Universität Stuttgart

Fachbereich Mathematik

Some Remarks on the Statistical Analysis of SVMs and Related Methods

Ingo Steinwart

Preprint 2013/015

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**

Some Remarks on the Statistical Analysis of SVMs and Related Methods

Ingo Steinwart

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 statistical learning is to find an $f_D : X \to \mathbb{R}$ 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 describing our learning goal. Probably the two best-known examples of such losses are the binary classification loss and 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

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

If this infimum is attained we denote a function that achieves $\mathscr{R}^*_{L,P}$ by $f^*_{L,P}$.

Now, a learning method $\mathscr{L}$ assigns to every finite data set D a function f_D. Such an $\mathscr{L}$ learns in the sense of L-risk consistency for P, if

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

for all $\varepsilon > 0$. Moreover, $\mathscr{L}$ is called universally L-risk consistent, if it is L-risk consistent for all distributions P on $X \times Y$.

Recall that the first results on universally consistent learning methods were shown by Stone [34] in a seminal paper. Since then, various learning methods have

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

been shown to be universally consistent. We refer to the books [10] and [16] 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, we say that $\mathscr{L}$ 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

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

In addition, we say that $\mathscr{L}$ learns with expected rate (ε_n) if $\mathbb{E}_{D\sim P^n}\mathscr{R}_{L,P}(f_D) \preceq \varepsilon_n$. Here, $a_n \preceq b_n$ means that there exists a constant $c \geq 0$ with $a_n \leq cb_n$ for all $n \geq 1$. Analogously, we sometimes write $a_n \sim b_n$ if $a_n \preceq b_n$ and $b_n \preceq a_n$.

Unlike consistency, learning rates usually require assumptions on P by the no-free-lunch theorem of Devroye, see [11] and [10, Thm. 7.2]. In Section 4 we will discuss such assumptions and the resulting rates for SVMs.

To recall the definition of SVMs and related methods, we fix a reproducing kernel Hilbert space (RKHS) H, a loss L that is convex in its third argument, and a $\lambda > 0$. Then, the optimization problem

$$f_{D,\lambda} \in \arg\min_{f\in H} \lambda \|f\|_H^2 + \mathscr{R}_{L,D}(f)\,, \tag{3}$$

where $\mathscr{R}_{L,D}(f)$ is the empirical risk of f, that is $\mathscr{R}_{L,D}(f) = \frac{1}{n}\sum_{i=1}^n L\big(x_i,y_i,f(x_i)\big)$, has a unique solution $f_{D,\lambda} \in H$, see [29, Lem. 5.1 & Thm. 5.2].

Let us briefly make some historical remarks: In 1992 V. Vapnik and co-workers, [6] presented the first SVM, namely the hard-margin SVM, which combined the generalized portrait algorithm from [38] with a kernel embedding inspired by [1]. Only a few years later, C. Cortes and V. Vapnik [8] proposed the first soft-margin SVMs, which are instances of (3) for which L is the (squared) hinge loss. Almost at the same time, the ε-insensitive loss for regression was proposed in [37, 12, 36]. However, approaches of the form (3) are actually significantly older. In 1971, for example, G. Kimeldorf and G. Wahba [17] showed a form of the representer theorem for the Sobolev space case $H = W^m([0,1]^d)$ with $m > d/2$ and the least squares loss L. Until the end of the 1980's a substantial amount of further research dealt with this and similar cases, see e.g. [25, 39]. Inspired by this work, [24] presented an approach called regularization network to the learning community in 1990, which basically considers (3) for the least squares loss.

Ideally, a learning method is automatic, i.e. no parameters need to be set by the user. In the SVM case, this means that λ and possible kernel parameters such as the width $\gamma > 0$ of the Gaussian kernel

$$k_\gamma(x,x') := \exp(-\gamma^{-2}\|x-x'\|)\,, \qquad x,x' \in \mathbb{R}^d,$$

are set automatically. In practice, such parameters are usually determined by cross-validation. Let us briefly describe a simplified version of this, see [29, Def. 6.28]. To this end, we split D in two (almost) equally sized parts D_1 and D_2. In addition, let Λ be a finite set of candidates for λ and, if necessary, Γ be a finite set of candidates for the kernel parameter. Then, for all combinations $(\lambda,\gamma)\in\Lambda\times\Gamma$, the optimization (3) is solved for the data set D_1, and the resulting *clipped* SVM solution, see (5), is validated on D_2, i.e., its empirical D_2-error is computed. Finally, the SVM solution with the smallest D_2-error is taken as the decision function f_D.

In the following, we try to give a brief survey on what is known about consistency and learning rates for SVMs. To this end, we first recall some key concepts related to their analysis in Section 2. We then consider consistency and learning rates in Sections 3 and 4, respectively. Due to limited space, these discussions are restricted to binary classification and least squares regression. However, most of the results we discuss are actually derived from generic oracle inequalities and thus they can be naturally extended to other losses. Here, differences usually only occur if assumptions on P are made to guarantee e.g. variance bounds or approximation properties. For an example we refer to quantile regression with the pinball loss in [30, 13].

2 Mathematical Prerequisites

In the following, let $(X,\mathscr{A})$ be a measurable space, $Y\subset\mathbb{R}$ be a closed subset, and P be a distribution on $X\times Y$ whose marginal distribution on X is denoted by P_X. In addition, we always assume that H is a separable reproducing kernel Hilbert space (RKHS) of a bounded measurable kernel k on X with $\|k\|_\infty\leq 1$. Finally, if not stated otherwise, L denotes a loss that satisfies $\mathscr{R}_{L,P}(0)<\infty$.

The goal of this section is to recall some concepts that describe interactions between P, L, and H, which are relevant for the analysis of SVMs.

Let us begin by recalling that the "inclusion" operator $I_k : H\to L_2(P_X)$ that maps an $f\in H$ to its equivalence $L_2(P_X)$-class $[f]_\sim$ is a Hilbert-Schmidt operator, see [29, Thm. 4.27]. Moreover, the usual integral operator $T_k : L_2(P_X)\to L_2(P_X)$ with respect to k is well-defined and given by $T_k = I_k\circ I_k^*$, where I_k^* denotes the adjoint operator of I_k. In particular, T_k is self-adjoint, positive and nuclear, see again [29, Thm. 4.27], and thus, the classical spectral theorem can be applied. This yields an at most countable family $(\mu_i)_{i\in I}\subset(0,\infty)$ of non-zero eigenvalues (with geometric multiplicities) of T_k, which, in case of infinite I, converges to zero. As usual, we assume without loss of generality that $I\subset\mathbb{N}$ and $\mu_1\geq\mu_2\geq\cdots>0$.

Some of the results we will review later make explicit assumptions on the decay of the eigenvalues, while other results make assumptions on the behavior of covering numbers or entropy numbers. Since the latter two are essentially the same concepts, let us only recall the latter. To this end, we first consider a compact metric space (M,d). Then, for $n\geq 1$, the n-th entropy number of an $A\subset M$ is defined by

$$\varepsilon_n(A,d) := \inf\left\{\varepsilon>0 : \exists t_1,\ldots,t_n\in M \text{ such that } A\subset\bigcup_{i=1}^n B(t_i,\varepsilon)\right\}$$

where $B(t,\varepsilon)$ denotes the closed ball with center t and radius ε. Moreover, if E and F are Banach spaces and $T : E\to F$ is a bounded linear operator, then the n-th (dyadic) entropy number of T is defined by $e_n(T) := \varepsilon_{2^{n-1}}(TB_E, \|\cdot\|_F)$, where B_E denotes the closed unit ball of E. In the Hilbert space case, eigenvalue and entropy number decays are closely related. For example, [31, Thm. 15] shows that

$$\mu_i(T_k) \preceq i^{-1/p} \qquad \Longleftrightarrow \qquad e_i(I_k : H \to L_2(P_X)) \preceq i^{-1/2p}\,. \tag{4}$$

Moreover, the latter is implied by $e_i(\mathrm{id} : H \to \ell_\infty(X)) \preceq i^{-1/2p}$.

Assumptions on the eigenvalue or entropy number decay are used to estimate the stochastic error of (3). To derive consistency and learning rates, however, we also need to bound the approximation error. To recall concepts in this direction, we first need the smallest possible L-risk in H, that is, $\mathscr{R}^*_{L,P,H} := \inf_{f\in H} \mathscr{R}_{L,P}(f)$. To achieve consistency, we obviously need zero approximation error, that is $\mathscr{R}^*_{L,P,H} = \mathscr{R}^*_{L,P}$. If H is universal, cf. [27] and [23], that is, X is a compact metric space and H is dense in $C(X)$, this equality can be guaranteed, see [29, Cor. 5.29]. For specific losses, however, weaker assumptions on H are sufficient. E.g., if L is the least squares loss, the equality $\mathscr{R}^*_{L,P,H} = \mathscr{R}^*_{L,P}$ holds, if and only if H is dense in $L_2(P_X)$. For many Lipschitz continuous losses including the hinge loss, the ε-insensitive loss, and the pinball loss, an analogous characterization holds in terms of $L_1(P_X)$-denseness, see [29, Cor. 5.37]. Finally, recall that for fixed $\gamma > 0$, the RKHS H_γ of the Gaussian kernel k_γ is dense in $L_p(P_X)$ for all $p \in [1,\infty)$, see [29, Thm. 4.63]. Once we have fixed an H with $\mathscr{R}^*_{L,P,H} = \mathscr{R}^*_{L,P}$, we need to consider the approximation error function (AEF)

$$A(\lambda) := \inf_{f\in H} \lambda \|f\|_H^2 + \mathscr{R}_{L,P}(f) - \mathscr{R}^*_{L,P}\,, \qquad \lambda \ge 0\,.$$

It can be shown that $\lim_{\lambda\to 0} A(\lambda) = 0$, see [29, Lem. 5.15]. In general, the speed of convergence cannot be faster than $O(\lambda)$ and this rate is achieved, if and only if there exists an $f \in H$ with $\mathscr{R}_{L,P}(f) = \mathscr{R}^*_{L,P}$, see [29, Cor. 5.18].

For the least squares loss, the behavior of the AEF can be described by interpolation spaces $[E,F]_{\theta,r}$ of the real method, see [4, 5]. Namely, [26] shows that $f^*_{L,P} \in [L_2(P_X), H]_{\beta,\infty}$, if and only if $A(\lambda) \in O(\lambda^\beta)$. Here we note that the latter condition is often imposed to derive learning rates. Other authors, however, assume $f^*_{L,P} \in \operatorname{ran} T_k^{\beta/2} = [L_2(P_X), [H]_\sim]_{\beta,2}$, where $\operatorname{ran} T_k^{\beta/2}$ denotes the image of the $\beta/2$-fractional power of T_k and the equality of this image to the interpolation space has been recently shown in [33, Thm. 4.6]. Finally, we always have the continuous embeddings $[L_2(\nu),[H]_\sim]_{\beta-\varepsilon,\infty} \hookrightarrow [L_2(\nu),[H]_\sim]_{\beta,2} \hookrightarrow [L_2(\nu),[H]_\sim]_{\beta,\infty}$ for all $\varepsilon > 0$.

Finally, one often knows in advance, that it suffices to look for decision functions of the form $f_D : X \to [-M,M]$ for some $M > 0$. In particular, this is the case if the loss is clippable at M, that is, for all $x \in X, y \in Y$, and $t \in \mathbb{R}$, we have

$$L(x,y,\widehat{t}\,) \le L(x,y,t)\,, \tag{5}$$

where $\widehat{t} := \max\{-M, \min\{M,t\}\}$. Note that for convex L this is satisfied if and only if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ has a global minimum that is contained in $[-M,M]$ for all $(x,y) \in X \times Y$, see [29, Lem. 2.23]. The latter is satisfied for many commonly used losses, and for such losses it is beneficial to clip the SVM decision function.

3 Universal Consistency

In this section we discuss several results concerning the universal consistency of learning methods of the form (3) for binary classification and least squares regression. Due to space constraints we restrict our considerations to a-priori chosen parameters. However Theorems 1 and 2 below and the results discussed for regression can also be formulated for data splitting approaches, cf. [29, Thm. 7.24 & 8.26].

Binary Classification

Let us first note that the binary classification loss, which defines the actual learning goal, is not even continuous, and hence cannot be used in the SVM optimization problem (3). This issue is resolved by using a surrogate loss L such as the (squared) hinge loss or the least squares loss. For these losses, the first consistency results can be found in [28] and [40]. To recall these results, we assume that $X \subset \mathbb{R}^d$ is compact and H is universal. Then [28] establishes universal *classification* consistency, if *a)* we use the hinge loss, *b)* we have $\varepsilon_i(X,d_k) \preceq i^{-1/\alpha}$ for some $\alpha > 0$, where d_k is the kernel metric in the sense of [29, Eq. (4.20)], and *c)* we use a sequence of regularization parameters (λ_n) with $\lambda_n \to 0$ and $n\lambda_n^\alpha \to \infty$. In addition, for the Gaussian kernel k_γ with fixed but arbitrary width γ we can choose $\alpha := d$. By completely different methods, [40] shows universal classification consistency for a variety of losses including the (squared) hinge and the least squares loss if $\lambda_n \to 0$ and $n\lambda_n \to \infty$. A key idea in both articles is to compare the excess L-risk $\mathscr{R}_{L,P}(f) - \mathscr{R}^*_{L,P}$ of arbitrary f to the excess classification risk of f. Namely, in [28] an asymptotic relationship is shown, while [40] goes one step further by establishing inequalities between these excess risks. This idea was picked up in [2], who showed for convex margin-based losses, i.e. for losses L of the form $L(y,t) = \varphi(yt)$, that we have an asymptotic relationship or an inequality between these excess risks, if and only if φ is differentiable at 0 with $\varphi'(0) < 0$. For such losses we have the following consistency result:

Theorem 1. *Let L be as above and $\varphi(t) \in O(t^q)$ for some $q \geq 1$ and $t \to \infty$. Moreover, let H be dense in $L_q(P_X)$ and $(\lambda_n) \subset (0,\infty)$ with $\lambda_n \to 0$. Then the clipped SVM is classification consistent for P, if one of these conditions is satisfied:*

i) $n\lambda_n / \ln n \to \infty$ *and* $n\lambda_n^{q/2} \to \infty$.
ii) $n\lambda_n^{q/2} \to \infty$ *and* $n\lambda_n^p \to \infty$ *for some* $p \in (0,1)$ *with* $\mu_i(T_k) \preceq i^{-1/p}$.

If X is compact and H is universal, then all assumptions involving q can be dropped.

Proof. The first assertion follows from [29, Lem. 5.15 & Thm. 5.31] together with a simple generalization of [29, Thm. 7.22]. The second result can be shown analogously by employing [29, Thm. 7.23] together with [29, Cor. 7.31] and (4). Now assume that X is compact and H is universal. We fix an $\varepsilon \in (0,2]$ and pick an $f : X \to \mathbb{R}$ with $\mathscr{R}_{L,P}(f) \leq \mathscr{R}^*_{L,P} + \varepsilon$. Since L is clippable, say at M, we may assume that f maps into $[-M,M]$. By [3, Thm. 29.14] we then find a $g \in C(X)$ with $\|f-g\|_{L_1(P_X)} \leq \varepsilon$. Again, we can assume that $\|g\|_\infty \leq M$. Since H is universal, there also exists an

$h_\varepsilon \in H$ with $\|h_\varepsilon - g\|_H \leq \varepsilon$. Here we note that we can additionally assume that the resulting function $\varepsilon \mapsto \|h_\varepsilon\|_H$ is decreasing. Our construction yields $\|h_\varepsilon\|_\infty \leq 2+M$ and $\|f-h_\varepsilon\|_{L_1(P_X)} \leq 2\varepsilon$. Since L is locally Lipschitz, see [29, Lem. 2.25], we find $\mathscr{R}_{L,P}(h_\varepsilon) - \mathscr{R}_{L,P}(f) \leq 2c_L\varepsilon$ by [29, Lem. 2.19], where $c_L \geq 1$ is a constant only depending on L. This gives $\mathscr{R}_{L,P}(h_\varepsilon) - \mathscr{R}^*_{L,P} \leq 3c_L\varepsilon$. For $\lambda \in (0,1]$ we now define $\varepsilon_\lambda := 2\inf\{\varepsilon \in (0,1] : \|h_\varepsilon\|_H^2 \leq \lambda^{-1/2}\}$. We then obtain $\|h_{\varepsilon_\lambda}\|_\infty \leq 2+M$ and

$$\lambda\|h_\varepsilon\|_H^2 + \mathscr{R}_{L,P}(h_\varepsilon) - \mathscr{R}^*_{L,P} \leq \lambda^{1/2} + 3c_L\varepsilon_\lambda \to 0 \qquad \lambda \to 0.$$

Choosing $f_0 := h_{\varepsilon_\lambda}$ in (the proof) of [29, Thm. 7.22 & 7.23] gives the assertions.

The result above yields universal classification consistency, if, e.g. $X = \mathbb{R}^d$ and $H = H_\gamma$ with *fixed* kernel width γ. For Gaussian kernels, it is, however, common practice to vary γ with the sample size, too. The following result covers this case:

Theorem 2. *Let L be convex, clippable, and margin-based with $\varphi'(0) < 0$. Furthermore, let $(\lambda_n) \subset (0,1]$ and $\gamma_n \subset (0,1]$ satisfy $\lambda_n\gamma_n^{-d} \to 0$. Then the clipped SVM is universally classification consistent if one of the following conditions holds:*

i) $X = \mathbb{R}^d$, $\varphi(t) \in O(t^q)$ for some $q \geq 1$ and $t \to \infty$, $n\lambda_n/\ln n \to \infty$ and $n\lambda_n^{q/2} \to \infty$.
ii) $X \subset \mathbb{R}^d$ is compact and $\lambda_n^\varepsilon\gamma_n^d n \to \infty$ for some $\varepsilon > 0$.

Proof. Using $\|\mathrm{id} : H_1 \to H_\gamma\| \leq \gamma^{-d/2}$, see [29, Prop. 4.46], it is easy to check that the AEFs A_γ and A_1 of the Gaussian RKHSs H_γ and H_1 satisfy $A_\gamma(\lambda) \leq A_1(\lambda\gamma^{-d})$. Then the first assertion follows as for Theorem 1. The second assertion can be shown using the arguments for compact X in the proof of Theorem 1.

Least Squares Regression

We already noted in the introduction that least squares regression methods of the form (3) had already been around when SVMs were proposed. Despite their earlier appearance, the first[1] universal consistency results in our sense seems to be shown relatively late by [18]. Under the moment condition $\mathscr{R}_{L,P}(0) = \mathbb{E}_{(x,y)\sim P}y^2 < \infty$, the authors obtain consistency for $H = W^m([0,1]^d)$ if $\lambda_n \to 0$, $n\lambda_n \to \infty$, and the decision functions f_{D,λ_n} are clipped at $\ln n$. In [16, Theorem 20.4] the condition $n\lambda_n \to \infty$ was relaxed to $n\lambda_n^p/(\ln n)^7 \to \infty$ with $p := d/(2m)$, and it seems plausible that their proof allows to remove the logarithmic factor at least partially, if Y is bounded and a more aggressive clipping is applied. In any case, for bounded Y the general theory tells us that the logarithmic factors can be removed. Indeed, for bounded Y, it is easy to check that the conditions ensuring consistency in Theorems 1 and 2 also ensure consistency for least squares regression if we set $q = 2$. In the case of Theorem 1, for example, we obtain consistency for generic H, if $\lambda_n \to 0$ and $n\lambda_n/\ln n \to \infty$, and the latter can be replaced by $n\lambda_n^p \to \infty$ for some $p \in (0,1)$, if X is compact, H is universal, and $\mu_i(T_k) \preceq i^{-1/p}$. Note that this covers the case $H = W^m([0,1]^d)$ for $p := d/(2m)$ by the well-known estimate $e_i(I_k : W^m([0,1]^d) \to \ell_\infty([0,1]^d)) \preceq i^{-\frac{m}{d}}$, see e.g. [14, p. 118].

[1] In [16] the authors actually give some credit to the 1987 paper [15] for the case $d = 1$.

4 Learning Rates

In this section we discuss some known learning rates for SVMs for binary classification and least squares regression.

Binary Classification

Probably the earliest established learning rates for SVMs with (squared) hinge loss can be found in [27]. To formulate this result we define $\eta(x) := P(Y=1|x)$, $x \in X$, as well as $X_- := \{\eta < 1/2\}$ and $X_+ := \{\eta > 1/2\}$. We say that P has zero noise, if $|2\eta - 1| = 1$ P_X-almost surely, and has strictly separated classes, if $d(X_-, X_+) > 0$ for a version of η and a metric d on X. Now assume that (X,d) is compact, H is universal and $\lambda_n = n^{-1}$. Then [27] shows that $P^n(D : \mathscr{R}_{L,P}(f_{D,\lambda_n}) = 0) \geq 1 - e^{-cn}$ for all $n \geq n_0$, where L is the classification loss and c and n_0 depend on P and H.

In [20] exponentially fast expected rates under similar but weaker conditions were shown. There the authors assume that (X,d) is compact, η has a Lipschitz continuous version and that P has Tsybokov's noise exponent $q = \infty$, see below. Note that together these assumptions imply that P has strictly separated classes. For universal kernels and the logistic loss for classification, they then show that there are constants $c_1, c_2 > 0$ with $\mathbb{E}_{D\sim P^n}\mathscr{R}_{L,P}(f_{D,\lambda_n}) - \mathscr{R}^*_{L,P} \leq \exp(-c_1 n \lambda_n)$ if $\lambda_n \leq c_2$ and $n\lambda_n^{1+p} \to \infty$. Here, $p \in (0,1)$ is a constant such that $\sup_\nu e_i(I_k : H \to L_2(\nu)) \leq c i^{-1/(2p)}$ for all $i \geq 1$, where the supremum is taken over all distributions ν on X.

For both results discussed so far, it seems fair to say that *a)* the assumptions on P are very strong and that *b)* similar rates can also be achieved without much effort for classical histogram rules. In the case of the hinge loss and Gaussian kernels with varying widths, more realistic assumptions on P have been proposed in [32], which, to some extent, generalize the assumptions above. To briefly describe them, we define the distance to the decision boundary by $\Delta(x) := d(x, X_+)$ if $x \in X_-$, $\Delta(x) := d(x, X_-)$ if $x \in X_+$, and $\Delta(x) = 0$ otherwise. Then P is said to have margin noise exponent $\beta \in (0,\infty]$, if $\mathbb{E}_{P_X} 1_{\{\Delta < t\}}|2\eta - 1| \leq (ct)^\beta$ for a constant $c \geq 1$ and all $t \geq 0$. A detailed discussion of this assumption can be found in [29, Sec. 8.2], so we only mention that β is large if there is not much mass and/or a lot of noise in the area $\{\Delta < t\}$ around the decision boundary. In addition, we need Tsybakov's noise condition [35] that bounds the total amount of noise by $P_X(|2\eta - 1| < t) \leq (ct)^q$ for constants $c > 0$ and $q \in [0,\infty]$, and all $t \geq 0$. Then [29, Thm. 8.26] shows that the data splitting approach with polynomially growing n^{-1}-nets Λ_n and $n^{-1/d}$-nets Γ_n of $(0,1]$ learns with rate $n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}+\varepsilon}$ for all $\varepsilon > 0$. Note that depending on β and q the exponent in the rate varies between 0 and 1, in particular, rates up to n^{-1} are possible in all dimensions d provided that β and q are large enough.

Finally, let us briefly discuss some rates for generic H and the hinge loss (the least squares case will be considered at the end of our discussions on least squares regression). To this end, we assume that P satisfies Tsybakov's noise condition for some $q \in [0,\infty]$, as well as $\mu_i(T_k) \preceq i^{-1/p}$ and $A(\lambda) \in O(\lambda^\beta)$ for some $p \in (0,1)$ and $\beta \in (0,1]$. Then we usually have to expect $\beta < 1$, since for $\beta = 1$ the Bayes decision function, which is a step function, must be contained in H and for commonly used H this is impossible. In addition, Tsybakov's noise condition gives a variance bound, which in turn can be used, e.g., in [29, Thm. 7.24]. The resulting learning rate is $n^{-\min\{\frac{2\beta}{\beta+1},\frac{\beta(q+1)}{\beta(q+2-p)+p(q+1)}\}}$ for the data splitting approach if (Λ_n) is a sequence of polynomially growing n^{-2}-nets of $(0,1]$.

Least Squares Regression

Similar to the case of consistency, the first learning rates were established for the space $H = W^m([0,1]^d)$. Indeed, based on some techniques from empirical processes pioneered by S. van de Geer, [19] showed expected rates of the form $(\ln n)^2 n^{-\frac{2s}{2s+d}}$ for a structural risk minimization procedure to choose the parameters m and λ. Here $s > d/2$ describes the unknown smoothness of the regression function in the sense of $f^*_{L,P} \in W^s([0,1]^d)$. The procedure is thus adaptive to the unknown smoothness s, and in addition, no assumptions except $\operatorname{supp} P_X \subset [0,1]^d$ are necessary.

Let us now turn to the generic case. Here, beginning with [9], various investigations have been made, so we only focus on the ones who established (nearly) optimal rates. To the best of our knowledge, the first result in this direction was established in [7] under the assumptions $\mu_i(T_k) \sim i^{-1/p}$ and $f^*_{L,P} \in \operatorname{ran} T^{\beta/2}$ for some $p \in (0,1)$ and $\beta \in [1,2]$. Note that $\beta \geq 1$ implies that $f^*_{L,P} \in H$. Then, modulo some logarithmic factor in the case $\beta = 1$, the authors establish the rate

$$n^{-\frac{\beta}{\beta+p}}\,, \tag{6}$$

and they also show that this rate is optimal. Especially remarkable is the fact, that the authors are able to deal with values $\beta > 1$, since for such values the classical approach that splits the analysis into a stochastic part and the AEF fails due to the fact that the AEF does not converge faster than linearly. To avoid this issue, the authors split quite differently with the help of spectral methods.

From a practical point of view, however, the case $\beta < 1$ is the more realistic one. For this case, the first essentially optimal rate was proved in [22] for a variant of (3) in which the exponent 2 in the regularization term is replaced by the smaller exponent $2p/(1+p)$, where $p \in (0,1)$ is chosen such that $\mu_i(T_k) \preceq i^{-1/p}$. Provided that the eigenvectors of T_k are *uniformly* bounded and $f^*_{L,P} \in [L_2(P_X),H]_{\beta,\infty}$ for some $\beta \in (0,1]$, [22] then establishes (6) modulo some logarithmic factors. A closer look at this assumption on the eigenvectors shows that it is solely used to establish the interpolation inequality $\|f\|_\infty \leq c\|f\|_H^p\,\|f\|_{L_2(P_X)}^{1-p}$ for all $f \in H$, where $c > 0$ is some constant. Interestingly, this inequality is equivalent to the continuous embedding $[L_2(P_X),H]_{p,1} \hookrightarrow L_\infty(P_X)$. Now, [31] shows that combining the interpolation inequality with [29, Thm. 7.23], also the original algorithm (3) learns with rate (6) and the additional logarithmic factors are superfluous. Moreover, if the eigenvalue assumption is two-sided, i.e. $\mu_i(T_k) \sim i^{-1/p}$, then (6) is also optimal for all $\beta \in (p,1]$.

In the Sobolev space case $H = W^m([0,1]^d)$ and $f^*_{L,P} \in W^s([0,1]^d)$ for some $m > d/2$ and $s \in (0,m]$ these generic results imply the above mentioned rates $n^{-\frac{2s}{2s+d}}$, if P_X is (essentially) the uniform distribution, see [31]. Moreover, [13] has recently shown that up to some arbitrarily small $\varepsilon > 0$ in the exponent, the rates can also be achieved by Gaussian RKHSs H_γ, if γ varies with the sample size, too. Note that the latter seems to be somewhat necessary, since for fixed γ and $f^*_{L,P} \notin C^\infty$, the AEF can only have logarithmic decay, see [26]. Finally, the rates of [31, 13] can also be achieved by the data splitting approach.

Let us finally return to binary classification with the least squares loss. To this end, we assume $\eta \in [L_2(P_X), H]_{\beta,\infty}$ and that Tsybakov's noise assumption is satisfied for some $q \in [0,\infty]$. Note that the latter implies a stronger calibration inequality between the excess least squares and the excess classification risk, see [2] and [29, Thm. 8.29]. Considering [31], we then obtain the rate $n^{-\frac{\beta q}{(\beta+p)(q+1)}}$, which at first glance seems to be fine, since for large β and q the exponent reaches 1. However, it may be the case that large values for β and q exclude each other. To illustrate this (see [21] for a similar observation), let us consider the Sobolev case $\eta \in W^s([0,1]^d)$ in which the rates in [31] become $n^{-\frac{2sq}{(2s+d)(q+1)}}$. To get rates close to n^{-1}, we need large s, say $s > 1 + d/2$. Then $\eta \in C^1$ by Sobolev's embedding theorem, which in turn excludes $q > 1$ by some geometric considerations, and hence rates arbitrarily close to n^{-1} are impossible. Finally, the same observation can be made for [13].

References

1. Aizerman, M., Braverman, E., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. Autom. Remote Control **25**, 821–837 (1964)
2. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Amer. Statist. Assoc. **101**, 138–156 (2006)
3. Bauer, H.: Measure and Integration Theory. De Gruyter, Berlin (2001)
4. Bennett, C., Sharpley, R.: Interpolation of Operators. Academic Press, Boston (1988)
5. Bergh, J., Löfström, J.: Interpolation Spaces, An Introduction. Springer-Verlag, New York (1976)
6. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152 (1992)
7. Caponnetto, A., De Vito, E.: Optimal rates for regularized least squares algorithm. Found. Comput. Math. **7**, 331–368 (2007)
8. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 273–297 (1995)
9. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Amer. Math. Soc. **39**, 1–49 (2002)
10. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
11. 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)
12. Drucker, H., Burges, C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: Advances in Neural Information Processing Systems 9, pp. 155–161 (1997)
13. Eberts, M., Steinwart, I.: Optimal regression rates for SVMs using Gaussian kernels. Electron. J. Stat. **7**, 1–42 (2013)

14. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press, Cambridge (1996)
15. van de Geer, S.: A new approach to least squares estimation, with applications. Ann. Statist. **15**, 587–602 (1987)
16. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002)
17. Kimeldorf, G.S., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**, 82–95 (1971)
18. Kohler, M., Krzyżak, A.: Nonparametric regression estimation using penalized least squares. IEEE Trans. Inform. Theory **47**, 3054–3058 (2001)
19. Kohler, M., Krzyżak, A., Schäfer, D.: Application of structural risk minimization to multivariate smoothing spline regression estimates. Bernoulli **4**, 475–489 (2002)
20. Koltchinskii, V., Beznosova, O.: Exponential convergence rates in classification. In: Proceedings of the 18th Annual Conference on Learning Theory, pp. 295–307 (2005)
21. Loustau, S.: Aggregation of SVM classifiers using Sobolev spaces. J. Mach. Learn. Res. **9**, 1559–1582 (2008)
22. Mendelson, S., Neeman, J.: Regularization in kernel learning. Ann. Statist. **38**, 526–565 (2010)
23. Micchelli, C.A., Xu, Y., Zhang, H.: Universal kernels. J. Mach. Learn. Res. **7**, 2651–2667 (2006)
24. Poggio, T., Girosi, F.: A theory of networks for approximation and learning. Proc. IEEE **78**, 1481–1497 (1990)
25. Silverman, B.: Some aspects of the spline smoothing approach to nonparametric regression. J. Royal Statist. Soc. Ser. B Stat. Methodol. **47**, 1–52 (1985)
26. Smale, S., Zhou, D.X.: Estimating the approximation error in learning theory. Anal. Appl. **1**, 17–41 (2003)
27. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2**, 67–93 (2001)
28. Steinwart, I.: Support vector machines are universally consistent. J. Complexity **18**, 768–791 (2002)
29. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
30. Steinwart, I., Christmann, A.: Estimating conditional quantiles with the help of the pinball loss. Bernoulli **17**, 211–225 (2011)
31. Steinwart, I., Hush, D., Scovel, C.: Optimal rates for regularized least squares regression. In: Proceedings of the 22nd Annual Conference on Learning Theory, pp. 79–93 (2009)
32. Steinwart, I., Scovel, C.: Fast rates for support vector machines using Gaussian kernels. Ann. Statist. **35**, 575–607 (2007)
33. Steinwart, I., Scovel, C.: Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. Constr. Approx. **35**, 363–417 (2012)
34. Stone, C.: Consistent nonparametric regression. Ann. Statist. **5**, 595–645 (1977)
35. Tsybakov, A.B.: Optimal aggregation of classifiers in statistical learning. Ann. Statist. **32**, 135–166 (2004)
36. Vapnik, V., Golowich, S., Smola, A.: Support vector method for function approximation, regression estimation, and signal processing. In: Advances in Neural Information Processing Systems 9, pp. 81–287 (1997)
37. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag, New York (1995)
38. Vapnik, V.N., Lerner, A.: Pattern recognition using generalized portrait method. Autom. Remote Control **24**, 774–780 (1963)
39. Wahba, G.: Spline Models for Observational Data. Series in Applied Mathematics 59, SIAM, Philadelphia (1990)
40. Zhang, T.: Statistical behaviour and consistency of classification methods based on convex risk minimization. Ann. Statist. **32**, 56–134 (2004)

Ingo Steinwart
Universität Stuttgart
Fachbereich Mathematik
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: Ingo.Steinwart@mathematik.uni-stuttgart.de
WWW: http://www.isa.uni-stuttgart.de/Steinwart

Erschienene Preprints ab Nummer 2007/2007-001

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

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