Universität
Stuttgart

Fachbereich
Mathematik

Measuring the Capacity of Sets of Functions in the Analysis of ERM

Ingo Steinwart

Preprint 2014/008

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

Measuring the Capacity of Sets of Functions in the Analysis of ERM

Ingo Steinwart

Abstract Empirical risk minimization (ERM) is a fundamental learning principle that serves as the underlying idea for various learning algorithms. Moreover, ERM appears in many hyper-parameter selection strategies. Not surprisingly, the statistical analysis of ERM has thus attracted a lot of attention during the last four decades. In particular, it is well-known that as soon as ERM uses an infinite set of hypotheses, the problem of measuring the size, or capacity, of this set is central in the statistical analysis. We provide a brief, incomplete, and subjective survey of different techniques for this problem, and illustrate how the concentration inequalities used in the analysis of ERM determine suitable capacity measures.

1 Introduction

Given a data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$ sampled from some unknown distribution P on $X \times Y$, the goal of supervised learning is to find a decision function $f_D : X \to \mathbb{R}$ whose L-risk

$$\mathscr{R}_{L,P}(f_D) := \int_{X \times Y} L\big(x, y, f_D(x)\big)\, dP(x,y)$$

is small. Here, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a loss function e.g. the binary classification loss or the least squares loss. However, other choices, e.g. for quantile regression, weighted classification, classification with reject option, are important, too. To formalize the concept of "learning", we also need the Bayes risk $\mathscr{R}^*_{L,P} := \inf \mathscr{R}_{L,P}(f)$, where the infimum runs over *all* $f : X \to \mathbb{R}$. If this infimum is attained we denote a function that achieves $\mathscr{R}^*_{L,P}$ by $f^*_{L,P}$. Clearly, no algorithm can construct a decision

Ingo Steinwart
Institute for Stochastics and Applications
University of Stuttgart, Germany
e-mail: ingo.steinwart@mathematik.uni-stuttgart.de

function f_D whose risk is smaller than the $\mathscr{R}^*_{L,P}$. On the other hand, having an f_D whose risk is close to the Bayes risk is certainly desirable.

To formalize this idea, let us fix a learning method $\mathscr{L}$, which assigns to every finite data set D a function f_D. Then $\mathscr{L}$ learns in the sense of L-risk consistency for P, if

$$\lim_{n\to\infty} P^n\Big(D \in (X\times Y)^n : \mathscr{R}_{L,P}(f_D) \leq \mathscr{R}^*_{L,P} + \varepsilon\Big) = 1 \tag{1}$$

for all $\varepsilon > 0$. Moreover, $\mathscr{L}$ is called universally L-risk consistent, if it is L-risk consistent for all distributions P on $X\times Y$ with, e.g. $\mathscr{R}^*_{L,P} < \infty$. Recall that the first results on universally consistent learning methods were shown by Stone [40] in a seminal paper. Since then, various learning methods have been shown to be universally consistent. We refer to the books [15] and [21] for binary classification and least squares regression, respectively.

Clearly, consistency does not specify the speed of convergence in (1). To address this, we fix a sequence $(\varepsilon_n) \subset (0,1]$ converging to 0. Then $\mathscr{L}$ learns with rate (ε_n), if there exists a family $(c_\tau)_{t\in(0,1]}$ such that, for all $n \geq 1$ and all $\tau \in (0,1]$, we have

$$P^n\Big(D \in (X\times Y)^n : \mathscr{R}_{L,P}(f_D) \leq \mathscr{R}^*_{L,P} + c_\tau\,\varepsilon_n\Big) \geq 1-\tau\,. \tag{2}$$

Recall that unlike consistency, learning rates usually require assumptions on P by famous the no-free-lunch theorem of Devroye, see [17] and [15, Thm. 7.2]. In other words, no quantitative, distribution independent a-priori guarantee against the Bayes risk can be made for any learning algorithm. The aim of learning rates is thus to understand for which distributions a learning algorithm learns sufficiently fast.

An important class of learning methods are empirical risk minimizers (ERMs). Motivated by the law of large numbers, the idea of ERM is to minimize the empirical risk

$$\mathscr{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^n L\big(x_i, y_i, f(x_i)\big)$$

instead of the unknown risk $\mathscr{R}_{L,P}(f)$. Unfortunately, if this is done in a naïve way, for example, by minimizing the empirical risk over all functions $f : X \to \mathbb{R}$, then the resulting learning method memorizes the data, but is, in general, not able to learn. Therefore, ERM methods fix a "small" set F of functions $f : X \to \mathbb{R}$ over which the empirical risk is minimized, that is, the resulting decision functions are given by

$$f_D \in \arg\min_{f\in F} \mathscr{R}_{L,D}(f)\,.$$

Here we note that in general such a minimizer does not need to exist. In the following, we therefore assume that it does exist. The possible non-uniqueness of the minimizer will not be a problem, so that no extra assumptions are required.

Clearly, ERM only produces decision functions contained in F, and hence it is never able to outperform the relatively best risk

$$\mathscr{R}^*_{L,P,F} := \inf\big\{\mathscr{R}_{L,P}(f)\,\big|\, f\in F\big\}\,.$$

In particular, if we have a non-zero *approximation error*, that is $\mathscr{R}^*_{L,P,F} - \mathscr{R}^*_{L,P} > 0$, then the corresponding ERM cannot be L-risk consistent for this P.

In the same spirit as L-risk consistency, it is an interesting question for ERM to ask for *oracle inequalities*, that is, for meaningful lower bounds of the probabilities

$$P^n\Big(D \in (X \times Y)^n : \mathscr{R}_{L,P}(f_D) \le \mathscr{R}^*_{L,P,F} + \varepsilon\Big)\,. \tag{3}$$

Clearly, if these probabilities converge to 1 for $n \to \infty$, then the corresponding ERM is L-risk consistent, if $\mathscr{R}^*_{L,P,F} - \mathscr{R}^*_{L,P} = 0$. Moreover, if $F = F_n$ changes with the number of samples, then bounds on (3) can be used to investigate L-risk consistency and convergence rates. Indeed, for ERM over such F_n, the analysis can be split into the deterministic approximation error $\mathscr{R}^*_{L,P,F_n} - \mathscr{R}^*_{L,P}$ and an estimation error described by bounds of the form (3). From a statistical point of view, oracle inequalities are thus a key element for determining *a-priori guarantees* such as L-risk consistency and learning rates. Note that for determining learning rates, the right-hand side of oracle inequalities may depend, up to a certain degree, on properties of P, since learning rates are always distribution dependent by the no-free-lunch theorem.

Another interesting task in the analysis of ERM is to seek *generalization error bounds*, which provide meaningful lower bounds of

$$\inf_P P^n\Big(D \in (X \times Y)^n : \mathscr{R}_{L,P}(f_D) \le \mathscr{R}_{L,D}(f_D) + \varepsilon\Big)\,,$$

where the infimum runs over *all* distributions P on $X \times Y$. Unlike oracle inequalities, generalization error bounds are provide *a-posteriori* guarantees by estimating the risks $\mathscr{R}_{L,P}(f_D)$ in terms of the achieved training error without knowing P. The latter explains why they need to be independent of P.

2 Prelude: ERM for Finite Hypothesis Classes

The simplest case, in which one can analyze ERM is the case of finite F. Although, this may seem be a rather artificial setting in view of e.g. consistency, it is of high practical relevance for hyper-parameter selection schemes that are based on a empirical validation error.

In the following, we restrict our considerations to bounded losses, i.e. losses L that satisfy $L(x,y,f(x)) \le B$ for all $(x,y) \in X \times Y$ and $f \in F$. Here one can show, see e.g. [45, p. 95] or [36, Prop. 6.18] that

$$P^n\left(D \in (X \times Y)^n : \mathscr{R}_{L,P}(f_D) < \mathscr{R}^*_{L,P,F} + B\sqrt{\frac{2\tau + 2\ln|F|}{n}}\,\right) \ge 1 - 2e^{-\tau} \tag{4}$$

holds for all distributions P on $X \times Y$, and all $\tau > 0$, $n \ge 1$. For later use, recall that the proof of this bound first employs the ERM property to establish

$$\mathscr{R}_{L,P}(f_D) - \mathscr{R}^*_{L,P,F} \le 2 \sup_{f\in F} \big| \mathscr{R}_{L,P}(f) - \mathscr{R}_{L,D}(f) \big| . \tag{5}$$

Then the union bound together with Hoeffding's inequality is used to show

$$P^n\left(D \in (X\times Y)^n : \sup_{f\in F} \big| \mathscr{R}_{L,P}(f) - \mathscr{R}_{L,D}(f) \big| \ge B\sqrt{\frac{\tau}{2n}} \right) \le 2\,|F| e^{-\tau} . \tag{6}$$

Since $\mathscr{R}_{L,P}(f_D) - \mathscr{R}_{L,D}(f_D) \le \sup_{f\in F} |\mathscr{R}_{L,P}(f) - \mathscr{R}_{L,D}(f)|$, it becomes clear that bounds on the probability of the right-hand side of (5) can also be used to obtain generalization bounds. For example, in the case above, we immediately obtain

$$P^n\left(D \in (X\times Y)^n : \mathscr{R}_{L,P}(f_D) < \mathscr{R}_{L,D}(f_D) + B\sqrt{\frac{\tau + \ln|F|}{2n}} \right) \ge 1 - 2e^{-\tau} .$$

There are situations in which the $O(n^{-1/2})$-bound (4) does not provide the best rate of convergence. For example, if there exists an $f \in F$ with $\mathscr{R}_{L,P}(f) = 0$, then we obviously have $\mathscr{R}^*_{L,P,F} = \mathscr{R}^*_{L,P} = 0$, and one can show, see e.g. [36, p. 241f],

$$P^n\left(D \in (X\times Y)^n : \mathscr{R}_{L,P}(f_D) < \frac{8B(\tau + \ln|F|)}{n} \right) \ge 1 - e^{-\tau} \tag{7}$$

for all $\tau > 0$, $n \ge 1$. Note that (7) gives an $O(n^{-1})$ convergence rate, which is significantly better than the rate $O(n^{-1/2})$ obtained by (4). Unfortunately, however, the approach above does not improve our a-posteriori guarantees. Indeed, to estimate the risk $\mathscr{R}_{L,P}(f_D)$ after training with the help of (7), we would need to know that our unknown data-generating distribution at hand satisfies $\mathscr{R}^*_{L,P,F} = 0$.

Since the proof of (7) is somewhat archetypal for later results, let us briefly recollect its main steps, too. The basic idea is to consider functions of the form

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r} , \qquad f \in F, \tag{8}$$

where h_f is defined by $h_f(x,y) := L(x,y,f(x))$ and $r > 0$ is chosen later in the proof. This gives $\mathbb{E}_P g_{f,r} = 0$, and using

$$\mathbb{E}_P h_f^2 \le B\, \mathbb{E}_P h_f , \tag{9}$$

which holds by the non-negativity of h_f, we find both $\mathbb{E}_P g_{f,r}^2 \le \frac{B}{2r}$ and $\|g_{f,r}\|_\infty \le \frac{B}{r}$. Consequently, Bernstein's inequality together with a union bound gives

$$P^n\left(D \in (X\times Y)^n : \sup_{f\in F} \mathbb{E}_D g_{f,r} \ge \sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr} \right) \le |F| e^{-\tau} .$$

Now, using $\mathscr{R}_{L,D}(f_D) = 0$, we find (7) by setting $r := \frac{4B\tau}{n}$. Note that the key idea in the proof above is the variance bound (9), which led to a non-trivial variance bound for $g_{f,r}$, which in turn made it possible to apply Bernstein's inequality.

Interestingly, for functions of the form $h_f(x,y) := L(x,y,f(x)) - L(x,y,f^*_{L,P}(x))$ we may still have a variance bound of the form (9). For example, for the least squares loss and $Y \subset [-M,M]$ it is well-known that (9) holds for all functions $f : X \to [-M,M]$, if B is replaced by $16M^2$, and for some other losses and certain distributions P we may have at least

$$\mathbb{E}_P h_f^2 \leq V \cdot \big(\mathbb{E}_P h_f\big)^{\vartheta} \tag{10}$$

for some constants $\vartheta \in (0,1]$ and $V \geq B^{2-\vartheta}$, see e.g. [42, 7, 4, 39, 6, 36, 37]. In these cases, it can then be shown by a technical but conceptionally simple modification of the argument above that

$$\mathscr{R}_{L,P}(f_D) - \mathscr{R}^*_{L,P} < 6\big(\mathscr{R}^*_{L,P,F} - \mathscr{R}^*_{L,P}\big) + 4\left(\frac{8V\big(\tau + \ln(1+|F|)\big)}{n}\right)^{\frac{1}{2-\vartheta}} \tag{11}$$

holds with probability P^n not less than $1 - e^{-\tau}$. We refer to e.g. [36, Thm. 7.2].

The drafts of the proofs we presented above indicate that the full proofs are rather elementary. Moreover, all proofs relied on a concentration inequality for quantities of the form $\mathbb{E}_D g - \mathbb{E}_P g$, that is, on a quantified version of the law of large numbers. In fact, as soon as we have such a concentration inequality we can easily apply the union bound and repeat the remaining parts of the proof of (4) to obtain a bound in the spirit of (4). Moreover, if our concentration inequality has a dominating variance term like Bernstein's inequality does, then improvements are possible by using the ideas that led to (7) and (11), respectively. These insights are in particular applicable when analyzing ERM for non-i.i.d. data, since for many classes of stochastic processes for which we have a law of large numbers, we actually have concentration inequalities, too. This has been used in e.g. [46, 49, 50, 35, 38, 22].

3 Binary Classification and VC-Dimension

Clearly, the union bound argument used above falls apart, if F is infinite, and hence a natural question is to ask for infinite sets F for which we can still bound the probability in (6). Probably the most classical result in this direction considers the binary classification loss L. In this case, each function $L \circ f$ defined by $L \circ f(x,y) := L(x, f(y))$ is an indicator function, so that one has to bound the probability of D satisfying

$$\sup_{g \in G} \big|\mathbb{E}_D g - \mathbb{E}_P g\big| \geq \varepsilon\,, \tag{12}$$

where $G := L \circ F := \{L \circ f : f \in F\}$ is a set of indicator functions. Note that, for indicator functions, the set $G_{|D} := \{g_{|D} : g \in G\}$ of restrictions onto D is always finite, independently of whether G is finite or not. Indeed, we have $|G_{|D}| \leq 2^n$, where n is the length of the data set D. Writing

$$\mathscr{H}(G,n) := \ln \mathbb{E}_{D\sim P^n} \left|G_{|D}\right|$$

for the so-called annealed entropy, it can then be shown, see e.g. [45, Thm. 4.1] that

$$P^n\left(D \in (X\times Y)^n : \sup_{g\in G}\left|\mathbb{E}_D g - \mathbb{E}_P g\right| \geq \sqrt{\frac{\tau + \mathscr{H}(G,2n)}{n}} + \frac{1}{n}\right) \leq 4e^{-\tau}. \qquad (13)$$

The proof of this inequality is rather complex but classical, and hence we only mentioned that it consists of: *a)* symmetrization by a ghost sample, *b)* conditioning and subsequent use of $|G_{|D}|$, and *c)* application of Hoeffding's inequality. Now, replacing (6) by (13) and using (5), we obtain the bound

$$P^n\left(D\in(X\times Y)^n : \mathscr{R}_{L,P}(f_D) < \mathscr{R}^*_{L,P,F} + 2\sqrt{\frac{\tau + \mathscr{H}(G,2n)}{n}} + \frac{2}{n}\right) \geq 1 - 4e^{-\tau} \quad (14)$$

for ERM with the binary classification loss over arbitrary F. Note that the conceptional difference to (4) is the replacement of $\ln|F|$ by the annealed entropy $\mathscr{H}(G,2n)$, which may provide non-trivial bounds even for infinite hypotheses sets F. Namely, it is not hard to conclude from (14) that $\mathscr{R}_{L,P}(f_D) \to \mathscr{R}^*_{L,P,F}$ holds in probability, if $\mathscr{H}(G,n)n^{-1} \to 0$. The latter holds, if, on "average" we have a significantly better bound than $|G_{|D}| \leq 2^n$.

The natural next question is to ask for sets G satisfying $\mathscr{H}(G,n)n^{-1} \to 0$ for *all* distributions P on $X \times Y$. To this end, let us consider the so-called growth-function

$$\mathscr{G}(G,n) := \ln \sup_{D\in(X\times Y)^n} \left|G_{|D}\right|.$$

Since $\mathscr{H}(G,n) \leq \mathscr{G}(G,n)$, we can always replace $\mathscr{H}(G,2n)$ by $\mathscr{G}(G,2n)$ in (13) and (14). Now the first fundamental combinatorial insight of VC-theory, see e.g. [45, Thm. 4.3], is that we either have $\mathscr{G}(G,n) = \ln 2^n$ for all $n \geq 1$, or there exists an $n_0 \geq 0$ such that for all $n > n_0$ we have $\mathscr{G}(G,n) < \ln 2^n$. This leads to the famous Vapnik-Chervonenkis dimension

$$\text{VC-dim}(G) := \max\left\{n \geq 0 : \mathscr{G}(G,n) = \ln 2^n\right\}.$$

In the case of $\text{VC-dim}(G) < \infty$, we thus have $\mathscr{G}(G,n) < \ln 2^n$ for all $n > \text{VC-dim}(G)$, while in the case $\text{VC-dim}(G) = \infty$ we never have a non-trivial bound for the growth function. Now, the second combinatorial insight is that in the first case, i.e. $d := \text{VC-dim}(G) < \infty$, we have by Sauer's lemma

$$\mathscr{G}(G,n) \leq d\left(1 + \ln\frac{n}{d}\right) \qquad (15)$$

for all $n > d$, see again [45, Thm. 4.3], and also [15, Ch. 13], [18, Ch. 4], and [16, Chapter 4]. Of course, the latter can be plugged into (14), which leads to

$$P^n\left(D \in (X\times Y)^n : \mathscr{R}_{L,P}(f_D) < \mathscr{R}^*_{L,P,F} + 2\sqrt{\frac{\tau + d + d\ln\frac{2n}{d}}{n}} + \frac{2}{n}\right) \geq 1 - 4e^{-\tau}$$

for ERM with the binary classification loss over hypotheses sets F with $d :=$ VC-dim$(L\circ F) < \infty$. In this case, we thus obtain $\mathscr{R}_{L,P}(f_D) \to \mathscr{R}^*_{L,P,F}$ in probability, and the rate is only by a factor of $\sqrt{\ln n}$ worse than that of (4) in the case of finite F. Conversely, if VC-dim$(L\circ F) = \infty$, then the probability of (12) cannot be bounded in a distribution independent way. Namely, for all $\varepsilon > 0$, there exists a distribution P such that (12) holds with probability one, see [45, Thm. 4.5] for details.

The above discussion shows that the VC-dimension is fundamental for understanding ERM for binary classification and i.i.d. data. For this reason, the VC-dimension has been bounded for various classes of hypotheses sets. We refer to [45, 3, 18, 16, 8, 43] and the many references mentioned therein. Finally, some generalizations to non i.i.d. data can be found in e.g. [1, 48].

4 Covering Numbers and Generalized Notions of Dimension

The results of Section 3 only apply to ERM with a loss L for which the induced set $L\circ F$ of functions consists of indicator functions. Unfortunately, the only common learning problem for which this is true is binary classification. In this section, we therefore consider more general losses.

One of the best-known means for analyzing ERM for general losses are covering numbers. To recall their definition, let us fix a set G of functions $Z\to\mathbb{R}$, where Z is an arbitrary, non-empty set. Let us assume that G is contained in some normed space $(E, \|\cdot\|_E)$, so that $\|g\|_E$ is explained for all $g\in G$. Then, for all $\varepsilon > 0$, the $\|\cdot\|_E$-covering numbers of G are defined by

$$\mathscr{N}(G, \|\cdot\|_E, \varepsilon) := \inf\left\{n \geq 1 : \exists g_1,\dots,g_n \in G \text{ such that } G \subset \bigcup_{i=1}^n (g_i + \varepsilon B_E)\right\},$$

where $\inf\emptyset := \infty$ and $B_E := \{g\in E : \|g\|_E \leq 1\}$ denotes the closed unit ball of E.

One way to bound the probability of (12) with the help of covering numbers is inspired by the proof of (13) and goes back to Pollard, see [32, p. 25ff] and [21, Thm. 9.1]. It leads to a bound of the form

$$P^n\left(D \in (X\times Y)^n : \sup_{g\in G}\left|\mathbb{E}_D g - \mathbb{E}_P g\right| > 8\varepsilon\right) \leq 8\mathbb{E}_{D\sim P^n}\mathscr{N}(G, \|\cdot\|_{L_1(D)}, \varepsilon)e^{-\frac{n\varepsilon^2}{2B^2}},$$

where $\|g\|_{L_1(D)} := \frac{1}{n}\sum_{i=1}^n |g(x_i, y_i)|$ denotes the empirical L_1-norm of $g\in G$.

To illustrate how to use this inequality let us assume for simplicity, that the loss L is Lipschitz with constant 1, that is $|L(x,y,t)-L(x,y,t')| \leq |t-t'|$ for all $x \in X$, $y \in Y$, and $t,t' \in \{f(x) : f \in F\}$. Then, for $G := L \circ F$, a simple consideration shows

$$\mathcal{N}(G, \|\cdot\|_{L_1(D)}, \varepsilon) \leq \mathcal{N}(F, \|\cdot\|_{L_1(D_X)}, \varepsilon), \tag{16}$$

where $D_X := (x_1, \ldots, x_n)$. Now assume that F is contained in the unit ball B_E of some d-dimensional normed space $(E, \|\cdot\|_E)$ of functions on X for which the identity map $\mathrm{id} : E \to L_1(D_X)$ is continuous for all $D_X \in X^n$. Let us additionally assume that $\|\mathrm{id} : E \to L_1(D_X)\| \leq M$ for a suitable M and all $D_X \in X^n$. Then using a volume comparison argument, see e.g. [12, Prop. 1.3.1], one finds

$$\mathcal{N}(F, \|\cdot\|_{L_1(D_X)}, \varepsilon) \leq 2\left(\frac{4M}{\varepsilon}\right)^d \tag{17}$$

for all $0 < \varepsilon \leq 4M$, and consequently, the concentration inequality above becomes

$$P^n\left(D \in (X \times Y)^n : \sup_{g \in G}\bigl|\mathbb{E}_D g - \mathbb{E}_P g\bigr| > 8\varepsilon\right) \leq 8e^{-\frac{n\varepsilon^2}{2B^2} + d \ln \frac{8M}{\varepsilon}}$$

for all $0 < \varepsilon \leq 4M$. Setting $\varepsilon := B\sqrt{\frac{(\tau+1+2\ln 8M)d\ln n}{n}}$ we then obtain

$$P^n\left(D \in (X \times Y)^n : \sup_{g \in G}\bigl|\mathbb{E}_D g - \mathbb{E}_P g\bigr| > 8B\sqrt{\frac{(\tau+1+2\ln 8M)d\ln n}{n}}\right) \leq 8e^{-\tau}$$

for all $n \geq 8$ satisfying $\frac{n}{\ln n} \geq \frac{(\tau+1+2\ln 8M)dB^2}{16M^2}$ and from the latter it is easy to find a bound for ERM over F, which is only by a factor of $\sqrt{\ln n}$ worse than that of (4).

Now note that this derivation did not actually need the assumptions on F made above, except the covering number bound (17). In other words, as soon as we have a polynomial covering number bound of the form (17), we get the same rate for ERM over F. Such polynomial bounds cannot only be obtained by the simple functional analytic approach taken above, but also by some more involved, combinatorial means. To briefly discuss some of these, recall that a $D = \{(z_1), \ldots, (z_n)\} \subset Z$ is ε-shattered, by a class G of functions on Z, if there exists a function $h : D \to \mathbb{R}$ such that, for all subsets $I \subset \{1, \ldots, n\}$, there exists a function $g \in G$ such that

$$\begin{aligned} g(z_i) &\leq h(z_i) - \varepsilon & \qquad & i \in I \\ g(z_i) &\geq h(z_i) - \varepsilon & \qquad & i \in \{1, \ldots, n\} \setminus I. \end{aligned}$$

Moreover, D is shattered by G, if it is ε-shattered by G for some $\varepsilon > 0$. Now, for $\varepsilon > 0$, the ε-fat-shattering dimension of G is defined to be size of the largest set D that can be ε-shattered by G, i.e.

$$\text{fat-dim}(G, \varepsilon) := \sup\bigl\{|D| : D \subset X \times Y \text{ is } \varepsilon\text{-shattered by } G\bigr\}.$$

Analogously, Pollard's pseudo-dimension, see [33, Sec. 4], is defined to be size of the largest set D that can be shattered by G. Clearly, for all $\varepsilon > 0$, the ε-fat-shattering dimension is dominated by the pseudo-dimension. Moreover, [30] shows that there exist absolute constants K and c such that

$$\mathscr{N}(G, \|\cdot\|_{L_2(D)}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{K\cdot\text{fat-dim}(G,c\varepsilon)} \tag{18}$$

for all $0 < \varepsilon < 1$ provided that $\|g\|_\infty \leq 1$ for all $g \in G$. In particular, since $\|\cdot\|_{L_2(D)}$-covering numbers dominate $\|\cdot\|_{L_1(D)}$-covering numbers, we easily see that the analysis based on (17) remains valid, if G, or F, have finite pseudo-dimension, and the same is true if fat-dim$\,G(\varepsilon)$, or fat-dim$(F(\varepsilon)$, are bounded by $c\varepsilon^{-p}$ for some constants $c > 0$ and $p > 0$ and all sufficiently small $\varepsilon > 0$.

A bound for the $\|\cdot\|_{L_\infty(D)}$-norms that is conceptionally similar to (18) was shown in [2] and later improved in [29], and the latter paper also contains several historical notes and links. Also, note that for sets G of indicator functions (15) always yields

$$\mathscr{N}(G, \|\cdot\|_{L_\infty(D)}, \varepsilon) \leq e^{\mathscr{G}(G,n)} \leq \left(\frac{en}{\text{VC-dim}(G)}\right)^{\text{VC-dim}(G)}.$$

Historically, one of the main motivations for considering the dimensions above is the characterization of uniform Glivenko-Cantelli classes G, that is, classes for which

$$\lim_{n\to\infty} \sup_P P^n \Big(D : \sup_{m\geq n} \sup_{g\in G} \big|\mathbb{E}_D g - E_P g\big| \geq \varepsilon) \Big) = 0 \tag{19}$$

holds, where the outer supremum is taken over all probability measures P on the underlying space. For sets of indicator functions, (19) holds, if and only if VC-dim$(G) < \infty$, see e.g. [2, Thm. 2.1] which, however, attributes this result to Assouad and Dudley, while general sets G of bounded functions satisfy (19), if and only if, fat-dim$(G,\varepsilon) < \infty$ for all $\varepsilon > 0$, see [2, Thm. 2.5].

So far all our estimates on the expected covering numbers are based on the implicit, intermediate step

$$\mathbb{E}_{D\sim P^n} \mathscr{N}(G_{|D}, \|\cdot\|_{L_1(D)}, \varepsilon) \leq \sup_{D\in(X\times Y)^n} \mathscr{N}(G_{|D}, \|\cdot\|_{L_1(D)}, \varepsilon)\,, \tag{20}$$

which, from a conceptional point of view, is not that surprising, since both (19) and generalization bounds require a sort of worst-case analysis. In addition, there is also a technical reason for this intermediate step, namely the plain difficulty of directly estimating the expectation on the left-hand side of (20). Now assume again that G consist of bounded functions. Then we can continue the right-hand side of (20) by

$$\sup_{D\in(X\times Y)^n} \mathscr{N}(G, \|\cdot\|_{L_1(D)}, \varepsilon) \leq \sup_{D\in(X\times Y)^n} \mathscr{N}(G, \|\cdot\|_{L_\infty(D)}, \varepsilon) \leq \mathscr{N}(G, \|\cdot\|_\infty, \varepsilon)\,.$$

In general, estimating the expected covering numbers on the left-hand side of (20) by $\mathscr{N}(G, \|\cdot\|_\infty, \varepsilon)$, is, of course, horribly crude. Indeed, $\mathscr{N}(G, \|\cdot\|_\infty, \varepsilon)$ may not even be finite although the expected covering numbers are. A classical example for such a phenomenon are the reproducing kernel Hilbert spaces H of the Gaussian kernels on $\mathbb{R}^d$, since for these $\mathrm{id} : H \to \ell_\infty(\mathbb{R}^d)$ is not compact and thus $\mathcal{N}(H, \|\cdot\|_\infty, \varepsilon) = \infty$ for all sufficiently small $\varepsilon > 0$, see [36, Examp. 4.32], while the expected covering numbers can e.g. be bounded by an approach similar to [36, Thm. 7.34]. On the other hand, there are also some advantages of considering $\|\cdot\|_\infty$-covering numbers: first, if F is the unit ball of a Banach space, then the asymptotic behavior of $\mathcal{N}(F, \|\cdot\|_\infty, \varepsilon)$ may be exactly known, see e.g. [19], and second, $\|\cdot\|_\infty$-covering numbers can be directly used to bound the probability of (12) by an elementary union bound argument in combination with a suitable ε-net of F and Hoeffding's inequality, cf. [36, Prop. 6.22] and its proof. More precisely, we have

$$\sup_{f \in \mathcal{F}} \bigl| \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f) \bigr| < B \sqrt{\frac{\tau + \ln \mathcal{N}(F, \|\cdot\|_\infty, \varepsilon)}{2n}} + 2\varepsilon \tag{21}$$

with probability P^n not less than $1 - 2e^{-\tau}$. Note that this inequality holds for all $\varepsilon > 0$, and hence we can pick an ε that minimizes the right-hand side of (21). Finding such an ε is feasible, as soon as we have a suitable upper bound on the covering numbers, e.g. a bound that behaves polynomially in ε. Moreover, it is not hard to see that the inequality yields both oracle inequalities and generalization bounds.

While (21) is, in general, looser than our previous estimates, its proof is more robust, when it comes to modifying it to non i.i.d. data. Indeed, as soon as we have a Hoeffding type inequality, we can easily derive a bound of the form (21). For some examples, when such an inequality holds, we refer to [20, 25, 14, 13] and the references therein. As a consequence, it seems fair to say that such bounds of the form (21) are certainly not useful for obtaining sharp learning rates, but they may be good enough for deriving "quick-and-dirty" generalization bounds and learning rates in situations, in which non-experts for the particular data-generating stochastic processes are otherwise lost. Moreover, if even Bernstein type inequalities such as the one in [31, 47] are available then $\|\cdot\|_\infty$-covering numbers of F can still be used, we refer to [22] for one of the sharpest known results for *regularized* ERM and the references mentioned therein.

5 More Sophisticated Inequalities: McDiarmid and Talagrand

So far, all of the results presented relied directly or indirectly on either Hoeffding's or Bernstein's inequality in combination with a union bound. In the last twenty years, this credo has been slowly shifted towards the use of concentration inequalities that do not require the union bound. The first of these inequalities is McDiarmid's inequality [26], see also [16, Ch. 2], which states, in a slightly simplified version, that

$$P^n\Bigl(D \in Z^n : h(D) - \mathbb{E}_{P^n} h(D) \geq \varepsilon\Bigr) \leq e^{-\frac{2n\varepsilon^2}{c^2}} \tag{22}$$

holds for all functions $h : Z^n \to \mathbb{R}$ satisfying the bounded difference assumption

$$\left|h(z_1,\ldots,z_n) - h(z_1,\ldots,z_{i-1},z',z_{i+1},\ldots z_n)\right| \leq \frac{c}{n} \tag{23}$$

for all $z_1,\ldots,z_n,z' \in Z$ and $i = 1,\ldots,n$. The example most interesting for our purposes is the function $h : Z^n \to \mathbb{R}$ defined by

$$h(D) := \sup_{g\in G}\left|\mathbb{E}_D g - \mathbb{E}_P g\right|,$$

where G consists of non-negative, bounded functions. It is easy to verify that h satisfies (23) for $c := \sup_{g\in G}\|g\|_\infty$, and plugging this into (22) shows that

$$\sup_{g\in G}\left|\mathbb{E}_D g - \mathbb{E}_P g\right| \leq \mathbb{E}_{D'\sim P^n}\sup_{g\in G}\left|\mathbb{E}'_D g - \mathbb{E}_P g\right| + c\sqrt{\frac{\tau}{2n}}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Consequently, it remains to bound the expectation on the right-hand side of this inequality. Fortunately, bounding such an expectation is a rather old problem from empirical process theory, and hence a couple of different techniques do exist. Usually, the first step in any attempt to bound such an expectation is symmetrization

$$\mathbb{E}_{D\sim P^n}\sup_{g\in G}\left|\mathbb{E}_D g - \mathbb{E}_P g\right| \leq 2\mathbb{E}_{D\sim P^n}\mathbb{E}_{\varepsilon\sim\nu^n}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(z_i)\right| =: 2\mathbb{E}_{D\sim P^n}\operatorname{Rad}_D(G),$$

where ν is the probability measure on $\{-1,1\}$ defined by $\nu(\{-1\}) = \nu(\{1\}) = 1/2$. Therefore, it suffices to bound the expectations of the *empirical Rademacher averages* $\mathbb{E}_{D\sim P^n}\operatorname{Rad}_D(G)$, and for this task there are several results available, so that we only highlight a few. For example, for singletons, Khintchine's inequality, see e.g. [24, Lem. 4.1] gives universal constants $c_1, c_2 > 0$ such that

$$c_1\|g\|_{L_2(D)}\, n^{-1/2} \leq \operatorname{Rad}_D(\{g\}) \leq c_2\|g\|_{L_2(D)}\, n^{-1/2} \tag{24}$$

for all functions g and all $D \in Z^n$. Moreover, if G is finite, then an application of Hoeffding's inequality, see e.g. [8, Thm. 3.3], gives

$$\operatorname{Rad}_D(G) \leq \sqrt{\frac{2\ln|G|}{n}}\max_{g\in G}\|g\|_{L_\infty(D)} \leq c\sqrt{\frac{2\ln|G|}{n}}$$

under the assumptions made above. Similarly, if G is a set of indicator functions with finite VC-dimension $d := \text{VC-dim}(G)$, then we have both

$$\operatorname{Rad}_D(G) \leq \sqrt{\frac{2d\ln(n+1)}{n}} \qquad \text{and} \qquad \operatorname{Rad}_D(G) \leq 36\sqrt{\frac{d}{n}},$$

where the latter holds for $n \geq 10$. The first result, which is rather classical, can be found in e.g. [8, p. 328], and the second result, with 36 replaced by universal constant, is also well-known, see e.g. [16, p. 31], [28, Cor. 2.32], and [8, Thm. 3.4]. We obtained the constant 36 by combining a variant of Dudley's integral, see [16, Thm. 3.2], with the bound

$$\mathcal{N}(G, L_2(D), k/n) \leq e(d+1)\Big(\frac{2en^2}{k^2}\Big)^d, \qquad k = 1, \dots, n,$$

proven by Haussler [23], but we admit that the value 36 is not very sharp, in particular not for larger values of d and n. Since Dudley's integral is also important for bounding Rademacher averages for real-valued function classes, let us recall, see e.g. [44, Ch. 2.2] and [18, Ch. 2], that it states

$$\mathrm{Rad}_D(G) \leq \frac{K}{\sqrt{n}} \int_0^\infty \sqrt{\ln \mathcal{N}(G, L_2(D), \varepsilon)}\, d\varepsilon\,, \tag{25}$$

where K is a universal constant, whose value can be explicitly estimated by a close inspection of the proof. In particular, for indicator functions we have $K \leq 12$, see [16, Thm. 3.2], and the same is true for general sets G, see [10, Cor. 3.2]. Moreover, (25) is almost tight, since Sudakov's minorization theorem gives

$$\frac{C}{\sqrt{n}} \sup_{\varepsilon>0} \varepsilon \sqrt{\ln \mathcal{N}(G, L_2(D), \varepsilon)} \leq \sqrt{\ln\Big(2 + \frac{1}{c_1 \|G\|_{L_2(D)}}\Big)}\, \mathrm{Rad}_D(G)\,, \tag{26}$$

where C is a universal constant, c_1 is the constant appearing in (24), and $\|G\|_{L_2(D)} := \sup_{g\in G} \|g\|_{L_2(D)}$. Here, we note that (26) was obtained by combining [24, Cor. 4.14] with (24). For a slightly different version we refer to [11, Cor. 1.5].

In view of (25) and (26), we are back to estimating empirical covering numbers, and hence the results from Section 4 can be applied. For example, if we have fat-dim$(G,\varepsilon) \leq c\varepsilon^{-p}$ for some constants $c, p > 0$ with $p \neq 2$ and all $\varepsilon > 0$, then combining (25) with (18) shows, cf. [28, Thm. 2.35] and [5, Thm. 10], that

$$\mathrm{Rad}_D(G) \leq C_p \ln c \sqrt{c}\, n^{-\frac{1}{2\wedge p}}\,, \qquad n \geq 1,$$

where C_p is a constant only depending on p, and for $p = 2$ the same is true with an additional $(\ln n)^2$-factor.

Since for ERM we are interested in classes of the form $G = L \circ F$, a natural next question is, whether one can relate the Rademacher averages of F to those of G. In some cases, see e.g. [7], this can addressed by the so-called contraction principle [24, Thm. 4.12], which shows

$$\mathrm{Rad}_D(\varphi \circ G) \leq 2\, \mathrm{Rad}_D(G) \tag{27}$$

for all 1−Lipschitz functions $\varphi : \mathbb{R} \to \mathbb{R}$ with $\varphi(0)$. In other cases, combining (25) with (16) does the better job, see e.g. [36, Ch. 7].

Let us recall that we are actually interested in bounding *expected* Rademacher averages, so that by (25) it suffices find upper bounds for

$$\mathbb{E}_{D\sim P^n}\sqrt{\ln \mathcal{N}(G, L_2(D), \varepsilon)}\,.$$

Like for the expected covering numbers in Section 4, the latter task is, in general, very difficult, and the arguments used so far, implicitly used a step analogous to (20). Another way to bound the expected covering numbers above is to follow the steps discussed after (20). Of course, in doing so, all issues regarding loose bounds can be expected here, too. There is, however, one case, in which these loose steps can be avoided. Indeed, [36, Thm. 7.13] shows that Dudley's entropy integral can also be expressed in terms of entropy numbers, which are, roughly speaking, the functional inverse of covering numbers. Then, instead of bounding expected covering numbers, the task is to bound expected entropy numbers. While in general, this seems to be as hopeless as the former task, for RKHS, it turns out to be possible, see [34].

Let us finally have a brief look at Talagrand's inequality [41]. Recall that in its improved version due to Bousquet [9], see also [36, Thm. 7.5 and A.9] for a complete and self-contained proof, it shows, for every $\gamma > 0$, that

$$\sup_{g\in G}\big|\mathbb{E}_D g - \mathbb{E}_P g\big| \leq (1+\gamma)\mathbb{E}_{D'\sim P^n}\sup_{g\in G}\big|\mathbb{E}'_D g - \mathbb{E}_P g\big| + \sqrt{\frac{2\tau\sigma^2}{n}} + \left(\frac{2}{3}+\frac{1}{\gamma}\right)\frac{\tau B}{n}$$

holds with probability P^n not less than $1-e^{-t}$, where $\|G\|_{L_2(P)} \leq \sigma$ and $\|G\|_\infty \leq B$.

One way of applying Talagrand's inequality in the analysis of ERM in the presence of a variance bound (10) is to consider functions of the form (8) with $h_f := L\circ f - L\circ f^*_{L,P}$. Then the first difficulty is to bound

$$\mathbb{E}_{D\sim P^n}\sup_{f\in F}\left|\frac{\mathbb{E}_D h_f - \mathbb{E}_P h_f}{\mathbb{E}_P h_f + r}\right|.$$

This is resolved by the so-called peeling argument, that estimates this expectation with the help of suitable upper bounds $\varphi(r)$ for the following, *localized* expectations

$$\mathbb{E}_{D\sim P^n}\sup_{\substack{f\in F\\ \mathbb{E}_P h_f \leq r}}\big|\mathbb{E}_D h_f - \mathbb{E}_P h_f\big| \leq \varphi(r)\,.$$

Using the variance bound (10), the localization $\mathbb{E}_P h_f \leq r$ can then replaced by the variance localization $\mathbb{E}_P h_f^2 \leq V r^\vartheta$, and hence the problem reduces to finding suitable upper bounds for the localized Rademacher averages $\mathrm{Rad}_D(G_r)$, where

$$G_r := \{h_f : \mathbb{E}_P h_f^2 \leq r\}\,.$$

In turn, these localized Rademacher averages can be estimated by a clever combination of the contraction principle and Dudley's entropy integral, see e.g. [27, Lem. 2.5]. A resulting, rather generic oracle inequality for (regularized) ERM can be found in [36, Thm. 20].

Finally, we note that there is another way to use Talagrand's inequality in the analysis of ERM, see e.g. [28, 4]. We decided to present the above one, since the approach can be more easily adapted to regularized empirical risk minimization, as it can be illustrated by comparing the analysis on support vector machines in [39] and [36, Ch. 8].

References

1. Adams, T.M., Nobel, A.B.: Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. Ann. Probab. **38**, 1345–1367 (2010)
2. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM **44**, 615–631 (1997)
3. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
4. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Amer. Statist. Assoc. **101**, 138–156 (2006)
5. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
6. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. J. Mach. Learn. Res. **9**, 1823–1840 (2008)
7. Blanchard, G., Bousquet, O., Massart, P.: Statistical performance of support vector machines. Ann. Statist. **36**, 489–531 (2008)
8. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. ESAIM Probab. Stat. **9**, 323–375 (2005)
9. Bousquet, O.: A Bennet concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris **334**, 495–500 (2002)
10. Bousquet, O.: Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms (2002). Ph.D. thesis, Ecole Polytechnique
11. Carl, B., Pajor, A.: Gelfand numbers of operators with values in a Hilbert space. Invent. Math. **94**, 479–504 (1988)
12. Carl, B., Stephani, I.: Entropy, Compactness and the Approximation of Operators. Cambridge University Press, Cambridge (1990)
13. Chazottes, J.R., Gouëzel, S.: Optimal concentration inequalities for dynamical systems. Comm. Math. Phys. **316**, 843–889 (2012)
14. Dedecker, J., Prieur, C.: New dependence coefficients. Examples and applications to statistics. Probab. Theory Related Fields **132**, 203–236 (2005)
15. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
16. Devroye, L., Lugosi, G.: Combinatorial Methods in Density Estimation. Springer, New York (2001)
17. Devroye, L.P.: Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. IEEE Trans. Pattern Anal. Mach. Intell. **4**, 154–157 (1982)
18. Dudley, R.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge (1999)
19. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press, Cambridge (1996)
20. van de Geer, S.A.: On Hoeffding's inequality for dependent random variables. In: H. Dehling, T. Mikosch, M. Sørensen (eds.) Empirical Process Techniques for Dependent Data, pp. 161–169. Birkhäuser, Boston, MA (2002)
21. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002)
22. Hang, H., Steinwart, I.: Fast learning from α-mixing observations. J. Multivariate Anal.
23. Haussler, D.: Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. J. Combin. Theory Ser. A **69**, 217–232 (1995)
24. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin (1991)

25. Maume-Deschamps, V.: Exponential inequalities and estimation of conditional probabilities. In: P. Bertail, P. Doukhan, P. Soulier (eds.) Dependence in Probability and Statistics, pp. 123–140. Springer, New York (2006)
26. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics (Norwich, 1989), *London Math. Soc. Lecture Note Ser.*, vol. 141, pp. 148–188. Cambridge Univ. Press, Cambridge (1989)
27. Mendelson, S.: Improving the sample complexity using global data. IEEE Trans. Inform. Theory **48**, 1977–1991 (2002)
28. Mendelson, S.: A few notes on statistical learning theory. In: S. Mendelson, A. Smola (eds.) Advanced Lectures on Machine Learning: Machine Learning Summer School 2002, Canberra, Australia, pp. 1–40. Springer, Berlin (2003)
29. Mendelson, S., Vershynin, R.: Entropy, combinatorial dimensions and random averages. In: J. Kivinen, R.H. Sloan (eds.) Computational Learning Theory (15th Annual Conference on Computational Learning Theory), pp. 14–28. Springer, Berlin (2002)
30. Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Invent. Math. **152**, 37–55 (2003)
31. Modha, D.S., Masry, E.: Minimum complexity regression estimation with weakly dependent observations. IEEE Trans. Inform. Theory **42**, 2133–2145 (1996)
32. Pollard, D.: Convergence of Stochastic Processes. Springer, New York (1984)
33. Pollard, D.: Empirical Processes: Theory and Applications. Institute of Mathematical Statistics & American Statistical Association, Hayward, CA & Alexandria, VA (1990)
34. Steinwart, I.: Oracle inequalities for SVMs that are based on random entropy numbers. J. Complexity **25**, 437–454 (2009)
35. Steinwart, I., Anghel, M.: An SVM approach for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. Ann. Statist. **37**, 841–875 (2009)
36. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
37. Steinwart, I., Christmann, A.: Estimating conditional quantiles with the help of the pinball loss. Bernoulli **17**, 211–225 (2011)
38. Steinwart, I., Hush, D., Scovel, C.: Learning from dependent observations. J. Multivariate Anal. **100**, 175–194 (2009)
39. Steinwart, I., Scovel, C.: Fast rates for support vector machines using Gaussian kernels. Ann. Statist. **35**, 575–607 (2007)
40. Stone, C.: Consistent nonparametric regression. Ann. Statist. **5**, 595–645 (1977)
41. Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**, 505–563 (1996)
42. Tsybakov, A.B.: Optimal aggregation of classifiers in statistical learning. Ann. Statist. **32**, 135–166 (2004)
43. van der Vaart, A., Wellner, J.A.: A note on bounds for VC dimensions. In: High Dimensional Probability V: the Luminy Volume, vol. 5, pp. 103–107. Inst. Math. Statist., Beachwood, OH (2009)
44. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, New York (1996)
45. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
46. Vidyasagar, M.: A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems, 2nd edn. Springer, London (2003)
47. Wintenberger, O.: Deviation inequalities for sums of weakly dependent time series. Electron. Commun. Probab. **15**, 489–503 (2010)
48. Yu, B.: Rates of convergence for empirical processes of stationary mixing sequences. Ann. Probab. **22**, 94–116 (1994)
49. Zou, B., Li, L.: The performance bounds of learning machines based on exponentially strongly mixing sequences. Comput. Math. Appl. **53**, 1050–1058 (2007)
50. Zou, B., Li, L., Xu, Z.: The generalization performance of ERM algorithm with strongly mixing observations. Mach. Learn. **75**, 275–295 (2009)

Ingo Steinwart
Universität Stuttgart
Fachbereich Mathematik
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: ingo.steinwart@mathematik.uni-stuttgart.de

Erschienene Preprints ab Nummer 2007/2007-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2014-008 *Steinwart, I.:* Measuring the Capacity of Sets of Functions in the Analysis of ERM

2014-007 *Steinwart, I.:* Convergence Types and Rates in Generic Karhunen-Loève Expansions with Applications to Sample Path Properties

2014-006 *Steinwart, I.; Pasin, C.; Williamson, R.; Zhang, S.:* Elicitation and Identification of Properties

2014-005 *Schmid, J.; Griesemer, M.:* Integration of Non-Autonomous Linear Evolution Equations

2014-004 *Markhasin, L.:* L_2- and $S^r_{p,q}B$-discrepancy of (order 2) digital nets

2014-003 *Markhasin, L.:* Discrepancy and integration in function spaces with dominating mixed smoothness

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-018 *Kimmerle, W.; Konovalov, A.:* On the Prime Graph of the Unit Group of Integral Group Rings of Finite Groups II

2012-017 *Stroppel, B.; Stroppel, M.:* Desargues, Doily, Dualities, and Exceptional Isomorphisms

2012-016 *Moroianu, A.; Pilca, M.; Semmelmann, U.:* Homogeneous almost quaternion-Hermitian manifolds

2012-015 *Steinke, G.F.; Stroppel, M.J.:* Simple groups acting two-transitively on the set of generators of a finite elation Laguerre plane

2012-014 *Steinke, G.F.; Stroppel, M.J.:* Finite elation Laguerre planes admitting a two-transitive group on their set of generators

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian

2011-028 *Spreer, J.:* Combinatorial 3-manifolds with cyclic automorphism group

2011-027 *Griesemer, M.; Hantsch, F.; Wellig, D.:* On the Magnetic Pekar Functional and the Existence of Bipolarons

2011-026 *Müller, S.:* Bootstrapping for Bandwidth Selection in Functional Data Regression

2011-025 *Felber, T.; Jones, D.; Kohler, M.; Walk, H.:* Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates

2011-024 *Jones, D.; Kohler, M.; Walk, H.:* Weakly universally consistent forecasting of stationary and ergodic time series

2011-023 *Györfi, L.; Walk, H.:* Strongly consistent nonparametric tests of conditional independence

2011-022 *Ferrario, P.G.; Walk, H.:* Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors

2011-021 *Eberts, M.; Steinwart, I.:* Optimal regression rates for SVMs using Gaussian kernels

2011-020 *Frank, R.L.; Geisinger, L.:* Refined Semiclassical Asymptotics for Fractional Powers of the Laplace Operator

2011-019 *Frank, R.L.; Geisinger, L.:* Two-term spectral asymptotics for the Dirichlet Laplacian on a bounded domain

2011-018 *Hänel, A.; Schulz, C.; Wirth, J.:* Embedded eigenvalues for the elastic strip with cracks

2011-017 *Wirth, J.:* Thermo-elasticity for anisotropic media in higher dimensions

2011-016 *Höllig, K.; Hörner, J.:* Programming Multigrid Methods with B-Splines

2011-015 *Ferrario, P.:* Nonparametric Local Averaging Estimation of the Local Variance Function

2011-014 *Müller, S.; Dippon, J.:* k-NN Kernel Estimate for Nonparametric Functional Regression in Time Series Analysis

2011-013 *Knarr, N.; Stroppel, M.:* Unitals over composition algebras

2011-012 *Knarr, N.; Stroppel, M.:* Baer involutions and polarities in Moufang planes of characteristic two

2011-011 *Knarr, N.; Stroppel, M.:* Polarities and planar collineations of Moufang planes

2011-010 *Jentsch, T.; Moroianu, A.; Semmelmann, U.:* Extrinsic hyperspheres in manifolds with special holonomy

2011-009 *Wirth, J.:* Asymptotic Behaviour of Solutions to Hyperbolic Partial Differential Equations

2011-008 *Stroppel, M.:* Orthogonal polar spaces and unitals

2011-007 *Nagl, M.:* Charakterisierung der Symmetrischen Gruppen durch ihre komplexe Gruppenalgebra

2011-006 *Solanes, G.; Teufel, E.:* Horo-tightness and total (absolute) curvatures in hyperbolic spaces

2011-005 *Ginoux, N.; Semmelmann, U.:* Imaginary Kählerian Killing spinors I

2011-004 *Scherer, C.W.; Köse, I.E.:* Control Synthesis using Dynamic D-Scales: Part II — Gain-Scheduled Control

2011-003 *Scherer, C.W.; Köse, I.E.:* Control Synthesis using Dynamic D-Scales: Part I — Robust Control

2011-002 *Alexandrov, B.; Semmelmann, U.:* Deformations of nearly parallel G_2-structures

2011-001 *Geisinger, L.; Weidl, T.:* Sharp spectral estimates in domains of infinite volume

2010-018 *Kimmerle, W.; Konovalov, A.:* On integral-like units of modular group rings

2010-017 *Gauduchon, P.; Moroianu, A.; Semmelmann, U.:* Almost complex structures on quaternion-Kähler manifolds and inner symmetric spaces

2010-016 *Moroianu, A.; Semmelmann,U.:* Clifford structures on Riemannian manifolds

2010-015 *Grafarend, E.W.; Kühnel, W.:* A minimal atlas for the rotation group $SO(3)$

2010-014 *Weidl, T.:* Semiclassical Spectral Bounds and Beyond

2010-013 *Stroppel, M.:* Early explicit examples of non-desarguesian plane geometries

2010-012 *Effenberger, F.:* Stacked polytopes and tight triangulations of manifolds

2010-011 *Györfi, L.; Walk, H.:* Empirical portfolio selection strategies with proportional transaction costs

2010-010 *Kohler, M.; Krzyżak, A.; Walk, H.:* Estimation of the essential supremum of a regression function

2010-009 *Geisinger, L.; Laptev, A.; Weidl, T.:* Geometrical Versions of improved Berezin-Li-Yau Inequalities

2010-008 *Poppitz, S.; Stroppel, M.:* Polarities of Schellhammer Planes

2010-007 *Grundhöfer, T.; Krinn, B.; Stroppel, M.:* Non-existence of isomorphisms between certain unitals

2010-006 *Höllig, K.; Hörner, J.; Hoffacker, A.:* Finite Element Analysis with B-Splines: Weighted and Isogeometric Methods

2010-005 *Kaltenbacher, B.; Walk, H.:* On convergence of local averaging regression function estimates for the regularization of inverse problems

2010-004 *Kühnel, W.; Solanes, G.:* Tight surfaces with boundary

2010-003 *Kohler, M; Walk, H.:* On optimal exercising of American options in discrete time for stationary and ergodic data

2010-002 *Gulde, M.; Stroppel, M.:* Stabilizers of Subspaces under Similitudes of the Klein Quadric, and Automorphisms of Heisenberg Algebras

2010-001 *Leitner, F.:* Examples of almost Einstein structures on products and in cohomogeneity one

2009-008 *Griesemer, M.; Zenk, H.:* On the atomic photoeffect in non-relativistic QED

2009-007 *Griesemer, M.; Moeller, J.S.:* Bounds on the minimal energy of translation invariant n-polaron systems

2009-006 *Demirel, S.; Harrell II, E.M.:* On semiclassical and universal inequalities for eigenvalues of quantum graphs

2009-005 *Bächle, A, Kimmerle, W.:* Torsion subgroups in integral group rings of finite groups

2009-004 *Geisinger, L.; Weidl, T.:* Universal bounds for traces of the Dirichlet Laplace operator

2009-003 *Walk, H.:* Strong laws of large numbers and nonparametric estimation

2009-002 *Leitner, F.:* The collapsing sphere product of Poincaré-Einstein spaces

2009-001 *Brehm, U.; Kühnel, W.:* Lattice triangulations of E^3 and of the 3-torus

2008-006 *Kohler, M.; Krzyżak, A.; Walk, H.:* Upper bounds for Bermudan options on Markovian data using nonparametric regression and a reduced number of nested Monte Carlo steps

2008-005 *Kaltenbacher, B.; Schöpfer, F.; Schuster, T.:* Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems

2008-004 *Leitner, F.:* Conformally closed Poincaré-Einstein metrics with intersecting scale singularities

2008-003 *Effenberger, F.; Kühnel, W.:* Hamiltonian submanifolds of regular polytope

2008-002 *Hertweck, M.; Höfert, C.R.; Kimmerle, W.:* Finite groups of units and their composition factors in the integral group rings of the groups $PSL(2, q)$

2008-001 *Kovarik, H.; Vugalter, S.; Weidl, T.:* Two dimensional Berezin-Li-Yau inequalities with a correction term

2007-006 *Weidl, T.:* Improved Berezin-Li-Yau inequalities with a remainder term

2007-005 *Frank, R.L.; Loss, M.; Weidl, T.:* Polya's conjecture in the presence of a constant magnetic field

2007-004 *Ekholm, T.; Frank, R.L.; Kovarik, H.:* Eigenvalue estimates for Schrödinger operators on metric trees

2007-003 *Lesky, P.H.; Racke, R.:* Elastic and electro-magnetic waves in infinite waveguides

2007-002 *Teufel, E.:* Spherical transforms and Radon transforms in Moebius geometry

2007-001 *Meister, A.:* Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions