quad + \mathbb{E}[\ell(z_1, \ldots, z_n; z_1) \ell(z_1, \ldots, z_n; z_2) - \mathbb{E}_z \ell(z_1, \ldots, z_n; z) \ell(z_1, \ldots, z_n; z_1)] \\
&\quad + \frac{1}{n} \mathbb{E}\ell(z_1, \ldots, z_n; z_1)^2,
\end{aligned}$$

and the last term is bounded by $\frac{M^2}{n}$. Similarly,

$$\begin{aligned}
\mathbb{E}\Phi_n^2 &\leq \mathbb{E}[\mathbb{E}_z \ell(z_1, \ldots, z_n; z) \mathbb{E}_{z'} \ell(z_1, \ldots, z_n; z') - \mathbb{E}_z \ell(z_1, \ldots, z_n; z) \ell(z_2, \ldots, z_n; z_1)] \\
&\quad + \mathbb{E}[\ell(z_2, \ldots, z_n; z_1) \ell(z_1, z_3, \ldots, z_n; z_2) - \mathbb{E}_z \ell(z_1, \ldots, z_n; z) \ell(z_2, \ldots, z_n; z_1)] \\
&\quad + \frac{1}{n} \mathbb{E}\ell(z_2, \ldots, z_n; z_1)^2,
\end{aligned}$$

and the last term is bounded by $\frac{M^2}{n}$.

In the proofs, we will use the following inequality for random variables X, X' and Y:

$$\mathbb{E}[XY - X'Y] \leq M\mathbb{E}|X - X'| \tag{4.1}$$

if $-M \leq Y \leq M$. The bounds on the second moments are already sums of terms of the type "$\mathbb{E}[XY - WZ]$", and we will find a way to use symmetry to change these terms into the type "$\mathbb{E}[XY - X'Y]$", where X and X' will be quantities over similar samples, and so $\mathbb{E}|X - X'|$ will be bounded by a certain stability of the algorithm.

4.1. *Leave-one-out (deleted) estimate*

We have seen that $\mathbb{E}\Phi_n = \mathbb{E}[I_{\exp}(z_1, \ldots, z_n) - I_{\exp}(z_2, \ldots, z_n)]$ and thus the bias decreases if and only if the expected errors are similar when learning on similar (one additional point) samples. Moreover, intuitively, these errors have to occur at the same places because otherwise evaluation of leave-one-out functions $\ell(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n; z)$ will not tell us about $\ell(z_1, \ldots, z_n; z)$. This implies that the L_1 distance between the functions on similar (one additional point) samples should be small. This connection between L_1 stability and the leave-one-out estimate has been observed by Devroye and Wagner [4] and further studied in [6]. We now define this stability notion:

Definition 4.1. L_1-Stability of an algorithm $\mathcal{A}$ is

$$\begin{aligned}\beta_1(n) &:= \|\ell(z_1, \ldots, z_n; \cdot) - \ell(z_2, \ldots, z_n; \cdot)\|_{L_1(\mu)} \\ &= \mathbb{E}_z|\ell(z_1, \ldots, z_n; z) - \ell(z_2, \ldots, z_n; z)|.\end{aligned}$$

The following theorem is proved in [4, 5] for classification algorithms. We give a similar proof for general learning algorithms. The result shows that the second moment (and therefore, both bias and variance) of the leave-one-out error estimate is bounded by the L_1 distance between loss functions on similar samples.

Theorem 4.2.

$$\mathbb{E}\Phi_n^2 \leq M(2\beta_1(n-1) + 4\beta_1(n)) + \frac{M^2}{n}.$$

Proof. The first term in the decomposition of the second moment of $\mathbb{E}\Phi_n^2$ can be bounded as follows:

$$\begin{aligned}&\mathbb{E}[\ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z') - \ell(z_1, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z_1)] \\ &\quad = \mathbb{E}[\ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z')] \\ &\quad = \mathbb{E}[\ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z') - \ell(z_2, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z')] \\ &\qquad + \mathbb{E}[\ell(z_2, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z')] \\ &\qquad + \mathbb{E}[\ell(z', z_2, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z')] \\ &\quad \leq 3M\beta_1(n).\end{aligned}$$

The first equality holds by renaming $z' \leftrightarrow z_1$. In doing this, we are using the fact that all the variables $z_1, \ldots, z_n, z, z'$ are identically distributed and independent. To obtain the inequality above, note that each of the three terms is bounded (using (4.1)) by $M\beta_1(n)$.

The second term in the decomposition is bounded similarly:

$$\begin{aligned}
&\mathbb{E}[\ell(z_2, \ldots, z_n; z_1)\ell(z_1, z_3, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z_1)] \\
&\quad = \mathbb{E}[\ell(z', z_3, \ldots, z_n; z)\ell(z, z_3, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z')] \\
&\quad = \mathbb{E}[\ell(z', z_3, \ldots, z_n; z)\ell(z, z_3, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z, z_3, \ldots, z_n; z')] \\
&\qquad + \mathbb{E}[\ell(z', z_2, \ldots, z_n; z)\ell(z, z_3, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_3, \ldots, z_n; z')] \\
&\qquad + \mathbb{E}[\ell(z', z_2, \ldots, z_n; z)\ell(z_3, \ldots, z_n; z') - \ell(z', z_2, \ldots, z_n; z)\ell(z_2, \ldots, z_n; z')] \\
&\quad \leq M\beta_1(n) + 2M\beta_1(n-1).
\end{aligned}$$

The first equality follows by renaming $z_2 \leftrightarrow z'$ as well as $z_1 \leftrightarrow z$ in the first term, and $z_1 \leftrightarrow z'$ in the second term. Finally, we bound the last term by M^2/n to obtain the result. □

4.2. *Empirical error (resubstitution) estimate: replacement case*

Recall that the bias of the resubstitution estimate is the Average Stability, $\mathbb{E}\Psi_n = \beta_{\text{bias}}$. However, this is not enough to bound the second moment $\mathbb{E}\Psi_n^2$ for general algorithms. Nevertheless, β_{bias} measures the average performance of in-sample and out-of-sample errors and this is inherently linked to the closeness of the resubstitution (in-sample) estimate and the expected error (out-of-sample performance). It turns out that it is possible to derive bounds on $\mathbb{E}\Psi_n^2$ by using a stronger version of the Average Stability. The natural strengthening is requiring that not only the first, but also the second moment of $\ell(z_1, \ldots, z_n; z_i) - \ell(z_1, \ldots, z_i', \ldots, z_n; z_i)$ is decaying to 0. We follow [8] in calling this type of stability *Cross-Validation* (CV) *Stability*:

Definition 4.3. CV (Replacement) Stability of an algorithm $\mathcal{A}$ is

$$\beta_{\text{cvr}} := \mathbb{E}|\ell(z_1, \ldots, z_n; z_1) - \ell(z, z_2, \ldots, z_n; z_1)|,$$

where the expectation is over a draw of $n+1$ points.

The following theorem was proven in [2]. Here, we give a version of the proof.

Theorem 4.4.

$$\mathbb{E}\Psi_n^2 \leq 6M\beta_{\text{cvr}}(n) + \frac{M^2}{n}.$$

Proof. The first term in the decomposition of $\mathbb{E}\Psi_n^2$ can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}_z\ell(z_1, \ldots, z_n; z)\mathbb{E}_{z'}\ell(z_1, \ldots, z_n; z') - \mathbb{E}_z\ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_2)] \\
&\quad = \mathbb{E}[\ell(z_1, z', z_3, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_2)] \\
&\quad = \mathbb{E}[\ell(z_1, z', z_3, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2)
\end{aligned}$$

$$
\begin{aligned}
&- \ell(z_1, z, z_3, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2)] \\
&+ \mathbb{E}[\ell(z_1, z, z_3, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2) \\
&- \ell(z_1, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2)] \\
&+ \mathbb{E}[\ell(z_1, \ldots, z_n; z)\ell(z_1, z', z_3, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_2)] \\
&\leq 3M\beta_{\mathrm{cvr}}(n).
\end{aligned}
$$

The first equality follows from renaming $z_2 \leftrightarrow z'$ in the first term. Each of the three terms in the sum above is bounded by $M\beta_{\mathrm{cvr}}(n)$.

The second term in the decomposition of $\mathbb{E}\Psi_n^2$ can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\ell(z_1, \ldots, z_n; z_1)\ell(z_1, \ldots, z_n; z_2) - \mathbb{E}_z\ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_1)\right] \\
&\quad = \mathbb{E}\left[\ell(z, z_2, \ldots, z_n; z)\ell(z, z_2, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_2)\right] \\
&\quad = \mathbb{E}\left[\ell(z, z_2, \ldots, z_n; z)\ell(z, z_2, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z, z_2, \ldots, z_n; z_2)\right] \\
&\qquad + \mathbb{E}\left[\ell(z_1, \ldots, z_n; z)\ell(z, z_2, \ldots, z_n; z_2)\right. \\
&\qquad \left. - \ell(z_1, \ldots, z_n; z)\ell(z_1, z, z_3, \ldots, z_n; z_2)\right] \\
&\qquad + \mathbb{E}\left[\ell(z_1, \ldots, z_n; z)\ell(z_1, z, z_3, \ldots, z_n; z_2) - \ell(z_1, \ldots, z_n; z)\ell(z_1, \ldots, z_n; z_2)\right] \\
&\quad \leq 3M\beta_{\mathrm{cvr}}(n).
\end{aligned}
$$

The first equality follows by renaming $z_1 \leftrightarrow z$ in the first term. Again, each of the three terms in the sum above can be bounded by $M\beta_{\mathrm{cvr}}(n)$. □

4.3. *Empirical error (resubstitution) estimate*

Mukherjee *et al.* [10] considered the "removal" version of the CV stability defined in Sec. 4.3, the motivation being that the addition of a new point z' complicates the cross-validation nature of the stability. Another motivation is the fact that $\ell(z_1, \ldots, z_n; z_1) - \ell(z_2, \ldots, z_n; z_1)$ is non-negative for Empirical Risk Minimization. It turns out that this "removal" version of the CV stability together with Expected and Empirical Stabilities upper-bound $\mathbb{E}\Psi_n$. Following [10], we have the following definition:

Definition 4.5. CV (Removal) Stability of an algorithm $\mathcal{A}$ is

$$\beta_{\mathrm{cv}}(n) := \mathbb{E}|\ell(z_1, \ldots, z_n; z_1) - \ell(z_2, \ldots, z_n; z_1)|.$$

The following theorem was proven in [10]. Here, we give a version of the proof.

Theorem 4.6.

$$\mathbb{E}\Psi_n^2 \leq M(\beta_{\mathrm{cv}}(n) + 4\beta_{\mathrm{exp}}(n) + 2\beta_{\mathrm{emp}}(n)) + \frac{M^2}{n}.$$

Proof. The first term in the decomposition of the second moment of $\mathbb{E}\Psi_n^2$ can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\ell(z_1,\ldots,z_n;z)\ell(z_1,\ldots,z_n;z') - \ell(z_1,\ldots,z_n;z)\ell(z_1,\ldots,z_n;z_1)\right] \\
&\quad = \mathbb{E}\left[\ell(z',z_2,\ldots,z_n;z)\ell(z',z_2,\ldots,z_n;z_1) - \ell(z_1,\ldots,z_n;z)\ell(z_1,\ldots,z_n;z_1)\right] \\
&\quad = \mathbb{E}\left[\ell(z',z_2,\ldots,z_n;z)\mathbb{E}_{z_1}\ell(z',z_2,\ldots,z_n;z_1)\right. \\
&\qquad \left. - \ell(z',z_2,\ldots,z_n;z)\mathbb{E}_{z_1}\ell(z_2,\ldots,z_n;z_1)\right] \\
&\qquad + \mathbb{E}\left[\mathbb{E}_z\ell(z',z_2,\ldots,z_n;z)\ell(z_2,\ldots,z_n;z_1) - \mathbb{E}_z\ell(z_2,\ldots,z_n;z)\ell(z_2,\ldots,z_n;z_1)\right] \\
&\qquad + \mathbb{E}\left[\mathbb{E}_z\ell(z_2,\ldots,z_n;z)\ell(z_2,\ldots,z_n;z_1) - \mathbb{E}_z\ell(z_1,\ldots,z_n;z)\ell(z_2,\ldots,z_n;z_1)\right] \\
&\qquad + \mathbb{E}\left[\ell(z_1,\ldots,z_n;z)\ell(z_2,\ldots,z_n;z_1) - \ell(z_1,\ldots,z_n;z)\ell(z_1,\ldots,z_n;z_1)\right] \\
&\quad \leq M(3\beta_{\exp}(n) + \beta_{\mathrm{cv}}(n)).
\end{aligned}$$

The first equality follows by renaming $z_1 \leftrightarrow z$ in the first term. In the sum above, the first three terms are each bounded by $M\beta_{\exp}(n)$, while the last one is bounded by $M\beta_{\mathrm{cv}}(n)$. Since the Expected (and Empirical) Error Stability has been defined in Sec. 3.2 as expectation of a square, we used the fact that $\mathbb{E}|X| \leq (\mathbb{E}X^2)^{1/2}$.

The second term in the decomposition of $\mathbb{E}\Psi_n^2$ is bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right)^2 - \mathbb{E}_z\ell(z_1,\ldots,z_n;z)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right] \\
&= \mathbb{E}\left[\ell(z_1,\ldots,z_n;z_1)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) - \ell(z_1,\ldots,z_n;z)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right] \\
&= \mathbb{E}\left[\ell(z_1,\ldots,z_n;z_1)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) - \ell(z_2,\ldots,z_n;z_1)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right] \\
&\quad + \mathbb{E}\left[\ell(z_2,\ldots,z_n;z_1)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) - \ell(z_2,\ldots,z_n;z)\frac{1}{n-1}\sum_{i=2}^n \ell(z_2,\ldots,z_n;z_i)\right] \\
&\quad + \mathbb{E}\left[\ell(z_2,\ldots,z_n;z)\frac{1}{n-1}\sum_{i=2}^n \ell(z_2,\ldots,z_n;z_i) - \ell(z_2,\ldots,z_n;z)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}_z\ell(z_2,\ldots,z_n;z)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) - \mathbb{E}_z\ell(z_1,\ldots,z_n;z)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)\right] \\
&\leq M(\beta_{\mathrm{cv}}(n) + 2\beta_{\mathrm{emp}}(n) + \beta_{\exp}(n)).
\end{aligned}$$

The first equality follows by symmetry:

$$\ell(z_1,\ldots,z_n;z_k)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) = \ell(z_1,\ldots,z_n;z_m)\frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i)$$

for all k, m. The first term in the sum above is bounded by $M\beta_{\mathrm{cv}}(n)$. The second term is bounded by $M\beta_{\mathrm{emp}}(n)$ (and $z_1 \leftrightarrow z$). The third term is also bounded by $M\beta_{\mathrm{emp}}(n)$, and the last term by $M\beta_{\exp}(n)$. □

4.4. *Resubstitution estimate for the Empirical Risk Minimization algorithm*

It turns out that for the ERM algorithm, Ψ_n is "almost positive". Intuitively, if one minimizes the empirical error, then the expected error is likely to be larger than the empirical estimate. Since Ψ_n is "almost positive", $\mathbb{E}\Psi_n \to 0$ implies $|\Psi_n| \xrightarrow{P} 0$. We now give a formal proof of this reasoning.

Recall, that an ERM algorithm searches in the function space $\mathcal{F}$. Let

$$f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_z \ell(f;z),$$

the minimizer of the expected error.[b] Consider the shifted loss class

$$\mathcal{L}'(\mathcal{F}) = \{\ell'(f;\cdot) = \ell(f;\cdot) - \ell(f^*;\cdot) | f \in \mathcal{F}\}$$

and note that $\mathbb{E}_z \ell'(f;z) \geq 0$ for any $f \in \mathcal{F}$. Trivially, if $\ell(z_1,\ldots,z_n;\cdot)$ is an empirical minimizer over the loss class $\mathcal{L}(\mathcal{F})$, then $\ell'(f;\cdot) = \ell(z_1,\ldots,z_n;\cdot) - \ell(f^*;\cdot)$ is an empirical minimizer over the shifted loss class $\mathcal{L}'(\mathcal{F})$

$$\begin{aligned}
&\mathbb{E}_z \ell'(z_1,\ldots,z_n;z) - \frac{1}{n}\sum_{i=1}^n \ell'(z_1,\ldots,z_n;z_i) \\
&\quad = \mathbb{E}_z \ell(z_1,\ldots,z_n;z) - \frac{1}{n}\sum_{i=1}^n \ell(z_1,\ldots,z_n;z_i) - \left(\mathbb{E}_z \ell(f^*;z) - \frac{1}{n}\sum_{i=1}^n \ell(f^*;z_i)\right).
\end{aligned}$$

Note that $\frac{1}{n}\sum_{i=1}^n \ell'(z_1,\ldots,z_n;z_i) \leq 0$ because $\mathcal{L}'(\mathcal{F})$ contains the zero function. Therefore, the left-hand side is non-negative and the second term on the right-hand side is small with high probability because f^* is non-random. We have

$$\mathbb{P}(\Psi_n(z_1,\ldots,z_n) < -\varepsilon) \leq \mathbb{P}\left(\mathbb{E}_z \ell(f^*;z) - \frac{1}{n}\sum_{i=1}^n \ell(f^*;z_i) < -\varepsilon\right) \leq e^{-2n\varepsilon^2/M^2}.$$

Therefore,

$$\mathbb{E}|\Psi_n| \leq \mathbb{E}\Psi_n + 2\varepsilon + 2Me^{-2n\varepsilon^2/M^2}.$$

If $\mathbb{E}\Psi_n \to 0$, the right-hand side can be made arbitrarily small for large enough n, thus proving $\mathbb{E}|\Psi_n| \to 0$. Clearly, $\mathbb{E}\Psi_n \to 0$ whenever $\mathbb{E}|\Psi_n| \to 0$. Hence, we have the following theorem:

Theorem 4.7. *For Empirical Risk Minimization,* $\beta_{\text{bias}}(n) \to 0$ *is equivalent to* $|\Psi_n| \xrightarrow{P} 0$.

Remark 4.8. With this approach, the rate of convergence of $I_{\text{emp}}(z_1,\ldots,z_n)$ to $I_{\text{exp}}(z_1,\ldots,z_n)$ is limited by the rate of convergence of $\frac{1}{n}\sum_{i=1}^n \ell(f^*;z_i)$ to $\mathbb{E}_z \ell(f^*;z)$, which is $O(n^{-1/2})$ without further assumptions.

[b]If the minimizer does not exist, we consider ε-minimizer.

For ERM, one can show that $|I_{\text{emp}}(z_1, \ldots, z_n) - I_{\text{emp}}(z_2, \ldots, z_n)| \leq \frac{M}{n}$. Hence, a "removal" version of Average Stability is closely related to Average Stability:

$$\begin{aligned}
&\mathbb{E}\left(\ell(z_1, \ldots, z_n; z_1) - \ell(z_2, \ldots, z_n; z_1)\right) \\
&\quad = \mathbb{E}\left(I_{\text{emp}}(z_1, \ldots, z_n) - I_{\text{exp}}(z_2, \ldots, z_n)\right) \\
&\quad = \beta_{\text{bias}}(n-1) + \mathbb{E}\left(I_{\text{emp}}(z_2, \ldots, z_n) - I_{\text{emp}}(z_1, \ldots, z_n)\right).
\end{aligned}$$

Thus, $\mathbb{E}\left(\ell(z_1, \ldots, z_n; z_1) - \ell(z_2, \ldots, z_n; z_1)\right) \to 0$ is also equivalent to $|\Psi_n| \xrightarrow{P} 0$.

Furthermore, one can show that

$$\ell(z_1, \ldots, z_n; z_1) - \ell(z_2, \ldots, z_n; z_1) \geq 0$$

for ERM (see [10]), and so CV (Removal) Stability, defined in Sec. 4.3, is equal to the above "removal" version of Average Stability. Hence, $\beta_{\text{cv}}(n) \to 0$ is equivalent to $|\Psi_n| \xrightarrow{P} 0$.

Since Empirical Risk Minimization over a uniform Glivenko–Cantelli class implies that $|\Psi_n| \xrightarrow{P} 0$, it also implies that $\beta_{\text{bias}}(n) \to 0$ and $\beta_{\text{cv}}(n) \to 0$. Thus, ERM over a UGC class is stable in these regards. By using techniques from the Empirical Process Theory, it can be shown (see [3]) that for ERM over a smaller family of classes, called Donsker classes, a much stronger stability in L_1 norm (see Definition 4.1) holds: $\beta_1(n) \to 0$. Donsker classes are classes of functions satisfying the Central Limit Theorem, and for binary classes of function this is equivalent to finiteness of the VC dimension.

5. Rates of Convergence

Previous sections focused on finding rather weak conditions for proving $\Psi_n \xrightarrow{P} 0$ and $\Phi_n \xrightarrow{P} 0$ via Markov's inequality. With stronger notions of stability, it is possible to use more sophisticated inequalities, which is the focus of this section.

5.1. *Uniform stability*

Uniform Stability (see Definition 2.1), is a very strong notion, and we would not expect, in general, that $\beta_\infty(n) \to 0$. Surprisingly, for *Tikhonov Regularization* algorithms

$$\mathcal{A}(z_1, \ldots, z_n) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f; z_i) + \lambda \|f\|_K^2,$$

it can be shown [2] that

$$\beta_\infty(n) \leq \frac{L^2 \kappa^2}{2\lambda n},$$

where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel K, $K(x, x) \leq \kappa^2 < \infty, \forall x \in \mathcal{X}$, and L is a Lipschitz constant relating norms between functions $f \in \mathcal{F}$ to norms between loss functions $\ell \in \mathcal{L}(\mathcal{F})$.

Clearly, β_∞ dominates all stabilities discussed in the previous sections, and so can be used to bound the mean and variance of the estimators. For this strong stability, a more powerful concentration inequality can be used instead of Markov's inequality. McDiarmid's bounded difference inequality states that if a function of many random variables does not change much when one variable is changed, then the function is almost a constant. This is exactly what we need to bound Ψ_n or Φ_n.

Theorem 5.1 (McDiarmid, [9]). *Let $\xi : \mathcal{Z}^n \mapsto \mathbb{R}$ be a measurable function, $\Gamma = \xi(z_1, \ldots, z_n)$, $\Gamma_i' = \xi(z_1, \ldots, z_i', \ldots, z_n)$, where $z_1, \ldots, z_n, z_1', \ldots, z_n'$ are i.i.d. random variables. If for all i,*

$$\sup_{z_1,\ldots,z_n,z_1',\ldots,z_n'} |\Gamma - \Gamma_i'| \leq \beta_n, \tag{5.1}$$

then for any $\varepsilon > 0$,

$$\mathbb{P}(|\Gamma - \mathbb{E}\Gamma| \geq \varepsilon) \leq 2\exp\left(\frac{-2\varepsilon^2}{n\beta_n^2}\right).$$

Bousquet and Elisseeff [2] applied this inequality to $\Gamma = \Psi_n$:

$$\begin{aligned}
&|\Psi_n(z_1, \ldots, z_n) - \Psi_n(z_1, \ldots, z, \ldots, z_n)| \\
&\quad \leq |I_{\text{emp}}(z_1, \ldots, z_n) - I_{\text{emp}}(z, z_2, \ldots, z_n)| \\
&\qquad + |I_{\text{exp}}(z_1, \ldots, z_n) - I_{\text{exp}}(z, z_2, \ldots, z_n)| \\
&\quad \leq \frac{1}{n}|\ell(z_1, \ldots, z_n; z_1) - \ell(z, z_2, \ldots, z_n; z)| \\
&\qquad + \frac{1}{n}\sum_{j=2}^{n} |\ell(z_1, \ldots, z_n; z_j) - \ell(z, z_2, \ldots, z_n; z_j)| \\
&\qquad + \mathbb{E}_z'|\ell(z_1, \ldots, z_n; z') - \ell(z, z_2, \ldots, z_n; z')| \\
&\quad \leq 2\beta_\infty(n) + \frac{M}{n} =: \beta_n.
\end{aligned}$$

If $\beta_\infty(n) = o(n^{-1/2})$, McDiarmid's inequality shows that Ψ_n is exponentially concentrated around $\mathbb{E}\Psi_n$, which is also small:

$$\mathbb{E}\Psi_n = \beta_{\text{bias}}(n) \leq \beta_\infty(n).$$

Therefore,

$$\forall\, \varepsilon > 0, \quad \mathbb{P}(\Psi_n \geq \beta_\infty(n) + \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{(2n\beta_\infty(n) + M)^2}\right).$$

Notice that for ERM, $|I_{\text{emp}}(z_1, \ldots, z_n) - I_{\text{emp}}(z, z_2, \ldots, z_n)| \leq \frac{M}{n}$ and so it is enough to require $\beta_{\text{bias}} \to 0$ and $|I_{\text{exp}}(z_1, \ldots, z_n) - I_{\text{exp}}(z, z_2, \ldots, z_n)| = o(n^{-1/2})$ to get exponential bounds. The last requirement is strong, as it requires expected errors on similar samples to be close for *every* sample. The next section deals with "almost-everywhere" stabilities (see [8]), i.e. when a stability quantity is small for most samples.

5.2. *Extending McDiarmid's inequality*

As one extreme, if we know that $\beta_\infty(n) = o(n^{-1/2})$, we can use exponential McDiarmid's inequality. As the other extreme, if we only have information about averages β_{emp} and β_{exp}, we are forced to use the second moment and Chebyshev's or Markov's inequality. What happens in between these extremes? What if we know more about the random variables $I_{\text{emp}}(z_1, \ldots, z_n) - I_{\text{emp}}(z, z_2, \ldots, z_n)$? One example is the case when we know that these random variables are almost always small. Unfortunately, assumptions of McDiarmid's inequality are no longer satisfied, so other ways of deriving exponential bounds are needed. This section elaborates on this situation.

Assume that for a given β_n, a measurable function $\xi : \mathcal{Z}^n \mapsto [-M, M]$ satisfies the bounded difference condition (5.1) on a subset $G \subseteq \mathcal{Z}^n$ of measure $1 - \delta_n$, while

$$\begin{aligned} &\forall\, (z_1, \ldots, z_n) \in \bar{G}, \exists\, z_i' \in \mathcal{Z} \\ &\qquad \text{s.t. } \beta_n < |\xi(z_1, \ldots, z_n) - \xi(z_1, \ldots, z_i', \ldots, z_n)| \le 2M, \end{aligned}$$

where $\bar{G}$ is the complement of the subset G. Again, denote $\Gamma = \xi(z_1, \ldots, z_n)$, $\Gamma_i' = \xi(z_1, \ldots, z_i', \ldots, z_n)$. A simple application of Efron–Stein inequality shows that

$$\begin{aligned} \operatorname{var}(\Gamma) &\le \frac{1}{2} n \mathbb{E} \left(\xi(z_1, \ldots, z_n) - \xi(z, z_2, \ldots, z_n) \right)^2 \\ &\le \frac{1}{2} n \mathbb{E} \left[I_{(z_1, \ldots, z_n) \in G} \left(\xi(z_1, \ldots, z_n) - \xi(z, z_2, \ldots, z_n) \right)^2 \right] \\ &\quad + \frac{1}{2} n \mathbb{E} \left[I_{(z_1, \ldots, z_n) \in \bar{G}} \left(\xi(z_1, \ldots, z_n) - \xi(z, z_2, \ldots, z_n) \right)^2 \right] \\ &\le \frac{1}{2} n (\beta_n^2 + 4M^2 \delta_n). \end{aligned} \tag{5.2}$$

This leads to a polynomial bound on $\mathbb{P}(|\Gamma - \mathbb{E}\Gamma| \ge \varepsilon)$. Kutin and Niyogi [7, 8] proved an inequality which is exponential when δ_n decays exponentially with n, thus extending McDiarmid's inequality to incorporate a small possibility of a large jump of ξ. A more general version of their bound is the following:

Theorem 5.2 (Kutin and Niyogi [8]). *Assume $\xi : \mathcal{Z}^n \mapsto [-M, M]$ satisfies the bounded difference condition* (5.1) *on a set of measure $1 - \delta_n$ and denote $\Gamma = \xi(z_1, \ldots, z_n)$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}(|\Gamma - \mathbb{E}\Gamma| \ge \varepsilon) \le 2 \exp\left(\frac{-\varepsilon^2}{8 n \beta_n^2} \right) + \frac{2Mn\delta_n}{\beta_n}. \tag{5.3}$$

Note that the bound tightens only if $\beta_n = o(n^{-1/2})$ and $\delta_n / \beta_n = o(n^{-1})$. Furthermore, the bound is exponential only if δ_n decays exponentially.[c]

While the variance bound in (5.2) is written in terms of the second moment, we can use powerful moment inequalities, recently developed by Boucheron *et al.* [1],

[c]By exponential rate, we mean decay $o(\exp(-n^r))$ for a fixed $r > 0$.

to bound the qth moment of Γ. Moreover, q can be optimized to get the tightest bounds.[d]

Define random variables V_+ and V_- as

$$V_+ = \mathbb{E}\left[\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2 I_{\Gamma \geq \Gamma_i'} | z_1, \ldots, z_n\right], \quad V_- = \mathbb{E}\left[\sum_{i=1}^{n}(\Gamma - \Gamma_i')^2 I_{\Gamma < \Gamma_i'} | z_1, \ldots, z_n\right].$$

Theorem 5.3 (Boucheron *et al.* [1]). *For $\xi : \mathcal{Z}^n \mapsto \mathbb{R}$, $\Gamma = \xi(z_1, \ldots, z_n)$, and any $q \geq 2$,*

$$\|(\Gamma - \mathbb{E}\Gamma)_+\|_q \leq \sqrt{2\kappa q}\|\sqrt{V_+}\|_q \quad \textit{and} \quad \|(\Gamma - \mathbb{E}\Gamma)_-\|_q \leq \sqrt{2\kappa q}\|\sqrt{V_+}\|_q,$$

where $x_+ = \max(0, x)$ and $\kappa \approx 1.271$ is a constant.

This result leads directly to the following theorem:

Theorem 5.4. *Assume $\xi : \mathcal{Z}^n \mapsto \mathbb{R}$ satisfies the bounded difference condition* (5.1) *on a set of measure $1 - \delta_n$, and denote $\Gamma = \xi(z_1, \ldots, z_n)$. Then for any $q \geq 2$ and $\varepsilon > 0$,*

$$\mathbb{P}(\Gamma - \mathbb{E}\Gamma > \varepsilon) \leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q},$$

where $\kappa \approx 1.271$.

Proof.

$$\mathbb{E}V_+^{q/2} = \mathbb{E}\{I_G V_+^{q/2} + I_{\bar{G}} V_+^{q/2}\} \leq (n\beta_n^2)^{q/2} + (nq(2M)^2)^{q/2}\delta_n.$$

By Theorem 5.3,

$$\mathbb{E}(\Gamma - \mathbb{E}\Gamma)_+^q \leq (2\kappa q)^{q/2}\mathbb{E}V_+^{q/2} \leq (n\beta_n^2 q 2\kappa)^{q/2} + (n(2M)^2)^{q/2}\delta_n.$$

Hence,

$$\mathbb{P}(\Gamma - \mathbb{E}\Gamma > \varepsilon) \leq \frac{\mathbb{E}(\Gamma - \mathbb{E}\Gamma)_+^q}{\varepsilon^q} \leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q}. \qquad \square$$

Note that the bound of Theorem 5.4 holds for any $q \geq 2$. To clarify the asymptotic behavior of the bound, assume $\beta_n = n^{-\gamma}$ for some $\gamma > 1/2$, and let $q = \varepsilon^2\beta_n^{-2}n^{-2\gamma+\eta} = \varepsilon^2 n^\eta$ for some η to be chosen later, $2\gamma - 1 > \eta > 0$. Assume $\delta_n = \exp(n^{-\theta})$ for some $\theta > 0$. The bound of Theorem 5.4 becomes

$$\begin{aligned}
\mathbb{P}(\Gamma - \mathbb{E}\Gamma > \varepsilon) &\leq \frac{(nq)^{q/2}((2\kappa)^{q/2}\beta_n^q + (2M)^q\delta_n)}{\varepsilon^q} \\
&\leq \left(\frac{2\kappa nq\beta_n^2}{\varepsilon^2}\right)^{q/2} + \delta_n\left(\frac{4M^2nq}{\varepsilon^2}\right)^{q/2} \\
&\leq \left(2\kappa n^{1+\eta-2\gamma}\right)^{\frac{\varepsilon^2}{2}n^\eta} + \left(4M^2n^{1+\eta}\right)^{\frac{\varepsilon^2}{2}n^\eta}\exp(-n^\theta)
\end{aligned}$$

[d]Thanks to Gábor Lugosi for suggesting this method.

$$\leq \exp\left((1+(1+\eta-2\gamma)\log n)n^{\eta}\frac{\varepsilon^2}{2}\right) + \exp\left((2\log(2M)+(1+\eta)\log n)\,n^{\eta}\frac{\varepsilon^2}{2}-n^{\theta}\right). \quad (5.4)$$

Since $1+\eta-2\gamma<0$, the first term is decaying exponentially with n. We can now choose $\eta<\min(\theta,2\gamma-1)$ for the second term to decay exponentially. In particular, let us compare our result to the result of Theorem 5.2. With $\delta_n=\exp(n^{-\theta})$ the bound in Eq. (5.3) becomes

$$\mathbb{P}(\Gamma-\mathbb{E}\Gamma>\varepsilon)\leq \exp\left(-\frac{\varepsilon^2}{8}n^{2\gamma-1}\right) + \exp\big((\log M+(\gamma+1)\log n)-n^{\theta}\big). \quad (5.5)$$

Depending on whether $\theta<2\gamma-1$ or not, the first or second term dominates convergence to zero, which coincides exactly with the asymptotic behavior of our bound. In fact, one can verify that the terms in the exponents of bounds (5.4) and (5.5) have the same order.

We have therefore recovered the result[e] *of Theorem* 5.2 *for the interesting case* $\delta_n=\exp(-n^{\theta})$ *by using moment inequality of Boucheron et al.* [1]. Note that the result of Theorem 5.4 is very general and different ways of picking q might prove useful. For instance, if $\delta_n=0$, i.e. the bounded difference condition (5.1) holds, we can choose $q=\frac{\varepsilon^2}{4n\beta_n^2}$ to recover McDiarmid's inequality.

Having proven extension to McDiarmid's inequality, we can use it in a straightforward way to derive bounds on $\mathbb{P}(|\Psi_n|>\varepsilon)$ and $\mathbb{P}(|\Phi_n|>\varepsilon)$ when expected and empirical quantities do not change "most of the time", when compared on similar samples (see [8] for examples).

6. Summary and Open Problems

We have shown how stability of algorithms provides an alternative to classical Statistical Learning Theory approach for controlling the behavior of empirical and leave-one-out estimates. The results presented are by no means a complete picture: one can come up with other notions of algorithmic stability, suited for the problem. Our goal was to present some results in a common framework and delineate important techniques for proving bounds.

One important (and largely unexplored) area of further research is looking at existing algorithms and proving bounds on their stabilities. For instance, work of Caponnetto and Rakhlin [3] showed that Empirical Risk Minimization (over certain classes) is L_1-stable. It might turn out that other algorithms are stable in this (or even stronger) sense when considered over restricted function classes, which are nevertheless used in practice. Can these results lead to faster learning rates for algorithms?

[e]This gives an answer to the open question 6.2 in [7].

Adding a regularization term for ERM leads to an extremely stable Tikhonov Regularization algorithm. How can regularization be used to stabilize other algorithms, and how does this affect the bias-variance trade-off of fitting the data versus having a simple solution?

Though the results presented in this paper are theoretical, there is a potential for estimating stability in practice. Can a useful quantity be computed by running the algorithm many times to determine its stability? Can this quantity serve as a measure of the performace of the algorithm?

Acknowledgments

This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Department of Brain and Cognitive Sciences, and which is affiliated with the Computer Sciences and Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co. Ltd., ITRI, Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co. Ltd.

References

[1] S. Boucheron, O. Bousquet, G. Lugosi and P. Massart, Moment inequalities for functions of independent random variables, *Ann. Probab.* **33**(2) (2005) 514–560.

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

[3] A. Caponnetto and A. Rakhlin, Some properties of Empirical Risk Minimization over Donsker classes, AI Memo 2005-018, Massachusetts Institute of Technology (May 2005).

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

[5] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, No. 31 (Springer, New York, 1996).

[6] M. J. Kearns and D. Ron, Algorithmic stability and sanity-check bounds for leave-one-out cross-validation, in *COLT* (1997), pp. 152–162.
[7] S. Kutin, Extensions to McDiarmid's inequality when differences are bounded with high probability, Technical report TR-2002-04, University of Chicago (2002).
[8] S. Kutin and P. Niyogi, Almost-everywhere algorithmic stability and generalization error, Technical report TR-2003-03, University of Chicago (2002).
[9] C. McDiarmid, On the method of bounded differences, *In Surveys in Combinatorics 1989* (1989), pp. 148–188.
[10] S. Mukherjee, P. Niyogi, T. Poggio and R. Rifkin, Statistical learning: Stability is necessary and sufficient for consistency of empirical risk minimization, CBCL Paper 2002-023, Massachusetts Institute of Technology (December 2002) (January 2004 revision).