Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Stability Properties of Empirical Risk Minimization over Donsker Classes

Andrea Caponnetto CAPONNET@MIT.EDU
Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, MA, 02139, USA
D.I.S.I., Università di Genova,
Via Dodecaneso 35, 16146 Genova, Italy

Alexander Rakhlin RAKHLIN@MIT.EDU
Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, MA, 02139, USA

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

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 (2004); Bartlett et al. (2004); Koltchinskii (2003) 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., 2004; 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, i.e. 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 proven by Bousquet and Elisseeff (2002), where a large family of algorithms (*Tikhonov regularization* based methods) has been shown to possess a strong L_∞ stability with respect to changes of single samples of the training set, and exponential bounds have been proven 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_∞-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. (1996): 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, a jump to a different part of the function class becomes less and less likely.

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

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 more restricted family of function classes: Donsker classes (see e.g. 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 e.g. 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 proven 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., 2004). 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 $(\mathcal{Z}, \mathcal{A})$ be a measurable space. Let P be a probability measure on $(\mathcal{Z}, \mathcal{A})$ and $Z_1, \ldots, Z_n$ be independent copies of Z with distribution P. Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ 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 $S = (Z_1, \ldots, Z_n)$. 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

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

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 ν_n indexed by $\mathcal{F}$ is defined as the map*

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

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

$$\nu_n \rightsquigarrow \nu$$

in $\ell^\infty(\mathcal{F})$, where the limit ν 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 ν must be a zero-mean Gaussian process with covariance function $\mathbb{E}\nu(f)\nu(f') = \langle f, f' \rangle$ (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, Eqn. 2.1.7 and van de Geer 2000, Thm. 6.3) in terms of entropy $\log \mathcal{N}$ and entropy with bracketing $\log \mathcal{N}_{[]}$.

Proposition 5 *If the envelope F of $\mathcal{F}$ is square integrable and*

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

then $\mathcal{F}$ is P-Donsker for every P, i.e. $\mathcal{F}$ is a universal Donsker class. Here the supremum is taken over all finitely discrete probability measures.

Proposition 6 *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 (Thm. 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 VC dimension is integrable.

3. Main Result

We now state the main result of this paper.

Theorem 7 *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 proof of Theorem 7 relies on the *almost sure representation theorem* (van der Vaart and Wellner, 1996, Thm. 1.10.4). Here we state the theorem applied to ν_n and ν.

Proposition 8 *Suppose $\mathcal{F}$ is P-Donsker. Let $\nu_n : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process. There exist a probability space $(\mathcal{Z}', \mathcal{A}', P')$ and maps $\nu_n' : \mathcal{Z}' \mapsto \ell^\infty(\mathcal{F})$ such that*

1. *$\nu_n' \overset{au}{\to} \nu'$*
2. *$\mathbb{E}^* f(\nu_n') = \mathbb{E}^* f(\nu_n)$ for every bounded $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ for all n.*

Lemma 9 is the main preliminary result used in the proof of Theorem 7 (and Theorem 13 in Section 5). We postpone its proof to Appendix A.

Lemma 9 *Let $\nu_n : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process. Fix n and assume that there exist a probability space $(\mathcal{Z}', \mathcal{A}', P')$ and a map $\nu_n' : \mathcal{Z}' \mapsto \ell^\infty(\mathcal{F})$ such that $\mathbb{E}^* f(\nu_n') = \mathbb{E}^* f(\nu_n)$ for every bounded $f : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$. Let ν' be a P-Brownian bridge defined on $(\mathcal{Z}', \mathcal{A}', P')$. Fix $C > 0$, $\epsilon = \min(C^3/128, C/4)$ and suppose $\delta \geq \xi\sqrt{n}$ for a given $\xi > 0$. Then, if $\mathcal{F}$ is P-Donsker, the following inequality holds*

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

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

Proof [Theorem 7] 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 8 states that for any $\delta > 0$

$$\Pr^*\left(\sup_{\mathcal{F}} |\nu_n' - \nu'| > \delta\right) \to 0.$$

By Lemma 9,

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

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

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

Corollary 10 *The result of Theorem 7 holds if the diameter is defined with respect to the $L_2(P)$ norm.*

4. Stability of almost-ERM

The main result of this section, Corollary 11, 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. (2004): 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 11 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 11 implies L_1-stability of ERM (or almost-ERM) algorithms on Donsker classes.

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

Corollary 11 *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 $f_n \in \mathcal{M}^{\xi(n)}_{S([n])}$ and $f'_n \in \mathcal{M}^{\xi'(n)}_{S(I_n)}$ for some $\xi(n) = o(n^{-1/2})$ and $\xi'(n) = o(n^{-1/2})$. Then*

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

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

Proof It is enough to show that $f'_n \in \mathcal{M}^{\xi''(n)}_{S([n])}$ for some $\xi''(n) = o(n^{-1/2})$ and result follows from the Theorem 7.

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

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 10 implies convergence in $L_2(P)$, and, therefore, in $L_1(P)$ norm. ■

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

The statement of Lemma 9 reveals that the rate of the decay of the diameter $\mathrm{diam}\mathcal{M}_S^{\xi(n)}$ is related to the rate at which $\Pr^*\left(\sup_{\mathcal{F}}|\nu - \nu_n| \geq \delta\right) \to 0$ for a fixed δ. 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 $\nu_n' : \mathcal{Z}^n \mapsto \ell^\infty(\mathcal{F})$ be the empirical process defined on the probability space $(\mathcal{Z}', \mathcal{A}', P')$.

Definition 12 *$\mathcal{F}$ is called a Komlos-Major-Tusnady class with respect to P and with the rate of convergence τ_n ($\mathcal{F} \in KMT(P;\tau_n)$) if $\mathcal{F}$ is P-pregaussian and for each $n \geq 1$ there is a version $\nu^{(n)}$ of P-Brownian bridge defined on $(\mathcal{Z}', \mathcal{A}', P')$ such that for all $t > 0$,*

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

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 13 *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*

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

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

Proof The result of Lemma 9 is stated for a fixed n. We now choose C, ξ, and δ 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 β 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 $\mathcal{F} \in KMT(P; n^{-\alpha})$,

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

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

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

Lemma 9 then implies

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

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

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

■

The entropy condition in Theorem 13 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 14 is a an application of Theorem 13 to this important family of classes. It follows in a straight-forward way from the remark above.

Corollary 14 *Assume $\mathcal{F}$ is a VC-subgraph class with VC-dimension V, and for some $\alpha > 0$ $\mathcal{F} \in KMT(P, n^{-\alpha})$. Let $\xi(n)\sqrt{n} = o(n^{-\eta})$, $\eta > 0$. Then*

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

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 16 relies on the following ratio inequality of Pollard (1995).

Proposition 15 *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*

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

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

Theorem 16 *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 $\kappa < \min\left(\frac{1}{6(2V+1)} \min(\alpha, \eta), 1/2\right)$*

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

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., 2004; 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.

8. Acknowledgements

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 9, which was used in the proof of Theorem 7, and Corollary 10. Let us start with some technical Lemmas.

Lemma 17 *Let $f_0, f_1 \in \mathcal{F}$, $\|f_0 - f_1\| \geq C/2$, $\|f_1\| \leq \|f_0\|$. Let $h : \mathcal{F} \to \mathbb{R}$ be defined as $h(f') = \frac{\langle f', f_0 \rangle}{\|f_0\|^2}$. Then for any $\epsilon \leq \frac{C^3}{128}$*

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

Proof

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

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

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

■

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

Lemma 18 *Let f_0, f_1, h be defined as in Lemma 17. Suppose $\epsilon \leq \frac{C^3}{128}$. Let ν_μ be a Gaussian process on $\mathcal{F}$ with mean μ and covariance $\mathrm{cov}(\nu_\mu(f), \nu_\mu(f')) = \langle f, f' \rangle$.*

Then for all $\delta > 0$

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

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

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

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

Notice that

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

Moreover Γ_0 and Γ_1 are convex and

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

by Lemma 17. Then $\Gamma_0 = \Gamma_1$ in a single point z_0 and

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

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

Furthermore,

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

and $\mathrm{var}(\nu_\mu(f_0)) = \|f_0\|^2 \geq C^2/16$, 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 9]

Consider the ϵ-covering $\{f_i | i = 1, \ldots, \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)\}$. 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 $\|f - f_k\| \leq \epsilon \leq C/4$, $\|f' - f_l\| \leq \epsilon \leq C/4$. By triangle inequality it follows that $\|f_k - f_l\| \geq C/2$.

Moreover

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

and

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

Therefore,

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

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

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

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

By union bound

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

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

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

where the first inequality results from a union bound argument while the second one results from Lemma 18 noticing that $-\nu'-\sqrt{n}P$ is a Gaussian process with covariance $\langle f,f'\rangle$ and mean $-\sqrt{n}P$, and since by construction $\epsilon\leq C^3/128$.

Finally, the claimed result follows from the two last relations. ■

We now prove, Corollary 10, the extension of Theorem 7 to L_2 diameters. The proof relies on the observation that a P-Donsker class is also Glivenko-Cantelli.

Proof [Corollary 10] Note that

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

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.

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

The first term can be bounded as

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

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 7. ■

Appendix B.

In this appendix we report the proof of Theorem 16 stated in Section 6. We first need to derive a preliminary lemma.

Lemma 19 *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 $\lambda(n) = o(n^{1/2})$*

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

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

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

Then

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

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 15 to the class $\mathcal{G}$, we obtain

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

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

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

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

$$\Pr^* \left(\sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} \big(|P(f - f')| - \xi(n)\big) > 26 \sup_{f,f' \in \mathcal{M}_S^{\xi(n)}} \big(\epsilon(P_n|f - f'| + P|f - f'|) + 5\gamma\big) \right) \leq 32\mathcal{N}(\gamma/2, \mathcal{F})^2 \exp(-n\epsilon\gamma). \tag{3}$$

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 $\lambda(n) = o(n^{1/2})$,

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

Thus,

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

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

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

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

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

Combining the relation above with Eqn. 3,

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

The result follows by the assumption on the entropy and by arbitrariness of δ. ■

We are now ready to prove Theorem 16.

Proof [Theorem 16] By Corollary 14,

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

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 19. 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 19. Proof of this fact is very similar to the proof of Corollary 10, except that C is replaced by $C\lambda(n)^{-1}$.

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

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

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

proving condition (1) in Lemma 19.

We now verify condition (2) in Lemma 19. Since $\mathcal{F}$ is a VC-subgraph class of dimension V, its entropy numbers $\log\mathcal{N}(\epsilon, \mathcal{F})$ behave like $V \log \frac{A}{\epsilon}$ (A is a constant), that is

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

Condition (2) of Lemma 19 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 19

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

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

$$\Pr^*\left(n^\kappa\sqrt{n}\sup_{f,f'\in\mathcal{M}_S^{\xi(n)}} |P(f-f')| \leq \sqrt{n}\xi(n)n^\kappa + 131 n^{\gamma(\rho-1)+\kappa}\right) \to 0. \tag{4}$$

Now fix a ρ such that $1/2 < \rho < 1 - \kappa/\gamma$. Because $0 < \kappa < \gamma/2$, there is always such a choice of ρ. Furthermore, $1 > \rho > 1/2$ so that the above convergence holds. Our choice of ρ implies that $\gamma(\rho-1)+\kappa < 0$ and so $n^{\gamma(\rho-1)+\kappa} \to 0$. Since $\kappa < \gamma/2 < \eta$, $\sqrt{n}\xi(n)n^\kappa \to 0$. 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 risk minimization. *The Annals of Probability*, 2004. submitted.

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. *Real analysis and probability*. Cambrige Studies in Advaced Mathematics. Cambrigde University Press, 2002.

Richard M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

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. Technical report, University of New Mexico, August 2003.

Vladimir I. 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.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI*, pages 275–282, 2002.

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. In *Computational Learing Theory*, pages 140–146, 1996.

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, 2004.

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, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 2005. To appear.

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 frequences 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.