10

Robustness

Overview. *So far, we have assumed that the pairs of input and output variables are independently generated by the same probability distribution* P*. In this chapter, we investigate robustness properties of support vector machines for classification and regression. Here we will determine the stability of SVMs if the distribution is not* P *but some other distribution* Q *close to* P*. For example,* Q *can be the empirical distribution generated by a data set or a mixture distribution* $\mathrm{Q} = (1-\varepsilon)\mathrm{P}+\varepsilon\tilde{\mathrm{P}}$*, where* $\tilde{\mathrm{P}}$ *can be any distribution. Using methods from robust statistics, we obtain conditions on the loss function and the kernel that guarantee that SVMs are stable in a neighborhood around* P*.*

Prerequisites. *Knowledge of loss functions, kernels, and the stability of infinite-sample SVMs is needed from Chapter 2, Section 3.9, Chapter 4, and Section 5.3. Some results from measure and integration theory, statistics, and functional analysis from the appendix are used.*

In this chapter, we argue that robustness is an important aspect for statistical machine learning. This chapter can be seen as a continuation of Section 5.3, which investigated the stability of infinite-sample SVMs. It will be shown that SVMs also have—besides other good properties—the advantage of being robust if the loss function L and the kernel k are carefully chosen. Weak conditions on L and k are derived that guarantee good robustness of SVM methods not only for some fixed parametric class of distributions—say Gaussian distributions—but for large classes of probability distributions. In the sense specified by Hadamard (1902), support vector machines are hence well-posed problems. Hadamard believed that well-posed mathematical problems should have the property that there exists a unique solution that additionally depends on the data continuously.

In previous chapters of the book, statistical properties of SVMs were derived such as existence, uniqueness, and L-risk consistency. These properties were derived under the model assumption that the pairs (X_i, Y_i), $i = 1, \dots, n$, are *independent* and *identically distributed* random variables on some space $X \times Y$ each with the (totally unknown) probability distribution $\mathrm{P} \in \mathcal{M}_1$, where $\mathcal{M}_1$ denotes the set of all probability distributions on $X \times Y$. This is surely a weak assumption if it is compared with parametric models that assume that P is an element of a specified *finite-dimensional* subset of $\mathcal{M}_1$.

From a theoretical and an applied point of view, it is nevertheless worthwhile to investigate the possible impact of model violations on $f_{\mathrm{P},\lambda}$, $f_{\mathrm{D},\lambda}$, and the corresponding L-risks. The results of this chapter can be used to choose the loss function L and the kernel k such that the corresponding SVM is robust.

If not otherwise stated, we assume in this chapter that $X = \mathbb{R}^d$, $Y = \mathbb{R}$, and $d \in \mathbb{N}$.

The rest of the chapter is organized as follows. Section 10.1 gives a motivation for investigating robustness properties of SVM methods. Section 10.2 describes general concepts of robust statistics: qualitative robustness, influence functions, the related notions of gross error sensitivity, the sensitivity curve, and maxbias, and breakdown points. In the following two sections, we investigate in detail robustness properties of SVMs for classification and regression and clarify the role of the loss function and the kernel. Section 10.5 treats a simple but powerful strategy based on independent subsampling for calculating SVMs from huge data sets for which current numerical algorithms might be too slow. It will be shown that this strategy offers robust and consistent estimations if the SVM used is itself robust and consistent.

10.1 Motivation

Why Is Robustness Important?

In almost all cases, statistical models are only approximations to the true random process that generated a given data set.[1] Hence it is necessary to investigate how deviations may influence 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 for further treatment of points deviating from this structure or deviating substructures, see Hampel *et al.* (1986). An informal description of a good robust method is as follows.

i) A good robust method offers results with a reasonably high quality when the data set was in fact generated by the assumed model.

[1] This is especially true in data mining; see Chapter 12.

ii) If the strict model assumptions are violated, then the results of a robust method are only influenced in a bounded way by a few data points that deviate grossly from the structure of the bulk of the data set or by many data points that deviate only mildly from the structure of the bulk of the data set.

Although these two considerations are for data sets (i.e., for the finite-sample case), it will become clear that asymptotic considerations are also helpful for investigating robustness properties.

Support vector machines are non-parametric methods and make no specific assumptions on the distribution P. Nevertheless, the robustness issue is important also for SVMs because the two classical assumptions that all data points are generated *independently* 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 that "*are far away ... from the pattern set by the majority of the data*"; see Hampel *et al.* (1986, p. 25). Let us consider a simple two-dimensional example. Figure 10.1 shows a simulated data set from a bivariate Gaussian distribution $N(\mu, \varSigma)$ with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \varSigma = \begin{pmatrix} 1.0 & 1.2 \\ 1.2 & 2.0 \end{pmatrix},$$

and with four artificial outliers. Of course, μ and $\varSigma$ will usually be unknown in applications. The ellipses cover 95 percent (dashed) and 99.99 percent (dotted) of the mass of this bivariate Gaussian distribution. The points $a = (4, 5)$ and $d = (5, -5)$ are extreme in the x- and y-directions, $b = (4, 0)$ is only extreme in the x-direction, and $c = (1.5, -2)$ is non-extreme in both directions but deviates strongly from the pattern set by the majority of the data. It is often relatively easy to detect outliers that are extreme in at least one component, in this case the points a, b, and d. However, outliers are often very hard to identify in high-dimensional data sets due to the curse of high dimensionality and because it is not feasible to consider all one-dimensional projections. One reason is that such outliers can be non-extreme in every component and in many lower-dimensional subspaces but are extreme if all components and the (unknown) pattern set by the majority of the data are taken into account at the same time; see the point c in Figure 10.1. With respect to this bivariate Gaussian distribution, c is even more extreme than a because c is not even an element of the ellipse containing 99.99 percent of the distribution mass of the specified bivariate Gaussian distribution. This can also be seen by computing the so-called *Mahalanobis distance*, which is defined by

$$\sqrt{(z-\mu)^{\mathsf{T}} \varSigma^{-1} (z-\mu)}, \quad z \in \mathbb{R}^2.$$

The Mahalanobis distance is 4.0 for a but equals 5.3 for c and 15.0 for d. In contrast to that, the Euclidean distance is not helpful here. The Euclidean distance of c to μ is only 2.5, which is smaller than the Euclidean distance of a to μ, which equals 6.4. Outliers like c and d usually have a large impact on non-robust methods and can make the whole statistical analysis useless if a non-robust method is used; see, e.g., Rousseeuw (1984), Rousseeuw and van Zomeren (1990), and Davies and Gather (1993).

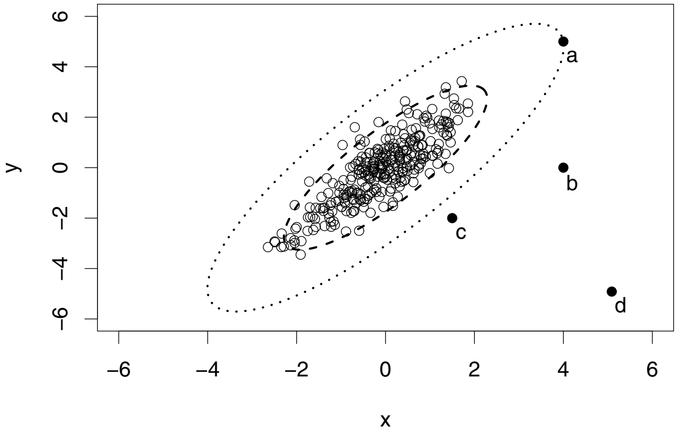

Fig. 10.1. Types of outliers.

There are many reasons for the occurrence of outliers or extreme values. *Typing errors* are often present if the data points are reported manually. *Gross errors* are errors due to a source of deviation that acts only occasionally but is quite powerful. Another reason might be that the whole data set or a part of it was actually generated by another distribution, say a Student's t-distribution or a generalized Pareto distribution, under which extreme values are much more likely than under the assumed model, say the class of Gaussian distributions. It can happen that outliers are even correlated, which contradicts the classical assumption that the observations in the data set are generated in an independent manner.

One might ask whether it is necessary to pay attention to outliers or gross errors. This is definitely true, as the number of patents on methods claiming reliable outlier detection shows. We would like to give three reasons for the importance of outliers:

i) Outliers do occur in practice. There are often no or virtually no gross errors in high-quality data, but 1 percent to 10 percent of gross errors in routine data seem to be more the rule than the exception; see Hampel *et al.* (1986, pp. 27ff.). The data quality is sometimes far from being optimal, especially in data mining problems, as will be explained in Chapter 12.

ii) Outliers may unfortunately have a high impact on the results if methods are used that do *not* bound the impact of outliers.

iii) Outliers may be interesting in their own right because they show a different behavior than the bulk of the data. This might even indicate some novelty worth considering in a detailed subsequent analysis. It is well-known that outlier identification based on robust methods is much safer than outlier identification based on non-robust methods.

It is worth mentioning that, from a robustness point of view, the occurrence of outliers is only one of several possible deviations from the assumed model. Obviously, it is in general *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 that will be collected in the future.

Let us summarize. The classical assumptions made by SVMs of independent random vectors (X_i, Y_i), $i = 1, \ldots, n$, each having the same distribution P, are weak but nevertheless not always fulfilled. Hence the question arises which impact small distortions of P may have on SVMs. Of course, the same question arises for empirical SVMs, where the unknown distribution P is replaced by the empirical distribution D. We will later use neighborhoods of P or D in the metric space of probability distributions to specify precisely what is meant by small distortions.

What Are the Goals of Robust Statistics?

In a nutshell, robust statistics investigates the impact that violations of the statistical model, outliers, and gross errors can have on the results of estimation, testing, or prediction methods and develops methods such that the impact is bounded.

Figure 10.2 sketches the idea of robustness from a somewhat more mathematical point of view. Assume that (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed random variables on some space $X \times Y \subset \mathbb{R}^d \times \mathbb{R}$ with unknown probability distribution P. Denote the empirical distribution corresponding to a data set $D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ by $\mathrm{D} = \mathrm{D}_n$. The sequence $(\mathrm{D}_n)_{n \in \mathbb{N}}$ converges almost surely by Theorem A.4.11 to the true data-generating distribution P if the sample size n converges to infinity. Assume that the statistical method of interest can be written as $S(\mathrm{D}_n)$ for any possible data set D_n and more general, as $S(\mathrm{P})$ for any distribution P, where

$$S : \mathrm{P} \mapsto S(\mathrm{P}) \tag{10.1}$$

is a measurable function specifying the statistical method. In this chapter, we will assume that the functions considered are measurable. In our case, we will have

$$S(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg\min_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 \,, \tag{10.2}$$

where $\lambda \geq 0$. A robust method S should be bounded and smooth in a neighborhood of P. Let us assume for a moment that continuity or a bounded directional derivative is meant by smoothness. Let $\mathcal{M}_1$ be the set of all probability measures on $X \times Y$ and a corresponding σ-algebra, d_1 be a metric on $\mathcal{M}_1$, and d_2 be a metric on the space of induced probability measures of $S(\mathrm{P})$, $\mathrm{P} \in \mathcal{M}_1$. Then we expect from a robust method that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$d_1(\mathrm{Q}, \mathrm{P}) < \delta \quad \Rightarrow \quad d_2(S(\mathrm{Q}), S(\mathrm{P})) < \varepsilon,$$

or that the derivative of $S(\mathrm{P})$ in the direction of $S(\mathrm{Q})$ is bounded. The smoothness property should in particular be valid for any sequence $(\mathrm{P}_n)_{n\in\mathbb{N}}$ of probability distributions converging to P. A special case are sequences of empirical distributions $(\mathrm{D}_n)_{n\in\mathbb{N}}$ converging to P. Recall that P is unknown. Hence it is essential that the function S has this smoothness property not only for one particular distribution P but for a large subclass of $\mathcal{M}_1$.

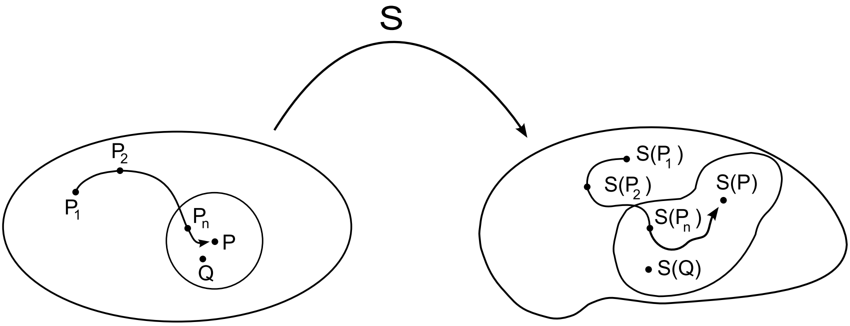

Fig. 10.2. Sketch: reasoning of robustness of $S(\mathrm{P})$. Left: P, a δ-neighborhood of P, and $\mathcal{M}_1$. Right: $S(\mathrm{P})$, an ε-neighborhood of $S(\mathrm{P})$, and the space of all probability measures of $S(\mathrm{P})$ for $\mathrm{P} \in \mathcal{M}_1$.

Let us consider two rather simple examples: the mean and the median. Let $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. The (empirical) mean is defined by

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i\,, \tag{10.3}$$

and the (empirical) median is given by

$$\mathrm{median}(y) := \begin{cases} y_{(\ell:n)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\big(y_{(\ell:n)} + y_{((\ell+1):n)}\big) & \text{if } n \text{ is even}\,, \end{cases} \tag{10.4}$$

where $\ell = \lfloor (n+1)/2 \rfloor$, and $y_{(1:n)} \le \ldots \le y_{(n:n)}$ denote the ordered values of $\{y_i, i = 1, \ldots, n\}$. Here we use the usual way of making the median unique for

n even. Now assume that the data points y_i are the observations of independent and identically distributed random variables Y_i on $(\mathbb{R}, \mathcal{B})$, $i = 1, \ldots, n$. Clearly, $\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the solution of

$$\bar{Y} = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - \theta)^2 , \tag{10.5}$$

and the median is a solution of

$$\operatorname{median}(Y) = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |Y_i - \theta| . \tag{10.6}$$

The mean and median are also solutions of a special class of empirical risk minimization problems. This can easily be seen if we use a reproducing kernel Hilbert space H defined via a *linear* kernel $k(x, x') = \langle x, x'\rangle$, the least squares loss or the L_1 loss function, which is identical to the pinball loss function for $\tau = \frac{1}{2}$, and a data set containing n data points $z_i = (x_i, y_i) \in \mathbb{R}^2$ with $x_i \equiv 1$, $i = 1, \ldots, n$. Assume that $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed random variables on $(\mathbb{R}^2, \mathcal{B}^2)$ with distribution $\mathrm{P} \in \mathcal{M}_1$ such that the marginal distribution P_X of X_i is equal to the Dirac distribution $\delta_{\{1\}}$. Under these assumptions, we obtain

$$S(\mathrm{D}) := f_{\mathrm{D},0} = \arg\min_{f\in H} \frac{1}{n}\sum_{i=1}^{n} L\bigl(y_i, f(x_i)\bigr) \tag{10.7}$$

with the corresponding infinite-sample problem

$$S(\mathrm{P}) := f_{\mathrm{P},0} = \arg\min_{f\in H} \mathbb{E}_{\mathrm{P}}\, L\bigl(Y, f(X)\bigr) . \tag{10.8}$$

Therefore, we can consider the mean and median either as function values $S_n((z_1, \ldots, z_n))$ of a function

$$S_n : \bigl(\mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n})\bigr) \to (\mathbb{R}, \mathcal{B}) \tag{10.9}$$

(see (10.3) and (10.4)) or as function values $S(\mathrm{D}_n)$ of a function S, where $\mathrm{D}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}$ denotes the empirical distribution and

$$S : \mathcal{M}_1(\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}) , \tag{10.10}$$

$$S(\mathrm{P}) = \arg\min_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2 , \tag{10.11}$$

where $\lambda = 0$ (see (10.7) and (10.8)). As far as we know, this functional approach was first used by von Mises (1937), and it is now a standard and powerful tool for investigating robustness properties of statistical methods. Robustness of an estimator S_n can then be defined via properties of the function S by requiring continuity, differentiability, or boundedness of S in P or in neighborhoods of P. This will be described in the next section.

10.2 Approaches to Robust Statistics

This section contains the necessary preliminaries to investigate robustness properties of SVMs in Sections 10.3 to 10.5. A reader familiar with robust statistics may skip this section and go directly to Section 10.3. A reader not familiar with robust statistics can find additional information also in Appendix A.4.

In the statistical literature, many different criteria have been proposed to define robustness in a mathematical way. The definitions of qualitative robustness and breakdown points are given, but we will mainly restrict attention to influence functions together with the related notions of gross error sensitivity, sensitivity curve, and maxbias in the subsequent sections.

Qualitative Robustness

As we explained in the previous section, neighborhoods and distances of probability measures are important in robust statistics. Let us therefore start with the definition of the Prohorov metric, which is needed for the definition of qualitative robustness. Let (Z, τ_Z) be a Polish space with complete metric d_Z. For any set $A \subset Z$, we define the *closed δ-neighborhood* of a set A by

$$A^\delta := \big\{x \in Z : \inf_{y \in A} d_Z(x,y) \leq \delta\big\}. \tag{10.12}$$

Definition 10.1. *Let (Z, τ_Z) be a Polish space, $\mathrm{P} \in \mathcal{M}_1(Z)$ be a probability measure on Z, and $\varepsilon > 0$, $\delta > 0$. Then the set*

$$N^{Pro}_{\varepsilon,\delta}(\mathrm{P}) := \big\{\mathrm{Q} \in \mathcal{M}_1(Z) : \mathrm{Q}(A) \leq \mathrm{P}(A^\delta) + \varepsilon \ \ \textit{for all } A \in \mathcal{B}(Z)\big\} \tag{10.13}$$

*is called a **Prohorov neighborhood** of P. We write $N^{Pro}_{\varepsilon}(\mathrm{P})$ instead of $N^{Pro}_{\varepsilon,\varepsilon}(\mathrm{P})$. The **Prohorov metric** between two probability distributions P_1, $\mathrm{P}_2 \in \mathcal{M}_1(Z)$ is defined by*

$$d_{\mathrm{Pro}}(\mathrm{P}_1, \mathrm{P}_2) := \inf\big\{\varepsilon > 0 : \mathrm{P}_1(A) \leq \mathrm{P}_2(A^\varepsilon) + \varepsilon \ \ \textit{for all } A \in \mathcal{B}(Z)\big\}. \tag{10.14}$$

The Prohorov neighborhood has the property that it has a component reflecting statistical uncertainty (i.e.; the dimensionless term ε) and a component reflecting the occurrence of rounding errors (i.e.; the term δ that, however, is not dimensionless). The Prohorov metric defines a metric on $\mathcal{M}_1(Z)$ and metricizes the weak* topology in $\mathcal{M}_1(Z)$; see Theorem A.4.19 and Theorem A.4.20. The weak* topology of $\mathcal{M}_1$ can also be metricized by other metrics; see Theorem A.4.22 for the bounded Lipschitz metric. For the special case $Z = \mathbb{R}$, the Lévy metric (see Problem 10.4) is often easier to use than the Prohorov metric because the Lévy metric is based on cumulative distribution functions.

Let (Z, τ_Z) and (W, τ_W) be Polish spaces with complete metrics d_Z and d_W that metricize the weak* topologies. Define a metric on Z^n by $d_{Z^n}(z, z') =$

$\max_{1\le i\le n} d_Z(z_i, z_i')$, where $z = (z_1, \ldots, z_n) \in Z^n$ and $z' = (z_1', \ldots, z_n') \in Z^n$. A data set consisting of n data points $z_i \in Z$, $i = 1, \ldots, n$, will be denoted by D_n, and the corresponding empirical measure will be denoted by D_n. Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables, each with probability distribution $\mathrm{P} \in \mathcal{M}_1(Z)$. We will denote the empirical distribution of a random sample from P of size n by P_n because it will often be important in this section to be precise which distribution generated the data set. The set of all empirical distributions based on n points is denoted by $\mathcal{M}_{1n}(Z) = \left\{\frac{1}{n}\sum_{i=1}^n \delta_{z_i} : z_i \in Z, i = 1, \ldots, n\right\}$. Furthermore, let $S_n : Z^n \to W$, $n \in \mathbb{N}$, be a sequence of random functions. The distribution of S_n will be denoted by P_{S_n} if Z_i has distribution P. Often there exists a measurable function $S : \mathcal{M}_1(Z) \to W$ such that $S_n = S(\mathrm{P}_n)$ for all $\mathrm{P}_n \in \mathcal{M}_{1n}(Z)$ and all $n \in \mathbb{N}$. The following definition goes back to Hampel (1968) and was generalized by Cuevas (1988) for Polish spaces.

Definition 10.2. *Let (Z, τ_Z) and (W, τ_W) be Polish spaces and $(S_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions, where $S_n : Z^n \to W$, $S_n(Z_1, \ldots, Z_n) \in W$, and $Z_1, \ldots, Z_n$ are independent and identically distributed according to $\mathrm{P} \in \mathcal{M}_1(Z)$. The sequence $(S_n)_{n\in\mathbb{N}}$ is called **qualitatively robust** at P if*

$$\forall\, \varepsilon > 0\; \exists\, \delta > 0 : \left\{d_{\mathrm{Pro}}(\mathrm{P}, \mathrm{Q}) < \delta \;\Rightarrow\; d_{\mathrm{Pro}}\big(\mathrm{P}_{S_n}, \mathrm{Q}_{S_n}\big) < \varepsilon, \forall\, n \in \mathbb{N}\right\}. \quad (10.15)$$

Qualitative robustness, which is defined as equicontinuity of the distributions of the statistic as the sample size changes, is hence closely related to continuity of the statistic viewed as a function in the weak* topology; see Theorems A.4.26 and A.4.27.

Remark 10.3. The arithmetic mean $S_n(Y_1, \ldots, Y_n) := \frac{1}{n}\sum_{i=1}^n Y_i$ of real-valued random variables can be written as $S_n(Y_1, \ldots, Y_n) = \mathbb{E}_{\mathrm{P}_n}(Y)$, where $\mathrm{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. By the law of large numbers (see Theorems A.4.8 and A.4.9), $\mathbb{E}_{\mathrm{P}_n}(Y)$ converges in probability or almost surely to $\mathbb{E}_{\mathrm{P}}(Y)$ if mild assumptions are satisfied. Furthermore, the mean is the uniformly minimum variance unbiased estimator of the expectation for Gaussian distributions. However, *the mean is neither continuous nor qualitatively robust at any distribution!* This follows from the fact that in every Prohorov neighborhood of every $\mathrm{P} \in \mathcal{M}_1(\mathbb{R})$ there are probability distributions that have an infinite expectation or for which the expectation does not exist. Consider for example the mixture distribution $\mathrm{Q} = (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$, where $\tilde{\mathrm{P}}$ is a Cauchy distribution and $\varepsilon > 0$ is positive but sufficiently small. Hence the mean is extremely sensitive with respect to such violations of the distributional assumption. This non-robustness property of the expectation operator has two consequences.

i) SVMs are in general not qualitatively robust if $X \times Y$ is unbounded.
ii) If $X \times Y$ is unbounded, certain P-integrability conditions for the loss function will be necessary for our results on robustness properties of SVMs given in Sections 10.3 and 10.4. ◁

Qualitative robustness has the disadvantage that it does not offer arguments on how to choose among different qualitative robust procedures. However, this can be done by the following approaches.

Influence Function and Related Measures

Let (X_i, Y_i), $i = 1, \ldots, n$, be independent and identically distributed random variables on some measurable space $(Z, \mathcal{B}(Z))$, for example $Z = X \times Y \subset \mathbb{R}^d \times \mathbb{R}$. We will consider a function S that assigns to every probability distribution P an element $S(\mathrm{P})$ of a given Banach space E. In the case of SVMs, we have $E = H$ and $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ or $E = \mathbb{R}$ and $S(\mathrm{P}) = \inf_{f \in H} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$.

Recall that qualitative robustness is related to *equicontinuity* of the sequence of estimators $(S_n)_{n \in \mathbb{N}}$ with respect to the Prohorov distance of the corresponding probability distributions. In contrast, the influence function proposed by Hampel (1968, 1974) is related to *differentiation* of S. Denote the Dirac distribution at z by δ_z.

Definition 10.4. *The* ***influence function*** $\mathrm{IF} : Z \to E$ *of* $S : \mathcal{M}_1(Z) \to E$ *at a point* z *for a distribution* $\mathrm{P} \in \mathcal{M}_1(Z)$ *is given by*

$$\mathrm{IF}(z; S, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{S\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\big) - S(\mathrm{P})}{\varepsilon} \tag{10.16}$$

in those $z \in Z$ *where the limit exists.*

If the influence function $\mathrm{IF}(z; S, \mathrm{P})$ exists and is continuous and linear, then the function $S : \mathrm{P} \mapsto S(\mathrm{P})$ is Gâteaux differentiable in the direction of the mixture distribution $\mathrm{Q} := (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$; see also Figure 10.2. Note that Gâteaux differentiation is weaker than Hadamard differentiation (or compact differentiation) and Fréchet differentiation; see Averbukh and Smolyanov (1967, 1968), Fernholz (1983), and Rieder (1994) for details.

The influence function has the interpretation that it measures the impact of an (infinitesimal) small amount of contamination of the original distribution P in the direction of a Dirac distribution located in the point z on the theoretical quantity of interest $S(\mathrm{P})$. Therefore, it is desirable that a statistical method has a *bounded* influence function. If different methods have a bounded influence function, the one with a lower bound is considered to be more robust.

If S fulfills some regularity conditions such as Fréchet differentiability (see Clarke, 1983, 1986; Bednarski *et al.*, 1991), it can be linearized near P in terms of the influence function via

$$S(\mathrm{P}^*) = S(\mathrm{P}) + \int \mathrm{IF}(z; S, \mathrm{P})\, d(\mathrm{P}^* - \mathrm{P})(z) + \ldots\ ,$$

where P^* is a probability measure in a neighborhood of P. According to Huber (1981, p. 72), *"it is not enough to look at the influence function at the model* *distribution only; we must also take into account its behavior in a neighborhood of the model."* Therefore, we will investigate the influence function of $f_{\mathrm{P},\lambda}$ of SVMs not only at a single fixed distribution P. We will show that the influence function of SVMs can be bounded for large sets of distributions (see Sections 10.3 and 10.4).

Definition 10.5. *The* ***sensitivity curve*** $SC_n : Z \to E$ *of an estimator* $S_n : Z^n \to E$ *at a point* $z \in Z$ *given a data set* $z_1, \dots, z_{n-1}$ *is defined by*

$$\mathrm{SC}_n(z; S_n) = n\big(S_n(z_1, \dots, z_{n-1}, z) - S_{n-1}(z_1, \dots, z_{n-1})\big) . \tag{10.17}$$

The sensitivity curve was proposed by J. W. Tukey and its properties are discussed by Hampel *et al.* (1986, p. 93). The sensitivity curve measures the impact of just one additional data point z on the empirical quantity of interest (i.e., on the estimate S_n).

Consider an estimator S_n defined via $S(\mathrm{D}_n)$, where $\mathrm{D}_n \in \mathcal{M}_1(Z)$ denotes the empirical distribution of the data points $z_1, \dots, z_n$. Denote the empirical distribution of $z_1, \dots, z_{n-1}$ by D_{n-1}. Then we have for $\varepsilon_n = \frac{1}{n}$ that

$$\mathrm{SC}_n(z; S_n) = \frac{S\big((1-\varepsilon_n)\mathrm{D}_{n-1} + \varepsilon_n \delta_z\big) - S(\mathrm{D}_{n-1})}{\varepsilon_n} . \tag{10.18}$$

Therefore, the sensitivity curve can be interpreted as a finite-sample version of the influence function. Let us consider for illustration purposes a univariate parametric location problem and a data set $z_i = (x_i, y_i)$, where $x_i = 1$ for $1 \le i \le n$. Figure 10.3 shows the sensitivity curve of the mean and that of the median for a univariate parametric location problem where P is set to the standard Gaussian distribution with Lebesgue density $f(y) = (2\pi)^{-1/2} e^{-y^2/2}$, $y \in \mathbb{R}$. It is obvious that the impact of a single extreme value y_i on the estimated location parameter increases linearly with $|y_i| \to \infty$ if the mean is used but that the median has a *bounded* sensitivity curve. Hence the median is more robust than the mean with respect to the sensitivity curve.

The following notion of (unstandardized) gross error sensitivity allows to compare the robustness of different statistical methods.

Definition 10.6. *Let E be a Banach space with norm $\|\cdot\|_E$. The* ***gross error sensitivity*** *of a function $S : \mathcal{M}_1(Z) \to E$ at a probability distribution* P *is defined by*

$$\gamma_u^*(S, \mathrm{P}) := \sup_{z \in Z} \|\mathrm{IF}\big(z; S, \mathrm{P}\big)\|_E \tag{10.19}$$

if the influence function exists.

Maxbias

Of theoretical as well as practical importance is the notion of maxbias, which measures the maximum bias $S(\mathrm{Q}) - S(\mathrm{P})$ within a neighborhood of probability distributions Q near P. There are several ways to define a neighborhood of P by using appropriate metrics on $\mathcal{M}_1(Z)$. Besides the Prohorov metric, the so-called contamination neighborhood, defined below, is quite common in robust statistics, although such neighborhoods do not metricize the weak* topology on $\mathcal{M}_1(Z)$.

Definition 10.7. *Let* $\mathrm{P} \in \mathcal{M}_1(Z)$ *and* $\varepsilon \in [0, 1/2)$. *A* ***contamination neighborhood*** *or* ***gross error neighborhood*** *of* P *is given by*

$$N_\varepsilon(\mathrm{P}) = \left\{(1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}} : \tilde{\mathrm{P}} \in \mathcal{M}_1(Z)\right\}.$$

Let S *be a function mapping from* $\mathcal{M}_1(Z)$ *into a Banach space* E *with norm* $\|\cdot\|_E$. *The* ***maxbias*** *(or supremum bias) of* S *at the distribution* P *with respect to the contamination neighborhood* $N_\varepsilon(\mathrm{P})$ *is defined by*

$$\mathrm{maxbias}(\varepsilon; S, \mathrm{P}) = \sup_{\mathrm{Q} \in N_\varepsilon(\mathrm{P})} \|S(\mathrm{Q}) - S(\mathrm{P})\|_E .$$

A contamination neighborhood has several nice properties. *(i)* It allows a good interpretation because it contains mixture distributions Q with respect to P and some other distribution $\tilde{\mathrm{P}}$ specifying the type of contamination. This can be seen as follows. Fix $\varepsilon \in [0, 1]$. Define n i.i.d. Bernoulli distributed random variables $\xi_1, \ldots, \xi_n$ such that $\varepsilon = \mathrm{P}(\xi_i = 1) = 1 - \mathrm{P}(\xi_i = 0)$. Then define n i.i.d. random functions $\zeta_1, \ldots, \zeta_n$, each with probability distribution P. Define also n i.i.d. random functions $\zeta_1^*, \ldots, \zeta_n^*$ each with probability distribution $\tilde{\mathrm{P}}$. It follows that the random functions

$$Z_i := \begin{cases} \zeta_i, & \text{if } \xi_i = 0, \\ \zeta_i^*, & \text{if } \xi_i = 1, \qquad 1 \le i \le n, \end{cases}$$

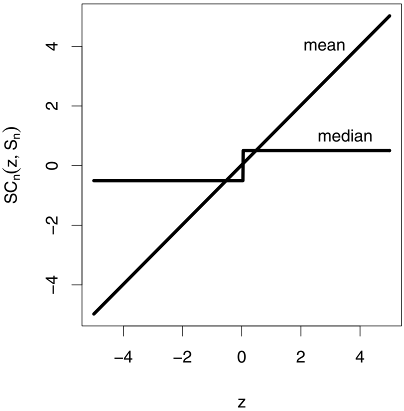

Fig. 10.3. Sensitivity curve of mean and median for a simulated data set with $n = 100$ data points generated from the standard Gaussian distribution $N(0, 1)$.

are i.i.d. each with distribution $\mathrm{Q} = (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$. If $\varepsilon \in [0, 0.5)$, we therefore expect that the pattern described by the majority of n data points generated by Q will follow P and the expected percentage of outliers with respect to P is at most ε. *(ii)* The contamination neighborhood is related to the influence function; see Definition 10.4. *(iii)* It is often easier to deal with this set of distributions than with other neighborhoods.

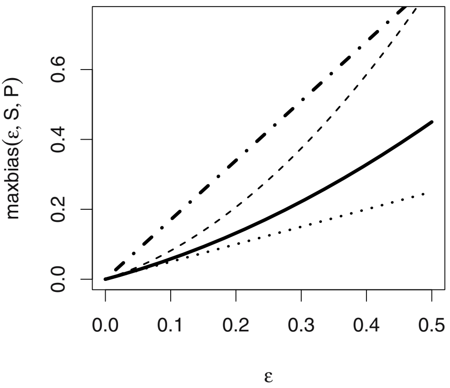

Fig. 10.4. Sketch of relationships between influence function, bias, and maxbias. Define $\mathrm{Q}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$, $\varepsilon \in [0, 0.5]$. Dotted line: $\varepsilon\,\|\mathrm{IF}(z; S, \mathrm{P})\|_E$; solid line: $\|S(\mathrm{Q}_\varepsilon) - S(\mathrm{P})\|_E$; dashed line: $\mathrm{maxbias}(\varepsilon; S, \mathrm{P})$; dotdashed line: linear upper bound for $\mathrm{maxbias}(\varepsilon; S, \mathrm{P})$.

Figure 10.4 illustrates the relationships between influence function, bias, and maxbias. Consider a mixture distribution $\mathrm{Q}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$, $\varepsilon \in [0, 0.5]$. The dotted line has the slope $\|\mathrm{IF}(z; S, \mathrm{P})\|_E$ and offers an approximation of the bias $\|S(\mathrm{Q}_\varepsilon) - S(\mathrm{P})\|_E$ for small values of ε. The bias is of course below the $\mathrm{maxbias}(\varepsilon; S, \mathrm{P})$. The gross error sensitivity $\gamma_u^*(S, \mathrm{P})$ offers—in contrast to $\|\mathrm{IF}(z; S, \mathrm{P})\|_E$—a *uniform* slope (i.e., the slope is valid for all points $z \in X \times Y$). The dotdashed line gives an upper bound for the maxbias where the bound is linear in ε. In the following two sections such quantities will be derived for SVMs for classification and regression.

Breakdown Points

Breakdown points measure the worst-case behavior of a statistical method. The following definition was proposed by Donoho and Huber (1983).

Definition 10.8. *Let E be a Banach space and $D_n = \{z_1, \ldots, z_n\}$ be a data set with values in $Z \subset \mathbb{R}^d$. The* ***finite-sample breakdown point*** *of an E-valued statistic $S(D_n)$ is defined by*

$$\varepsilon_n^*(S, D_n) = \max\left\{\frac{m}{n} \; : \mathrm{Bias}(m; S, D_n) \text{ is finite}\right\},$$

where

$$\mathrm{Bias}(m; S, D_n) = \sup_{D_n'} \|S(D_n') - S(D_n)\|_E$$

and the supremum is over all possible samples D_n' that can be obtained by replacing any m *of the original data points by* arbitrary *values in Z.*

Remark 10.9. It is possible to define a variant of the breakdown point via the maxbias by

$$\varepsilon^*(S, \mathrm{P}) := \sup\{\varepsilon > 0 : \; \mathrm{maxbias}(\varepsilon; S, \mathrm{P}) < \infty\}\,. \tag{10.20}$$

There are variants of the asymptotic breakdown point where the Prohorov metric is replaced by other metrics (e.g., Lévy, bounded Lipschitz, Kolmogorov, total variation). One can also define an asymptotic breakdown point based on gross error neighborhoods. ◁

10.3 Robustness of SVMs for Classification

In this section, we consider robustness properties of classifiers

$$S(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg \inf_{f\in H} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f),$$

where L is a margin-based loss function; see Definition 2.24. Throughout this section, we define $Y := \{-1, +1\}$. First, we give sufficient conditions for the existence of the influence function. Then we show that the influence function is bounded under weak conditions. Most of our results in this section are valid for *any* distribution $\mathrm{P} \in \mathcal{M}_1(X \times Y)$, where $X = \mathbb{R}^d$, $d \in \mathbb{N}$. Therefore, they are also valid for the special case of empirical distributions $\mathrm{D} = \mathrm{D}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$; i.e.; for any given data set consisting of n data points and for the empirical regularized risks defined by

$$S(\mathrm{D}) := f_{\mathrm{D},\lambda} = \arg \inf_{f\in H} \mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f)\,.$$

The proofs given in this and the following section make use of some general results from functional analysis and probability theory. Before we formulate the robustness results, let us therefore recall some basic notions from calculus in (infinite-dimensional) Banach spaces (see Lemma A.5.15). Let $G : E \to F$ be a function between two Banach spaces E and F. Then 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 $\varphi(x)/\|x\| \to 0$ for $x \to 0$ such that

$$G(x_0 + x) - G(x_0) \;=\; Ax + \varphi(x) \tag{10.21}$$

for all $x \in E$. It turns out that A is uniquely determined by (10.21). Hence we write $G'(x) := \frac{\partial G}{\partial E}(x) := A$. The function G is called *continuously differentiable* if the function $x \mapsto G'(x)$ exists on E and is continuous. Analogously we define continuous differentiability on open subsets of E.

We also have to recall the notion of *Bochner integrals*; see Section A.5.4. We restrict attention to the reproducing kernel Hilbert space H since this is the only space we need in this section. Let H be a separable RKHS of a bounded, measurable kernel k on X with canonical feature map $\Phi : X \to H$; i.e.; $\Phi(x) = k(x, \cdot)$. Note that Φ is measurable due to Lemma 4.25. Furthermore, let $\mathrm{P} \in \mathcal{M}_1$ and $h : Y \times X \to \mathbb{R}$ be a measurable and P-integrable function. Then the Bochner integral $\mathbb{E}_{\mathrm{P}} h(Y, X)\Phi(X)$ is an element of H. In our special situation, we can also interpret this integral as an element of the dual space H' by the Fréchet-Riesz Theorem A.5.12; i.e.; $\mathbb{E}_{\mathrm{P}} h(Y, X)\Phi(X)$ acts as a bounded linear functional on H via $w \mapsto \langle \mathbb{E}_{\mathrm{P}} h(Y, X)\Phi(X), w\rangle$. Finally, we will consider Bochner integrals of the form $\mathbb{E}_{\mathrm{P}} h(Y, X)\langle \Phi(X), \cdot\rangle\Phi(X)$ that define bounded linear operators on H by the function $w \mapsto \mathbb{E}_{\mathrm{P}} h(Y, X)\langle \Phi(X), w\rangle\Phi(X)$. We sometimes write $L \circ f$ instead of $L(Y, f(X))$ and Φ instead of $\Phi(X)$ to shorten the notation if misunderstandings are unlikely. We use this kind of notation also for derivatives of L. The Dirac distribution in z is denoted by δ_z.

Existence of the Influence Function

We can now establish our first two results for smooth margin-based loss functions and a bounded continuous kernel. The first theorem covers, for example, the Gaussian RBF kernel.

Theorem 10.10. *Let $Y = \{-1, +1\}$ and $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable margin-based loss function with representing function φ. Furthermore, let $X \subset \mathbb{R}^d$ be a closed subset, H be an RKHS of a bounded continuous kernel on X with canonical feature map Φ, and $\mathrm{P} \in \mathcal{M}_1$. Then the influence function of $S(\mathrm{P}) := f_{\mathrm{P},\lambda}$ exists for all $z = (x, y) \in X \times Y$ and is given by*

$$\mathrm{IF}(z; S, \mathrm{P}) = \mathbb{E}_{\mathrm{P}}[\varphi'\bigl(Y f_{\mathrm{P},\lambda}(X)\bigr)K^{-1}\Phi(X)] - \varphi'\bigl(y f_{\mathrm{P},\lambda}(x)\bigr)K^{-1}\Phi(x)\,, \quad (10.22)$$

where $K : H \to H$ defined by $K = 2\lambda\,\mathrm{id}_H + \mathbb{E}_{\mathrm{P}}\varphi''\bigl(Y f_{\mathrm{P},\lambda}(X)\bigr)\langle \Phi(X), \cdot\rangle\Phi(X)$ denotes the Hessian of the regularized risk.

Proof. Our analysis relies heavily on the function $G : \mathbb{R} \times H \to H$ defined by

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z}\varphi'\bigl(Y f(X)\bigr)\Phi(X) \quad (10.23)$$

and $K = \frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$. Note that for $\varepsilon \notin [0, 1]$ the H-valued expectation is with respect to a signed measure; see Definition A.3.2.

Let us first recall that the solution $f_{\mathrm{P},\lambda}$ exists by Theorem 5.2 and Corollary 5.3. Additionally, we have $\|f_{\mathrm{P},\lambda}\|_H \leq \bigl(\varphi(0)/\lambda\bigr)^{1/2}$, where $\varphi : \mathbb{R} \to [0,\infty)$ is the function representing L by $L(y,t) = \varphi(yt)$, $y \in Y$, $t \in \mathbb{R}$. By using Lemma 2.21 and (5.7), we obtain

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)}{\partial H}, \qquad \varepsilon \in [0,1]. \tag{10.24}$$

Furthermore, the function $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)$ is convex for all $\varepsilon \in [0,1]$, and (10.24) shows that we have $G(\varepsilon, f) = 0$ if and only if $f = f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}$. 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

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

For the existence of $\varepsilon \mapsto f_\varepsilon$, we have to check by the implicit function theorem in Banach spaces (see Theorem A.5.17) that G is continuously differentiable and that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible. Let us start with the first: an easy computation shows

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathrm{P}}\varphi'\bigl(Yf(X)\bigr)\Phi(X) + \mathbb{E}_{\delta_z}\varphi'\bigl(Yf(X)\bigr)\Phi(X)\,. \tag{10.25}$$

Note that $\varphi'' \circ f$ is bounded because φ'' is continuous and $f \in H$ is bounded. Similar to (10.24), we thus find

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda \,\mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z}\varphi''\bigl(Yf(X)\bigr)\langle\Phi(X), \cdot\rangle\Phi(X)\,. \tag{10.26}$$

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; see Theorem A.5.16.

In order to show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible, it suffices to show by the Fredholm Alternative (see Theorem A.5.5) that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is *injective* and that $A : H \mapsto H$ defined by

$$Ag := \mathbb{E}_{\mathrm{P}}\varphi''\bigl(Yf_{\mathrm{P},\lambda}(X)\bigr)g(X)\Phi(X)\,, \qquad g \in H,$$

is a *compact operator*. To show the compactness, we need some facts from measure theory. Since $X \subset \mathbb{R}^d$ is assumed to be closed, it is a Polish space; see the examples listed after Definition A.2.11. Furthermore, Borel probability measures on Polish spaces are regular by Ulam's theorem; see Theorem A.3.15. Therefore, such probability measures can be approximated from inside by compact sets; see Definition A.3.14. In our situation, this means that for all $n \geq 1$ there exists a compact measurable subset $X_n \subset X$ with $\mathrm{P}_X(X_n) \geq 1 - \frac{1}{n}$, where P_X denotes the marginal distribution of P with respect to X. Now define a sequence of operators $A_n : H \to H$ by

$$A_n g := \int_{X_n \times Y} \varphi''\big(y f_{\mathrm{P},\lambda}(x)\big) g(x) \Phi(x)\, d\mathrm{P}(x,y), \qquad g \in H.$$

Note that, if X is compact, we can of course choose $X_n = X$, which implies $A_n = A$. Let us now show that all operators A_n are compact.

By the definition of A_n and Theorem A.5.22, there exists a constant $c > 0$ depending on λ, φ'' and k such that for all $g \in B_H$ we have

$$A_n g \in c \cdot \overline{\operatorname{aco} \Phi(X_n)}\,, \tag{10.27}$$

where $\operatorname{aco} \Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$ and the closure is with respect to $\|\cdot\|_H$. This shows that A_n is compact, $n \in \mathbb{N}$. In order to see that the operator A is compact, it therefore suffices to show $\|A_n - A\|_H \to 0$ w.r.t. the operator norm for $n \to \infty$. However, the latter convergence can be easily checked using $\mathrm{P}_X(X_n) \geq 1 - \frac{1}{n}$.

It remains to prove that A is injective. For $g \neq 0$, we find

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

Here the last equality is due to the fact that $B\mathbb{E}_{\mathrm{P}} h = \mathbb{E}_{\mathrm{P}} B h$ for all E-valued functions h and bounded linear operators B due to (A.32). The last inequality is true since the second derivative of a convex and twice-differentiable function is nonnegative. Obviously, the estimate above shows that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda}) = 2\lambda \operatorname{id}_H + A$ is injective.

Finally, we use the equality $\mathrm{IF}(z; S, \mathrm{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0)$ to derive a formula for the influence function, where $\varepsilon \mapsto f_\varepsilon$ is the function implicitly defined by $G(\varepsilon, f) = 0$ such that the implicit function theorem A.5.17 gives

$$\mathrm{IF}(z; S, \mathrm{P}) = -K^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda})\,, \tag{10.28}$$

where $K := \frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$. Now combine (10.28) with (10.25) and (10.26). □

The next remark follows immediately from Theorem 10.10.

Remark 10.11. If the assumptions of Theorem 10.10 are fulfilled, then a convex margin-based loss function with φ' and φ'' bounded in combination with a bounded continuous kernel yields a bounded influence function of $f_{\mathrm{P},\lambda}$. Hence, the class of Lipschitz-continuous loss functions is of primary interest from the viewpoint of robust statistics. SVMs based on the logistic loss yield a bounded influence function, if k is bounded and continuous. SVMs based on the least squares loss or the AdaBoost loss yield unbounded influence functions.

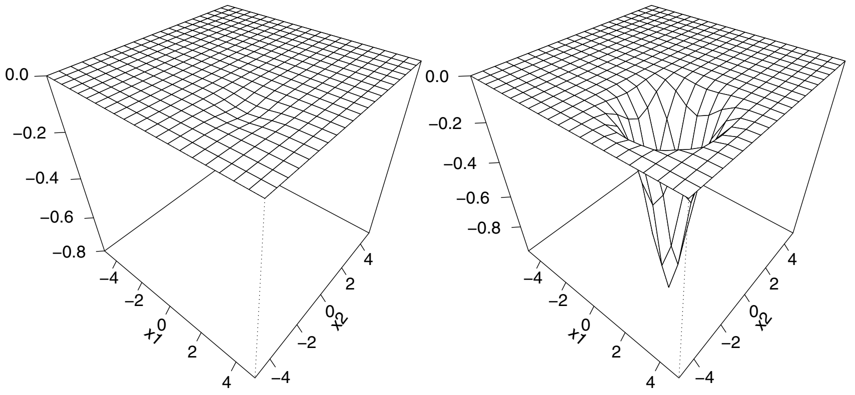

Fig. 10.5. Plot of the function $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ for $L = L_{\text{c-logist}}$, where the point mass contamination is δ_z, $z := (x, y) = (2, -2, y)$ and $\mathrm{P}(Y = +1 \mid X = x) = 0.982$. Left subplot: $y = +1$. Right subplot: $y = -1$.

Remark 10.12. The influence function derived in Theorem 10.10 depends on the point $z = (x, y)$, where the point mass contamination takes place only via the term

$$\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)\,. \tag{10.29}$$

This function is illustrated in Figure 10.5 for the special case of kernel logistic regression (i.e.; $L = L_{\text{c-logist}}$) in combination with a Gaussian RBF kernel and $\mathrm{P}(Y = 1|X = x) = 1/\big(1 + e^{-f(x)}\big)$, $f(x) := -x_1 + x_2$, $(x_1, x_2) \in \mathbb{R}^2$. The left subplot clearly shows that the quantity $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ is approximately zero for this combination of L and k if the highly probable value $z = (x, y) = (2, -2, 1)$ is considered. Of special interest is the right subplot, which shows that the improbable value $z = (x, y) = (2, -2, -1)$ affects the influence function via the quantity $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ only in a *smooth, bounded, and local* manner. Smoothness is achieved by choosing L and k continuous and differentiable. Boundedness is achieved by using a bounded kernel in combination with a loss function having a bounded first derivative φ'. A local impact instead of a more global impact due to a Dirac distribution is achieved by an appropriate bounded and non-linear kernel such as the Gaussian RBF kernel. Note that polynomial kernels with $m \geq 1$ are unbounded if $X = \mathbb{R}^d$. $\triangleleft$

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

Corollary 10.13. *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable margin-based loss function. Furthermore, let $X \subset \mathbb{R}^d$ be* *compact, H be an RKHS of a continuous kernel on X, and* $\mathrm{P} \in \mathcal{M}_1$. *Then the influence function of $f_{\mathrm{P},\lambda}$ exists for all $z \in X \times Y$.*

Proof. Every compact subset of $\mathbb{R}^d$ is closed, and continuous kernels on compact subsets are bounded. Hence the assertion follows directly from Theorem 10.10. □

Remark 10.14. By a straightforward modification of the proof of Corollary 10.13, we actually find that the special Gâteaux derivative of $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ exists for *every* direction; i.e.,

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

exists for all $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ provided that the assumptions of Theorem 10.10 hold. This is interesting from the viewpoint of applied statistics because a point mass contamination is just one kind of contamination that can occur in practice. ◁

Bounds for the Bias and the Influence Function

As explained in Section 10.2, a desirable property of a robust statistical method $S(\mathrm{P})$ is that S has a bounded influence function. In this section, we investigate whether it is possible to obtain an SVM $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ with a bounded influence function where the bound is *independent* of $z \in X \times Y$ and $\mathrm{P} \in \mathcal{M}_1$. We will also consider the question of whether the sensitivity curve or the maxbias can be bounded in a similar way.

For the formulation of our results, we need to recall that the norm of total variation of a signed measure μ on a Banach space E is defined by

$$\|\mu\|_{\mathcal{M}} := |\mu|(E) := \sup \left\{ \sum_{i=1}^{n} |\mu(E_i)| : E_1, \dots, E_n \text{ is a partition of } E \right\} .$$

The following theorem bounds the difference quotient in the definition of the influence function for classifiers based on $f_{\mathrm{P},\lambda}$. For practical applications, this result is especially important because it also gives an upper bound for the bias in gross error neighborhoods. In particular, it states that the influence function of such classifiers is uniformly bounded whenever it exists and that the sensitivity curve is uniformly bounded, too.

Theorem 10.15. *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex margin-based loss function. Furthermore, let $X \subset \mathbb{R}^d$ be closed and H be an RKHS of a bounded, continuous kernel k with canonical feature map $\Phi : X \to H$.*

i) For all $\lambda > 0$, there exists a constant $c(L, k, \lambda) > 0$ explicitly given by (10.33) such that for all distributions $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ *we have*

$$\left\| \frac{f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}}{\varepsilon} \right\|_H \le c(L, k, \lambda) \, \|\mathrm{Q} - \mathrm{P}\|_{\mathcal{M}} , \quad \varepsilon > 0 . \qquad (10.30)$$

ii) An upper bound for the maxbias of $f_{\mathrm{P},\lambda}$ is given by

$$\operatorname{maxbias}(\varepsilon; f, \mathrm{P}) \le \varepsilon\, c(L,k,\lambda) \sup_{\mathrm{Q}\in\mathcal{M}_1} \|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}} \le 2c(L,k,\lambda)\cdot\varepsilon\,,$$

where $\varepsilon \in (0, 1/2)$.

iii) If the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists, then it is bounded by

$$\|\mathrm{IF}(z; S, \mathrm{P})\|_H \le 2c(L,k,\lambda)\,, \qquad z \in X \times Y\,, \tag{10.31}$$

and the gross error sensitivity is bounded by

$$\gamma_u^*(S, \mathrm{P}) \le 2c(L,k,\lambda)\,. \tag{10.32}$$

Proof. i). Recall that every convex function on $\mathbb{R}$ is locally Lipschitz-continuous (see Lemma A.6.5). Combining Lemma 2.25 with $\mathcal{R}_{L,\mathrm{P}}(0) = \varphi(0) < \infty$, where φ satisfies $L(y,t) = \varphi(yt)$ for all $y \in Y$ and $t \in \mathbb{R}$, shows that the assumptions of Corollary 5.10 are fulfilled. Let us define $B_\lambda := \|k\|_\infty \sqrt{\mathcal{R}_{L,\mathrm{P}}(0)/\lambda}$, where $\|k\|_\infty := \sup_{x\in X}\sqrt{k(x,x)} < \infty$. Let us fix $\mathrm{P} \in \mathcal{M}_1$. Then Corollary 5.10 guarantees the existence of a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that for all $\tilde{\mathrm{P}} \in \mathcal{M}_1$ we have

$$\|f_{\mathrm{P},\lambda} - f_{\tilde{\mathrm{P}},\lambda}\|_H \le \lambda^{-1}\|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\tilde{\mathrm{P}}} h\Phi\|_H\,.$$

Now define $\tilde{\mathrm{P}} = (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$, where $\mathrm{Q} \in \mathcal{M}_1$. Let $|L_{|Y\times[-c,c]}|_1$ denote the Lipschitz constant of L restricted to $Y \times [-c, c]$, $c > 0$. Hence, we have

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

where

$$c(L,k,\lambda) = \lambda^{-1}\,\|k\|_\infty\,|L_{|Y\times[-B_\lambda,B_\lambda]}|_1\,. \tag{10.33}$$

We obtain (10.30). The parts *ii)* and *iii)* follow from $\|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}} \le 2$. □

Note that Theorem 10.15 applies to almost all margin-based loss functions of practical interest because differentiability of φ is not assumed. Special cases are the loss functions hinge, kernel logistic, modified Huber, least squares, truncated least squares, and AdaBoost.

Remark 10.16. The preceding theorem also gives uniform bounds for Tukey's sensitivity curve. Consider the special case where P is equal to the empirical distribution of $(n-1)$ data points (i.e.; $\mathrm{D}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\delta_{(x_i,y_i)}$) and that Q is equal to the Dirac measure $\delta_{(x,y)}$ in some point $(x,y) \in X \times Y$. Let $\varepsilon = \frac{1}{n}$. Combining (10.17) and (10.18), we obtain for $\varepsilon > 0$ under the assumptions of Theorem 10.15 the inequality

$$n\|f_{(1-\varepsilon)\mathrm{D}_{n-1}+\varepsilon\delta_{(x,y)},\lambda} - f_{\mathrm{D}_{n-1},\lambda}\|_H \le c(L,k,\lambda)\ \|\delta_{(x,y)} - \mathrm{D}_{n-1}\|_{\mathcal{M}}\,. \quad (10.34)$$

Note that the bound of the bias presented in Corollary 5.10 relies on the Hilbert norm $\big\|\,\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\,\big\|_H$, whereas the bounds of the bias given by Theorem 10.15 and by (10.34) use the norm of total variation $\|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}}$.

Furthermore, Theorem 10.15 shows for empirical distributions D_n that the maxbias of $f_{\mathrm{D}_n,\lambda}$ in an ε-contamination neighborhood $N_\varepsilon(\mathrm{D}_n)$ is at most $2c(L,k,\lambda)\varepsilon$, where $\varepsilon \in (0,1/2)$. In other words, the upper bound for the maxbias increases *at most linearly* with slope $2c(L,k,\lambda)$ with respect to the mixing proportion ε; see also Figure 10.4. $\triangleleft$

Remark 10.17. Consider $\mathrm{P},\mathrm{Q} \in \mathcal{M}_1$ having densities p, q with respect to some σ-finite measure ν. Then, Theorem 10.15 also gives bounds of the influence function and the sensitivity curve in terms of the Hellinger metric $H(\mathrm{P},\mathrm{Q}) = (\int (p^{1/2}-q^{1/2})^2\,d\nu)^{1/2}$ because we have $\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} \le 2\,H(\mathrm{P},\mathrm{Q}) \le 2\,\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}^{1/2}$; see Witting (1985). $\triangleleft$

Note that the *bounds* for the difference quotient and the influence function in Theorem 10.15 converge to infinity if λ converges to 0 and $\|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}} > 0$. However, λ converging to 0 has the interpretation that misclassifications are penalized by constants proportional to $\frac{1}{\lambda}$ tending to ∞. Decreasing values of λ therefore correspond to a decreasing amount of robustness, which is to be expected. The regularizing quantity λ hence has two roles.

i) It controls the *penalization of misclassification errors.*
ii) It controls the *robustness properties* of $f_{\mathrm{P},\lambda}$.

We would like to mention that for the robustness properties of $f_{\mathrm{P},\lambda}$, the ratio $\frac{1}{\lambda}$ plays a role similar to the tuning constant for Huber-type M-estimators. Let us 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 (X_i, Y_i) with some cumulative distribution function $\mathrm{P}\big(Y_i \le y\,|\,X_i = x\big) = F(y-\theta)$, $y \in \mathbb{R}$, where $\theta \in \mathbb{R}$ is unknown. Huber's robust M-estimator with tuning constant $M \in (0,\infty)$ has an influence function proportional to $\varphi_M(z) = \max\big\{-M, \min\{M, z\}\big\}$; see Hampel *et al.* (1986, pp. 104ff.). For all $M \in (0,\infty)$, the influence function is bounded by $\pm M$. However, the bound tends to $\pm\infty$ if $M \to \infty$, and Huber's M-estimator with $M = \infty$ is equal to the non-robust mean, which has an unbounded influence function for Gaussian distributions. Therefore, the quantity $\frac{1}{\lambda}$ for SVMs plays a role similar to the tuning constant M for Huber-type M-estimators. The same argumentation is true for Huber-type M-estimators in classification and in regression.

Christmann and Steinwart (2004) derived some results for the influence function for the more general case where not only $f_{\mathrm{P},\lambda}$ but also a real-valued offset term $b_{\mathrm{P},\lambda}$ must be estimated.

Empirical Results for SVMs

The question arises as to how the theoretical results for $f_{\mathrm{P},\lambda}$ given above are related to empirical results for finite sample sizes.

Let us therefore complement the previous theoretical results by a few numerical results for a training data set with a relatively small sample size. We consider the slightly more general case where an offset term also must be estimated (i.e.; $S(\mathrm{P}) = (f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda}) \in H \times \mathbb{R}$) because many software tools for SVMs use an offset term and because the theoretical results given above were for the case without an intercept term; see Christmann and Steinwart (2004) for additional theoretical results. Let $\mathrm{D} = \mathrm{D}_n$ be the empirical distribution of a training data set with n data points $(x_i, y_i) \in \mathbb{R} \times \{-1, +1\}$, $i = 1, \dots, n$. We will investigate the impact that an additional data point can have on the support vector machine with an offset term $b \in \mathbb{R}$ for pattern recognition. The replacement of one of the n data points can be treated in the same manner. An analogous investigation for the case without offset gave results similar 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 covariances were set to 0.5. The responses y_i were generated from a classical logistic regression model with $\theta = (-1, 1)$ and $b = 0.5$ such that

$$\mathrm{P}(Y = 1|x) = \frac{1}{1 + e^{-(\langle x, \theta \rangle + b)}}, \quad x \in \mathbb{R}^2.$$

We consider two popular kernels: a Gaussian RBF kernel with parameter $\gamma > 0$ 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. The computations were done using the software SVM^{light} developed by Joachims (1999). A cross-validation based on the leave-one-out error for the training data set was carried out by a two-dimensional grid search consisting of 13×10 grid points. As a result of the cross-validation, the tuning parameters for the SVM with RBF kernel were set to $\gamma = 2$ and $\lambda = \frac{1}{4n}$. 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 $\lambda = \frac{1}{2n}$ in the computations for the linear kernel. For $n = 500$, this results in $\lambda = 5 \times 10^{-4}$ for the RBF kernel and $\lambda = 0.001$ for the linear kernel. Please note that such small values of λ will result in relatively large bounds.

Figure 10.6 shows the sensitivity curves of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ if we add a single point $z = (x, y)$ to the original data set, where $x := (x_1, x_2) = (6, 6)$ and $y = +1$. The additional data point has a local and smooth impact on $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ with a peak in a neighborhood of x 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 10.7, where plots of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ are given for the original data set and for the modified data set, which contains the additional data point z. Please note

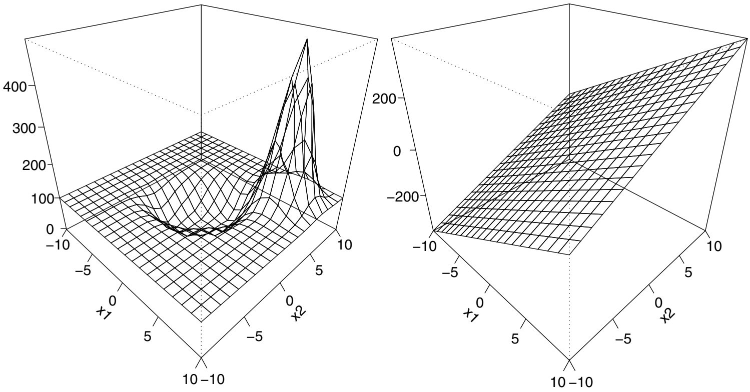

Fig. 10.6. Sensitivity function of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ if the additional data point z is located at $z = (x, y)$, where $x = (6, 6)$ and $y = +1$. Left: RBF kernel. Right: linear kernel.

that the RBF kernel yields $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ approximately equal to zero outside a central region, as almost all data points are lying inside the central region. Comparing the plots of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ 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 $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ for the original data set and for the modified data set based on the SVM with the linear kernel look quite similar, the sensitivity curve is similar to an affine hyperplane that is affected by the value of z. This allows the interpretation that just a single data point can have an impact on $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ estimated by an 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 percentage of classification errors and on the fitted y-value for z. We vary z over a grid in the x-coordinates. Figure 10.8 shows that the percentage of classification errors is approximately constant outside the central region, which 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 affine 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 Gaussian RBF kernel is used; see Figure 10.9. In contrast, 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 $\hat{\theta}$ of the parametric logistic regression model; see Figure 10.10. We vary z over a grid in the

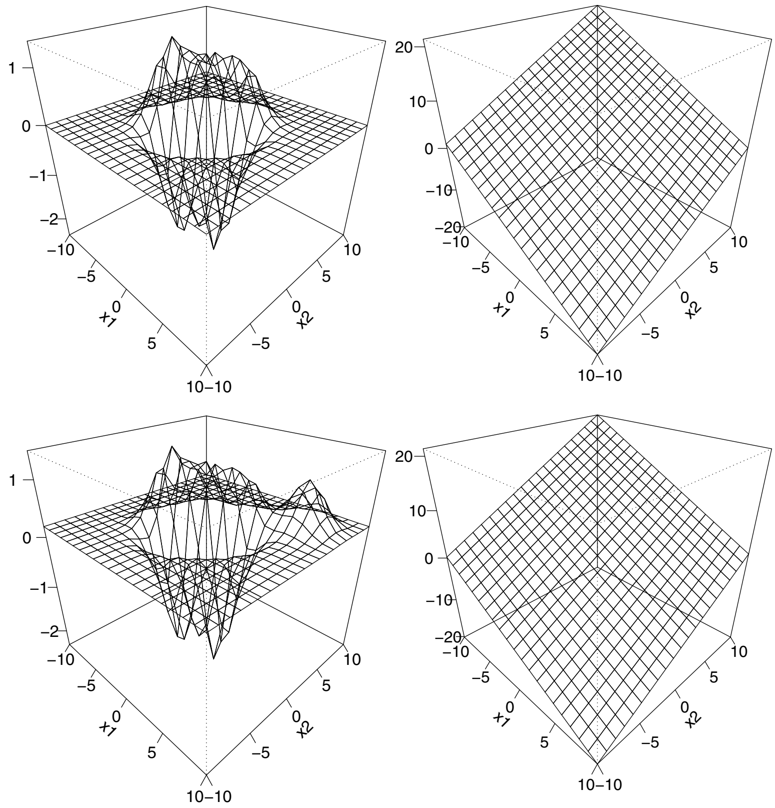

Fig. 10.7. Plot of $f_{D,\lambda} + b_{D,\lambda}$. Upper left: RBF kernel, original data set. Upper right: linear kernel, original data set. Lower left: RBF kernel, modified data set. Lower right: linear kernel, modified data set. The modified data set contains the additional data point $z = (x, y)$, where $x = (6, 6)$ and $y = +1$.

x-coordinates in the same manner as before. As the plots for $\hat{\theta}_1$ and $\hat{\theta}_2$ look very similar, we only show the latter. Note that the axes are not identical in Figure 10.10 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 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, the sensitivity curve of the offset based on an SVM with an RBF kernel shows a

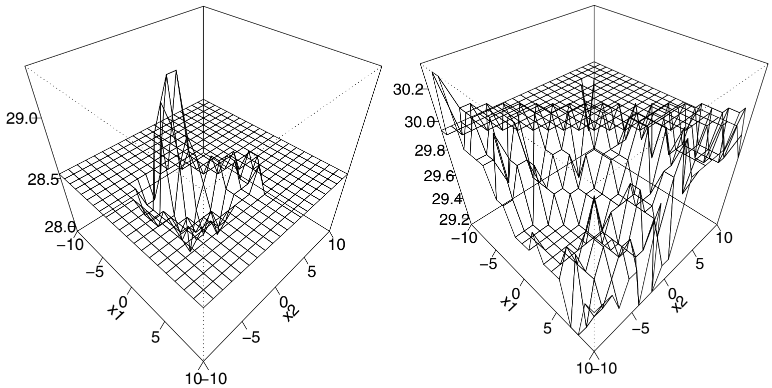

Fig. 10.8. Percentage of classification errors if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

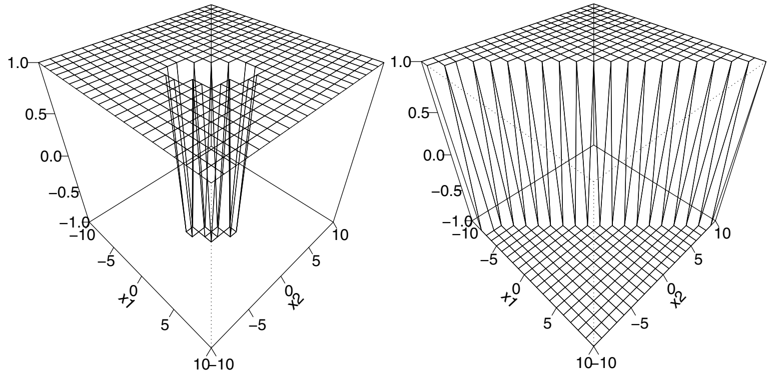

Fig. 10.9. Fitted y-value for a new observation if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

smooth but curved shape outside the region containing the majority of the data points.

10.4 Robustness of SVMs for Regression (*)

In this section, we will investigate the existence and boundedness of the influence function of the mapping $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ for the regression case. Some results for the related robustness measures sensitivity curve, gross error sensitivity, and maxbias will also be given. This section has relationships with

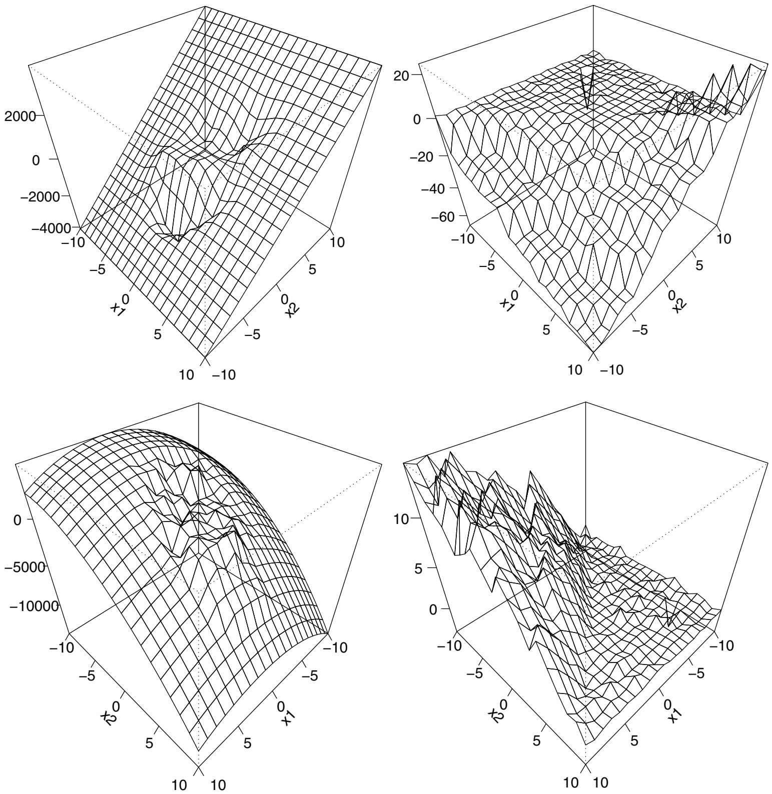

Fig. 10.10. Sensitivity functions for $\hat{\theta}$ and $\hat{\beta}$. Upper left: sensitivity function for $\hat{\theta}_2$, RBF kernel. Upper right: sensitivity function for $\hat{\theta}_2$, linear kernel. Lower left: sensitivity function for $\hat{\beta}$, RBF kernel. Lower right: sensitivity function for $\hat{\beta}$, linear kernel. Note that the axes differ in the four subplots due to improved visibility.

Section 5.3 on the stability of infinite sample versions of SVMs. We assume in this section that $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$.

Existence of the Influence Function

The first result shows that the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists if the loss function is convex and twice continuously differentiable and if the kernel is bounded and continuous. The proof of this result is based on the application of the implicit function theorem A.5.17. In contrast to the proof of Theorem 10.10 for the classification case, we are now usually faced with *unbounded* label sets Y such that we need some additional arguments. In particular, the relationship between the growth type of the loss function and the tail behavior of the distribution P will turn out to be important. This was one reason why we investigated Nemitski losses in Chapter 2.

Theorem 10.18. *Let $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be closed, $\mathrm{P} \in \mathcal{M}_1$, H be an RKHS of a bounded continuous kernel k on X with canonical feature map $\Phi : X \to H$, and $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex* P*-integrable Nemitski loss function that has partial derivatives $L' = \partial_2 L$ and $L'' = \partial_{22} L$ such that $|L'|$ and L'' are* P*-integrable Nemitski loss functions. Then the influence function of $f_{\mathrm{P},\lambda}$ exists for all $z := (x, y) \in X \times Y$ and we have*

$$\mathrm{IF}(z; S, \mathrm{P}) = \mathbb{E}_{\mathrm{P}} L'\bigl(Y, f_{\mathrm{P},\lambda}(X)\bigr) K^{-1} \Phi(X) - L'\bigl(y, f_{\mathrm{P},\lambda}(x)\bigr) K^{-1} \Phi(x), \quad (10.35)$$

where $K : H \to H$, $K = 2\lambda \,\mathrm{id}_H + \mathbb{E}_{\mathrm{P}} L''\bigl(Y, f_{\mathrm{P},\lambda}(X)\bigr) \langle \Phi(X), \cdot \rangle \Phi(X)$ denotes the Hessian of the regularized risk.

Proof. Define the function $G : \mathbb{R} \times H \to H$ by

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

for all $\varepsilon \in \mathbb{R}$, $f \in H$. Let us first check that G is well-defined. Since $\|k\|_\infty < \infty$, we have $\|f\|_H < \infty$ for all $f \in H$. As in the proof of Lemma 2.17, we get $\mathbb{E}_{\mathrm{P}} |L'(Y, f(X))| < \infty$ for all $f \in H$. Note that $\sup_{x \in X} \|\Phi(x)\|_H = \|k\|_\infty < \infty$. Therefore, the H-valued integral used in the definition of G is defined for all $\varepsilon \in \mathbb{R}$ and all $f \in H$. Note that for $\varepsilon \notin [0, 1]$ the H-valued integral is with respect to a signed measure. Now we obtain

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P} + \varepsilon \delta_z, \lambda}}{\partial H}(f), \quad \varepsilon \in [0, 1]; \quad (10.36)$$

see Lemma 2.21 and the discussion at the beginning of Section 5.2. Since the mapping $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P} + \varepsilon \delta_z, \lambda}(f)$ is convex and continuous for all $\varepsilon \in [0, 1]$, equation (10.36) shows that we have

$$G(\varepsilon, f) = 0 \quad \Longleftrightarrow \quad f = f_{(1-\varepsilon)\mathrm{P} + \varepsilon \delta_z, \lambda}$$

for such values of 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 $\varepsilon \in (-\delta, \delta)$ because the existence of such a function guarantees

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

To this end, recall that the implicit function theorem A.5.17 for Banach spaces guarantees the existence of this function provided that *(i)* G is continuously differentiable and that *(ii)* $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible.

Let us start with part *(i)*. A straightforward calculation shows that

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathrm{P}} L'\bigl(Y, f(X)\bigr) \Phi(X) + L'\bigl(y, f(x)\bigr) \Phi(x), \quad (10.37)$$

and a computation slightly more involved than in the proof of Lemma 2.21 gives

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda \, \mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z} L''\big(Y, f(X)\big)\langle \Phi(X), \cdot \rangle \Phi(X) =: K. \quad (10.38)$$

In order to prove that $\frac{\partial G}{\partial \varepsilon}$ is continuous, we fix $\varepsilon \in \mathbb{R}$ and a sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n \in H$, $n \in \mathbb{N}$, and $\lim_{n\to\infty} f_n = f \in H$. Since $\|k\|_\infty < \infty$, the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded. By the continuity of L' and because $|L'|$ is a P-integrable Nemitski loss function, there exists a bounded measurable function $g : Y \to \mathbb{R}$ with $L'(y, f_n(x)) \leq L'(y, g(y))$ for all $n \geq 1$ and all $(x, y) \in X \times Y$. For the mapping $v(y) := L(y, g(y))$, $y \in Y$, we get $v \in L_1(\mathrm{P})$, and therefore an application of the dominated convergence theorem A.5.21 for Bochner integrals gives the continuity of $\frac{\partial G}{\partial \varepsilon}$. The continuity of G and $\frac{\partial G}{\partial H}$ can be shown analogously. Using Theorem A.5.16, we conclude that G is continuously differentiable, and *(i)* is shown.

Now let us prove *(ii)*. In order to show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible it suffices to show by the Fredholm Alternative A.5.5 that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is *injective* and that

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

defines a *compact operator* on H.

To show the compactness of the operator A, recall that X and Y are Polish spaces since we assumed that X and Y are closed subsets of $\mathbb{R}^d$ and $\mathbb{R}$, respectively; see the examples listed after Definition A.2.11. Furthermore, Borel probability measures on Polish spaces are regular by Ulam's Theorem A.3.15. Therefore, they can be approximated from inside by compact sets; see Definition A.3.14. Hence there exists a sequence of compact subsets $X_n \times Y_n \subset X \times Y$ with $\mathrm{P}(X_n \times Y_n) \geq 1 - \frac{1}{n}$, $n \in \mathbb{N}$. Let us also define a sequence of operators $A_n : H \to H$ by

$$A_n g := \int_{X_n} \int_{Y_n} L''\big(y, f_{\mathrm{P},\lambda}(x)\big) \, \mathrm{P}(dy|x) \, g(x) \Phi(x) \, d\mathrm{P}_X(x), \quad g \in H. \quad (10.39)$$

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

$$\begin{aligned} h_g(x) &:= \int_{Y_n} L''\big(y, f_{\mathrm{P},\lambda}(x)\big) |g(x)| \, \mathrm{P}(dy|x) \\ &\leq \|k\|_\infty \int_{Y_n} \big|L''\big(y, f_{\mathrm{P},\lambda}(x)\big)\big| \, \mathrm{P}(dy|x) =: h(x), \quad n \in \mathbb{N}, \end{aligned}$$

because L'' is a P-integrable Nemitski loss function. Therefore, we have $h \in L_1(\mathrm{P}_X)$, which implies $h_g \in L_1(\mathrm{P}_X)$ with $\|h_g\|_1 \leq \|h\|_1$ for all $g \in B_H$. Consequently, $d\mu_g := h_g d\mathrm{P}_X$ and $d\mu := h d\mathrm{P}_X$ are finite measures. By Theorem A.5.22, we hence obtain

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

where $\operatorname{aco}\Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$, and the closure is with respect to $\|\cdot\|_H$. Now, using the continuity of Φ, we see that $\Phi(X_n)$ is compact and hence so is the closure of $\operatorname{aco}\Phi(X_n)$. This shows that A_n is a compact operator. In order to see that A is compact, it therefore suffices to show that $\|A_n - A\| \to 0$ with respect to the operator norm for $n \to \infty$. Recalling that the convexity of L implies $L'' \geq 0$, the desired convergence follows from $\mathrm{P}(X_n \times Y_n) \geq 1 - \frac{1}{n}$, $L'' \circ f_{\mathrm{P},\lambda} \in L_1(\mathrm{P})$, and

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

Let us now show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda}) = 2\lambda\, \mathrm{id}_H + A$ is injective. For $g \in H\setminus\{0\}$, we obtain

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

which shows the injectivity and hence *(ii)*.

The implicit function theorem A.5.17 guarantees that the function $\varepsilon \mapsto f_\varepsilon$ is differentiable on $(-\delta, \delta)$ if $\delta > 0$ is small enough. Furthermore, (10.37) and (10.38) yield for $z = (x, y) \in X \times Y$ that

$$\begin{aligned} \mathrm{IF}(z; S, \mathrm{P}) &= \frac{\partial f_\varepsilon}{\partial \varepsilon}(0) = -K^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda}) \\ &= K^{-1}\big(\mathbb{E}_{\mathrm{P}}\big(L'(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)\big)\big) - L'\big(y, f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x). \quad \square \end{aligned}$$

Remark 10.19. Theorem 10.18 contains a tail assumption on P to ensure that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) < \infty$. Recall that $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0, \infty)$ is well-defined and continuous if L is a P-integrable Nemitski loss function of order $p \in [1, \infty)$, see Lemma 2.17. Taking Remark 10.3 into account, a tail assumption on P seems to be quite natural for *unbounded* sets Y. We see three ways to avoid this tail assumption on P, but all of them seem to be unsatisfactory. *(i)* One can restrict attention to bounded sets Y, but this is often unrealistic for regression problems and does not make sense from the viewpoint of robust statistics. *(ii)* One can replace the convex loss function by a bounded non-convex loss function, but in general this yields non-convex risk functions and hence problems with respect to the existence and uniqueness of $f_{\mathrm{P},\lambda}$. Further, the advantage of computational efficiency often vanishes because the numerical problems can turn from convex problems into NP-hard ones. *(iii)* The redefinition of the minimization problem $\inf_{f\in H}\mathcal{R}_{L,\mathrm{P}}(f)+\lambda\|f\|_H^2$ into

$$\inf_{f\in H}\mathbb{E}_{\mathrm{P}}L^*(X,Y,f(X))+\lambda\|f\|_H^2,$$

where $L^*: X\times Y\times\mathbb{R}$, $L^*(x,y,t):=L(x,y,t)-L(x,y,0)$, seems to be helpful for Hölder-continuous loss functions, which includes of course the important class of Lipschitz-continuous loss functions, but $L^*(x,y,f(x))$ can be *negative* for some values of $(x,y,f(x))$. ◁

Remark 10.20. The proof of Theorem 10.18 can be modified in order to replace point mass contaminations δ_z by arbitrary contaminations $\mathrm{Q}\in\mathcal{M}_1$ if L, $|L'|$, and L'' are Q-integrable Nemitski loss functions. ◁

Remark 10.21. From a robustness point of view, one is mainly interested in statistical methods with *bounded influence functions*. It is worth mentioning that Theorem 10.18 not only ensures the existence of the influence function $\mathrm{IF}(z;S,\mathrm{P})$ but also indicates how to guarantee its boundedness. Indeed, (10.35) shows that the only term of the influence function that depends on the point mass contamination δ_z is

$$-L'\big(y,f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x)\,. \tag{10.40}$$

Hence a combination of a bounded continuous kernel with a convex loss function with L' being bounded assures a bounded influence function. Hence, we have, analogous to the results obtained in Section 10.3, that the class of Lipschitz-continuous loss functions are of primary interest from the viewpoint of robust statistics. ◁

The next result gives influence functions for distance-based loss functions.

Corollary 10.22. *Let $X=\mathbb{R}^d$, $Y=\mathbb{R}$, and H be an RKHS of a bounded continuous kernel k on X with canonical feature map $\Phi: X\to H$. Further, let $L: Y\times\mathbb{R}\to[0,\infty)$ be a convex distance-based loss function with representing function $\psi:\mathbb{R}\to[0,\infty)$ having partial derivatives $L'=\partial_2 L$ and $L''=\partial_{22}L$ such that L, $|L'|$, and L'' are* P*-integrable Nemitski loss functions and $\|b\|_{L_1(\mathrm{P})}<\infty$. Then the following statements hold.*

i) The influence function of $f_{\mathrm{P},\lambda}$ exists for all $z:=(x,y)\in X\times Y$ and

$$\mathrm{IF}(z;S,\mathrm{P})=-\mathbb{E}_{\mathrm{P}}\psi'\big(Y-f_{\mathrm{P},\lambda}(X)\big)K^{-1}\Phi(X)+\psi'\big(y-f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x), \tag{10.41}$$

where $K: H\to H$, $K=2\lambda\,\mathrm{id}_H+\mathbb{E}_{\mathrm{P}}\psi''\big(Y-f_{\mathrm{P},\lambda}(X)\big)\langle\Phi(X),\cdot\rangle\Phi(X)$.

ii) The influence function $\mathrm{IF}(z; S, \mathrm{P})$ *is bounded in* z *if* L *is Lipschitz-continuous.*

Proof. Theorem 10.18 yields that $\mathrm{IF}(z; S, \mathrm{P})$ exists and is bounded provided $L'(\cdot, f_{\mathrm{P},\lambda}(x)) : \mathbb{R} \to \mathbb{R}$ is bounded for all $x \in X$. For distance-based loss functions, we hence immediately obtain the assertions. □

Remark 10.23.

i) Let us emphasize that it is *not* sufficient to choose a bounded continuous kernel alone to obtain a bounded influence function: the loss function also must be chosen appropriately.

ii) Corollary 10.22 shows that the SVM based on the least squares loss, which has some nice computational properties, as will be shown in Chapter 11, is a method with an *unbounded* influence function. Therefore, an ad hoc one-step reweighted version of the SVM based on the least squares loss can be interesting from a robustness point of view, see Suykens *et al.* (2002) and Debruyne *et al.* (2007).

iii) In contrast, the logistic loss function provides a robust method with a *bounded* influence function if we use it in combination with a Gaussian RBF kernel. ◁

Example 10.24. For illustration purposes, let us consider a Gaussian RBF kernel and the loss functions $L_{\text{r-logist}}$ and L_{LS}. Note that the partial derivative with respect to the last argument of $L_{\text{r-logist}}$ is bounded, whereas the corresponding partial derivative of L_{LS} is unbounded. We expect by (10.40) and (10.41) that SVMs based on $L_{\text{r-logist}}$ will give more robust results than SVMs using L_{LS}. Figure 10.11 shows that this is indeed true. The small data set listed in Table 10.1 contains 100 data points $(x_i, y_i) \in \mathbb{R}^2$ of a baby's daily milk consumption.[2] The upper subplot shows that both SVMs offer similar fits to the original data set. Now let us assume that the tired father made a typing error when he reported the daily measurement during the night for day number 40: he reported $y_{40} = 7.4$ instead of $y_{40} = .74$, which is an obvious mistake. This single typing error has a strong impact on $f_{\mathrm{D},\lambda}$ based on L_{LS} in a broad interval. On the other hand, the impact of this extreme y-value on $f_{\mathrm{D},\lambda}$ based on $L_{\text{r-logist}}$ is much smaller due to the boundedness of L'. ◁

Bounds for the Influence Function

Unfortunately, Theorem 10.18 and Corollary 10.22 require a twice continuously differentiable loss function and therefore they cannot be used to investigate SVMs based on, for example, the ϵ-insensitive loss, Huber's loss, or the pinball loss function, which are not Fréchet differentiable in one or two

[2] From one of the authors

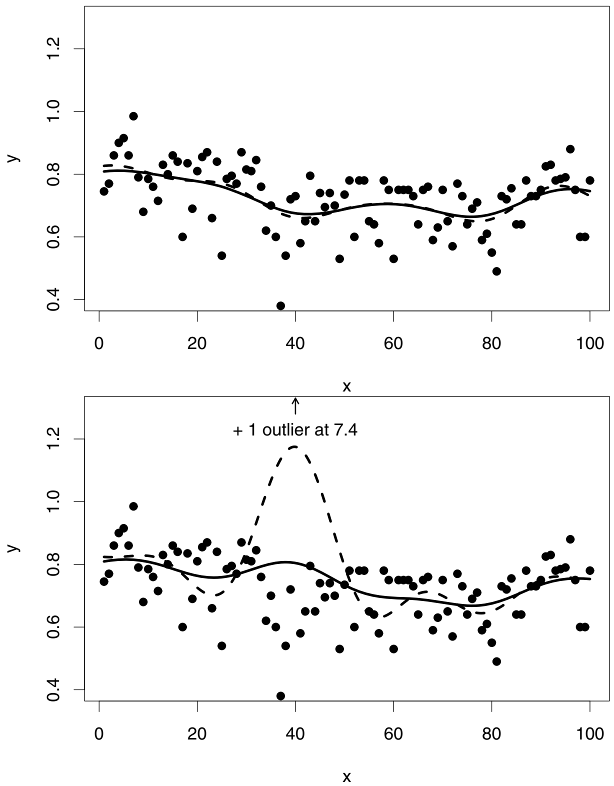

Fig. 10.11. Daily milk consumption. The upper subplot displays the fitted regression curves $f_{\mathrm{D},\lambda}(x)$ for the original data set. The lower subplot displays the fitted regression curves $f_{\mathrm{D},\lambda}(x)$ for the data set containing one extreme value. The curves were fitted using the loss functions $L_{\text{r-logist}}$ (solid) and L_{LS} (dashed), respectively.

points.[3] The next three theorems give bounds for the difference quotient used in the definition of the influence function and apply to convex Nemitski loss functions of some order p. Hence these results partially resolve the problem above for non-Fréchet differentiable loss functions. For practical purposes, the following results may even be more interesting than the results for the influence function because they give bounds for the bias and do not consider an infinitesimally small amount of contamination. Note that the following three results show that *upper bounds for the bias* under gross error contamination models increase *at most linearly* with respect to the mixing proportion ε because the constants c used in the bounds do not depend on ε.

[3] Christmann and Van Messem (2008) derived an analogon to Theorem 10.18 using Bouligand derivatives for SVMs based on non-smooth Lipschitz-continuous losses.

Table 10.1. Data set: daily milk consumption of a baby. The measurements are listed in liters.

Day	Milk	Day	Milk	Day	Milk	Day	Milk	Day	Milk
1	0.745	2	0.770	3	0.860	4	0.900	5	0.915
6	0.860	7	0.985	8	0.790	9	0.680	10	0.785
11	0.760	12	0.715	13	0.830	14	0.800	15	0.860
16	0.840	17	0.600	18	0.835	19	0.690	20	0.810
21	0.855	22	0.870	23	0.660	24	0.840	25	0.540
26	0.785	27	0.795	28	0.770	29	0.870	30	0.815
31	0.810	32	0.845	33	0.760	34	0.620	35	0.700
36	0.600	37	0.380	38	0.540	39	0.720	40	0.740
41	0.580	42	0.650	43	0.795	44	0.650	45	0.740
46	0.695	47	0.740	48	0.700	49	0.530	50	0.735
51	0.780	52	0.600	53	0.780	54	0.780	55	0.650
56	0.640	57	0.580	58	0.780	59	0.750	60	0.530
61	0.750	62	0.750	63	0.750	64	0.730	65	0.640
66	0.750	67	0.760	68	0.590	69	0.630	70	0.750
71	0.650	72	0.570	73	0.770	74	0.730	75	0.640
76	0.690	77	0.710	78	0.590	79	0.610	80	0.550
81	0.490	82	0.730	83	0.720	84	0.755	85	0.640
86	0.640	87	0.780	88	0.730	89	0.730	90	0.750
91	0.825	92	0.830	93	0.780	94	0.785	95	0.790
96	0.880	97	0.750	98	0.600	99	0.600	100	0.780

Theorem 10.25. *Let* $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ *and* $L : Y \times \mathbb{R} \to [0, \infty)$ *be a convex* P*-integrable and* Q*-integrable Nemitski loss function of order* $p \in [1, \infty)$*. Furthermore, let* k *be a bounded, measurable kernel on* X *with separable RKHS* H *and canonical feature map* $\Phi : X \to H$*. Then, for all* $\lambda > 0$ *and* $\varepsilon \in [0, 1]$*, we have*

$$\left\| f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} \right\|_H \leq c_{\mathrm{P},\mathrm{Q}}\, \varepsilon\,, \tag{10.42}$$

where

$$c_{\mathrm{P},\mathrm{Q}} := \frac{\|b\|_{L_1(\mathrm{P})} + \|b\|_{L_1(\mathrm{Q})} + 2^{2p+1} c B_\lambda^p}{(\lambda\, \mathcal{R}_{L,\mathrm{P}}(0))^{1/2}}\,,$$

$B_\lambda := \|k\|_\infty (\mathcal{R}_{L,\mathrm{P}}(0)/\lambda)^{1/2}$*, and* $b : X \times Y \to [0, \infty)$ *is a function satisfying the Nemitski condition* (2.9)*. If* $\mathrm{Q} = \delta_{(x,y)}$ *and if* $\mathrm{IF}((x, y); S, \mathrm{P})$ *exists, then*

$$\|\mathrm{IF}((x, y); S, \mathrm{P})\|_H \leq c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x, y) \in X \times Y \tag{10.43}$$

and

$$\gamma_u^*(S, \mathrm{P}) \leq \frac{\|b\|_{L_1(\mathrm{P})} + \sup_{(x,y)\in X\times Y} b(x, y) + 2^{2p+1} c B_\lambda^p}{(\lambda\, \mathcal{R}_{L,\mathrm{P}}(0))^{1/2}}\,.$$

Proof. Fix $\varepsilon \in [0, 1]$ and define $\tilde{\mathrm{P}} := (1 - \varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$. By Theorem 5.9, there exists a bounded, measurable function $h : X \times Y \to \mathbb{R}$ independent of ε and Q such that

$$\begin{aligned}\|f_{\mathrm{P},\lambda} - f_{\tilde{\mathrm{P}},\lambda}\|_H &\le \lambda^{-1} \,\|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} h\Phi\|_H \\ &= \varepsilon\,\lambda^{-1} \,\|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi\|_H \\ &\le \varepsilon\,\lambda^{-1} \,\|k\|_\infty \big(\|h\|_{L_1(\mathrm{P})} + \|h\|_{L_1(\mathrm{Q})}\big) \\ &\le \varepsilon\,\lambda^{-1} \,\|k\|_\infty \big(\|b\|_{L_1(\mathrm{P})} + \|b\|_{L_1(\mathrm{Q})} + 2c|4B_\lambda|^p\big)/B_\lambda\,,\end{aligned}$$

where (5.16) is used in the last inequality. This gives the assertion (10.42). The inequality (10.43) follows from the definition of the influence function and the fact that the constant $c_{\mathrm{P},\mathrm{Q}}$ does not depend on the mixture proportion ε. For the special case $\mathrm{Q} = \delta_{(x,y)}$ we have $\|b\|_{L_1(\mathrm{Q})} = b(x,y)$ and hence we obtain bounds for the difference quotient used in the definition of the influence function if we divide the bound by ε. □

If we restrict our attention to distance-based loss functions of growth type $p \ge 1$, and many loss functions used in practice are distance-based, we are able to obtain stronger results.

Theorem 10.26. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p > 1$ with representing function $\psi : \mathbb{R} \to [0,\infty)$ and P, $\mathrm{Q} \in \mathcal{M}_1$ such that L is P-integrable and Q-integrable. Furthermore, let k be a bounded, measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$ and for $\varepsilon \in [0,1]$, we have*

$$\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\|_H \le c_{\mathrm{P},\mathrm{Q}}\,\varepsilon\,, \tag{10.44}$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \tilde{c}_3\,\lambda^{-1}\|k\|_\infty\big(|\mathrm{P}-\mathrm{Q}|_{p-1}^{p-1} + \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}\,\big(1 + \lambda^{(1-p)/2}\|k\|_\infty^{p-1}|\mathrm{P}|_p^{p(p-1)/2}\big)\big)$$

and the constant $\tilde{c}_3 = \tilde{c}_3(L,p,k,\lambda) \ge 0$. If $\mathrm{Q} = \delta_{(x,y)}$ and if the influence function exists, then

$$\|\mathrm{IF}((x,y);S,\mathrm{P})\|_H \le c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x,y) \in X \times Y, \tag{10.45}$$

and the gross error sensitivity fulfills

$$\gamma_u^*(S,\mathrm{P}) \le \sup_{(x,y)\in X\times Y} c_{\mathrm{P},\delta_{(x,y)}}\,. \tag{10.46}$$

Proof. Corollary 5.11 gives the existence of a measurable function $h : X \times Y \to \mathbb{R}$ with

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \frac{\varepsilon}{\lambda}\|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi\|_H\,.$$

As L is a loss function of upper growth type $p > 1$, there exists $c > 0$ such that

$$\lambda\|f_{\mathrm{P},\lambda}\|_H^2 \le \mathcal{R}_{L,\mathrm{P},\lambda}^{reg}(f_{\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P},\lambda}^{reg}(0) \le c\big(|\mathrm{P}|_p^p + 1\big)\,.$$

This yields for some $\tilde{c}_1 > 0$ the inequalities

$$\|f_{\mathrm{P},\lambda}\|_\infty \le \|k\|_\infty \|f_{\mathrm{P},\lambda}\|_H \le (\tilde c_1/\lambda)^{1/2}\|k\|_\infty |\mathrm{P}|_p^{p/2} .$$

Using Corollary 5.11 and (5.28) from its proof, we obtain for some $\tilde c_2 > 0$ that

$$\begin{aligned}|h(x,y)| &\le 4^p c_L \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr\}\\ &\le 4^p c_L \bigl(1 + |y|^{p-1} + |f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr)\\ &\le \tilde c_2 \left(1 + |y|^{p-1} + \lambda^{(1-p)/2}\|k\|_\infty^{p-1}\, |\mathrm{P}|_p^{p(p-1)/2}\right), \quad (x,y) \in X \times Y.\end{aligned}$$

It follows that

$$\begin{aligned}&\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\|_H \;\le\; \varepsilon\,\lambda^{-1}\,\|k\|_\infty\,\mathbb{E}_{|\mathrm{P}-\mathrm{Q}|}|h|\\ &\le \varepsilon\,\tilde c_3\,\lambda^{-1}\|k\|_\infty\bigl(|\mathrm{P}-\mathrm{Q}|_{p-1}^{p-1} + \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}\left(1 + \lambda^{(1-p)/2}\|k\|_\infty^{p-1}|\mathrm{P}|_p^{p(p-1)/2}\right)\bigr),\end{aligned}$$

where $\tilde c_3$ depends on L but not on ε. This gives the assertion in (10.44). The inequalities (10.45) and (10.46) follow immediately from the fact that $c_{\mathrm{P},\mathrm{Q}}$ does not depend on ε. □

Recall that Lipschitz-continuous, distance-based, convex loss functions are of upper growth type $p = 1$ due to Lemma 2.36*ii)*. Such loss functions are therefore *not* covered by the previous theorem, but by the next one. Important special cases of the next theorem are the ϵ-insensitive loss and the pinball loss.

Theorem 10.27. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a Lipschitz-continuous, convex, distance-based loss function with representing function $\psi : \mathbb{R} \to [0,\infty)$, and P, $\mathrm{Q} \in \mathcal{M}_1$ with $|\mathrm{P}|_1 < \infty$ and $|\mathrm{Q}|_1 < \infty$. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$ and $\varepsilon \in [0,1]$, we have*

$$\bigl\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\bigr\|_H \le c_{\mathrm{P},\mathrm{Q}}\,\varepsilon\,,$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}.$$

If $\mathrm{Q} = \delta_{(x,y)}$ and if the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists, then

$$\|\mathrm{IF}((x,y);S,\mathrm{P})\|_H \le c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x,y) \in X \times Y,$$

and the gross error sensitivity fulfills

$$\gamma_u^*(S,\mathrm{P}) \le \lambda^{-1}\,\|k\|_\infty\,|\psi|_1 \sup_{(x,y)\in X\times Y} \|\mathrm{P} - \delta_{(x,y)}\|_{\mathcal{M}} \le 2\lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,.$$

In particular, the H-norm of the sensitivity curve is bounded by a uniform constant independent of the sample size n, independent of (x,y), and valid for any empirical distribution $\mathrm{D}_n \in \mathcal{M}_1$:

$$\|\mathrm{SC}_n((x,y);f_{\mathrm{D}_n,\lambda})\|_H \le 2\lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,, \quad (x,y) \in X \times Y. \qquad (10.47)$$

Proof. Corollary 5.10 guarantees that there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that $\|h\|_\infty \le |L|_{B_\lambda,1} \le |\psi|_1$ and

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \varepsilon\lambda^{-1}\|\mathbb{E}_\mathrm{P} h\Phi - \mathbb{E}_\mathrm{Q} h\Phi\|_H .$$

It follows that

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \varepsilon\lambda^{-1}\|k\|_\infty \mathbb{E}_{|\mathrm{P}-\mathrm{Q}|}|h| \le \varepsilon\lambda^{-1}\|k\|_\infty|\psi|_1 \,\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} .$$

This gives the assertion. □

Remark 10.28.

i) The combination of a Lipschitz-continuous loss function with $|\psi|_1 = 1$ and a Gaussian RBF kernel with $\|k\|_\infty = 1$ is of particular interest for applications. The H-norm of the sensitivity curve is then uniformly bounded by $\frac{2}{\lambda}$ for all data sets and for all points (x, y) by Theorem 10.27.

ii) The previous three theorems also offer bounds for $\mathrm{maxbias}(\varepsilon; S, \mathrm{P})$ over contamination neighborhoods provided attention is restricted to distributions $\mathrm{Q} \in N_\varepsilon(\mathrm{P})$ satisfying the tail conditions given in those theorems; see Problem 10.7. ◁

Comparison of SVMs with M-Estimation (*)

Let us now compare the influence function of SVMs with the influence function of M-estimators in *linear* regression models. A linear regression model assumes that (X_i, Y_i) are independent and identically distributed random variables according to $\mathrm{P} \in \mathcal{M}_1(X \times Y)$, $X = \mathbb{R}^d$, $Y = \mathbb{R}$, with regular conditional distribution such that

$$\mathbb{E}_\mathrm{P}(Y|x) := \int_Y y\, d\mathrm{P}(y|x) = x^\mathsf{T}\theta ,$$

where $\theta \in \Theta := \mathbb{R}^d$ is unknown. Such linear models are quite popular in many areas of applied statistics and in data mining. Obviously, linear regression is a special case of SVMs: we use $\lambda = 0$ in combination with a linear kernel $k(x, x') := \langle x, x' \rangle$, $x, x' \in \mathbb{R}^d$. Let us assume for reasons of simplicity that the scale parameter $\sigma \in (0, \infty)$ of the linear regression model is known, say $\sigma = 1$. Note that $\sigma^2 = \mathrm{Var}_\mathrm{P}(Y|x)$ for Gaussian distributions. The function $S : \mathcal{M}_1 \to \mathbb{R}^d$ corresponding to an M-estimator is the solution of

$$\mathbb{E}_\mathrm{P}\, \eta(X, Y - X^\mathsf{T}S(\mathrm{P}))X = 0 , \tag{10.48}$$

where the odd function $\eta(x, \cdot)$ is continuous for $x \in \mathbb{R}^d$ and $\eta(x, u) \ge 0$ for all $x \in \mathbb{R}^d$, $u \in [0, \infty)$. Almost all M-estimators for linear regression proposed in the literature may be written in the form $\eta(x, u) = \psi\big(v(x)u\big)w(x)$, where $\psi : \mathbb{R} \to \mathbb{R}$ is usually continuous, bounded, and increasing and $w : \mathbb{R}^d \to [0, \infty)$, $v : \mathbb{R}^d \to [0, \infty)$ are weight functions. The functions ψ, w, and v are chosen in advance by the user or may result as solutions to certain robust optimization problems. An important subclass of M-estimators are those of Mallows type, where η is of the form $\eta(x,u) = \psi(u)w(x)$. Note that defining M-estimators as solutions of (10.48) is more general than defining M-estimators via optimization problems. The influence function of $S(\mathrm{P}) = \theta$ in the point $z = (x,y)$ at a distribution $\mathrm{P} \in \mathcal{M}_1$ is given by

$$\mathrm{IF}(z; S, \mathrm{P}) = K^{-1}(\eta, \mathrm{P})\, \eta(x, y - x^{\mathsf{T}} S(\mathrm{P}))\, x \in \mathbb{R}^d\,, \tag{10.49}$$

where $K(\eta, \mathrm{P}) := \mathbb{E}_{\mathrm{P}}\, \eta'(X, Y - X^{\mathsf{T}} S(\mathrm{P}))XX^{\mathsf{T}}$. An important difference between SVMs and M-estimators in linear regression is that $\mathrm{IF}(z; S, \mathrm{P}) \in \mathbb{R}^d$ in (10.49), but $\mathrm{IF}(z; S, \mathrm{P}) \in H$ in (10.35). In other words, the influence function of an M-estimator is only a function if z varies, whereas the influence function of SVMs is an H-valued function already for any fixed value of z. For linear kernels, there exists an isomorphism between the RKHS H and $\mathbb{R}^d$, but this is not true for many other kernels (e.g., for Gaussian RBF kernels).

A comparison of the influence functions for SVMs (see (10.35) and (10.41)) with the influence function of M-estimators given by (10.49) yields that both influence functions nevertheless have a similar structure. The function $K = K(L'', k, \mathrm{P})$ for SVMs and the matrix $K(\eta, \mathrm{P})$ for M-estimation do not depend on z. The terms in the influence functions depending on $z = (x, y)$, where the point mass contamination δ_z occurs, are a product of two factors. The first factors, measuring the outlyingness in the y-direction, are

$-L'(y, f_{\mathrm{P},\lambda}(x))$	for SVMs,
$\psi(v(x)(y - x^{\mathsf{T}}\theta))$	for general M-estimation,
$\psi'(y - f_{\mathrm{P},\lambda}(x))$	for SVMs with distance-based loss, and
$\psi(y - x^{\mathsf{T}}\theta)$	for Mallows type M-estimation.

SVMs based on a distance-based loss function and Mallows type M-estimators use first factors that only depend on the residuals. The second factors are

$K^{-1}\Phi(x)$	for SVMs, and
$w(x)x$	for M-estimation.

Therefore, the second factors do not depend on y and measure the outlyingness in the x-direction. Note that $K^{-1}\Phi(x)$ takes values in the RKHS H in the case of SVMs, whereas $w(x)x \in \mathbb{R}^d$ for M-estimation.

Concluding, one can say that there is a natural connection between SVMs and M-estimators for linear regression, although M-estimators have no penalty term (i.e., $\lambda = 0$) and are estimating a vector in $\mathbb{R}^d$, whereas SVMs with non-linear kernels estimate a function in a reproducing kernel Hilbert space that can have an infinite dimension.

10.5 Robust Learning from Bites (*)

In this section, we investigate a simple method called robust learning from bites (RLB) based on independent subsampling. The main goal of the method is to broaden the applicability of robust SVMs for huge data sets where classical algorithms to compute $f_{\mathrm{D},\lambda}$ for the whole data set might be too slow.

The idea of RLB is quite simple. Consider a huge data set $D = D_n$ containing data points $(x_i, y_i) \in X \times Y$, $i = 1, \dots, n$. First split the huge data set D randomly into disjoint subsets S_b with $|S_b| = n_b$, where $1 \le b \le B$ and $B \in \mathbb{N}$ much smaller than n. Then use robust SVMs for each subset and aggregate the SVM results in a robust manner.

In this section, we will assume that $X \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $Y \subset \mathbb{R}$, and n is large. We will further assume that $\min_{1 \le b \le B} n_b \to \infty$ as $n \to \infty$.

Definition 10.29. *Let $D = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in X \times Y^n$, $n \in \mathbb{N}$. Consider a random partition of D into B non-empty and disjoint subsets, i.e.,*

$$D = S_1 \uplus \ldots \uplus S_B \,,$$

where $S_b \subset D$, $n_b := |S_b|$, $n = \sum_{b=1}^{B} n_b$, $b \in \{1, \dots, B\}$, $B \in \{1, \dots, n\}$, $B \ll n$. Let $f_{\mathrm{S}_b,\lambda}$ be the SVM decision functions based on the subsamples S_b, $b = 1, \dots, B$, and let $g : H^B \to H$ and $g^* : \mathbb{R}^B \to \mathbb{R}$ be measurable functions.

*i) An **RLB estimator of type I** is defined by*

$$f_{\mathrm{D},\lambda,B}^{RLB,I} = g(f_{\mathrm{S}_1,\lambda}, \dots, f_{\mathrm{S}_B,\lambda}) \,.$$

*ii) An **RLB estimator of type II** is given by*

$$f_{\mathrm{D},\lambda,B}^{RLB,II}(x) = g^*(f_{\mathrm{S}_1,\lambda}(x), \dots, f_{\mathrm{S}_B,\lambda}(x)) \,, \qquad \forall\, x \in X.$$

Remark 10.30.

i) Obviously, RLB estimates can be defined not only based on SVMs.

ii) An RLB estimator of type I can obviously be used to define an RLB estimator of type II.

iii) An RLB estimator of type II does not necessarily define an RLB estimator of type I because the related function g^* does not necessarily correspond to a measurable and H-valued function g.

iv) The class of RLB estimators of type I and, due to *ii)*, the class of RLB estimators of type II are non-empty because for g equal to the mean we have $g(f_{\mathrm{S}_1,\lambda}, \dots, f_{\mathrm{S}_B,\lambda}) = \frac{1}{B} \sum_{b=1}^{B} f_{\mathrm{S}_b,\lambda} \in H$. $\lhd$

From our point of view, RLB is mainly a numerical algorithm to allow the computation of $f_{\mathrm{D},\lambda}$ or the predictions $f_{\mathrm{D},\lambda}(x)$ for huge data sets. Nevertheless, it will turn out that RLB estimators have some nice properties. Clearly, an RLB estimator of type II is of interest if only predictions are necessary.

Table 10.2 summarizes the three main steps of RLB. Let us now restrict attention to location estimators in the aggregation step belonging to the flexible class of L-estimators that use non-negative weights $a_{B,b} \in [0, 1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$; see Huber (1981, Section 3.3) for the general case. L-estimators are *linear* combinations of order statistics (hence their name has nothing to

Table 10.2. Principle of RLB.

```
Step 1: Construct bites.
  Split data set $D$ randomly into $B$ disjoint subsets $S_b$ of sample sizes
  $n_b \approx \lfloor n/B \rfloor$, $n_b > 1$, $b = 1, \dots, B$, $B \ll n$.
Step 2: Fit bites.
  for $(b = 1, \dots, B)$
   { Compute the (robust) estimator $f_{\mathrm{S}_b,\lambda}$ based on bite $S_b$. }
Step 3: Aggregate predictions.
  Fix weights $a_{B,b} \in [0,1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$.
  RLB type I: compute $f_{\mathrm{D},\lambda,B}^{RLB,I} = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}$.
  RLB type II: compute $f_{\mathrm{D},\lambda,B}^{RLB,II}(x_i) = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}(x_i)$, $x_i \in X$.
  If RLB type II and $g^*$ is the median:
  (i) Compute $f_{\mathrm{D},\lambda,B}^{RLB,II}(x_i) = \mathrm{median}_{1 \le b \le B} f_{\mathrm{S}_b,\lambda}(x_i)$, $x_i \in X$.
  (ii) Compute distribution-free $(1-\alpha)$ confidence intervals for
       $f_{\mathrm{P},\lambda,B}^{RLB,II}(x_i)$ based on the pair $(r,s)$ of order statistics; i.e.
       $[f_{\mathrm{S}_{(r:B)},\lambda}(x), f_{\mathrm{S}_{(s:B)},\lambda}(x)]$.
```

do with the loss function L). L-estimators in the aggregation step are defined by

$$f_{\mathrm{D},\lambda,B}^{RLB,II}(x) = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_{(\mathrm{b:B})},\lambda}(x)\,, \tag{10.50}$$

where

$$\{f_{\mathrm{S}_{(\mathrm{b:B})},\lambda}(x) : b = 1, \dots, B\} = \{f_{\mathrm{S}_b,\lambda}(x) : b = 1, \dots, B\}$$

and

$$f_{\mathrm{S}_{(1:B)},\lambda}(x) \le f_{\mathrm{S}_{(2:B)},\lambda}(x) \le \ldots \le f_{\mathrm{S}_{(B:B)},\lambda}(x)\,, \quad x \in X, \tag{10.51}$$

denote the order statistics of $f_{\mathrm{S}_b,\lambda}(x)$, $b = 1, \dots, B$. Such L-estimators are obviously convex combinations of the B estimators computed for the bites. Important L-estimators for location are given in the following example.

Example 10.31.

i) The *α-trimmed means*, $\alpha \in [0, 1/2)$, use

$$a_{B,b} = \begin{cases} \frac{1}{B - 2\lfloor \alpha B \rfloor} & \text{if } b \in \{\lfloor \alpha B \rfloor + 1, \dots, B - \lfloor \alpha B \rfloor\}, \\ 0 & \text{otherwise.} \end{cases}$$

ii) The *mean* is obtained for $\alpha = 0$ (i.e., $a_{B,b} = \frac{1}{B}$) and we obtain an RLB estimator of type I.

iii) The median uses

$$a_{B,b} = \begin{cases} \frac{1}{2} & \text{if } B \text{ is even and } b \in \left\{\frac{B}{2}, \frac{B}{2}+1\right\} \\ 0 & \text{if } B \text{ is even and } b \notin \left\{\frac{B}{2}, \frac{B}{2}+1\right\} \\ 1 & \text{if } B \text{ is odd and } b = \frac{B+1}{2} \\ 0 & \text{if } B \text{ is odd and } b \neq \frac{B+1}{2}\,, \end{cases}$$

which is the limiting case as $\alpha \to 1/2$. ◁

Our leading examples will be convex combinations of $f_{\mathrm{S}_b,\lambda}$ and the median. Of course, other robust estimators can be used instead (e.g., M-estimators, S-estimators, or Hodges-Lehmann-type R-estimators); see Huber (1981, p. 63).

If B is large, precision estimates can additionally be obtained by computing standard deviations of the predictions $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ using the central limit theorem (see Theorem A.4.10) or by applying versions of the law of the iterated logarithm (see, e.g., Einmahl and Li, 2008). However, in general we favor an alternative distribution-free method based on the median. If B is small or if it is unknown whether $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ has a finite variance, one can construct distribution-free confidence intervals for the median of $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ based on special order statistics, as the following well-known result shows.

Theorem 10.32. *Let $B \in \mathbb{N}$, $0 \le r < s \le B$, and $\tau \in (0,1)$. Let $Z_1, \ldots, Z_B$ be independent and identically distributed real-valued random variables and denote the τ-quantile by $q_\tau := \inf\{z \in \mathbb{R} : \mathrm{P}(Z_1 \le z) \ge \tau\}$. Then the corresponding order statistics $Z_{(1:B)} \le \ldots \le Z_{(B:B)}$ satisfy*

$$\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) \ge \sum_{i=r}^{s-1} \binom{B}{i} \tau^i (1-\tau)^{B-i} \ge \mathrm{P}(Z_{(r:B)} < q_\tau < Z_{(s:B)})\,.$$

Proof. Let $b \in \{0, \ldots, B\}$, $z \in \mathbb{R}$, and define $p_z := \mathrm{P}(Z_1 \le z)$. We have

$$\mathrm{P}(Z_{(b:B)} \le z) = \mathrm{P}\Big(\sum_{i=1}^{B} \mathbf{1}_{(-\infty,z]}(Z_i) \ge b\Big) = \sum_{i=b}^{B} \binom{B}{i} p_z^i (1-p_z)^{B-i}\,.$$

Recall that the incomplete beta function is given by

$$I_\tau(a,c) := \int_0^\tau t^{a-1}(1-t)^{c-1}\,dt \Big/ \int_0^1 t^{a-1}(1-t)^{c-1}\,dt, \quad a, c \in [0,\infty).$$

Using partial integration we obtain for $b \in \{0, \ldots, B\}$ that $I_\tau(b, B-b+1) = \sum_{i=b}^{B} \binom{B}{i} \tau^i (1-\tau)^{B-i}$. Using $\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) = \mathrm{P}(Z_{(r:B)} \le q_\tau) - \mathrm{P}(Z_{(s:B)} < q_\tau)$ and the definition of q_τ, we obtain

$$\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) \ge I_\tau(r, B-r+1) - I_\tau(s, n-s+1)\,.$$

The second inequality follows by similar arguments. □

Table 10.3 lists some values of B and the corresponding pair (r, s) of order statistics determining the confidence interval $[f_{\mathrm{S}_{(r:B)},\lambda}(x), f_{\mathrm{S}_{(s:B)},\lambda}(x)]$. The lower bound of the actual confidence level which is $0.5^B \sum_{j=r}^{s} \binom{B}{j}$ is also listed. The actual level of the confidence intervals can differ from $1 - \alpha$ for small values of B, see Table 10.3. The last column in Table 10.3 lists the value of $\min\{r - 1, B - s\}/B$, which is the finite-sample breakdown point for the distribution-free confidence interval for the median. For example, if $B = 17$, the fifth and the thirteenth order statistics yield a distribution-free confidence interval at the 95 percent level for the median without any further distributional assumption. Because the results of the four lowest and the four highest predictions are not considered, the breakdown point of this confidence interval is $4/17 = 0.235$.

Table 10.3. Selected pairs (r, s) of order statistics for non-parametric confidence intervals for the median.

Confidence Level $1-\alpha$	B	r	s	Lower Bound of Actual Confidence Level	$\min\{r-1, B-s\}/B$
0.90	8	2	7	0.930	0.125
	10	2	9	0.979	0.100
	18	6	13	0.904	0.278
	30	11	20	0.901	0.333
	53	21	33	0.902	0.377
	71	29	43	0.904	0.394
	104	44	61	0.905	0.413
0.95	9	2	8	0.961	0.111
	10	2	9	0.979	0.100
	17	5	13	0.951	0.235
	37	13	25	0.953	0.324
	51	19	33	0.951	0.353
	74	29	46	0.953	0.378
	101	41	61	0.954	0.396
0.99	10	1	10	0.998	0.000
	12	2	11	0.994	0.083
	26	7	20	0.991	0.231
	39	12	28	0.991	0.282
	49	16	34	0.991	0.306
	73	26	48	0.990	0.342
	101	38	64	0.991	0.366

If the robust estimator is based on hyperparameters (e.g., kernel parameters or the constant ϵ for SVMs based on the ϵ-insensitive loss) and if their values must be determined from the data set itself, a common approach is to split huge data sets into three parts for training, validation, and testing.

The training data set is used to estimate the quantity of interest for a given set of hyperparameters. The validation data set is used to determine good values for the hyperparameters by optimizing an appropriate goodness-of-fit criterion or by minimizing the generalization error. Finally, the test data set is used to estimate the goodness-of-fit criterion or the generalization error for new data points.

General Properties of RLB

The estimators $f_{\mathrm{S}_b,\lambda}$ from the B bites are stochastically independent because they are computed from disjoint parts of the data set. Denote the number of available CPUs by c and let k_B be the smallest integer that is not smaller than B/c. The computation time and the memory space for RLB can be obviously approximated in the following way.

i) Computation time, c CPUs. Assume that the computation time of the estimator $f_{\mathrm{D},\lambda}$ for a data set with n observations and d explanatory variables is of order $O(g(n,d))$, where g is some positive function. Then the computation time of RLB with B bites of subsample size $n_b \approx n/B$ is approximately of order $O(k_B g(n/B,d))$.

ii) Memory space, c CPUs. Assume that the estimator $f_{\mathrm{D},\lambda}$ for a data set with n observations and d explanatory variables needs memory space and hard disk space of order $O(g_1(n,d))$ and $O(g_2(n,d))$, respectively, where g_1 and g_2 are positive functions. Then the computation of RLB with B bites of subsample size $n_b \approx n/B$ needs memory space and hard disk space approximately of order $O(cg_1(n/B,d))$ and $O(cg_2(n/B,d))$, respectively.

The next result shows that RLB estimators inherit the usual consistency properties from the original estimators.

Lemma 10.33 (Convergence). *Consider an RLB estimator $f_{\mathrm{D},\lambda,B}^{RLB,I}$ of type I based on a convex combination with $a_{B,b} \in [0,1]$ and $\sum_{b=1}^{B} a_{B,b} = 1$.*

i) If $\mathbb{E}_\mathrm{P}(f_{\mathrm{S}_b,\lambda}) = \mathbb{E}_\mathrm{P}(f_{\mathrm{D}_n,\lambda})$, $b \in \{1,\dots,B\}$, then $\mathbb{E}_\mathrm{P}(f_{\mathrm{D},\lambda,B}^{RLB,I}) = \mathbb{E}_\mathrm{P}(f_{\mathrm{D}_n,\lambda})$.

ii) Assume that $f_{\mathrm{D}_n,\lambda}$ converges in probability (or almost surely) to $f_{\mathrm{P},\lambda}$ if $n \to \infty$, B is fixed, and $\min_{1\le b\le B} n_b \to \infty$. If g is continuous, then $f_{\mathrm{D}_n,\lambda,B}^{RLB,I}$ converges in probability (or almost surely) to $f_{\mathrm{P},\lambda}$ if $n \to \infty$.

iii) Assume that $n_b^{1/2}(f_{\mathrm{S}_b,\lambda} - f_{\mathrm{P},\lambda})$ converges in distribution to a multivariate Gaussian distribution $N(0,\Sigma)$ if $\min_{1\le b\le B} n_b \to \infty$, where $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite, and that B is fixed. If g is continuous, then $n^{1/2}(f_{\mathrm{D}_n,\lambda,B}^{RLB,I} - f_{\mathrm{P},\lambda})$ converges in distribution to a multivariate Gaussian distribution $N(0,\Sigma)$, $n \to \infty$.

Proof. *i)* follows from the linearity of the expectation operator. The parts *ii)* and *iii)* follow immediately from the continuity of g. □

The next result shows that an RLB estimator based on SVMs is a kernel-based estimator, too.

Theorem 10.34. *Consider the estimators* $f_{\mathrm{D}_n,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ *and* $f_{\mathrm{S}_b,\lambda}(x) = \sum_{(x_i,y_i)\in S_b} \alpha_{i,b} k(x, x_i)$, $x \in X$, $b = 1, \ldots, B$. *Then the RLB estimator* $f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x)$ *with weights* $a_{B,b} \in [0, 1]$, $\sum_{b=1}^{B} a_{B,b} = 1$, *and* B *fixed, is a kernel-based estimator and satisfies*

$$f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \sum_{i=1}^{n} \alpha_{i,RLB}\, k(x, x_i) \tag{10.52}$$

$$= \sum_{i\,\in\, SV(S_1)\,\cup\,\ldots\,\cup\, SV(S_B)} \alpha_{i,RLB}\, k(x, x_i), \qquad x \in X, \tag{10.53}$$

where $SV(S_b)$ *denotes the indexes of support vectors in* $f_{\mathrm{S}_b,\lambda}$ *and* $\alpha_{i,RLB} = \sum_{b=1}^{B} a_{B,b}\, \alpha_{i,b}$, $(x_i, y_i) \in D_n$.

Proof. By assumption, each bite S_b is fitted with an SVM having the decision function

$$f_{\mathrm{S}_b,\lambda}(x) = \sum_{i\in S_b} \alpha_{i,b} k(x, x_i), \qquad x \in X.$$

Because the bites S_b, $b = 1, \ldots, B$, are disjoint and B is fixed, the RLB estimator with weights $a_{B,b}$ has the representation

$$f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \sum_{b=1}^{B} a_{B,b} \sum_{i\in S_b} \alpha_{i,b}\, k(x, x_i) \tag{10.54}$$

$$= \sum_{i=1}^{n} \sum_{b=1}^{B} a_{B,b}\, \alpha_{i,b}\, k(x, x_i)\,, \quad x \in X, \tag{10.55}$$

which gives the assertion. □

If all support vectors in $S_1, \ldots, S_B$ are different, we have $\alpha_{i,RLB} = a_{B,b}\, \alpha_{i,b}$ in (10.53). Now, we investigate the number of support vectors of RLB estimators based on SVMs.

Theorem 10.35. *Under the assumptions of Theorem 10.34, an RLB estimator* $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ *has the following properties.*

i) The number of support vectors (i.e., $\alpha_{i,RLB} \neq 0$*) of* $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ *is given by*

$$\#SV(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}) = \Bigg| \bigcup_{b\in\{1,\ldots,B:\, a_{B,b}>0\}} SV(S_b) \Bigg|\,. \tag{10.56}$$

ii) Let $Y = \{-1, +1\}$, $B \in \mathbb{N}$ *be fixed,* $\min\{n_1, \ldots, n_B\} \to \infty$, *and consider weights* $a_{B,b} \in (0,1)$ *with* $\sum_{b=1}^{B} a_{B,b} = 1$. *Under the assumptions of Theorem 8.34, we have the probabilistic lower bound on the sparseness:*

$$\mathrm{P}^n\left(D_n \in (X \times Y)^n : \#SV(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}) \geq \frac{\mathcal{S}_{L,\mathrm{P}} - \varepsilon}{n}\right) \geq \prod_{b=1}^{B}\left(1 - e^{-\frac{\delta^2 n_b}{18h^2(\lambda)B}}\right). \tag{10.57}$$

Proof. i) follows immediately from (10.53). Now let us consider *ii)*. Using the probabilistic lower bound on the sparseness given in Theorem 8.34, we see that $f_{\mathrm{D}_n,\lambda}$ satisfies

$$\mathrm{P}^n\Big(D_n \in (X \times Y)^n : \#SV(f_{\mathrm{D}_n,\lambda}) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n\Big) \;\geq\; 1 - e^{-\frac{\delta^2 n}{18h^2(\lambda)}} . \tag{10.58}$$

The bites S_b, $b = 1, \ldots, B$, are independent and identically distributed by the construction of RLB and we have $n = Bn_b$. Hence

$$\begin{aligned}
&\mathrm{P}^n\left(D_n \in (X \times Y)^n : \#\mathrm{SV}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n\right)\\
&= \mathrm{P}^n\Big((S_1, \ldots, S_B) \in (X \times Y)^n : \#\mathrm{SV}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \geq \sum_{b=1}^{B}(\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n_b\Big)\\
&\geq \mathrm{P}^n\left(\forall\, S_b \in (X \times Y)^{n_b}, b = 1, \ldots, B \,:\, \#\mathrm{SV}\big(f_{\mathrm{S}_b,\lambda_{n_b}}\big) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n_b\right)\\
&= \prod_{b=1}^{B}\mathrm{P}^{n_b}\left(S_b \in (X \times Y)^{n_b} : \#\mathrm{SV}\big(f_{\mathrm{S}_b,\lambda_{n_b}}\big) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n_b\right)\\
&\geq \prod_{b=1}^{B}\left(1 - e^{-\frac{\delta^2 n_b}{18h^2(\lambda)B}}\right) \to 1,\; n \to \infty\,,
\end{aligned}$$

which completes the proof. □

The result given in (10.57) has the following interpretation: with probability exponentially fast tending to one if the total sample size $n = Bn_b$ converges to ∞ but B is fixed, the fraction of support vectors of the kernel based RLB estimator $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ in a binary classification problem is essentially greater than the average of the Bayes risks for the bites. If we compare the rates of convergence in (10.57) and (10.58), we obtain the expected result that the probabilistic lower bound in both results is the same, but that the rate of convergence is better in (10.58) than in its counterpart in (10.57) for $B > 1$. This can be interpreted as the price we have to pay if we use RLB because of its computational advantages for huge data sets instead of $f_{\mathrm{D}_n,\lambda}$.

Let us now investigate conditions to guarantee that RLB estimators using SVMs are L-risk consistent; i.e.,

$$\mathcal{R}_{L,\mathrm{P}}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \to \mathcal{R}_{L,\mathrm{P}}$$

in probability for $n \to \infty$.

Theorem 10.36. *Let $f_{\mathrm{D},\lambda,b}$ be an L-risk consistent support vector machine with a convex loss function. Consider an RLB estimator $f^{RLB,I}_{\mathrm{D},\lambda,B}$ with weights $a_{B,b} \in (0,1)$, $\sum_{b=1}^{B} a_{B,b} = 1$, $B \geq 1$ fixed, and $\min_{1\leq b\leq B} n_b \to \infty$. Then $f^{RLB,I}_{\mathrm{D},\lambda,B}$ is L-risk consistent.*

Proof. The RLB estimator $f^{RLB,I}_{\mathrm{D},\lambda,B}$ is a convex combination of $f_{\mathrm{S}_b,\lambda}$, $b = 1,\ldots,B$, because $a_{B,b} \in (0,1)$ and $\sum_{b=1}^{B} a_{B,b} = 1$. Therefore,

$$\begin{aligned}
0 &\leq \int L\big(Y, f^{RLB,I}_{\mathrm{D},\lambda,B}(X)\big)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \int L\Big(Y, \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}(X)\Big)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \int \sum_{b=1}^{B} a_{B,b}\, L\big(Y, f_{\mathrm{S}_b,\lambda}(X)\big)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} && (10.59) \\
&= \sum_{b=1}^{B} a_{B,b} \Big(\int L\big(Y, f_{\mathrm{S}_b,\lambda}(X)\big)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} \Big), && (10.60)
\end{aligned}$$

which converges in probability to zero if $\min_{1\leq b\leq B} n_b \to \infty$, $n \to \infty$. Here we used the convexity of L in (10.59) and the L-risk consistency of $f_{\mathrm{D}_n,\lambda}$ in (10.60). □

From the no-free-lunch theorem (see Theorem 6.6), the proof given above cannot be modified in a simple way to cover the case where the number of bites $B = B(n)$ depends on the sample size because we have no uniform rate of consistency without restricting the class of probability measures.

Robustness Properties of RLB

Now we derive some results that show that certain robustness properties are inherited from the original estimator $f_{\mathrm{D}_n,\lambda}$ to the RLB estimator. Let us start with an investigation of the influence function of the RLB estimator $f^{RLB,I}_{\mathrm{P},\lambda,B}$.

Theorem 10.37 (Influence function of RLB). *Assume that the influence function of $f_{\mathrm{P},\lambda}$ exists for the distribution $\mathrm{P} \in \mathcal{M}_1$. Further assume that the weights satisfy $a_{B,b} \in [0,1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$ and B fixed. Then the influence function of $f^{RLB,I}_{\mathrm{P},\lambda,B}$ using these weights exists and equals the influence function of $f_{\mathrm{P},\lambda}$.*

Proof. Let $z \in X \times Y$ and define $S_1(\mathrm{P}) := f^{RLB,I}_{\mathrm{P},\lambda,B}$ and $S_2(\mathrm{P}) := f_{\mathrm{P},\lambda}$, $\mathrm{P} \in \mathcal{M}_1$. It follows that

$$\begin{aligned}
\mathrm{IF}(z; S_1, \mathrm{P}) &= \lim_{\varepsilon \downarrow 0} \frac{f^{RLB,I}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda,B} - f^{RLB,I}_{\mathrm{P},\lambda,B}}{\varepsilon} \\
&= \lim_{\varepsilon \downarrow 0} \frac{\sum_{b=1}^{B} a_{B,b}\, f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda} - \sum_{b=1}^{B} a_{B,b}\, f_{\mathrm{P},\lambda}}{\varepsilon} \\
&= \sum_{b=1}^{B} a_{B,b} \lim_{\varepsilon \downarrow 0} \frac{f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda} - f_{\mathrm{P},\lambda}}{\varepsilon} \\
&= \mathrm{IF}(z; S_2, \mathrm{P}),
\end{aligned}$$

which gives the assertion. □

Hence, if $f_{\mathrm{P},\lambda}$ has a bounded influence function, the same is true for an RLB estimator of type I using a convex combination of weights as specified above. Recall that the existence and boundedness of the influence function of $f_{\mathrm{P},\lambda}$ were proven in Sections 10.3 and 10.4 under weak conditions on the loss function and on the kernel for classification problems and regression problems.

We have not considered breakdown points for $f_{\mathrm{P},\lambda}$ or $f_{\mathrm{D},\lambda}$. However, if we specify $\lambda = 0$ in combination with a linear kernel, then SVMs are just M-estimators in linear regression for which the breakdown points of M-estimators are well-known.

Theorem 10.38 (Finite-sample breakdown point of RLB). *Let $n_b \equiv n/B$, $b = 1, \ldots, B$. Assume that the finite-sample breakdown points $\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})$ of the estimators $f_{\mathrm{S}_b,\lambda}$ are all identical. Denote the finite-sample breakdown point of the estimator $\hat{\mu} = \hat{\mu}(f_{S_1,\lambda}, \ldots, f_{S_B,\lambda})$ in the aggregation step by $\varepsilon^*_B(\hat{\mu})$. Then the finite-sample breakdown point of an RLB estimator is given by*

$$\varepsilon^*_{RLB,n,B} = \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})\Big(\varepsilon^*_B(\hat{\mu}) + \frac{1}{B}\Big) + \frac{B}{n}\varepsilon^*_B(\hat{\mu})\,. \tag{10.61}$$

Proof. The minimum number of points needed to modify $f_{\mathrm{S}_b,\lambda}$ in bite S_b such that a breakdown occurs is given by $n_b\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda}) + 1$, $b = 1, \ldots, B$. The RLB estimator breaks down if at least $B\varepsilon^*_B(\hat{\mu}) + 1$ of the estimators $f_{\mathrm{S}_1,\lambda}$, …, $f_{\mathrm{S}_B,\lambda}$ break down. This gives the assertion. □

Remark 10.39.

i) If B and $n_b = B/n$ are large, we obtain from (10.61) the lower bound

$$\varepsilon^*_{RLB,n,B} \geq \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})\,\varepsilon^*_B(\hat{\mu})\,. \tag{10.62}$$

ii) For the median, we obtain $\varepsilon^*_B(\hat{\mu}) = \frac{1}{2} - \frac{1}{B}$ if B is even and $\varepsilon^*_B(\hat{\mu}) = \frac{1}{2}$ if B is odd.

iii) For α-trimmed means, $\alpha \in (0, \frac{1}{2})$, we obtain $\varepsilon^*_B(\hat{\mu}) = \lfloor \alpha B \rfloor / B$.

iv) If the mean or any other estimator with $\varepsilon^*_B(\hat{\mu}) = 0$ is used in the aggregation step, RLB has a finite-sample breakdown point of $\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})/B \to 0$ if $B \to \infty$. ◁

Example 10.40 (Univariate location model). Consider the univariate location problem, where $x_i \equiv 1$ and $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, $n = 55$. Assume that all output values are different. The finite-sample breakdown point of the median is $\lfloor n/2 \rfloor / n = 0.49$. The mean has a finite-sample breakdown point of 0. Now let us investigate the robustness of the RLB approach with $B = 5$ and $n_b = 11$, $b = 1, \ldots, B$. *(i)* If the median is used as the location estimator in each bite and if the median is used in the aggregation step, the finite-sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = 0.309$. This value is reasonably high but lower than the finite-sample breakdown point of the median for the whole data set, which is 0.49. Note that in a *fortunate* situation the impact of up to $(2 \cdot 11 + 5 \cdot 3)/55 = 0.672$ extremely large data points (say equal to $+\infty$) is still bounded for the RLB estimator in this setup: modify all data points in $B\varepsilon^*_B(\hat{\mu}) = 2$ bites and up to $n_b\, \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda}) = 5$ data points in the remaining $B(1 - \varepsilon^*_B(\hat{\mu})) = 3$ bites. This is no contradiction to (10.61) because the breakdown point measures the *worst-case* behavior. *(ii)* If the median is used as the location estimator in each bite and if the mean is used in the aggregation step, the finite-sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = (1/B)\varepsilon^*_{n_b}(\hat{f}) = 0.09$. *(iii)* If the mean is used as the location estimator in each bite and also in the aggregation step, we of course obtain $\varepsilon^*_{RLB,n,B} = 0$. ◁

Determination of the Number of Bites

The number of bites obviously has some impact on the statistical behavior of the RLB estimator and also on the computation time and the necessary computer memory. An optimal choice of the number of bites B will in general depend on the unknown distribution P. But some general arguments are given on how to determine B in an appropriate manner.

One should take the sample size n, the computer resources (number of CPUs, RAM, hard disk), and the acceptable computation time into account. The quantity B should be much lower than n because otherwise there is not much hope of obtaining useful estimators from the bites. Further, B should depend on the dimensionality d of the input values. For example, a rule of thumb for *linear* regression is that n/d should be at least 5. Because the function f is completely unknown in non-parametric regression, assumptions on the complexity of f are crucial. The sample size n_b for each bite should converge to infinity if $n \to \infty$ to obtain consistency of RLB. The results from some numerical experiments not given here can be summarized as follows.

i) If B is too large, the computational overhead increases and the danger of bad fits increases because n_b is too small to provide reasonable estimators.

ii) A major decrease in computation time and memory saving is often already present if B is chosen in a way such that the numerical algorithms to fit each bite fit nicely into the computer (CPU, RAM, hard disk). Nowadays, robust estimators can often be computed for sample sizes up to $n_b = 10^4$ or $n_b = 10^5$. In this case, $B = \lfloor n/n_b \rfloor$ can be a reasonable choice.

iii) If distribution-free confidence intervals at the $(1-\alpha)$ level for the median of the predictions (i.e., $f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \mathrm{median}_{1\le b\le B}\, f_{\mathrm{S}_b,\lambda}(x)$, $x \in X$) are needed, one should take into account that the actual confidence level of such confidence intervals based on order statistics can be conservative (i.e., higher than the specified level) for some pairs (r,s) of order statistics due to the discreteness of order statistics.

Numerical Example

Now we apply the RLB approach to kernel logistic regression (i.e., $L = L_{\text{c-logist}}$). All computations were done with the program `myKLR` (Rüping, 2003), which is an implementation of the algorithm proposed by Keerthi *et al.* (2005) to solve the dual problem. We choose KLR for two reasons. First, the computation of KLR needs much more time than SVMs based on the hinge loss because the latter solves a quadratic instead of a convex program in dual space and because for the hinge loss, the number of support vectors is usually much smaller than n. Therefore, the need for computational improvements is greater for $L_{\text{c-logist}}$ than for L_{hinge} and the potential gains of RLB can be more important. Second, the number of support vectors of KLR is approximately equal to n, which slows down the computation of predictions.

The simulated data sets contain n data points $(x_i, y_i) \in \mathbb{R}^8 \times \{-1,+1\}$ simulated in the following way. All eight components of $x_i = (x_{i,1},\dots,x_{i,8})$ are simulated independently from a uniform distribution on $(0,1)$. The responses y_i are simulated independently from a logistic regression model according to $\mathrm{P}(Y_i = +1|X_i = x_i) = 1/\big(1+e^{-f(x_i)}\big)$ and $\mathrm{P}(Y_i = -1|X_i = x_i) = 1 - \mathrm{P}(Y_i = +1|X_i = x_i)$. We set

$$f(x_i) = \sum_{j=1}^{8} x_{i,j} - x_{i,1}x_{i,2} - x_{i,2}x_{i,3} - x_{i,4}x_{i,5} - x_{i,1}x_{i,6}x_{i,7}\,.$$

The data points were saved as ASCII files, where $x_{i,j}$ is stored with four decimal places. The numerical results of fitting kernel logistic regression to such data sets are given in Table 10.4. It is obvious that in this situation RLB can save a lot of computation time. If the whole data set has $n = 10^5$ observations, approximately 10 hours is needed to compute KLR on a PC. If RLB with $B = 10$ bits are used, each with a subsample size of $n_b = 10^4$, one needs approximately 16 minutes if there is 1 GB of kernel cache available. This is a reduction by a factor of 38. If there are five CPUs available and each processor can use up to 200 MB kernel cache, RLB with $B = 10$ will need approximately 11 minutes, which is a reduction by a factor of 55.

When RLB was applied to a data set from a union of German insurance companies ($n \approx 4.5 \times 10^6$), the computation time decreased from months to days, see Christmann (2005) and Christmann *et al.* (2007) for details.

Table 10.4. Computation times for kernel logistic regression using `myKLR`.

Sample Size n	CPU Time	Used Cache in MB	Available Cache in MB
2000	4 sec	33	200
5000	25 sec	198	200
10000	5 min, 21 sec	200	200
10000	1 min, 33 sec	787	1000
20000	24 min, 11 sec	1000	1000
20000	14 min, 35 sec	1000	1000
100000	9 h, 56 min, 46 sec	1000	1000

10.6 Further Reading and Advanced Topics

Many different criteria have been proposed to define robustness or stability in a mathematical way, such as the minimax approach (Huber, 1964), qualitative robustness (Hampel, 1968, 1971), the sensitivity curve (Tukey, 1977), the approach based on least favorable local alternatives (Huber, 1981; Rieder, 1994), the approach based on influence functions (Hampel, 1974; Hampel *et al.*, 1986), the maxbias curve (Hampel *et al.*, 1986; Huber, 1964), the global concept of min-max bias robust estimation (He and Simpson, 1993; Martin *et al.*, 1989), the breakdown point (Donoho and Huber, 1983; Hampel, 1968), depth-based methods (Tukey, 1975), and configural polysampling (Morgenthaler and Tukey, 1991).

Section 10.2 described the main approaches of robust statistics. It is mainly based on Huber (1981), Hampel *et al.* (1986), Rousseeuw and Leroy (1987), Jurečková and Sen (1996), Davies and Gather (2004), and Maronna *et al.* (2006). For qualitative robustness, we refer also to Hampel (1968, 1971) and Cuevas (1988). The robustness results of SVMs for classification and regression treated in Sections 10.3 and 10.4 are based on Christmann (2002, 2004, 2005) and Christmann and Steinwart (2004, 2007). The robust learning from bites approach discussed in Section 10.5 was proposed by Christmann *et al.* (2007). Using Bouligand derivatives instead of Gâteaux derivatives, it was shown that SVMs based on Lipschitz-continuous loss functions have under weak assumptions a bounded Bouligand influence function provided that a bounded kernel is used; see Christmann and Van Messem (2008). Special cases are SVMs based on the ϵ-insensitive loss, Huber's loss, logistic loss for regression, and pinball loss for quantile regression. Debruyne *et al.* (2007) gave robustness results for a reweighted form of the SVM based on the least squares loss and Debruyne (2007) and proposed tools for model selection.

For a detailed description of outliers and approaches to detect, identify or test for outliers, we refer to Beckmann and Cook (1983), Barnett and Lewis (1994), Gather (1984, 1990), Davies and Gather (1993), Becker and Gather (1999), Gather and Pawlitschko (2004), Hampel *et al.* (1986), and Christmann (1992). Steinwart *et al.* (2005) and Scott and Nowak (2006), among many others, used SVMs for anomaly detection and for learning minimum volume sets that are related to outlier regions proposed by Davies and Gather (1993).

Choice of a Metric

Neighborhoods around a probability distribution play an important role in robust statistics. Such neighborhoods are usually constructed by specifying a metric on the space of probability measures. However, the robustness properties of a statistical procedure can depend on the metric. The choice of a suitable metric is not always a simple task; see Hampel (1968), Huber (1981, Chapter 2), and Davies (1993, pp. 1851ff.) for nice treatments of this topic. For a recent controversial discussion of the question of which metrics are especially suitable for robust statistics, see Davies and Gather (2005, 2006) and Hampel (2005).

Sometimes one restricts attention to investigating robustness properties of statistical methods only in gross-error neighborhoods because such problems are often easier to solve. At first glance-neighborhoods defined via the norm of total variation seem to be an attractive alternative to gross-error neighborhoods. However, Hampel (1968, p. 43) showed the interesting fact that neighborhoods defined by the norm of total variation do not perform much better than gross-error neighborhoods. Consider two probability measures P and Q defined on the same measurable space, and let $\varepsilon \in (0,1)$. Then Q is an element of a neighborhood of radius ε around P with respect to the metric defined by the norm of total variation if and only if $\mathrm{Q} = (1-\varepsilon)\mathrm{P}+\varepsilon(\mathrm{P}+\mathrm{Q}_1-\mathrm{Q}_2)$, where $\mathrm{Q}_1 := \varepsilon^{-1}(\mathrm{Q} - \min\{\mathrm{P},\mathrm{Q}\})$ and $\mathrm{Q}_2 := \varepsilon^{-1}(\mathrm{P} - \min\{\mathrm{P},\mathrm{Q}\})$. Note that Q_1 and Q_2 are not necessarily probability measures but Q is a mixture of P and $\mathrm{P} + \mathrm{Q}_1 - \mathrm{Q}_2$ with mixing proportion ε.

Qualitative Robustness

Hampel (1968, 1971) proposed not only qualitative robustness but also the notion of Π-robustness. Informally, a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called Π-robust if it is qualitatively robust but also insensitive to "small" deviations from the assumption of independence. Hampel (1971, Theorem 3) proved that Π-robustness at some distribution P implies qualitative robustness but not vice versa. Boente *et al.* (1987) generalized Hampel's concept of Π-robustness for a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ in Euclidean spaces to Polish spaces. Cuevas (1988) showed that qualitative robustness is incompatible with consistency of multivariate density estimates if the L_1-metric is used. Cuevas also investigated the case of qualitative robustness for certain stochastic processes with continuous trajectories on $[0,1]$ and proved that the simple mean function is not robust in this sense, but robust estimation is possible by generalizing M-estimators and α-trimmed means.

Robustness of SVMs

The boundedness of the sensitivity curve of SVMs for classification was essentially already established by Bousquet and Elisseeff (2002). The results given in Sections 10.3 and 10.4 and also some results for the more general case, where not only $f_{\mathrm{P},\lambda}$ but also an offset term $b_{\mathrm{P},\lambda}$ must be estimated, were derived by Christmann and Steinwart (2004, 2007, 2008). If attention is restricted to SVMs based on a Lipschitz-continuous loss function for regression and a bounded kernel, Christmann and Van Messem (2008) showed that $f_{\mathrm{P},\lambda}$ has even a bounded Bouligand influence function.

Robust Learning from Bites

The RLB approach proposed by Christmann (2005) and Christmann *et al.* (2007) has connections to the remedian proposed by Rousseeuw and Bassett (1990) for univariate location estimation, Rvote proposed by Breiman (1999a), DRvote with classification trees using majority voting (Chawla *et al.*, 2004), and subsampling (Politis *et al.*, 1999). Bootstrapping computer-intensive robust methods for huge data sets is often impossible due to computation time and memory limitations of the computer, but see Salibian-Barrera *et al.* (2008). RLB has some similarity to the algorithms FAST-LTS and FAST-MCD proposed by Rousseeuw and Van Driessen (1999, 2000) for robust estimation in linear regression or multivariate location and scatter models for large data sets. FAST-LTS and FAST-MCD split the data set into subsamples, optimize the objective function in each subsample, and use these solutions as starting values to optimize the objective function for the whole data set. This is in contrast to RLB, which aggregates estimation results from the bites to obtain robust confidence intervals.

Stability and Learnability

There is currently some interest in proving connections between stability, learnability, and predictivity for broad classes of ERM methods. We would like to cite Bousquet and Elisseeff (2002), Poggio *et al.* (2004), Mukherjee *et al.* (2006), and Elisseeff *et al.* (2005). Although different notions of stability are used in these papers, their notions of stability have a meaning similar to robustness. Several of these notions of stability only measure the impact of just *one* data point such that the connection to Tukey's sensitivity curve is obvious. Caponnetto and Rakhlin (2006) consider stability properties of empirical risk minimization over Donsker classes.

Robustness in Parametric and Non-parametric Models

There is a large body of literature on robust estimators and tests for outliers in *parametric* models, especially for linear regression, binary regression, and multivariate location and scatter problems. From the viewpoint of robustness, the relationships between SVMs based on a convex loss function and classical M-estimation, say of Huber type, are strong. However, there are three important differences between the two approaches. First, classical M-estimation for linear regression or binary classification has no penalty term in its optimization problem (i.e., $\lambda = 0$). Second, SVMs often use non-linear kernels such as the Gaussian RBF kernel, which already can improve robustness properties (see, e.g., Theorems 10.10 and 10.18) whereas classical M-estimation uses only the unbounded linear kernel. Third, non-convex loss functions are sometimes used in classical M-estimation in linear models to improve robustness properties of such methods, although one is then often faced with the problem of non-uniqueness of the estimates and with algorithmic problems to find a global optimum and not only a local one.

Downweighting extreme points in the design space X is automatically done by SVMs using, for example, the Gaussian RBF kernel. Generalized M-estimators for regression are able to downweight not only data points extreme in the residual $y - f(x)$ but also extreme points x, so-called *leverage points*. For additional information on M-estimation and related methods in linear models, we refer to Huber (1964, 1981), Hampel *et al.* (1986), Fernholz and Morgenthaler (2005), Gather and Hilker (1997), Kent and Tyler (1991, 1996, 2001), and Maronna and Yohai (1981). Mendes and Tyler (1996) investigated constrained M-estimation for linear regression. Simpson *et al.* (1987) gave theoretical arguments in favor of M-estimators based on a twice Fréchet differentiable loss function. Rieder (1994) and Ruckdeschel and Rieder (2004) investigated robust asymptotic statistics based on shrinking neighborhoods. Rieder *et al.* (2008) proposed the radius-minimaxity concept.

Many robust alternatives to M-estimators were proposed in the literature. We did not describe such methods here because the connection between SVMs and M-estimation is much closer. However, we would like to mention the following alternatives to M-estimation in parametric models.

For L-estimators, which are linear combinations of order statistics, we refer to Serfling (1980) and Huber (1981). Regression quantiles are generalizations of L-estimators to the regression context and are treated, for example, by Koenker and Bassett (1978), Portnoy and Koenker (1997), and Koenker (2005). Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006) investigated SVMs based on the Lipschitz-continuous pinball loss function to estimate quantile functions in a nonparametric way. Some results on L-risk consistency, rates of convergence, and various robustness properties of such SVMs were derived by Christmann and Steinwart (2007, 2008) and Steinwart and Christmann (2008).

Statistical properties of R-estimators were investigated, for example, by Serfling (1980), Coakley and Hettmansperger (1993), Hettmansperger *et al.* (1997), Ollila *et al.* (2002, 2003), and Hallin and Paindaveine (2004, 2005).

Of special interest in robust estimation in parametric models are methods with a positive or high-breakdown point to overcome the lack of a low breakdown point of M-, L-, and R-estimators in high dimensions d. Probably the most important and most often used high-breakdown estimators for linear regression are the estimators least median of squares (LMS) and least trimmed squares (LTS), which were investigated in detail by Rousseeuw (1984, 1994, 1997a,b) and Rousseeuw and Leroy (1987). Both estimators were also used in many recent proposals to construct two-step estimators, which inherit from the starting estimators LMS or LTS a high-breakdown point and from the second step improved asymptotic behavior. LMS and LTS belong to the class of S-estimators and their generalizations. These estimators were investigated by Rousseeuw and Yohai (1984), Rousseeuw and Van Driessen (2000), Rousseeuw and van Zomeren (1990, 1991), Davies (1990, 1993, 1994), and Croux *et al.* (1994). Yohai (1987) proposed high-breakdown-point and high-efficiency robust estimates for linear regression; see also Yohai *et al.* (1991). Yohai and Zamar (1988) investigated high-breakdown point estimates of regression by means of the minimization of an efficient scale.

For recent advances of robust methods for online monitoring data we refer to Gather *et al.* (2002), Gather and Fried (2004), and Gather and Schettlinger (2007).

For robust estimation in generalized linear models, we refer to Pregibon (1982) for resistant estimation and Stefanski *et al.* (1986), Künsch *et al.* (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996), and Cantoni and Ronchetti (2001) for M-estimation. Christmann (1992, 1994, 1998) proposed high-breakdown point estimators in generalized linear models. Rousseeuw and Christmann (2003) proposed robust estimates for logistic regression and other binary regression problems and solved the problem of non-existence of the classical non-robust maximum likelihood estimates and almost all robust estimates proposed earlier. The mathematical reason why many estimators do not have a solution for all possible data sets turned out to be the complete or quasi-complete separation of data points with positive weights; see Albert and Anderson (1984) and Santner and Duffy (1986).

For depth-related methods, see Tukey (1975) for the halfspace depth, Oja (1983) for Oja's median, Liu (1990, 1999) for the simplicial depth, Rousseeuw and Hubert (1999) for regression depth, Mizera (2002) for tangent depth, He and Wang (1997) for depth contours for multivariate data sets, and Zuo and Serfling (2000) and Mosler (2002) for various variants of statistical depth functions. For a numerical comparison between the support vector machine and the regression depth method proposed by Rousseeuw and Hubert (1999), see Christmann and Rousseeuw (2001) and Christmann *et al.* (2002).

Projection pursuit and its relation to robust estimation were investigated by Huber (1985, 1993) and Donoho and Johnstone (1989).

10.7 Summary

This chapter showed that support vector machines have—besides many other nice mathematical and algorithmic properties—good robustness properties if the loss function and the kernel are appropriately chosen. There is a natural connection between SVMs and M-estimation.

Robust statistics investigates how small violations of the model assumptions influence the results of the methods used. Many researchers still ignore deviations from ideal models because they hope that such deviations will not matter. However, J. W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 that: *"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"*; see Hampel *et al.* (1986, p. 21).

Small model violations are unavoidable in almost all practical applications because statistical models are usually simplifications of the true process that generated a data set and because typing errors, outliers, and many types of contamination occur in practice. In particular, this is true for data mining projects in which huge data sets with moderate data quality often must be analyzed; see Chapter 12.

It is therefore important to study the robustness properties of SVMs, although the model assumption of independent and identically distributed pairs of random variables without further restriction of the probability distribution is weak. Many approaches of robust statistics were developed for estimators of an unknown vector $\theta \in \mathbb{R}^d$ in parametric models. From a mathematical point of view, it is also interesting to investigate how these approaches can be used for SVMs where a function f in a (usually) infinite-dimensional Hilbert space H must be estimated.

A motivation of robust statistics was given in Section 10.1. Different approaches of robustness were described in Section 10.2, including topological neighborhoods in the space of probability distributions, qualitative robustness, influence functions with the related notions of gross error sensitivity, maxbias and sensitivity curve, and breakdown points.

Robustness properties of support vector machines for binary classification and regression were investigated in Sections 10.3 and 10.4. It was shown that the influence function of SVMs exists and is bounded if the loss function has a bounded first derivative and if the kernel is bounded and continuous, e.g., the Gaussian RBF kernel. This shows that SVMs based on a logistic loss function have nice robustness properties for classification and regression problems. In contrast, SVMs based on the least squares loss yield less robust results than SVMs based on the classical hinge loss or the ϵ-insensitive loss. Bounds for the maxbias, gross error sensitivity, and the sensitivity curve were derived; some of the bounds even turned out to be uniform. For the regression case with unbounded output space Y, it turned out that good robustness properties of SVMs are only available if the tail behavior of the probability distribution P and the growth type of the loss function L are appropriately related to each other. Especially Lipschitz-continuous loss functions of growth type 1 offer good robustness properties. There are relationships between these results on robustness and stability results obtained by Poggio *et al.* (2004) and Mukherjee *et al.* (2006).

In Section 10.5, a simple but powerful subsampling strategy called robust learning from bites (RLB) was described to make SVMs usable for huge data sets (e.g., in data mining projects). RLB is designed for situations under which the original robust method cannot be applied due to excessive computation time or memory space problems. In these situations, RLB offers robust estimates and additionally robust confidence intervals. Although RLB estimators will in general not fulfill certain optimality criteria, the method has the advantages of scalability (the number of bites can be chosen according to the data set and the available hardware), performance (the computational steps for different bites can easily be distributed on several processors), robustness (RLB inherits robustness properties from the SVMs used by this strategy), and confidence intervals (no complex formulas are needed to obtain distribution-free (componentwise) confidence intervals for the estimates or the predictions).

10.8 Exercises

10.1. Mean and median ($\star$)
Prove that the mean and median are solutions of the optimization problems (10.5) and (10.6).

10.2. Property of Prohorov metric ($\star$)
Let $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(Z)$. Define $\mathrm{P}_\delta := (1-\delta)\mathrm{P} + \delta\mathrm{Q}$ and $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$ for $\delta, \varepsilon \in [0,1]$. Show that $d_{\mathrm{Pro}}(\mathrm{P}_\delta, \mathrm{P}_\varepsilon) \leq |\delta - \varepsilon|$ if $\delta \to \varepsilon$.

Hint: Dudley (2002).

10.3. Bounded Lipschitz metric ($\star\star\star$)
Prove Theorem A.4.22*iii)*.

Hint: The direction (b) ⇒ (a) follows easily. One can use the Lagrange approach, the Kuhn and Tucker (1951) theorem, and Strassen's Theorem A.4.16 to prove the converse direction. See Huber (1981, pp. 30ff.) for details.

10.4. Lévy metric ($\star\star$)
The Lévy distance $d_{\text{Lévy}}(\mathrm{P}_1, \mathrm{P}_2)$ of two probability measures P and Q on $\mathbb{R}$ is defined by

$$\inf\bigl\{\varepsilon > 0 : \mathrm{P}_1((-\infty, x-\varepsilon]) - \varepsilon \leq \mathrm{P}_2((-\infty, x]) \leq \mathrm{P}_1((-\infty, x+\varepsilon] + \varepsilon\bigr\}\,.$$

i) Show that $d_{\text{Lévy}}$ is a metric that metricizes the weak* topology.
ii) Define $\mathrm{P}_\delta := (1-\delta)\mathrm{P} + \delta\mathrm{Q}$ and $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$, $\delta, \varepsilon \in [0,1]$. Show that the Lévy metric shares with the Prohorov metric and the bounded Lipschitz metric the property $d_{\text{Lévy}}(\mathrm{P}_\delta, \mathrm{P}_\varepsilon) = O(|\delta - \varepsilon|)$, $\delta \to \varepsilon$.

Hint: See Huber (1981, pp. 25ff.).

10.5. Kolmogorov metric and total variation distance ($\star\star$)

i) Prove that the Kolmogorov distance defined by

$$d_{\mathrm{Kol}}(\mathrm{P},\mathrm{Q}) := \sup_{x\in\mathbb{R}} |\mathrm{P}((-\infty,x]) - \mathrm{Q}((-\infty,x])|, \quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_1(\mathbb{R}),$$

is a metric.

ii) Prove that the total variation distance defined by

$$d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q}) := \sup_{A\in\mathcal{B}(\mathbb{R}^n)} |\mathrm{P}(A) - \mathrm{Q}(A)|, \quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_1(\mathbb{R}^n),$$

is a metric.

iii) Clarify the connections between d_{Pro}, $d_{\text{Lévy}}$, d_{Kol}, and d_{tv}.

iv) Do d_{tv} and d_{Kol} generate the weak* topology?

Hints: See Huber (1981, pp. 34ff.). *iii).* We have $d_{\text{Lévy}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{Pro}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q})$ and $d_{\text{Lévy}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{Kol}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q})$. *iv).* No.

10.6. Asymmetric logistic loss ($\star$)
Consider the asymmetric logistic loss $L : Y\times\mathbb{R}\to[0,\infty)$,

$$L(x,y,t) = \Big(\frac{c_2}{2} - c_1\Big)r - \frac{c_2}{2}\ln\Big(4\Lambda(r-c_3)\big(1-\Lambda(r-c_3)\big)\Big) + c_4\,,$$

where $r = y - t$, $0 < c_1 < c_2 < \infty$, $\Lambda(r) := 1/(1+e^{-r})$, $c_3 = -\Lambda^{-1}(c_1/c_2)$, and $c_4 = (c_2/2)\ln(4\frac{c_1}{c_2}(1-\frac{c_1}{c_2}))$.

i) Show that $(c_1,c_2) = (1,2)$ gives $L = L_{\text{r-logist}}$.

ii) Show that $L' = \partial_3 L$ and $L'' = \partial_{33}L$ are continuous and bounded.

iii) Is L a Nemitski loss function of some order $p\in(0,\infty)$?

Hint: iii). Yes.

10.7. Maxbias ($\star\star$)
Derive bounds for the maxbias$(\varepsilon; S, \mathrm{P})$ of SVMs over contamination neighborhoods $N_\varepsilon(\mathrm{P})$.

Hint: Use Theorems 10.25, 10.26, and 10.27.

10.8. Kernel-based LMS ($\star\star$)
Consider a non-parametric regression problem treated in Section 10.4. Define an estimator for $f\in H$ by $f^*_{\mathrm{P},\lambda} := \arg\min_{f\in H} \mathrm{Median}_{\mathrm{P}} L(Y, f(X)) + \lambda\|f\|_H^2$.

i) Explain why this estimator does *not* fit into the framework of SVMs for regression treated in Section 10.4.

ii) Show that this estimator corresponds to the least median of squares estimator for the special case of a linear kernel and $\lambda = 0$. Summarize the statistical properties of LMS from the literature.

Hint: See, e.g., Hampel (1975), Rousseeuw (1984), Rousseeuw and Leroy (1987), and Davies (1990, 1993).

10.9. Robustness properties of kernel-based LMS ($\star\star\star\star$)
Derive the robustness properties of $f^*_{\mathrm{P},\lambda}$ defined in Exercise (10.8).

Hint: Research.