

# Stability Properties of Empirical Risk Minimization over Donsker Classes

**Andrea Caponnetto**

**Alexander Rakhlin**

*Center for Biological and Computational Learning*

*Massachusetts Institute of Technology*

*Cambridge, MA, 02139, USA*

CAPONNET@UCHICAGO.EDU

RAKHLIN@ALUM.MIT.EDU

**Editor:** Leslie Pack Kaelbling

## Abstract

We study some stability properties of algorithms which minimize (or almost-minimize) empirical error over Donsker classes of functions. We show that, as the number  $n$  of samples grows, the  $L_2$ -diameter of the set of almost-minimizers of empirical error with tolerance  $\xi(n) = o(n^{-\frac{1}{2}})$  converges to zero in probability. Hence, even in the case of multiple minimizers of expected error, as  $n$  increases it becomes less and less likely that adding a sample (or a number of samples) to the training set will result in a large jump to a new hypothesis. Moreover, under some assumptions on the entropy of the class, along with an assumption of Komlos-Major-Tusnady type, we derive a power rate of decay for the diameter of almost-minimizers. This rate, through an application of a uniform ratio limit inequality, is shown to govern the closeness of the expected errors of the almost-minimizers. In fact, under the above assumptions, the expected errors of almost-minimizers become closer with a rate strictly faster than  $n^{-1/2}$ .

**Keywords:** empirical risk minimization, empirical processes, stability, Donsker classes

## 1. Introduction

The empirical risk minimization (ERM) algorithm has been studied in learning theory to a great extent. Vapnik and Chervonenkis (1971, 1991) showed necessary and sufficient conditions for its consistency. In recent developments, Bartlett and Mendelson (2006); Bartlett et al. (2004); Koltchinskii (2006) proved sharp bounds on the performance of ERM. Tools from empirical process theory have been successfully applied, and, in particular, it has been shown that the *localized Rademacher averages* play an important role in studying the behavior of the ERM algorithm.

In this paper we are not directly concerned with rates of performance of ERM. Rather, we prove some properties of ERM algorithms, which, to our knowledge, do not appear in the literature. The analysis of this paper has been motivated by the study of *algorithmic stability*: the behavior of a learning algorithm with respect to perturbations of the training set. Algorithmic stability has been studied in the recent years as an alternative to the classical (complexity-oriented) approach to deriving generalization bounds (Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Mukherjee et al., 2006; Poggio et al., 2004; Rakhlin et al., 2005). Motivation for studying algorithmic stability comes, in part, from the work of Devroye and Wagner (1979). Their results indicate that for any algorithm, the performance of the leave-one-out estimator of expected error is bounded by  $L_1$ -stability of the algorithm, that is, by the average  $L_1$  distance between hypotheses on similar samples. This

result can be used to derive bounds on the performance of the leave-one-out estimate for algorithms such as  $k$ -Nearest Neighbors. It is important to note that no class of finite complexity is searched by algorithms like  $k$ -NN, and so the classical approach of using complexity of the hypothesis space fails.

Further important results were proved by Bousquet and Elisseeff (2002), where a large family of algorithms (*Tikhonov regularization* based methods) has been shown to possess a strong  $L_\infty$  stability with respect to changes of single samples of the training set, and exponential bounds have been proved for the generalization error in terms of empirical error. Tikhonov regularization based algorithms minimize the empirical error plus a stabilizer, and are closely related to ERM. Though ERM is not, in general,  $L_\infty$ -stable, it is  $L_1$ -stable over certain classes of functions, as one of the results of this paper shows. To the best of our knowledge, the outcomes of the present paper do not follow directly from results available in the machine learning literature. In fact we had to turn to empirical process theory for the mathematical tools necessary for studying stability of ERM.

Various assumptions on the function class, over which ERM is performed, have been considered recently to obtain fast rates on the performance of ERM. The importance of having a unique best function in the class has been shown by Lee et al. (1998): the difficult learning problems seem to be the ones where two minimizers of the expected error exist and are far apart. Although the present paper does not address the question of performance rates, it does shed some light on the behavior of ERM when two (or more) minimizers of expected error exist. Our results imply that, under a certain weak condition on the class, as the expected performance of empirical minimizers approaches the best in the class with the addition of new samples, a jump to a different part of the function class becomes less and less likely.

Since ERM minimizes empirical error instead of expected error, it is reasonable to require that the two quantities become close uniformly over the class, as the number of examples grows. Hence, ERM is a sound strategy only if the function class is uniform Glivenko-Cantelli, that is, it satisfies the uniform law of large numbers. In this paper we focus our attention on a more restricted family of function classes: Donsker classes (see for example, Dudley, 1999). These are classes satisfying not only the law of large numbers, but also a version of the central limit theorem. Though a more restricted family of classes, Donsker classes are still quite general. In particular, uniform Donsker and uniform Glivenko-Cantelli properties are equivalent in the case of binary-valued functions (and also equivalent to finiteness of VC dimension). The central limit theorem for Donsker classes states a form of convergence of the empirical process to a Gaussian process with a specific covariance structure (see for example, Dudley, 1999; van der Vaart and Wellner, 1996). This structure is used in the proof of the main result of the paper to control the correlation of the empirical errors of ERM minimizers on similar samples.

The paper is organized as follows. In Section 2 we introduce the notation and background results. Section 3 presents the main result of the paper, which is proved in the appendix using tools from empirical process theory. In Section 4, we show  $L_1$ -stability of ERM over Donsker classes as an application of the main result of Section 3. In Section 5 we show an improvement (in terms of the rates) of the main result under a suitable Komlos-Major-Tusnady condition and an assumption on entropy growth. Section 6 combines the results of Sections 4 and 5 and uses a uniform ratio limit theorem to obtain fast rates of decay on the deviations of expected errors of almost-ERM solutions, thus establishing *strong expected error stability* of ERM (see Mukherjee et al., 2006). Section 7 is a final summary of the results of the paper. Most of the proofs are postponed to the Appendix.

## 2. Notation and Background Results

Let  $(Z, \mathcal{A})$  be a measurable space. Let  $P$  be a probability measure on  $(Z, \mathcal{A})$  and  $Z_1, \dots, Z_n$  be independent copies of  $Z$  with distribution  $P$ . Let  $\mathcal{F}$  be a class of functions from  $Z$  to  $\mathbb{R}$ . In the setting of learning theory, samples  $Z$  are input-output pairs  $(X, Y)$  and for  $f \in \mathcal{F}$ ,  $f(Z)$  measures how well the relationship between  $X$  and  $Y$  is captured by  $f$ . The goal is to minimize  $Pf = \mathbb{E}f(Z)$  where information about the unknown  $P$  is given only through the finite sample  $S = (Z_1, \dots, Z_n)$ . Define the empirical measure as  $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i}$ .

**Definition 1** Given a sample  $S$ ,

$$f_S := \operatorname{argmin}_{f \in \mathcal{F}} P_n f = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(Z_i)$$

is a minimizer of the empirical risk (empirical error), if the minimum exists.

Since an exact minimizer of the empirical risk might not exist, as well as for algorithmic reasons, we consider the set of almost-minimizers of empirical risk.

**Definition 2** Given  $\xi \geq 0$  and  $S$ , define the set of almost empirical minimizers

$$\mathcal{M}_S^\xi = \{f \in \mathcal{F} : P_n f - \inf_{g \in \mathcal{F}} P_n g \leq \xi\}$$

and define its diameter as

$$\operatorname{diam} \mathcal{M}_S^\xi = \sup_{f, g \in \mathcal{M}_S^\xi} \|f - g\|.$$

The  $\|\cdot\|$  in the above definition is the seminorm on  $\mathcal{F}$  induced by symmetric bilinear product

$$\langle f, f' \rangle = P((f - Pf)(f' - Pf')),$$

hence  $\|f\|$  is the standard deviation of  $f$  relative to  $P$ .

This is a natural measure of distance between functions, as will become apparent later, because of the central role of the covariance structure of Brownian bridges in our proofs. The results obtained for the seminorm  $\|\cdot\|$  will be easily extended to the  $L_2(P)$  norm, thanks to the close relation of these two notions of distance.

**Definition 3** The empirical process  $v_n$  indexed by  $\mathcal{F}$  is defined as the map

$$f \mapsto v_n(f) = \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(Z_i) - Pf).$$

**Definition 4** A class  $\mathcal{F}$  is called  $P$ -Donsker if

$$v_n \rightsquigarrow v$$

in  $\ell^\infty(\mathcal{F})$ , where the limit  $v$  is a tight Borel measurable element in  $\ell^\infty(\mathcal{F})$  and " $\rightsquigarrow$ " denotes weak convergence, as defined on p. 17 of van der Vaart and Wellner (1996).

In fact, it follows that the limit process  $v$  must be a zero-mean Gaussian process with covariance function  $\mathbb{E}v(f)v(f') = \langle f, f' \rangle$  (i.e., a Brownian bridge).

Various Donsker theorems provide sufficient conditions for a class being  $P$ -Donsker. Here we mention a few known results (see van der Vaart and Wellner 1996, Equation 2.1.7 and van de Geer 2000, Theorem 6.3) in terms of entropy and entropy with bracketing, which we define below (see van der Vaart and Wellner, 1996).

**Definition 5** *The covering number  $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$  is the minimal number of balls  $\{g : \|g - f\| < \epsilon\}$  of radius  $\epsilon$  needed to cover the set  $\mathcal{F}$ . The centers of the balls need not belong to  $\mathcal{F}$ , but they should have finite norms. The entropy is the logarithm of the covering number.*

**Definition 6** *Given two functions  $l$  and  $u$ , the bracket  $[l, u]$  is the set of all functions  $f$  with  $l \leq f \leq u$ . An  $\epsilon$ -bracket is a bracket  $[l, u]$  with  $\|u - l\| < \epsilon$ . The bracketing number  $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$  is the minimum number of  $\epsilon$ -brackets needed to cover  $\mathcal{F}$ . The upper and lower bounds  $u$  and  $l$  need not belong to  $\mathcal{F}$  but are assumed to have finite norms. The entropy with bracketing is the logarithm of the bracketing number.*

**Definition 7** *An envelope function of a class  $\mathcal{F}$  is any function  $x \mapsto F(x)$  such that  $|f(x)| \leq F(x)$  for every  $x$  and  $f \in \mathcal{F}$ .*

**Proposition 8** *If the envelope  $F$  of  $\mathcal{F}$  is square integrable and*

$$\int_0^\infty \sup_Q \sqrt{\log \mathcal{N}(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon < \infty,$$

*then  $\mathcal{F}$  is  $P$ -Donsker for every  $P$ , that is,  $\mathcal{F}$  is a universal Donsker class. Here the supremum is taken over all finitely discrete probability measures, and the  $L_2(Q)$ -norm is defined as  $\|f\|_{Q,2} = (\int |f|^2)^{1/2}$ .*

**Proposition 9** *If  $\int_0^\infty \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$ , then  $\mathcal{F}$  is  $P$ -Donsker.*

From the learning theory perspective, however, the most interesting theorems are probably those relating the Donsker property to the VC-dimension. For example, if  $\mathcal{F}$  is a  $\{0, 1\}$ -valued class, then  $\mathcal{F}$  is *universal* Donsker if and only if its VC dimension is finite (Theorem 10.1.4 of Dudley (1999) provides a more general result involving Pollard’s entropy condition). As a corollary of their Proposition 3.1, Giné and Zinn (1991) show that under the Pollard’s entropy condition, the  $\{0, 1\}$ -valued class  $\mathcal{F}$  is in fact *uniform* Donsker. Finally, Rudelson and Vershynin extended these results to the real-valued case: a class  $\mathcal{F}$  is *uniform* Donsker if the square root of its scale-sensitive VC dimension is integrable.

### 3. Main Result

We now state the main result of this paper.

**Theorem 10** *Let  $\mathcal{F}$  be a  $P$ -Donsker class. For any sequence  $\xi(n) = o(n^{-1/2})$ ,*

$$\text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0.$$

The outer probability  $P^*$  above is due to measurability issues. Definitions and results on various types of convergence, as well as ways to deal with measurability issues arising in the proofs, are based on the rigorous book of van der Vaart and Wellner (1996).

The following corollary, whose proof is given in Appendix A, extends the above result to  $L_2$  (and thus  $L_1$ ) diameters.

**Corollary 11** *The result of Theorem 10 holds if the diameter is defined with respect to the  $L_2(P)$  norm.*

It is easy to verify that the dependence  $\xi(n) = o(n^{-1/2})$  of the tolerance, assumed in Theorem 10, is not improvable. In fact a simple example can show that if  $\xi(n) \asymp n^{-1/2}$  the set of  $\xi(n)$ -almost minimizers may not shrink in probability.

**Example 1** *Consider  $Z = \{x_1, x_2\}$  with  $x_1 \neq x_2$ , and  $P = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ . Moreover let  $\mathcal{F}$  be the set of functions  $\{f_1, f_2\}$ , with*

$$f_i(x) = \begin{cases} 0 & \text{if } x = x_i, \\ 1 & \text{otherwise.} \end{cases}$$

*Then it is clear that, given the finite sample  $S = (Z_1, \dots, Z_n)$ ,  $\mathcal{M}_S^\xi = \mathcal{F}$  (and hence  $\text{diam} \mathcal{M}_S^{\xi(n)} = 1$ ) whenever  $P_n f_1 - P_n f_2 = \frac{2}{n}|q - \mathbb{E}q| \leq \xi(n)$ , where  $q$  is the binomial random variable*

$$q = \#\{i | Z_i = x_1\}.$$

*Now since the variance of  $q$  is  $\frac{n}{4}$ , it is clear that*

$$\forall C > 0 \quad \Pr \left\{ |q - \mathbb{E}q| \leq Cn^{\frac{1}{2}} \right\} = \Omega(1),$$

*which shows that, if  $\xi(n) \asymp n^{-1/2}$ , with probability bounded away from zero,  $\text{diam} \mathcal{M}_S^{\xi(n)} = 1$ .*

The above example is very basic, yet provides important intuition. A class can contain two quite different functions with the smallest expectation, but it is unlikely that they both almost-minimize the empirical error to within  $o(n^{-1/2})$ . In fact, the above example suggests that the fluctuations of the difference in empirical performance of two functions is of the order  $n^{-1/2}$ . The extension of this result to more general function classes with possibly infinite number of expected minima is the main goal of Theorem 10.

Before diving into the proof of Theorem 10, let us state a few notions of stochastic convergence.

**Definition 12 (Definition 1.9.1 in van der Vaart and Wellner (1996))** *Let  $(Z, \mathcal{A}, P)$  be a probability space. Let  $Z_n, Z : Z \mapsto \mathcal{D}$  be arbitrary maps and  $(\mathcal{D}, d)$  be a metric space.*

- *$Z_n$  converges in outer probability to  $Z$  if  $d(Z_n, Z)^* \rightarrow 0$  in probability; this means that  $P(d(Z_n, Z)^* > \varepsilon) = P^*(d(Z_n, Z) > \varepsilon) \rightarrow 0$ , for every  $\varepsilon > 0$ , and is denoted by  $Z_n \xrightarrow{P^*} 0$ .*
- *$Z_n$  converges almost uniformly to  $Z$  if, for every  $\varepsilon > 0$ , there exists a measurable set  $A$  with  $P(A) \geq 1 - \varepsilon$  and  $d(Z_n, Z) \rightarrow 0$  uniformly on  $A$ ; this is denoted  $Z_n \xrightarrow{au} Z$ .*

The proof of Theorem 10 relies on the *almost sure representation theorem* (van der Vaart and Wellner, 1996, Theorem 1.10.4). Here we state the theorem applied to  $v_n$  and  $v$ .

**Proposition 13** *Suppose  $\mathcal{F}$  is  $P$ -Donsker. Let  $v_n : Z^n \mapsto \ell^\infty(\mathcal{F})$  be the empirical process. There exist a probability space  $(Z', \mathcal{A}', P')$  and maps  $v', v'_n : Z' \mapsto \ell^\infty(\mathcal{F})$  such that*

1.  $v'_n \xrightarrow{au} v'$ ,
2.  $\mathbb{E}^* f(v'_n) = \mathbb{E}^* f(v_n)$  for every bounded  $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$  for all  $n$ .

Lemma 14 is the main preliminary result used in the proof of Theorem 10 (and Theorem 17 in Section 5). We postpone its proof to Appendix A.

**Lemma 14** *Let  $v_n : Z^n \mapsto \ell^\infty(\mathcal{F})$  be the empirical process. Fix  $n$  and assume that there exist a probability space  $(Z', \mathcal{A}', P')$  and a map  $v'_n : Z' \mapsto \ell^\infty(\mathcal{F})$  such that  $\mathbb{E}^* f(v'_n) = \mathbb{E}^* f(v_n)$  for every bounded  $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ . Let  $v'$  be a  $P$ -Brownian bridge defined on  $(Z', \mathcal{A}', P')$ . Fix  $C > 0$ ,  $\varepsilon = \min(C^3/128, C/4)$  and suppose  $\delta \geq \xi\sqrt{n}$  for a given  $\xi > 0$ . Then, if  $\mathcal{F}$  is  $P$ -Donsker, the following inequality holds*

$$\Pr^* \left( \text{diam} \mathcal{M}_S^\xi > C \right) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)^2 \left( \frac{128\delta}{C^3} + \Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| \geq \delta/2 \right) \right).$$

We are now ready to prove the main result of this section.

**Proof** [Theorem 10] Lemma 1.9.3 in van der Vaart and Wellner (1996) shows that when the limiting process is Borel measurable, almost uniform convergence implies convergence in outer probability. Therefore, the first implication of Proposition 13 states that for any  $\delta > 0$

$$\Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| > \delta \right) \rightarrow 0.$$

By Lemma 14,

$$\Pr^* \left( \text{diam} \mathcal{M}_S^{\xi(n)} > C \right) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)^2 \left( \frac{128\delta}{C^3} + \Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| \geq \delta/2 \right) \right)$$

for any  $C > 0$ ,  $\varepsilon = \min(C^3/128, C/4)$ , and any  $\delta \geq \xi(n)\sqrt{n}$ . Since  $\xi(n) = o(n^{-1/2})$ ,  $\delta$  can be chosen arbitrarily small, and so  $\Pr^* \left( \text{diam} \mathcal{M}_S^{\xi(n)} > C \right) \rightarrow 0$ . ■

#### 4. Stability of almost-ERM

The main result of this section, Corollary 15, shows  $L_2$ -stability of almost-ERM on Donsker classes. It implies that, in probability, the  $L_2$  (and thus  $L_1$ ) distance between almost-minimizers on similar training sets (with  $o(\sqrt{n})$  changes) goes to zero when  $n$  tends to infinity.

This result provides a partial answer to the questions raised in the machine learning literature by Kutin and Niyogi (2002); Mukherjee et al. (2006): is it true that when one point is added to the training set, the ERM algorithm is less and less likely to jump to a far (in the  $L_1$  sense) hypothesis? In fact, since binary-valued function classes are uniform Donsker if and only if the VC dimension is finite, Corollary 15 proves that almost-ERM over binary VC classes possesses  $L_1$ -stability. For the

real-valued classes, uniform Glivenko-Cantelli property is weaker than uniform Donsker property, and therefore it remains unclear if almost-ERM over uGC but not uniform Donsker classes is stable in the  $L_1$  sense.

Use of  $L_1$ -stability goes back to Devroye and Wagner (1979), who showed that this stability is sufficient to bound the difference between the leave-one-out error and the expected error of a learning algorithm. In particular, Devroye and Wagner show that nearest-neighbor rules possess  $L_1$ -stability (see also Devroye et al., 1996). Our Corollary 15 implies  $L_1$ -stability of ERM (or almost-ERM) algorithms on Donsker classes.

In the following  $[n]$  denotes the set  $\{1, 2, \dots, n\}$  and  $A \triangle B$  is the symmetric difference of sets  $A$  and  $B$ .

**Corollary 15** *Assume  $\mathcal{F}$  is  $P$ -Donsker and uniformly bounded with envelope  $F \equiv 1$ . For  $I \subset \mathbb{N}$ , define  $S(I) = (Z_i)_{i \in I}$ . Let  $I_n \subset \mathbb{N}$  such that  $M_n := |I_n \triangle [n]| = o(n^{1/2})$ . Suppose  $f_n \in \mathcal{M}_{S([n])}^{\xi(n)}$  and  $f'_n \in \mathcal{M}_{S(I_n)}^{\xi'(n)}$  for some  $\xi(n) = o(n^{-1/2})$  and  $\xi'(n) = o(n^{-1/2})$ . Then*

$$\|f_n - f'_n\| \xrightarrow{P^*} 0.$$

The norm  $\|\cdot\|$  can be replaced by  $L_2(P)$  or  $L_1(P)$  norm.

**Proof** It is enough to show that  $f'_n \in \mathcal{M}_{S([n])}^{\xi''(n)}$  for some  $\xi''(n) = o(n^{-1/2})$  and result follows from Theorem 10.

$$\begin{aligned} \frac{1}{n} \sum_{i \in [n]} f'_n(Z_i) &\leq \frac{M_n}{n} + \frac{1}{n} \sum_{i \in I_n} f'_n(Z_i) \\ &\leq \frac{M_n}{n} + \frac{|I_n|}{n} \left( \xi'(n) + \inf_{g \in \mathcal{F}} \frac{1}{|I_n|} \sum_{i \in I_n} g(Z_i) \right) \\ &\leq \frac{M_n}{n} + \frac{|I_n|}{n} \xi'(n) + \frac{1}{n} \sum_{i \in I_n} f_n(Z_i) \\ &\leq 2 \frac{M_n}{n} + \frac{|I_n|}{n} \xi'(n) + \frac{1}{n} \sum_{i \in [n]} f_n(Z_i) \\ &\leq 2 \frac{M_n}{n} + \frac{|I_n|}{n} \xi'(n) + \xi(n) + \inf_{g \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} g(Z_i). \end{aligned}$$

Define

$$\xi''(n) := 2 \frac{M_n}{n} + \frac{|I_n|}{n} \xi'(n) + \xi(n).$$

Because  $M_n = o(n^{\frac{1}{2}})$ , it follows that  $\xi''(n) = o(n^{-1/2})$ . Corollary 11 implies convergence in  $L_2(P)$ , and, therefore, in  $L_1(P)$  norm.  $\blacksquare$

## 5. Rates of Decay of $\text{diam} \mathcal{M}_S^{\xi(n)}$

The statement of Lemma 14 reveals that the rate of the decay of the diameter  $\text{diam} \mathcal{M}_S^{\xi(n)}$  is related to the rate at which  $\Pr^*(\sup_{\mathcal{F}} |v - v_n| \geq \delta) \rightarrow 0$  for a fixed  $\delta$ . A number of papers studied this

rate of convergence, and here we refer to the notion of *Komlos-Major-Tusnady class* (KMT class), as defined by Koltchinskii (1994). Let  $v'_n : Z^n \mapsto \ell^\infty(\mathcal{F})$  be the empirical process defined on the probability space  $(Z', \mathcal{A}', P')$ .

**Definition 16**  $\mathcal{F}$  is called a *Komlos-Major-Tusnady class* with respect to  $P$  and with the rate of convergence  $\tau_n$  ( $\mathcal{F} \in KMT(P; \tau_n)$ ) if  $\mathcal{F}$  is  $P$ -pregaussian and for each  $n \geq 1$  there is a version  $v^{(n)}$  of a  $P$ -Brownian bridge defined on  $(Z', \mathcal{A}', P')$  such that for all  $t > 0$ ,

$$\Pr^* \left( \sup_{\mathcal{F}} |v^{(n)} - v'_n| \geq \tau_n(t + K \log n) \right) \leq \Lambda e^{-\theta t}$$

where  $K > 0$ ,  $\Lambda > 0$  and  $\theta > 0$  are constants, depending only on  $\mathcal{F}$ .

Sufficient conditions for a class to be  $KMT(P; n^{-\alpha})$  have been investigated in the literature; some results of this type can be found in Koltchinskii (1994); Rio (1993) and Dudley (2002), Section 9.5(B).

The following theorem shows that for KMT classes fulfilling a suitable entropy condition, it is possible to give explicit rates of decay for the diameter of ERM almost-minimizers.

**Theorem 17** Assume  $\mathcal{F}$  is  $P$ -Donsker and  $\mathcal{F} \in KMT(P; n^{-\alpha})$  for some  $\alpha > 0$ . Assume  $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \left(\frac{A}{\epsilon}\right)^V$  for some constants  $A, V > 0$ . Let  $\xi(n)\sqrt{n} = o(n^{-\eta})$ ,  $\eta > 0$ . Then

$$n^\gamma \text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$$

for any  $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$ .

**Proof** The result of Lemma 14 is stated for a fixed  $n$ . We now choose  $C$ ,  $\xi$ , and  $\delta$  depending on  $n$  as follows. Let  $C(n) = Bn^{-\gamma}$ , where  $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$  and  $B > 0$  is an arbitrary constant. Let  $\xi = \xi(n)$ . Let  $\delta(n) = n^{-\beta}$ , where  $\beta = \frac{1}{2}(\min(\alpha, \eta) + 3(2V + 1)\gamma)$ . When  $\beta$  is defined this way, we have

$$\min(\alpha, \gamma) > \beta > 3(2V + 1)\gamma$$

because  $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$  by assumption. In particular,  $\beta < \eta$  and, hence, eventually  $\delta(n) > \xi(n)\sqrt{n} = o(n^{-\eta})$ .

Since  $C(n)$  decays to zero and  $\epsilon(n) = \min(C(n)^3/128, C(n)/4)$ , eventually  $\epsilon(n) = C(n)^3/128 = n^{-3\gamma}B^3/128$ .

Since  $\mathcal{F} \in KMT(P; n^{-\alpha})$ ,

$$\Pr^* \left( \sup_{\mathcal{F}} |v^{(n)} - v_n| \geq n^{-\alpha}(t + K \log n) \right) \leq \Lambda e^{-\theta t}$$

for any  $t > 0$ , choosing  $t = n^\alpha \delta(n)/2 - K \log n$  we obtain

$$\Pr^* \left( \sup_{\mathcal{F}} |v^{(n)} - v_n| \geq \delta(n)/2 \right) \leq \Lambda e^{-\theta(n^{\alpha-\beta}/2 - K \log n)}.$$

Lemma 14 then implies

$$\begin{aligned} \Pr^* \left( \text{diam} \mathcal{M}_S^{\xi(n)} > C(n) \right) &\leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)^2 \left( \frac{128\delta}{C(n)^3} + \Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| \geq \delta/2 \right) \right) \\ &\leq \left( \frac{128A}{B^3} n^{3\gamma} \right)^{2V} \frac{128}{B^3} n^{-\beta} n^{3\gamma} + \left( \frac{128A}{B^3} n^{3\gamma} \right)^{2V} \Lambda e^{-\theta(n^{\alpha-\beta}/2 - K \log n)} \\ &= \left( \frac{128A}{B^3} \right)^{2V} \frac{128}{B^3} n^{3\gamma(2V+1)-\beta} + \Lambda \left( \frac{128A}{B^3} \right)^{2V} n^{k\theta+6\gamma V} e^{-\frac{\theta}{2}n^{\alpha-\beta}}. \end{aligned}$$

Since  $\alpha > \beta > 3\gamma(2V + 1)$ , both terms above go to zero, that is,

$$\Pr^* \left( n^\gamma \text{diam} \mathcal{M}_S^{\xi(n)} > B \right) \rightarrow 0 \text{ for any } B > 0.$$

■

The entropy condition in Theorem 17 is clearly verified by VC-subgraph classes of dimension  $V$ . In fact, since  $L_2$  norm dominates  $\|\cdot\|$  seminorm, upper bounds on  $L_2$  covering numbers of VC-subgraph classes induce analogous bounds on  $\|\cdot\|$  covering numbers. Corollary 18 is an application of Theorem 17 to this important family of classes. It follows in a straight-forward way from the remark above.

**Corollary 18** *Assume  $\mathcal{F}$  is a VC-subgraph class with VC-dimension  $V$ , and for some  $\alpha > 0$   $\mathcal{F} \in \text{KMT}(P, n^{-\alpha})$ . Let  $\xi(n)\sqrt{n} = o(n^{-\eta})$ ,  $\eta > 0$ . Then*

$$n^\gamma \text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$$

for any  $\gamma < \frac{1}{3(2V+1)} \min(\alpha, \eta)$ .

## 6. Expected Error Stability of almost-ERM

In the previous section, we proved bounds on the rate of decay of the diameter of almost-minimizers. In this section, we show that given such a bound, as well as some additional conditions on the class, the differences between *expected errors* of almost-minimizers decay faster than  $n^{-1/2}$ . This implies a form of *strong expected error stability* for ERM.

The proof of Theorem 20 relies on the following ratio inequality of Pollard (1995).

**Proposition 19** *Let  $\mathcal{G}$  be a uniformly bounded function class with the envelope function  $G \equiv 2$ . Assume  $\mathcal{N}(\gamma, \mathcal{G}) = \sup_Q \mathcal{N}(2\gamma, \mathcal{G}, L_1(Q)) < \infty$  for  $0 < \gamma \leq 1$  and  $Q$  ranging over all discrete probability measures. Then*

$$\Pr^* \left( \sup_{f \in \mathcal{G}} \frac{|P_n f - P f|}{\varepsilon(P_n |f| + P|f|) + 5\gamma} > 26 \right) \leq 32 \mathcal{N}(\gamma, \mathcal{G}) \exp(-n\varepsilon\gamma).$$

The next theorem gives explicit rates for expected error stability of ERM over VC-subgraph classes fulfilling a KMT type condition.

**Theorem 20** *If  $\mathcal{F}$  is a VC-subgraph class with VC-dimension  $V$ ,  $\mathcal{F} \in KMT(P; n^{-\alpha})$  and  $\sqrt{n}\xi(n) = o(n^{-\eta})$ , then for any  $\kappa < \min\left(\frac{1}{\delta(2V+1)} \min(\alpha, \eta), 1/2\right)$*

$$n^{1/2+\kappa} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \xrightarrow{P^*} 0.$$

## 7. Conclusions

We presented some new results establishing stability properties of ERM over certain classes of functions. This study was motivated by the question, raised by some recent papers, of  $L_1$ -stability of ERM under perturbations of a single sample (Mukherjee et al., 2006; Kutin and Niyogi, 2002; Rakhlin et al., 2005). We gave a partially positive answer to this question, proving that, in fact, ERM over Donsker classes fulfills  $L_2$ -stability (and hence also  $L_1$ -stability) under perturbations of  $o(n^{\frac{1}{2}})$  among the  $n$  samples of the training set. This property follows directly from the main result of the paper which shows decay (in probability) of the diameter of the set of solutions of almost-ERM with tolerance function  $\xi(n) = o(n^{-\frac{1}{2}})$ . We stress that for classification problems (i.e., for binary-valued functions) no generality is lost in assuming the Donsker property, since for ERM to be a sound algorithm, the equivalent Glivenko-Cantelli property has to be assumed anyway. On the other hand, in the real-valued case many complexity-based characterizations of Donsker property are available in the literature.

In the perspective of possible algorithmic applications, we analyzed some additional assumptions implying uniform rates on the decay of the  $L_1$  diameter of almost-minimizers. It turned out that an explicit rate of this type can be given for VC-subgraph classes satisfying a suitable Komlos-Major-Tusnady type condition. For this condition, many independent characterizations are known.

Finally, using a suitable ratio inequality we showed how  $L_1$ -stability results can induce strong forms of expected error stability, providing a further insight into the behavior of the Empirical Risk Minimization algorithm.

Results of this paper can be used to analyze stability of a class of clustering algorithms by casting them in the empirical risk minimization framework (see Rakhlin and Caponnetto, 2006).

Algorithmic implications of our results would require further investigation. For example, in the context of on-line learning, when a point is added to the training set, with high probability one would only have to search for empirical minimizers in a small  $L_1$ -ball around the current hypothesis, which might be a tractable problem. Moreover,  $L_1$ -stability might have consequences for computational complexity of ERM. While it has been shown that ERM is NP-hard even for simple function classes (see for example, Ben-David et al., 2003), our results could allow more optimistic average-case analysis.

## Acknowledgments

We would like to thank S. Mukherjee, T. Poggio and S. Smale for useful discussions and suggestions.

This report describes research done at the Center for Biological & 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), as well as in the Dipartimento di Informatica e Scienze dell'Informazione (DISI) at University of Genoa, Italy. 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/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 (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation. This research has also been partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

## Appendix A.

In this appendix we derive some results presented in Section 3. In particular, Lemma 14, which was used in the proof of Theorem 10, and Corollary 11. Let us start with some technical Lemmas.

**Lemma 21** *Let  $f_0, f_1 \in \mathcal{F}$ ,  $\|f_0 - f_1\| \geq C/2$ ,  $\|f_1\| \leq \|f_0\|$ . Let  $h : \mathcal{F} \rightarrow \mathbb{R}$  be defined as  $h(f') = \frac{\langle f', f_0 \rangle}{\|f_0\|^2}$ . Then for any  $\varepsilon \leq \frac{C^3}{128}$*

$$\inf_{\mathcal{B}(f_0, \varepsilon)} h - \sup_{\mathcal{B}(f_1, \varepsilon)} h \geq \frac{C^2}{16}.$$

### Proof

$$\begin{aligned} \Delta &:= \inf_{\mathcal{B}(f_0, \varepsilon)} h - \sup_{\mathcal{B}(f_1, \varepsilon)} h \\ &= h(f_0) - h(f_1) + \inf\{h(f' - f_0) + h(f_1 - f'') \mid f' \in \mathcal{B}(f_0, \varepsilon), f'' \in \mathcal{B}(f_1, \varepsilon)\} \\ &\geq h(f_0) - h(f_1) - \frac{2\varepsilon}{\|f_0\|} \geq h(f_0) - h(f_1) - \frac{8\varepsilon}{C}, \end{aligned}$$

since  $\|f_0\| \geq C/4$ .

Finally

$$2\langle f_0 - f_1, f_0 \rangle = \|f_0 - f_1\|^2 - \|f_1\|^2 + \|f_0\|^2 \geq \|f_0 - f_1\|^2 \geq \frac{C^2}{4},$$

then

$$h(f_0) - h(f_1) \geq \frac{C^2}{8\|f_0\|^2} \geq \frac{C^2}{8},$$

which proves that

$$\Delta \geq \frac{C^2}{8} - \frac{8\varepsilon}{C} \geq \frac{C^2}{16}.$$

■

The following Lemma is an adaptation of Lemma 2.3 of Kim and Pollard (1990).

**Lemma 22** *Let  $f_0, f_1, h$  be defined as in Lemma 21. Suppose  $\varepsilon \leq \frac{C^3}{128}$ . Let  $v_\mu$  be a Gaussian process on  $\mathcal{F}$  with mean  $\mu$  and covariance  $\text{cov}(v_\mu(f), v_\mu(f')) = \langle f, f' \rangle$ . Then for all  $\delta > 0$*

$$\Pr^* \left( \left| \sup_{\mathcal{B}(f_0, \varepsilon)} v_\mu - \sup_{\mathcal{B}(f_1, \varepsilon)} v_\mu \right| \leq \delta \right) \leq \frac{64\delta}{C^3}.$$

**Proof** Define the Gaussian process  $Y(\cdot) = v_\mu(\cdot) - h(\cdot)v_\mu(f_0)$ . Since  $\text{cov}(Y(f'), v_\mu(f_0)) = \langle f', f_0 \rangle - h(f')\|f_0\|^2 = 0$ ,  $v_\mu(f_0)$  and  $Y(\cdot)$  are independent.

We now reason conditionally with respect to  $Y(\cdot)$ . Define

$$\Gamma_i(z) = \sup_{\mathcal{B}(f_i, \varepsilon)} \{Y(\cdot) + h(\cdot)z\} \quad \text{with } i = 0, 1.$$

Notice that

$$\Pr^* \left( \left| \sup_{\mathcal{B}(f_0, \varepsilon)} v_\mu - \sup_{\mathcal{B}(f_1, \varepsilon)} v_\mu \right| \leq \delta \mid Y \right) = \Pr^* (|\Gamma_0(v_\mu(f_0)) - \Gamma_1(v_\mu(f_0))| \leq \delta).$$

Moreover  $\Gamma_0$  and  $\Gamma_1$  are convex and

$$\inf \partial_- \Gamma_0 - \sup \partial_+ \Gamma_1 \geq \inf_{\mathcal{B}(f_0, \varepsilon)} h - \sup_{\mathcal{B}(f_1, \varepsilon)} h \geq \frac{C^2}{16},$$

by Lemma 21. Then  $\Gamma_0 = \Gamma_1$  in a single point  $z_0$  and

$$\Pr^* (|\Gamma_0(v_\mu(f_0)) - \Gamma_1(v_\mu(f_0))| \leq \delta) \leq \Pr^* (v_\mu(f_0) \in [z_0 - \Delta, z_0 + \Delta]),$$

with  $\Delta = 16\delta/C^2$ .

Furthermore,

$$\Pr^* (v_\mu(f_0) \in [z_0 - \Delta, z_0 + \Delta]) \leq \frac{32\delta}{C^2 \sqrt{2\pi \text{var}(v_\mu(f_0))}},$$

and  $\text{var}(v_\mu(f_0)) = \|f_0\|^2 \geq C^2/16$ , which completes the proof. ■

The reasoning in the proof of the next lemma goes as follows. We consider a finite cover of  $\mathcal{F}$ . Pick any two almost-minimizers which are far apart. They belong to two covering balls with centers far apart. Because the two almost-minimizers belong to these balls, the infima of the empirical risks over these two balls are close. This is translated into the event that the suprema of the shifted empirical process over these two balls are close. By looking at the Gaussian limit process, we are able to exploit the covariance structure to show that the suprema of the Gaussian process over balls with centers far apart are unlikely to be close.

**Proof** [Lemma 14]

Consider the  $\varepsilon$ -covering  $\{f_i | i = 1, \dots, \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)\}$ . Such a covering exists because  $\mathcal{F}$  is totally bounded in  $\|\cdot\|$  norm (see page 89, van der Vaart and Wellner, 1996). For any  $f, f' \in \mathcal{M}_S^\xi$  s.t.  $\|f - f'\| > C$ , there exist  $k$  and  $l$  such that  $\|f - f_k\| \leq \varepsilon \leq C/4$ ,  $\|f' - f_l\| \leq \varepsilon \leq C/4$ . By triangle inequality it follows that  $\|f_k - f_l\| \geq C/2$ .

Moreover

$$\inf_{\mathcal{F}} P_n \leq \inf_{\mathcal{B}(f_k, \varepsilon)} P_n \leq P_n f \leq \inf_{\mathcal{F}} P_n + \xi$$

and

$$\inf_{\mathcal{F}} P_n \leq \inf_{\mathcal{B}(f_l, \varepsilon)} P_n \leq P_n f' \leq \inf_{\mathcal{F}} P_n + \xi.$$

Therefore,

$$\left| \inf_{\mathcal{B}(f_k, \varepsilon)} P_n - \inf_{\mathcal{B}(f_l, \varepsilon)} P_n \right| \leq \xi.$$

The last relation can be restated in terms of the empirical process  $v_n$ :

$$\left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v_n - \sqrt{n}P\} \right| \leq \xi \sqrt{n} \leq \delta.$$

$$\begin{aligned} \Pr^* \left( \text{diam} \mathcal{M}_S^\xi > C \right) &= \Pr^* \left( \exists f, f' \in \mathcal{M}_S^\xi, \|f - f'\| > C \right) \leq \\ \Pr^* \left( \exists l, k \text{ s.t. } \|f_k - f_l\| \geq C/2, \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v_n - \sqrt{n}P\} \right| \leq \delta \right). \end{aligned}$$

By union bound

$$\begin{aligned} \Pr^* \left( \text{diam} \mathcal{M}_S^\xi > C \right) &\leq \sum_{\substack{k, l=1 \\ \|f_k - f_l\| \geq C/2}}^{\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)} \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v_n - \sqrt{n}P\} \right| \leq \delta \right). \end{aligned}$$

We now want to bound the terms in the sum above. Assuming without loss of generality that  $\|f_k\| \geq \|f_l\|$ , we obtain

$$\begin{aligned}
 & \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v_n - \sqrt{n}P\} \right| \leq \delta \right) \\
 &= \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v'_n - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v'_n - \sqrt{n}P\} \right| \leq \delta \right) \\
 &= \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v' - \sqrt{n}P + v' - v'_n\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v' - \sqrt{n}P + v' - v'_n\} \right| \leq \delta \right) \\
 &\leq \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v' - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v' - \sqrt{n}P\} \right| - 2 \sup_{\mathcal{F}} |v'_n - v'| \right| \leq \delta \right) \\
 &\leq \Pr^* \left( 2 \sup_{\mathcal{F}} |v'_n - v'| \geq \delta \vee \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v' - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v' - \sqrt{n}P\} \right| \leq 2\delta \right) \\
 &\leq \Pr^* \left( \left| \sup_{\mathcal{B}(f_k, \varepsilon)} \{-v' - \sqrt{n}P\} - \sup_{\mathcal{B}(f_l, \varepsilon)} \{-v' - \sqrt{n}P\} \right| \leq 2\delta \right) + \Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| \geq \delta/2 \right) \\
 &\leq \frac{128\delta}{C^3} + \Pr^* \left( \sup_{\mathcal{F}} |v'_n - v'| \geq \delta/2 \right),
 \end{aligned}$$

where the first inequality results from a union bound argument while the second one results from Lemma 22 noticing that  $-v' - \sqrt{n}P$  is a Gaussian process with covariance  $\langle f, f' \rangle$  and mean  $-\sqrt{n}P$ , and since by construction  $\varepsilon \leq C^3/128$ .

Finally, the claimed result follows from the two last relations.  $\blacksquare$

We now prove, Corollary 11, the extension of Theorem 10 to  $L_2$  diameters. The proof relies on the observation that a  $P$ -Donsker class is also Glivenko-Cantelli.

**Proof** [Corollary 11] Note that

$$\|f - f'\|_{L_2}^2 = \|f - f'\|^2 + (P(f - f'))^2.$$

The expected errors of almost-minimizers over a Glivenko-Cantelli (and therefore over Donsker) class are close because empirical averages uniformly converge to the expectations.

$$\begin{aligned}
 & \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad \|f - f'\|_{L_2} > C \right) \\
 &\leq \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad |Pf - Pf'| > C/\sqrt{2} \right) + \Pr^* \left( \text{diam} \mathcal{M}_S^{\xi(n)} > C/\sqrt{2} \right).
 \end{aligned}$$

The first term can be bounded as

$$\begin{aligned}
 & \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad |Pf - Pf'| > C/\sqrt{2} \right) \\
 &\leq \Pr^* \left( \exists f, f' \in \mathcal{F}, |P_n f - P_n f'| \leq \xi(n), |Pf - Pf'| > C/\sqrt{2} \right) \\
 &\leq \Pr^* \left( \sup_{f, f' \in \mathcal{F}} |(P_n - P)(f - f')| > |C/\sqrt{2} - \xi(n)| \right)
 \end{aligned}$$

which goes to 0 because the class  $\{f - f' | f, f' \in \mathcal{F}\}$  is Glivenko-Cantelli. The second term goes to 0 by Theorem 10.  $\blacksquare$

## Appendix B.

In this appendix we report the proof of Theorem 20 stated in Section 6. We first need to derive a preliminary lemma.

**Lemma 23** *Let  $\mathcal{F}$  be  $P$ -Donsker class with envelope function  $G \equiv 1$ . Assume  $\mathcal{N}(\gamma, \mathcal{F}) = \sup_Q \mathcal{N}(\gamma, \mathcal{F}, L_1(Q)) < \infty$  for  $0 < \gamma \leq 1$  and  $Q$  ranging over all discrete probability measures. Let  $\mathcal{M}_S^{\xi(n)}$  be defined as above with  $\xi(n) = o(n^{-1/2})$  and assume that for some sequence of positive numbers  $\lambda(n) = o(n^{1/2})$*

$$\lambda(n) \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} P|f - f'| \xrightarrow{P^*} 0. \quad (1)$$

Suppose further that for some  $1/2 < \rho < 1$

$$\lambda(n)^{2\rho-1} - \log \mathcal{N}(\frac{1}{2}n^{-1/2}\lambda(n)^{\rho-1}, \mathcal{F}) \rightarrow +\infty. \quad (2)$$

Then

$$\Pr^* \left( \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \leq \sqrt{n}\xi(n) + 131\lambda(n)^{\rho-1} \right) \rightarrow 0.$$

**Proof** Define  $\mathcal{G} = \{f - f' : f, f' \in \mathcal{F}\}$  and  $\mathcal{G}' = \{|f - f'| : f, f' \in \mathcal{F}\}$ . By Example 2.10.7 of van der Vaart and Wellner (1996),  $\mathcal{G} = (\mathcal{F}) + (-\mathcal{F})$  and  $\mathcal{G}' = |\mathcal{G}| \subseteq (\mathcal{G} \wedge 0) \vee (-\mathcal{G} \wedge 0)$  are Donsker as well. Moreover,  $\mathcal{N}(2\gamma, \mathcal{G}) \leq \mathcal{N}(\gamma, \mathcal{F})^2$  and the envelope of  $\mathcal{G}$  is  $G \equiv 2$ . Applying Proposition 19 to the class  $\mathcal{G}$ , we obtain

$$\Pr^* \left( \sup_{f, f' \in \mathcal{F}} \frac{|P_n(f - f') - P(f - f')|}{\varepsilon(P_n|f - f'| + P|f - f'|) + 5\gamma} > 26 \right) \leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\varepsilon\gamma).$$

The inequality therefore holds if the sup is taken over a smaller (random) subclass  $\mathcal{M}_S^{\xi(n)}$ .

$$\Pr^* \left( \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} \frac{|P(f - f')| - \xi(n)}{\varepsilon(P_n|f - f'| + P|f - f'|) + 5\gamma} > 26 \right) \leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\varepsilon\gamma).$$

$$\text{Since } \sup_x \frac{A(x)}{B(x)} \geq \sup_x \frac{A(x)}{\sup_x B(x)} = \frac{\sup_x A(x)}{\sup_x B(x)},$$

$$\Pr^* \left( \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} (|P(f - f')| - \xi(n)) > 26 \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} (\varepsilon(P_n|f - f'| + P|f - f'|) + 5\gamma) \right) \quad (3)$$

$$\leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\varepsilon\gamma).$$

By assumption,

$$\lambda(n) \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} P|f - f'| \xrightarrow{P^*} 0.$$

Because  $\mathcal{G}'$  is Donsker and  $\lambda(n) = o(n^{1/2})$ ,

$$\lambda(n) \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P_n|f - f'| - P|f - f'| \xrightarrow{P^*} 0.$$

Thus,

$$\lambda(n) \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} P_n|f - f'| + P|f - f'| \xrightarrow{P^*} 0.$$

Letting  $\varepsilon = \varepsilon(n) := n^{-1/2}\lambda(n)^\rho$ , this implies that for any  $\delta > 0$ , there exist  $N_\delta$  such that for all  $n > N_\delta$ ,

$$\Pr^* \left( \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} 26\varepsilon(n) (P_n|f - f'| + P|f - f'|) > \lambda(n)^{\rho-1} \right) < \delta.$$

Now, choose  $\gamma = \gamma(n) := n^{-1/2}\lambda(n)^{\rho-1}$  (note that since  $\rho < 1$ , eventually  $0 < \gamma(n) < 1$ ), the last inequality can be rewritten in the following form

$$\Pr^* \left( \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} 26 (\varepsilon(n) (P_n|f - f'| + P|f - f'|) + 5\gamma(n)) > 131\lambda(n)^{\rho-1} \right) < \delta.$$

Combining the relation above with Equation 3,

$$\begin{aligned} & \Pr^* \left( \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \leq \sqrt{n}\xi(n) + 131\lambda(n)^{\rho-1} \right) \\ & \geq 1 - 32\mathcal{N} \left( \frac{1}{2}n^{-1/2}\lambda(n)^{\rho-1}, \mathcal{F} \right)^2 \exp(-\lambda(n)^{2\rho-1}) - \delta. \end{aligned}$$

The result follows by the assumption on the entropy and by arbitrariness of  $\delta$ . ■

We are now ready to prove Theorem 20.

**Proof** [Theorem 20] By Corollary 18,

$$n^\gamma \text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$$

for any  $\gamma < \min \left( \frac{1}{3(2V+1)} \min(\alpha, \eta), 1/2 \right)$ . Let  $\lambda(n) = n^\gamma$  and note that  $\lambda(n) = o(\sqrt{n})$ , which is a condition in Lemma 23. First, we show that a power decay of the  $\|\cdot\|$  diameter implies the same rate

of decay of the  $L_1$  diameter, hence verifying condition (1) in Lemma 23. Proof of this fact is very similar to the proof of Corollary 11, except that  $C$  is replaced by  $C\lambda(n)^{-1}$ .

$$\begin{aligned} & \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad \|f - f'\|_{L_2} > C\lambda(n)^{-1} \right) \\ & \leq \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad |Pf - Pf'| > C\lambda(n)^{-1}/\sqrt{2} \right) \\ & + \Pr^* \left( \text{diam} \mathcal{M}_S^{\xi(n)} > C\lambda(n)^{-1}/\sqrt{2} \right). \end{aligned}$$

The second term goes to zero since  $\lambda(n)\text{diam} \mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0$ . Moreover, since  $\lambda(n) = o(\sqrt{n})$  and  $\mathcal{G}$  is Donsker, the first term can be bounded as

$$\begin{aligned} & \Pr^* \left( \exists f, f' \in \mathcal{M}_S^{\xi(n)} \quad \text{s.t.} \quad |Pf - Pf'| > C\lambda(n)^{-1}/\sqrt{2} \right) \\ & \leq \Pr^* \left( \exists f, f' \in \mathcal{F}, |P_n f - P_n f'| \leq \xi(n), |Pf - Pf'| > C\lambda(n)^{-1}/\sqrt{2} \right) \\ & \leq \Pr^* \left( \sup_{f, f' \in \mathcal{F}} |P(f - f') - P_n(f - f')| > \left| \frac{C}{\sqrt{2}} \lambda(n)^{-1} - \xi(n) \right| \right) \\ & = \Pr^* \left( \lambda(n) \sup_{g \in \mathcal{G}} |Pg - P_n g| > \left| \frac{C}{\sqrt{2}} - \xi(n)\lambda(n) \right| \right) \rightarrow 0, \end{aligned}$$

proving condition (1) in Lemma 23.

We now verify condition (2) in Lemma 23. Since  $\mathcal{F}$  is a VC-subgraph class of dimension  $V$ , its entropy numbers  $\log \mathcal{N}(\varepsilon, \mathcal{F})$  behave like  $V \log \frac{A}{\varepsilon}$  ( $A$  is a constant), that is

$$\log \mathcal{N} \left( \frac{1}{2} n^{-1/2} \lambda(n)^{\rho-1}, \mathcal{F} \right) \leq \text{const} + \frac{1}{2} V \log n + (1 - \rho) V \log \lambda(n).$$

Condition (2) of Lemma 23 will therefore hold whenever  $\lambda(n)$  grows faster than  $(\log n)^{\frac{1}{2\rho-1}}$ , for any  $1 > \rho > \frac{1}{2}$ . In our problem,  $\lambda(n)$  grows polynomially, so condition (2) is satisfied for any fixed  $1 > \rho > 1/2$ .

Hence, by Lemma 23

$$\Pr^* \left( \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \leq \sqrt{n} \xi(n) + 131 n^{\gamma(\rho-1)} \right) \rightarrow 0.$$

Choose any  $0 < \kappa < \gamma/2$  and multiply both sides of the inequality by  $n^\kappa$ . We obtain

$$\Pr^* \left( n^\kappa \sqrt{n} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \leq \sqrt{n} \xi(n) n^\kappa + 131 n^{\gamma(\rho-1)+\kappa} \right) \rightarrow 0.$$

Now fix a  $\rho$  such that  $1/2 < \rho < 1 - \kappa/\gamma$ . Because  $0 < \kappa < \gamma/2$ , there is always such a choice of  $\rho$ . Furthermore,  $1 > \rho > 1/2$  so that the above convergence holds. Our choice of  $\rho$  implies that  $\gamma(\rho - 1) + \kappa < 0$  and so  $n^{\gamma(\rho-1)+\kappa} \rightarrow 0$ . Since  $\kappa < \gamma/2 < \eta$ ,  $\sqrt{n} \xi(n) n^\kappa \rightarrow 0$ . Hence,

$$n^{1/2+\kappa} \sup_{f, f' \in \mathcal{M}_S^{\xi(n)}} |P(f - f')| \xrightarrow{P^*} 0$$

for any  $\kappa < \min\left(\frac{1}{6(2V+1)} \min(\alpha, \eta), 1/2\right)$ . ■

## References

- P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.
- P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 270–284. Springer, 2004.
- S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.
- O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.
- L. P. Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.
- R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.
- R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.
- E. Giné and J. Zinn. Gaussian characterization of uniform Donsker classes of functions. *The Annals of Probability*, 19:758–782, 1991.
- J. Kim and D. Pollard. Cube root asymptotics. *Annals of Statistics*, 18:191–219, 1990.
- V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 2006. To appear.
- V. Koltchinskii. Komlós-Major-Tusnády approximation for the general empirical process and Haar expansion of classes of functions. *Journal of Theoretical Probability*, 7:73–118, 1994.
- S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 275–28, San Francisco, CA, 2002. Morgan Kaufmann.
- W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.
- S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

- T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.
- D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.
- A. Rakhlin and A. Caponnetto. Stability of  $k$ -means clustering. In *Proceedings of Neural Information Processing Systems Conference*, 2006. To appear.
- A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–417, 2005.
- E. Rio. Strong approximation for set-indexed partial sum processes via KMT constructions I. *The Annals of Probability*, 21(2):759–790, 1993.
- M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*. To appear.
- S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
- A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.
- V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Th. Prob. and its Applications*, 17(2):264–280, 1971.
- V. N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.