

# CONSISTENCY AND ROBUSTNESS OF KERNEL BASED REGRESSION

BY ANDREAS CHRISTMANN<sup>1</sup> AND INGO STEINWART

*University of Dortmund and Los Alamos National Laboratory*

We investigate properties of kernel based regression (KBR) methods which are inspired by the convex risk minimization method of support vector machines. We first describe the relation between the used loss function of the KBR method and the tail of the response variable  $Y$ . We then establish a consistency result for KBR and give assumptions for the existence of the influence function. In particular, our results allow to choose the loss function and the kernel to obtain computational tractable and consistent KBR methods having bounded influence functions. Furthermore, bounds for the sensitivity curve which is a finite sample version of the influence function are developed, and some numerical experiments are discussed.

## 1. Introduction

In regression problems the goal is to estimate an approximated functional relationship  $Y \approx f(X)$ , where  $(X, Y)$  is a pair consisting of an  $\mathbb{R}^d$ -valued *observation* random variable  $X$  and an  $\mathbb{R}$ -valued *outcome* or *response* random variable  $Y$ . In the simplest case one assumes that this relationship is linear plus some noise and the goal is then to estimate the linear term having only a set of observations  $(x_i, y_i)$  from independent and identically distributed (i.i.d.) random variables  $(X_i, Y_i)$ ,  $1 \leq i \leq n$ . A classical method for this problem is the least squares estimator which is known to solve the problem for  $n \rightarrow \infty$  under parametric assumptions. Unfortunately, this method fails if the assumption on the linear relationship is violated and, even worse, it is well-known that the least squares estimator is non-robust against mild model violations. However there is a vast literature of good robust estimation methods for linear regression models and for nonlinear parametric regression models, see e.g. Huber (1981), Hampel *et al.* (1986), Rousseeuw (1984), Rousseeuw and Yohai (1984), Yohai (1987), Davies (1993), and Mendes and Tyler (1996).

It is the nature of estimation methods for parametric regression models that they require strong assumptions on the distribution of  $(X, Y)$ . If such knowledge on  $(X, Y)$  is not available, one consequently has to use non-parametric methods instead. Many of these methods rely on the least squares loss function because that *a)* simplifies the mathematical treatment and *b)* leads to efficient algorithms, see Györfi *et al.* (2002). Unfortunately, using the least squares loss function also means that one cannot expect these methods to be robust. Alternatively, one can deploy recently developed kernel based regression (KBR) methods like support vector machines (SVMs) which lead to efficient algorithms for a variety of loss functions, see Vapnik (1998) or Schölkopf and Smola (2002) for an introduction. However, almost nothing is known for these methods with respect to both consistency and robustness, so that they lack a theoretical foundation. The aim of this paper is to provide important

- 
1. Supported in part by the Deutsche Forschungsgemeinschaft (SFB-475) and by DoMuS.
  2. *AMS 2000 subject classification.* Primary 62G08, 62G35; secondary 68Q32, 62G20.
  3. *Keywords and Phrases.* Non-parametric regression, convex risk minimization, consistency, influence function, robustness, sensitivity curve, support vector regression.

aspects of such a foundation: we first show consistency of KBR methods requiring only mild tail conditions on the distribution of the response variable  $Y|X = x$ . Then we provide results that describe the influence of both the kernel and the loss function on the robustness. In particular, we establish the existence of the influence function for a broad class of KBR methods and present conditions under which the influence function is bounded. Here it turns out that, depending on the kernel and the loss function, some KBR methods are robust while others are not, and consequently, our results show how to choose both quantities to obtain consistent KBR estimators with good robustness properties. Interestingly, but in some sense not surprising, it turns out that the robust KBR methods are exactly the ones that require the mildest tail conditions on  $Y$  for consistency.

The rest of the paper is organized as follows: Section 2 introduces kernel based regression methods. Then, in Section 3, we present some important notions describing the growth behavior of loss functions. Subsequently, we consider properties of the associated risk functionals. In particular we discuss the relation between the growth of loss functions and the tails of  $Y$ . We then establish a stability result for infinite-sample KBR estimators which will be used for both our consistency analysis in Section 4 and our robustness discussion in Section 5. Besides the above mentioned results on the influence function we also give bounds for the sensitivity curve of KBR methods in the latter section. Furthermore a connection to M-estimation for parametric regression becomes visible. Then, some numerical examples are presented in Section 6, and Section 7 contains a discussion. Finally, all proofs are given in the appendix.

## 2. Kernel based regression

The goal of nonparametric regression is to estimate an approximated functional relationship between an observation random variable  $X$  and a response random variable using  $n$  observations  $(x_i, y_i) \in X \times Y$  drawn independently from the same *unknown* distribution  $P$  of the pair  $(X, Y)$ . For technical reasons we assume throughout this work that  $X$  and  $Y$  are closed subsets of  $\mathbb{R}^d$  and  $\mathbb{R}$ , respectively<sup>1</sup>. Recall that in this case  $P$  can be split up into the marginal distribution  $P_X$  and the regular conditional probability  $P(\cdot | x)$ ,  $x \in X$ , on  $Y$ .

Now let  $L : Y \times \mathbb{R} \rightarrow \mathbb{R}$  be a function which is convex with respect to its second argument. Then the KBR methods considered in this work minimize the empirical regularized risk

$$\hat{f}_{n,\lambda} := \arg \min_{f \in H} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \lambda \|f\|_H^2, \quad (1)$$

where  $\lambda > 0$  is a regularization parameter and  $H$  is a reproducing kernel Hilbert space (RKHS) of a kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . Throughout this work we write  $\Phi : X \rightarrow H$  for the canonical feature map of  $k$  which is defined by  $\Phi(x) := k(\cdot, x)$ ,  $x \in X$ . Recall that the reproducing property gives  $f(x) = \langle f, k(\cdot, x) \rangle$  for all  $f \in H$  and  $x \in X$ .

Obviously, problem (1) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk

$$f_{P,\lambda} := \arg \min_{f \in H} \mathbb{E}_P L(Y, f(X)) + \lambda \|f\|_H^2. \quad (2)$$

---

1. For  $X$  it suffices to assume that it is a locally compact Polish space.

Loss Function	$L(y, t)$	$L'(y, t)$	$L''(y, t)$
Least Squares	$r^2$	$-2r$	2
$\varepsilon$ -insensitive, $\varepsilon > 0$	$\max( r  - \varepsilon, 0)$	0, if $ r  < \varepsilon$ $\text{sgn}(-r)$ , if $ r  > \varepsilon$	0, if $r \notin \{\pm\varepsilon\}$
Huber, $c > 0$	$r^2/2$ , if $ r  \leq c$ $c r  - c^2/2$ , if $ r  > c$	$-r$ , if $ r  < c$ $c \text{sgn}(-r)$ , if $ r  > c$	1, if $ r  < c$ 0, if $ r  > c$
Logistic	$-\log(4\Lambda(r)[1 - \Lambda(r)])$	$1 - 2\Lambda(r)$	$2\Lambda(r)[1 - \Lambda(r)]$

Table 1: Some loss functions and their derivatives with respect to  $t$ . We use the shorthands  $r := y - t$  and  $\Lambda(r) := 1/[1 + e^{-r}]$ .

The objective function in (2) is denoted by  $R_{L,P,\lambda}^{reg}(\cdot)$  in the sequel. Note that in practice one usually solves the dual problem of (1) numerically, since in the dual problem instead of the RKHS itself only its kernel occur. In particular the choice of the kernel  $k$  enables efficient estimation of linear and also of non-linear functions. Of special importance is the Gaussian radial basis function (RBF) kernel

$$k(x, x') = \exp(-\gamma\|x - x'\|^2), \quad \gamma > 0, \quad (3)$$

which is a universal kernel on every compact subset of  $\mathbb{R}^d$ , see Definition 14. This kernel is a bounded kernel as  $|k(x, x')| \leq 1$  for all  $x, x' \in \mathbb{R}^d$ . Polynomial kernels  $k(x, x') = (c + \langle x, x' \rangle)^m$ ,  $m \geq 1$ ,  $c \geq 0$ ,  $x, x' \in \mathbb{R}$ , are also popular in practice, but obviously they are neither universal nor bounded.

Popular convex loss functions for regression problems depend on  $y$ ,  $x$ , and  $f$  via the residual  $r := y - f(x)$ , and are based on the implicit “signal + noise” assumption  $y_i = f(x_i) + \varepsilon_i$ . We will call such loss functions invariant, cf. Definition 1. Three important examples are the least squares loss function, Vapnik’s  $\varepsilon$ -insensitive loss function, and Huber’s loss function, see Table 1. Another invariant loss function is the logistic loss function (see again Table 1) which is a compromise between the former three loss functions: it is twice continuously differentiable with  $L'' > 0$  which is true for the least squares loss function and it increases approximately linearly if  $|r|$  tends to infinity which is true for Vapnik’s and Huber’s loss functions. These four loss functions are even symmetric, because of  $L(y, t) = L(t, y)$  for  $y, t \in \mathbb{R}$ . Asymmetric loss functions may be interesting in some applications where extremely skewed distributions occur, e.g. in analyzing the claim sizes in insurance data (Christmann, 2004). As an example for a smooth, invariant, and asymmetric loss function we mention

$$L(y, t) = \left(\frac{c_2}{2} - c_1\right)r - \frac{c_2}{2} \log\left(4\Lambda(r - c_3)(1 - \Lambda(r - c_3))\right) + c_4,$$

where  $r = y - t$ ,  $0 < c_1 < c_2 < \infty$ ,  $c_3 = -\Lambda^{-1}(c_1/c_2)$  and  $c_4 = (c_2/2) \log(4\frac{c_1}{c_2}(1 - \frac{c_1}{c_2}))$ . For  $(c_1, c_2) = (1, 2)$  we obtain the logistic loss function. Note that  $L'$  and  $L''$  are continuous and bounded for the logistic loss function and its asymmetric modification.

### 3. Loss functions, risks, and stability of infinite-sample KBR methods

In this section we first introduce some important concepts for loss functions which are used throughout this work. Then in Subsection 3.2 we introduce their associated risks and discuss the interplay between growth behavior of the loss functions and the tail of the response variable. Finally, in Subsection 3.3 we establish existence and stability results for the infinite-sample KBR methods given by (2). These results are needed to obtain both the consistency results in Section 4 and some of the robustness results in Section 5.

#### 3.1 Loss functions and their growth behavior

The main goal of this subsection is to describe the growth behavior of loss functions. To this end let us begin with some basic definitions for loss functions.

**Definition 1** *Let  $Y \subset \mathbb{R}$  be a non-empty closed subset. Then a continuous function  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  is called a loss function. Furthermore, we say that  $L$  is*

- i) convex if  $L(y, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$  is convex for all  $y \in Y$ .*
- ii) Lipschitz continuous if there exists a constant  $c > 0$  such that*

$$|L(y, t) - L(y, t')| \leq c \cdot |t - t'| \tag{4}$$

*for all  $y \in Y, t, t' \in \mathbb{R}$ . In this case we denote the smallest possible  $c$  in (4) by  $|L|_1$ .*

- iii) invariant if there exists a function  $l : \mathbb{R} \rightarrow [0, \infty)$  with  $l(0) = 0$  and  $L(y, t) = l(y - t)$  for all  $y \in Y, t \in \mathbb{R}$ .*

Obviously, all loss functions listed in Table 1 are invariant. Moreover, an invariant loss function  $L$  is convex if and only if the corresponding  $l$  is convex. Analogously,  $L$  is Lipschitz continuous if and only if  $l$  is Lipschitz continuous and in this case we have  $|L|_1 = |l|_1$ , where  $|l|_1$  denotes the Lipschitz constant of  $l$ . In particular, all loss functions listed in Table 1 are convex and besides the least squares loss function they are also Lipschitz continuous.

As already mentioned the growth behavior of loss functions plays an important role in both consistency and robustness results. Therefore we now introduce some basic concepts which describe the growth behavior of  $L$ . We begin with

**Definition 2** *Let  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a loss function,  $a : Y \rightarrow [0, \infty)$  be a measurable function, and  $p \in [0, \infty)$ . We say that  $L$  is a loss function of type  $(a, p)$  if there exists a constant  $c > 0$  such that*

$$L(y, t) \leq c \left( a(y) + |t|^p + 1 \right)$$

*for all  $y \in Y$  and all  $t \in \mathbb{R}$ . Furthermore, we say that  $L$  is of strong type  $(a, p)$  if the first two partial derivatives  $L' := \partial_2 L$  and  $L'' := \partial_{22} L$  of  $L$  with respect to the second argument of  $L$  exist and  $L, L'$  and  $L''$  are of  $(a, p)$ -type.*

For invariant loss functions it turns out that there is an easy way to determine their type. In order to describe the corresponding results we need the following definition.

**Definition 3** Let  $L$  be an invariant loss function with corresponding function  $l : \mathbb{R} \rightarrow \mathbb{R}$ . We say that  $L$  is of upper order  $p$ ,  $p \geq 0$ , if there exists a constant  $c > 0$  such that for all  $r \in \mathbb{R}$  we have

$$\|l_{|[-r,r]}\|_\infty \leq c(|r|^p + 1).$$

Here,  $\|\cdot\|_\infty$  denotes the supremum norm. Analogously, we say that  $L$  is of lower order  $p$ ,  $p \geq 0$ , if there exists a constant  $c > 0$  such that for all  $r \in \mathbb{R}$  we have

$$\|l_{|[-r,r]}\|_\infty \geq c(|r|^p - 1).$$

Recalling that convex functions are locally Hölder continuous we see that for invariant loss functions  $L$  the corresponding  $l$  is Hölder continuous on every interval  $[-r, r]$ . Consequently,

$$H(r) := |l_{|[-r,r]}|_1, \quad r \geq 0 \tag{5}$$

defines a non-decreasing function  $H : [0, \infty) \rightarrow [0, \infty)$ . We denote its symmetric extension also by  $H$ , so that we have  $H(-r) = H(r)$  for all  $r \in \mathbb{R}$ . Now we can establish the following simple properties (see the appendix for a proof) of convex, invariant loss functions.

**Lemma 4** Let  $L$  be an invariant loss function with corresponding  $l : \mathbb{R} \rightarrow \mathbb{R}$  and  $p \geq 0$ . Then the following is true:

- i) if  $L$  is convex and satisfies  $\lim_{|r| \rightarrow \infty} l(r) = \infty$  then it is of lower order 1.
- ii) if  $L$  is Lipschitz continuous then it is of upper order 1.
- iii) if  $L$  is the least squares loss then it is of lower and upper order 2.
- iv) if  $L$  is convex then for all  $r > 0$  we have

$$H(r) \leq \frac{2}{r} \|l_{|[-2r,2r]}\|_\infty \leq 4H(2r).$$

- v) if  $L$  is of upper order  $p$  then  $L$  is of type  $(a, p)$  with  $a : Y \rightarrow [0, \infty)$  defined by  $a(y) := |y|^p$ ,  $y \in Y$ .

With the help of the above lemma it is easy to see that the least squares loss function is of strong type  $(y^2, 2)$ . Furthermore, the logistic loss function is of strong type  $(|y|, 1)$  since it is twice continuously differentiable with respect to its second variable and both derivatives are bounded, namely:  $|\partial_2 L(y, t)| \leq 1$  and  $|\partial_{22} L(y, t)| \leq \frac{1}{2}$ ,  $t \in \mathbb{R}$ . The other two loss functions of Table 1 are of upper and lower order 1 since they are convex and Lipschitz continuous, however they are not of any strong type since they are not twice continuously differentiable.

### 3.2 Risks

With the help of a loss function  $L$  one can assign a risk to a measurable map  $f : X \rightarrow \mathbb{R}$ . Namely, if  $\mathbb{P}$  is a distribution on  $X \times Y$  then the  $L$ -risk of  $f$  with respect to  $\mathbb{P}$  is defined by

$$\mathcal{R}_{L,\mathbb{P}}(f) := \int_{X \times Y} L(y, f(x)) d\mathbb{P}(x, y) = \int_X \int_Y L(y, f(x)) \mathbb{P}(dy|x) \mathbb{P}_X(dx),$$

where we recall that the regular conditional probability  $\mathbb{P}(\cdot|x)$  exists because  $Y$  is closed (and thus Polish). Note that the above integral—although it may be not finite—is always

defined since  $L$  is non-negative and continuous. In order to find a condition which ensures  $\mathcal{R}_{L,P}(f) < \infty$  we need the following definition which for later purpose is formulated in a rather general way (see Brown and Percy (1977) for signed measures).

**Definition 5** *Let  $\mu$  be a signed measure on  $X \times Y$  with total variation  $|\mu|$  and  $a : Y \rightarrow [0, \infty)$  be a measurable function. Then we write*

$$|\mu|_a := \int_{X \times Y} a(y) d|\mu|(x, y).$$

*Furthermore, if  $a(y) = |y|^p$  for some  $p > 0$  and all  $y \in Y$ , we write  $|\mu|_p := |\mu|_a$  whenever no confusion can arise.*

Now we can formulate the announced sufficient condition ensuring  $\mathcal{R}_{L,P}(f) < \infty$ .

**Proposition 6** *Let  $L$  be an  $(a, p)$ -type loss function,  $P$  be a distribution on  $X \times Y$  with  $|P|_a < \infty$  and  $f : X \rightarrow \mathbb{R}$  be a function with  $f \in L_p(P)$ . Then we have  $\mathcal{R}_{L,P}(f) < \infty$ .*

The above proposition shows in particular that for  $p$ -integrable functions  $f : X \rightarrow \mathbb{R}$  we have  $\mathcal{R}_{L,P}(f) < \infty$  if  $L$  is an invariant loss function of upper order  $p$  and  $P$  satisfies  $|P|_p < \infty$ . The next result is somehow an inversion of this fact.

**Lemma 7** *Let  $L$  be an invariant loss function of lower order  $p$ ,  $f : X \rightarrow \mathbb{R}$  be a measurable function and  $P$  be a distribution  $X \times Y$  with  $\mathcal{R}_{L,P}(f) < \infty$ . Then we have  $|P|_p < \infty$  if and only if  $f \in L_p(P)$ .*

**Remark 8** *If  $L$  is an invariant loss function of lower and upper order  $p$  and  $P$  is a distribution with  $|P|_p = \infty$  the above lemma shows  $\mathcal{R}_{L,P}(f) = \infty$  for all  $f \in L_p(P)$ . This suggests that we may even have  $\mathcal{R}_{L,P}(f) = \infty$  for all measurable  $f : X \rightarrow Y$ . However, this is in general not the case. For example, let  $P_X$  be a distribution on  $X$  and  $g : X \rightarrow \mathbb{R}$  be a measurable function with  $g \notin L_p(P_X)$ . Furthermore, let  $P$  be the distribution on  $X \times \mathbb{R}$  whose marginal distribution on  $X$  is  $P_X$  and whose conditional probability satisfies  $P(Y = g(x)|x) = 1$ . Then we have  $|P|_p = \int_X |g(x)|^p dP_X(x) = \infty$ , but  $\mathcal{R}_{L,P}(g) = \int_X l(g(x) - g(x)) dP_X(x) = 0$ .*

### 3.3 Stability of infinite-sample KBR methods

As already discussed we are mainly interested in the optimization problem (2). To make any meaningful investigation of the solution  $f_{P,\lambda}$  we first have to ensure its existence. This is done in the following proposition.

**Proposition 9** *Let  $L$  be a convex loss function which is of  $(a, p)$ -type,  $P$  be a distribution on  $X \times Y$  with  $|P|_a < \infty$ ,  $H$  be an RKHS of a bounded kernel  $k$ , and  $\lambda > 0$ . Then there exists a unique minimizer  $f_{P,\lambda} \in H$  of*

$$f \mapsto \mathcal{R}_{L,P,\lambda}^{reg}(f) := \mathcal{R}_{L,P}(f) + \lambda \|f\|_H^2$$

*and we have  $\|f_{P,\lambda}\|_H \leq \sqrt{\mathcal{R}_{L,P}(0)/\lambda} := \delta_{P,\lambda}$ .*

**Remark 10** *If  $H$  is an RKHS of a bounded kernel and  $L$  is a convex and invariant loss function of lower and upper order  $p$  then it is easy to see by Lemma 7 that exactly for the distributions  $\mathbb{P}$  with  $|\mathbb{P}|_p < \infty$  the minimizer  $f_{\mathbb{P},\lambda}$  is uniquely determined. Furthermore, if  $|\mathbb{P}|_p = \infty$  we have  $\mathcal{R}_{L,\mathbb{P},\lambda}^{reg}(f) = \infty$  for all  $f \in H$ . In the following we will therefore use the definition  $f_{\mathbb{P},\lambda} := 0$  for such distributions.*

Our next aim is to establish a representation of  $f_{\mathbb{P},\lambda}$ . To this end we define for  $p \in [1, \infty]$  the conjugate  $p' \in [1, \infty]$  by  $1/p + 1/p' = 1$ . Furthermore we have to recall the notion of subdifferentials (cf. e.g. Phelps (1986)).

**Definition 11 (Subdifferential)** *Let  $H$  be a Hilbert space,  $F : H \rightarrow \mathbb{R} \cup \{\infty\}$  be a convex function and  $w \in H$  with  $F(w) \neq \infty$ . Then the subdifferential of  $F$  at  $w$  is defined by*

$$\partial F(w) := \{w^* \in H : \langle w^*, v - w \rangle \leq F(v) - F(w) \text{ for all } v \in H\}.$$

Furthermore, if  $L$  is a convex loss function, we denote the subdifferential of  $L$  with respect to the second variable by  $\partial_2 L$ .

The robustness approach based on influence functions (see Definition 16) is based on a special Gâteaux-derivative. Therefore, we mention that if  $F$  is Gâteaux-differentiable at  $w$  then  $\partial F(w)$  contains only the derivative of  $F$  at  $w$ , see (Phelps, 1986, Prob. 1.8).

With the help of the subdifferential  $\partial_2 L$  we can now recall the following result of DeVito *et al.* (2004) which is a generalization of a representation shown in Steinwart (2003).

**Proposition 12** *Let  $p \geq 1$ ,  $L$  be a convex loss function of type  $(a, p)$ , and  $P$  be a distribution on  $X \times Y$  with  $|\mathbb{P}|_a < \infty$ . Let  $H$  be the RKHS of a bounded, continuous kernel  $k$  over  $X$ , and  $\Phi : X \rightarrow H$  be the canonical feature map of  $H$ . Then there exists an  $h \in L_{p'}(P)$  such that  $h(x, y) \in \partial_2 L(y, f_{\mathbb{P},\lambda}(x))$  for all  $(x, y) \in X \times Y$  and*

$$f_{\mathbb{P},\lambda} = -\frac{1}{2\lambda} \mathbb{E}_P h \Phi. \tag{6}$$

With the help of Proposition 12 we can now state the following stability result, see Zhang (2001) and Steinwart (2003) for similar results for classification problems.

**Theorem 13** *Let  $p$ ,  $L$ ,  $\mathbb{P}$ ,  $H$ ,  $\Phi$ , and  $h$  be as in Proposition 12. Then for all distributions  $\mathbb{Q}$  on  $X \times Y$  with  $|\mathbb{Q}|_a < \infty$  we have  $h \in L_{p'}(\mathbb{P}) \cap L_1(\mathbb{Q})$  and*

$$\|f_{\mathbb{P},\lambda} - f_{\mathbb{Q},\lambda}\|_H \leq \frac{1}{\lambda} \|\mathbb{E}_P h \Phi - \mathbb{E}_Q h \Phi\|_H. \tag{7}$$

Furthermore, if  $L$  is actually an invariant loss function of upper order  $p$  and  $P$  satisfies  $|\mathbb{P}|_p < \infty$  then we  $h \in L_{p'}(\mathbb{P}) \cap L_{p'}(\mathbb{Q})$ .



#### 4. Consistency of kernel based regression

In this section we establish  $L$ -risk consistency of KBR methods, i.e. we show that

$$\mathcal{R}_{L,P}(\hat{f}_{n,\lambda_n}) \rightarrow \mathcal{R}_{L,P} := \inf\{\mathcal{R}_{L,P}(f) \mid f : X \rightarrow \mathbb{R} \text{ measurable}\}$$

holds in probability for  $n \rightarrow \infty$  for suitable chosen regularization sequences  $(\lambda_n)$ . Of course, such convergence can only hold if the used RKHS is rich enough. One way of describing the richness of  $H$  is the following definition taken from Steinwart (2001).

**Definition 14** *Let  $X \subset \mathbb{R}^d$  be compact and  $k : X \times X \rightarrow \mathbb{R}$  be a continuous kernel with RKHS  $H$ . We say that  $k$  is universal if  $H$  is dense in the space of continuous functions  $C(X)$  equipped with  $\|\cdot\|_\infty$ .*

It is well-known that many popular kernels including the Gaussian RBF kernels are universal, cf. e.g. Steinwart (2001) for a simple proof of universality of the latter kernel. With the above definition we can now formulate our consistency result.

**Theorem 15** *Let  $X \subset \mathbb{R}^d$  be a compact subset,  $L$  be an invariant, convex loss function of lower and upper order  $p \geq 1$ , and  $H$  be a RKHS of a universal kernel on  $X$ . We write  $p^* := \max\{2p, p^2\}$  and fix a sequence  $(\lambda_n)$  of positive numbers with  $\lambda_n \rightarrow 0$  and  $\lambda_n^{p^*} n \rightarrow \infty$ . Then KBR based on (1) using  $\lambda_n$  for sample sets of length  $n$  is  $L$ -risk consistent for all distributions  $P$  with  $|P|_p < \infty$ .*

Note that Theorem 15 in particular shows that KBR using the least squares loss function is *weakly universally consistent* in the sense of Györfi *et al.* (2002). Furthermore, it is worthwhile to note that under the above assumptions on  $L$ ,  $H$ , and  $(\lambda_n)$  we can even *characterize* the distributions  $P$  for which KBR estimates based on (1) are  $L$ -risk consistent. Indeed, if  $|P|_p = \infty$  then KBR is trivially  $L$ -risk consistent for  $P$  whenever  $\mathcal{R}_{L,P} = \infty$ . Conversely, if  $|P|_p = \infty$  and  $\mathcal{R}_{L,P} < \infty$  then KBR cannot be  $L$ -risk consistent for  $P$  since Lemma 7 shows  $\mathcal{R}_{L,P}(f) = \infty$  for all  $f \in H$ .

In some sense it seems natural to consider only consistency for distributions satisfying the tail assumption  $|P|_p < \infty$  as this was done e.g. in Györfi *et al.* (2002) for least squares methods. In this sense Theorem 15 gives consistency for all reasonable distributions. However, it is important to note that the above characterization shows that our KBR methods are *not* robust against small violations of this tail assumption. Indeed, let  $P$  be a distribution with  $|P|_p < \infty$ , and  $\tilde{P}$  be a distribution with  $|\tilde{P}|_p = \infty$  and  $\mathcal{R}_{L,\tilde{P}}(f^*) < \infty$  for some  $f^* \in L_p(P)$ . Then every mixture distribution  $Q_\varepsilon := (1 - \varepsilon)P + \varepsilon\tilde{P}$ ,  $\varepsilon \in (0, 1)$ , satisfies both  $|Q_\varepsilon|_p = \infty$  and  $\mathcal{R}_{L,Q_\varepsilon} < \infty$  and thus KBR is not consistent for any of the small perturbation  $Q_\varepsilon$  of  $P$  while it is consistent for original distribution  $P$ . From a robustness point of view, this is of course a negative result.

#### 5. Robustness of kernel based regression

In the statistical literature different criteria have been proposed to define the notion of robustness in a mathematical way, e.g. Huber (1964), Hampel (1974), Hampel *et al.* (1986), Tukey (1977), Donoho and Huber (1983), and Rousseeuw and Hubert (1999).



In this paper, we mainly use Hampel's approach based on the influence function. We will consider a map  $T$  which assigns to every distribution  $P$  on a given set  $Z$  an element  $T(P)$  of a given Banach space  $E$ . For the case of the convex risk minimization problem given in (2) we have  $E = H$  and  $T(P) = f_{P,\lambda}$ .

**Definition 16 (Influence function)** *The influence function of  $T$  at a point  $z$  for a distribution  $P$  is the special Gâteaux derivative (if it exists)*

$$IF(z; T, P) = \lim_{\varepsilon \downarrow 0} \frac{T((1 - \varepsilon)P + \varepsilon\Delta_z) - T(P)}{\varepsilon}, \quad (8)$$

where  $\Delta_z$  is the Dirac distribution at the point  $z$ , i.e.  $\Delta_z(\{z\}) = 1$ .

The influence function has the interpretation, that it measures the impact of an (infinitesimal) small amount of contamination of the original distribution  $P$  in direction of a Dirac distribution located in the point  $z$  on the theoretical quantity of interest  $T(P)$ . Therefore, in the robustness approach based on influence functions it is desirable that a statistical method  $T(P)$  has a *bounded* influence function. We also use Tukey's sensitivity curve which can be interpreted as a finite sample version of the influence function. The sensitivity curve measures the impact of a single data point  $z$ .

**Definition 17 (Sensitivity curve)** *The sensitivity curve of an estimator  $T_n$  at a point  $z$  given a data set  $z_1, \dots, z_{n-1}$  is defined by*

$$SC_n(z; T_n) = n(T_n(z_1, \dots, z_{n-1}, z) - T_{n-1}(z_1, \dots, z_{n-1})).$$

If the estimator  $T_n$  is defined via  $T(P_n)$ , where  $P_n$  denotes the empirical distribution of the data points  $z_1, \dots, z_n$ , then we have for  $\varepsilon_n = 1/n$ :

$$SC_n(z; T_n) = \frac{T((1 - \varepsilon_n)P_{n-1} + \varepsilon_n\Delta_z) - T(P_{n-1})}{\varepsilon_n}. \quad (9)$$

In the following we give sufficient conditions for the existence of the influence function for the kernel based regression methods based on (2). Further, we establish conditions on the kernel  $k$  and on the loss function  $L$  ensuring that the influence function and the sensitivity curve are bounded. To this end we need to recall some notions from Banach space calculus. We say that a map  $G : E \rightarrow F$  between Banach spaces  $E$  and  $F$  is (Fréchet)-differentiable in  $x_0 \in E$  if there exists a bounded linear operator  $A : E \rightarrow F$  and a function  $\varphi : E \rightarrow F$  with  $\frac{\varphi(x)}{\|x\|} \rightarrow 0$  for  $x \rightarrow 0$  such that

$$G(x_0 + x) - G(x_0) = Ax + \varphi(x) \quad (10)$$

for all  $x \in E$ . Furthermore, since  $A$  is uniquely determined by (10) we write  $G'(x) := \frac{\partial G}{\partial E}(x) := A$ . The map  $G$  is called continuously differentiable if the map  $x \mapsto G'(x)$  exists on  $E$  and is continuous. Analogously we define continuous differentiability on open subsets of  $E$ . For further information we refer to Akerkar (1999), Brown and Pearcy (1977), and Yosida (1974).

The next result shows that the influence function of  $T(P) = f_{P,\lambda}$  based on (2) exists, if the loss function is convex and twice continuously differentiable and if the kernel is bounded and continuous. The proof of this theorem as well as the proofs of the following results can be found in the appendix.

**Theorem 18** *Let  $H$  be a RKHS of a bounded continuous kernel  $k$  on  $X$  with canonical feature map  $\Phi : X \rightarrow H$ , and  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a convex loss function of some strong type  $(a, p)$ . Furthermore, let  $\mathbb{P}$  be a distribution on  $X \times Y$  with  $|\mathbb{P}|_a < \infty$ . Then the influence function of  $f_{\mathbb{P}, \lambda}$  exists for all  $z := (x, y) \in X \times Y$  and we have*

$$IF(z; T, \mathbb{P}) = S^{-1}(\mathbb{E}_{\mathbb{P}}(L'(Y, f_{\mathbb{P}, \lambda}(X))\Phi(X))) - L'(y, f_{\mathbb{P}, \lambda}(x))S^{-1}\Phi(x) , \quad (11)$$

where  $S : H \rightarrow H$  is defined by  $S = 2\lambda \text{id}_H + \mathbb{E}_{\mathbb{P}}L''(Y, f_{\mathbb{P}, \lambda}(X))\langle \Phi(X), \cdot \rangle \Phi(X)$ .

It is worth mentioning that the proof can easily be modified in order to replace point mass contaminations  $\Delta_z$  by arbitrary contaminations  $\tilde{\mathbb{P}}$  satisfying  $|\tilde{\mathbb{P}}|_a < \infty$ . As the discussion after Theorem 15 shows we cannot omit this tail assumption on  $\tilde{\mathbb{P}}$  in general.

From a robustness point of view, one is mainly interested in bounded influence functions. Interestingly, for some kernel based regression methods based on (2) Theorem 18 not only ensures the existence of the influence function but also indicates how to guarantee its boundedness. Indeed, (11) shows that the only term of the influence function that depends on the point mass contamination  $\Delta_z$  is

$$-L'(y, f_{\mathbb{P}, \lambda}(x))S^{-1}\Phi(x) . \quad (12)$$

Now let us assume that the used kernel is a Gaussian RBF kernel. Then we have  $\Phi(x) \neq 0$  for all  $x \in X$  and consequently, the influence function is bounded if and only if  $L'(\cdot, f_{\mathbb{P}, \lambda}(x)) : \mathbb{R} \rightarrow \mathbb{R}$  is bounded for all  $x \in X$ . For invariant loss functions we hence immediately obtain the following corollary.

**Corollary 19** *Let  $X := \mathbb{R}^d$ ,  $Y := \mathbb{R}$ , and  $k$  be a Gaussian RBF kernel on  $X$  with canonical feature map  $\Phi : X \rightarrow H$ . Furthermore, let  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a convex and invariant loss function of some strong type  $(a, p)$ , and  $\mathbb{P}$  be a distribution on  $X \times Y$  with  $|\mathbb{P}|_a < \infty$ . Then the influence function of  $f_{\mathbb{P}, \lambda}$  exists for all  $z := (x, y) \in X \times Y$  and we have*

$$IF(z; T, \mathbb{P}) = -S^{-1}(\mathbb{E}_{\mathbb{P}}(l'(Y - f_{\mathbb{P}, \lambda}(X))\Phi(X))) + l'(y - f_{\mathbb{P}, \lambda}(x))S^{-1}\Phi(x) , \quad (13)$$

where  $l : \mathbb{R} \rightarrow \mathbb{R}$  is the function representing  $L$  and  $S : H \rightarrow H$  is defined by  $S = 2\lambda \text{id}_H + \mathbb{E}_{\mathbb{P}}l''(Y, f_{\mathbb{P}, \lambda}(X))\langle \Phi(X), \cdot \rangle \Phi(X)$ . Consequently, the influence function is bounded in  $z$  if and only if  $L$  is Lipschitz continuous.

The above corollary shows that the least squares loss function leads to a method with an unbounded influence function. In contrast to that, using the logistic loss function or its asymmetric generalization provides robust methods with bounded influence functions if used in combination with the Gaussian RBF kernel.

Unfortunately, the above results require a twice continuously differentiable loss function and therefore they cannot be used to investigate methods based on e.g. the  $\varepsilon$ -insensitive loss or Huber's loss. Our next results which in particular bound the difference quotient used in the definition of the influence function applies to all convex loss functions of some type  $(a, p)$  and hence partially resolves the above problem for non-differentiable loss functions.

**Theorem 20** *Let  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a convex loss function of some type  $(a, p)$  and  $\mathbb{P}, \tilde{\mathbb{P}}$  be distributions  $X \times Y$  with  $|\mathbb{P}|_a < \infty$  and  $|\tilde{\mathbb{P}}|_a < \infty$ . Furthermore, let  $H$  be a RKHS of a bounded, continuous kernel on  $X$ . Then for all  $\lambda > 0, \varepsilon > 0$  we have*

$$\|f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}\|_H \leq \frac{2c\varepsilon (|\mathbb{P}|_a + |\tilde{\mathbb{P}}|_a + 2^{p+1}\delta_{\mathbb{P},\lambda}^p \|k\|_\infty^p + 2)}{\sqrt{\lambda\mathcal{R}_{L,\mathbb{P}}(0)}},$$

where  $c$  is the constant of the type  $(a, p)$ -inequality.

For the special case that  $\tilde{\mathbb{P}}$  is the Dirac distribution  $\Delta_z$  concentrated in  $z = (x, y)$  we have  $|\Delta_z|_a = a(y)$  and hence we obtain bounds for the difference quotient which occurs in the definition of the influence function if we divide the bound by  $\varepsilon$ . Unfortunately it then turns out that we can almost never bound the difference quotient *uniformly* in  $z$  by the above result. The reason for this problem is that the  $(a, p)$ -type is a rather loose concept for describing the growth behavior of loss functions. However, if we consider only *invariant* loss functions—and many loss functions used in practice are invariant—we are able to obtain stronger results.

**Theorem 21** *Let  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a convex invariant loss function of upper order  $p \geq 1$ , and  $\mathbb{P}, \tilde{\mathbb{P}}$  be distributions  $X \times Y$  with  $|\mathbb{P}|_p < \infty$  and  $|\tilde{\mathbb{P}}|_p < \infty$ . Furthermore, let  $H$  be a RKHS of a bounded, continuous kernel on  $X$ . Then for all  $\lambda > 0, \varepsilon > 0$  we have*

$$\|f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}\|_H \leq c\varepsilon \|k\|_\infty \frac{|\mathbb{P} - \tilde{\mathbb{P}}|_{p-1} + |\mathbb{P} - \tilde{\mathbb{P}}|_0 (\|k\|_\infty^{p-1} |\mathbb{P}|_p^{(p-1)/2} \lambda^{-(p-1)/2} + 1)}{\lambda},$$

where the constant  $c$  only depends on  $L$  and  $p$ .

Recall that Lipschitz continuous invariant loss functions have upper order  $p = 1$ , and thus for such loss functions only the  $0^{\text{th}}$ -moments  $|\cdot|_0$  occurs in the above theorem. Furthermore for all finite, signed measures  $\mu$  we have  $|\mu|_0 = \|\mu\|_{\mathcal{M}}$ , where  $\|\mu\|_{\mathcal{M}}$  denotes the norm of total variation (see e.g. Brown and Percy (1977)), and hence we immediately obtain

**Corollary 22** *Let  $L : Y \times \mathbb{R} \rightarrow [0, \infty)$  be a Lipschitz continuous, convex and invariant loss function, and  $\mathbb{P}, \tilde{\mathbb{P}}$  be distributions  $X \times Y$  with  $|\mathbb{P}|_1 < \infty$  and  $|\tilde{\mathbb{P}}|_1 < \infty$ . Furthermore, let  $H$  be a RKHS of a bounded, continuous kernel on  $X$ . Then for all  $\lambda > 0, \varepsilon > 0$  we have*

$$\|f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}\|_H \leq \frac{\varepsilon |l|_1 \|k\|_\infty \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}}{\lambda},$$

where  $l : \mathbb{R} \rightarrow [0, \infty)$  is the function associated with  $L$ . In particular considering (1), we have

$$\|SC_n(z; T_n)\|_H \leq 2\lambda^{-1} \|k\|_\infty |l|_1$$

for all  $z \in X \times Y$ .

Finally, let us compare the influence function of kernel based regression methods with the influence function of M-estimators in *linear* regression models with  $f(x_i) = x_i' \theta$ , where  $\theta \in \mathbb{R}^d$  denotes the unknown parameter vector. Let us assume for reasons of simplicity that

the scale parameter  $\sigma \in (0, \infty)$  of the linear regression model is known. For more details about such M-estimators see Hampel *et al.* (1986). The functional  $T(\mathbb{P})$  corresponding to an M-estimator is the solution of

$$\mathbb{E}_{\mathbb{P}} \eta(X, [Y - X'T(\mathbb{P})]/\sigma) X = 0, \quad (14)$$

where the odd function  $\eta(x, \cdot)$  is continuous for  $x \in \mathbb{R}^d$  and  $\eta(x, u) \geq 0$  for all  $x \in \mathbb{R}^d$ ,  $u \in [0, \infty)$ . Almost all proposals of  $\eta$  may be written in the form  $\eta(x, u) = \psi(v(x) \cdot u) \cdot w(x)$ , where  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  is a suitable user-defined function (often continuous, bounded and increasing), and  $w : \mathbb{R}^d \rightarrow [0, \infty)$ ,  $v : \mathbb{R}^d \rightarrow [0, \infty)$  are weight functions. An important subclass of M-estimators are of Mallows-type, i.e.  $\eta(x, u) = \psi(u) \cdot w(x)$ . The influence function of  $T(\mathbb{P}) = \theta$  in the point  $z = (x, y)$  at a distribution  $\mathbb{P}$  for  $(X, Y)$  on  $\mathbb{R}^d \times \mathbb{R}$  is given by

$$\text{IF}(z; T, \mathbb{P}) = M^{-1}(\eta, \mathbb{P}) \cdot \eta(x, [y - x'T(\mathbb{P})]/\sigma) \cdot x, \quad (15)$$

where  $M(\eta, \mathbb{P}) := \mathbb{E}_{\mathbb{P}} \eta'(X, [Y - X'T(\mathbb{P})]/\sigma) X X'$ . An important difference between kernel based regression and M-estimation is that  $\text{IF}(z; T, \mathbb{P}) \in \mathbb{R}^d$  in (15), but  $\text{IF}(z; T, \mathbb{P}) \in H$  in (11) for point mass contamination in the point  $z$ .

A comparison of the influence function of KBR given in (11) and (13) with the influence function of M-estimators given in (15) yields that both influence functions have nevertheless a similar structure. The function  $S = S(L'', k, \mathbb{P})$  for KBR and the matrix  $M(\eta, \mathbb{P})$  for M-estimation do not depend on  $z$ . The terms in the influence functions depending on  $z = (x, y)$ , where the point mass contamination  $\Delta_z$  occurs, are a product of two factors. The first factors are  $-L'(y, f_{\mathbb{P}, \lambda}(x))$  for general KBR,  $\psi(v(x) \cdot (y - x'\theta)/\sigma)$  for general M-estimation,  $l'(y - f_{\mathbb{P}, \lambda}(x))$  for KBR with an invariant loss function, and  $\psi((y - x'\theta)/\sigma)$  for M-estimation of Mallows-type. Hence the first factors are measuring the outlyingness in  $y$ -direction. KBR with an invariant loss function and M-estimators of Mallows-type use first factors which only depend on the residuals. The second factors are  $S^{-1}\Phi(x)$  for the kernel based methods and  $w(x)x$  for M-estimation. Therefore, they do not depend on  $y$  and measure the outlyingness in  $x$ -direction.

Concluding one can say that there is a natural connection between KBR estimation and M-estimation in the sense of the influence function approach. The main difference between the influence functions is of course, that the map  $S^{-1}\Phi(x)$  takes values in the RKHS  $H$  in the case of KBR whereas  $w(x)x \in \mathbb{R}^d$  for M-estimation.

## 6. Examples

In this section we give simple numerical examples to show that:

- KBR with the  $\varepsilon$ -insensitive loss function is indeed more robust than KBR based on the least squares loss function if there are outliers in  $y$ -direction.
- In general there is no hope to obtain robust predictions  $\hat{f}(x)$  with KBR if  $x$  belongs to a subset of the design space  $X$  where no or almost no data points are in the training data set, i.e. if  $x$  is a leverage point.

We constructed a data set with  $n = 101$  points in the following way. There is one explanatory variable  $x_i$  with values from  $-5$  to  $5$  in steps of order  $0.1$ . The responses  $y_i$  are

simulated by  $y_i = x_i + e_i$ , where  $e_i$  is a random number from a normal distribution with expectation 0 and variance 1. The  $\varepsilon$ -SVR and LS-SVR with similar hyperparameters give almost the same fitted curves, see Figure 1(a).

Figure 1(b) shows that  $\varepsilon$ -SVR is much less influenced by outliers in  $y$ -direction (one data point is move to  $(x, y) = (-2, 100)$ ) than LS-SVR due to the different behavior of the first derivative of the loss functions.

Now we add to the original data set sequentially three data points all equal to  $(x, y) = (100, 0)$  which are bad leverage points with respect to a *linear* regression model. The number of such data points has a large impact on KBR with a linear kernel, but the predictions of KBR with a Gaussian RBF kernel are stable but nonlinear, see Figure 1(c).

Now we study the impact of adding to the original data set two data points  $z_1 = (100, 100)$  and  $z_2 = (0, 100)$  on the predictions of KBR, see Figure 1(d). By construction  $z_1$  is a good leverage point and  $z_2$  is a bad leverage point with respect to a *linear* regression model which follows e.g. by computing the highly robust LTS estimator (Rousseeuw, 1984), whereas the roles of these data points are switched for a quadratic model. There is no regression model which can fit all data points well because the  $x$ -components of  $z_1$  and  $z_2$  are equal by construction. This toy example shows that in general one can not hope to obtain robust predictions  $\hat{f}(x)$  for  $\mathbb{E}_P(Y|X = x)$  with KBR if  $x$  belongs to a subset of  $X$  where no or almost no data points are in the training data set because the addition of a single data point can have a big impact on KBR if the RKHS  $H$  is rich enough. Note that the hyperparameters  $\varepsilon$ ,  $\gamma$ , and  $C = 1/(2\lambda)$  were specified in these examples to illustrate certain aspects of KBR and were therefore not determined by a grid search, a Nelder-Mead search or by cross-validation.

## 7. Discussion

In this paper properties of kernel based regression methods including support vector machines were investigated. Consistency of kernel based regression methods was derived and results for the influence function, its difference quotient and the sensitivity curve were established. Our theoretical results show that KBR methods using a loss function with bounded first derivative (*e.g.* logistic loss) in combination with a bounded and rich enough continuous kernel (*e.g.* a Gaussian RBF kernel) are not only consistent and computational tractable, but also offer attractive robustness properties.

Most of our results have analogues in the theory of kernel based classification methods, see e.g. Steinwart (2005), and Christmann and Steinwart (2004). However, since in the classification scenario  $Y$  is only  $\{-1, 1\}$ -valued, many effects of the regression scenario with unbounded  $Y$  do not occur in the above papers. Consequently, we had to develop a variety of new techniques and concepts: one central issue here was to find notions for loss functions which on the one hand are mild enough to cover a wide range of reasonable loss functions, and on the other hand are strong enough to allow meaningful results for both consistency and robustness under minimal conditions on  $Y$ . In our analysis it turned out that the relation between the growth behaviour of the loss function and the tail behaviour of  $Y$  play a central role for both types of results. Interestingly, similar tail properties of  $Y$  are widely used for obtaining consistency of non-parametric regression estimators and for establishing robustness properties of M-estimators in linear regression. For example, Györfi *et al.* (2002) assume  $\mathbb{E}_P Y^2 < \infty$  for the least squares loss, Hampel *et al.* (1986,

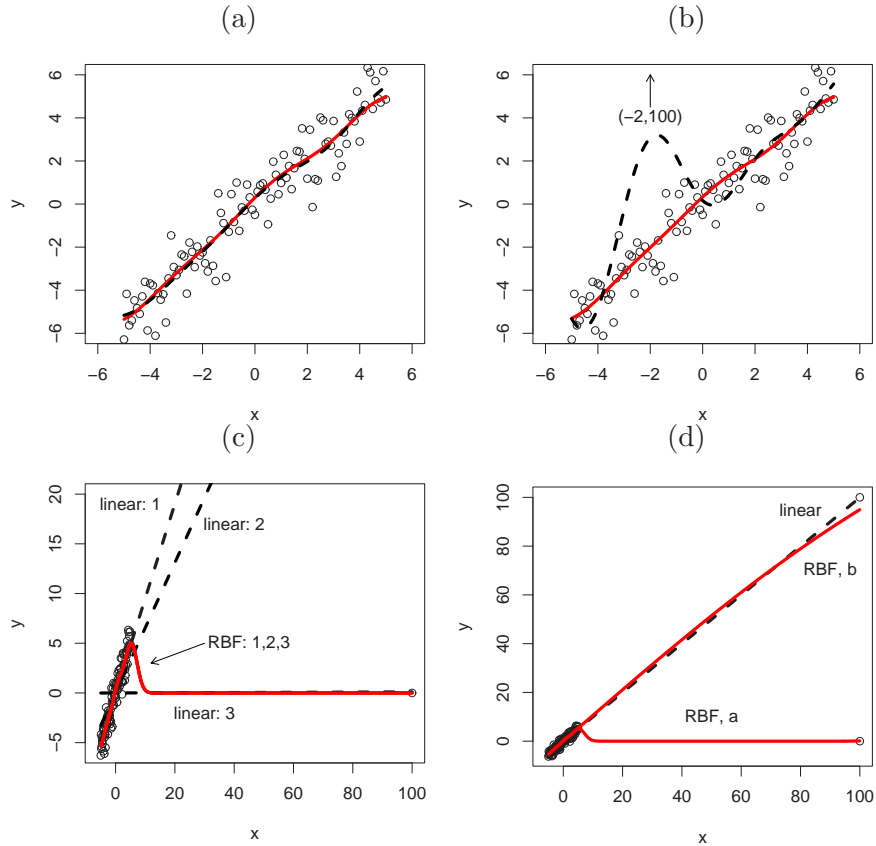


Figure 1: (a) Simulated data set.  $\varepsilon$ -SVR with  $k$ =RBF (solid); LS-SVR with  $k$ =RBF (dashed). (b) one outlier in  $y$ -direction at  $(x, y) = (-2, 100)$ .  $\varepsilon$ -SVR with  $k$ =RBF (solid); LS-SVR with  $k$ =RBF (dashed). (c) Simulated data set with additional 1, 2, or 3 data points in  $(x, y) = (100, 0)$ .  $\varepsilon$ -SVR with  $k$ =linear;  $\varepsilon$ -SVR with  $k$ =RBF. (d) two additional data points in  $(x, y) = (100, 0)$  and  $(x, y) = (100, 100)$ .  $\varepsilon$ -SVR with  $k$ =linear;  $\varepsilon$ -SVR with  $k$ =RBF (RBF,a) and  $k$ =RBF with  $(\varepsilon, \gamma, C) = (0.1, 0.00001, 10000)$  (RBF,b), respectively. The hyperparameters were  $(\varepsilon, \gamma, C) = (0.1, 0.1, 10)$  for  $\varepsilon$ -SVR with  $k$ =RBF,  $(\varepsilon, C) = (0.1, 10)$  for  $\varepsilon$ -SVR with  $k$ =linear, and  $(\gamma, C) = (0.1, 10)$  for LS-SVR with  $k$ =RBF.

p.315) assume existence and non-singularity of  $\mathbb{E}_P \eta'(X, [Y - X'T(P)]/\sigma) XX'$ , and Davies (1993, p.1876) assumes  $\mathbb{E}_P \|X\|(\|X\| + |Y|) < \infty$ . Another important issue was to deal with the estimation error in the consistency analysis. We decided to use a stability approach in order to avoid truncation techniques, so that the proof of our consistency result became surprisingly short. An additional benefit of this approach was that it revealed an interesting connection between the robustness and the consistency of KBR methods. Note that a somewhat similar observation was recently made by Poggio *et al.* (2004) and Mukherjee *et al.* (2004) for a wide class of learning algorithms. However, they assume that the loss function or  $Y$  is *bounded* and hence their results cannot be used in our more general setting.

Our results concerning the influence function of kernel based regression (Theorem 18 and Corollary 19) are valid under the assumption that the loss function is twice continuously differentiable, whereas our other robustness results are valid for more general loss functions. The strong differentiability assumption was made because our proof is based on a classical theorem of implicit functions. We have not investigated whether similar results hold true for continuous but not differentiable loss functions. It may be possible to obtain such results by using an implicit function theorem for non-smooth functions based on a weaker concept than Fréchet differentiability. However, there are indications why a smooth loss function may even be desirable. The function  $-L'$  has a similar role for kernel based regression than the  $\psi$ -function for M-estimators. Huber (1981, p. 51) considered robust estimation in parametric models and investigated the case that the underlying distribution is a mixture of a smooth distribution and a point mass. He showed that an M-estimator has a non-normal limiting behavior if the point mass is at a discontinuity of the derivative of the score function. Since distributions with point masses are not excluded by nonparametric regression methods such as KBR his results indicate that a twice continuously differentiable loss function may guard against such phenomena.

Theorems 18 and 21 and the comments after Theorem 15 show that KBR estimators based on appropriate choices of  $L$  and  $k$  have a bounded influence function if the distribution  $P$  has the tail property  $|P|_a < \infty$ , but are non-robust against small violations of this tail assumption. The deeper reason for this instability is that the theoretical regularized risk itself is defined via  $\mathbb{E}_P L(Y, f(X))$ , which is a non-robust location estimator for the distribution of the losses. This location estimator can be infinite for mixture distributions  $(1 - \varepsilon)P + \varepsilon\tilde{P}$  no matter how small  $\varepsilon > 0$  is. Following general rules of robust estimation in linear regression models, one might replace this non-robust location estimator by a robust alternative like an  $\alpha$ -trimmed mean or the median (Rousseeuw, 1984), which results in

$$f_{P,\lambda}^* = \arg \min_{f \in H} \text{Median}_P L(Y, f(X)) + \lambda \|f\|_H^2, \quad (16)$$

or might use a bounded, non-convex loss function  $L$  (Rousseeuw and Yohai, 1984). We conjecture that  $f_{P,\lambda}^*$  offers additional robustness, but sacrifices computational efficiency. However, such methods are beyond the scope of this paper.

To our best knowledge there are no results on robustness properties of KBR which are comparable to those presented here. However, Suykens *et al.* (2002) proposed the WLS-SVR algorithm which is based on KBR with least squares loss function (LS-SVR) in the following way: a LS-SVR is fitted, then data points with large absolute residuals divided by a robust scale estimate are downweighted, and finally a second LS-SVR is done taking these weights into account. Our theoretical results given in Section 5 show that a robustification of LS-SVR is indeed necessary because its loss function increases too fast. However, it is doubtful whether the WLS-SVR approach completely resolves that problem since a general result of robust statistics is that such weighted estimators already need a robust estimator in the first step, see *e.g.* Rousseeuw (1984) and Yohai *et al.* (1991) for linear regression.

## Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB-475) and of DoMuS (University of Dortmund, “Model building and simulation”) are gratefully acknowledged.



## Appendix: Proofs of the Results

### A.1 Proofs of Section 3

**Proof of Lemma 4.** Assertion *iii)* is trivial and the left inequality of assertion *iv)* is well known from convex analysis. Furthermore, the right inequality of *iv)* easily follows from  $l(r) = |l(r) - l(0)| \leq H(r)|r - 0| = |r|H(r)$  for all  $r \in \mathbb{R}$ . Moreover, the right inequality of *iv)* directly implies *ii)*. Furthermore *i)* can be easily by the left inequality of *vi)* since  $\lim_{|r| \rightarrow \infty} l(r) = \infty$  implies  $H(r) > 0$  for all  $r \neq 0$ . Finally, the last assertion follows from

$$L(y, t) = l(y - t) \leq c(|y - t|^p + 1) \leq c(|y|^p + |t|^p + 1). \quad \square$$

**Proof of Proposition 6.** For bounded measurable functions  $f : X \rightarrow \mathbb{R}$  we have

$$\begin{aligned} \mathcal{R}_{L,P}(f) &= \int_{X \times Y} L(y, f(x)) d\mathbb{P}_X(x, y) \leq c \int_{X \times Y} (a(y) + |f(x)|^p + 1) d\mathbb{P}_X(x, y) \\ &\leq c \|a\|_{L_1(P)} + c \|f\|_{L_p(P)}^p + c < \infty. \quad \square \end{aligned}$$

**Proof of Lemma 7.** For all  $a, b \in \mathbb{R}$  we have  $(|a| + |b|)^p \leq 2^{p-1}(|a|^p + |b|^p)$  if  $p \geq 1$ , and  $(|a| + |b|)^p \leq |a|^p + |b|^p$  otherwise. This obviously implies  $|a|^p \leq 2^{p-1}(|a - b|^p + |b|^p)$  and  $|a|^p \leq |a - b|^p + |b|^p$ , respectively. Now let us assume that we know  $f \in L_p(P)$ . By our preliminary considerations we then obtain

$$\infty > \mathcal{R}_{L,P}(f) \geq c \int_{X \times Y} (|y - f(x)|^p - 1) d\mathbb{P}(x, y) \geq c \int_{X \times Y} (|y|^p - c_p |f(x)|^p - 1) d\mathbb{P}(x, y)$$

for some finite constants  $c > 0$  and  $c_p > 0$ . From this we immediately  $|P|_p < \infty$ . The converse implication can be shown analogously.  $\square$

**Proof of Proposition 9.** Our proof follows DeVito *et al.* (2004). However, the assertion can also be shown elementarily by modifying the proof of Lemma 3.1 in Steinwart (2005). Since  $L$  is of type  $(a, p)$  we observe by (Ekeland and Turnbull, 1983, Prop. III.5.1) that  $\mathcal{R}_{L,P} : L_p(P) \rightarrow \mathbb{R}$  is continuous (actually, their result is only stated for  $X \subset \mathbb{R}$ , but it is straightforward to check that it holds for arbitrary measure spaces). Furthermore,  $\text{id} : H \rightarrow L_p(P)$  is continuous since  $k$  is bounded and hence  $\mathcal{R}_{L,P,\lambda}^{reg} : H \rightarrow \mathbb{R}$  is continuous. This map is also convex, and the set  $\{f \in H : \mathcal{R}_{L,P,\lambda}^{reg}(f) \leq \delta_{P,\lambda}\}$  is non-empty (it contains  $0 \in H$ ) and bounded. Therefore, (Ekeland and Turnbull, 1983, Prop. II.4.6) ensures the existence of  $f_{P,\lambda}$ . The uniqueness follows from the strict convexity of  $\mathcal{R}_{L,P,\lambda}^{reg}$ . The last assertion is trivial.  $\square$

**Proof of Theorem 13.** We begin with proving the general case. By Proposition 12 there exists an  $h \in L_{p'}(P)$  with  $h(x, y) \in \partial_2 L(y, f_{P,\lambda}(x))$  for all  $(x, y) \in X \times Y$ , and  $f_{P,\lambda} = -\frac{1}{2\lambda} \mathbb{E}_P h \Phi$ . Let us first show that  $h$  is integrable for  $\mathbb{Q}$ . To this end using the

shorthand  $K := \|k\|_\infty$  we find

$$\begin{aligned}
 |h(x, y)| \leq |\partial_2 L(y, f_{P,\lambda}(x))| &\leq |L(y, \cdot)|_{[-f_{P,\lambda}(x), f_{P,\lambda}(x)]}|_1 \\
 &\leq |L(y, \cdot)|_{[-\delta_{P,\lambda}K, \delta_{P,\lambda}K]}|_1 \\
 &\leq \frac{2}{\delta_{P,\lambda}K} \|L(y, \cdot)|_{[-2\delta_{P,\lambda}K, 2\delta_{P,\lambda}K]}\|_\infty \\
 &\leq \frac{2c|a(y) + |2\delta_{P,\lambda}K|^p + 1|}{\delta_{P,\lambda}K}.
 \end{aligned} \tag{17}$$

From this we immediately deduce  $h \in L_1(Q)$ .

Now, by the definition of the subdifferential we have

$$h(x, y)(f_{Q,\lambda}(x) - f_{P,\lambda}(x)) \leq L(y, f_{Q,\lambda}(x)) - L(y, f_{P,\lambda}(x)),$$

and hence

$$\mathbb{E}_{(x,y) \sim Q} L(y, f_{P,\lambda}(x)) + \langle f_{Q,\lambda} - f_{P,\lambda}, \mathbb{E}_Q h \Phi \rangle \leq \mathbb{E}_{(x,y) \sim Q} L(y, f_{Q,\lambda}(x)). \tag{18}$$

Moreover an easy calculation shows

$$\lambda \|f_{P,\lambda}\|_H^2 + 2\lambda \langle f_{Q,\lambda} - f_{P,\lambda}, f_{P,\lambda} \rangle + \lambda \|f_{P,\lambda} - f_{Q,\lambda}\|_H^2 = \lambda \|f_{Q,\lambda}\|_H^2. \tag{19}$$

Combining (18) and (19) it follows

$$\begin{aligned}
 \mathcal{R}_{L,Q,\lambda}^{reg}(f_{P,\lambda}) + \langle f_{Q,\lambda} - f_{P,\lambda}, \mathbb{E}_Q h \Phi + 2\lambda f_{P,\lambda} \rangle + \lambda \|f_{P,\lambda} - f_{Q,\lambda}\|_H^2 &\leq \mathcal{R}_{L,Q,\lambda}^{reg}(f_{Q,\lambda}) \\
 &\leq \mathcal{R}_{L,Q,\lambda}^{reg}(f_{P,\lambda}).
 \end{aligned}$$

Therefore by using  $f_{P,\lambda} = -\frac{1}{2\lambda} \mathbb{E}_P h \Phi$  we obtain

$$\begin{aligned}
 \lambda \|f_{P,\lambda} - f_{Q,\lambda}\|_H^2 &\leq \langle f_{P,\lambda} - f_{Q,\lambda}, \mathbb{E}_Q h \Phi - \mathbb{E}_P h \Phi \rangle \\
 &\leq \|f_{P,\lambda} - f_{Q,\lambda}\|_H \cdot \|\mathbb{E}_Q h \Phi - \mathbb{E}_P h \Phi\|_H,
 \end{aligned}$$

which shows the assertion in the general case.

Now let us assume that  $L$  is invariant. As usual we denote the function that represents  $L$  by  $l : \mathbb{R} \rightarrow [0, \infty)$ . Then we easily check that  $h$  satisfies  $h(x, y) \in \partial_2 L(y, f_{P,\lambda}(x)) = -\partial l(y - f_{P,\lambda}(x))$  for all  $(x, y) \in X \times Y$ . Now for  $p = 1$  we see by *iv*) of Lemma 4 that  $l$  is Lipschitz continuous and hence the function  $H$  is constant. Since this gives, cf. (Phelps, 1986, Prop. 1.11),

$$|h(x, y)| \leq H(y - f_{P,\lambda}(x)) \leq \|H\|_\infty$$

we find  $h \in L_\infty(Q)$  which is the assertion for  $p = 1$ . Therefore let us finally consider the case  $p > 1$ . Then for  $(x, y) \in X \times Y$  with  $r := |y - f_{P,\lambda}(x)| \geq 1$  we have

$$|h(x, y)| \leq |\partial l(y - f_{P,\lambda}(x))| \leq H(r) \leq \frac{2}{r} \|l|_{[2r, 2r]}\|_\infty \leq c r^{p-1}$$

for a suitable constant  $c > 0$ . Furthermore, for  $(x, y) \in X \times Y$  with  $|y - f_{P,\lambda}(x)| \leq 1$  we have  $|h(x, y)| \leq |\partial l(y - f_{P,\lambda}(x))| \leq H(y - f_{P,\lambda}(x)) \leq H(1)$ . Together, these estimates show

$$|h(x, y)| \leq \tilde{c} \max\{1, |y - f_{P,\lambda}(x)|^{p-1}\} \leq \tilde{c} \hat{c}_p (1 + |y|^{p-1} + |f_{P,\lambda}(x)|^{p-1})$$

for some constant  $\tilde{c}$  only depending on the loss function  $L$  and  $\hat{c}_p := \max\{1, 2^{p-2}\}$ . Now, using  $p'(p-1) = p$  we obtain  $h \in L_{p'}(\mathbb{Q})$  with

$$\|h\|_{L_{p'}(\mathbb{Q})} \leq \tilde{c} \hat{c}_p (|\mathbb{Q}|_p + \|k\|_\infty^{p-1} \|f_{\mathbb{P},\lambda}\|_H^{p-1} + 1). \quad (20)$$

Finally, for later purpose we note that our previous considerations for  $p = 1$  showed that (20) also holds in this case.  $\square$

## A.2 Proofs of Section 4

In order to prove Theorem 15 we need some preliminary results. Our first lemma shows that the influence of the regularization term  $\lambda \|f_{\mathbb{P},\lambda}\|_H^2$  used in the definition of kernel based regression methods vanishes for  $\lambda \rightarrow 0$ .

**Lemma 23** *Let  $L$  be a loss function,  $H$  be a RKHS over  $X$  with continuous kernel  $k$  and  $\mathbb{P}$  be a distribution on  $X \times Y$ . Suppose that the minimizer  $f_{\mathbb{P},\lambda}$  of (2) exists for all  $\lambda > 0$ . Then we have*

$$\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,\mathbb{P},\lambda}^{reg}(f_{\mathbb{P},\lambda}) = \inf_{f \in H} \mathcal{R}_{L,\mathbb{P}}(f) := \mathcal{R}_{L,\mathbb{P},H}.$$

**Proof of Lemma 23.** Let  $\varepsilon > 0$  and  $f_\varepsilon \in H$  with  $\mathcal{R}_{L,\mathbb{P}}(f_\varepsilon) \leq \mathcal{R}_{L,\mathbb{P},H} + \varepsilon$ . Then for all  $\lambda < \varepsilon \|f_\varepsilon\|_H^{-2}$  we have

$$\mathcal{R}_{L,\mathbb{P},H} \leq \lambda \|f_{\mathbb{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda}) \leq \lambda \|f_\varepsilon\|_H^2 + \mathcal{R}_{L,\mathbb{P}}(f_\varepsilon) \leq 2\varepsilon + \mathcal{R}_{L,\mathbb{P},H}. \quad \square$$

The next lemma shows that universal kernels have zero approximation error with respect to the  $L$ -risk.

**Lemma 24** *Let  $L$  be an convex and invariant loss function of lower and upper order  $p \geq 1$ , and  $H$  be a RKHS of a universal kernel. Then for all distributions  $\mathbb{P}$  on  $X \times Y$  with  $|\mathbb{P}|_p \leq \infty$  we have*

$$\mathcal{R}_{L,\mathbb{P},H} = \mathcal{R}_{L,\mathbb{P}}.$$

**Proof of Lemma 24.** We split the proof into two parts by first showing

$$\mathcal{R}_{L,\mathbb{P},H} = \mathcal{R}_{L,\mathbb{P},L_\infty(\mathbb{P})} := \inf\{\mathcal{R}_{L,\mathbb{P}}(f) \mid f \in L_\infty(\mathbb{P})\} \quad (21)$$

and then establishing

$$\mathcal{R}_{L,\mathbb{P},L_\infty(\mathbb{P})} = \mathcal{R}_{L,\mathbb{P}}. \quad (22)$$

In order to prove (21) let us choose an  $\varepsilon > 0$  and a  $g \in L_\infty(\mathbb{P})$ . Then by the universality of  $H$  there exists a function  $f \in H$  with  $\|f\|_\infty \leq \|g\|_\infty$  and  $\mathbb{P}_X(|f-g| \geq \varepsilon) \leq \varepsilon$ . Furthermore, with the arguments used in the proof of Theorem 13 we find

$$H(|y| + \|g\|_\infty) \leq c(|y|^{p-1} + \|g\|_\infty^{p-1} + 1)$$

for all  $y \in Y$  and a suitable constant  $c > 0$  only depending on  $L$  and  $p$ . This shows  $(y \mapsto H(|y| + \|g\|_\infty)) \in L_1(\mathbb{P})$  and hence we obtain

$$\begin{aligned} |\mathcal{R}_{L,\mathbb{P}}(f) - \mathcal{R}_{L,\mathbb{P}}(g)| &\leq \int_{X \times Y} |l(y - f(x)) - l(y - g(x))| d\mathbb{P}(x, y) \\ &\leq \int_{X \times Y} H(|y| + \|g\|_\infty) |f(x) - g(x)| d\mathbb{P}(x, y) \\ &\leq \left\| (y \mapsto H(|y| + \|g\|_\infty)) \right\|_{L_1(\mathbb{P})} (\varepsilon + 2\varepsilon \|g\|_\infty). \end{aligned}$$

From this we easily get (21). In order to show (22) let us first recall that for all  $f : X \rightarrow Y$  with  $\mathcal{R}_{L,\mathbb{P}}(f) < \infty$  we have  $f \in L_p(\mathbb{P})$  by Lemma 7. Therefore we can restrict the infimum used in the definition of  $\mathcal{R}_{L,\mathbb{P}}$  to functions contained in  $L_p(\mathbb{P})$ . Now for a fixed  $f \in L_p(\mathbb{P})$  we define  $h(x, y) := (|y|^{p-1} + |f(x)|^{p-1} + 1) |f(x)|$ ,  $(x, y) \in X \times Y$ . By Hölder's inequality and  $p'(p-1) = p$  we then find

$$\begin{aligned} &\int_{X \times Y} (|y|^{p-1} + |f(x)|^{p-1}) |f(x)| d\mathbb{P}(x, y) \\ &\leq 2 \left( \int_{X \times Y} |y|^{p'(p-1)} + |f(x)|^{p'(p-1)} d\mathbb{P}(x, y) \right)^{1/p'} \|f\|_{L_p(\mathbb{P})} \\ &< \infty, \end{aligned}$$

where in the last estimate we used  $f \in L_p(\mathbb{P})$  and  $((x, y) \mapsto y) \in L_p(\mathbb{P})$ . This shows  $h \in L_1(\mathbb{P})$ . Let us now define  $f_n := \mathbf{1}_{\{|f| \leq n\}} f$ ,  $n \geq 1$ , where  $\mathbf{1}_A$  denotes the indicator function of  $A$ . Then we obtain

$$\begin{aligned} |\mathcal{R}_{L,\mathbb{P}}(f_n) - \mathcal{R}_{L,\mathbb{P}}(f)| &\leq \int |l(y - f_n(x)) - l(y - f(x))| d\mathbb{P}(x, y) \\ &\leq \int_{|f| \geq n} H(|y| + |f(x)|) |f(x)| d\mathbb{P}(x, y) \\ &\leq c \int_{|f| \geq n} h d\mathbb{P}(x, y), \end{aligned}$$

and hence we get  $\lim_{n \rightarrow \infty} \mathcal{R}_{L,\mathbb{P}}(f_n) = \mathcal{R}_{L,\mathbb{P}}$  since  $h \in L_1(\mathbb{P})$ .  $\square$

Under the assumptions of Lemma 23 and Lemma 24 we immediately see that  $\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) \rightarrow \mathcal{R}_{L,\mathbb{P}}$  holds for  $\lambda_n \rightarrow 0$ . Therefore, we obtain  $L$ -consistency whenever we can show that  $|\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) - \mathcal{R}_{L,\mathbb{P}}(\hat{f}_{n,\lambda_n})| \rightarrow 0$  holds in probability for  $n \rightarrow \infty$  and suitable null sequences  $(\lambda_n)$ . Our main tool for ensuring this convergence will be Theorem 13 which in particular describes the behavior of  $\|f_{\mathbb{P},\lambda_n} - \hat{f}_{n,\lambda_n}\|_\infty$  if we let  $\mathbb{Q}$  be an empirical measure based on a sample set of length  $n$ . The next result shows how the norm of this difference can be used to estimate  $|\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) - \mathcal{R}_{L,\mathbb{P}}(\hat{f}_{n,\lambda_n})|$ .

**Lemma 25** *Let  $L$  be a convex invariant loss function of some type  $p \geq 1$  and  $\mathbb{P}$  be a distribution on  $X \times Y$  with  $|\mathbb{P}|_p < \infty$ . Then there exists a constant  $c_p > 0$  only depending on  $L$  and  $p$  such that for all bounded measurable functions  $f, g : X \rightarrow Y$  we have*

$$|\mathcal{R}_{L,\mathbb{P}}(f) - \mathcal{R}_{L,\mathbb{P}}(g)| \leq c_p (|\mathbb{P}|_{p-1} + \|f\|_\infty^{p-1} + \|g\|_\infty^{p-1} + 1) \|f - g\|_\infty.$$

**Proof of Lemma 25.** Again we have  $H(|y| + |a|) \leq \tilde{c}_p(|y|^{p-1} + |a|^{p-1} + 1)$  for all  $a \in \mathbb{R}$ ,  $y \in Y$ , and a suitable constant  $\tilde{c}_p > 0$  depending on  $L$  and  $p$ . Furthermore, we find

$$\begin{aligned} |\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}(g)| &\leq \int |l(y - f(x)) - l(y - g(x))| dP(x, y) \\ &\leq \int H(|y| + \|f\|_\infty + \|g\|_\infty) |f(x) - g(x)| dP(x, y). \end{aligned}$$

Now we easily obtain the assertion by combining both estimates.  $\square$

Let us now deal with the stochastic analysis of  $|\mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P}(\hat{f}_{n,\lambda})| \rightarrow 0$ . To this end we need the following lemma.

**Lemma 26** *Let  $Z$  be a measurable space,  $P$  be a distribution on  $Z$ ,  $H$  a Hilbert space and  $f : Z \rightarrow H$  be a measurable function with  $\|f\|_q := (\mathbb{E}_P \|f\|_H^q)^{1/q} < \infty$  for some  $q \in (1, \infty)$ . We write  $q^* := \min\{1/2, 1/q'\}$ . Then there exists a universal constant  $c_q > 0$  such that for all  $\varepsilon > 0$  and all  $n \geq 1$  we have*

$$P^n \left( (z_1, \dots, z_n) \in Z^n : \left\| \frac{1}{n} \sum_{i=1}^n f(z_i) - \mathbb{E}_P f \right\| \geq \varepsilon \right) \leq c_q \left( \frac{\|f\|_q}{\varepsilon n^{q^*}} \right)^q.$$

For the proof of Lemma 26 we have to recall some basics from local Banach space theory. To this end we call a sequence of independent, symmetric  $\{-1, 1\}$ -valued random variables  $(\varepsilon_i)$  a *Rademacher sequence*. Now let  $E$  be a Banach space,  $(X_i)$  be an i.i.d. sequence of  $E$ -valued, centered random variables and  $(\varepsilon_i)$  be a Rademacher sequence which is independent to  $(X_i)$ . Then for all  $1 \leq p < \infty$  and all  $n \geq 1$  we have (see Hoffmann-Jørgensen (1974, Cor. 4.2))

$$\mathbb{E} \left\| \sum_{i=1}^n X_i \right\|^p \leq 2^p \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i X_i \right\|^p. \quad (23)$$

Furthermore, a  $E$  is said to have type  $p$ ,  $1 \leq p \leq 2$ , if there exists a constant  $c_p(E) > 0$  such that for all  $n \geq 1$  and all finite sequence  $x_1, \dots, x_n \in E$  we have

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^p \leq c_p(E) \sum_{i=1}^n \|x_i\|^p.$$

Since in the following we are only interested in Hilbert spaces  $H$  we note that these spaces always have type 2 with constant  $c_2(H) = 1$  by orthogonality. Furthermore, they also have type  $p$  for all  $1 \leq p < 2$  by Kahane's inequality (see e.g. (Diestel *et al.*, 1995, p. 211)) which ensures

$$\left( \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^p \right)^{1/p} \leq c_{p,q} \left( \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^q \right)^{1/q}$$

for all  $0 < p, q < \infty$ , all Banach spaces  $E$ , all finite sequence  $x_1, \dots, x_n$  and constants  $c_{p,q} > 0$  only depending on  $p$  and  $q$ . For more information we refer to Ledoux and Talagrand (1991) and Diestel *et al.* (1995). Now we can proceed with

**Proof of Lemma 26.** Let us write  $h(z_1, \dots, z_n) := \frac{1}{n} \sum_{i=1}^n f(z_i) - \mathbb{E}_P f$ . Then a standard calculation shows

$$\mathbb{P}^n(\|h\| \geq \varepsilon) = \int \mathbf{1}_{\{\|h\|^q \geq \varepsilon^q\}} d\mathbb{P} \leq \frac{\mathbb{E}_P \|h\|_H^q}{\varepsilon^q},$$

and hence it remains to estimate  $\mathbb{E}_P \|h\|_H^q$ . To this end recall that by (23) we have

$$\mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \left\| \sum_{i=1}^n f(z_i) - \mathbb{E}_P f \right\|_H^q \leq 2^q \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i (f(z_i) - \mathbb{E}_P f) \right\|_H^q, \quad (24)$$

where the inner expectation on the right hand side is with respect to the Rademacher sequence  $(\varepsilon_i)$ . For  $1 < q \leq 2$  we hence obtain

$$\begin{aligned} \mathbb{E}_P \|h\|_H^q &= n^{-q} \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \left\| \sum_{i=1}^n f(z_i) - \mathbb{E}_P f \right\|_H^q \\ &\leq 2^q n^{-q} \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i (f(z_i) - \mathbb{E}_P f) \right\|_H^q \\ &\leq 2^q c_q n^{-q} \sum_{i=1}^n \mathbb{E}_{z_i \sim \mathbb{P}} \|f(z_i) - \mathbb{E}_P f\|_H^q \\ &\leq 4^q c_q n^{1-q} \mathbb{E}_P \|f\|_H^q, \end{aligned}$$

where  $c_q$  is the type  $q$  constant of Hilbert spaces. From this we easily obtain the assertion for  $1 < q \leq 2$ . Now let us assume that  $2 < q < \infty$ . Then using Kahane's inequality there is a universal constant  $c_q > 0$  with

$$\begin{aligned} \mathbb{E}_P \|h\|_H^q &\leq 2^q n^{-q} \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i (f(z_i) - \mathbb{E}_P f) \right\|_H^q \\ &\leq c_q n^{-q} \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \left( \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i (f(z_i) - \mathbb{E}_P f) \right\|_H^2 \right)^{q/2} \\ &\leq c_q n^{-q} \mathbb{E}_{(z_1, \dots, z_n) \sim \mathbb{P}^n} \left( \sum_{i=1}^n \|f(z_i) - \mathbb{E}_P f\|_H^2 \right)^{q/2} \\ &\leq c_q n^{-q} \left( \sum_{i=1}^n \left( \mathbb{E}_{z_i \sim \mathbb{P}} \|f(z_i) - \mathbb{E}_P f\|_H^q \right)^{2/q} \right)^{q/2} \\ &\leq 2^q c_q n^{-q/2} \mathbb{E}_P \|f\|_H^q, \end{aligned}$$

where we used that Hilbert spaces have type 2 with constant 1. From this estimate we easily obtain the assertion for  $2 < q < \infty$ .  $\square$

**Proof of Theorem 15.** To avoid handling with too many constants let us assume  $\|k\|_\infty = 1$ ,  $|\mathbb{P}|_p = 1$ , and  $c = 2^{-(p+2)}$  for the upper order constant of  $L$ . Then an easy calculation shows  $\mathcal{R}_{L, \mathbb{P}}(0) \leq 1$ . Furthermore, we assume without loss of generality that  $\lambda_n \leq 1$  for all  $n \geq 1$ . Obviously this implies  $\|f_{\mathbb{P}, \lambda_n}\|_\infty \leq \|f_{\mathbb{P}, \lambda_n}\|_H \leq \lambda_n^{-1/2}$ . Now for  $n \in \mathbb{N}$

and regularization parameter  $\lambda_n$  let  $h_n : X \times Y \rightarrow \mathbb{R}$  be the function obtained by Theorem 13. Then our assumptions and (20) give  $\|h_n\|_{L_{p'}(\mathbb{P})} \leq 3 \cdot 2^{p^*/p-2} \lambda_n^{-(p-1)/2}$ . Furthermore, Lemma 25 provides a constant  $c_p > 0$  such that for all  $g \in H$  with  $\|f_{\mathbb{P},\lambda_n} - g\|_H \leq 1$  we have

$$\begin{aligned} & |\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) - \mathcal{R}_{L,\mathbb{P}}(g)| \\ & \leq c_p \left( |\mathbb{P}|_{p-1} + \|f_{\mathbb{P},\lambda_n}\|_\infty^{p-1} + \|g\|_\infty^{p-1} + 1 \right) \|f_{\mathbb{P},\lambda_n} - g\|_\infty \\ & \leq c_p \left( 2 + \lambda_n^{-(p-1)/2} + (\|f_{\mathbb{P},\lambda_n}\|_\infty + \|f_{\mathbb{P},\lambda_n} - g\|_\infty)^{p-1} \right) \|f_{\mathbb{P},\lambda_n} - g\|_H \\ & \leq \tilde{c}_p \lambda_n^{-(p-1)/2} \|f_{\mathbb{P},\lambda_n} - g\|_H, \end{aligned} \quad (25)$$

where  $\tilde{c}_p \geq 1$  is a suitable constant only depending on  $p$  and  $L$ . Now let  $0 < \varepsilon \leq 1$  and  $T$  be a training set of length  $n$  with

$$\|\mathbb{E}_{\mathbb{P}} h_n \Phi - \mathbb{E}_T h_n \Phi\|_H \leq \frac{\lambda_n^{(p+1)/2} \varepsilon}{\tilde{c}_p}. \quad (26)$$

Then Theorem 13 gives  $\|f_{\mathbb{P},\lambda_n} - f_{T,\lambda_n}\|_H \leq \tilde{c}_p^{-1} \lambda_n^{(p-1)/2} \varepsilon \leq 1$  and hence (25) yields

$$|\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) - \mathcal{R}_{L,\mathbb{P}}(f_{T,\lambda_n})| \leq \tilde{c}_p \lambda_n^{-(p-1)/2} \|f_{\mathbb{P},\lambda_n} - f_{T,\lambda_n}\|_H \leq \varepsilon. \quad (27)$$

Let us now estimate the probability of  $T$  satisfying (26). To this end we define  $q := p'$ . Then we have  $q^* := \min\{1/2, 1/q'\} = \min\{1/2, 1/p\} = p/p^*$  and by Lemma 26 we obtain

$$\begin{aligned} P^n \left( T \in (X \times Y)^n : \|\mathbb{E}_{\mathbb{P}} h_n \Phi - \mathbb{E}_T h_n \Phi\|_H \leq \frac{\lambda_n^{(p+1)/2} \varepsilon}{\tilde{c}_p} \right) & \geq 1 - \hat{c}_p \left( \frac{\|h\|_{p'}}{\varepsilon \lambda_n^{(p+1)/2} n^{q^*}} \right)^{p'} \\ & \geq 1 - \hat{c}_p \left( \frac{3 \cdot 2^{p^*/p-2}}{\varepsilon \lambda_n^p n^{p/p^*}} \right)^{p'}, \end{aligned}$$

where  $\hat{c}_p$  is a constant only depending on  $L$  and  $p$ . Now using  $\lambda_n^p n^{p/p^*} = (\lambda_n^{p^*} n)^{p/p^*} \rightarrow \infty$  we find that the probability of samples sets  $T$  satisfying (26) converges to 1 if  $n = |T| \rightarrow \infty$ . As we have seen above this implies that (27) holds true with probability tending to 1. Now, since  $\lambda_n \rightarrow 0$  we additionally have  $|\mathcal{R}_{L,\mathbb{P}}(f_{\mathbb{P},\lambda_n}) - \mathcal{R}_{L,\mathbb{P}}| \leq \varepsilon$  for all sufficiently large  $n$  and hence we finally obtain the assertion.

### A.3 Proofs of Section 5

In order to shorten notations we sometimes write  $L(f)$  instead of  $L(y, f(x))$ . Moreover we also use this kind of notation for derivatives of  $L$ .

The following proofs heavily rely on the implicit function theorem in Banach spaces. Therefore, we recall a simplified version of this theorem (*cf.* Akerkar, 1999; Zeidler, 1986). Here and throughout the rest of the appendix  $B_E$  denotes the open unit ball of a Banach space  $E$ .



**Theorem 27 (Implicit function theorem)** *Let  $E, F$  be Banach spaces and  $G : E \times F \rightarrow F$  be a continuously differentiable map. Suppose that we have  $(x_0, y_0) \in E \times F$  such that  $G(x_0, y_0) = 0$  and  $\frac{\partial G}{\partial F}(x_0, y_0)$  is invertible. Then there exists a  $\delta > 0$  and a continuously differentiable map  $f : x_0 + \delta B_E \rightarrow y_0 + \delta B_F$  such that for all  $x \in x_0 + \delta B_E$ ,  $y \in y_0 + \delta B_F$  we have*

$$G(x, y) = 0 \quad \text{if and only if} \quad y = f(x).$$

Moreover, the derivative of  $f$  is given by

$$f'(x) = - \left( \frac{\partial G}{\partial F}(x, f(x)) \right)^{-1} \frac{\partial G}{\partial E}(x, f(x)).$$

For the application of the implicit function theorem we have to show that certain operators are invertible. For this the following theorem which is known as the Fredholm Alternative (cf. Cheney, 2001) turns out to be helpful:

**Theorem 28 (Fredholm Alternative)** *Let  $E$  be a Banach space and  $S : E \rightarrow E$  be a compact operator. Then  $\text{id}_E + S$  is surjective if and only if it is injective.*

**Proof of Theorem 18.** The key ingredient of our analysis is the map  $G : \mathbb{R} \times H \rightarrow H$  defined by

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)P + \varepsilon\Delta_z} L'(Y, f(X))\Phi(X)$$

for all  $\varepsilon \in \mathbb{R}$ ,  $f \in H$ . Let us first check that its definition makes sense. To this end recall that every  $f \in H$  is a bounded function since we assumed that  $H$  has a bounded kernel  $k$ . As in the proof of Proposition 6 we then find  $\mathbb{E}_P |L'(Y, f(X))| < \infty$  for all  $f \in H$ . Since the boundedness of  $k$  also ensures that  $\Phi$  is a bounded map, we then see that the  $H$ -valued expectation used in the definition of  $G$  is defined for all  $\varepsilon \in \mathbb{R}$  and all  $f \in H$  (Note that for  $\varepsilon \notin [0, 1]$  the  $H$ -valued expectation is with respect to a signed measure, cf. Dudley (2002)). Now for  $\varepsilon \in [0, 1]$  we obtain (see Christmann and Steinwart (2004) for a detailed derivation)

$$G(\varepsilon, f) = \frac{\partial R_{L, (1-\varepsilon)P + \varepsilon\Delta_z, \lambda}^{reg}}{\partial H}(f). \quad (28)$$

Since  $f \mapsto R_{L, (1-\varepsilon)P + \varepsilon\Delta_z, \lambda}^{reg}(f)$  is convex and continuous (cf. proof of Prop. 9) for all  $\varepsilon \in [0, 1]$  equation (28) shows that we have  $G(\varepsilon, f) = 0$  if and only if  $f = f_{(1-\varepsilon)P + \varepsilon\Delta_z, \lambda}$  for such  $\varepsilon$ . Our aim is to show the existence of a differentiable function  $\varepsilon \mapsto f_\varepsilon$  defined on a small interval  $(-\delta, \delta)$  for some  $\delta > 0$  that satisfies  $G(\varepsilon, f_\varepsilon) = 0$  for all  $\varepsilon \in (-\delta, \delta)$ . Once we have shown the existence of this function we immediately obtain

$$IF(z; T, P) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0).$$

For the existence of  $\varepsilon \mapsto f_\varepsilon$  we only have to check by Theorem 27 that  $G$  is continuously differentiable and that  $\frac{\partial G}{\partial H}(0, f_{P, \lambda})$  is invertible.

Let us start with the first. To this end we find by an easy calculation

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_P L'(Y, f(X))\Phi(X) + \mathbb{E}_{\Delta_z} L'(Y, f(X))\Phi(X), \quad (29)$$

and a slightly more involved computation (cf. Christmann and Steinwart (2004)) shows

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda \text{id}_H + \mathbb{E}_{(1-\varepsilon)\mathbb{P} + \varepsilon\Delta_z} L''(Y, f(X)) \langle \Phi(X), \cdot \rangle \Phi(X) = S. \quad (30)$$

In order to prove that  $\frac{\partial G}{\partial \varepsilon}$  is continuous we fix an  $\varepsilon$  and a convergent sequence  $f_n \rightarrow f$  in  $H$ . Since  $H$  has a bounded kernel the sequence of functions  $(f_n)$  is then uniformly bounded. By the continuity of  $L'$  we thus find a measurable bounded function  $g : Y \rightarrow \mathbb{R}$  with  $L'(y, f_n(x)) \leq L'(y, g(y))$  for all  $n \geq 1$  and all  $(x, y) \in X \times Y$ . As in the proof of Proposition 6 we find  $(y \mapsto L(y, g(y))) \in L_1(\mathbb{P})$  and therefore an application of Lebesgue's theorem for Bochner integrals gives the continuity of  $\frac{\partial G}{\partial \varepsilon}$ . Since the continuity of  $G$  and  $\frac{\partial G}{\partial H}$  can be shown analogously we obtain that  $G$  is continuously differentiable (cf. Akerkar, 1999).

In order to show that  $\frac{\partial G}{\partial H}(0, f_{\mathbb{P}, \lambda})$  is invertible it suffices to show by the Fredholm Alternative that  $\frac{\partial G}{\partial H}(0, f_{\mathbb{P}, \lambda})$  is injective and that

$$Ag := \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P}, \lambda}(X)) g(X) \Phi(X), \quad g \in H,$$

defines a compact operator on  $H$ .

To show the compactness of the operator  $A$  recall that  $X$  and  $Y$  are Polish spaces (cf. Dudley (2002)) since we assumed that  $X$  and  $Y$  are closed. Furthermore, Borel probability measures on Polish spaces are *regular* by Ulam's theorem, i.e. they can be approximated from inside by compact sets, cf. (Bauer, 1990, p. 180). In our situation, this means that for all  $n \geq 1$  there exists a compact subset  $X_n \times Y_n \subset X \times Y$  with  $\mathbb{P}(X_n \times Y_n) \geq 1 - \frac{1}{n}$ . Now we define a sequence of operators  $A_n : H \rightarrow H$  by

$$A_n g := \int_{X_n} \int_{Y_n} L''(y, f_{\mathbb{P}, \lambda}(x)) \mathbb{P}(dy|x) g(x) \Phi(x) d\mathbb{P}_X(x) \quad (31)$$

for all  $g \in H$ . Note that if  $X \times Y$  is compact we can choose  $X_n \times Y_n := X \times Y$  which implies  $A = A_n$ . Let us now show that all  $A_n$  are compact operators. To this end we first observe for  $g \in B_H$ ,  $n \geq 1$ , and  $x \in X$  that

$$h_g(x) := \int_{Y_n} L''(y, f_{\mathbb{P}, \lambda}(x)) |g(x)| \mathbb{P}(dy|x) \leq c \|k\|_{\infty} \int_{Y_n} (a(y) + |f_{\mathbb{P}, \lambda}(x)|^p + 1) \mathbb{P}(dy|x) =: h(x)$$

for  $a : Y \rightarrow \mathbb{R}$ ,  $p \geq 1$  and  $c > 0$  according to the  $(a, p)$ -type of  $L''$ . Obviously, we have  $h \in L_1(\mathbb{P}_X)$  which implies  $h_g \in L_1(\mathbb{P}_X)$  with  $\|h_g\|_1 \leq \|h\|_1$  for all  $g \in B_H$ . Consequently  $d\mu_g := h_g d\mathbb{P}_X$  and  $d\mu := h d\mathbb{P}_X$  are finite measures and by (Diestel and Uhl, 1977, Cor. 8 on p. 48) we hence obtain

$$\begin{aligned} A_n g &= \int_{X_n} \text{sign } g(x) \Phi(x) h_g(x) d\mathbb{P}_X(x) = \int_{X_n} \text{sign } g(x) \Phi(x) d\mu_g(x) \\ &\in \mu_g(X_n) \overline{\text{aco } \Phi(X_n)} \\ &\subset \mu(X_n) \overline{\text{aco } \Phi(X_n)}, \end{aligned}$$

where  $\text{aco } \Phi(X_n)$  denotes the absolute convex hull of  $\Phi(X_n)$ , and the closure is with respect to  $\|\cdot\|_H$ . Now using the continuity of  $\Phi$  we see that  $\Phi(X_n)$  is compact and hence so is the

closure of  $\text{aco } \Phi(X_n)$ . This shows that  $A_n$  is a compact operator. In order to see that  $A$  is compact, it therefore suffices to show  $\|A_n - A\| \rightarrow 0$  for  $n \rightarrow \infty$ . Recalling that the convexity of  $L$  implies  $L'' \geq 0$  the latter convergence follows from  $\mathbb{P}(X_n \times Y_n) \geq 1 - \frac{1}{n}$ ,  $L'' \circ f_{\mathbb{P},\lambda} \in L_1(P)$ , and

$$\begin{aligned} \|A_n g - A g\| &= \left\| \int_{(X \times Y) \setminus (X_n \times Y_n)} L''(y, f_{\mathbb{P},\lambda}(x)) g(x) \Phi(x) d\mathbb{P}(x, y) \right\| \\ &\leq \int_{(X \times Y) \setminus (X_n \times Y_n)} L''(y, f_{\mathbb{P},\lambda}(x)) |g(x)| \|\Phi(x)\| d\mathbb{P}(x, y) \\ &\leq \|k\|_\infty^2 \|g\|_H \int_{(X \times Y) \setminus (X_n \times Y_n)} L''(y, f_{\mathbb{P},\lambda}(x)) d\mathbb{P}(x, y). \end{aligned}$$

Let us now show that  $\frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda}) = 2\lambda \text{id}_H + A$  is injective. To this end let us choose a  $g \in H$  with  $g \neq 0$ . Then we find

$$\begin{aligned} \langle (2\lambda \text{id}_H + A)g, (2\lambda \text{id}_H + A)g \rangle &= 4\lambda^2 \langle g, g \rangle + 4\lambda \langle g, Ag \rangle + \langle Ag, Ag \rangle \\ &> 4\lambda \langle g, Ag \rangle \\ &= 4\lambda \langle g, \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X)) g(X) \Phi(X) \rangle \\ &= 4\lambda \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X)) g^2(X) \\ &\geq 0, \end{aligned}$$

which shows the injectivity.

As already described we can now apply the implicit function theorem to see that  $\varepsilon \mapsto f_\varepsilon$  is differentiable on a small interval  $(-\delta, \delta)$ . Furthermore, (29) and (30) yield

$$\begin{aligned} IF(z; T, \mathbb{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0) &= -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}) \\ &= S^{-1} (\mathbb{E}_{\mathbb{P}}(L'(Y, f_{\mathbb{P},\lambda}(X)) \Phi(X))) - L'(y, f_{\mathbb{P},\lambda}(x)) S^{-1} \Phi(x). \quad \square \end{aligned}$$

**Proof of Theorem 20.** Let us write  $\mathbb{Q} := (1 - \varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}}$ . By Theorem 13 there then exists a bounded, measurable function  $h : X \times Y \rightarrow \mathbb{R}$  independent of  $\varepsilon$  and  $\tilde{\mathbb{P}}$  such that we have

$$\begin{aligned} \|f_{\mathbb{P},\lambda} - f_{(1-\varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}},\lambda}\|_H &\leq \lambda^{-1} \|\mathbb{E}_{\mathbb{P}} h \Phi - \mathbb{E}_{(1-\varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}}} h \Phi\|_H \\ &= \varepsilon \lambda^{-1} \|\mathbb{E}_{\mathbb{P}} h \Phi - \mathbb{E}_{\tilde{\mathbb{P}}} h \Phi\|_H \\ &\leq \varepsilon \lambda^{-1} \|k\|_\infty (\|h\|_{L_1(\mathbb{P})} + \|h\|_{L_1(\tilde{\mathbb{P}})}) \\ &\leq \frac{2\varepsilon c (|\mathbb{P}|_a + |\tilde{\mathbb{P}}|_a + 2^{p+1} |\delta_{\mathbb{P},\lambda}| \|k\|_\infty^p + 2)}{\lambda \delta_{\mathbb{P},\lambda}} \end{aligned}$$

where in the last estimate we used (17).  $\square$

**Proof of Theorem 21.** Let us use the notations of the previous proof. Using  $\|f_{\mathbb{P},\lambda}\|_\infty \leq \|k\|_\infty \sqrt{\mathcal{R}_{L,\mathbb{P}}(0)/\lambda}$  we then obtain analogously to the proof of Theorem 13 that

$$|h(x, y)| \leq \tilde{c} (|y|^{p-1} + |f_{\mathbb{P},\lambda}(x)|^{p-1} + 1) \leq c \left( |y|^{p-1} + \|k\|_\infty^{p-1} |\mathbb{P}|_p^{(p-1)/2} \lambda^{-(p-1)/2} + 1 \right),$$

where  $\tilde{c}, c > 0$  are constants only depending on  $L$  and  $p$ . With this estimate we find

$$\begin{aligned} & \|f_{\mathbb{P},\lambda} - f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda}\|_H \\ & \leq \varepsilon \lambda^{-1} \|\mathbb{E}_{\mathbb{P}} h\Phi - \mathbb{E}_{\tilde{\mathbb{P}}} h\Phi\|_H \\ & \leq \varepsilon \lambda^{-1} \|k\|_{\infty} \mathbb{E}_{|\mathbb{P}-\tilde{\mathbb{P}}|} |h| \\ & \leq c\varepsilon \|k\|_{\infty} \frac{|\mathbb{P} - \tilde{\mathbb{P}}|_{p-1} + |\mathbb{P} - \tilde{\mathbb{P}}|_0 \left( \|k\|_{\infty}^{p-1} |\mathbb{P}|_p^{(p-1)/2} \lambda^{-(p-1)/2} + 1 \right)}{\lambda}. \quad \square \end{aligned}$$

**Proof of Corollary 22.** Using the notations of the previous proof we have  $|h(x, y)| \leq |l|_1$  for all  $(x, y) \in X \times Y$ . Now the first assertion can be shown analogously to the previous proof. Using (9) the second assertion is a direct consequence of the first one.  $\square$

## References

- AKERKAR, R. (1999). *Nonlinear Functional Analysis*. Narosa Publishing House, New Dehli.
- BAUER, H. (1990). *Maß- und Integrationstheorie*. De Gruyter, Berlin.
- BROWN, A. AND PEARCY, C. (1977). *Introduction to Operator Theory I*. Springer, New York.
- CHRISTMANN, A. (2004). An approach to model complex high-dimensional insurance data. *To appear in: Allg. Statist. Archiv*.
- CHRISTMANN, A. AND STEINWART, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, **5**, 1007–1034.
- DAVIES, P. (1993). Aspects of robust linear regression. *Ann. Statist.*, **21**, 1843–1899.
- DEVITO, E., ROSASCO, L., CAPONNETTO, A., PIANA, M., AND VERRI, A. (2004). Some properties of regularized kernel methods. *Journal of Machine Learning Research*, **5**, 1363–1390.
- DIESTEL, J. AND UHL, J. (1977). *Vector Measures*. American Mathematical Society, Providence.
- DIESTEL, J., JARCHOW, H., AND TONGE, A. (1995). *Absolutely summing operators*. Cambridge University Press.
- DONOHO, D. AND HUBER, P. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, and J. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Jr., Belmont, California, Wadsworth.
- DUDLEY, R. (2002). *Real Analysis and Probability*. Cambridge University Press.
- EKELAND, I. AND TURNBULL, T. (1983). *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press.

- GYÖRFI, L., KOHLER, M., KRZYŻAK, A., AND WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
- HAMPEL, F. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.
- HAMPEL, F., RONCHETTI, E., ROUSSEEUW, P., AND STAHEL, W. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley, New York.
- HOFFMANN-JØRGENSEN, J. (1974). Sums of independent Banach space valued random variables. *Studia Math.*, **52**, 159–186.
- HUBER, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.
- HUBER, P. (1981). *Robust Statistics*. Wiley, New York.
- LEDOUX, M. AND TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer, Berlin.
- MENDES, B. AND TYLER, D. (1996). Constrained M-estimation for regression. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Hubers 60th Birthday*, pages 299–320, Lecture Notes in Statistics, Springer, New York.
- MUKHERJEE, S., NIYOGI, P., POGGIO, T., AND RIFKIN, R. (2004). Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. CBCL paper 223, MIT, Cambridge, MA.
- PHELPS, R. (1986). *Convex Functions, Monotone Operators and Differentiability*, volume 1364 of *Lecture Notes in Math*. Springer.
- POGGIO, T., RIFKIN, R., MUKHERJEE, S., AND NIYOGI, P. (2004). General conditions for predictivity in learning theory. *Nature*, **428**, 419–422.
- ROUSSEEUW, P. AND HUBERT, M. (1999). Regression depth. *Journal of the American Statistical Association*, **94**, 388–433.
- ROUSSEEUW, P. AND YOHAI, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, and D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, volume 26 of *Lecture Notes in Statistics*, pages 256–272, New York. Springer.
- ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.
- SCHÖLKOPF, B. AND SMOLA, A. (2002). *Learning with Kernels*. MIT Press.
- STEINWART, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.
- STEINWART, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, **4**, 1071–1105.

- STEINWART, I. (2005). Consistency of support vector machines and other regularized kernel machines. *To appear in: IEEE Trans. Inform. Theory.*
- SUYKENS, J., GESTEL, T. V., BRABANTER, J. D., MOOR, B. D., AND VANDEWALLE, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.
- TUKEY, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.
- VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley, New York.
- YOHAI, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, **15**, 642–656.
- YOHAI, V., STAHEL, W., AND ZAMAR, R. (1991). A procedure for robust estimation and inference in linear regression. In W. Stahel and S. Weisberg, editors, *Directions in robust statistics and diagnostics, Part II*, pages 365–374, Springer, New York.
- YOSIDA, K. (1974). *Functional Analysis*. Springer, Berlin, 4<sup>th</sup> edition.
- ZHANG, T. (2001). Convergence of large margin separable linear classification. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press.

ANDREAS CHRISTMANN  
UNIVERSITY OF DORTMUND  
DEPARTMENT OF STATISTICS  
44221 DORTMUND  
GERMANY

E-MAIL:  
christmann@statistik.uni-dortmund.de

INGO STEINWART  
CCS-3  
MAIL STOP B256  
LOS ALAMOS NATIONAL LABORATORY  
LOS ALAMOS, NM 87545  
USA

E-MAIL: ingo@lanl.gov