On Stability and Concentration of Measure

Alexander Rakhlin, Sayan Mukherjee and Tomaso Poggio

Center for Biological Computation and Learning, McGovern Institute, Computer Science and Artificial Intelligence Lab, Brain Sciences Department, Massachusetts Institute of Technology

CBCL Paper 239

June 2004

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2004		**00-06-2004 to 00-06-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
On Stability and Concentration of Measure	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Center for Biological and Computational Learning,77 Massachusetts Avenue,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**13**	

Abstract

Stability conditions can be thought of as a way of controlling the variance of the learning process. Strong stability conditions additionally imply concentration of certain quantities around their expected values. It was shown recently that stability of learning algorithms is closely related to their generalization and consistency. In this paper we examine stability conditions from this point of view, complementing the results of [6, 5].

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 Dept. of Brain Cognitive Sciences, and which is affiliated with the Computer Sciences Artificial Intelligence Laboratory (CSAIL).

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, 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), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda RD Co., Ltd., ITRI, Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph and Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co., Ltd..

1 Introduction

This paper is motivated by the results of [5] (see also [6]). Mukherjee et al. [5], building on work by [2] and by [1], showed that a certain form of double-sided *cross-validation leave-one-out* stability is not only necessary and sufficient for generalization and consistency of ERM but it is also sufficient, when *expected and empirical leave-one-out stability* hold, for generalization of any symmetric learning algorithm. In this paper, we describe a few additional results that will hopefully illuminate better the role of stability in generalization. We work in the *change-one* framework instead of the *leave-one* framework. We show that a weak form of stability called pseudostability (see [5], definition 3.9 for the leave-one-out case) is not only necessary and sufficient for ERM algorithms but is also sufficient for generalization, if *expected and empirical change-one-out stability* hold with sufficiently fast rates. We also show that by using a stronger definition of CV stability than [5] we are able to ensure generalization by using only expected stability, without empirical stability.

2 Extending McDiarmid's Inequality

McDiarmid's inequality has been used in the past few years to obtain concentration results from stability conditions. These stability conditions can be thought of as Lipschitz conditions on the map from sets to functions (i.e. a change in the training set does not affect the output function by more than β = Lipschitz constant). When the Lipschitz constant can be shown to be decreasing in the number of points faster than $O(1/\sqrt{n})$, concentration results follow from McDiarmid's inequality.

Theorem 2.1 (McDiarmid, [4]) *Let $\Omega_1, ..., \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^{n} \Omega_k$ and let X be a random variable on Ω which is uniformly difference-bounded by β_n (i.e for any k if $\omega, \omega' \in \Omega$ differ only in the k-th coordinate, then $|X(\omega) - X(\omega')| \leq \beta_n$), then for any $\epsilon > 0$,*

$$\mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{n\beta_n^2}\right)$$

Kutin and Niyogi [2] extended McDiarmid's inequality to include a possibility of a bad event:

Theorem 2.2 (Kutin,Niyogi) *Let X be a random variable $(|X| \leq 1)$ on Ω which is strongly difference-bounded by $(\frac{\lambda}{n}, \exp(-Kn))$ (i.e. there is a bad subset $B \subset \Omega$ of measure $\exp(-Kn)$ s.t. for any k if $\omega, \omega' \in \Omega$ differ only in the k-th coordinate, then $|X(\omega) - X(\omega')|$ is bounded by $\frac{\lambda}{n}$ if $\omega \notin B$ and by 1 otherwise), then for any $0 \leq \epsilon \leq 2\lambda\sqrt{K}$ and $n \geq \max\{\frac{1}{\lambda}, 3(\frac{6}{K} + 3)\ln(\frac{6}{K} + 3)\}$,*

$$\mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 n}{8\lambda^2}\right)$$

This theorem is a special case of a more general theorem proved by Kutin and Niyogi [2]:

Theorem 2.3 (Kutin,Niyogi) *Let X be a random variable ($|X| \leq 1$) on Ω which is strongly difference-bounded by (β_n, δ), $1 \geq \beta_n > 0$. Then for any ϵ,*

$$\mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \leq 2\left(\exp\left(\frac{-\epsilon^2}{8n\beta_n^2}\right) + \frac{n\delta_n}{\beta_n}\right)$$

For the above bound to decrease with n, β_n has to decrease faster than $O(1/\sqrt{n})$. Additionally, δ_n has to decrease faster than β_n/n, i.e. faster than $n^{-3/2}$.
We now give an example of a random variable which is strongly difference bounded by $(0, n^{-1/2})$, but is not concentrated:

Example Let $\omega = (\omega_1, ..., \omega_n) \in [0,1]^n$. Let $X(\omega) = 0$ if the number of ω_i's greater than $1/2$ is larger than $\lceil n/2 \rceil$ and $X(\omega) = 1$ otherwise. In other words, X takes values 0 or 1 depending on whether the majority of the points falls to the left or to the right of $1/2$. Note that a change of one point does not change the value of X unless the set ω of points is balanced. The probability of this event is $\Theta(1/\sqrt{n})$. Even though the measure of the "bad event" decreases, X is not concentrated: $\mathbb{E}X = 1/2$ by symmetry.

The above example shows that McDiarmid's inequality cannot be extended to give a useful result for bad sets of measure $\delta = O(1/\sqrt{n})$, while the extension by Kutin and Niyogi shows that for fast enough rates ($\delta = o(n^{-3/2})$ and appropriate β_n), X is concentrated around its mean.

3 Concentration and Stability

Let $S = (z_1, ..., z_n)$ and $S^{i,z} = (z_1, ..., z_{i-1}, z, z_{i+1}, ..., z_n)$. For brevity of notation, let f_S be the *loss* function when trained on the set S (i.e. $V(f_S, z)$ in the notation of [5]). Assume that such functions are upper bounded by M.
Consider the following stability definitions for *change-one* that is replacement of one point (compare with the analog EE_{loo} and E_{loo} definitions of [5] for the leave-one-out case):

Definition 3.1 *We say that an algorithm is $(\beta_{emp}, \delta_{emp})$ Empirical Error stable if with probability $1 - \delta_{emp}$ (over the choice of S),*

$$\forall z, \quad \left| \frac{1}{n} \sum_{z_j \in S} f_S(z_j) - \frac{1}{n} \sum_{z_j \in S^{i,z}} f_{S^{i,z}}(z_j) \right| \leq \beta_{emp}$$

Definition 3.2 *We say that an algorithm is $(\beta_{exp}, \delta_{exp})$ Expected Error stable if with probability $1 - \delta_{exp}$ (over the choice of S),*

$$\forall z, \quad |\mathbb{E}_u f_S(u) - \mathbb{E}_u f_{S^{i,z}}(u)| \leq \beta_{exp}$$

By the concentration result of the previous section, if $\beta_{emp} = o(1/\sqrt{n})$, $\delta_{emp} = o(\beta_{emp}/n)$, the empirical errors are concentrated around their expected value $\mathbb{E}_S \sum_{z_j \in S} f_S(z_j) = \mathbb{E}_S f_S(z_1)$ and if $\beta_{exp} = o(1/\sqrt{n})$, $\delta_{exp} = o(\beta_{exp}/n)$, the expected errors are concentrated around their expected value $\mathbb{E}_{S,z} f_S(z)$:

Proposition 3.1 *If an algorithm is $(\beta_{emp}, \delta_{emp})$ Empirical Error stable, then with probability at least $1 - \left(2\exp\left(-\frac{\epsilon^2}{8n\beta_{emp}^2}\right) + \frac{n\delta_{emp}M}{\beta_{emp}}\right)$,*

$$|\mathbb{E}_S f_S(z_1) - \frac{1}{n}\sum_{z_j \in S} f_S(z_j)| \leq \epsilon.$$

Proposition 3.2 *If an algorithm is $(\beta_{exp}, \delta_{exp})$ Expected Error stable, then with probability at least $1 - \left(2\exp\left(-\frac{\epsilon^2}{8n\beta_{exp}^2}\right) + \frac{n\delta_{exp}M}{\beta_{exp}}\right)$,*

$$|\mathbb{E}_{S,z} f_S(z) - \mathbb{E}_z f_S(z)| \leq \epsilon.$$

If the above two stability conditions hold,

$$\begin{aligned} \mathbb{E}_T(\mathbb{E}_z f_T(z) - \sum_{z_j \in T} f_T(z_j))^2 &\approx \mathbb{E}_T \left[\mathbb{E}_{S,z} f_S(z) - \mathbb{E}_S f_S(z_1)\right]^2 \\ &= \mathbb{E}_T \left[\mathbb{E}_{S,z} \left(f_S(z) - f_{S^{i,z}}(z)\right)\right]^2 \end{aligned}$$

and therefore for the second moment to decrease, there must be a condition forcing

$$\mathbb{E}_{S,z} \left(f_S(z) - f_{S^{i,z}}(z)\right) \to 0.$$

We will call this condition *CV-Pseudostability*.

Definition 3.3 *We say that an algorithm has β_{ps} CV-Pseudostability if*

$$|\mathbb{E}_{S,z} \left(f_S(z) - f_{S^{i,z}}(z)\right)| \leq \beta_{ps}$$

This is the analog of the leave-one-out pseudostability defined by [5] in definition 3.9, which is weaker than their CV_{loo} stability because because $f_S(z) - f_{S^{i,z}}(z)$ has to be small only on *average*.
Now note that Empirical Error Stability for the removal case implies Empirical Error Stability for replacement (with appropriate rates), and the same holds for the Expected Error. Therefore, CV-Pseudostability gives a weak condition which together with Error Stability and Empirical Stability with the rates $\beta_n = o(1/\sqrt{n})$, $\delta_n = o(\beta_n/n)$, imply convergence of the empirical error to the expected error for any symmetric algorithm.
To elaborate more on this point, assume the Error Stability (removal) and the empirical Stability (removal). Because of the Error Stability, $\mathbb{E}_{S,z} f_S(z) \approx \mathbb{E}_S f_{S^i}(z_i)$. Also, $\mathbb{E}_{S,z} f_{S^{i,z}}(z) = \mathbb{E}_S f_S(z_i)$. Therefore, translated into the removal case, the CV-Pseudostability condition becomes

$$\mathbb{E}_S \left(f_{S^i}(z_i) - f_S(z_i)\right) \to 0.$$

This is exactly CV_{loo} stability without absolute values and was called pseudoPH stability in definition 3.9 of [5]. We therefore conclude that for the removal case, Error Stability together with Empirical Stability (with rates $\beta_n = o(1/\sqrt{n})$, $\delta_n = o(\beta_n/n)$) and pseudoPH stability are enough for generalization. This result should be compared with Theorem 3.1 of [5]): here we also obtain generalization by assuming a *weaker* CV stability but *stronger* empirical and expected stability.

4 Lower Bounds Using Stability

In this section we will lower-bound the second moment

$$\mathbb{E}_S(\mathbb{E}_z f_S(z) - \sum_{z_j \in S} f_S(z_j))^2.$$

Clearly,

$$\begin{aligned}
\mathbb{E}_S(\mathbb{E}_z f_S(z) - \sum_{z_j \in S} f_S(z_j))^2 &\geq (\mathbb{E}_{S,z} f_S(z) - \mathbb{E}_S \sum_{z_j \in S} f_S(z_j))^2 \\
&= \left[\mathbb{E}_{S,z}\left(f_S(z) - f_S(z_1)\right)\right]^2 \\
&= \left[\mathbb{E}_{S,z}\left(f_S(z) - f_{S^{i,z}}(z)\right)\right]^2
\end{aligned}$$

Therefore, convergence of $\mathbb{E}_{S,z}\left(f_S(z) - f_{S^{i,z}}(z)\right)$ to zero (CV-Pseudostability) is a necessary condition for the convergence of the empirical to the expected. For ERM, this condition is also sufficient (see next section).
We now examine the question of necessity of all three stability conditions posed by [5], but with CV-Pseudostability instead of CV_{loo} as the first condition. Assume that CV-Pseudostability holds ($\beta_{ps} \to 0$). Assume additionally that Error Stability holds ($\beta_{err} = o\left(\frac{1}{\sqrt{n}}\right)$, $\delta_{err} = o(\beta_{err}/n)$). Then

$$\begin{aligned}
\mathbb{E}_z f_S(z) - \sum_{z_j \in S} f_S(z_j) &= (\mathbb{E}_z f_S(z) - \mathbb{E}_{S,z} f_S(z)) \\
&+ (\mathbb{E}_{S,z} f_S(z) - \mathbb{E}_{S,z} f_{S^{i,z}}(z)) \\
&+ \left(\mathbb{E}_{S,z} f_{S^{i,z}}(z) - \sum_{z_j \in S} f_S(z_j)\right)
\end{aligned}$$

The first term is bounded by the concentration of expected values around their mean (follows from Expected Stability and the results of the previous section). The second term is bounded by CV-Pseudostability. Therefore,

$$\begin{aligned}
\mathbb{E}_z f_S(z) - \sum_{z_j \in S} f_S(z_j) &\approx \mathbb{E}_{S,z} f_{S^{i,z}}(z) - \sum_{z_j \in S} f_S(z_j) \\
&= \mathbb{E}_T \sum_{z_j \in T} f_T(z_j) - \sum_{z_j \in S} f_S(z_j)
\end{aligned}$$

Therefore, for the empirical to converge to expected, we must require concentration of empiricals around their mean. Empirical Stability does imply this concentration, but it might be possible to have a weaker requirement. Similarly, if we have CV-Pseudostability and Empirical Stability, we must require a concentration of expected values around their mean. This is implied by Expected Stability, but, again, there might be a weaker condition.

5 Empirical Risk Minimization

Proposition 5.1 *CV-Pseudostability is necessary and sufficient for consistency and generalization of any Emprical Risk Minimization algorithm.*

PROOF:Empirical Risk Minimization searches in the function class $\mathcal{F}$ for a function which minimizes (or ϵ-minimizes) empirical risk. Assume f^* is the loss function with the smallest *expected* error, i.e. $\mathbb{E}_z f^*(z) \leq \inf_{g \in \mathcal{F}} \mathbb{E}_z g(z)$. Consider the shifted loss class $\mathcal{G} = \mathcal{F} - f^* = \{g' = f - f^* | f \in \mathcal{F}\}$. Let $g_S = f_S - f^*$. Note that if f_S is an empirical minimizer in $\mathcal{F}$ w.r.t. set S, then g_S is the empirical minimizer in $\mathcal{G}$ w.r.t. S.

$$\begin{aligned} \mathbb{E}_z f_S(z) - \frac{1}{n} \sum_{z_j \in S} f_S(z_j) &= \mathbb{E}_z g_S(z) - \frac{1}{n} \sum_{z_j \in S} g_S(z_j) \\ &+ \mathbb{E}_z f^*(z) - \frac{1}{n} \sum_{z_j \in S} f^*(z_j) \end{aligned}$$

The second term tends to zero by Hoeffding's inequality. Therefore, generalization over class $\mathcal{F}$ is equivalent to generalization over $\mathcal{G}$. Moreover, note that $\frac{1}{n} \sum_{z_j \in S} g_S(z_j) \leq 0$ because the zero function is in the class $\mathcal{G}$ and that $\mathbb{E}_z g_S(z) \geq 0$ because f^* attains the smallest expected error. Therefore, $\mathbb{E}_z g_S(z) - \frac{1}{n} \sum_{z_j \in S} g_S(z_j) \geq 0$ and so convergence $\mathbb{E}_z g_S(z) - \frac{1}{n} \sum_{z_j \in S} g_S(z_j) \to 0$ is equivalent to

$$\mathbb{E}_S \left(\mathbb{E}_z g_S(z) - \frac{1}{n} \sum_{z_j \in S} g_S(z_j) \right) \to 0.$$

Rewriting,

$$\begin{aligned} \mathbb{E}_S \left(\mathbb{E}_z g_S(z) - \frac{1}{n} \sum_{z_j \in S} g_S(z_j) \right) &= \mathbb{E}_{S,z} \left(g_S(z) - g_S(z_1) \right) \\ &= \mathbb{E}_{S,z} \left(g_S(z) - g_{S^{i,z}}(z) \right) \\ &= \mathbb{E}_{S,z} \left(f_S(z) - f_{S^{i,z}}(z) \right) \end{aligned}$$

As shown in Theorem 3.4 of [5], CV_{loo}-Pseudostability is also equivalent to generalization and consistency of ERM (for ERM CV_{loo} pseudostability and CV_{loo} stability are equivalent).

6 Bounding Generalization Error without Strong Concentration Results

Mukherjee et al [5] showed that for the replacement case, CV_{loo} stability together with Expected and Empirical stabilities, is sufficient for generalization. Their method (similar to that of Devroye and Wagner [3]) bounds the second moment of the difference of the expected and empirical errors. No concentration of the errors around their average values is required for this method. We now prove a similar result for the replacement case and show that by using a somewhat stronger definition of CV stability (which is different from but consistent with our definition of pseudostability) we can prove sufficiency for generalization using only the expected stability (without empirical stability).

Definition 6.1 *We say that an algorithm is $(\beta_{cv}, \delta_{cv})$ strongly CV stable when*

$$\mathbb{P}\left(\forall i,\ |f_S(z) - f_{S^{i,z}}(z)| > \beta_{cv}\right) \leq \delta_{cv}$$

Alternative form (by symmetry):

$$\mathbb{P}\left(\forall i,\ |f_{S^{i,z}}(z_i) - f_S(z_i)| > \beta_{cv}\right) \leq \delta_{cv}$$

It is crucial that the quantifier $\forall i$ is inside of the probability. Thus this definition is stronger than CV$_{loo}$ stability (or its change-one analog). Also note that the probabilities can be taken over $n+1$ points or over n points with a fixed z.

Proposition 6.1 *For any $i \neq j$, with probability at least $1 - \delta_{cv}$,*

$$|f_S(z_j) - f_{S^{i,z}}(z_j)| \leq 2\beta_{cv}.$$

PROOF:

$$|f_S(z_j) - f_{S^{i,z}}(z_j)| \leq |f_S(z_j) - f_{S^{j,z}}(z_j)| + |f_{S^{j,z}}(z_j) - f_{S^{i,z}}(z_j)|$$

Both terms above are bounded by CV stability. Indeed, in the first term, we're starting with the set $S^{j,z}$ which does not contain z_j and replacing it by the set $(S^{j,z})^{j,z_j} = S$ which contains it. In the second term, we're starting with the set $S^{j,z}$ which does not contain z_j and replacing it by the set $(S^{j,z})^{i,z_j} = S^{i,z}$ which contains it.

Proposition 6.2 *Strong CV stability implies that*

$$\left|\frac{1}{n}\sum_{z_i \in S} f_S(z_i) - \frac{1}{n}\sum_{z_i \in S} f_{S^{i,z}}(z_i)\right| \leq \beta_{cv}$$

with probability $1 - \delta_{cv}$.

PROOF:

$$\left|\frac{1}{n}\sum_{z_i\in S} f_S(z_i) - \frac{1}{n}\sum_{z_i\in S} f_{S^{i,z}}(z_i)\right| \leq \frac{1}{n}\sum_{z_i\in S} |f_S(z_i) - f_{S^{i,z}}(z_i)| \leq \beta_{cv}$$

The reason the probability of this event does not increase is due to the way we defined *strong CV* stability. The pair S, z is "good" with probability $1-\delta_{cv}$ and then *any coordinate i* can be changed.

Proposition 6.3 *Strong CV stability and Expected Error stability imply generalization. More precisely,*

$$\mathbb{E}_S(\mathbb{E}_z f_S(z) - \frac{1}{n}\sum_{z_i\in S} f_S(z_i))^2 \leq M(4\beta_{cv} + 3M\delta_{cv} + 3\beta_{exp} + 2M\delta_{exp} + 1/n)$$

PROOF:

$$\begin{aligned}\mathbb{E}_S(\mathbb{E}_z f_S(z) \quad - \quad & \frac{1}{n}\sum_{z_i\in S} f_S(z_i)))^2 = \mathbb{E}_S\left[\mathbb{E}_z f_S(z)\mathbb{E}_{z'} f_S(z')\right] - \mathbb{E}_S\left[\mathbb{E}_z f_S(z)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right] \\ + \quad & \mathbb{E}_S\left[\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right] - \mathbb{E}_S\left[(\mathbb{E}_z f_S(z))\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right]\end{aligned}$$

First,

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

Second,

$$\begin{aligned}\mathbb{E}_S\mathbb{E}_z f_S(z)\mathbb{E}_{z'} f_S(z') &= \mathbb{E}_S\left(\mathbb{E}_z f_S(z)\mathbb{E}_{z'} f_S(z')\right) - \mathbb{E}_S\left(\mathbb{E}_z f_S(z)\mathbb{E}_{z'} f_{S^{i,z'}}(z')\right) \\ &+ \mathbb{E}_S\left(\mathbb{E}_z f_S(z)\mathbb{E}_{z'} f_{S^{i,z'}}(z')\right) - \mathbb{E}_{S,z'}\left(\mathbb{E}_z f_{S^{i,z'}}(z) f_{S^{i,z'}}(z')\right) \\ &+ \mathbb{E}_{S,z',z}\left(f_{S^{i,z'}}(z) f_{S^{i,z'}}(z')\right) \\ &\leq M(\beta_{cv} + M\delta_{cv}) + M(\beta_{exp} + M\delta_{exp}) + \mathbb{E}_S f_S(z) f_S(z_i)\end{aligned}$$

We bound the first term using CV stability, second using Expected Error stability, and use symmetry at the last step.
Finally, by Proposition 6.2

$$\mathbb{E}_S\left[\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right] \leq M(\beta_{cv}+M\delta_{cv})+\mathbb{E}_S\left[\left(\frac{1}{n}\sum_{z_i\in S} f_{S^{i,z}}(z_i)\right)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right].$$

Furthermore, for $i \neq j$ by symmetry,

$$\begin{aligned}\mathbb{E}_S\left[\left(\frac{1}{n}\sum_{z_i\in S} f_{S^{i,z}}(z_i)\right)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right] &= \frac{n^2-n}{n^2}\mathbb{E}_{S,z}\left(f_{S^{i,z}}(z_i)f_S(z_j)\right) + \frac{n}{n^2}\mathbb{E}_S\left(f_{S^{i,z}}(z_i)f_S(z_i)\right) \\ &\leq \mathbb{E}_{S,z}\left(f_{S^{i,z}}(z_i)f_S(z_j)\right) + M/n.\end{aligned}$$

Now, by symmetry $\mathbb{E}_{S,z} f_{S^{i,z}}(z_i)f_S(z_j) = \mathbb{E}_{S,z} f_S(z)f_{S^{i,z}}(z_j)$. By Proposition 6.1, with probability $1-\delta_{cv}$, $|f_{S^{i,z}}(z_j) - f_S(z_j)| \leq 2\beta_{cv}$. Therefore,

$$\mathbb{E}_{S,z}\left(f_{S^{i,z}}(z_i)f_S(z_j)\right) \leq M(2\beta_{cv} + M\delta_{cv}) + \mathbb{E}_S\left(f_S(z)f_S(z_j)\right).$$

Putting it together,

$$\mathbb{E}_S\left(\mathbb{E}_S\left[\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\left(\frac{1}{n}\sum_{z_i\in S} f_S(z_i)\right)\right]\right) \leq \mathbb{E}_{S,z}\left(f_S(z)f_S(z_j)\right) + M(3\beta_{cv} + 2M\delta_{cv}) + M/n$$

The grand total is:

$$\mathbb{E}_S(\mathbb{E}_z f_S(z) - \frac{1}{n}\sum_{z_i\in S} f_S(z_i))^2 \leq M(4\beta_{cv} + 3M\delta_{cv} + 3\beta_{exp} + 2M\delta_{exp} + 1/n)$$

7 Remarks

This paper clarifies a few questions left open by previous work on stability and specifically by [6, 5]. In particular, it clarifies that if pseudostability holds neither empirical nor expected error alone is sufficient to ensure generalization. More importantly, it shows that there exist several alternative stability conditions which are sufficient for generalization in general and are all equivalent to generalization and consistency of ERM.

Acknowledgments We would like to thank Shie Mannor and Shahar Mendelson for useful discussions.

References

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

[2] S. Kutin. Extensions to mc diarmid's inequality when differences are bounded with high probability. Technical report TR-2002-04, University of Chicago, 2002.

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

[4] C. McDiarmid. On the method of bounded differences. *In Surveys in Combinatorics 1989*, pages 148–188, 1989.

[5] 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].

[6] T. Poggio, T. Rifkin, R. Mukherjee S., and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, pages 419–422, 2004.

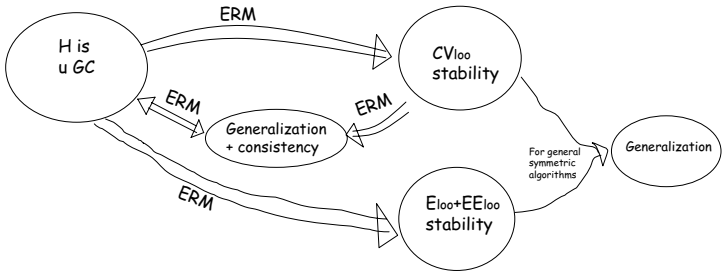

Figure 1: *An overall view of some of the properties discussed in [5].*

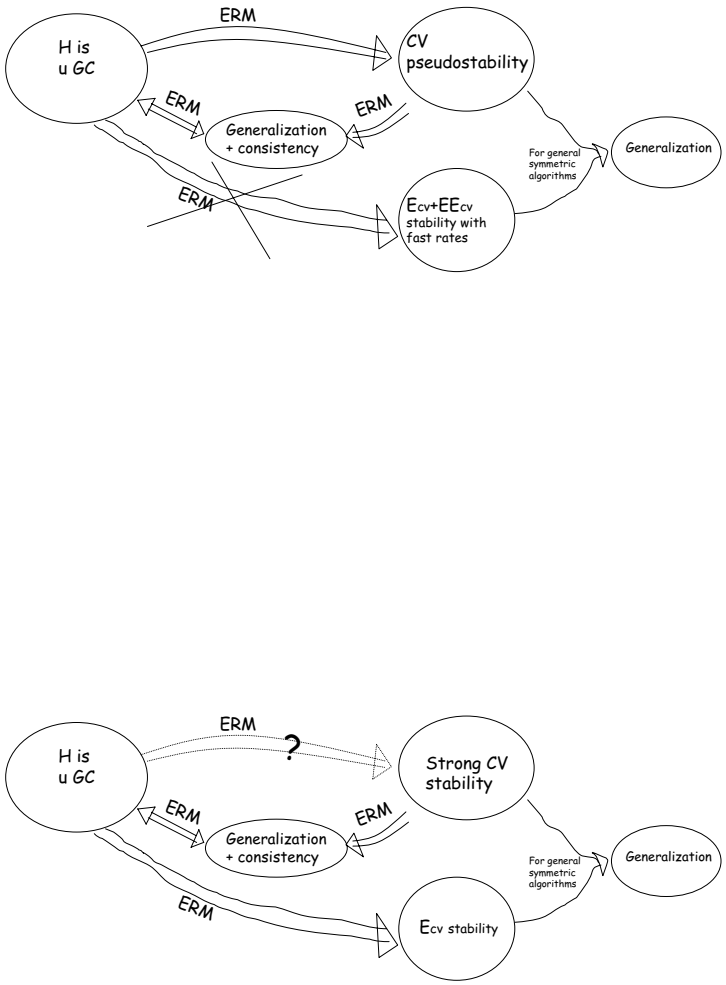

Figure 2: *The two main new results of this paper.*