Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-033
AIM-2005-018
CBCL-250

May 17, 2005

Some Properties of Empirical Risk Minimization over Donsker Classes

Andrea Caponnetto and Alexander Rakhlin

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Abstract. We study properties of algorithms which minimize (or almost-minimize) empirical error over a Donsker class of functions. We show that the L_2-diameter of the set of almost-minimizers is converging to zero in probability. Therefore, as the number of samples grows, it is becoming unlikely that adding a point (or a number of points) to the training set will result in a large jump (in L_2 distance) to a new hypothesis. We also show that under some conditions the expected errors of the almost-minimizers are becoming close with a rate faster than $n^{-1/2}$.

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 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.

1. Introduction

Let $(\mathcal{Z}, \mathcal{A})$ be a measurable space. Let P be (an unknown) measure on $(\mathcal{Z}, \mathcal{A})$ and $Z_1, \dots, Z_n$ be independent copies of Z with distribution P. Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ 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 $S = (Z_1, \dots, Z_n)$. Define the empirical measure as $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$.

Definition 1. *Given a sample S and class $\mathcal{F}$,*

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

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

The Empirical Risk Minimization (ERM) algorithm above has been studied in Learning Theory to a great extent. In this paper we prove some properties of almost-ERM algorithms, which, to our knowledge, do not appear in the literature. ERM is a reasonable strategy only if the class $\mathcal{F}$ is uniform Glivenko-Cantelli, that is, $\mathcal{F}$ satisfies the uniform law of large numbers. In this paper we focus our attention on more restricted classes: Donsker classes. These are classes satisfying not only the law of large numbers, but also a version of the central limit theorem. The specific structure of the limit of this convergence will allow us to control correlation of the empirical means of the minimizers of empirical error.

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

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

This norm is a natural measure of distance between functions, as will become apparent later, because the dot product above is the covariance of the limiting gaussian process. Due to a close relation of the $\|\cdot\|$ norm to the $L_2(P)$ norm, the results of this paper will hold for the $L_2(P)$ norm as well.

Definition 3. *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 [10].

In fact, it follows that the limit process ν must be a zero-mean Gaussian process with covariance function $\mathbb{E}\nu(f)\nu(f') = \|f - f'\|^2$.

Various Donsker Theorems provide sufficient conditions for checking if a class is P-Donsker. Here we mention a few known results (see e.g. [10]) in terms of entropy $\log\mathcal{N}$ and entropy with bracketing $\log\mathcal{N}_{[]}$.

Proposition 1. *If $\int_0^\infty \sqrt{\log\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2(P))}d\epsilon < \infty$, then $\mathcal{F}$ is P-Donsker.*

Definition 5. *An envelope F of the function class $\mathcal{F}$ is a measurable function with $F > |f|$ $\forall f \in \mathcal{F}$.*

Proposition 2. *If the envelope 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, i.e. $\mathcal{F}$ is* universal *Donsker class. Here the supremum is taken over all finitely discrete probability measures.*

If $\mathcal{F}$ is a $\{0,1\}$-valued class, then $\mathcal{F}$ is *uniform* Donsker class if and only if its VC dimension is finite (see [3]). Rudelson and Vershynin [7] extend Dudley's result: a class $\mathcal{F}$ is *uniform* Donsker if the square root of its VC dimension is integrable.

2. Main Result

We now state the main result of this paper:

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

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

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 [10].

Corollary 1. *The result of Theorem 1 holds if the diameter is defined with respect to the $L_2(P)$ norm.*

We start the proofs with two technical Lemmata.

Lemma 1. *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 [4].

Lemma 2. *Let f_0, f_1, h be defined as in Lemma 1. 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 1. 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 proof of our main theorem relies on the *Almost Sure Representation Theorem* (Thm 1.10.4 in [10]). Here we state the theorem applied to ν_n and ν.

Proposition 3. *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 1.9.3 in [10] in turn shows that when the limiting process is Borel measurable, almost uniform convergence implies convergence in outer probability. Therefore, the first implication of the theorem above states that for any $C > 0$

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

We are now ready to prove Theorem 1. The reasoning in the proof 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 an 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 of Theorem 1. Fix $C > 0$ and let $\epsilon = \min(C^3/128, C/4)$. 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 in [10]). For any $f, f' \in \mathcal{M}_S^{\xi(n)}$ 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

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

and

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

Therefore,

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

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

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

Now choose an arbitrary $\delta > 0$ and fix n_δ s.t. for n greater that n_δ the l.h.s. in the above relation is less than δ. Then $\forall n > n_\delta$

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

By union bound

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

We now want to bound the terms in the sum above. By the Almost Sure Representation Theorem, there exist a probability space $(\mathcal{Z}', \mathcal{A}', P')$ and maps $\nu_n' : \mathcal{Z}' \mapsto \ell^\infty(\mathcal{F})$ such that $\Pr^* \left(\sup_{\mathcal{F}} |\nu_n' - \nu'| \right) \to 0$ and ν_n and ν_n' have the same distribution. Assuming without loss of generality that $\|f_k\| \geq \|f_l\|$, we obtain

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

where the first inequality results from a union bound argument while the second one results from Lemma 2 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 we have

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

and the thesis follows from the arbitrariness of δ. □

Proof of Corollary 1. Note that

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

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

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

The first term can be bounded as

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

which goes to 0 because the class $\{f - f' | f, f' \in \mathcal{F}\}$ is P-Donsker. The second term goes to 0 by Theorem 1. □

3. Stability of almost-ERM

Corollary 2 shows *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) is decreasing.

This result provides a partial answer to the questions raised in the Machine Learning literature by [6, 8]: 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 2 proves that almost-ERM over binary VC classes possesses L_1 stability. For the real-valued classes, the uniform Glivenko-Cantelli property is strictly more general than the 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. This provides a partial answer to the question raised in [8], where L_1 stability over uGC classes was conjectured.

Use of L_1 stability goes back to Devroye and Wagner [2], who showed that it 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 [1]). Our Corollary 2 implies L_1 stability of ERM (or almost-ERM) algorithms on Donsker classes.

It is known that exact empirical risk minimization is an NP-hard problem even for simple function classes. An interesting further direction of research is to see whether the result of Corollary 2 can have algorithmic consequences.

Corollary 2. *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 $M_n := |I_n \triangle [1:n]| = o(n^{1/2})$. Suppose $f_n \in \mathcal{M}_{S([1:n])}^{\xi(n)}$ 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 $f'_n \in \mathcal{M}_{S([1:n])}^{\xi''(n)}$ for some $\xi''(n) = o(n^{-1/2})$ and result follows from the Theorem 1.

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

Define $\xi''(n) := 2\frac{M_n}{n} + \frac{|I_n|}{n}\xi'(n) + \xi(n)$. Because $M_n = o(\sqrt{n})$, i.e. the two sets are not very different, it follows that $\xi''(n) = o(n^{-1/2})$. Corollary 1 implies convergence in $L_2(P)$, and, therefore, in $L_1(P)$ norm. □

4. Expected Error Stability of almost-ERM

We show that if a bound on the rate of decrease of the diameter in Theorem 1 is available, then, under some conditions on the class, the difference between expected errors of almost-minimizers decays faster than $n^{-1/2}$. Similarly to the previous section, this implies that ERM is *stable* in the sense that when the training set is perturbed, the difference of expected errors decays faster than $n^{-1/2}$.

From the proof of Theorem 1, the rate of decrease of the diameter is bounded by the rate of convergence of the empirical process to the gaussian process. Some results on the rate of such convergence can be found in [5]. In the following Corollary, we will assume the rate of decay of the diameter is known and a condition on the metric entropy growth is satisfied.

Corollary 3. *Let $\mathcal{F}$ be a uniformly bounded function class with the envelope function $\mathcal{F} \equiv 1$. Assume $\mathcal{N}(\mathcal{F},\gamma) = \sup_Q \mathcal{N}_1(\mathcal{F},Q,\gamma) < \infty$ 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 $\lambda(n) = o(n^{1/2})$*

$$\lambda(n)\mathrm{diam}\mathcal{M}_S^{\xi(n)} \xrightarrow{P^*} 0. \tag{1}$$

Suppose further that

$$\lambda(n)^{1/2} - \log\mathcal{N}(\mathcal{F}, n^{-1/2}\lambda(n)^{-1/4}) \to +\infty. \tag{2}$$

Then

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

In particular, if $\mathcal{F}$ is a VC-subgraph class, the condition (2) is satisfied whenever the diameter decays faster than $\log^2 n$, i.e. $\lambda(n)/\log^2 n \to \infty$.

The proof relies on the following ratio inequality of Pollard [9]:

Proposition 4. *Let $\mathcal{G}$ be a uniformly bounded function class with the envelope function $G \equiv 2$. Assume $\mathcal{N}(\mathcal{G},\gamma) = \sup_Q \mathcal{N}_1(\mathcal{G},Q,2\gamma) < \infty$ for $0 < \gamma \leq 1$ and Q ranging over all discrete probability measures. Then*

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

Proof of Corollary 3. Define $\mathcal{G} = \{f - f' : f, f' \in \mathcal{F}\}$ and $\mathcal{G}' = \{|f - f'| : f, f' \in \mathcal{F}\}$. It can be shown that $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{G}'$ are Donsker classes (see [10]). In particular, $\mathcal{N}(\mathcal{G}, 2\gamma) \leq \mathcal{N}(\mathcal{F}, \gamma)^2$ and the envelope of $\mathcal{G}$ is $G \equiv 2$. Apply Proposition 4 to the class $\mathcal{G}$:

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

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

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

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

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

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,

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

Now choose $\epsilon_n = n^{-1/2}\lambda(n)^\alpha$ and $\gamma_n = n^{-1/2}\lambda(n)^{-\beta}$ for any $0 < \beta < \alpha < 1$. Then $n\epsilon_n\gamma_n = \lambda(n)^{\alpha-\beta}$ and

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

For the sake of simplicity, set $\alpha = 3/4$ and $\beta = 1/4$.

By definition of limit, for any $\delta > 0$, there exist N_δ such that for all $n > N_\delta$,

$$\Pr^*\left(\sqrt{n} \sup_{f,f'\in\mathcal{M}_S^{\xi(n)}} 26\left(\epsilon_n\left(P_n|f-f'| + P|f-f'|\right) + 5\gamma_n\right) > 2\lambda(n)^{-1/4}\right) < \delta.$$

Thus,

$$\Pr^*\left(\sqrt{n} \sup_{f,f'\in\mathcal{M}_S^{\xi(n)}} |P(f-f')| \leq \sqrt{n}\xi(n) + 2\lambda(n)^{-1/4}\right) \geq 1 - 32\mathcal{N}(\mathcal{F}, \frac{1}{2}n^{-1/2}\lambda(n)^{-1/4})^2 \exp(-\lambda(n)^{1/2}) - \delta.$$

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

If $\mathcal{F}$ is a VC subgraph class of dimension V, its entropy numbers $\log\mathcal{N}(\mathcal{F}, \epsilon)$ behave like $V \log\frac{1}{\epsilon}$, i.e. $\log\mathcal{N}(\mathcal{F}, n^{-1/2}\lambda(n)^{-1/4})$ behaves like $V\log n + V\log\lambda(n)$. Condition (2) will therefore hold whenever $\lambda(n)$ grows faster than $\log^2 n$. □

References

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

[2] 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.

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

[4] J. Kim and D. Pollard. Cube root asymptotics. *Annals of Statistics*, 18:191–219, 1990.

[5] Vladimir I. Koltchinskii. Komlós-Major-Tusnády approximation for the general empirical process and Haar expansion of classes of functions. 7:73–118, 1994.

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

[7] M.Rudelson and R.Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*. To appear.

[8] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is necessary and sufficient for consistency of empirical risk minimization. CBCL Paper 2002-023, Massachusetts Institute of Technology, December 2002 [January 2004 revision].

[9] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.

[10] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.

Andrea Caponnetto, CBCL, Massachusetts Institute of Technology

E-mail address: caponnet@mit.edu

Alexander Rakhlin, CBCL, Massachusetts Institute of Technology

E-mail address: rakhlin@mit.edu