Journal of Machine Learning Research ( ) Submitted 6/03; Published ??

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: Leslie Pack Kaelbling

Abstract

The paper brings together methods from two disciplines: machine learning theory and robust statistics. 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 of the influence function. Kernel logistic regression, support vector machines, least squares and the AdaBoost loss function are treated as special cases. 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 pattern recognition and statistical machine learning two major goals are the estimation of a functional relationship $y \approx f(x)$ between an outcome y and a vector of explanatory variables $x = (x_1, \ldots, x_k)' \in \mathbb{R}^d$ and the prediction of an unobserved outcome y_{new} based on an observed value x_{new}. The function f is unknown. One needs the implicit assumption that the relationship between x_{new} and y_{new} is $-$ at least almost $-$ the same than 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, y) are independent and identically generated from an underlying unknown distribution $\mathbb{P}$ for a pair of random variables (X, Y). In practical applications the training data set is often quite large, high dimensional and complex. The quality of the predictor $f(x)$ is measured by some loss function $L(y, f(x))$. The goal is to find a predictor $f_{\mathbb{P}}(x)$ which minimizes the expected loss, i.e.

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

In this paper we are interested in binary classification, where $y \in Y := \{-1, +1\}$. The straightforward prediction rule is: predict $y = +1$ if $f(x) \geq 0$, and predict $y = -1$ otherwise. The loss function for the classification error is given by $I(y, f(x)) = \mathbb{I}(yf(x) < 0) + \mathbb{I}(f(x) = 0)\mathbb{I}(y = -1)$, where $\mathbb{I}$ denotes the indicator function. Inspired by the law of large numbers one might estimate $f_{\mathbb{P}}$ by 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 overfitting 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. Hoeffgen et al. (1995). To circumvent this problem, one minimizes a convex upper bound of the classification error function $I(y, f)$, cf. Schölkopf and Smola (2002) and Vapnik (1998). If $L : Y \times \mathbb{R} \to \mathbb{R}$ is an appropriate convex function, one considers the (approximate) minimization of the empirical risks. Consider the following estimation problems:

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

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

where $\lambda > 0$ is a small regularization parameter, H is a reproducing kernel Hilbert space (RKHS) of a kernel k, and b is an unknown real-valued offset. The decision functions are $\operatorname{sign}(\hat{f}_{n,\lambda})$ or $\operatorname{sign}(\hat{f}_{n,\lambda} + \hat{b}_{n,\lambda})$. Note, that in practice usually (4) is solved while many theoretical papers deal with (3) since the unregularized offset b often causes technical difficulties. Problems (3) and (4) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk given in (5) or (6), respectively (cf. Vapnik, 1998, Zhang, 2001; Steinwart, 2002b):

$$f_{\mathbb{P},\lambda} = \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathbb{E}_{\mathbb{P}}\, L(Y, f(X)) \tag{5}$$

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

The objective functions in (5) and (6) are denoted by $R^{reg}_{L,\mathbb{P},\lambda}(.)$ and $R^{reg}_{L,\mathbb{P},\lambda}(.,.)$ in the following. 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 and plotted in Figure 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 penalizes misclassifications in a similar way than 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, Hastie and Tibshirani (2000), and Hastie, Tibshirani and Friedman (2001). The modified Huber's loss function, cf. Zhang (2001), changes the modified least squares loss such that misclassified points with $v < -1$ are penalized only linearly.

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

Table 1: Loss functions, $v = yf(x)$.

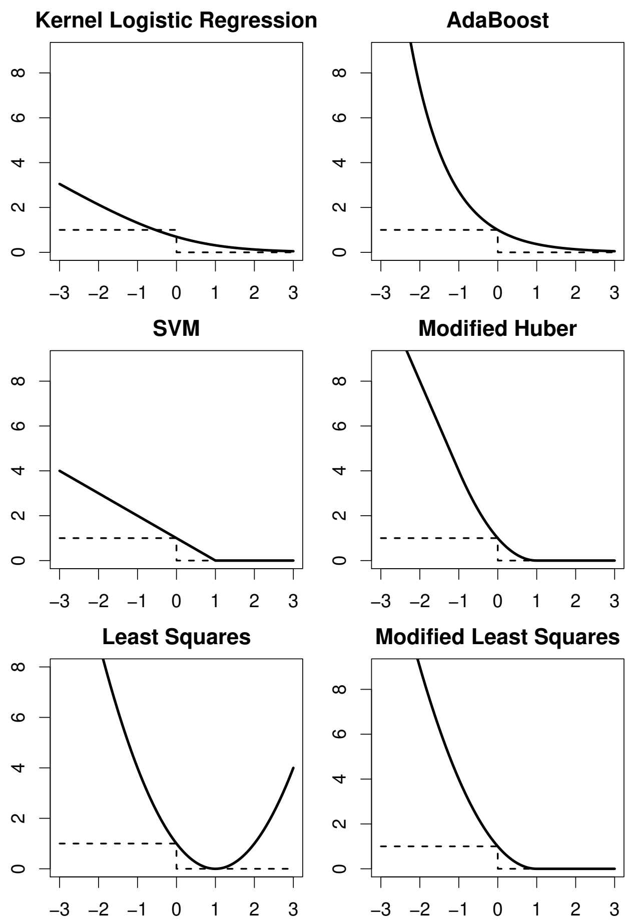

Figure 1: Illustration of different loss functions.

Steinwart (2002a) 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 (2001) 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 (2002b) 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 the minimization problems of the theoretical and of the empirical problems. 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. Finally, it turns out as a by-product, that the solutions of (3) and (5) are unique. The same holds for the RKHS part of the solutions of (4) and (6). Schölkopf und Smola (2002) describe other support vector machines and give an overview on algorithms to solve the minimization problems corresponding to SVMs.

Obviously, the proof that many classifiers based on convex loss functions are universally consistent under weak conditions is a strong argument in favor of these statistical learning methods. Nevertheless, it is important to investigate robustness properties for such statistical learning methods for the following reasons. In practice one has to apply the methods to a data set with a finite sample size. 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). 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. 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. Hipp et al. (2001). Obviously, it is *not* the goal to *model* the occurrence of typing errors or gross errors. Goals of robust statistics are to investigate the impact such data points can have on the results of estimation or testing methods and the development of methods such that the impact of such data points is bounded. Main strategies of robust statistics are Huber's minimax approach (Huber, 1964; Huber, 1981), Hampel's influence function (Hampel, 1974; Hampel et al., 1986), the finite sample breakdown point proposed by Donoho and Huber (1983), Rieder's approach based on least favourable local alternatives (Rieder, 1994), and the regression depth method proposed by Rousseeuw and Hubert (1999).

Here, we will 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. A method is called robust in the theory of robust statistics based on influence functions, if the method is based on a functional with a bounded influence function. From the viewpoint of robust statistics it is therefore 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 functionals defined by $R^{reg}_{L,\mathbb{P},\lambda}(.)$ and $R^{reg}_{L,\mathbb{P},\lambda}(.,.)$. Hence, this paper investigates robustness properties of statistical learning methods based on convex risk minimization.

The rest of the paper is organized as follows. Section 2 gives the definitions of the influence function and the sensitivity curve, which are the two robustness concepts we are dealing with. Section 3 and Section 4 contain the main results. In Section 3 sufficient conditions are given for the existence of the influence function for classifiers based on (5) and (6). In Section 4 it is shown that the influence function of the functional in (6) and the difference quotient used in the definition of the influence function for (5) can be bounded independently of z and $\mathbb{P}$. Section 5 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 6 contains the conclusion. Finally, the Appendix gives the proofs of the main theorems discussed in this paper.

2. Influence function and sensitivity curve

Goals of robust statistics are the investigation of robustness properties of statistical methods and the development of methods with good robustness properties. One major approach of robust statistics is the influence function of functionals proposed by Hampel (1974) and Hampel et al. (1986). Here, 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 is called a functional. In the case of the convex risk minimization methods (5) and (6) 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.

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

$$IF(z;T,\mathbb{P}) = \lim_{\varepsilon\downarrow 0} \frac{T\big((1-\varepsilon)\mathbb{P} + \varepsilon\Delta_z\big) - T(\mathbb{P})}{\varepsilon} \quad , \tag{7}$$

where Δ_z is the Dirac distribution at the point z.

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 is based on a functional with a *bounded* influence function.

The sensitivity curve SC_n proposed by J.W. Tukey (cf. Hampel et al., 1986, p. 93) can be interpreted as a finite sample version of the influence function (see (9)). 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 $z_1, \ldots, z_{n-1}$ is defined by*

$$SC_n(z;T_n) = n\big(T_n(z_1,\ldots,z_{n-1},z) - T_{n-1}(z_1,\ldots,z_{n-1})\big)\,. \tag{8}$$

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

$$SC_n(z;T_n) = \frac{T\big((1-\varepsilon_n)\mathbb{P}_{n-1} + \varepsilon_n\Delta_z\big) - T(\mathbb{P}_{n-1})}{\varepsilon_n}\,. \tag{9}$$

For many estimators the sensitivity curve converges to the influence function, as n tends to infinity. Counterexamples are given e.g. in Davies (1993).

3. Existence of the influence function

In this section we give sufficient conditions for the existence of the influence function for classifiers based on (5) and (6). Throughout this section B_E denotes the closed unit ball of a Banach space E. We first recall a simplified version of the implicit function theorem in Banach spaces (cf. Akerkar, 1999; Zeidler, 1986):

Theorem 3 *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*

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

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 (cf. Cheney, 2001) turns out to be helpful:

Theorem 4 *Let E be a Banach space and $K : E \to E$ be a compact operator. Then $\mathrm{id}_E + K$ is surjective if and only if it is injective.*

We first establish a result for classifiers based on (5) with smooth loss function:

Theorem 5 *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 (5) exists for all $z \in X \times Y$.*

Remark 6 *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 5 hold. This is an interesting result from the view 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 (6):

Theorem 7 *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 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 8 *As in the case of problem (5) 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 9 *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.*

4. Bounds on the influence function

As mentioned in Section 2, a desirable property of a robust statistical method is that its corresponding functional 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 (5) and (6). 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

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

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 (5). In particular, it states that the influence function of these classifiers is uniformly bounded whenever it exists. Please note, that the following theorem based on Steinwart (2002b) applies to all six loss functions given in Table 1 because differentiability of L is not assumed. Furthermore, this theorem shows that the sensitivity curves of all six methods are uniformly bounded if we set $\varepsilon = 1/n$, see (9).

Theorem 10 *Let* $L : Y \times \mathbb{R} \to [0, \infty)$ *be a continuous and convex loss function. Furthermore, let* $X \subset \mathbb{R}^d$ *be compact and* H *be a RKHS of a continuous kernel on* X*. Then for all* $\lambda > 0$ *there exists a constant* $c_L(\lambda) > 0$ *explicitly given in (20) such that for all distributions* $\mathbb{P}$ *and* $\tilde{\mathbb{P}}$ *on* $X \times Y$ *we have*

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

Unfortunately, using the estimate of Steinwart (2002b) does not give any meaningful result for classifiers based on (6). Therefore, the approach of the following theorem is to apply the formula for the derivative given by the implicit function theorem.

Theorem 11 *Let* $L : Y \times \mathbb{R} \to [0, \infty)$ *be a convex and twice continuously differentiable loss function with* $a \le L'' \le b$ *for some* $a, b > 0$*. Furthermore, let* $X \subset \mathbb{R}^d$ *be compact,* 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 (6). Then for all* $\lambda > 0$ *there exists a constant* $c_L(\lambda) > 0$ *such that for all distributions* $\mathbb{P}$ *on* $X \times Y$ *and all* $z \in X \times Y$ *we have*

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

Remark 12 *Theorem 11 applies to (6) with the least squares loss function. However, Theorem 11 covers neither the logistic regression 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*

$$a \;\leq\; \int L''(Y, f_{\mathbb{P},\lambda}(X) + b_{\mathbb{P},\lambda}) d\mathbb{P} \;\leq\; b \tag{10}$$

for some $b \geq a > 0$. A simple sufficient condition for the latter can be derived by the proof of Steinwart (2002b, Lemma II.6): let $A_y^\rho := \{x \in X : \mathbb{P}(y|x) > \rho\}$, $y \in Y$, $\rho > 0$, and $\alpha_{\mathbb{P}}(\rho) := \rho \min\{\mathbb{P}_X(A_1^\rho), \mathbb{P}_X(A_{-1}^\rho)\}$. Fixing $\lambda > 0$, a twice continuously differentiable L and a threshold $\alpha > 0$ there exists $b \geq a > 0$ such that every $\mathbb{P}$ with $\alpha_{\mathbb{P}}(\rho) \geq \alpha$ for some $\rho > 0$ satisfies (10). Note, that the assumption $\alpha_{\mathbb{P}}(\rho) \geq \alpha$ guarantees that the two classes of $\mathbb{P}$ are "balanced".

Remark 13 *As mentioned in Remark 8 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 11 shows that the special Gâteaux derivative of T can be uniformly bounded.*

Remark 14 *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 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:*

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

5. 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

$$\begin{array}{ll} \arg\min_{f \in H,\, b \in \mathbb{R}} & \frac{1}{2Cn}\, ||f||_H^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \\ \text{such that} & y_i(f(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0\,. \end{array} \tag{11}$$

We consider two popular kernels: a Gaussian radial basis function (RBF) kernel $f(x, x') = \exp(-\gamma \|x - x'\|^2)$ 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 $\gamma \in$

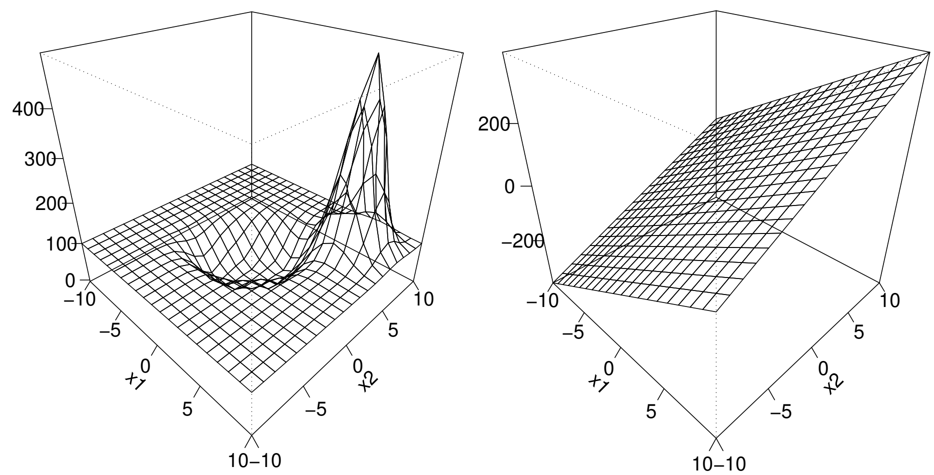

Figure 2: Sensitivity function of $\hat{f}+\hat{b}$, if the additional data point z is located at $z=(x,y)$, where $x=(6,6)$ and $y=1$. Left: RBF kernel. Right: linear kernel.

$\{0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 10, 20\}$ 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 2 shows the sensitivity curves of $\hat{f}+\hat{b} := \hat{f}_{n,\lambda}+\hat{b}$, 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 neighorhood 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 3 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

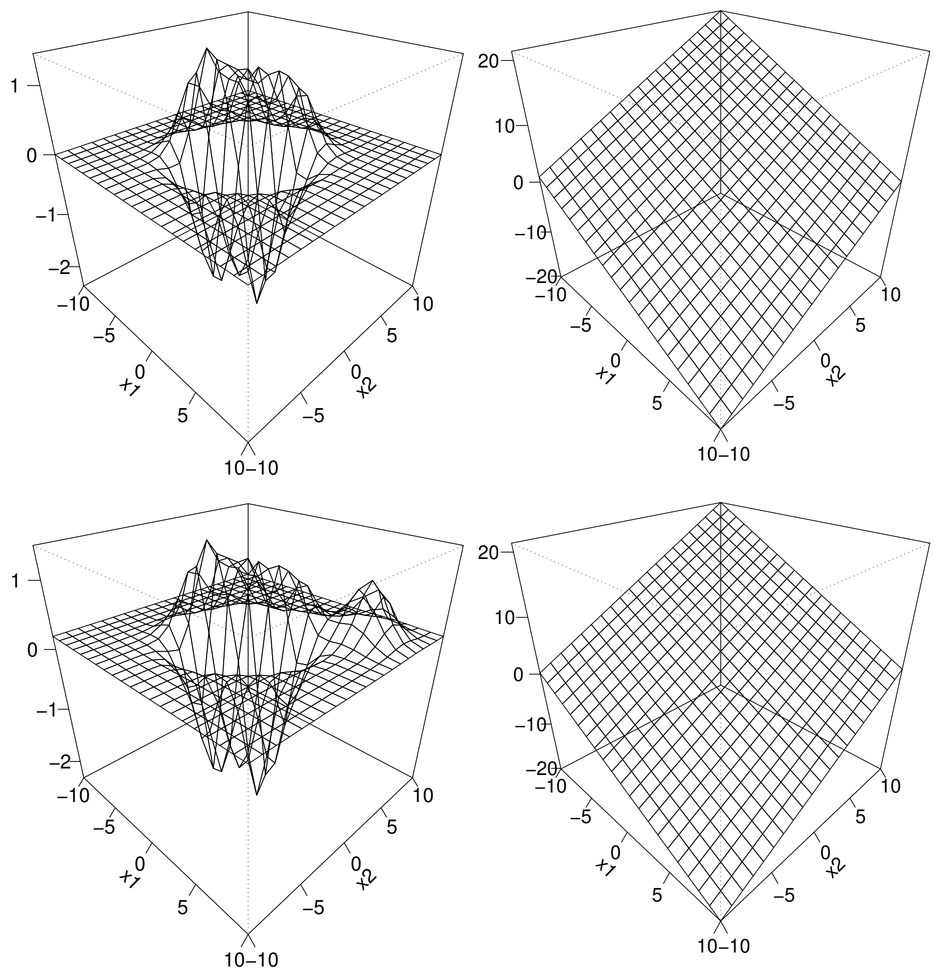

Figure 3: Plot of $\hat{f} + \hat{b}$. Upper left: RBF kernel, original data set. Upper right: linear kernel, original data set. Lower left: RFB kernel, modified data set. Lower right: linear kernel, modified data set. The modified data set contains the additional data point $z = (x, y)$, where $x = (6, 6)$ and $y = 1$.

$x-$coordinates. Figure 4 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 halfspace but changes in the other halfspace. 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 5. 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 halfspace 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 6. We vary z over a grid in the $x-$coordinates in the same manner as before.

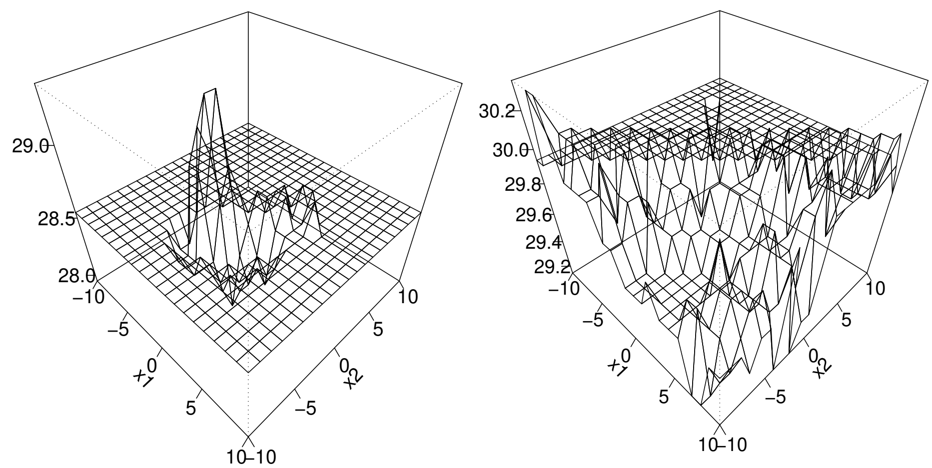

Figure 4: Percent of classification errors if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel

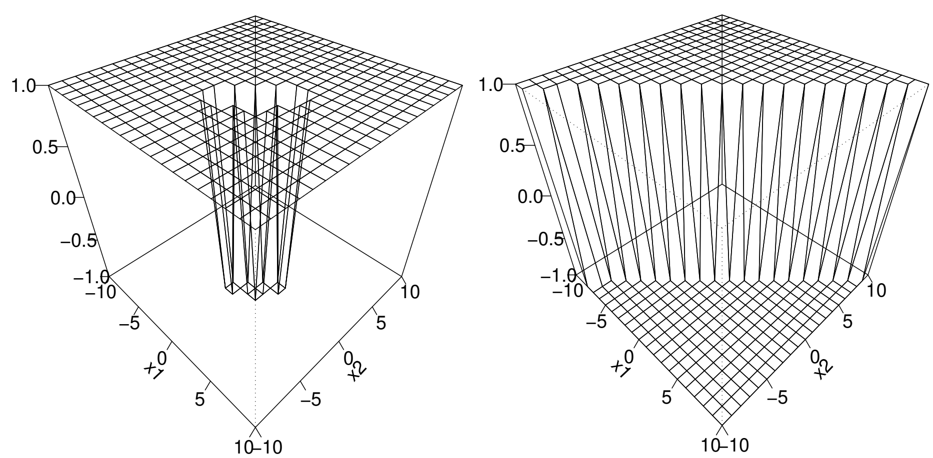

Figure 5: Fitted $y-$value for new observation if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel

As the plots for $\hat{\theta}_1$ and $\hat{\theta}_2$ are looking very similar, we only show the latter. Please note, that the axes are not identically in Figure 6 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 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 halfspace, but change approximately linear in the other affine halfspace. This behavior also occurs for the sensitivity curve of the offset by using a linear kernel.

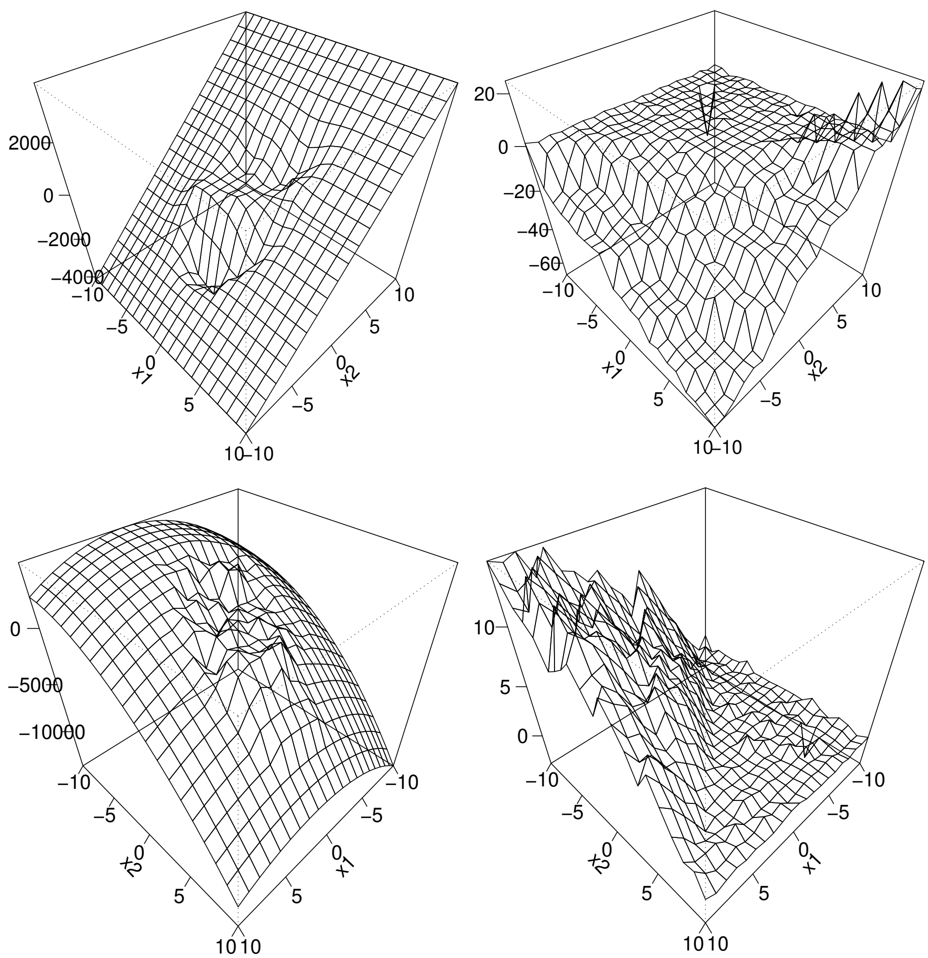

Figure 6: Sensitivity function for $\hat{\theta}$ and $\hat{b}$, respectively. Upper left: Sensitivity function for $\hat{\theta}_2$, RBF kernel. Upper right: Sensitivity function for $\hat{\theta}_2$, linear kernel. Lower left: Sensitivity function for $\hat{b}$, RBF kernel. Lower right: Sensitivity function for $\hat{b}$, linear kernel.

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.

6. 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. 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 the classifiers and also for bounds of the influence function which hold uniformally 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. 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. Hence, we gave some numerical results for the sensitivity curve, which can be interpreted as a final 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.

We briefly like to mention that the sensitivity curves for the slope parameters of a support vector machine with a RBF kernel $k(x, x') = \exp(-\gamma\|x - x'\|^2)$ are looking quite similar to rotated sensitivity curves or to influence functions of robust S-estimators based on a smooth $\rho-$function in the linear regression model. This might be a consequence of a relationship between the SVM using a RBF kernel and robust S-estimators based on a smooth $\rho-$function fulfilling the usual properties (cf. Davies, 1990): (a) $\rho(u) = \rho(-u)$, $u \in \mathbb{R}$, (b) $\rho(u)$, $u > 0$, is nonincreasing, continuous at 0 and continuous on the left, and (c) for some $c > 0$, $\rho(u) > 0$ if $|u| \leq c$, and $\rho(u) = 0$ if $|u| > c$, which is true e.g. for $\rho_c(u) = \left(1 - u^2/c^2\right)^2$, if $|u| <= c$, $\rho_c(u) = 0$ else. The RBF kernel k of a SVM considered as a function of $u = \|x - x'\|$ has similar properties than the $\rho-$function used by S-estimators. Consider the linear regression model $y_i = x_i'\theta + \varepsilon_i$, $1 \leq i \leq n$, where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$, and $\theta \in \mathbb{R}^p$. Further, assume ε_i, $1 \leq i \leq n$, are independently and identically distributed random variables with respect to some distribution $\mathbb{P}$ such that $\mathbb{P}(\varepsilon_i \leq u) = F(u/\sigma)$, $u \in \mathbb{R}$, where $\sigma \in (0, \infty)$ is a scale parameter and $F : \mathbb{R} \to [0, 1]$ is a nondegenerate distribution function. An S-estimate $(\hat{\theta}, \hat{\sigma})$ of (θ, σ) is implicitly defined by minimizing a scale parameter σ subject to an inequality constraint, i.e.

$$\arg\min_{\theta \in \mathbb{R}^p,\, \sigma \in (0,\infty)} \sigma \tag{12}$$

$$\text{s.t.} \qquad \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i'\theta}{\sigma}\right) \geq 1 - \varepsilon, \tag{13}$$

cf. Rousseeuw and Yohai (1984) and Davies (1990). The constraint guarantees that at least $n(1 - \varepsilon)$ of the residuals $(y_i - x_i'\theta)/\sigma$ have absolute values less than of equal to c due to property (c) of the $\rho-$function. Formula (11) allows the interpretation that the SVM minimizes an average plus a regularized squared norm (and hence a measure for variability) with respect to several inequality constraints.

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, Fischer and Joachims (2002).

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 proposed by Huber (1981) and Donoho and Huber (1983), 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") is gratefully acknowledged.

Appendix A.

In this appendix we prove the theorems from Section 3 and Section 4.

Proof of Theorem 5. Let $\Phi : X \to H$ be the feature map of H, i.e. $\Phi(x) := k(x, .)$, where k is the kernel of H. Let us consider 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 the above expectation is actually a Bochner integral in H. Furthermore, for $\varepsilon \notin [0, 1]$ the expectation is with respect to a signed measure. Obviously, for $\varepsilon \in [0, 1]$ we obtain

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

Since $R^{reg}_{L,(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z,\lambda}$ is convex for all $\varepsilon \in [0, 1]$ we have $G(\varepsilon, f) = 0$ 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 3 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

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathbb{P}} L'(Y, f(X))\Phi(X) + \mathbb{E}_{\Delta_z} L'(Y, f(X))\Phi(X) . \tag{14}$$

Moreover, using the reproducing property $\langle \Phi(x), g\rangle = g(x)$, $g \in H, x \in X$ we find

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda\, \mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathbb{P}+\varepsilon\Delta_z} L''(Y, f(X))\langle \Phi(X), .\rangle \Phi(X) . \tag{15}$$

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 recall that $\Phi(X)$ is compact by the continuity of Φ. Therefore, there exists a $c > 0$ such that

$$Ag \in c \cdot \overline{\operatorname{aco} \Phi(X)}$$

for all $g \in B_H$. Since the closure of the absolute convex hull $\operatorname{aco}\Phi(X)$ is compact the desired compactness of the operator A follows. Furthermore, for $g \neq 0$ we find

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

since the second derivative of a convex function is always nonnegative. Therefore, $\frac{\partial G}{\partial H}(0, f_{\mathbb{P},\lambda}) = 2\lambda \operatorname{id}_H + A$ is injective. ■

Proof of Theorem 7. We sometimes write $L(f + b)$ 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. The proof is similar to that of Theorem 5. However, due to the extra variable b we have to modify our approach: we 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 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 5. However, this time the implicit function theorem will also ensure the uniqueness of the solution of (6). Obviously, this is necessary for the existence of the influence function. In order to apply Theorem 3 we need the partial derivatives of G. By an easy computation we find

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

and

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

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 (6). Existence of a solution follows from Zhang (2001), Steinwart (2002b), and Steinwart (2003). 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 $\frac{\partial G}{\partial(H\times\mathbb{R})}(0, f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda})$ is injective and that

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

is a compact operator on $H \times \mathbb{R}$. The latter can be seen using the argument of the proof of Theorem 5. For the former let us suppose that we have an element $(g, t) \in H \times \mathbb{R}$ with $(2\lambda \mathrm{id}_{H\times\mathbb{R}} + K)(g, t) = 0$. This is equivalent to the following linear system of equations

$$2\lambda g + \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) g\, \Phi + t\, \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})\Phi = 0 \tag{16}$$

$$\mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) g + t\, \mathbb{E}_{\mathbb{P}} L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) = 0\,. \tag{17}$$

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 5 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 $\mu := L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda}) d\mathbb{P}$. Note, that $L'' > 0$ implies $\mu \neq 0$. Now, (17) yields

$$\mu(g) = -\mu(\mathbf{1})\,, \tag{18}$$

where $\mathbf{1}$ denotes the constant function with value 1. Hence, by (16) we find

$$0 = 2\lambda\langle g, g\rangle + \mu(g^2) + \mu(g) = 2\lambda\langle g, g\rangle + \mu(g^2) - \mu(\mathbf{1})\;. \tag{19}$$

Furthermore, (18) implies

$$0 \leq \mu((g + \mathbf{1})^2) = \mu(g^2) + 2\mu(g) + \mu(\mathbf{1}) = \mu(g^2) - \mu(\mathbf{1})\,.$$

This together with (19) yields $2\lambda\langle g, g\rangle \leq 0$ and hence $g = 0$. However, the latter contradicts (18) and hence there is no non-trivial solution of the system (16), (17).
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 (6) is convex. The rest of the proof follows the ideas of the proof of Theorem 5. ■

Proof of Theorem 10. 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 (cf. Steinwart, 2002b) shows $\|f_{\mathbb{P},\lambda}\|_\infty \leq \delta_\lambda K$. Now, by Theorem 3.15 in Steinwart (2003) there exists a measurable function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \leq \left|L_{|Y\times[-\delta_\lambda K, \delta_\lambda K]}\right|_1$ such that for all distributions $\hat{\mathbb{P}}$ we have

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

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

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

where

$$c_L(\lambda) = \lambda^{-1}\,K\big|L_{|Y\times[-\delta_\lambda K,\delta_\lambda K]}\big|_1\,. \tag{20}$$

This shows the assertion. ■

Proof of Theorem 11. By rescaling problem (6) 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 7 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

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

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 (21). We begin with

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

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 7 we write $\mu := L''(f_{\mathbb{P},\lambda} + b_{\mathbb{P},\lambda})d\mathbb{P}$. Then we find

$$\langle S(g,t), (g,t)\rangle = 2\lambda\langle g, g\rangle + \mu(g^2) + 2t\mu(g) + t^2\mu(\mathbf{1})\,. \tag{22}$$

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 $t = \pm\sqrt{1-s^2}$. Note, that since $K \leq 1$ we have $|\mu(w)| \leq \mu(\mathbf{1})$. Therefore, (22) yields

$$\begin{aligned}
\langle S(g,t), (g,t)\rangle &\geq 2\lambda s^2 - 2s\sqrt{1-s^2}\mu(\mathbf{1}) + (1-s^2)\mu(\mathbf{1}) \\
&\geq 2\lambda s^2 - 2s(1-s)\mu(\mathbf{1}) + (1-s^2)\mu(\mathbf{1}) \\
&= 2\lambda s^2 + s^2\mu(\mathbf{1}) - 2s\mu(\mathbf{1}) + \mu(\mathbf{1}) \\
&\geq \frac{\lambda\mu(\mathbf{1})}{2\lambda + \mu(\mathbf{1})}\,,
\end{aligned}$$

where the last estimate is based on a simple minimization with respect to s. By the proof of Pedersen (1989, Prop. 3.2.12) we hence find

$$\|S(g,t)\| \geq \frac{\lambda\mu(\mathbf{1})}{2\lambda+\mu(\mathbf{1})}\|(g,t)\|$$

for all $(g,t) \in H \times \mathbb{R}$. Hence we obtain

$$\|S^{-1}\| \leq \left(\frac{\lambda\mu(\mathbf{1})}{2\lambda+\mu(\mathbf{1})}\right)^{-1} = \frac{1}{\lambda} + \frac{2}{\mu(\mathbf{1})} \ .$$

Since $L'' \geq a$ implies $\mu(\mathbf{1}) \geq a > 0$ we have shown the assertion. ■

References

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

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

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

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

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

Davies, P.L. The asymptotics of S-estimators in the linear regression model. *Ann. Statist.*, 18:1651-1675, 1990.

Davies, P.L. Aspects of robust linear regression. *Ann. Statist.*, 21:1843-1899, 1993.

Donoho, D.L. and Huber, P.J. (1983). The Notion of Breakdown Point. In *A Festschrift for Erich L. Lehmann*, eds. P.J. Bickel, K.A. Doksum, and J.L. Hodges, Jr., pages 157-184, Belmont, California, Wadsworth, 1983.

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

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

Hampel, F.R. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383-393, 1974.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. *Robust statistics: The Approach Based on Influence Functions.* Wiley, New York, 1986.

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

Hipp, J., Güntzer, Grimmer, U. (2001). Data Quality Mining - Making a Virtue of Necessity. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD. Santa Barbara, CA. Available electronically at `http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf`

Höffgen, K.U., Simon, H.-U. and van Horn, K.S. Robust Trainability of Single Neurons. *J. Computer and System Sciences*, 50:114-125, 1995.

Huber, P.J. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73-101, 1964.

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

Joachims, T. *Making large-Scale SVM Learning Practical.* In *Advances in Kernel Methods - Support Vector Learning*, eds: B. Schölkopf, C. Burges, A. Smola. MIT Press, Cambridge, Massachusetts, 1999.

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

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

Rousseeuw, P.J. and Yohai, V. Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.*, 26:256-272, 1984.

Rousseeuw, P.J. and Hubert, M. Regression Depth. *J. Amer. Statist. Assoc.*, 94:388-433, 1999.

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

Steinwart, I. Support vector machines are universally consistent. *J. Complexity*, 18:768-791, 2002a.

Steinwart, I. Consistency of support vector machines and other regularized kernel machine. Preprint. Available electronically at `http://www.c3.lanl.gov/~ingo/publications/info-02.ps`, 2002b.

Steinwart, I. Sparseness of support vector machines. Preprint. Available electronically at `http://www.c3.lanl.gov/~ingo/publications/jmlr-03.ps`, 2003.

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

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

Zhang, T. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist*, forthcoming, 2001.