Journal of Machine Learning Research 10 (2009) 2115-2132 Submitted 12/06; Revised 2/08; Published 9/09

Deterministic Error Analysis of Support Vector Regression and Related Regularized Kernel Methods

Christian Rieger RIEGER@INS.UNI-BONN.DE
Institute for Numerical Simulation & Hausdorff Center for Mathematics
University of Bonn
Wegelerstr. 6, 53115 Bonn, Germany

Barbara Zwicknagl BARBARA.ZWICKNAGL@HCM.UNI-BONN.DE
Institute for Applied Mathematics & Hausdorff Center for Mathematics
University of Bonn
Endenicher Allee 60, 53115 Bonn, Germany

Editor: Bernhard Schölkopf

Abstract

We introduce a new technique for the analysis of kernel-based regression problems. The basic tools are sampling inequalities which apply to all machine learning problems involving penalty terms induced by kernels related to Sobolev spaces. They lead to explicit deterministic results concerning the worst case behaviour of ε- and ν-SVRs. Using these, we show how to adjust regularization parameters to get best possible approximation orders for regression. The results are illustrated by some numerical examples.

Keywords: sampling inequality, radial basis functions, approximation theory, reproducing kernel Hilbert space, Sobolev space

1. Introduction

Support Vector (SV) machines and related kernel-based algorithms are modern learning systems motivated by results of statistical learning theory as introduced by Vapnik (1995). The concept of SV machines is to provide a prediction function which is accurate on the given training data and which is sparse in the sense that it can be written in terms of a typically small subset of all samples, called the support vectors, as stated by Schölkopf et al. (1995). Therefore, SV regression and classification algorithms are closely related to regularized problems from classical approximation theory as pointed out by Girosi (1998) and Evgeniou et al. (2000) who had applied techniques from functional analysis to derive probabilistic error bounds for SV regression.
This paper provides a theoretical framework to derive deterministic error bounds for some popular SV machines. We show how a sampling inequality by Wendland and Rieger (2005) can be used to bound the worst-case generalization error for the ν- and the ε-regression without making any statistical assumptions on the inaccuracy of the training data. In contrast to the literature, our error bounds explicitly depend on the pointwise noise in the data. Thus they can be used for any subsequent probabilistic analysis modelling certain assumptions on the noise distribution.
The paper is organized as follows. In the next section we recall some basic facts about reproducing kernels in Hilbert spaces. Section 3 deals with regularized approximation problems in Hilbert spaces with reproducing kernels and outlines the connection to classical SV regression (SVR) al-

gorithms. We provide a deterministic error analysis for the ν- and the ε-SVR for both exact and inexact training data. Our analytical results showing optimal convergence orders in Sobolev spaces are illustrated by numerical experiments.

2. Reproducing Kernels in Hilbert Spaces

We suppose that K is a positive definite kernel on some domain $\Omega \subset \mathbb{R}^d$ which should contain at least one point. To start with, we briefly recall the well known definition of a reproducing kernel in a Hilbert space. In the following we shall use the notation that bold letters denote vectors, that is $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$.

Definition 1 *Let $\mathcal{H}(\Omega)$ be a Hilbert space of functions $f : \Omega \to \mathbb{R}$. A function $K : \Omega \times \Omega \to \mathbb{R}$ is called reproducing kernel of $\mathcal{H}(\Omega)$, if*

- $K(\mathbf{y}, \cdot) \in \mathcal{H}(\Omega)$ *for all* $\mathbf{y} \in \Omega$ *and*
- $f(\mathbf{y}) = (f, K(\mathbf{y}, \cdot))_{\mathcal{H}(\Omega)}$ *for all* $f \in \mathcal{H}(\Omega)$ *and all* $\mathbf{y} \in \Omega$.

For each positive definite kernel $K : \Omega \times \Omega \to \mathbb{R}$ there exists a unique Hilbert space $\mathcal{N}_K(\Omega)$ of functions $f : \Omega \to \mathbb{R}$, such that K is the reproducing kernel of $\mathcal{N}_K(\Omega)$ (see Wendland, 2005, Theorems 10.1 and 10.11). This Hilbert space $\mathcal{N}_K(\Omega)$ is called the *native space of K*. Though this definition of a native space is rather abstract, it can be shown that in some cases the native spaces coincide with classical function spaces.
From now on we shall only consider *radial* kernels K, that is,

$$K(\mathbf{x}, \mathbf{y}) = K(\|\mathbf{x} - \mathbf{y}\|) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d ,$$

where we use the same notation for the kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and for the function $K : \mathbb{R}^d \to \mathbb{R}$. We hope that this does not cause any confusion. We shall mainly focus on continuous kernels $K \in L_1(\Omega)$, that is,

$$\|K\|_{L_1(\Omega)} := \int_\Omega |K(\mathbf{x})| \, d\mathbf{x} < \infty .$$

For $K \in L_1(\mathbb{R}^d)$, we define the Fourier transform $\hat{K}$ by

$$\hat{K}(\omega) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} K(\mathbf{x}) e^{-i\mathbf{x}\cdot\omega} d\mathbf{x} , \ \omega \in \mathbb{R}^d .$$

For the case $\Omega = \mathbb{R}^d$ there is the following characterization of native spaces of certain radial kernels $K : \Omega \to \mathbb{R}^d$ (Wendland, 2005, Theorem 10.12).

Theorem 2 *Suppose that $K \in C(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ is a real-valued and positive definite radial kernel. Then the native space of K is given by*

$$\begin{aligned} \mathcal{N}_K(\mathbb{R}^d) &= \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \, : \, \frac{\hat{f}}{\sqrt{\hat{K}}} \in L_2(\mathbb{R}^d) \right\} , \\ (f, g)_{\mathcal{N}_K(\mathbb{R}^d)} &= (2\pi)^{-d/2} \left(\frac{\hat{f}}{\sqrt{\hat{K}}}, \frac{\hat{g}}{\sqrt{\hat{K}}} \right)_{L_2(\mathbb{R}^d)} , \end{aligned}$$

where $\hat{f}$ denotes the Fourier transform of f.

We recall that the Sobolev spaces $W_2^s(\mathbb{R}^d)$ on $\mathbb{R}^d$ with $s \geq 0$ are given by

$$W_2^s(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) \ : \ \hat{f}(\cdot)(1+\|\cdot\|_2^2)^{s/2} \in L_2(\mathbb{R}^d) \right\} . \tag{1}$$

Therefore for a radial kernel function K whose Fourier transform decays like

$$c_1(1+\|\cdot\|_2^2)^s \leq \hat{K} \leq c_2(1+\|\cdot\|_2^2)^s \ \ , s > d/2 \tag{2}$$

for some constants $c_1, c_2 > 0$, the associated native space $\mathcal{N}_K(\mathbb{R}^d)$ is $W_2^s(\mathbb{R}^d)$ with an equivalent norm. There are several examples of kernels satisfying the condition (2). One famous example for fixed $s \in (d/2, \infty)$ is the *Matern kernel* (Wendland, 2005)

$$K_s(\mathbf{x}) := \frac{2^{1-s}}{\Gamma(s)} \|\mathbf{x}\|_2^{s-d/2} \mathcal{K}_{d/2-s}(\|\mathbf{x}\|_2) ,$$

where $\mathcal{K}$ denotes the Bessel function of the third kind. In our examples, however, we focus on *Wendland's functions* (Wendland, 2005). They are very convenient to implement since they are compactly supported and piecewise polynomials. Such nice reproducing kernels are so far only available for certain choices of the space dimension d and the decay parameter s (see Wendland, 2005), but a recent result by Schaback (2009) covers almost all cases of practical interest. We shall explain some more properties of these kernels in the experimental part, see Section 10, and refer to the recent monograph by Wendland (2005) for details.
In order to establish the equivalence of native spaces and Sobolev spaces on bounded domains one needs certain extension theorems for Sobolev functions on bounded domains (see Wendland, 2005).

Definition 3 *Let $\Omega \subset \mathbb{R}^d$ be a domain. We define the Sobolev spaces of integer orders $k \in \mathbb{N}$ as*

$$W_2^k(\Omega) = \{ f \in L_2(\Omega) \ : \ f \text{ has weak derivatives } D^\alpha f \in L_2(\Omega) \text{ of order } |\alpha| \leq k \}$$

with the norm

$$\|u\|_{W_2^k(\Omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L_2(\Omega)}^2 \right)^{1/2} .$$

For fractional smoothness $s = k + \sigma$ with $0 < \sigma < 1$ and $k \in \mathbb{N}$ we define the semi-norm

$$|u|_{W_2^s(\Omega)} := \left(\sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha u(\mathbf{x}) - D^\alpha u(\mathbf{y})|^2}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2\sigma}} d\mathbf{x} d\mathbf{y} \right)^{1/2} ,$$

and set

$$W_2^s(\Omega) := \left\{ u \in L_2(\Omega) \ : \ \left(\|u\|_{W_2^k(\Omega)}^2 + |u|_{W_2^s(\Omega)}^2 \right)^{1/2} < \infty \right\} .$$

In the case $\Omega = \mathbb{R}^d$ this space is known to be equivalent to the space given by (1) in terms of Fourier transforms (for more details on these spaces, see Wloka, 1982). Finally, Wendland (2005) proves the following equivalence for domains having Lipschitz boundaries. Roughly speaking, a set $\Omega \subset \mathbb{R}^d$ has a Lipschitz boundary if its boundary is locally (in a suitable direction) the graph of a Lipschitz function such that Ω lies completely on one hand-side of this graph (see Brenner and Scott, 1994). Then there is the following theorem (see Wendland, 2005, Cor. 10.48).

Theorem 4 *Suppose that $K \in L_1(\mathbb{R}^d)$ has a Fourier transform that decays as $(1+\|\cdot\|_2^2)^{-s}$ for $s > d/2$. Suppose that Ω has a Lipschitz boundary. Then*

$$\mathcal{N}_K(\Omega) \cong W_2^s(\Omega)$$

with equivalent norms.

3. Regularized Problems in Native Hilbert Spaces

In the native Hilbert spaces we consider the following learning or recovery problem. We assume that we are given (possibly only approximate) function values $y_1,\dots,y_N \in \mathbb{R}$ of an unknown function $f \in \mathcal{N}_K(\Omega)$ on some scattered points $X := \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(N)}\} \subset \Omega$, that is $f(\mathbf{x}^{(j)}) \approx y_j$ for $j = 1,\dots,N$. In the following we shall use the notation that bold letters denote vectors, that is $\mathbf{v} = (v_1,\dots,v_d)^T \in \mathbb{R}^d$.
To control accuracy and complexity of the reconstruction simultaneously, we use the optimization problem

$$\min_{\substack{s\in\mathcal{N}_K(\Omega)\\ \varepsilon\in\mathbb{R}^+}} \frac{1}{N}\sum_{j=1}^{N} V_\varepsilon\left(\left|s\left(\mathbf{x}^{(j)}\right)-y_j\right|\right) + \frac{1}{2C}\|s\|^2_{\mathcal{N}_K(\Omega)}\ , \tag{3}$$

where $C > 0$ is a positive parameter and V_ε denotes a positive function which may be parametrized by a positive real number ε. We point out that V_ε need not be a classical loss function. Therefore we shall give some proofs of results which were formulated by Schölkopf and Smola (2002) in the case of V_ε being a loss function.

Theorem 5 (Representer theorem) *If $(s_{X,\mathbf{y}},\varepsilon^*)$ is a solution of the optimization problem (3), then there exists a vector $\mathbf{w} \in \mathbb{R}^N$ such that*

$$s_{X,\mathbf{y}}(\cdot) = \sum_{j=1}^{N} w_j K\left(\mathbf{x}^{(j)},\cdot\right)\ ,$$

that is $s_{X,\mathbf{y}} \in \operatorname{span}\{K(\mathbf{x}^{(1)},\cdot),\dots,K(\mathbf{x}^{(N)},\cdot)\}$.

Proof For the readers' convenience, we repeat the proof from Schölkopf and Smola (2002) in our specific situation. Every $s \in \mathcal{N}_K(\Omega)$ can be decomposed into two parts $s = s_{||} + s_\perp$, where $s_{||}$ is contained in the linear span of $\{K(\mathbf{x}^{(1)},\cdot),\dots,K(\mathbf{x}^{(N)},\cdot)\}$, and $s_\perp$ is contained in the orthogonal complement, that is $\langle s_{||}, s_\perp\rangle_{\mathcal{N}_K(\Omega)} = 0$. By the reproducing property of the kernel K in the native space, the problem (3) can be rewritten as

$$\min_{\substack{s=s_{||}+s_\perp\\ \varepsilon\in\mathbb{R}^+}} \frac{1}{N}\sum_{j=1}^{N} V_\varepsilon\left(\left|\left\langle s_{||}, K\left(\mathbf{x}^{(j)},\cdot\right)\right\rangle - y_j\right|\right) + \frac{1}{2C}\left\|s_{||}\right\|^2_{\mathcal{N}_K(\Omega)} + \frac{1}{2C}\|s_\perp\|^2_{\mathcal{N}_K(\Omega)}\ .$$

Therefore a solution $(s_{X,\mathbf{y}},\varepsilon^*)$ of the optimization problem (3) satisfies $(s_{X,\mathbf{y}})_\perp = 0$, which implies $s_{X,\mathbf{y}} \in \operatorname{span}\{K(\mathbf{x}^{(1)},\cdot),\dots,K(\mathbf{x}^{(N)},\cdot)\}$. ■

Since the proof of Theorem 5 does not depend on the minimality with respect to ε this result holds also true if ε is a fixed parameter instead of a primal variable. To be precise we state this result as a corollary.

Corollary 6 *If $s_{X,\mathbf{y}}$ is a solution of the optimization problem*

$$\min_{s\in\mathcal{N}_K(\Omega)} \frac{1}{N}\sum_{j=1}^{N} V_\varepsilon\left(\left|s\left(\mathbf{x}^{(j)}\right)-y_j\right|\right)+\frac{1}{2C}\|s\|^2_{\mathcal{N}_K(\Omega)}\ , \tag{4}$$

with $\varepsilon\in\mathbb{R}^+$ being a fixed parameter, then $s_{X,\mathbf{y}}\in span\left\{K\left(\mathbf{x}^{(1)},\cdot\right),\ldots,K\left(\mathbf{x}^{(N)},\cdot\right)\right\}$.

The representer theorems can be used to reformulate infinite-dimensional optimization problems of the forms (3) or (4) in a finite-dimensional setting (see Schölkopf and Smola, 2002).

4. Support Vector Regression

As a first optimization problem of the form (3) we consider the ν-SVR which was introduced by Schölkopf et al. (2000). The function $V_\varepsilon(\mathbf{x})=|\mathbf{x}|_\varepsilon+\varepsilon\nu$ is related to Vapnik's ε-intensive loss function (Vapnik, 1995)

$$|\mathbf{x}|_\varepsilon=\begin{cases}0 & if\ |\mathbf{x}|\le\varepsilon\\ |\mathbf{x}|-\varepsilon & if\ |\mathbf{x}|>\varepsilon\end{cases}\ ,$$

but has an additional term with a positive parameter ν. The associated optimization problem is called ν-SVR and takes the form

$$\min_{\substack{s\in\mathcal{N}_K(\Omega)\\ \varepsilon\in\mathbb{R}^+}} \frac{1}{N}\sum_{j=1}^{N}\left|s\left(\mathbf{x}^{(j)}\right)-y_j\right|_\varepsilon+\varepsilon\nu+\frac{1}{2C}\|s\|^2_{\mathcal{N}_K(\Omega)}\ . \tag{5}$$

Theorem 7 *The optimization problem (5) possesses a solution $\left(s^{(\nu)}_{X,\mathbf{y}},\varepsilon^*\right)$.*

Proof This follows from a general result by Micchelli and Pontil (2005). The problem (5) is equivalent to the optimization problem

$$\min_{\substack{s\in\mathcal{N}_K(\Omega)\\ \delta\in\mathbb{R}}} \frac{1}{N}\sum_{j=1}^{N}\left|s\left(\mathbf{x}^{(j)}\right)-y_j\right|_{\delta^2}+\delta^2\nu+\frac{1}{2C}\|s\|^2_{\mathcal{N}_K(\Omega)}\ . \tag{6}$$

If we set $\mathcal{H}:=\mathcal{N}_K(\Omega)\times\mathbb{R}$ we can define an inner product on $\mathcal{H}$ by

$$\langle h_1,h_2\rangle_{\mathcal{H}}:=\langle f_1,f_2\rangle_{\mathcal{N}_K(\Omega)}+2C\nu\langle r_1,r_2\rangle_{\mathbb{R}}$$

for $h_j=(f_j,r_j)$, $j=1,2$. To make $\mathcal{H}$ a space of functions we use the canonical identification of $\mathbb{R}$ with the space of constant functions $\mathbb{R}\to\mathbb{R}$. The Hilbert space $\mathcal{H}$ then has the reproducing kernel $\tilde{K}:=\left(K,\frac{1}{2C\nu}\mathbf{1}\right)$ where $\mathbf{1}$ denotes the constant function which maps everything to 1, that is $\tilde{K}((\mathbf{x},r),(\mathbf{y},s))=K(\mathbf{x},\mathbf{y})+1/(2C\nu)$ for all $r,s\in\mathbb{R}$. With this notation the problem (6) can be rewritten as

$$\min_{(s,\delta)\in\mathcal{H}} Q^{\mathbf{y}}\left(I_X(s,\delta)\right)+\frac{1}{2C}\|(s,\delta)\|^2_{\mathcal{H}}\ , \tag{7}$$

where

$$I_X(s,\delta):=\left(s(\mathbf{x}^{(1)}),\ldots,s(\mathbf{x}^{(N)}),\delta\right)^T\in\mathbb{R}^{N+1}$$

and

$$Q^{\mathbf{y}} : \mathbb{R}^{N+1} \to \mathbb{R}, \quad Q^{\mathbf{y}}(\mathbf{p}, \delta) = \frac{1}{N} \sum_{j=1}^{N} \left| p_j - y_j \right|_{\delta^2}.$$

Since $Q^{\mathbf{y}}$ is continuous on $\mathbb{R}^{N+1}$ for all $\mathbf{y} \in \mathbb{R}^N$, the problem (7) possesses a solution as shown by Micchelli and Pontil (2005). ■

If we introduce the slack variables $\xi, \xi^* \in \mathbb{R}^N$, the representer theorem gives us an equivalent finite-dimensional problem which was considered by Schölkopf et al. (2000).

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^{\mathbf{N}} \\ \xi^*, \xi \in \mathbb{R}^N \\ \varepsilon \in \mathbb{R}^+}} \frac{1}{2} \mathbf{w}^T \mathbf{K} \mathbf{w} + C \left(\nu \varepsilon + \frac{1}{N} \sum_{j=1}^{N} \left(\xi_j + \xi_j^* \right) \right)$$

$$\begin{aligned} \text{subject to} \quad (\mathbf{K}\mathbf{w})_j - y_j &\leq \varepsilon + \xi_j \,, \\ (-\mathbf{K}\mathbf{w})_j + y_j &\leq \varepsilon + \xi_j^* \,, \\ \xi_j^*, \xi_j \geq 0 \,, \qquad \varepsilon &\geq 0 \quad \text{for } 1 \leq j \leq N \,, \end{aligned} \tag{8}$$

where

$$\mathbf{K} = \left(K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right) \right)_{i,j=1\ldots N}$$

denotes the Gram matrix of the kernel K. We will use this equivalent problem for implementation and our numerical tests.

A particularly interesting problem arises if we skip the parameter ν and let ε be fixed. Then the optimization problem (8) takes the form

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^{\mathbf{N}} \\ \xi^*, \xi \in \mathbb{R}^N}} \frac{1}{2} \mathbf{w}^T \mathbf{K} \mathbf{w} + C \frac{1}{N} \sum_{j=1}^{N} \left(\xi_j + \xi_j^* \right)$$

$$\begin{aligned} \text{subject to} \quad (\mathbf{K}\mathbf{w})_j - y_j &\leq \varepsilon + \xi_j \,, \\ (-\mathbf{K}\mathbf{w})_j + y_j &\leq \varepsilon + \xi_j^* \,, \\ \xi_j^*, \xi_j &\geq 0 \quad \text{for } 1 \leq j \leq N \,. \end{aligned} \tag{9}$$

Schölkopf et al. (2000) called this problem ε-SVR. Similarly to the ν-SVR, the problem (9) can be formulated as a regularized minimization problem in a Hilbert space (Evgeniou et al., 2000), namely

$$\min_{s \in \mathcal{N}_K(\Omega)} \frac{1}{N} \sum_{j=1}^{N} \left| s\left(\mathbf{x}^{(j)} \right) - y_j \right|_{\varepsilon} + \frac{1}{2C} \left\| s \right\|_{\mathcal{N}_K(\Omega)}^2 \,. \tag{10}$$

Like the ν-SVR, this optimization problem possesses a solution (see Micchelli and Pontil, 2005, Lemma 1).

5. A Sampling Inequality

We shall employ a special case of a *sampling inequality* introduced by Wendland and Rieger (2005). It requires the following assumptions which we need from now on. Let $\Omega \subset \mathbb{R}^d$ be a bounded

domain with Lipschitz boundary that satisfies an interior cone condition. A domain Ω is said to satisfy an interior cone condition with radius $r > 0$ and angle $\theta \in \left(0, \frac{\pi}{2}\right)$ if for every $\mathbf{x} \in \Omega$ there is a unit vector $\xi(\mathbf{x})$ such that the cone

$$C(\mathbf{x}, \xi(\mathbf{x}), \theta, r) := \left\{ \mathbf{x} + \lambda \mathbf{y} \ : \ \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_2 = 1, \mathbf{y}^T \xi(\mathbf{x}) \geq \cos(\theta), \lambda \in [0, r] \right\}$$

is contained in Ω. In particular, a domain which satisfies an interior cone condition cannot have any outward cusps. We shall assume for the rest of this paper that Ω satisfies an interior cone condition with radius $R_{\max}$ and angle θ. We shall derive estimates that are valid only if the training points are sufficiently dense in Ω. To make this condition precise, we will need a slightly unhandy constant which depends only on the geometry of Ω, namely (see Wendland, 2005)

$$C_\Omega := \frac{\sin\left(2\arcsin\left(\frac{\sin\theta}{4(1+\sin\theta)}\right)\right)\sin\theta}{8\left(1+\sin\left(2\arcsin\left(\frac{\sin\theta}{4(1+\sin\theta)}\right)\right)\right)(1+\sin\theta)} R_{\max} \ .$$

Suppose that K is a radial kernel function such that the native Hilbert space of K is norm-equivalent to a Sobolev space, that is $\mathcal{N}_K(\Omega) = W_2^\tau(\Omega)$. Here we assume that $\lfloor \tau - \frac{1}{2} \rfloor > d/2$, where we use the notation $\lfloor t \rfloor := \max\{n \in \mathbb{N}_0 : n \leq t\}$ for $t \geq 0$. Furthermore, let $X = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\} \subset \Omega$ be a finite set with sufficiently small fill distance

$$h := h_{X,\Omega} := \sup_{\mathbf{x} \in \Omega} \min_{\mathbf{x}^{(j)} \in X} \left\| \mathbf{x} - \mathbf{x}^{(j)} \right\|_2 \ .$$

The fill distance can be interpreted geometrically as the radius of the largest ball with center in $\overline{\Omega}$ that does not contain any of the points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$. It is a useful quantity for the deterministic error analysis in Sobolev spaces. The case $h = 0$ implies that $X = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\}$ is dense in Ω, and therefore convergence is studied for the limit $h \to 0$ which means that the domain Ω is equally filled with points from X. Let us explain the relation to the usual error bounds in terms of the number of points N. In the case of regularly distributed points we have that $h = cN^{-\frac{1}{d}}$ with some constant $c > 0$ (Wendland, 2005). Therefore the limit $h \to 0$ is equivalent to the limit $N \to \infty$ which is the more intuitive meaning of asymptotic convergence. But there is a drawback, since the error bounds in terms of N depend crucially on the space dimension d, while error bounds in terms of the fill distance h are dominated by the smoothness of the function to be learned. We will comment on this again later for the special error bounds we consider here. We shall use the following result by Wendland and Rieger (2005).

Theorem 8 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary that satisfies an interior cone condition. Let τ be a positive real number with $\lfloor \tau - \frac{1}{2} \rfloor > \frac{d}{2}$, and let $1 \leq q \leq \infty$. Then there exists a positive constant $C > 0$ such that for all discrete sets $X \subset \Omega$ with sufficiently small fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$ the inequality*

$$\|u\|_{L_q(\Omega)} \leq C \left(h^{\tau - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|u\|_{W_2^\tau(\Omega)} + \|u|_X\|_{\ell_\infty(X)} \right)$$

holds for all $u \in W_2^\tau(\Omega)$, where we use the notation $(t)_+ := \max\{0, t\}$.

We shall apply this theorem to the residual function $f - s_{X,\mathbf{y}}$ of the function $f \in W_2^\tau(\Omega)$ to be recovered and a solution $s_{X,\mathbf{y}} \in W_2^\tau(\Omega)$ of the regression problem. In our applications we shall focus on the two main cases $q = \infty$ and $q = 2$. Other cases can be treated analogously. It will turn out that we get optimal convergence rates in the noiseless case. In presence of noise the resulting error will explicitly be bounded in terms of the noise in the data.

6. ν-SVR with Exact Data

In order to derive error bounds for the ν-SVR optimization problem (5) we shall apply Theorem 8 to the residual $f - s^{(\nu)}_{X,\mathbf{y}}$, where $\left(s^{(\nu)}_{X,\mathbf{y}}, \varepsilon^*\right)$ denotes a solution to the problem (5) for $X := \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\} \subset \Omega$ and $\mathbf{y} \in \mathbb{R}^N$. In this section we consider exact data, that is

$$f\left(\mathbf{x}^{(j)}\right) = y_j \quad \text{for } j = 1, \ldots, N \tag{11}$$

for a function $f \in W_2^\tau(\Omega) \cong \mathcal{N}_K(\Omega)$. As pointed out by Wendland and Rieger (2005) we first need a stability and a consistency estimate for the solution $s^{(\nu)}_{X,\mathbf{y}}$.

Lemma 9 *Under the assumption (11) concerning the data, we find that for every X a solution $\left(s^{(\nu)}_{X,\mathbf{y}}, \varepsilon^*\right)$ to problem (5) satisfies*

$$\begin{aligned} \left\|s^{(\nu)}_{X,\mathbf{y}}\right\|_{\mathcal{N}_K(\Omega)} &\leq \|f\|_{\mathcal{N}_K(\Omega)} \quad \textit{and} \\ \left\|s^{(\nu)}_{X,\mathbf{y}}|_X - \mathbf{y}\right\|_{\ell_\infty(X)} &\leq \frac{N}{2C}\|f\|^2_{\mathcal{N}_K(\Omega)} + \varepsilon^* \cdot (1 - N\nu) \ . \end{aligned}$$

Proof We denote the objective function of the optimization problem (5) by

$$H^{\mathbf{y}}_{C,\nu}(s,\varepsilon) := \frac{1}{N}\sum_{j=1}^{N}\left|s\left(\mathbf{x}^{(j)}\right) - y_j\right|_\varepsilon + \nu\varepsilon + \frac{1}{2C}\|s\|^2_{\mathcal{N}_K(\Omega)} \ , \tag{12}$$

and the interpolant to f with respect to X and K with I_f, that is $I_f|_X = \mathbf{y}$ and $I_f \in \operatorname{span}\left\{K\left(\mathbf{x}^{(1)}, \cdot\right), \ldots, K\left(\mathbf{x}^{(N)}, \cdot\right)\right\}$. With this notation we have

$$\frac{1}{2C}\left\|s^{(\nu)}_{X,\mathbf{y}}\right\|^2_{\mathcal{N}_K(\Omega)} \leq H^{\mathbf{y}}_{C,\nu}\left(s^{(\nu)}_{X,\mathbf{y}}, \varepsilon^*\right) \leq H^{\mathbf{y}}_{C,\nu}(I_f, 0) = \frac{1}{2C}\left\|I_f\right\|^2_{\mathcal{N}_K(\Omega)} \leq \frac{1}{2C}\|f\|^2_{\mathcal{N}_K(\Omega)}$$

since $\left\|I_f\right\|_{\mathcal{N}_K(\Omega)} \leq \|f\|_{\mathcal{N}_K(\Omega)}$ (Wendland, 2005), which implies the first claim.
Furthermore we have for $i = 1, \ldots, N$

$$\begin{aligned} \left|s^{(\nu)}_{X,\mathbf{y}}\left(\mathbf{x}^{(i)}\right) - y_i\right| &\leq \sum_{j=1}^{N}\left|s^{(\nu)}_{X,\mathbf{y}}\left(\mathbf{x}^{(j)}\right) - y_j\right|_{\varepsilon^*} + \varepsilon^* \leq N H^{\mathbf{y}}_{C,\nu}\left(s^{(\nu)}_{X,\mathbf{y}}, \varepsilon^*\right) + \varepsilon^*(1 - N\nu) \\ &\leq N H^{\mathbf{y}}_{C,\nu}(I_f, 0) + \varepsilon^*(1 - N\nu) \leq \frac{N}{2C}\left\|I_f\right\|^2_{\mathcal{N}_K(\Omega)} + \varepsilon^*(1 - N\nu) \\ &\leq \frac{N}{2C}\|f\|^2_{\mathcal{N}_K(\Omega)} + \varepsilon^*(1 - N\nu) \ , \end{aligned}$$

which finishes the proof. ■

With Theorem 8 we find immediately the following result.

Theorem 10 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary that satisfies an interior cone condition. Let τ be a positive real number with $\lfloor \tau - \frac{1}{2} \rfloor > \frac{d}{2}$ and $1 \leq q \leq \infty$. We suppose*

$f \in W_2^\tau(\Omega)$ with $f(\mathbf{x}^{(i)}) = y_i$. Let $\left(s_{X,\mathbf{y}}^{(\nu)}, \varepsilon^*\right)$ be a solution of the ν-SVR. Then there is a constant $\tilde{C} > 0$, which depends on τ, d and Ω but not on f or X, such that the approximation error can be bounded by

$$\left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{L_q(\Omega)} \leq \tilde{C} \left(2h^{\tau - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|f\|_{W_2^\tau(\Omega)} + \frac{N}{2C} \|f\|^2_{W_2^\tau(\Omega)} + (1 - N\nu) \cdot \varepsilon^* \right)$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$.

Proof Combining Lemma 9 and Theorem 8 leads to

$$\begin{aligned}
\left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{L_q(\Omega)} &\leq \tilde{C} \left(h^{\tau - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{W_2^\tau(\Omega)} + \left\| \mathbf{y} - s_{X,\mathbf{y}}^{(\nu)}|_X \right\|_{\ell_\infty(X)} \right) \\
&\leq \tilde{C} \left(h^{\tau - \left(\frac{d}{2} - \frac{d}{q}\right)_+} \left(\|f\|_{W_2^\tau(\Omega)} + \left\| s_{X,\mathbf{y}}^{(\nu)} \right\|_{W_2^\tau(\Omega)} \right) + \left\| \mathbf{y} - s_{X,\mathbf{y}}^{(\nu)}|_X \right\|_{\ell_\infty(X)} \right) \\
&\leq \tilde{C} \left(2h^{\tau - \left(\frac{d}{2} - \frac{d}{q}\right)_+} \|f\|_{W_2^\tau(\Omega)} + \frac{N}{2C} \|f\|^2_{W_2^\tau(\Omega)} + (1 - N\nu)\, \varepsilon^* \right) .
\end{aligned}$$

■

At first glance the term containing ε^* seems to be odd because it could be uncontrollable. But according to Chang and Lin (2002) we can at least assume ε^* to be bounded by

$$\varepsilon^* \leq \frac{1}{2} \left(\max_{i=1,\ldots,N} y_i - \min_{i=1,\ldots,N} y_i \right) .$$

If this inequality is not satisfied, the problem (8) possesses only the trivial solution $s \equiv 0$ which is not interesting. Furthermore, we see that the ε^*-term occurs with a factor $(1 - N\nu)$, which can be used to control this term. If we choose $\nu \geq \frac{1}{N}$, the term $(1 - N\nu)\varepsilon^*$ vanishes or is even negative. The parameter ν is a lower bound on the fraction of support vectors (see Schölkopf et al., 2000), and hence $\nu = 1/N$ means to get at least one support vector, that is a non-trivial solution. Since we are not interested in the case of trivial solutions, the condition $\nu \geq 1/N$ is a reasonable assumption. On the other hand, we can use the results from Lemma 9 to derive a more explicit upper bound on $\varepsilon^* = \varepsilon^*(C, \nu, f)$ by

$$0 \leq \left\| s_{X,\mathbf{y}}^{(\nu)}|_X - \mathbf{y} \right\|_{\ell_\infty(X)} \leq \frac{N}{2C} \|f\|^2_{\mathcal{N}_k(\Omega)} + \varepsilon^* (1 - N\nu) .$$

If we assume $\nu > 1/N$, this leads to

$$\varepsilon^* = \varepsilon^*(C, \nu, f) \leq \frac{N}{2C(N\nu - 1)} \|f\|^2_{\mathcal{N}_k(\Omega)} .$$

Note that these bounds cannot be used for a better parameter choice, since we would need to rearrange this inequality and solve for C or ν. This would only be possible if there were lower bounds on ε^* as well. Moreover, the parameter C appears in our error bound as a factor $\frac{N}{2C}$ which implies that we expect convergence only in the case $C \to \infty$. In this case ε^* will be small, as can be deduced from problem (8).

We shall now make our bounds more explicit for the case of quasi-uniformly distributed points. In this case the number of points N and the fill distance h are related to each other by

$$c_1 N^{-1/d} \leq h \leq c_2 N^{-1/d} \ , \tag{13}$$

where c_1 and c_2 denote positive constants (see Wendland, 2005, Proposition 14.1).

Corollary 11 *In case of quasi-uniform exact data we can choose the problem parameters as*

$$C = \frac{N \|f\|_{W_2^\tau(\Omega)}}{2h^\tau} \approx h^{-(\tau+d)} \|f\|_{W_2^\tau(\Omega)} \text{ and } \nu \geq \frac{1}{N}$$

to get

$$\left\| f - s^{(\nu)}_{X,\mathbf{y}} \right\|_{L_2(\Omega)} \leq \tilde{C} h^\tau \|f\|_{W_2^\tau(\Omega)} \leq \tilde{C} N^{-\frac{\tau}{d}} \|f\|_{W_2^\tau(\Omega)} \ ;$$

or as

$$C = \frac{N \|f\|_{W_2^\tau(\Omega)}}{2h^{\tau-\frac{d}{2}}} \approx h^{-\left(\tau+\frac{d}{2}\right)} \|f\|_{W_2^\tau(\Omega)} \text{ and } \nu \geq \frac{1}{N}$$

to get

$$\left\| f - s^{(\nu)}_{X,\mathbf{y}} \right\|_{L_\infty(\Omega)} \leq \tilde{C} h^{\tau-\frac{d}{2}} \|f\|_{W_2^\tau(\Omega)} \leq \tilde{C} N^{-\frac{\tau}{d}+\frac{1}{2}} \|f\|_{W_2^\tau(\Omega)}$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$, with generic positive constants $\tilde{C}$ which depend on τ, d, Ω but not on f or X.

Note that these bounds yield arbitrarily high convergence orders, provided that the functions are smooth enough, that is τ is large enough. Therefore they are in this setting better than the usual minimax rate $N^{-\frac{2\tau}{2\tau+d}}$ (see Stone, 1982). In the following we shall only give our error estimates in terms of the fill distance h rather than in terms of the number of points N. This is due to the fact that the approximation rate τ in h is independent of the space dimension d. However it should be clear how the approximation rates translate into error estimates in terms of N in the case of quasi-uniform data due to the inequality (13). Note that the parameter choice in the case of arbitrary, non-uniformly distributed data can be treated analogously.
Corollary 11 shows, that the solution of the ν-SVR leads to the same approximation orders with respect to the fill distance h as classical kernel-based interpolation (see Wendland, 2005). But the ν-SVR allows for much more flexibility and less complicated solutions. Our numerical results will confirm these convergence rates.

7. ν-SVR with Inexact Data

In this section we denote again by $\left(s^{(\nu)}_{X,\mathbf{y}}, \varepsilon^*\right)$ the solution to the problem (5) for a set of points $X := \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\} \subset \Omega$ and $\mathbf{y} \in \mathbb{R}^N$, but we allow the given data to be corrupted by some additive error $\mathbf{r} = (r_1, \ldots, r_N)$, that means

$$f\left(\mathbf{x}^{(j)}\right) = y_j + r_j \quad \text{for } j = 1, \ldots, N, \tag{14}$$

where is $f \in W_2^\tau(\Omega) \cong \mathcal{N}_K(\Omega)$. Note that there are no assumptions concerning the error distribution. As in the previous section we have to show a stability and a consistency estimate of the following form.

Lemma 12 *Under the assumption (14) concerning the data* $\mathbf{y}$*, a solution* $\left(s_{X,\mathbf{y}}^{(\nu)}, \varepsilon^*\right)$ *to the optimization problem (5) satisfies for every* X *and for all* $\varepsilon \geq 0$

$$\begin{aligned} \left\| s_{X,\mathbf{y}}^{(\nu)} \right\|_{\mathcal{N}_K(\Omega)} &\leq \sqrt{\frac{2C}{N} \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + 2C\nu\varepsilon + \|f\|_{\mathcal{N}_K(\Omega)}^2} \quad \textit{and} \\ \left\| s_{X,\mathbf{y}}^{(\nu)} - \mathbf{y} \right\|_{\ell_\infty(X)} &\leq \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + \nu N\varepsilon + (1 - N\nu)\,\varepsilon^* + \frac{N}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2 . \end{aligned}$$

Proof Again, we denote the interpolant to f with respect to X and K by I_f and use $H_{C,\nu}^{\mathbf{y}}$ as defined in Equation (12). Then we have for all $\varepsilon > 0$

$$\frac{1}{2C} \left\| s_{X,\mathbf{y}}^{(\nu)} \right\|_{\mathcal{N}_K(\Omega)}^2 \leq H_{C,\nu}^{\mathbf{y}} \left(s_{X,\mathbf{y}}^{(\nu)}, \varepsilon^* \right) \leq H_{C,\nu}^{\mathbf{y}} \left(I_f, \varepsilon \right) \leq \frac{1}{N} \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + \nu\varepsilon + \frac{1}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2$$

which implies

$$\left\| s_{X,\mathbf{y}}^{(\nu)} \right\|_{\mathcal{N}_K(\Omega)} \leq \sqrt{\frac{2C}{N} \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + 2C\nu\varepsilon + \|f\|_{\mathcal{N}_K(\Omega)}^2} .$$

Moreover we have for all $i = 1, \ldots, N$ and all $\varepsilon > 0$

$$\begin{aligned} \left| s_{X,\mathbf{y}}^{(\nu)} \left(\mathbf{x}^{(i)} \right) - y_i \right| &\leq \sum_{j=1}^{N} \left| s_{X,\mathbf{y}}^{(\nu)} \left(\mathbf{x}^{(j)} \right) - y_j \right|_{\varepsilon^*} + \varepsilon^* \\ &\leq N H_{C,\nu}^{\mathbf{y}} \left(s_{X,\mathbf{y}}^{(\nu)}, \varepsilon^* \right) + (1 - N\nu)\,\varepsilon^* \\ &\leq \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + \nu N\varepsilon + (1 - N\nu)\,\varepsilon^* + \frac{N}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2 . \end{aligned}$$

■

Again we can use the results from Lemma 12 to derive a more explicit upper bound on $\varepsilon^* = \varepsilon^*(C, \nu, f, \varepsilon)$. Note that ε^* depends now also on the free parameter ε.

$$0 \leq \left\| s_{X,\mathbf{y}}^{(\nu)}|_X - \mathbf{y} \right\|_{\ell_\infty(X)} \leq \frac{N}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2 + \varepsilon^* (1 - N\nu) + \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + \nu N\varepsilon .$$

If we assume $\nu > 1/N$, this leads to

$$\varepsilon^*(C, \nu, f, \varepsilon) \leq \frac{1}{N\nu - 1} \left(\frac{N}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2 + \sum_{j=1}^{N} \left| r_j \right|_\varepsilon + \nu N\varepsilon \right) .$$

Using the sampling inequality as in the case of exact data leads to the following result on L_q-norms.

Theorem 13 *We suppose $f \in W_2^\tau(\Omega)$ with $f\left(\mathbf{x}^{(i)}\right) = y_i + r_i$. Let $\left(s_{X,\mathbf{y}}^{(\nu)}, \varepsilon^*\right)$ be a solution of the ν-SVR, that is the optimization problem (5). Then there is a constant $\tilde{C} > 0$, which depends on τ, d and Ω but not on f or X, such that for all $\varepsilon > 0$ the approximation error can be bounded by*

$$\left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{L_q(\Omega)} \leq \tilde{C} \left(h^{\tau - \left(\frac{d}{2} - \frac{d}{q}\right)_+} \left(\|f\|_{W_2^\tau(\Omega)} + \sqrt{\frac{2C}{N} \sum_{j=1}^N \left|r_j\right|_\varepsilon + 2C\nu\varepsilon + \|f\|_{W_2^\tau(\Omega)}^2} \right) + \sum_{j=1}^N \left|r_j\right|_\varepsilon + \nu N \varepsilon + \varepsilon^* (1 - N\nu) + \frac{N}{2C} \|f\|_{W_2^\tau(\Omega)}^2 + \|\mathbf{r}\|_{\ell_\infty(X)} \right)$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$.

Note that the choice of the "optimal" ε leading to the best bound, depends dramatically on the problem. We now want to assume that the data errors do not exceed the data itself. For this we suppose

$$\|\mathbf{r}\|_{\ell_\infty(X)} \leq \delta \leq \|f\|_{W_2^\tau(\Omega)}$$

for a parameter $\delta > 0$.

Corollary 14 *If we choose the parameters as*

$$\begin{aligned} C &= \frac{N \|f\|_{W_2^\tau(\Omega)}^2}{2\delta}, \\ \varepsilon &= \delta, \quad \textit{and} \quad \nu = \frac{1}{N}, \end{aligned}$$

we get

$$\left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{L_2(\Omega)} \leq \tilde{C} \left(h^\tau \|f\|_{W_2^\tau(\Omega)} + \delta \right)$$

and

$$\left\| f - s_{X,\mathbf{y}}^{(\nu)} \right\|_{L_\infty(\Omega)} \leq \tilde{C} \left(h^{\tau - d/2} \|f\|_{W_2^\tau(\Omega)} + \delta \right)$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$, with a generic positive constant $\tilde{C}$ which depends on τ, d and Ω but not on f or X.

8. ε-SVR with Exact Data

Since our arguments for the ν-SVR apply similarly to the ε-SVR, we skip over details and just state the results. Note that in this case the non-negative parameter ε is fixed in contrast to the free variable in the ν-SVR. Analogously to the notation introduced in the previous sections, we denote by $s_{X,\mathbf{y}}^{(\varepsilon)}$ the solution to the problem (10) for $X := \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\} \subset \Omega$ and $\mathbf{y} \in \mathbb{R}^N$. The stability and consistency estimates take the following form.

Lemma 15 *Under the assumption (11) concerning the data, we find that for every X and every fixed $\varepsilon \in \mathbb{R}^+$ a solution $s_{X,\mathbf{y}}^{(\varepsilon)}$ to problem (10) satisfies*

$$\begin{aligned} \left\| s_{X,\mathbf{y}}^{(\varepsilon)} \right\|_{\mathcal{N}_K(\Omega)} &\leq \|f\|_{\mathcal{N}_K(\Omega)} \quad \textit{and} \\ \left\| s_{X,\mathbf{y}}^{(\varepsilon)}|_X - \mathbf{y} \right\|_{\ell_\infty(X)} &\leq \frac{N}{2C} \|f\|_{\mathcal{N}_K(\Omega)}^2 + \varepsilon. \end{aligned}$$

Again this leads to the following result on continuous L_q-norms.

Theorem 16 *We suppose $f \in W_2^\tau(\Omega)$ with $f\left(\mathbf{x}^{(i)}\right) = y_i$. Let $s_{X,\mathbf{y}}^{(\varepsilon)}$ be a solution of the ε-SVR, that is the optimization problem (10). Then there is a constant $\tilde{C} > 0$, which depends on τ, d and Ω but not on ε, f or X, such that the approximation error can be bounded by*

$$\left\| f - s_{X,\mathbf{y}}^{(\varepsilon)} \right\|_{L_q(\Omega)} \leq \tilde{C} \left(2h^{\tau - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|f\|_{W_2^\tau(\Omega)} + \frac{N}{2C} \|f\|^2_{W_2^\tau(\Omega)} + \varepsilon \right) \tag{15}$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$.

Applying the same arguments as in the ν-SVR case we obtain the following corollary.

Corollary 17 *If we choose*

$$C = \frac{N \|f\|_{W_2^\tau(\Omega)}}{2h^\tau} \quad \text{, respectively } \; C = \frac{N \|f\|_{W_2^\tau(\Omega)}}{2h^{\tau - d/2}}$$

the inequality (15) turns into

$$\left\| f - s_{X,\mathbf{y}}^{(\varepsilon)} \right\|_{L_2(\Omega)} \leq \tilde{C} \left(3h^\tau \|f\|_{W_2^\tau(\Omega)} + \varepsilon \right) ,$$

respectively

$$\left\| f - s_{X,\mathbf{y}}^{(\varepsilon)} \right\|_{L_\infty(\Omega)} \leq \tilde{C} \left(3h^{\tau - \frac{d}{2}} \|f\|_{W_2^\tau(\Omega)} + \varepsilon \right)$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$, with a generic positive constant $\tilde{C}$ which depends on τ, d and Ω but not on $f \in W_2^\tau(\Omega)$ or X.

The rôle of the parameter C is similar to the one in case of the ν-SVR. But unlike in the case of the ν-SVR we are free to choose the parameter ε. We see that exact data implies that we should choose $\varepsilon \approx 0$. The case $C \to \infty$ and $\varepsilon \to 0$ leads to exact interpolation, and the well known error bounds for kernel-based interpolation (see Wendland, 2005) are attained.
We point out that the ε-SVR is closely related to the squared ε-loss,

$$\min_{s \in \mathcal{N}_K(\Omega)} \frac{1}{N} \sum_{j=1}^{N} \left| s\left(\mathbf{x}^{(j)}\right) - y_j \right|_\varepsilon^2 + \frac{1}{2C} \|s\|^2_{\mathcal{N}_K(\Omega)} \; . \tag{16}$$

This is important because for $\varepsilon = 0$ we get the square loss. Proceeding along the lines of this section, we find for a solution $s_{X,\mathbf{y}}^{(s\ell\varepsilon)}$ of (16) for exact data the stability bound

$$\left\| s_{X,\mathbf{y}}^{(s\ell\varepsilon)} \right\|_{\mathcal{N}_K(\Omega)} \leq \|f\|_{\mathcal{N}_K(\Omega)}$$

and the consistency estimate

$$\left\| s_{X,\mathbf{y}}^{(s\ell\varepsilon)}|_X - \mathbf{y} \right\|_{\ell_\infty(X)} \leq \sqrt{2} \left(\frac{N}{2C} \|f\|^2_{\mathcal{N}_K(\Omega)} + \varepsilon^2 \right)^{1/2} \leq \frac{\sqrt{N}}{\sqrt{C}} \|f\|_{\mathcal{N}_K(\Omega)} + \sqrt{2}\varepsilon \; .$$

Therefore, we obtain similar approximation results for the ε-squared loss as for the ε-SVR by inserting the estimates into the sampling inequalities. Similarly, the results of Section 9 can be adapted to the ε-squared loss. For the special case $\varepsilon = 0$, we obtain the usual least squares, which was analyzed by Wendland and Rieger (2005) in the case of exact data, and by Riplinger (2007) in the case of inexact data.

9. ε-SVR with Inexact Data

In this section we denote again by $s^{(\varepsilon)}_{X,\mathbf{y}}$ the solution to the problem (10) for a set of points $X := \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(N)}\} \subset \Omega$ and $\mathbf{y} \in \mathbb{R}^N$, but we allow the given data to be corrupted by some additive error according to assumption (14).

Lemma 18 *Under the assumption (14) concerning the data, for every X and every fixed $\varepsilon \in \mathbb{R}^+$ a solution $s^{(\varepsilon)}_{X,\mathbf{y}}$ to problem (10) satisfies*

$$\begin{aligned} \left\| s^{(\varepsilon)}_{X,\mathbf{y}} \right\|_{\mathcal{N}_K(\Omega)} &\leq \sqrt{\|f\|^2_{\mathcal{N}_K(\Omega)} + \frac{2C}{N}\sum_{i=1}^N |r_i|_\varepsilon} \quad \text{and} \\ \left\| s^{(\varepsilon)}_{X,\mathbf{y}}|_X - \mathbf{y} \right\|_{\ell_\infty(X)} &\leq \frac{N}{2C}\|f\|^2_{\mathcal{N}_K(\Omega)} + \sum_{i=1}^N |r_i|_\varepsilon + \varepsilon\,. \end{aligned}$$

These bounds shall now be plugged into the sampling inequality.

Theorem 19 *We suppose $f \in W_2^\tau(\Omega)$ with $f\left(\mathbf{x}^{(i)}\right) = y_i$. Let $s^{(\varepsilon)}_{X,\mathbf{y}}$ be a solution of the ε-SVR, that is the optimization problem (10). Then there is a constant $\tilde{C} > 0$, which depends on τ, d and Ω but not on ε, f or X, such that the approximation error can be bounded by*

$$\begin{aligned} \left\| f - s^{(\varepsilon)}_{X,\mathbf{y}} \right\|_{L_q(\Omega)} \leq\ & \tilde{C}\ \left(2h^{\tau - d\left(\frac{1}{2}-\frac{1}{q}\right)_+} \left(\|f\|_{W_2^\tau(\Omega)} + \sqrt{\|f\|^2_{W_2^\tau(\Omega)} + \frac{2C}{N}\sum_{i=1}^N |r_i|_\varepsilon} \right) \right. \\ &+ \left. \frac{N}{2C}\|f\|^2_{W_2^\tau(\Omega)} + \sum_{i=1}^N |r_i|_\varepsilon + \varepsilon + \|\mathbf{r}\|_{\ell_\infty(X)} \right) \end{aligned}$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$.

If we again assume that the error level δ does not overrule the native space norm of the generating function,

$$\|\mathbf{r}\|_{\ell_\infty(X)} \leq \delta \leq \|f\|_{W_2^\tau(\Omega)}\ ,$$

we get the following convergence orders, for our specific choices of the parameters.

Corollary 20 *Again we assume that the error satisfies (14). If we choose $\varepsilon = \delta$ and $C = \frac{N\|f\|_{W_2^\tau}}{2h^\tau}$ respectively $C = \frac{N\|f\|_{W_2^\tau}}{2h^{\tau-d/2}}$ then we find*

$$\begin{aligned} \left\| f - s^{(\varepsilon)}_{X,\mathbf{y}} \right\|_{L_2(\Omega)} &\leq \tilde{C}\left(h^\tau \|f\|_{W_2^\tau(\Omega)} + \delta \right) \quad \text{and} \\ \left\| f - s^{(\varepsilon)}_{X,\mathbf{y}} \right\|_{L_\infty(\Omega)} &\leq \tilde{C}\left(h^{\tau-d/2} \|f\|_{W_2^\tau(\Omega)} + \delta \right) \end{aligned}$$

for all discrete sets $X \subset \Omega$ with fill distance $h := h_{X,\Omega} \leq C_\Omega \tau^{-2}$, with a generic positive constant $\tilde{C}$ which depends on τ, d, and Ω but not on f or X.

10. Numerical Results

In this section we present some numerical examples to support our analytical results, in particular the rates of convergence in case of exact training data, and the detection of the error levels in case of noisy data.

10.1 Exact Training Data

Figure 1 illustrates the approximation orders in case of exact given data as considered in Sections 6 and 8. For that, we used regular data sets generated by the respective functions to be reconstructed and employed the ε- and the ν-SVR with the parameter choices provided in Corollaries 17 and 11, respectively. We implemented the finite dimensional formulations of the associated optimization problems as described in Equations (9) and (8). As kernel functions we used Wendland's functions for two reasons: On the one hand side they yield rather sparse kernel matrices $\mathbf{K}$ due to their compact support, on the other hand side they are easy to implement since they are piecewise polynomials. Furthermore Wendland's functions may be scaled to improve their numerical behaviour. An unscaled function K has support $\mathrm{supp}(K) \subset B(0,1) \subset \mathbb{R}^d$. The scaling is done in such a way that the decay of the Fourier transform is preserved, that is,

$$K^{(c)}(\mathbf{x}) = c^{-d} K\left(\frac{\mathbf{x}}{c}\right) , \quad \mathbf{x} \in \mathbb{R}^d . \tag{17}$$

By construction we have $supp\left(K^{(c)}\right) \subset B(0,c)$, such that small choices of the scaling parameter c imply rather sparse kernel matrices $\mathbf{K^{(c)}} = \left(K^{(c)}\left(\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_2\right)\right)_{i,j=1\ldots N}$. On the other hand side it is known that the constant factor in our error estimates increases with decreasing c. This is a typical trade-off situation between good approximation properties and good condition numbers of the kernel matrices $\mathbf{K^{(c)}}$ (Wendland, 2005). We chose a scaling $c = 0.1$ in all one-dimensional examples and a scaling $c = 2$ in all two-dimensional examples. Since these standard choices already work well, there was no need for a more careful choice. To our knowledge, there are so far no theoretical results on the optimal scaling.

The double logarithmic plots in Figure 1 visualize the convergence orders in terms of the fill distance. For that, the L_∞-approximation error $\|f - s_{X,\mathbf{y}}\|_{L_\infty}$ is plotted versus the fill distance h. The convergence rates can be found as the slopes of the lines.

In subfigure 1(a) the data was generated by

$$f(x) = (x - 0.5)_+^{2.5+eps} \in W_2^3([0,1]) ,$$

where eps denotes the relative machine precision in the sense of MATLAB. We use the notation $(t)_+ := \max\{0,t\}$ for all $t \in \mathbb{R}$. This function f is sampled on regular grids in the unit interval $I := [0,1]$ with 30 to 96 points. Note that in this case the fill distance is given by $h \approx 1/N$. We use two different kernel functions, namely (see Wendland, 2005)

- $K_1(x) = (1 - |x|)_+^3 (3|x| + 1)$ with native space $W_2^2([0,1])$, and
- $K_2(x) = (1 - |x|)_+^5 \left(8|x|^2 + 5|x| + 1\right)$ with native space $W_2^3([0,1])$.

The scaling parameter according to Equation (17) is chosen as $c = 0.1$. We employed the ε- and the ν- SVR with the parameter choices provided in Corollaries 17 and 11. The respective corollaries

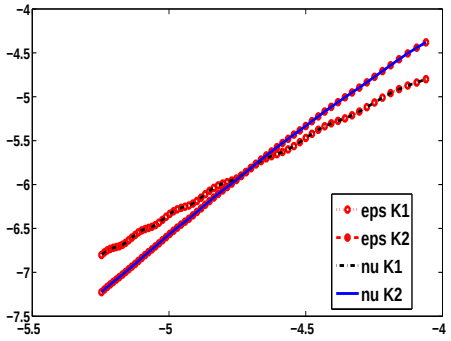
predict convergence rates of 1.5 for K_1, and 2.5 for K_2. In subfigure 1(a) the plots for the ε- and ν-SVR (almost identical) both show orders 1.7 for K_1 and 2.4 for K_2.
Subfigure 1(b) shows a 2-dimensional example. The data is generated by the smooth function

$$f(\mathbf{x}) = \sin(x_1 + x_2) \ .$$

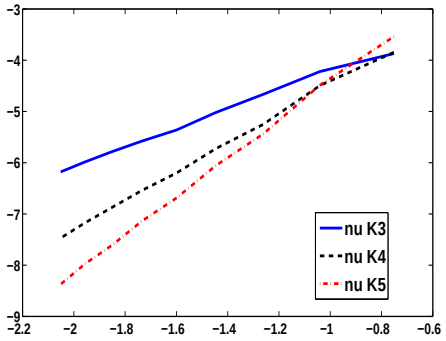
This function f is sampled on regular grids in the unit interval $I := [0,1]^2$ with 16 to 144 points. Note that in this case the fill distance is given by $h \approx \frac{1}{\sqrt{N}}$. We use three different kernel functions, namely (see Wendland, 2005)

- $K_3(\mathbf{x}) = (1 - \|\mathbf{x}\|)_+^4 (4\|\mathbf{x}\| + 1)$ with native space $W_2^{2.5}([0,1]^2)$,
- $K_4(\mathbf{x}) = (1 - \|\mathbf{x}\|)_+^6 \left(35\|\mathbf{x}\|^2 + 18\|\mathbf{x}\| + 3\right)$ with native space $W_2^{3.5}([0,1]^2)$, and
- $K_5(\mathbf{x}) = (1 - \|\mathbf{x}\|)_+^8 \left(32\|\mathbf{x}\|^3 + 25\|\mathbf{x}\|^2 + 8\|\mathbf{x}\| + 1\right)$ with native space $W_2^{4.5}([0,1]^2)$.

The kernel functions were scaled by $c = 2$ according to Equation (17). For the sake of simplicity we employed only the ν-SVR with the parameter choices provided in Corollary 11. The predicted convergence rates in the fill distance h are 1.5 for K_3, 2.5 for K_4 and 3.5 for K_5. The numerical experiments show orders 1.8 for K_3, 2.8 for K_4 and 3.7 for K_5. Therefore, the numerical examples support our analytical results.



(a) Data generated by $f \in W_2^3(I)$ on regular grids in I. ν- and ε-SVR yield orders 1.7 for K_1, and 2.4 for K_2. Scaling parameter $c = 0.1$.



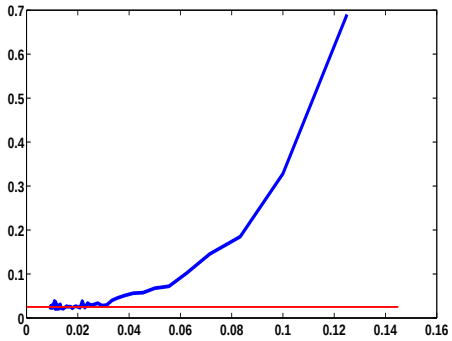
(b) Data generated by smooth function on regular grids in I^2. ν-SVR yields orders 1.8 for K_3, 2.8 for K_4, and 3.7 for K_5. Scaling parameter $c = 2$.

Figure 1: Logarithm of the L_∞-approximation error plotted versus the logarithm of the fill distance h for exact training data.

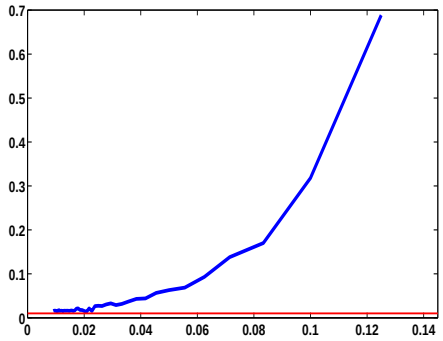
10.2 Inexact Data

Figure 2 shows examples for the case of noisy data. The plots show the L_∞-approximation error $\|f - s_{X,\mathbf{y}}\|_{L_\infty}$ versus the fill distance h. For simplicity we concentrated on the case of the ν-SVR in the one dimensional setting. We used the noise model given by Equation (14), that is $y = f + r$. In Subfigure 2(a) the function $f(x) = \sin(10x)$ is sampled on regular grids of 2 to 56 points in $[0,1]$. The data is disturbed by an error r which is normally distributed with mean zero and standard deviation 0.01. As kernel function we use K_1, and the parameters of the ν-SVR are chosen as in

Corollary 14. The plot shows that for $h \to 0$ the error remains of the same order of magnitude as the error level $\|r\|_{\ell_\infty}$.
In Subfigure 2(b) the function $f(x) = \sin(10x)$ is sampled on regular grids of 5 to 56 points in the unit interval $I = [0,1]$. Here, the data is corrupted by an error of ± 0.01, where the sign of the error is chosen randomly with equal likelihood for plus and minus. As kernel function we use K_1 with $c = 0.3$, and the parameters of the ν-SVR are chosen as in Corollary 14. The plot shows that the L_∞-approximation error converges to a constant of the order of magnitude of the error level for $h \to 0$.



(a) Data disturbed by random error with mean zero and standard deviation 0.01. Approximation error for $h \to 0$ reaches the error level and remains bounded of the same order of magnitude as the error level.



(b) Data disturbed by random sign deterministic error ± 0.01. Approximation error converges to a constant of the order of magnitude of the error level for $h \to 0$.

Figure 2: L_∞-approximation error versus fill distance in case of inexact data.

11. Summary and Outlook

We proved deterministic worst-case error estimates for kernel-based regression algorithms. The main ingredient are sampling inequalities. We provided a detailed analysis only for the ν- and the ε-SVR for both exact and inexact training data. However, the same techniques apply to all machine learning problems involving penalty terms induced by kernels related to Sobolev spaces. If the function to be reconstructed lies in the reproducing kernel Hilbert space (RKHS) of an infinitely smooth kernel such as the Gaussian or an infinite dot product kernel, a similar analysis based on sampling inequalities can be done, leading to exponential convergence rates (see Rieger and Zwicknagl 2008 and Zwicknagl 2009 for first results in this direction).
So far, our error estimates depend explicitly on the pointwise noise in the data, and we do not make any assumptions on the noise distribution. Future work should incorporate probabilistic models on the noise distribution to yield estimates for the expected error.

Acknowledgments

We thank Professor Robert Schaback for helpful discussions and his continued support. Further thanks go to the referees for several valuable comments. CR was supported by the Deutsche Forschungsgemeinschaft through the Graduiertenkolleg 1023 *Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods*. BZ would like to thank the German National Academic Foundation (Studienstiftung des deutschen Volkes) for their support.

References

S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, 1994.

C-C. Chang and C-L. Lin. Training ν-support vector regression: Theory and algorithms. *Neural Computation*, 14(8):1959–1977, 2002.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10 (8):1455–1480, 1998.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

C. Rieger and B. Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. To appear in *Advances in Computational Mathematics*, 2008.

M. Riplinger. Lernen als inverses Problem und deterministische Fehlerabschätzung bei Support Vektor Regression. Diplomarbeit, Universität des Saarlandes, 2007.

R. Schaback. The missing wendland functions. To appear in *Advances in Computational Mathematics*, 2009.

B. Schölkopf and A.J. Smola. *Learning with kernels - Support Vector Machines, Regularisation, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.

B. Schölkopf, C. Burges, and V.Vapnik. Extracting support data for a given task. In *Proceedings, First International Conference on Knowledge Discovery and Data Mining*. CA:AAAI Press., Menlo Park, 1995.

B. Schölkopf, R.C. Wiliamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005.

H. Wendland and C. Rieger. Approximate interpolation. *Numerische Mathematik*, 101:643–662, 2005.

J. Wloka. *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*. Mathematische Leitfäden. Teubner, Stuttgart, 1982.

B. Zwicknagl. Power series kernels. *Constructive Approximation*, 29(1):61–84, 2009.