Journal of Machine Learning Research 5 (2004) 1007–1034 Submitted 6/03; Revised 05/04; Published 8/04

On Robustness Properties of Convex Risk Minimization Methods for Pattern Recognition

Andreas Christmann CHRISTMANN@STATISTIK.UNI-DORTMUND.DE
University of Dortmund
Department of Statistics
44221 Dortmund, Germany

Ingo Steinwart INGO@LANL.GOV
Modeling, Algorithms and Informatics Group, CCS-3
Mail Stop B256
Los Alamos National Laboratory
Los Alamos, NM 87545, USA

Editor: Peter Bartlett

Abstract

The paper brings together methods from two disciplines: machine learning theory and robust statistics. We argue that robustness is an important aspect and we show that many existing machine learning methods based on the convex risk minimization principle have − besides other good properties − also the advantage of being robust. Robustness properties of machine learning methods based on convex risk minimization are investigated for the problem of pattern recognition. Assumptions are given for the existence of the influence function of the classifiers and for bounds on the influence function. Kernel logistic regression, support vector machines, least squares and the AdaBoost loss function are treated as special cases. Some results on the robustness of such methods are also obtained for the sensitivity curve and the maxbias, which are two other robustness criteria. A sensitivity analysis of the support vector machine is given.

Keywords: AdaBoost loss function, influence function, kernel logistic regression, robustness, sensitivity curve, statistical learning, support vector machine, total variation

1. Introduction

In statistical learning theory the principle of regularized empirical risk minimization based on convex loss functions plays an important role, see Vapnik (1998). One strong argument in favor of such methods is that many classifiers based on convex loss functions are universally consistent under weak conditions. Nevertheless, it is important to investigate robustness properties for such statistical learning methods for the following reasons. In almost all cases statistical models are only approximations to the true random process which generated a given data set and for which a method for analyzing the data is designed. Hence the natural question arises what impact such deviations may have on the results. J.W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 (Hampel et al., 1986, p. 21):

> *A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that*

> *this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.*

The main aims of robust statistics are the description of the structure best fitting the *bulk* of the data and the identification of points deviating from this structure or deviating substructures for further treatment, *cf.* Hampel et al. (1986). Very briefly, a good robust method can be described as follows.

- If the strict model assumptions are violated, than the results of a robust method are only influenced in a moderate way by a few data points, which deviate grossly from the structure of the bulk of the data set, or by many data points, which deviate only mildly from the structure of the bulk of the data set.

- A robust method should have a reasonable high efficiency when the data set was in fact generated by the assumed model.

In practice one has to apply machine learning methods to a data set with a finite sample size. Machine learning methods are nonparametric tools. Nevertheless, the robustness issue is important, because the classical assumption that all data points were *independently* generated by the *same* distribution can be violated in practice. One reason is that outliers often occur in real data sets. Outliers can be described as data points which "are far away ... from the pattern set by the majority of the data", see Hampel et al. (1986, p. 25). Sometimes outliers are even correlated, which contradicts the classical assumption that the observations in the data set were generated in an independent manner. There are many reasons for the occurrence of outliers, *e.g.* typing errors and gross errors, which are errors due to a source of deviations which acts only occasionally but is quite powerful. *E.g.* undetected outliers may have an extreme impact on the estimation of insurance tariffs computed by motor vehicle insurance companies. From a robustness point of view the occurrence of outliers is only one of several possible deviations from the assumed model. There are often no or virtually no gross errors in high-quality data, but 1% to 10% of gross errors in routine data seem to be more the rule than the exception, *cf.* Hampel et al. (1986, p. 27f). Especially in large data mining problems the data quality is sometimes far from being optimal, *cf.* Hand et al. (2001) or Hipp et al. (2001). Obviously, it is *not* the goal to *model* the occurrence of typing errors or gross errors, because it is unlikely that they will occur in the same manner for other data sets which will be collected in the future. Goals of robust statistics are to investigate the impact such data points can have on the results of estimation, testing or prediction methods and the development of methods such that the impact of such data points is bounded.

We like to give an example showing that robustness properties of statistical methods can be very important in practice. A data set from 15 German motor vehicle insurance companies from the Verband öffentlicher Versicherer in Düsseldorf, Germany, is investigated by Christmann (2004). The main goals are the estimation of the expected claim amount, which is the primary response variable, and the probability that a customer has at least one claim within one year, which is the secondary response variable. It is well-known that even the very weak assumptions typically made by machine learning methods, see Section 2, may be violated for such data sets for the following reasons. The true values of the claim amount, *i.e.* the primary response variable, is not known exactly for all customers. *E.g.* if a major accident occurs in November, the exact claim size will often not be known at the end of the year and perhaps not even at the end of the following year. Possible reasons are law-suits or the case of physical injuries. In this case, a statistician will have to use more or less appropriate estimations of the exact claim size to construct a new insurance tariff for the next year. Hence, the empirical distribution of the claim amount is in general a mixture of really observed values and of estimated claim amounts. Further, some explanatory variables may have imprecise values. *E.g.* there is a variable describing how many kilometers a customer is driving with the car within one year. The customer has to choose between some categories, *e.g.* below 9000 kilometers, between 9000 and 12000 kilometers, between 12000 and 15000 kilometers, and so on. There are reasons making it plausible, that a percentage of these values in the data set are too small, *e.g.* it is well-known to the customers that the premium of an insurance tariff increases for increasing values of this variable. The data set contains data from more than four million customers with more than 70 variables. Hence there is a high probability that some data points are typing errors, although the data set is of high quality.

There are different approaches to robustness in the statistical literature. For statistics representable as a functional of the empirical distribution, qualitative robustness, which is defined as equicontinuity of the distributions of the statistic as the sample size changes, is closely related to continuity of the statistic viewed as a functional in the weak(-star) topology, *cf.* Huber (1981, p. 7f) or Hampel et al. (1986, p. 41). The concept of qualitative robustness is a rather weak robustness criterion. It has the disadvantage that it does not offer arguments how to choose among different qualitative robust procedures. Huber's minimax approach of robust statistics (Huber, 1964, 1981) is to minimize the maximum asymptotic variance of the estimator within a neighborhood of the model. Other strategies of robust statistics are Hampel's influence function (Hampel, 1974; Hampel et al., 1986), the finite sample breakdown point proposed by Donoho and Huber (1983), the approach based on least favourable local alternatives (Rieder, 1994), and the regression depth method proposed by Rousseeuw and Hubert (1999).

Here, we will mainly use the approach based on the influence function. This approach can be applied to quite general models and the influence function has a nice interpretation, because it is a special Gâteaux derivative, see Section 3. A map T is called robust in the theory of robust statistics based on influence functions, if T has as a *bounded* influence function. From the viewpoint of robust statistics it is important to investigate the impact a small amount of contamination of the 'true' probability measure $\mathbb{P}$ can have on the statistical learning process which is specified via the regularized theoretical risk, *i.e.* the objective functions $R^{reg}_{L,\mathbb{P},\lambda}(.)$ and $R^{reg}_{L,\mathbb{P},\lambda}(.,.)$ given in (6) and (7).

This paper investigates robustness properties of statistical learning methods based on convex risk minimization and is organized as follows. Section 2 gives some notions on convex risk minimization methods. Section 3 gives the definitions of the influence function, the sensitivity curve, and the maxbias, which are the robustness concepts we are dealing with. Section 4 and Section 5 contain the main results. For practical applications Theorem 12 is our most important result and it covers a broad class of loss functions including the ones used by SVM, kernel logistic regression, AdaBoost and least squares. In Section 4 sufficient conditions are given for the existence of the influence function for classifiers based on (6) and (7). In Section 5 it is shown that the influence function of the solution of (7) and the difference quotient used in the definition of the influence function for (6) can be bounded independently of z and $\mathbb{P}$. Bounds for the sensitivity curve and for the maxbias are also given. Section 6 describes the results of some simulation experiments to gain insight into the robustness properties of the SVM for finite sample sizes and investigates the impact a single data point can have if a radial basis function kernel or a linear kernel is used. Section 7 contains the conclusion. Finally, the Appendix gives the proofs of the main theorems discussed in this paper and lists some mathematical facts which are used in our proofs.

2. Convex Risk Minimization in Machine Learning

In pattern recognition and statistical machine learning the major goal is the estimation of a functional relationship $y_i \approx f(x_i)$ between an outcome y_i and a vector of explanatory variables $x_i = (x_{i,1}, \ldots, x_{i,k})' \in \mathbb{R}^d$. The function f is unknown. The estimate for f is used to get predictions of an unobserved outcome y_{new} based on an observed value x_{new}. One needs the implicit assumption that the relationship between x_{new} and y_{new} is—at least almost—the same as in the training data set (x_i, y_i), $i = 1, \ldots, n$. Otherwise, it is useless to extract knowledge on f from the training data set. The classical assumption in machine learning is that the training data (x_i, y_i) are independent and identically generated from an underlying unknown distribution $\mathbb{P}$ for a pair of random variables (X_i, Y_i). In practical applications the training data set is often quite large, high dimensional and complex. The quality of the predictor $f(x_i)$ is measured by some loss function $L(y_i, f(x_i))$. The goal is to find a predictor $f_{\mathbb{P}}(x_i)$ that minimizes the expected loss, *i.e.*

$$\mathbb{E}_{\mathbb{P}} L(Y, f_{\mathbb{P}}(X)) = \min_f \mathbb{E}_{\mathbb{P}} L(Y, f(X)), \tag{1}$$

where $\mathbb{E}_{\mathbb{P}} L(Y, f(X)) = \int L(y, f(x)) d\mathbb{P}(x,y)$ denotes the expectation of L with respect to $\mathbb{P}$. We sometimes write $L(f)$ instead of $L(y, f(x))$ and $L(f+b)$ instead of $L(y, f(x)+b)$ to shorten the notation, if misunderstandings are unlikely. We use this kind of notation also for derivatives of L.

In this paper we are interested in binary classification, where $y_i \in Y := \{-1, +1\}$. The straightforward prediction rule is: predict $y_i = +1$ if $f(x_i) \geq 0$, and predict $y_i = -1$ otherwise. The loss function for the classification error is given by $I(y_i, f(x_i)) = \mathbb{I}(y_i f(x_i) < 0) + \mathbb{I}(f(x_i) = 0)\mathbb{I}(y_i = -1)$, where $\mathbb{I}$ denotes the indicator function. Inspired by the law of large numbers one might estimate $f_{\mathbb{P}}$ with the minimizer f_{emp} of the empirical classification error, that is

$$f_{emp} = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} I(y_i, f(x_i)) \;. \tag{2}$$

To avoid over-fitting one usually has to restrict the class of functions f considered in (2). Unfortunately, the classification function I is not convex and the minimization of (2) is often NP-hard, *cf.* Höffgen et al. (1995). To circumvent this problem, one can replace the classification error function $I(y_i, f(x_i))$ in (2) by a convex upper bound $L : Y \times \mathbb{R} \to \mathbb{R}$ *cf.* Vapnik (1998) and Schölkopf and Smola (2002). Furthermore, using reproducing kernel Hilbert spaces and an additional regularization term have some algorithmic advantages. These modifications lead to the following empirical regularized risks:

$$\hat{f}_{n,\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)), \tag{3}$$

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg \min_{f \in H,\, b \in \mathbb{R}} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i) + b), \tag{4}$$

where $\lambda > 0$ is a small regularization parameter, H is a reproducing kernel Hilbert space (RKHS) of a kernel k, and b is called *offset*. The decision functions are $\text{sign}(\hat{f}_{n,\lambda})$ or $\text{sign}(\hat{f}_{n,\lambda} + \hat{b}_{n,\lambda})$. Note that in practice usually (4) is solved while many theoretical papers deal with (3) since the unregularized offset b often causes technical difficulties in the analysis.

In practice the dual problems of (3) and (4) are solved. In these problems the RKHS does not occur explicitly, instead the corresponding kernel is involved. The choice of the kernel k enables

Method	L	L'
Support Vector Machine	$\max(1-v,0)$	-1, if $v<1$
		0, if $v>1$
Kernel Logistic Regression	$\ln(1+\exp(-v))$	$-1/(1+\exp(v))$
AdaBoost	$\exp(-v)$	$-\exp(-v)$
Least Squares	$(1-v)^2$	$2(v-1)$
Modified Least Squares	$\max(1-v,0)^2$	$-2\max(0,1-v)$
Modified Huber	$-4v$, if $v<-1$	-4, if $v<-1$
	$\max(1-v,0)^2$, else	$-2\max(0,1-v)$, otherwise

Table 1: Loss functions, where $v=yf(x)$ or $v=y[f(x)+b]$, respectively.

the above methods to efficiently estimate not only linear, but also non-linear functions. Of special importance is the Gaussian radial basis function (RBF) kernel

$$k(x,x')=\exp(-\gamma\|x-x'\|^2), \quad \gamma>0, \tag{5}$$

which is a universal kernel on every compact subset of $\mathbb{R}^d$ in the sense of Steinwart (2001). Furthermore, this kernel is a bounded kernel, as $|k(x,x')|\leq 1$ for all $x,x'\in\mathbb{R}^d$. Polynomial kernels $k(x,x')=(c+\langle x,x'\rangle)^m$ are also popular in practice, but are unbounded for $m\geq 1$ and $X=\mathbb{R}^d$.

Popular loss functions depend on y and f via $v=yf(x)$ or $v=y(f(x)+b)$. Some important specifications of L are given in Table 1. The support vector machine (SVM) penalizes points linearly if $v<1$. Kernel logistic regression and AdaBoost use twice continuously differentiable loss functions. The loss function used by kernel logistic regression (Wahba, 1999) penalizes misclassifications in a similar way to the SVM, *i.e.* approximately linearly if $v\to-\infty$. The loss function used by AdaBoost increases exponentially for $v\to-\infty$, *cf.* Freund and Schapire (1996), Friedman et al. (2000), and Hastie et al. (2001). The modified Huber's loss function, *cf.* Zhang (2004), changes the modified least squares loss such that misclassified points with $v<-1$ are penalized only linearly.

Two major benefits of using a convex loss function are known:

- For convex loss functions the resulting problems (3) and (4) are *computationally tractable* in the sense that they can be approximately solved in polynomial time. In fact, for many loss functions fast algorithms do exist. For bounded loss functions the convexity is lost and to our best knowledge almost nothing is known whether the problems are computational tractable. For applications of non-convex loss functions in the context of weighted least squares support vector machines for regression problems see Suykens et al. (2002).

- In the last two years an exciting observation almost revolutionized considerations on the learning performance of classification algorithms. The standard approach for bounding the estimation error of (regularized) empirical risk minimization (ERM) algorithms is to apply a uniform deviation bound. With this technique no learning rates faster than $n^{-\frac{1}{2}}$, where n is the sample size, can be obtained for nontrivial function classes and noisy distributions. However, if one "quantifies" the amount of noise (Tsybakov, 2004) and considers ERM-type algorithms with *convex* loss functions then learning rates up to n^{-1} are possible! For more information of this recent development we refer to Bartlett et al. (2002), Bartlett et al. (2003) and Bartlett and Mendelson (2002). In particular, this program has been successively applied to support vector machines by Scovel and Steinwart (2003). Learning rates for boosting methods are investigated by Blanchard et al. (2003).

These two benefits of convex loss functions are the major reasons why the convexity plays a very important role in recent machine learning algorithms. This is in contrast to robust statistics, where often non-convex loss functions are used, although such robust statistics are based on objective functions with more than one local optimum, *c.f.* Hampel et al. (1986) and Christmann (1994, 1998).

Problems (3) and (4) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk given in (6) or (7), respectively (Vapnik, 1998; Zhang, 2004; Steinwart, 2002a):

$$f_{\mathbb{P},\lambda} = \arg\min_{f\in H} \lambda \|f\|_H^2 + \mathbb{E}_{\mathbb{P}} L(Y, f(X)), \tag{6}$$

$$(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg\min_{f\in H,\, b\in\mathbb{R}} \lambda \|f\|_H^2 + \mathbb{E}_{\mathbb{P}} L(Y, f(X)+b). \tag{7}$$

The objective functions in (6) and (7) are denoted by $R^{reg}_{L,\mathbb{P},\lambda}(.)$ and $R^{reg}_{L,\mathbb{P},\lambda}(.,.)$ in the sequel.

Steinwart (2002b) shows that SVM's are universally consistent, *i.e.* the classification error of $\hat{f}_{n,\lambda}(.)$ converges to the optimal Bayes error $\mathbb{E}_{\mathbb{P}} I(Y, f_{\mathbb{P}}(X))$ in probability, provided that the reproducing kernel Hilbert space is dense in the space $C(X)$, $X \subset \mathbb{R}^d$ compact, and $\lambda = \lambda_n$ tends "slowly" to 0 for $n \to \infty$. Zhang (2004) improves this result by showing that for many convex loss functions the classifiers based on (3) are universally consistent if $\lambda_n \to 0$ and $\lambda_n n \to \infty$. Steinwart (2002a) characterizes the loss functions which lead to universally consistent classifiers and establishes universal consistency for classifiers based on (3) and (4). Furthermore, he shows that there exist solutions of both the theoretical and the empirical problems. Under certain assumptions on the data generating distribution one can even establish rates on the learning speed of SVM's. Such results can be found in Steinwart (2001), Chen et al. (2003), and Scovel and Steinwart (2003). Moreover, Steinwart (2003) gives lower asymptotical bounds on the number of support vectors, *i.e.* on the data points with non-vanishing coefficients, and investigates the asymptotic behavior of $\hat{f}_{n,\lambda}(.)$ in terms of the loss function L. As a by-product it also turns out that the solutions of (3) and (6) are unique. The same holds for the RKHS part of the solutions of (4) and (7). Upper bounds on the number of support vectors can be found in Steinwart (2004). Schölkopf and Smola (2002) describe other support vector machines and give an overview on algorithms to solve the minimization problems corresponding to SVMs.

3. Robustness

In the statistical literature different criteria have been proposed to define the notion of robustness in a mathematical way, *e.g.* the minimax approach (Huber, 1964), the sensitivity curve (Tukey, 1977), the approach based on influence functions (Hampel, 1974; Hampel et al., 1986), the maxbias curve (Huber, 1964; Hampel et al., 1986), and the finite sample breakdown point (Donoho and Huber, 1983).

In this paper, we mainly use the approach based on the influence function proposed by Hampel (1974) and Hampel et al. (1986). We will consider a map T which assigns to every distribution $\mathbb{P}$ on a given set Z an element $T(\mathbb{P})$ of a given Banach space E. In the case of the convex risk minimization methods given in (6) and (7) E equals the RKHS and $T(\mathbb{P}) = f_{\mathbb{P},\lambda}$ or $T(\mathbb{P}) = (f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$, respectively. The book by Huber (1981, p. 34ff) is a standard reference for Gâteaux and Fréchet derivatives in the context of robust statistics.

Definition 1 (Influence function) *The influence function of T at a point z for a distribution $\mathbb{P}$ is the special Gâteaux derivative (if it exists)*

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

where Δ_z is the Dirac distribution at the point z such that $\Delta_z(\{z\}) = 1$.

The influence function has the interpretation, that it measures the impact of an (infinitesimal) small amount of contamination of the original distribution $\mathbb{P}$ in direction of a Dirac distribution located in the point z on the theoretical quantity of interest $T(\mathbb{P})$. Therefore, in the robustness approach based on influence functions it is desirable that a statistical method which can be written as $T(\mathbb{P})$ has a *bounded* influence function. If T fulfills some regularity conditions, it can be linearized near $\mathbb{P}$ in terms of the influence function via

$$T(\mathbb{P}^*) = T(\mathbb{P}) + \int IF(z;T,\mathbb{P}) \, [\mathbb{P}^*(dz) - \mathbb{P}(dz)] + \dots ,$$

where $\mathbb{P}^*$ is a probability distribution near $\mathbb{P}$, *cf.* Huber (1981, p. 14). If different methods have a bounded influence function, the one with a lower bound is the more robust one.

The sensitivity curve SC_n proposed by J.W. Tukey and discussed in detail by Hampel et al. (1986, p. 93) can be interpreted as a finite sample version of the influence function, see (10). The sensitivity curve measures the impact of just one additional data point z on the empirical quantity of interest, *i.e.* on the estimate T_n.

Definition 2 (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*

$$SC_n(z;T_n) = n\big(T_n(z_1,\dots,z_{n-1},z) - T_{n-1}(z_1,\dots,z_{n-1})\big) \, . \tag{9}$$

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

$$SC_n(z;T_n) = \frac{T\big((1-\varepsilon_n)\mathbb{P}_{n-1} + \varepsilon_n \Delta_z\big) - T(\mathbb{P}_{n-1})}{\varepsilon_n} \, . \tag{10}$$

Of theoretical as well as of practical importance is also the notion of maxbias, which measures the maximum bias $T(Q) - T(\mathbb{P})$ within a neighborhood of probability distributions Q near $\mathbb{P}$. In robust statistics the so-called contamination (or gross-error) neighborhood (defined below), *cf.* Huber (1981), is quite common for the following three reasons. It allows a good interpretation because it contains mixture distributions with respect to the 'true' distribution $\mathbb{P}$ and some other distribution. The contamination neighborhood has some relationship to the influence function and to breakdown points, and finally it is often easier to deal with this set of distributions than with other neighborhoods. Note that the contamination neighborhood is not a neighborhood in the topological sense.

Definition 3 (Maxbias) *Let $\mathbb{P}$ be a fixed probability distribution on $X \times Y$. A contamination neighborhood of $\mathbb{P}$ is given by*

$$N_\varepsilon(\mathbb{P}) = \left\{ Q = (1-\varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}};\ \tilde{\mathbb{P}} \text{ is any distribution on } X \times Y,\ 0 \le \varepsilon < \frac{1}{2} \right\}.$$

The maxbias (or supremum bias) of T at the distribution $\mathbb{P}$ with respect to the contamination neighborhood $N_\varepsilon(\mathbb{P})$ is defined by

$$\operatorname{maxbias}(\varepsilon; T, \mathbb{P}) = \sup_{Q \in N_\varepsilon(\mathbb{P})} \| T(Q) - T(\mathbb{P}) \| .$$

4. Existence of the Influence Function

In this section we give sufficient conditions for the existence of the influence function for classifiers based on (6) and (7), whereas the next section will show that the influence function is bounded under weak conditions. Most of our results are valid for *any* distribution $\mathbb{P}$ on $X \times Y$. Therefore, they are also valid for the special case of the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \Delta_{(x_i,y_i)}$, *i.e.* for a given data set, and for the empirical regularized risks defined in (3) and (4).

Since our robustness results in this section are based on the calculus in (infinite dimensional) Banach spaces we first recall some basic notions; more details are given in the appendix. To this end let $G : E \to F$ be a map between two Banach spaces E and F. We say that G is *(Fréchet)-differentiable in* $x_0 \in E$ if there exists a bounded linear operator $A : E \to F$ and a function $\varphi : E \to F$ with $\frac{\varphi(x)}{\|x\|} \to 0$ for $x \to 0$ such that

$$G(x_0 + x) - G(x_0) \ = \ Ax + \varphi(x) \tag{11}$$

for all $x \in E$. It turns out that A is uniquely determined by (11). We hence 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.

We also have to introduce the notion of Bochner-integrals. For simplicity we restrict ourselves to the RKHS case, since this is the only one we actually need. To this end let H be a RKHS of a bounded, continuous kernel k on X with feature map $\Phi : X \to H$, i.e. $\Phi(x) = k(x, .)$. Furthermore, let $\mathbb{P}$ be a probability measure on $X \times Y$ and $h : Y \times X \to \mathbb{R}$ be a function which is continuous in its second variable $x \in X$. Then the Bochner-integral $\mathbb{E}_{\mathbb{P}} h(Y,X)\Phi(X)$ is an element of H which in our case can be computed by a simple Riemann approach, i.e. by partitioning the underlying space $X \times Y$. For a precise definition of Bochner-integrals we refer to Diestel and Uhl (1977). Note that in our special situation we can also interpret $\mathbb{E}_{\mathbb{P}} h(Y,X)\Phi(X)$ as an element of the dual space H' by the Fréchet-Riesz theorem, i.e. $\mathbb{E}_{\mathbb{P}} h(Y,X)\Phi(X)$ acts as a functional on H via $w \mapsto \langle \mathbb{E}_{\mathbb{P}} h(Y,X)\Phi(X), w \rangle$. Finally, we have to consider Bochner-integrals of the form $\mathbb{E}_{\mathbb{P}} h(Y,X)\langle \Phi(X), . \rangle \Phi(X)$ which define bounded linear operators on H by the map $w \mapsto \mathbb{E}_{\mathbb{P}} h(Y,X)\langle \Phi(X), w \rangle \Phi(X)$.

We can now establish our first two results which treat classifiers based on (6) with a smooth loss function and a bounded continuous kernel. The first theorem covers *e.g.* the Gaussian RBF kernel. The Dirac distribution in the point z is denoted by Δ_z.

Theorem 4 *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable loss function. Furthermore, let $X \subset \mathbb{R}^d$ be a closed or open subset, H be a RKHS of a bounded continuous kernel* *on X and $\mathbb{P}$ be a distribution on $X \times Y$. We define $G : \mathbb{R} \times H \to H$ by*

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

which implies

$$\frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}) = -\mathbb{E}_{\mathbb{P}}[L'(Y, f_{\mathbb{P},\lambda}(X))\Phi(X)] + L'(z_y, f_{\mathbb{P},\lambda}(z_x))\Phi(z_x).$$

Furthermore, we define $S : H \to H$ by

$$S := \frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda}) = 2\lambda \, \mathrm{id}_H + \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X)) \langle \Phi(X), . \rangle \Phi(X) \, .$$

Then the influence function of the classifiers based on (6) exists for all $z = (z_x, z_y) \in X \times Y$ and is given by

$$IF(z; T, \mathbb{P}) = -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}) \, . \tag{12}$$

Remark 5 *The influence function derived in Theorem 4 depends on the point $z = (z_x, z_y)$, where the point mass contamination takes place, only by the term $L'(z_y, f_{\mathbb{P},\lambda}(z_x))\Phi(z_x)$.*

In practice the set X is usually a bounded and closed subset of $\mathbb{R}^d$ and hence compact. In this case existence of the influence function can be shown without the assumption that the kernel is bounded, and hence the following theorem covers also polynomial kernels.

Theorem 6 *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable loss function. Furthermore, let $X \subset \mathbb{R}^d$ be compact, H be a RKHS of a continuous kernel on X and $\mathbb{P}$ be a distribution on $X \times Y$. Then the influence function of the classifiers based on (6) exists for all $z \in X \times Y$.*

Remark 7 *By a simple modification of the proof of the above theorem we actually find that the special Gâteaux derivative of $T : \mathbb{P} \mapsto f_{\mathbb{P},\lambda}$ exists for every direction,* i.e.

$$\lim_{\varepsilon \downarrow 0} \frac{f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}}{\varepsilon}$$

exists for all distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ on $X \times Y$ provided that the assumptions of Theorem 4 hold. This is an interesting result from the viewpoint of applied statistics, because a point mass contamination is just one possible kind of contamination which can occur in practice.

The following theorem shows the existence of the influence function for classifiers based on (7). Since the offset $b_{\mathbb{P},\lambda}$ can be infinite for certain loss functions if $\mathbb{P}$ is degenerate, i.e $\mathbb{P}(\{(y,x) : x \in X\}) = 1$ for $y = +1$ or $y = -1$, we have to exclude these probability measures. Note that these measures almost surely produce training sets of the form $((1,x_1), \ldots, (1,x_n))$, or $((-1,x_1), \ldots, (-1,x_n))$, respectively.

Theorem 8 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex and twice continuously differentiable loss function with $L'' > 0$. Furthermore, let $X \subset \mathbb{R}^d$ be open or closed, H be a RKHS of a continuous kernel on X and $\mathbb{P}$ be a non-degenerate distribution on $X \times Y$. We define $G : \mathbb{R} \times H \times \mathbb{R} \to H \times \mathbb{R}$ by*

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

which implies

$$\frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = -\mathbb{E}_{\mathbb{P}}[L'(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda})\Phi(X)] + L'(z_y, f_{\mathbb{P},\lambda}(z_x) + b_{\mathbb{P},\lambda})\Phi(z_x)\,.$$

Furthermore, for $S := \frac{\partial G}{\partial (H\times\mathbb{R})}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ we have

$$S = \begin{pmatrix} 2\lambda\,\mathrm{id}_H + \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda})\langle \Phi(X), .\rangle \Phi(X) & \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda})\Phi(X) \\ \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda})\Phi(X) & \mathbb{E}_{\mathbb{P}} L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda}) \end{pmatrix}.$$

Then the influence function of the classifiers based on (7) exists for all $z = (z_x, z_y) \in X \times Y$ and is given by

$$IF(z; T, \mathbb{P}) = -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})\,. \tag{13}$$

Remark 9 *The influence function derived in Theorem 8 depends on the point $z = (z_x, z_y)$, where the point mass contamination takes place, only by the term $L'(z_y, f_{\mathbb{P},\lambda}(z_x) + b_{\mathbb{P},\lambda})\Phi(z_x)$. Hence loss functions L and kernels k such that L' and the feature map Φ are bounded are of special interest from the view point of robust statistics.*

Remark 10 *As in the case of problem (6) a slight modification of the proof gives that $T : \mathbb{P} \mapsto (f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ is special Gâteaux differentiable.*

Remark 11 *Considering the loss functions in Table 1 we immediately see that the above theorems apply to the kernel logistic regression, the least squares and the AdaBoost loss function. The second derivatives of the modified least squares and the modified Huber loss function fail to exists in only one point. For the loss function of the standard SVM, even the first derivative does not exist in one point.*

5. Bounds on the Influence Function, Sensitivity Curve and Maxbias

As mentioned in Section 1, a desirable property of a robust statistical method is that T has a bounded influence function. In this section we show that for certain loss functions the influence function can be bounded *independently* of z and $\mathbb{P}$ for classifiers based on (6) and (7). For the formulation of our results we need to recall that the norm of total variation of a signed measure μ on a space X is defined by

$$\|\mu\|_{\mathcal{M}} := |\mu|(X) := \sup\Big\{\sum_{i=1}^{n} |\mu(A_i)| : A_1, \ldots, A_n \text{ is a partition of } X\Big\}.$$

For more information on this norm we refer to Brown and Pearcy (1977).

Our first result bounds the difference quotient in the definition of the influence function for classifiers based on (6). For practical applications Theorem 12 is our most important result. In particular, it states that the influence function of these classifiers is uniformly bounded whenever it exists, and that the sensitivity curve is uniformly bounded too. Please note that the following theorem based on Steinwart (2003) applies to all six loss functions given in Table 1 because differentiability of L is not assumed.

Theorem 12 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a continuous and convex loss function. Furthermore, let $X \subset \mathbb{R}^d$ and H be a RKHS of a bounded, continuous kernel on X. Then for all $\lambda > 0$ there exists a constant $c_L(\lambda) > 0$ explicitly given in (27) such that for all distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ on $X \times Y$ we have*

$$\left\| \frac{f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}}{\varepsilon} \right\|_H \leq c_L(\lambda) \; \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}, \quad \varepsilon > 0\,.$$

Remark 13 *The above theorem also gives uniform bounds for Tukey's sensitivity curve of f. Consider the special case that $\mathbb{P}$ is equal to the empirical probability measure of $(n-1)$ data points,* i.e. $\mathbb{P}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} \Delta_{(x_i,y_i)}$*, and that $\tilde{\mathbb{P}}$ is equal to the Dirac measure $\Delta_{(x,y)}$ in some point $(x,y) \in X \times Y$. Let $\varepsilon = \frac{1}{n}$. Under the assumptions of Theorem 12 it follows from (10), that*

$$n \|f_{(1-\varepsilon)\mathbb{P}_{n-1}+\varepsilon\Delta_{(x,y)},\lambda} - f_{\mathbb{P}_{n-1},\lambda}\|_H \leq c_L(\lambda) \; \|\mathbb{P}_{n-1} - \Delta_{(x,y)}\|_{\mathcal{M}}, \quad \varepsilon > 0\,.$$

Essentially, this result has already been established by Bousquet and Elisseeff (2002).

Remark 14 *Because no assumptions on $\tilde{\mathbb{P}}$ are made in Theorem 12, an upper bound for the maxbias of $f_{\mathbb{P},\lambda}$ (see Definition 3) for such machine learning methods is given by*

$$\text{maxbias}(\varepsilon; f, \mathbb{P}) = \sup_{Q \in N_\varepsilon} \|f_{Q,\lambda} - f_{\mathbb{P},\lambda}\| \leq \varepsilon\, c_L(\lambda) \sup_{\tilde{\mathbb{P}} \in \mathcal{P}} \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}} \leq 2c_L(\lambda)\varepsilon\,,$$

where $\varepsilon \in (0, 1/2)$, and $\mathcal{P}$ denotes the set of all probability measures on $X \times Y$. As no assumptions are made for $\mathbb{P}$, this result is valid for empirical distributions too. Consider the empirical distribution $\mathbb{P}_n$ defined by given data set $(x_i, y_i) \in X \times Y$ with n data points. Then the maxbias of $f_{\mathbb{P}_n,\lambda}$ in a contamination neighborhood $N_\varepsilon(\mathbb{P}_n)$ is at most $2c_L(\lambda)\varepsilon$, where $\varepsilon \in (0, 1/2)$.

Unfortunately, using the estimate of Steinwart (2003) does not give any meaningful result for classifiers based on (7). However, under some additional assumptions on L we can still bound the influence function.

Theorem 15 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex and twice continuously differentiable loss function with $a \leq L'' \leq b$ for some $a, b > 0$. Furthermore, let $X \subset \mathbb{R}^d$ be open or closed, H be a RKHS of a continuous kernel on X, and $T_\lambda(\mathbb{P}) = (f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ be given by (7). Then for all $\lambda > 0$ there exists a constant $c_L(\lambda) > 0$ such that for all non-degenerated distributions $\mathbb{P}$ on $X \times Y$ and all $z \in X \times Y$ we have*

$$\|IF(z; T, \mathbb{P})\|_{H \times \mathbb{R}} \leq c_L(\lambda) \; \|\mathbb{P} - \Delta_z\|_{\mathcal{M}}\,.$$

Remark 16 *Theorem 15 applies to (7) with the least squares loss function. However, Theorem 15 covers neither the logistic regression loss function as we only have $L'' \geq 0$ nor the AdaBoost loss function which satisfies $L'' = L = \exp(-\,.)$ However, we get the same bound of the influence function if we restrict our considerations to distributions $\mathbb{P}$ with*

$$a \;\leq\; \int L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda})\, d\mathbb{P} \;\leq\; b \tag{14}$$

for some $b \geq a > 0$. A simple sufficient condition for the latter can be derived by the proof of Steinwart (2002a, Lemma II.6): let $A_y^\rho := \{x \in X : \mathbb{P}(y|x) > \rho\}$, $y \in Y$, $\rho > 0$, and $\alpha_{\mathbb{P}}(\rho) := \rho \min\{\mathbb{P}_X(A_1^\rho), \mathbb{P}_X(A_{-1}^\rho)\}$. Then fixing $\lambda > 0$, a twice continuously differentiable L and a threshold $\alpha > 0$ there exist $b \geq a > 0$ such that every $\mathbb{P}$ with $\alpha_{\mathbb{P}}(\rho) \geq \alpha$ for some $\rho > 0$ satisfies (14). Note that the assumption $\alpha_{\mathbb{P}}(\rho) \geq \alpha$ guarantees that the two classes of $\mathbb{P}$ are "balanced".

Remark 17 *As mentioned in Remark 10 the map $T : \mathbb{P} \mapsto (f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ is special Gâteaux differentiable. A simple modification of the proof of Theorem 15 shows that the special Gâteaux derivative of T can be uniformly bounded.*

Remark 18 *Consider the case that $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are probability measures with densities p and $\tilde{p}$ with respect to some dominating measure ν. Then, the last two theorems also give bounds of the influence functions and the sensitivity curve in terms of the Hellinger metric $H(\mathbb{P}, \tilde{\mathbb{P}}) = [\int (\sqrt{p} - \sqrt{\tilde{p}})^2 \, d\nu]^{1/2}$. This follows from a relationship between the norm of total variation and the Hellinger metric:*

$$\|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}} \leq 2H(\mathbb{P}, \tilde{\mathbb{P}}) \leq 2\,\|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}^{1/2} .$$

Note that the bounds for the difference quotient in Theorem 12 and for the influence function in Theorem 15 converge to infinity, if λ converges to 0 and $\|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}} > 0$ or $\|\mathbb{P} - \Delta_z\|_{\mathcal{M}} > 0$. However, λ converging to 0 has the interpretation that misclassifications are penalized by constants C tending to ∞. Therefore, decreasing values of λ correspond to a decreasing amount of robustness, which was to be expected. The quantity λ can be interpreted as a tuning constant controlling the robustness properties of the method in a similar way than it is well-known for many robust methods, *e.g.* Huber-type M-estimators in location or regression models. Consider the Huber-type M-estimator (Huber, 1964) in a univariate location model, where all data points are realizations from n independent and identically distributed random variables with some distribution function $F(\cdot - \theta)$, where $\theta \in \mathbb{R}$ is unknown. Huber's robust M-estimator with tuning constant $b \in (0, \infty)$ has an influence function proportional to $\psi_b(z) = \min\{b, \max\{z - b\}\}$, *cf.* Hampel et al. (1986, p. 104f). For all $b \in (0, \infty)$ the influence function is bounded by $\pm b$. However, the bounds tend to $\pm\infty$ if $b \to \infty$, and Huber's M-estimator with $b = \infty$ is equal to the non-robust maximum likelihood estimator which has an unbounded influence function. Therefore, the quantity $1/\lambda$ (or the cost C) in the machine learning methods we are dealing with has a similar role to the tuning constant b in Huber-type M-estimators.

6. Empirical Results for the SVM

In this section we study the impact an additional data point can have on the SVM with offset b for pattern recognition. An analogous investigation for the case without offset gave similar results to those described in this section. We generated a training data set with $n = 500$ data points x_i from a bivariate normal distribution with expectation $\mu = (0, 0)$ and covariance matrix Σ. The variances were set to 1, whereas the covariance was set to 0.5. The responses y_i were generated from a classical logistic regression model with $\theta = (-1, 1)'$, $b = 0.5$, such that $P(Y_i = +1) = [1 + \exp(-(x_i'\theta + b))]^{-1}$ and $P(Y_i = -1) = 1 - P(Y_i = +1)$. The computations were done using the software SVM^{light} developed by Joachims (1999). SVM^{light} solves the dual program corresponding to the primal optimization problem

$$\begin{array}{ll} \arg\min_{f \in H, b \in \mathbb{R}} & \frac{1}{2Cn} \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \\ \text{such that} & y_i(f(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 . \end{array} \tag{15}$$

We consider two popular kernels: a Gaussian radial basis function kernel with parameter γ, see (5) and a linear kernel. Appropriate values for γ and for the constant C (or λ) are important for the SVM and are often determined by cross validation, *cf.* Schölkopf and Smola (2002, p. 217). A cross validation based on the leave-one-out error for the training data set was carried out by a two-dimensional grid search on

$$\gamma \in \{0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 10, 20\}$$

and

$$C \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 5, 10, 20\}.$$

As a result of the cross validation, the tuning parameters for the SVM with RBF kernel were set to $\gamma = 0.25$ and $C = 2$. The leave-one-out error for the SVM with a linear kernel turned out to be stable over a broad range of values for C. We used $C = 1$ in the computations for the linear kernel. For $n = 500$ this results in $\lambda = (2Cn)^{-1} = 5 \times 10^{-4}$ for the RFB kernel and $\lambda = (2Cn)^{-1} = 0.001$ for the linear kernel. Please note that such small values of λ will result in relatively large bounds.

Figure 1 shows the sensitivity curves of $\hat{f} + \hat{b} := \hat{f}_{n,\lambda} + \hat{b}_{n,\lambda}$, if we add a single point $z = (x, y)$ to the original data set, where $x_1 = 6$, $x_2 = 6$, and $y = +1$. The additional data point has a local and smooth impact on $\hat{f} + \hat{b}$ with a peak in a neighborhood of (x_1, x_2), if one uses the RBF kernel. For a linear kernel, the impact is approximately linear. The reason for this different behavior of the SVM with different kernels becomes clear from Figure 2 where plots of $\hat{f} + \hat{b}$ are given for the original data set and for the modified data set, which contains the additional data point z. Please note that the RBF kernel yields $\hat{f} + \hat{b}$ approximately equal to zero outside a central region, as almost all data points are lying inside the central region. Comparing the plots of $\hat{f} + \hat{b}$ based on the RBF kernel for the modified data set with the corresponding plot for the original data set, it is obvious that the additional smooth peak is due to the new data point located at $x = (6, 6)$ with $y = +1$. It is interesting to note that although the estimated functions $\hat{f} + \hat{b}$ for the original data set and for the modified data set based on the SVM with the linear kernel are looking quite similar, the sensitivity curve is similar to an affine hyperplane which is affected by the value of z. This allows the interpretation, that just a single data point can have an impact on $\hat{f} + \hat{b}$ estimated by a SVM with a linear kernel over a broader region than for an SVM with an RBF kernel.

Now, we study the impact of an additional data point $z = (x, y)$, where $y = +1$, on the percent of classification errors and on the fitted $y-$value for z. We vary z over a grid in the $x-$coordinates. Figure 3 shows that the percentage of classification errors is approximately constant outside the central region that contains almost all data points if a Gaussian RBF kernel was used. For the SVM with a linear kernel, the percentage of classification errors tends to be approximately constant in one half-space but changes in the other half-space. The response of the additional data point was correctly estimated by $\hat{y} = +1$ outside the central region, if a RBF kernel is used, see Figure 4. In contrast to that, using a linear kernel results in estimated responses $\hat{y} = +1$ or $\hat{y} = -1$ of the additional data point depending on the affine half-space in which the $x-$value of z is lying.
Finally, let us study the impact of an additional data point located at $z = (x, y)$, where $y = +1$, on the estimated parameters $\hat{b}$ and $\hat{\theta}$, see Figure 5. We vary z over a grid in the $x-$coordinates in the same manner as before. As the plots for $\hat{\theta}_1$ and $\hat{\theta}_2$ are looking very similar, we only show the latter. Note that the axes are not identical in Figure 5 due to the kernels. The sensitivity curves for the slopes estimated by the SVM with an RBF kernel are similar to a hyperplane outside the central region, which contains almost all data points. In the central region, there is a smooth transition between

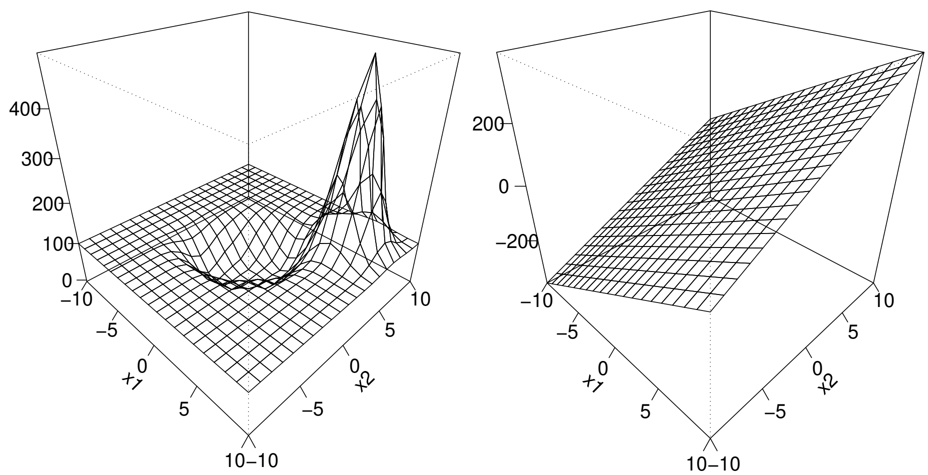

Figure 1: Sensitivity function of $\hat{f}+\hat{b}$, if the additional data point z is located at $z=(x,y)$, where $x=(6,6)$ and $y=+1$. Left: RBF kernel. Right: linear kernel.

regions with higher sensitivity values and regions with lower sensitivity values. The sensitivity curves for the slopes of the SVM with a linear kernel are flat in one affine half-space, but change approximately linearly in the other affine half-space. This behavior also occurs for the sensitivity curve of the offset by using a linear kernel. In contrast to that, the sensitivity curve of the offset based on a SVM with a RBF kernel shows a smooth but curved shape outside the region containing the majority of the data points.

7. Concluding Remarks

In this paper, we used the influence function approach of robust statistics (Hampel et al., 1986) for recent statistical learning methods based on convex risk minimization methods for the problem of pattern recognition. The influence function has the interpretation that it measures the impact of an infinitesimal amount of contamination of the original distribution $\mathbb{P}$ in direction of a Dirac distribution located in the point z on the theoretical quantity of interest $T(\mathbb{P})$. Special cases of such convex risk minimization methods are the support vector machine, kernel logistic regression, AdaBoost, and least squares. Assumptions were derived for the existence of the influence function of f or (f,b) used by the classifiers and also for uniform bounds on the influence function which hold with respect to the distribution $\mathbb{P}$ and the point z of the Dirac distribution Δ_z describing the contamination. For the case without offset b one can uniformly bound the difference quotient considered by the influence function under weak conditions which also yields uniform bounds for Tukey's sensitivity curve and uniform upper bounds for the maxbias. In particular, the influence function for these classifiers is uniformly bounded if it exists. Some of the results are not limited to the special Gâteaux derivative used in the definition of the influence function. The assumptions of some of our results exclude the support vector machine because the SVM uses a loss function which is not differentiable in one point, but Theorem 12 covers the SVM as a special case. We gave some numerical results for

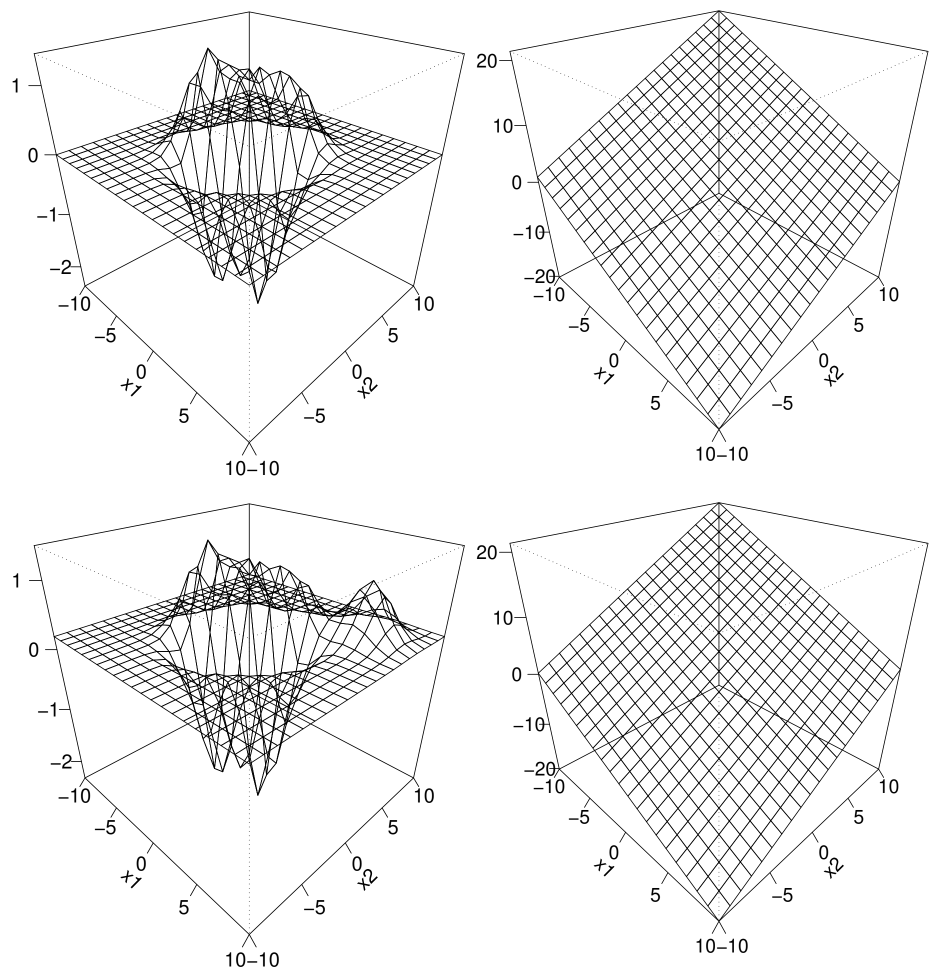

Figure 2: Plot of $\hat{f}+\hat{b}$. Upper left: RBF kernel, original data set. Upper right: linear kernel, original data set. Lower left: RFB kernel, modified data set. Lower right: linear kernel, modified data set. The modified data set contains the additional data point $z=(x,y)$, where $x=(6,6)$ and $y=+1$.

the sensitivity curve, which can be interpreted as a finite sample version of the influence function, of the SVM classifier. It turned out, that the popular exponential radial basis function kernel resulted in smooth sensitivity curves for $\hat{f}+\hat{b}$ and for the estimated coefficients $(\hat{\theta},\hat{b})$. Varying the position of one additional data point had a smooth and local impact on $\hat{f}+\hat{b}$, if one uses an RBF kernel. For the linear kernel the impact of varying one additional data point behaves also in a relatively smooth manner, but the impact seems to be more globally than locally.

For a numerical comparison between the support vector machine and the regression depth method recently proposed by Rousseeuw and Hubert (1999) see Christmann and Rousseeuw (2001) and Christmann et al. (2002).

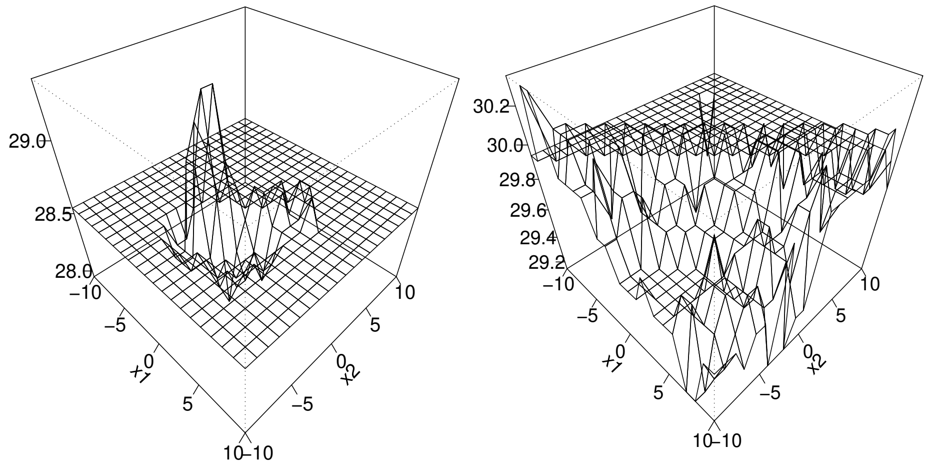

Figure 3: Percent of classification errors if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

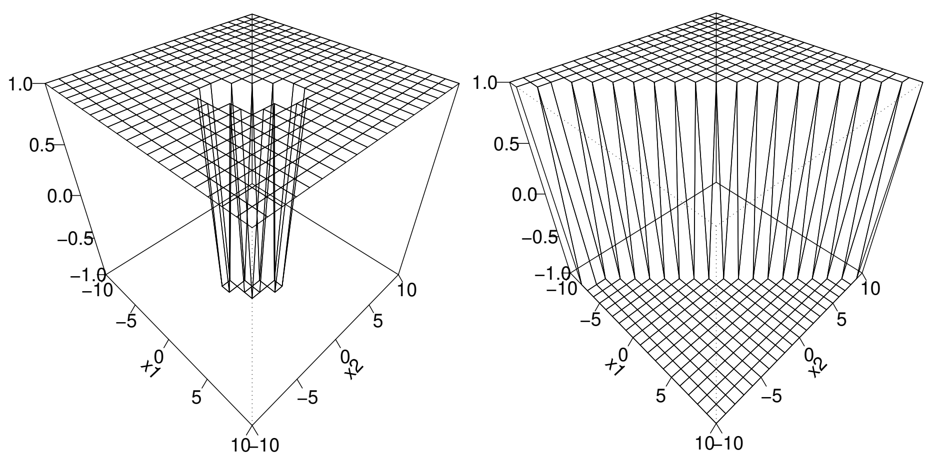

Figure 4: Fitted $y-$value for new observation if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

It would be interesting to study the influence function of convex risk minimization methods for other problems, *e.g.* $\varepsilon-$regression or kernel principal component analysis, or to consider other robustness concepts, but this is beyond the scope of this paper.

Acknowledgments

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and of DoMuS (University of Dortmund, "Model building and

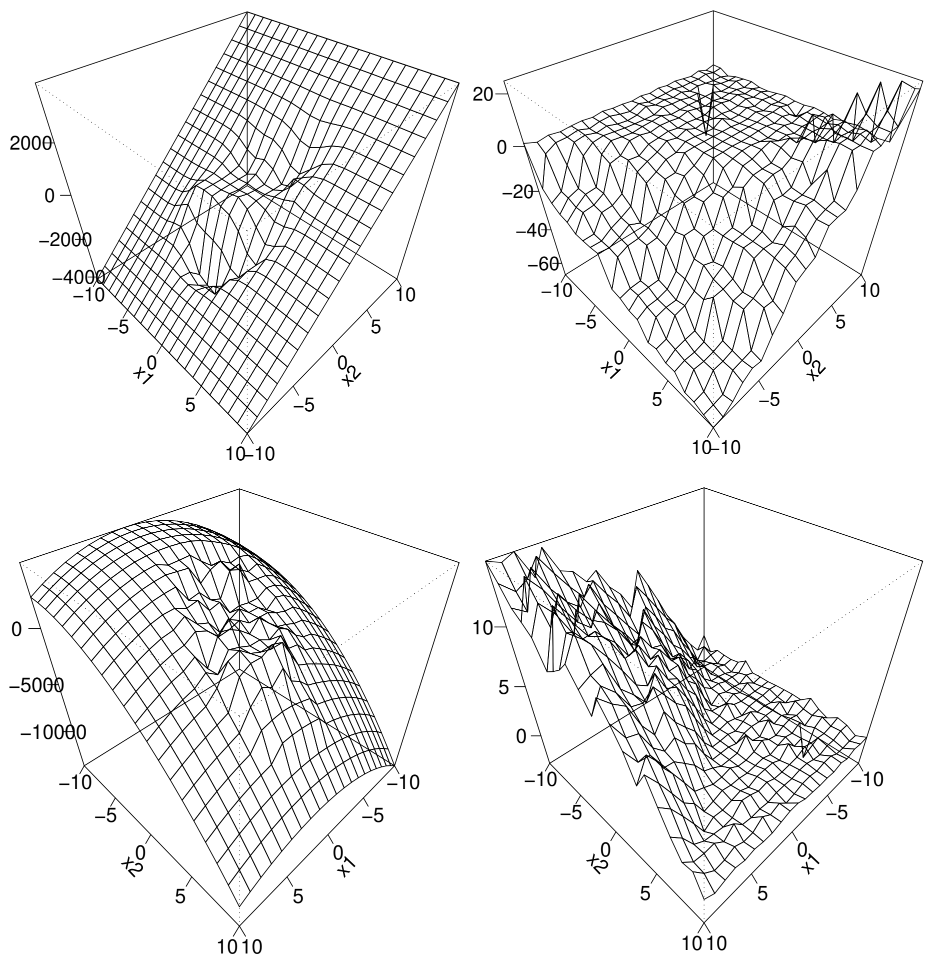

Figure 5: Sensitivity function for $\hat{\theta}$ and $\hat{b}$, respectively. Upper left: Sensitivity function for $\hat{\theta}_2$, RBF kernel. Upper right: Sensitivity function for $\hat{\theta}_2$, linear kernel. Lower left: Sensitivity function for $\hat{b}$, RBF kernel. Lower right: Sensitivity function for $\hat{b}$, linear kernel.

simulation") are gratefully acknowledged. We thank three anonymous referees and the editor for helpful remarks on an earlier version of the paper.

Appendix A. Mathematical Background

In Appendix A we list some facts from functional analysis, which are used in Appendix B to prove our theorems.

Since our proofs are heavily based on the calculus in (infinite dimensional) Banach spaces we first recall the basic facts, see *e.g.* Akerkar (1999), Brown and Pearcy (1977), and Yosida (1974). To this end let $G : E \to F$ be a map between two Banach spaces E and F. Recall, that we say that G is (Fréchet)-differentiable in $x_0 \in E$ if there exists a bounded linear operator $A : E \to F$ and a function $\varphi : E \to F$ with $\frac{\varphi(x)}{\|x\|} \to 0$ for $x \to 0$ such that

$$G(x_0 + x) - G(x_0) \;=\; Ax + \varphi(x) \tag{16}$$

for all $x \in E$. It turns out that A is uniquely determined by (16). As in Section 4 we hence write $G'(x) := \frac{\partial G}{\partial E}(x) := A$. Again, 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.

Unlike 1-dimensional derivates general Fréchet derivates suffer from some notational difficulties. For example, the derivate $G'(x)$ itself is a *map* for every x and thus $G'(x)$ is described by $y \mapsto G'(x)y$. Furthermore, considering partial derivates can cause notational problems too. Indeed, if *e.g.* id_E is the identity of E we have

$$x \;=\; \mathrm{id}'(x)x \;=\; \frac{\partial \mathrm{id}}{\partial E}(x)x \;=\; \frac{\partial \mathrm{id}(x)}{\partial x}x$$

where the right expression uses standard notation. We feel that the latter can cause problems for the unexperienced reader.

As in the finite dimensional case the differential operator satisfies basic calculus, that is linearity and a chain rule

$$\big(G_2 \circ G_1\big)'(x) \;=\; G_2'(G_1(x)) \circ G_1'(x)$$

for $G_1 : E_1 \to E_2$, $G_2 : E_2 \to E_3$ whenever all derivates exist in the above equation. Furthermore, for a bounded linear map $A : E \to F$ we have $A'(x) = A$ for all $x \in E$. If $G(f) := \|f\|_H^2$ for all elements f of a Hilbert space H we find $G' = 2\mathrm{id}_H$, where id_H denotes the identity on H.

Let us consider an example that helps to understand the differentiation steps in the following proofs. To this end let $\mathbb{P}$ be a probability measure on a subset $X \subset \mathbb{R}^d$ and H be a RKHS of bounded continuous functions over X with feature map $\Phi : X \to H$, *i.e.* $\Phi(x) := k(x, .)$, where k is the kernel of H. We consider the map $G : H \to \mathbb{R}$ which is defined by $Gf := \mathbb{E}_{\mathbb{P}} L \circ f$ for all $f \in H$ and a twice continuously differentiable function $L : \mathbb{R} \to \mathbb{R}$. In order to compute the derivate of this map we decompose G into $G = B \circ A \circ I$, where $I : H \to C_b(X)$ is the canonical embbeding $w \mapsto \langle w, \Phi(.)\rangle$ of H into the space of all bounded continuous functions $C_b(X)$, $A : C_b(X) \to C_b(X)$ is defined by $f \mapsto L \circ f$ and $B : C_b(X) \to \mathbb{R}$ is the functional $f \mapsto \mathbb{E}_{\mathbb{P}} f$. For the chain rule we need to compute the derivatives of these factors. Since I is linear we have $I'(v)w = Iw$ for all $v, w \in H$. Analogously, the linearity of B gives $B'(f)g = \mathbb{E}_{\mathbb{P}} g$ for all $f, g \in C_b(X)$. As shown in the book of Akerkar (1999), we also have $A'(f)g = g \cdot (L' \circ f)$ for all $f, g \in C_b(X)$, *i.e.* A' is the multiplication operator with respect to $L' \circ f$. Applying the chain rule to $A \circ I$ we hence find

$$(A \circ I)'(v)w \;=\; (A'(I(v)) \circ I'(v))w \;=\; L' \circ (Iv) \cdot (Iw)$$

for all $v, w \in H$. As we see the brackets play an important role for the mechanic evaluation of these derivatives. Another application of the chain rule gives

$$\begin{aligned} G'(v)w \;=\; \big(B'(A \circ I(v)) \circ (A \circ I)'(v)\big)w \;&=\; B'(A \circ I(v))\big((A \circ I)'(v)w\big) \\ &=\; \mathbb{E}_{\mathbb{P}}(A \circ I)'(v)w \\ &=\; \mathbb{E}_{\mathbb{P}} L' \circ (Iv) \cdot (Iw) \end{aligned}$$

for all $v, w \in H$. Note that by definition $G'(v) : H \to \mathbb{R}$ is a bounded functional. By the theorem of Fréchet-Riesz such functionals can be represented by elements of H via the mapping $v \mapsto \langle v, .\rangle$. In our situation we can directly compute this representation: for $v, w \in H$ we have

$$\langle \mathbb{E}_{\mathbb{P}} L' \circ (Iv)\Phi, w\rangle \;=\; \mathbb{E}_{\mathbb{P}} L' \circ (Iv)\langle Iv, w\rangle \;=\; \mathbb{E}_{\mathbb{P}} L' \circ (Iv) \cdot (Iw)\,,$$

i.e. $\mathbb{E}_{\mathbb{P}} L' \circ (Iv)\Phi$ is a representation of $G'(v)$. Note that $\mathbb{E}_{\mathbb{P}} L' \circ (Iv)\Phi$ is a H-valued *Bochner-integral*. For finite dimensional spaces $\mathbb{R}^n$ the $\mathbb{R}^n$-valued Bochner-integral can be computed by the integrals of the n components. In general Banach spaces some problems can occur by different notions of measurability. Since in our case H is separable by the continuity of Φ and all our functions are continuous we do not have these difficulties. In fact for compact X our Bochner-integrals can be even computed using a simple Riemann approach. For more information about Bochner-integrals we refer to Diestel and Uhl (1977) or Yosida (1974).

Since in our proofs we also have to compute the second derivate of maps of the form of G, let us now treat G''. For convenience we use the Fréchet-Riesz representation of G'. Analogously to the above considerations we decompose G'. To this end $B : C_b(X) \to H$ denotes the operator defined by $Bf := \mathbb{E}_{\mathbb{P}} f\Phi$ for all $f \in C_b(X)$. Furthermore, $A : C_b(X) \to C_b(X)$ is the operator $Af := L' \circ f$, $f \in C_b(X)$. Using the Fréchet-Riesz representation of G' we then find $G' = B \circ A \circ I$. Now observe that B is linear. Therefore we have $B'(f)g = \mathbb{E}_{\mathbb{P}} g\Phi$ for all $g \in C_b(X)$. Moreover, we find $(A\circ I)'(v)w = L'' \circ (Iv)Iw$ for all $v, w \in H$ as above. Using the chain rule this gives

$$G''(v)w \;=\; B''(A\circ I(v))\big((A\circ I)'(v)w\big) \;=\; \mathbb{E}_{\mathbb{P}}(A\circ I)'(v)w\Phi \;=\; \mathbb{E}_{\mathbb{P}} L'' \circ (Iv)Iw\Phi$$

for all $v, w \in H$.

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

Theorem 19 (Implicit function theorem) *Let E, F be Banach spaces and $G : E \times F \to F$ 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 $f : x_0 + \delta B_E \to y_0 + \delta B_F$ such that for all $x \in x_0 + \delta B_E$, $y \in y_0 + \delta B_F$ we have*

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

Moreover, the derivative of f is given by

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

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, (Cheney, 2001) turns out to be very helpful:

Theorem 20 (Fredholm Alternative) *Let E be a Banach space and $S : E \to E$ be a compact operator. Then $\mathrm{id}_E + S$ is surjective if and only if it is injective.*

We also need the Krein-Milman theorem, see Yosida (1974, p. 363) or Brown and Pearcy (1977, p. 309).

Theorem 21 (Krein-Milman theorem) *Let K be a non-void compact convex subset of a locally convex real linear topological space. Then K is equal to the closure of the convex hull of the set of all extreme points of K.*

Appendix B. Proofs of the Theorems

In this appendix we prove the theorems from Section 4 and Section 5. We sometimes write $L(f)$ instead of $L(y, f(x))$ and $L(f+b)$ instead of $L(y, f(x)+b)$ to shorten the notation, if misunderstandings are unlikely. We use this kind of notation also for derivatives of L.

PROOF OF THEOREM 4. Let us first check that the solution $f_{\mathbb{P},\lambda}$ exists in our situation. Indeed, in the proof of the existence statement by Steinwart (2002a) the compactness of X is only used to ensure $K := \|I : H \to \ell_\infty(X)\| < \infty$, where $\ell_\infty(X)$ denotes the space of all bounded functions $f : X \to \mathbb{R}$ equipped with the supremum norm and I is the canonical embbeding. The finiteness of this norm, however, *characterizes* bounded kernels. Therefore, the existence statement by Steinwart (2002a) is true in our case too.

Now, let $\Phi : X \to H$ be the feature map of H as in the above example. Our analysis heavily rely on the map $G : \mathbb{R} \times H \to H$ that is defined by

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

Note that for $\varepsilon \notin [0,1]$ the H-valued expectation is with respect to a signed measure. For these measures we refer to Dudley (2002). Now as in the above example, for $\varepsilon \in [0,1]$ we obtain

$$G(\varepsilon, f) = \frac{\partial R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}}{\partial H}(f)\,. \tag{17}$$

Since $f \mapsto R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}(f)$ is convex for all $\varepsilon \in [0,1]$ Equation (17) shows that we have $G(\varepsilon, f) = 0$ if and only if $f = f_{(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}$ for such ε. 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

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

For the existence of $\varepsilon \mapsto f_\varepsilon$ we only have to check by Theorem 19 that G is continuously differentiable and that $\frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda})$ is invertible. Let us start with the first: an easy computation shows

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathbb{P}} L'(Y, f(X))\Phi(X) + \mathbb{E}_{\Delta_z} L'(Y, f(X))\Phi(X)\,. \tag{18}$$

Moreover, as in the above example we find

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda\,\mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z} L''(Y, f(X))\langle \Phi(X), .\rangle \Phi(X)\,. \tag{19}$$

Since H has a bounded kernel it is a simple routine to check that both partial derivatives are continuous. This together with the continuity of G ensures 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

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

defines a compact operator on H. To show the compactness we have to recall some measure theory, see Dudley (2002). Since X is assumed to be open or closed, it is a Polish space. 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 $X_n \subset X$ with $\mathbb{P}_X(X_n) \geq 1 - 1/n$, where $\mathbb{P}_X$ denotes the marginal distribution of $\mathbb{P}$ with respect to X. We define a sequence of operators $A_n : H \to H$ by

$$A_n g := \int_{X_n \times Y} L''(y, f_{\mathbb{P},\lambda}(x)) g(x) \Phi(x) d\mathbb{P}(x,y)$$

for all $g \in H$. Let us now show that all A_n are compact. By the definition of A_n there exists a constant $c > 0$ depending on λ, L'' and K such that for all $g \in B_H$ we have

$$A_n g \in c \cdot \overline{\operatorname{aco} \Phi(X_n)}\,, \tag{20}$$

where $\operatorname{aco}\Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$. Indeed, for discrete probability measures $\mathbb{P}$ relation (20) follows directly from the definition using the fact $\|g\|_\infty \leq K$ for $g \in B_H$. To see the general case recall that by Krein-Milman's theorem the set of discrete probability measures is weak*-dense in the set of probability measures. Then (20) follows since H has the approximation property (Lindenstrauss and Tzafriri, 1977). Furthermore, since Φ is continuous $\Phi(X_n)$ is compact and hence so is $\overline{\operatorname{aco}\Phi(X_n)}$. This shows that A_n is compact. In order to see that A is compact, it therefore suffices to show $\|A_n - A\| \to 0$ for $n \to \infty$. The latter convergence can be easily checked using $\mathbb{P}_X(X_n) \geq 1 - 1/n$.

It remains to prove that A is injective. To this end for $g \neq 0$ we find

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

Here, the last equation is due to the fact that $B\mathbb{E}_{\mathbb{P}} h = \mathbb{E}_{\mathbb{P}} Bh$ for all E-valued functions h and all bounded linear operators $B : E \to F$ between Banach spaces E and F. A much stronger result is given in Diestel and Uhl (1977, p. 47). The last inequality is true since the second derivative of a convex function is always nonnegative. Obviously, the above estimate shows that $\frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda}) = 2\lambda \operatorname{id}_H + A$ is injective.

We use $IF(z; T, \mathbb{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0)$ to derive a formula for the influence function, where $\varepsilon \mapsto f_\varepsilon$ is the function implicitly defined by $G(\varepsilon, f) = 0$. The implicit function theorem hence gives

$$IF(z; T, \mathbb{P}) = -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda})\,, \tag{21}$$

where $S := \frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda})$. ■

PROOF OF THEOREM 6. Since every compact subset of $\mathbb{R}^d$ is closed and continuous kernels on compact subsets are bounded the assertion directly follows from Theorem 4. ■

PROOF OF THEOREM 8. The proof is similar to that of Theorem 4. However, due to the extra variable b we have to adapt our approach. As in the proof of Theorem 4 we first point out that the solutions $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) \in H \times \mathbb{R}$ exist. Again, this can be seen by a slight modification of the argument used by Steinwart (2002a). Now, let us define the map $G : \mathbb{R} \times H \times \mathbb{R} \to H \times \mathbb{R}$ by

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

Again, for $\varepsilon \in [0,1]$ the definition of G ensures

$$G(\varepsilon, f, b) = \frac{\partial R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}}{\partial(H \times \mathbb{R})}(f)\ ,$$

if we apply the Fréchet-Riesz identification $(H \times \mathbb{R})' = H \times \mathbb{R}$. Since $R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}$ is convex for all $\varepsilon \in [0,1]$ we have $G(\varepsilon, f, b) = 0$ if and only if (f,b) minimizes $R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}$ for such ε. Our aim is to apply the implicit function theorem in the way we did it in the proof of Theorem 4. However, this time the implicit function theorem will also ensure the uniqueness of the solution of (7). Obviously, this is necessary for the existence of the influence function. In order to apply Theorem 19 we need the partial derivatives of G. By an easy computation we find

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f, b) = -\mathbb{E}_{\mathbb{P}} L'(f+b)\Phi(X) + \mathbb{E}_{\Delta_z} L'(f+b)\Phi(X)$$

and

$$\frac{\partial G}{\partial(H \times \mathbb{R})}(\varepsilon, f, b) = \begin{pmatrix} 2\lambda \mathrm{id}_H + \mathbb{E}_\varepsilon L''(f+b)\langle \Phi, .\rangle \Phi & \mathbb{E}_\varepsilon L''(f+b)\Phi \\ \mathbb{E}_\varepsilon L''(f+b)\Phi & \mathbb{E}_\varepsilon L''(f+b) \end{pmatrix},$$

where we use the abbreviation $\mathbb{E}_\varepsilon := \mathbb{E}_{(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z}$. A routine check shows that both G and the partial derivatives are continuous and hence G is continuously differentiable.

Now, let us fix a solution $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ of (7). In order to show that the operator $\frac{\partial G}{\partial(H\times\mathbb{R})}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ is invertible it suffices to show by the Fredholm Alternative that it is injective and that

$$A := \begin{pmatrix} \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})\langle \Phi, .\rangle \Phi & \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})\Phi \\ \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})\Phi & \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) - 2\lambda \end{pmatrix}$$

is a compact operator on $H \times \mathbb{R}$. The latter can be seen using the argument of the proof of Theorem 4. For the former let us suppose that we have an element $(g,t) \in H \times \mathbb{R}$ with $(2\lambda \mathrm{id}_{H\times\mathbb{R}} + A)(g,t) = 0$. This is equivalent to the following linear system of equations

$$2\lambda g + \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) g\,\Phi + t\,\mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})\Phi = 0 \tag{22}$$

$$\mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) g + t\,\mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) = 0\,. \tag{23}$$

Let us first assume that $t = 0$. Then the above system yields

$$2\lambda g + \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) g \Phi = 0\,.$$

Using the techniques of the proof of Theorem 4 we easily find that this implies $g = 0$. Therefore, we may assume without loss of generality that $t = 1$. In order to avoid long notations we introduce the measure $d\mu := L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})d\mathbb{P}$. Note that $L'' > 0$ implies $\mu \neq 0$. Now, (23) yields

$$\mu(g) = -\mu(1)\,, \tag{24}$$

where 1 denotes the constant function with value 1. Hence, by (22) we find

$$0 = 2\lambda\langle g,g\rangle + \mu(g^2) + \mu(g) = 2\lambda\langle g,g\rangle + \mu(g^2) - \mu(1) \; . \tag{25}$$

Furthermore, (24) implies

$$0 \le \mu((g+1)^2) = \mu(g^2) + 2\mu(g) + \mu(1) = \mu(g^2) - \mu(1) \, .$$

This together with (25) yields $2\lambda\langle g,g\rangle \le 0$ and hence $g = 0$. However, the latter contradicts (24) and hence there is no non-trivial solution of the system (22), (23).
Now, the implicit function theorem states in particular, that the solution $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ is unique in a small neighborhood of $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$. Hence it is globally unique since the set of solutions of (7) is convex. The rest of the proof follows the ideas of the proof of Theorem 4.

We use $IF(z;T,\mathbb{P}) = \frac{\partial(f_\varepsilon,b_\varepsilon)}{\partial\varepsilon}(0)$ to derive a formula for the influence function, where $\varepsilon \mapsto (f_\varepsilon, b_\varepsilon)$ is the function implicitly defined by $G(\varepsilon, f, b) = 0$. The implicit function theorem hence gives

$$IF(z;T,\mathbb{P}) = -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) \; , \tag{26}$$

where $S := \frac{\partial G}{\partial(H\times\mathbb{R})}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$. ■

PROOF OF THEOREM 12. Recall that every convex function on $\mathbb{R}$ is locally Lipschitz continuous. Let $|L_{|Y\times[-c,c]}|_1$ denote the Lipschitz constant of L restricted to $Y \times [-c,c]$, $c > 0$. We define $\delta_\lambda := \sqrt{(L(-1,0)+L(1,0))/\lambda}$ and $K := \sup_{x\in X}\sqrt{k(x,x)}$. We fix a distribution $\mathbb{P}$. An easy estimate (Steinwart, 2002a) shows $\left\|f_{\mathbb{P},\lambda}\right\|_\infty \le \delta_\lambda K$. We will apply Theorem 28 in Steinwart (2003). Since this result is only formulated for compact subsets X we first have to check that it is also true in our situation. Indeed Propositions 26 and 27 in Steinwart (2003) only uses the boundedness of the kernel. Furthermore, the proof of Theorem 28 itself only uses the boundedness, too. Hence this theorem actually holds in our case! Thus, there exists a measurable function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \le \left|L_{|Y\times[-\delta_\lambda K,\delta_\lambda K]}\right|_1$ such that for all distributions $\hat{\mathbb{P}}$ we have

$$\|f_{\mathbb{P},\lambda} - f_{\hat{\mathbb{P}},\lambda}\|_H \le \frac{1}{\lambda}\left\|\mathbb{E}_{\mathbb{P}} h\Phi - \mathbb{E}_{\hat{\mathbb{P}}} h\Phi\right\|_H ,$$

where $\Phi : X \to H$ is the feature map of H. Let $\hat{\mathbb{P}} := (1-\varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}}$. Then the above inequality yields

$$\begin{aligned} \varepsilon^{-1}\,\|f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}\|_H &\le (\varepsilon\lambda)^{-1}\,\|\mathbb{E}_{\mathbb{P}} h\Phi - \mathbb{E}_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}}} h\Phi\|_H \\ &= \lambda^{-1}\,\|\mathbb{E}_{\mathbb{P}} h\Phi - \mathbb{E}_{\tilde{\mathbb{P}}} h\Phi\|_H \\ &\le c_L(\lambda)\,\|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}} , \end{aligned}$$

where

$$c_L(\lambda) = \lambda^{-1} K \left|L_{|Y\times[-\delta_\lambda K,\delta_\lambda K]}\right|_1 . \tag{27}$$

This shows the assertion. ■

For the proof of Theorem 15 we need the following result.

Lemma 22 *Let $\lambda,\mu \in (0,\infty)$ be fixed constants. Then the function $g:(0,1)\to\mathbb{R}$, $g(s) = 2\lambda s^2 - 2\mu s[1-s^2]^{1/2}+\mu$ has a global minimum in the point*

$$s_{min} = 2^{-1/2}\left(1-\frac{\lambda}{(\mu^2+\lambda^2)^{1/2}}\right)^{1/2}$$

and

$$g(s_{min}) = \lambda\left(1-\frac{\lambda}{(\mu^2+\lambda^2)^{1/2}}\right)+\mu\left(1-\left(1-\frac{\lambda^2}{\mu^2+\lambda^2}\right)^{1/2}\right) > 0\,.$$

PROOF OF LEMMA 22. Let $s\in(0,1)$. We have $g'(s) = 4\lambda s - 2\mu(1-2s^2)[1-s^2]^{-1/2}$ and

$$g''(s) = 4\lambda + \left(4s - \frac{s(1-2s^2)}{1-s^2}\right)\frac{2\mu}{\sqrt{1-s^2}} \geq 4\lambda + 3s\frac{2\mu}{\sqrt{1-s^2}} > 0$$

for all $s\in(0,1)$. Define $a := \lambda/\left[\mu^2+\lambda^2\right]^{1/2}$. We have $4\lambda s_{min} > 0$ and

$$\begin{aligned}\frac{2\mu(1-2s_{min}^2)[1-s_{min}^2]^{-1/2}}{4\lambda s_{min}} &= \frac{\mu}{\lambda}\,\frac{a}{[(1+a)(1-a)]^{1/2}}\\ =\frac{\mu}{\lambda}\left[a^{-2}-1\right]^{-1/2} &= \frac{\mu}{\lambda}\left[\frac{\mu^2+\lambda^2}{\lambda^2}-1\right]^{-1/2} = 1\,.\end{aligned}$$

This yields $g'(s_{min}) = 0$ and we obtain the expression $g(s_{min}) = \lambda\left[1-\frac{\lambda}{(\mu^2+\lambda^2)^{1/2}}\right]+\mu\left[1-(1-\frac{\lambda^2}{\mu^2+\lambda^2})^{1/2}\right]$. ■

PROOF OF THEOREM 15. By rescaling problem (7) we may assume without loss of generality that $K := \sup_{x\in X}\sqrt{k(x,x)} \leq 1$. Recall, that in the proof of Theorem 8 we used

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

where $\varepsilon\mapsto(f_\varepsilon,b_\varepsilon)$ was the function implicitly defined by $G(\varepsilon,f,b)=0$. The implicit function theorem hence gives

$$IF(z;T,\mathbb{P}) = -S^{-1}\circ\frac{\partial G}{\partial\varepsilon}(0,f_{\mathbb{P},\lambda},b_{\mathbb{P},\lambda})\,, \tag{28}$$

where $S := \frac{\partial G}{\partial(H\times\mathbb{R})}(0,f_{\mathbb{P},\lambda},b_{\mathbb{P},\lambda})$. Therefore, it suffices to bound the norms of the operators on the right side of (28). We begin with

$$\begin{aligned}\left\|\frac{\partial G}{\partial\varepsilon}(0,f_{\mathbb{P},\lambda},b_{\mathbb{P},\lambda})\right\| &= \|\mathbb{E}_{\mathbb{P}}L'(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\Phi - \mathbb{E}_{\Delta_z}L'(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\Phi\|\\ &\leq b_{\mathbb{P},\lambda}\,\|\mathbb{P}-\Delta_z\|_{\mathcal{M}}\,.\end{aligned}$$

Furthermore, for $(g,t)\in H\times\mathbb{R}$ we have

$$\begin{aligned}S(g,t) &= \begin{pmatrix} 2\lambda\,\mathrm{id}_H + \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\langle\Phi,.\rangle\Phi & \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\Phi\\ \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\Phi & \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\end{pmatrix}\begin{pmatrix} g\\ t\end{pmatrix}\\ &= \begin{pmatrix} 2\lambda g + \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})g\Phi + t\mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\Phi\\ \mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})g + t\mathbb{E}_{\mathbb{P}}L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})\end{pmatrix}.\end{aligned}$$

As in the proof of Theorem 8 we write $d\mu := L''(f_{\mathbb{P},\lambda}+b_{\mathbb{P},\lambda})d\mathbb{P}$. Then we find

$$\langle S(g,t),(g,t)\rangle = 2\lambda\langle g,g\rangle + \mu(g^2) + 2t\mu(g) + t^2\mu(1)\,. \tag{29}$$

Let us suppose that $\|(g,t)\| = 1$. Then there exist $w \in H$ with $\|w\| = 1$ and $s \in [0,1]$ such that $g = sw$ and $t = \pm\sqrt{1-s^2}$. If $\mu(w) \geq 0$ Equation (29) then yields

$$\langle S(g,t),(g,t)\rangle \;\geq\; 2\lambda s^2 + \frac{s^2\mu^2(w)}{\mu(1)} - 2s\sqrt{1-s^2}\mu(w) + (1-s^2)\mu(1)\,. \tag{30}$$

Here, we used $\mu(w) \leq \sqrt{\mu(1)}\sqrt{\mu(w)}$. It is easy to check that (30) also holds if $\mu(w) \leq 0$. Now, recall that by the assumption of the theorem we have $L'' \geq a$. This implies $\mu(1) \geq a > 0$. For the special case $s = 0$ it follows from (30), that

$$\langle S(g,t),(g,t)\rangle \geq \mu(1) > a > 0. \tag{31}$$

For the special case $s = 1$ it follows from (30), that

$$\langle S(g,t),(g,t)\rangle \geq 2\lambda + \frac{\mu^2(w)}{\mu(1)} \geq 2\lambda > 0\,. \tag{32}$$

Now, we will consider the case $s \in (0,1)$. We first minimize the right hand side of (30) with respect to $\mu(w)$. To this end recall that $K \leq 1$ implies $|\mu(w)| \leq \mu(1)$. Therefore, the function $\mu(w) \mapsto \frac{s^2\mu^2(w)}{\mu(1)} - 2s\sqrt{1-s^2}\mu(w)$ is minimal if

$$\mu(w) = \min\{\mu(1), \mu(1)s^{-1}\sqrt{1-s^2}\}\,. \tag{33}$$

Using (30) it follows for $\mu(w) = \mu(1)$ that

$$\langle S(g,t),(g,t)\rangle \;\geq\; 2\lambda s^2 + \mu(1)\big(1-2s\sqrt{1-s^2}\big) = 2\lambda s^2 - 2\mu(1)s\sqrt{1-s^2} + \mu(1)\,. \tag{34}$$

Hence, Lemma 22 and (34) yield for the case $s \in (0,1)$ and $\mu(w) = \mu(1)$ that

$$\langle S(g,t),(g,t)\rangle \;\geq\; c_{low} \in (0,\infty)\,, \tag{35}$$

where

$$c_{low} = \lambda\left(1 - \frac{\lambda}{(\mu^2(1)+\lambda^2)^{1/2}}\right) + \mu(1)\left(1 - \left(1 - \frac{\lambda^2}{\mu^2(1)+\lambda^2}\right)^{1/2}\right).$$

By (33) we finally have to treat the case $s \in (0,1)$ and $\mu(w) = \mu(1)s^{-1}\sqrt{1-s^2}$. In this case we have $s^2 \geq \frac{1}{2}$ by $\mu(w) \leq \mu(1)$. Hence we obtain from (30) that

$$\langle S(g,t),(g,t)\rangle \;\geq\; 2\lambda s^2 + (1-s^2)\mu(1) - 2(1-s^2)\mu(1) + (1-s^2)\mu(1) \;=\; 2\lambda s^2 \;\geq\; \lambda. \tag{36}$$

Combining (31), (32), (35), and (36) we obtain

$$\langle S(g,t),(g,t)\rangle \;\geq\; \min\{a, 2\lambda, c_{low}, \lambda\} > 0\,. \tag{37}$$

Therefore, by the proof of Pedersen (1989, Prop. 3.2.12) it follows

$$\|S(g,t)\| \geq \min\{a,\lambda,c_{low}\}\,\|(g,t)\|$$

for all $(g,t) \in H \times \mathbb{R}$, and

$$\|S^{-1}\| \leq \frac{1}{\min\{a,\lambda,c_{low}\}} \in (0,\infty)\,.$$

This shows the assertion. ■

References

R. Akerkar. *Nonlinear Functional Analysis*. Narosa Publishing House, New Dehli, 1999.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. `http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf`, 2002.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. `http://stat-www.berkeley.edu/tech-reports/638.pdf`, 2003.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

H. Bauer. *Maß- und Integrationstheorie*. De Gruyter, Berlin, 1990.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

A. Brown and C. Pearcy. *Introduction to Operator Theory I*. Springer, New York, 1977.

D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou. Support vector machine soft margin classifiers. City University of Hong Kong, 2003.

W. Cheney. *Analysis for Applied Mathematics*. Springer, New York, 2001.

A. Christmann. Least median of weighted squares in logistic regression with large strata. *Biometrika*, 81:413–417, 1994.

A. Christmann. *On positive breakdown point estimators in regression models with discrete response variables*. 1998. Habilitation thesis, University of Dortmund, Department of Statistics.

A. Christmann. On a strategy to develop robust and simple tariffs from motor vehicle insurance data. Technical Report 16/04, University of Dortmund, SFB-475, 2004. `http://www.statistik.uni-dortmund.de/download/mitarbeiter/christmann/Christmann-insurance04.pdf`.

A. Christmann, P. Fischer, and T. Joachims. Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, 17:273–287, 2002.

A. Christmann and P. J. Rousseeuw. Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*, 37:65–75, 2001.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, RI, 1977.

D. L. Donoho and P. J. Huber. The notion of breakdown point. In P. J. Bickel, K. A. Doksum, and J. L. Hodges Jr, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184, Belmont, California, Wadsworth, 1983.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the 13th International Conference*, pages 148–156. Morgan Kauffman Publishers, San Francisco, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28:337–407, 2000.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics. The approach based on influence functions*. Wiley, New York, 1986.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, Massachusetts, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

J. Hipp, U. Güntzer, and U. Grimmer. Data quality mining - making a virtue of necessity. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD, Santa Barbara, CA, `http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf`, 2001.

K. U. Höffgen, H.-U. Simon, and K. S. van Horn. Robust trainability of single neurons. *Journal Computer and System Sciences*, 50:114–125, 1995.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35: 73–101, 1964.

P. J. Huber. *Robust statistics*. Wiley, New York, 1981.

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 41–56, MIT Press, Cambridge, Massachusetts, 1999.

J. Lindenstrauss and L. Tzafriri. *Classicial Banach spaces I*. Springer, Berlin, 1977.

G. K. Pedersen. *Analysis Now*. Springer, New York, 1989.

H. Rieder. *Robust Asymptotic Statistics*. Springer, New York, 1994.

P. J. Rousseeuw and M. Hubert. Regression depth. *Journal of the American Statistical Association*, 94:388–433, 1999.

B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.

J. C. Scovel and I. Steinwart. Fast rates for support vector machines. Los Alamos Technical Report LA-UR-03-9117, `http://www.c3.lanl.gov/~ingo/publications/ann-03.ps`, 2003.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart. Consistency of support vector machines and other regularized kernel machine. *IEEE Transactions on Information Theory*, to appear, 2002a.

I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002b.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

I. Steinwart. Sparseness of support vector machines – some asymptotically sharp bounds. *Proceedings of Neural Information Processing Systems*, in press, 2004.

J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48:85–105, 2002.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32: 135–166, 2004.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88, MIT Press, Cambridge, Massachusetts, 1999.

K. Yosida. *Functional Analysis*. Springer, Berlin, 4^{th} edition, 1974.

E. Zeidler. *Nonlinear Functional Analysis and its Applications I*. Springer, New York, 1986.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–134, 2004.