Journal of Machine Learning Research 4 (2003) 1071-1105 Submitted 01/03; Published 11/03

Sparseness of Support Vector Machines

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

Editor: Nello Christianini

Abstract

Support vector machines (SVMs) construct decision functions that are linear combinations of kernel evaluations on the training set. The samples with non-vanishing coefficients are called support vectors. In this work we establish lower (asymptotical) bounds on the number of support vectors. On our way we prove several results which are of great importance for the understanding of SVMs. In particular, we describe to which "limit" SVM decision functions tend, discuss the corresponding notion of convergence and provide some results on the stability of SVMs using subdifferential calculus in the associated reproducing kernel Hilbert space.

Keywords: Computational learning theory, Pattern recognition, PAC model, Support vector machines, Sparseness

1. Introduction

Consider the binary classification problem where X is a set, $Y := \{-1,1\}$ and P is an *unknown* distribution on $X \times Y$. Let $T = ((x_1,y_1),\ldots,(x_n,y_n)) \in (X \times Y)^n$ be a sequence of i.i.d. pairs generated by P. The goal is to use the information of the *training set* T to predict the label y for any new observation x. To accomplish this goal a classifier is used to construct a decision function $f_T : X \to \mathbb{R}$. The prediction of $f_T(x)$ is sign $f_T(x)$.

The type of classifiers that we treat is based on one of the following optimization problems

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

or

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

Here, $T = \big((x_1,y_1),\ldots,(x_n,y_n)\big) \in (X \times Y)^n$ is a training set, $\lambda > 0$ is a *regularization parameter*, H is a reproducing kernel Hilbert space (RKHS) of a kernel k and L is a suitable convex loss function (cf. the following section for precise definitions). The additional term b in (2) is called the *offset*. The corresponding decision functions of these classifiers are $f_{T,\lambda}$ or $\tilde{f}_{T,\lambda} + \tilde{b}_{T,\lambda}$, respectively, where $f_{T,\lambda} \in H$ and $(\tilde{f}_{T,\lambda}, \tilde{b}_{T,\lambda}) \in H \times \mathbb{R}$ are *arbitrary* solutions of (1) and (2).

Common choices for L are the hinge loss function $L(y,t) := \max\{0, 1-yt\}$, the squared hinge loss function $L(y,t) := (\max\{0,1-yt\})^2$ and the least square loss function $L(y,t) := (1-yt)^2$. Figure 1 shows the shape of these loss functions. The corresponding classifiers are called L1-SVM,

L2-SVM, and LS-SVM, respectively. Note, that the former two are also called 1-norm and 2-norm soft margin classifiers while the latter is also known as regularization network if no offset is used. A thorough treatment of these classifiers can be found in the books of Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), and Suykens et al. (2002).

In practice the solutions to (1) and (2) are usually obtained indirectly by solving the duals. Recall that in these dual formulations the RKHS H only occurs implicitly via its kernel. The most popular kernels are the Gaussian RBF $k(x,x') = \exp(-\sigma^2\|x-x'\|_2^2)$ for $x,x' \in \mathbb{R}^d$ and fixed $\sigma > 0$ and polynomial kernels $k(x,x') = (\langle x,x'\rangle + c)^m$ for $x,x' \in \mathbb{R}^d$ and fixed $c \geq 0$, $m \in \mathbb{N}$. The latter will not be considered in this work as we shall justify below.

By the well-known representer theorem (see Kimeldorf and Wahba, 1971, Cox and O'Sullivan, 1990, Schölkopf et al., 2001) the solutions $f_{T,\lambda}$ and $\tilde{f}_{T,\lambda}$ of (1) and (2) are of the form

$$\sum_{i=1}^{n} \alpha_i k(x_i, .) \,, \tag{3}$$

where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ are suitable coefficients. Obviously, only *samples* x_i with $\alpha_i \neq 0$ have an influence on $f_{T,\lambda}$ or $\tilde{f}_{T,\lambda}$, respectively. Such samples are called *support vectors*. The major aim of this work is to establish lower bounds on the fraction of support vectors. Note, that in general the above representation is not unique but depends on the used optimization algorithm. In this paper we are interested in lower bounds which are independent of the specific representation. In particular, we bound the number of support vectors for optimization algorithms which produce the sparsest possible representations. Some standard algorithms such as all solvers for dual problems of the L1-SVM and the L2-SVM produce less sparse representations in general.

If k is a polynomial kernel it is easy to see that the representation (3) is not unique whenever the sample size n is too large. Namely, we can always find a representation such that the number of its support vectors does not exceed the (finite!) dimension of the RKHS of the polynomial kernel. Hence, depending on the algorithm used to find $f_{T,\lambda}$ or $\tilde{f}_{T,\lambda}$ it can happen that the fraction of support vectors tends to 0. For establishing nontrivial results we therefore restrict the class of considered kernels. It shall turn out that *universal* kernels (Steinwart, 2001), i.e. kernels whose RKHS is dense in the space of continuous functions over X, are an appropriate choice. Recall that among several others the Gaussian RBF kernel is universal (Steinwart, 2001). Besides the fact that universal kernels are often used in practice they also enjoy the property that classifiers based on (1) or (2) can "learn" under specific conditions on L and the behaviour of $\lambda = \lambda_n$ as shown by Steinwart (2002, 2003a), and Zhang (2004). Here "learning" is in the sense of universal consistency which guarantees that the probability for misclassifying a new sample (x,y) generated by P tends to the smallest possible value. To make this precise the *misclassification risk* of a measurable function $f : X \to \mathbb{R}$ is defined by

$$\mathcal{R}_P(f) \;:=\; P\big(\{(x,y) \in X \times Y : \operatorname{sign} f(x) \neq y\}\big) \,.$$

The smallest achievable misclassification risk $\mathcal{R}_P := \inf\big\{\mathcal{R}_P(f) \mid f : X \to \mathbb{R} \text{ measurable}\big\}$ is called the *Bayes risk* of P. A classifier is called universally consistent if the risks of its decision functions converge to the Bayes risk in probability for all P.

Now let us assume that the kernel used in (1) or (2) is universal. We shall see in Lemma 1 that in this case the representation (3) is almost surely unique under some mild conditions on the data generating distribution P. Furthermore, if the representation is unique it is obvious that the number of its support vectors is independent of the specific optimization algorithm used to solve (1) or (2).

Hence let us assume for a moment that the representation is almost surely unique. Under some conditions on L and λ_n our first main result (Theorem 9) then informally states

With probability tending to 1 when $n \to \infty$ the fraction of support vectors is essentially greater than the Bayes risk $\mathcal{R}_P$.

Note that the above mentioned assumptions on L and λ_n will almost coincide with the conditions ensuring universal consistency. Therefore, when a noisy classification problem is learned we should expect that the number of support vectors increases linearly in the sample size. Our second main result (Theorem 10) improves the lower bound if the loss function is differentiable. Informally, it states:

Let L be differentiable. With probability tending to 1 when $n \to \infty$ the fraction of support vectors can be essentially bounded from below by the probability of the set on which the labels are not noise free.

Note, that the probability of the set on which the labels are not noise free is always greater than or equal to $2\mathcal{R}_P$. In the extreme case where the noise does not vanish on the entire set (e.g. when the conditional class densities are Gaussian) the above probability even equals one, while the Bayes risk can be arbitrarily close to 0. In particular, the above statement shows that L2-SVMs—although using a margin—can produce decision functions that are far from being sparse.

In proving the above statements we shall establish several other important properties of SVMs and related algorithms. Namely, we shall show a general form of the representer theorem which in particular gives lower and upper bounds on the coefficients $\alpha_1, \ldots, \alpha_n$ in (3). Furthermore, we treat convergence issues for $f_{T,\lambda}$ and $\tilde{f}_{T,\lambda}$ interpreted both as an element of the RKHS and as a function. These results provide a better understanding of the asymptotic behaviour of support vector machines.

This work is organized as follows: In Section 2 we introduce all basic and advanced notions and discuss the latter. We then explain the main ideas of this work and informally present some auxiliary results such as the above mentioned quantified representer theorem and the convergence results. Finally, we state our main results and apply them to many common classifiers. The main work of this article is done in Section 3 which contains the proofs. For clarity's sake we have divided this section into several subsections most of them containing auxiliary results: Subsection 3.1 gives a brief overview on the subdifferential calculus of convex functions defined on Hilbert spaces. In Subsection 3.2 we show some basic properties of convex loss functions. A first convergence result which is independent of the considered algorithms is proved in Subsection 3.3. The following two subsections are devoted to a refinement of this convergence result: In Subsection 3.4 we prove a stability result for the classifiers based on (1) or (2). The main tool is the application of the subdifferential calculus to infinite-sample versions of (1) and (2). As a by-product, the quantified representer theorem is proved. In Subsection 3.5 we establish refined versions of the convergence result of Subsection 3.3. Finally, in Subsection 3.6 we prove our main results.

2. Bounding the Number of Support Vectors from Below

The aim of this section is to give an informal idea of the techniques of this work and to present the main results including several examples. We begin with a subsection which introduces some

basic concepts such as RKHS's, support vectors, and loss functions. For the latter we also discuss the behaviour of functions approximately minimizing the corresponding risk. In Subsection 2.2 we explain the main ideas of our work and informally state some auxiliary results which are of their own interest. In the last subsection we present our main results and apply them to many well-known classifiers including SVMs.

2.1 Kernels, Support Vectors, and Loss Functions

We will assume throughout this work that X is a compact metric space and P is a Borel probability measure on $X \times Y$, where Y is equipped with the discrete topology. We write $\overline{\mathbb{R}} := [-\infty,\infty]$, $\mathbb{R}^+ := [0,\infty)$ and $\overline{\mathbb{R}}^+ := [0,\infty]$. Given two functions $g,h : (0,\infty) \to (0,\infty)$ let $g \preceq h$ if there exists a constant $c > 0$ with $g(\varepsilon) \le c\,h(\varepsilon)$ for all sufficiently small $\varepsilon > 0$. We write $g \sim h$ if both $g \preceq h$ and $h \preceq g$.

For a positive definite kernel $k : X \times X \to \mathbb{R}$ we denote the corresponding RKHS (see Aronszajn, 1950, Berg et al., 1984) by H_k or simply H. For its closed unit ball we write B_H. Recall that the *feature map* $\Phi : X \to H$, $x \mapsto k(x,.)$ satisfies $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_H$ for all $x,x' \in X$ by the reproducing property. Moreover, k is continuous if and only if Φ is. In this case, H can be continuously embedded into the space of all continuous functions $C(X)$ via $I : H \to C(X)$ defined by $Iw := \langle w, \Phi(.) \rangle_H$, $w \in H$. Since we always assume that k is continuous, we sometimes identify elements of H as continuous functions on X. If the embedding $I : H \to C(X)$ has a dense image we call k a *universal* kernel (see Steinwart, 2001, Sect. 3). Recall that this is equivalent to the condition that for all $g \in C(X)$ and all $\varepsilon > 0$ there exists an element $f \in H$ with

$$\|f - g\|_\infty \;\le\; \varepsilon,$$

where $\|.\|_\infty$ denotes the supremum norm. For some kernels the RKHS is explicitely known and hence the universality for these kernels is easily checked (see Saitoh, 1997, Ritter, 2000). An elementary proof of the universality of the Gaussian RBF kernel was provided by Steinwart (2001, Sect. 3).

Let H be a RKHS with kernel k. For a function $f \in H$ the *minimal number of support vectors* is defined by

$$\#SV(f) := \min\Bigl\{ n \in \mathbb{N} \cup \{\infty\} : \exists \alpha_1,\ldots,\alpha_n \neq 0 \text{ and } x_1,\ldots,x_n \in X \text{ with } f = \sum_{i=1}^n \alpha_i k(x_i,.) \Bigr\}.$$

Note that we have $\#SV(f_{T,\lambda}) < \infty$ and $\#SV(\tilde{f}_{T,\lambda}) < \infty$ by the representer theorem. A representation of f is called *minimal* if it has $\#SV(f)$ support vectors. The next lemma which is proved in Subsection 3.6 characterizes minimal representations:

Lemma 1 *Let k be a universal kernel and $f = \sum_{i=1}^n \alpha_i k(x_i,.)$ be a representation of f. Then $\#SV(f) = n$ if and only if $x_1,\ldots,x_n$ are mutually different and $\alpha_i \neq 0$ for all $i = 1,\ldots,n$. Furthermore, minimal representations are unique up to permutations of indexes.*

In particular, if k is a universal kernel, $T = ((x_1,y_1),\ldots,(x_n,y_n))$ is a training set with mutually different $x_1,\ldots,x_n$ and $\sum_{i=1}^n \alpha_i k(x_i,.)$ is a representation of $f_{T,\lambda}$ or $\tilde{f}_{T,\lambda}$ with m support vectors then $\#SV(f_{T,\lambda}) = m$ or $\#SV(\tilde{f}_{T,\lambda}) = m$, respectively. If T contains repeated sample values, i.e. $x_i = x_j$ for some $i \neq j$, it can happen that the representation of the solution found by a specific algorithm is not minimal. Indeed, the dual optimization problems for the hinge loss or the squared hinge loss lead

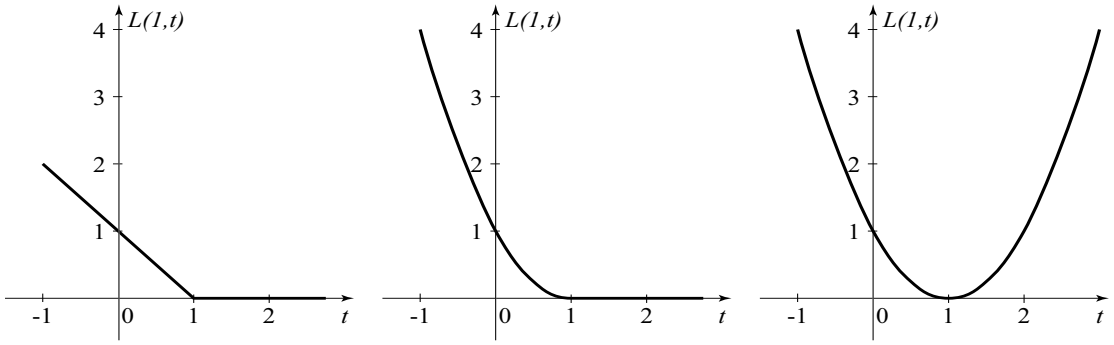

Figure 1: Some admissible loss functions for $y = 1$. Left: the hinge loss $L(y,t) := \max\{0, 1-yt\}$. Middle: the squared hinge loss $L(y,t) := (\max\{0, 1-yt\})^2$. Right: the least square loss $L(y,t) := (1-yt)^2$.

to algorithms which do not construct minimal representations in the presence of repeated sample values. However, the above lemma gives a simple way for minimizing a given representation: for all sample values x of T add all coefficients α_i with $x_i = x$ and call the sum α'_l. Then choose one sample x_j with $x_j = x$ as a representative, use α'_l as coefficient for $x'_l := x_j$, and remove all other samples x_i with $x_i = x$ from T. After this loop has been completed eliminate all samples x'_l with zero coefficient.

Downs et al. (2001) proposed a technique which finds samples that are linearly dependent in the RKHS in order to construct representations that are more sparse than the ones found by optimizing the dual of the L1-SVM optimization problem. For some data sets significant reductions were reported using a Gaussian RBF kernel. The above discussion shows that these reductions could only be achieved by either the existence of repeated sample values or numerical errors!

Let us now consider loss functions and their associated risks. We begin with a basic definition:

Definition 2 *A continuous function $L : Y \times \overline{\mathbb{R}} \to \overline{\mathbb{R}}^+$ with $L(Y,\mathbb{R}) \subset \mathbb{R}^+$ is called a* loss function. *Given a measurable function $f : X \to \overline{\mathbb{R}}$ and a Borel probability measure P on $X \times Y$ the* L-risk *of f is defined by*

$$\mathcal{R}_{L,P}(f) := \mathbb{E}_{(x,y)\sim P} L(y, f(x)) .$$

If P is the empirical measure corresponding to $T \in (X \times Y)^n$ we write $\mathcal{R}_{L,T}(.)$. Furthermore, we denote the smallest possible L-risk by

$$\mathcal{R}_{L,P} := \inf\{\mathcal{R}_{L,P}(f) \mid f : X \to \overline{\mathbb{R}} \text{ measurable}\} .$$

It is straightforward to see that not every loss function interacts well with the misclassification risk in the sense that an L-risk $\mathcal{R}_{L,P}(f)$ close to $\mathcal{R}_{L,P}$ guarantees that the misclassification risk $\mathcal{R}_P(f)$ is close to the Bayes risk. In the following we describe the class of loss functions which ensure that $\mathcal{R}_P(f_n) \to \mathcal{R}_P$ whenever $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}$. To this end write

$$C(\alpha,t) := \alpha L(1,t) + (1-\alpha) L(-1,t)$$

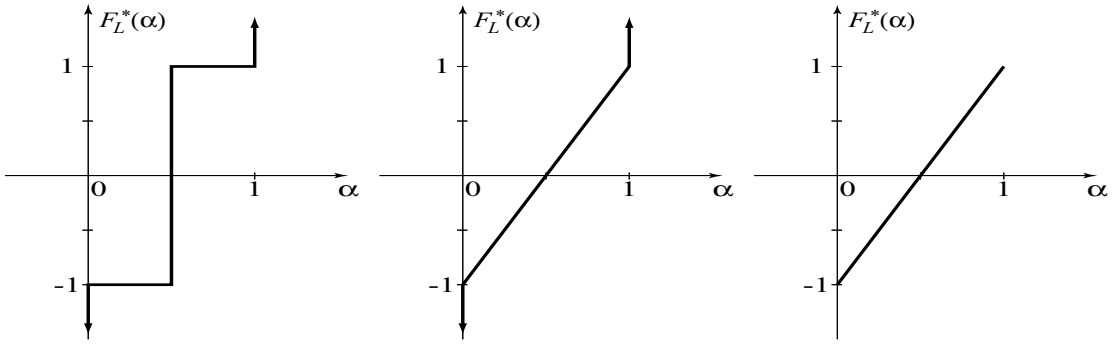

Figure 2: The minimizer F_L^* for some different loss functions. Left: for the hinge loss. Middle: for the squared hinge loss. Right: for the least square loss. Note that for the hinge loss function $F_L^*(\alpha)$ is not a singleton for $\alpha \in \{0, 1/2, 1\}$. For the squared hinge loss function this holds for $\alpha \in \{0, 1\}$.

for $\alpha \in [0,1]$ and $t \in \overline{\mathbb{R}}$. This function can be used to compute the L-risk of a measurable function $f : X \to \overline{\mathbb{R}}$ by

$$\begin{aligned} \mathcal{R}_{L,P}(f) &= \mathbb{E}_{(x,y)\sim P} L(y, f(x)) \\ &= \int_X P(1|x)L(1, f(x)) + P(-1|x)L(-1, f(x))\, P_X(dx) \\ &= \int_X C\big(P(1|x), f(x)\big)\, P_X(dx)\,. \end{aligned} \tag{4}$$

Here, P_X is the marginal distribution of P on X and $P(y|x)$ denotes a regular conditional probability (see Dudley, 2002). Equation (4) shows that we have to minimize the function $C(\alpha, .)$ for every $\alpha \in [0,1]$ in order to minimize the L-risk. This leads to the *set-valued* function F_L^* defined by

$$F_L^*(\alpha) := \Big\{ t \in \overline{\mathbb{R}} : C(\alpha, t) = \min_{s \in \overline{\mathbb{R}}} C(\alpha, s) \Big\}$$

for all $\alpha \in [0,1]$. For some standard loss functions F_L^* is illustrated in Figure 2. Obviously, given a *measurable* selection f^* of F_L^* the function $f^*(P(1|.))$ actually minimizes the L-risk, i.e.

$$\mathcal{R}_{L,P}\big(f^*(P(1|.))\big) = \mathcal{R}_{L,P}\,.$$

Recall that such a measurable selection f^* always exists (see Steinwart, 2003a). Furthermore, it was shown by Steinwart (2003a) that a loss function L interacts well with the misclassification loss in the above sense if $F_L^*(\alpha)$ only contains elements with a "correct" sign. The latter is formalized in the following definition:

Definition 3 *A loss function L is called* admissible *if for every $\alpha \in [0,1]$ we have*

$$\begin{aligned} F_L^*(\alpha) &\subset [-\infty, 0) \qquad \text{if} \quad \alpha < 1/2\,, \\ F_L^*(\alpha) &\subset (0, \infty,] \qquad \text{if} \quad \alpha > 1/2\,. \end{aligned}$$

Furthermore we say that L is strongly admissible *if it is admissible and* $\operatorname{card} F_L^*(\alpha) = 1$ *for all* $\alpha \in (0,1)$ *with* $\alpha \neq 1/2$.

It was proved by Steinwart (2003a) that L is admissible if and only if there are universally consistent classifiers based on (1). For classifiers based on (2) the admissibility is also necessary and sufficient apart from some technical conditions. A major step in establishing this characterization was to show that the admissibility of L is a sufficient (and necessary) condition for $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}$ implying $\mathcal{R}_P(f_n) \to \mathcal{R}_P$ for all P. In a recent paper of Bartlett et al. (2003) this implication is quantified in terms of an inequality between $\mathcal{R}_P(f_n) - \mathcal{R}_P$ and $\mathcal{R}_{L,P}(f_n) - \mathcal{R}_{L,P}$. For our aims we need a different kind of improvement. To this end let us assume for a moment that $F_L^*(\alpha)$ is a singleton for all but countably many $\alpha \in [0,1]$. Furthermore, let us suppose that $F_L^*(\alpha)$ is a—possibly degenerate—interval in $\mathbb{R}$ for all $\alpha \in [0,1]$. Then we shall prove the following result:

For all $\varepsilon > 0$ *and all sequences* (f_n) *of* $\mathbb{R}$*-valued measurable functions with* $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}$ *we have*

$$P_X\left(\left\{x \in X : \rho\big(f_n(x), F_L^*(P(1|x))\big) \geq \varepsilon\right\}\right) \;\to\; 0\,,$$

where $\rho\big(f_n(x), F_L^*(P(1|x))$ *denotes the distance of* $f_n(x)$ *to the set* $F_L^*(P(1|x))$.

The exact formulation which also gets rid of the assumption $F_L^*(\alpha) \subset \mathbb{R}$ is presented in Theorem 22. If $F_L^*(\alpha)$ is a singleton for all $\alpha \in [0,1]$ then the above statement shows that $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}$ implies $f_n \to F_L^*(P(1|.))$ in probability. In the general case, it states that with probability converging to 1 the functions f_n map into an ε-tube around the minimizer $F_L^*(P(1|.))$. In particular, since $\mathcal{R}_{L,P}(f_{T,\lambda_n}) \to \mathcal{R}_{L,P}$ and $\mathcal{R}_{L,P}(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}) \to \mathcal{R}_{L,P}$ (see Steinwart, 2003a) whenever k is universal and (λ_n) converges "slowly enough" to 0, this holds for the solutions of (1) and (2). The latter was already claimed by Lin (2002) and Zhang (2004) in order to explain the learning ability of SVMs.

We need some further definitions for loss functions: A loss function L is called *convex* if $L(y,.)$ is convex for $y = \pm 1$. A loss function is said to be *Lipschitz-continuous* if

$$|L|_1 \;:=\; \sup\left\{\frac{|L(y,t) - L(y,t')|}{|t - t'|} : y \in Y,\; t, t' \in \mathbb{R},\; t \neq t'\right\} \;<\; \infty\,.$$

Analogously, L is *locally* Lipschitz-continuous if $L_{|Y\times[-a,a]}$ is Lipschitz-continuous for all $a > 0$. Recall that convex loss functions are always locally Lipschitz-continuous (cf. Lemma 29). In order to treat classifiers that are based on (2) we need the following definition of Steinwart (2003a):

Definition 4 *An admissible loss function L is called* regular *if L is locally Lipschitz-continuous, $L(1,.)_{|(-\infty,0]}$ is monotone decreasing and unbounded, $L(-1,.)_{|[0,\infty)}$ is monotone increasing and unbounded and for all $\gamma > 0$ there exists a constant $c_\gamma > 0$ such that for all $a > 0$ we have*

$$\big|L_{|Y\times[-\gamma a,\gamma a]}\big|_1 \;\leq\; c_\gamma \big|L_{|Y\times[-a,a]}\big|_1 \tag{5}$$

$$\big\|L_{|Y\times[-\gamma a,\gamma a]}\big\|_\infty \;\leq\; c_\gamma \big\|L_{|Y\times[-a,a]}\big\|_\infty\,. \tag{6}$$

Note that convex admissible loss function are regular if (5) and (6) hold (cf. Subsection 3.2). Furthermore, it is easily checked that most of the loss functions considered in practice are regular (see the examples at the end of this section).

Finally, let H be a RKHS and L be a loss function. We define the *regularized L-risks* by

$$\begin{aligned} \mathcal{R}^{reg}_{L,P,\lambda}(f) &:= \lambda\|f\|^2_H + \mathcal{R}_{L,P}(f) \\ \mathcal{R}^{reg}_{L,P,\lambda}(f,b) &:= \lambda\|f\|^2_H + \mathcal{R}_{L,P}(f+b) \end{aligned}$$

for all $f \in H$, $b \in \mathbb{R}$ and all $\lambda > 0$. If P is the empirical measure corresponding to $T \in (X \times Y)^n$ we write $\mathcal{R}^{reg}_{L,T,\lambda}(.)$ and $\mathcal{R}^{reg}_{L,T,\lambda}(.,.)$, respectively. Note that $\mathcal{R}^{reg}_{L,T,\lambda}(.)$ is the objective function of (1) and $\mathcal{R}^{reg}_{L,T,\lambda}(.,.)$ coincides with the objective function of (2).

It was shown by Steinwart (2003a) that the regularized risks can always be minimized and that the minimizers are actually in a ball of a certain radius. In order to recall the exact formulation we define

$$\begin{aligned} \delta_\lambda &:= \sqrt{\frac{L(1,0)+L(-1,0)}{\lambda}} \\ L_\lambda &:= L_{|Y\times[-\delta_\lambda K, \delta_\lambda K]} , \end{aligned}$$

where k is a kernel, $K := \sup\{\sqrt{k(x,x)} : x \in X\}$, and L is a loss function. These quantities will be used throughout the text. Now, the results of Steinwart (2003a) are:

Lemma 5 *Let L be an admissible loss function and H be a RKHS of continuous functions. Then for all Borel probability measures P on $X \times Y$ and all $\lambda > 0$ there exists an element $f_{P,\lambda} \in H$ with*

$$\mathcal{R}^{reg}_{L,P,\lambda}(f_{P,\lambda}) \;=\; \inf_{f\in H} \mathcal{R}^{reg}_{L,P,\lambda}(f) \,.$$

Moreover, for all such minimizing elements $f_{P,\lambda} \in H$ we have $\|f_{P,\lambda}\| \le \delta_\lambda$.

For classifiers based on (2) we have to exclude *degenerate Borel* probability measures, i.e. measures with

$$P_X\big(x \in X : P(y|x) = 1\big) \;=\; 1$$

for $y = 1$ or $y = -1$ in order to ensure that the offset is finite:

Lemma 6 *Let L be a regular loss function and H be a RKHS of continuous functions. Then for all non-degenerate Borel probability measures P on $X \times Y$ and all $\lambda > 0$ there exists a pair $(\tilde f_{P,\lambda}, \tilde b_{P,\lambda}) \in H \times \mathbb{R}$ with*

$$\mathcal{R}^{reg}_{L,P,\lambda}(\tilde f_{P,\lambda}, \tilde b_{P,\lambda}) \;=\; \inf_{\substack{f\in H \\ b\in\mathbb{R}}} \mathcal{R}^{reg}_{L,P,\lambda}(f,b) \,.$$

Moreover, for all such minimizing pairs $(\tilde f_{P,\lambda}, \tilde b_{P,\lambda}) \in H \times \mathbb{R}$ we have $\|\tilde f_{P,\lambda}\| \le \delta_\lambda$.

2.2 Towards a Proof of the Bounds: Subdifferentials, Stability and a Quantified Representer Theorem

In this subsection we recall the definition of subdifferentials. We then outline the roadmap of our proofs stating the main steps informally and discussing their relation.

As already mentioned we are mainly interested in convex loss functions. Unfortunately, not all of the convex loss functions used in practice are differentiable. In particular, the hinge loss function which is used in the L1-SVM is not. In order to treat non-differentiable convex loss functions we need the notion of subdifferentials:

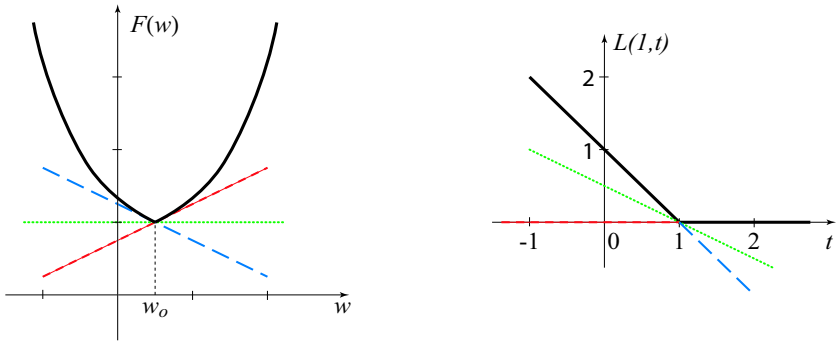

Figure 3: Left: The subdifferential of F at the point w_0 describes the affine hyperplanes that are dominated by F and are equal to it at w_0. In particular the left (long-dashed/blue) and right (short-dashed/red) derivates are included. At a minimum, the 0-hyperplane (dotted/green) is also included. Right: The subdifferential of the hinge loss function at $y = t = 1$ describes all hyperplanes with linear coefficient between -1 (dashed/blue) and 0 (solid/red)

Definition 7 *Let H be a Hilbert space, $F : H \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in H$ with $F(w) \neq \infty$. Then the* subdifferential *of F at w is defined by*

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

If F is (Gâteaux) differentiable at w then $\partial F(w)$ contains only the derivate of F at w (see Phelps, 1986, Prop. 1.8.). For a geometric interpretation of subdifferentials we refer to Figure 3. Furthermore, the subdifferential enjoys all properties from basic calculus such as linearity and (certain) chain rules. These properties are listed in Subsection 3.1. Given a subset A of H we often use the notation

$$\partial F(A) := \bigcup_{w \in A} \partial F(w) .$$

The following definition plays an important role in the investigation of subdifferentials:

Definition 8 *A set-valued function $F : H \to 2^H$ on a Hilbert space H is said to be a* monotone operator *if for all $v, w \in H$ and all $v^* \in F(v)$, $w^* \in F(w)$ we have*

$$\langle v^* - w^*, v - w \rangle \geq 0 .$$

It is an easy exercise to show that the subdifferential map $w \mapsto \partial F(w)$ of a continuous convex function $F : H \to \mathbb{R}$ on a Hilbert space H is a monotone operator. Slightly more difficult is the following result that we shall establish in Lemma 20:

If L is a convex admissible loss function then $\alpha \mapsto F_L^*(\alpha)$ is a monotone operator, i.e. $\alpha \leq \alpha'$ implies $t \leq t'$ for all $t \in F_L^*(\alpha)$, $t' \in F_L^*(\alpha')$. In particular, $F_L^*(\alpha)$ is a singleton for all but at most countably many α.

The main reason for considering subdifferentials is to "differentiate" the objective functions $\mathcal{R}^{reg}_{L,T,\lambda}(.)$ and $\mathcal{R}^{reg}_{L,T,\lambda}(.,.)$ of (1) and (2) and their infinite sample versions $\mathcal{R}^{reg}_{L,P,\lambda}(.)$ and $\mathcal{R}^{reg}_{L,P,\lambda}(.,.)$. In the empirical case this immediately leads to the announced quantified representer theorem (cf. Remark 32):

There exists a representation $f_{T,\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, .)$ *of the solution of (1) with*

$$\alpha_i \in -\frac{1}{2n\lambda} \partial_2 L(y_i, f_{T,\lambda}(x_i)) \tag{7}$$

for all $i = 1, \ldots, n$*. Here,* $\partial_2 L$ *denotes the subdifferential operator of* L *with respect to the second variable. The same holds for the solution* $\tilde{f}_{T,\lambda}$ *of (2).*

Note that for differentiable loss functions this result is trivial. The fundamental observation from the quantified representer theorem is that x_i must be a support vector of the described representation whenever $0 \notin \partial_2 L(y_i, f_{T,\lambda}(x_i))$. Now recall from the previous subsection that $f_{T,\lambda}$ converges to $F^*_L(P(1|.))$. Hence a natural idea suggests that x_i must be a support vector of the above representation whenever $0 \notin \partial_2 L\big(y_i, \partial_2 L(y_i, F^*_L(P(1|x_i)) \cap \mathbb{R})\big)$. Of course this is only a vague reasoning and several issues have to be resolved when making a rigorous proof from this line. In particular we mention the following questions:

- Does $0 \notin \partial_2 L(y_i, t)$ imply $0 \notin \partial_2 L(y_i, s)$ for all s suitable close to t?
- Recall that $f_{T,\lambda}$ only converges to $F^*_L(P(1|.))$ in "probability". Hence there can always be a small "bad" set B_T on which $f_{T,\lambda}$ is not close to $F^*_L(P(1|.))$. Even if the first question can be positively answered we cannot apply it for samples contained in B_T. Furthermore, the bad set B_T depends on T and hence there is no straightforward method to show that the fraction of samples of T in B_T is small (cf. Figure 4). Therefore we have to answer: How likely is it that many samples of T are contained in the bad set B_T?

Besides these issues there occur others due to the set valued nature of F^*_L. Although all of these can be resolved their solutions sometimes require some technical analysis as we will see in the last section. In order to get a smooth description of the main ansatz of this work we do not go into the details here.

As we shall see in Lemma 21 there exists a positive answer to the first question raised above which meets our requirements. Furthermore note that the first question can immediately be positively answered whenever L is differentiable since convex differentiable functions are always continuously differentiable. In the general case we shall use the fact that the subdifferential mapping of a convex function is semi-continuous (cf. Proposition 15).

The solution of the second problem gives a deeper insight of the asymptotic behaviour of SVMs and hence we briefly describe the major idea: If the bad sets $B = B_T$ were independent of T it would be rather easy to see that the T's that have a large fraction of samples in B are unlikely for large sample sizes n. In order to find such an data-independent bad set we first show the following stability result (cf. Theorem 28):

For all $\lambda > 0$ *and all probability measures* P *and* Q *their exists a bounded function* $h : X \times Y \to \mathbb{R}$ *independent of* Q *which satisfies*

$$\| f_{P,\lambda} - f_{Q,\lambda} \| \leq \frac{\| \mathbb{E}_P h\Phi - \mathbb{E}_Q h\Phi \|}{\lambda} \,,$$

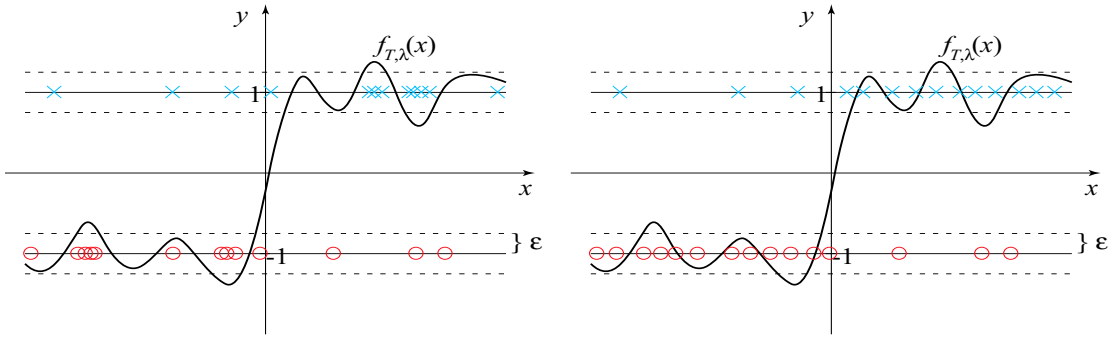

Figure 4: Asymptotic behaviour of the L1-SVM for a simple noisy problem with minimizer $F_L^*(P(1|x)) = \operatorname{sign} x$. Left: although the decision function $f_{T,\lambda}$ essentially maps into an ε-tube around the minimizer, most samples are in B_T, i.e. they are mapped outside the ε-tube. We will see that this situation is not typically. Right: the "good case" which we (have to) ensure with high probability. The decision function $f_{T,\lambda}$ maps most of the samples into an ε-tube around $F_L^*(P(1|x))$.

where Φ is the feature map of the used kernel. A similar but weaker result holds for the solutions of the optimization problems with offset.

This stability result is proved using the subdifferential of $\mathcal{R}^{reg}_{L,P,\lambda}$. Letting Q be empirical measures the right side can be bounded by a concentration inequality for Hilbert space valued random variables (cf. Proposition 33 and Lemma 34). Informally speaking, this yields:

The empirically found solutions $f_{T,\lambda}$ converge to the infinite sample solution $f_{P,\lambda}$ in the RKHS in probability. Under some additional assumptions on P and L (cf. Proposition 37) the same holds for the solutions of (2).

Now, let us assume that our kernel is universal. Then it was shown by Steinwart (2003a) that $\mathcal{R}_{L,P}(f_{P,\lambda_n}) \to \mathcal{R}_{L,P}$ whenever $\lambda_n \to 0$. Hence f_{P,λ_n} converges to $F_L^*(P(1|.))$ "in probability". Furthermore, if $\lambda_n \to 0$ "slowly" then $\|f_{T,\lambda_n} - f_{P,\lambda_n}\|_\infty \to 0$ in probability by our first step. Therefore we find (cf. Proposition 35 and Proposition 38):

Let us assume that (λ_n) tends "slowly" to 0. Then the bad sets B_T of f_{T,λ_n} are contained in the bad sets B_n of f_{P,λ_n} with high probability. The probability of the latter tends to 0.

As a consequence the probability of training sets T which have more than $\varepsilon|T|$ samples in B_T tends to 0 for all $\varepsilon > 0$. Figure 4 illustrates this fact. Figure 5 shows the situation in the feature space.

Let $X_{cont} := \{x \in X : P_X(\{x\}) = 0\}$. Continuing our motivation the previous considerations show that we should expect that most of the samples in

$$S := \left\{(x,y) \in X_{cont} \times Y : 0 \notin \partial_2 L\big(y, F_L^*(P(1|x)) \cap \mathbb{R}\big)\right\} \tag{8}$$

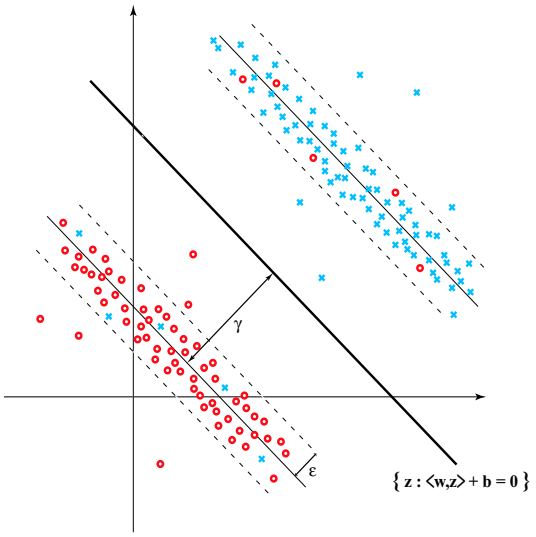

Figure 5: Asymptotic behaviour of the L1-SVM in the feature space of a universal kernel for a noisy problem: the samples are concentrated around the functional margin $\gamma = 1$. Note that the graphic is only a low-dimensional projection of the situation in the RKHS since actually the samples are linearly independent in the RKHS. If the distribution P is not everywhere noisy it may happen that a larger fraction of samples are far beyond the margin. This case corresponds to the illustrations usually presented in text books on SVMs.

are support vectors. Recall that in view of Lemma 1 we have to exclude the set $X \setminus X_{cont}$ in which repeated sample values can occur. Furthermore, for typical training sets the fraction of samples in S is close to the probability of S. Hence, for a convex loss function L and a Borel probability measure P on $X \times Y$ we define

$$\mathcal{S}_{L,P} := \begin{cases} P(S) & \text{if } 0 \notin \partial_2 L(1, F_L^*(1/2)) \cap \partial_2 L(-1, F_L^*(1/2)) \\ P(S) + \frac{1}{2} P_X(X_0 \cap X_{cont}) & \text{otherwise.} \end{cases} \tag{9}$$

Here, we write $X_0 := \{x \in X : P(1|x) = 1/2\}$. Note that for convex admissible loss functions we have $0 \notin \partial_2 L(1, F_L^*(\alpha)) \cap \partial_2 L(-1, F_L^*(\alpha))$ for all $\alpha \neq 1/2$ (cf. Lemma 20). However, for the hinge loss function $0 \in \partial_2 L(1, F_L^*(1/2)) \cap \partial_2 L(-1, F_L^*(1/2))$ holds and this is the major source of many technical problems handled in Section 3, which hide the outlined main idea of the proofs.

2.3 Results and Examples

In this subsection we state our main results, which asymptotically bound the number of support vectors from below by the quantity $\mathcal{S}_{L,P}$ introduced in the previous subsection. Then we give lower bounds for $\mathcal{S}_{L,P}$ both in the general case and for differentiable loss functions. Finally we apply our results to several well known examples.

The first theorem treats classifiers based on (1). Its proof as well as the proofs of the following results can be found in Subsection 3.6.

Theorem 9 *Let L be an admissible and convex loss function, k be a universal kernel and $\lambda_n > 0$ be a regularization sequence with $\lambda_n \to 0$ and $n\lambda_n^2/|L_{\lambda_n}|_1^2 \to \infty$. Then for all Borel probability measures P on $X \times Y$ and all $\varepsilon > 0$ the classifier based on (1) with respect to k, L and (λ_n) satisfies*

$$\Pr^*\Big(T \in (X \times Y)^n : \#SV(f_{T,\lambda_n}) \geq (\mathcal{S}_{L,P} - \varepsilon)n\Big) \to 1 \,.$$

Here, $\Pr^*$ *denotes the outer probability measure of P^n in order to avoid measurability considerations.*

The next theorem establishes an analogous result for classifiers based on (2). Because of the offset we have to exclude degenerate probability measures P which have been introduced in Subsection 2.1. It is obvious that for such probability measures $\tilde{f}_{T,\lambda} = 0$ holds for almost all T. In particular, we have $\#SV(\tilde{f}_{T,\lambda_n}) = 0$ in this case.

Theorem 10 *Let L be a strongly admissible, regular and convex loss function, k be a universal kernel and $\lambda_n > 0$ be a regularization sequence with $\lambda_n \to 0$, $n\lambda_n^3/|L_{\lambda_n}|_1^2 \to \infty$ and $n\lambda_n/\big(\|L_{\lambda_n}\|_\infty^2 |L_{\lambda_n}|_1^2 \log n\big) \to \infty$. Then for all non-degenerate Borel probability measures P on $X \times Y$ and all $\varepsilon > 0$ the classifier based on (2) with respect to k, L and (λ_n) satisfies*

$$\Pr^*\Big(T \in (X \times Y)^n : \#SV(\tilde{f}_{T,\lambda_n}) \geq (\mathcal{S}_{L,P} - \varepsilon)n\Big) \to 1 \,.$$

We like to remark that in the previous theorem it is not necessary to require *strongly* admissible loss functions. Indeed, the result holds for regular convex loss functions, too. However, the proof for the latter is even more technical than the proof of Theorem 10. This, together with the fact that every loss function of practical interest (cf. the examples below) is strongly admissible motivated us to state the above theorem in its less general form.

The following propositions provide lower bounds on $\mathcal{S}_{L,P}$ for important types of loss functions. We begin with:

Proposition 11 *Let L be a convex admissible loss function and P be a Borel probability measure on $X \times Y$. Then we have*

$$\mathcal{S}_{L,P} \;\geq\; \inf\Big\{P\big((x,y) \in X_{cont} \times Y : f(x) \neq y\big) \;\big|\; f : X \to Y \text{ measurable}\Big\} \,.$$

In particular, $\mathcal{S}_{L,P} \geq \mathcal{R}_P$ holds whenever $X_{cont} = X$.

Roughly speaking, the above result together with Theorem 9 and Theorem 10 gives lower bounds for the number of support vectors for uniformly consistent classifiers based on (1) or (2), respectively. Namely, the proposition shows that we cannot expect less than $n\mathcal{R}_P$ support vectors for such classifiers if $X_{cont} = X$. Recall that it is also well-known from experiments that the sparseness of SVMs heavily depends on the noise of the underlying distribution. The next proposition improves the lower bound on $\mathcal{S}_{L,P}$ for differentiable loss functions:

Proposition 12 *Let L be a convex admissible and differentiable loss function and P be a Borel probability measure on $X \times Y$. Then we have*

$$\mathcal{S}_{L,P} \;\geq\; P_X\big(x \in X_{cont} : 0 < P(1|x) < 1\big) \,.$$

Roughly speaking, this proposition shows that for differentiable loss functions the fraction of support vectors is essentially lower bounded by the probability of the set of points in which noise occurs. In particular, even if we have a small Bayes risk we cannot expect sparse representations in general.

Together with our main theorems Proposition 12 also throws new light on the role of the margin in SVMs: namely, it is not only the margin that gives sparse decision functions but the *whole* shape of the loss function. Indeed, comparing the squared hinge loss function (cf. the examples below) and the least square loss function we obtain the same bad lower bounds on the number of support vectors. Only in noiseless regions sparse representations seem to be more likely using the squared hinge loss function since unlike the squared loss function this loss function does not penalize samples with margin > 1.

We conclude this section by some important examples of classifiers based on (1) and (2):

Example 1 L1-SVMs without offset *are based on the minimization problem (1) with the* hinge loss function $L(y,t) := \max\{0, 1-yt\}$. *The conditions on* (λ_n) *formulated in Theorem 9 reduce to* $\lambda_n \to 0$ *and* $n\lambda_n^2 \to \infty$. *Then, applying Proposition 12 yields lower bounds on the number of support vectors. In particular, the number of support vectors is asymptotically greater than* $n\mathcal{R}_P$ *in the case of* $X_{cont} = X$. *We conjecture that this lower bound can be replaced by* $2n\mathcal{R}_P$. *In order to explain this conjecture recall that L1-SVMs produce the same set of decision functions as the so-called* ν*-SVMs (see Schölkopf et al., 2000)). Furthermore, as shown by Steinwart (2003b) an asymptotically optimal value for the regularization parameter* ν *is* $2\mathcal{R}_P$. *Recalling that* ν *is also a lower bound on the fraction of support vectors (see Schölkopf et al., 2000) leads to our conjecture.*

Example 2 L1-SVMs with offset *are based on (2) and the hinge loss function. The corresponding conditions on* (λ_n) *of Theorem 10 can be unified to* $\lambda_n \to 0$ *and* $n\lambda_n^3 \to \infty$. *Of course, applying Proposition 12 yields the same lower bound as for the L1-SVM without offset. However, if the distribution is in a certain sense unbalanced this bound can be improved: for simplicity we suppose* $X_{cont} = X$ *and* $X_0 = \emptyset$. *We define* $X_1 := \{x \in X : P(1|x) > 1/2\}$ *and* $X_{-1} := \{x \in X : P(1|x) < 1/2\}$. *Recall that these sets are the classes which have to be approximated by the classifier. Furthermore, we define* $X_i^j := X_i \times \{j\}$ *for* $i, j \in \{-1, 1\}$. *Under the assumptions of Theorem 10 we then obtain (cf. the end of Section 3.6 for a sketch of the proof)*

$$\Pr^* \Big(T \in (X \times Y)^n : \#SV(\tilde{f}_{T,\lambda_n}) \geq \big(\mathcal{R}_{L,P} + |P(X_{-1}^1) - P(X_1^{-1})| - \varepsilon\big) n \Big) \to 1 \qquad (10)$$

for L1-SVMs with offset. In particular, if -1*-noise and* 1*-noise do not have the same probability, i.e.* $|P(X_{-1}^1) - P(X_1^{-1})| > 0$ *then (10) improves the result of Theorem 10. If either* $P(X_{-1}^1) = 0$ *or* $P(X_1^{-1}) = 0$ *the lower bound in (10) becomes* $2n\mathcal{R}_P$ *which also corroborates our belief described in the previous example.*

Example 3 L2-SVMs without offset *are based on the minimization problem (1) with the squared hinge loss function, i.e.* $L(y,t) := \big(\max\{0, 1-yt\}\big)^2$. *The conditions on* (λ_n) *formulated in Theorem 9 are* $\lambda_n \to 0$ *and* $n\lambda_n^3 \to \infty$. *The value of* $\mathcal{S}_{L,P}$ *can be estimated by Proposition 12.*

Example 4 L2-SVMs with offset *are based on the minimization problem (2) with the squared hinge loss function. The conditions on* (λ_n) *of Theorem 10 can be unified to* $\lambda_n \to 0$ *and* $n\lambda_n^4 / \log n \to \infty$. *If k is a* C^∞*-kernel the latter can be replaced by the slightly weaker condition* $n\lambda_n^4 \to \infty$ *(see Steinwart, 2003a, for details). Again, the value of* $\mathcal{S}_{L,P}$ *can be estimated by Proposition 12.*

Example 5 Least square support vector machines *are based on (2) with the squared loss function, i.e.* $L(y,t) := (1-yt)^2$. *The conditions on* (λ_n) *are the same as for L2-SVMs with offset. As above, the value of* $\mathcal{S}_{L,P}$ *can be estimated by Proposition 12.*

Example 6 Regularization networks or kernel ridge regression classifiers *are based on the minimization problem (1) with the squared loss function. The conditions on the regularization sequence coincide with the conditions for the L2-SVMs without offset. Again, the value of* $\mathcal{S}_{L,P}$ *can be estimated by Proposition 12.*

Example 7 R1-SVMs for classification *are based on either (2) or (1) using the* ε*-insensitive loss function* $L_\varepsilon(y,t) := \max\{0, |y-t| - \varepsilon\}$ *for some* $0 \le \varepsilon < 1$. *Our results coincide with the results for the L1-SVM with or without offset, respectively.*

Example 8 R2-SVMs for classification *are based on either (2) or (1) using the squared* ε*-insensitive loss function* $L_\varepsilon(y,t) := \big(\max\{0, |y-t| - \varepsilon\}\big)^2$ *for some* $0 \le \varepsilon < 1$. *Our results coincide with the results for the L2-SVM with or without offset, respectively.*

Example 9 *One can also consider classifiers based on (1) or (2) using the logistic loss function* $L(y,t) := \log(1+\exp(-yt))$. *With the help of Remark 32 we easily see that the lower bounds of Theorem 9 and Theorem 10 hold with* $\mathcal{S}_{L,P} = P_X(X_{cont})$ *for* all *regularization sequences* (λ_n). *In particular, if* $X_{cont} = X$ *we have* $\#SV(f_{T,\lambda}) = \#SV(\tilde{f}_{T,\lambda}) = n$ *for almost all training sets* T *of length* n *and all* $\lambda > 0$.

Remark 13 *In a recent paper of Steinwart (2004, to appear) some of the above results were significantly improved. In particular, the conjecture for the L1-SVM was proved. Furthermore, it was shown that this bound is in some sense optimal in many situations while the bound for the LS-SVM is, in general, too loose.*

3. Proofs

In this part of our article the main work is done. In particular we give exact formulations of the informally stated results and rigorously prove all results. The proofs are self contained, however, the some basic knowledge in functional analysis and probability theory is required.

3.1 Subdifferentials

In this subsection we collect some important properties of subdifferentials. Throughout this subsection H denotes a Hilbert space. We begin with a proposition that provides some elementary facts of the subdifferential (see Phelps, 1986, Prop. 1.11.):

Proposition 14 *The subdifferential* $\partial F(w)$ *of a convex function* $F : H \to \mathbb{R} \cup \{\infty\}$ *is a non-empty, convex and weak*-compact subset of* H *for all* $w \in H$ *where* F *is continuous and finite. If* F *is Lipschitz-continuous we also have* $\|w^*\| \le |F|_1$ *for all* $w^* \in \partial F(w)$ *and all* $w \in H$.

The next proposition shows that the subdifferential is in some sense semi-continuous (see Phelps, 1986, Prop. 2.5, for a proof):

Proposition 15 *If $F : H \to \mathbb{R}$ is continuous and convex then the subdifferential map $w \mapsto \partial F(w)$ is norm-to-weak* upper semi-continuous. In particular, if $\dim H < \infty$ then for all $w \in H$ and all $\varepsilon > 0$ there exists a $\delta > 0$ with*

$$\partial F(w + \delta B_H) \subset \partial F(w) + \varepsilon B_H .$$

The following result characterizes minima of convex functions (see Phelps, 1986, Prop. 1.26, for a proof):

Proposition 16 *The function F has a global minimum at $w \in H$ if and only if $0 \in \partial F(w)$.*

We are mainly interested in the calculus of subdifferentials. We begin with the linearity of subdifferentials (see, for example, Phelps, 1986, Thm. 3.16):

Proposition 17 *Let $\lambda \geq 0$ and $F, G : H \to \mathbb{R}$ be convex lower-semicontinuous functions such that G is continuous in at least one point. Then for all $w \in H$ we have:*

i) $\partial(\lambda F)(w) = \lambda \partial F(w)$

ii) $\partial(F + G)(w) = \partial F(w) + \partial G(w)$.

The following proposition provides a chain rule for subdifferentials (see Romano, 1995, for a discussion):

Proposition 18 *Let H_1, H_2 be Hilbert spaces, $A : H_1 \to H_2$ be a bounded and linear operator and $F : H_2 \to \mathbb{R} \cup \{\infty\}$ be a convex function that is finite and continuous in 0. Then for all $w \in H_1$ we have*

$$\partial(F \circ A)(w) = A^* \partial F(Aw),$$

where A^* denotes the adjoint operator of A.

3.2 Some Technical Lemmas

The following lemma collects some simple but nevertheless useful facts about convex admissible loss functions:

Lemma 19 *Let L be a convex admissible loss function. Then L is locally Lipschitz-continuous and*

i) $\partial_2 L(1,0) \subset (-\infty, 0)$ *and* $\partial_2 L(-1,0) \subset (0, \infty)$.

ii) for all $t \in \mathbb{R}$ we have

$$0 \notin \partial_2 L(1,t) \cap \partial_2 L(-1,t) . \tag{11}$$

iii) for all bounded subsets $A \subset \mathbb{R}$ there exists an $\varepsilon > 0$ such that for all $t \in A$ we have

$$0 \notin \partial_2 L(1, t + \varepsilon B_{\mathbb{R}}) \cap \partial_2 L(-1, t + \varepsilon B_{\mathbb{R}}) . \tag{12}$$

Proof *i*): Let us suppose that there exist an $s \in \partial_2 L(1,0)$ with $s \geq 0$. If $s = 0$ then $0 \in \partial_2 C(1,0)$ and hence $0 \in F_L^*(1)$ which contradicts the admissibility. If $s > 0$ then $s' > 0$ for all $s' \in \partial_2 L(1,t)$, $t > 0$, by the monotony of the subdifferential. Therefore $L(1,.)$ is monotonously increasing on $(0, \infty)$. This yields $F_L^*(1) \cap (0, \infty] = \emptyset$ which also contradicts the admissibility. The second assertion is proved

analogously.
ii): Let us suppose that there exist a $t \in \mathbb{R}$ with $0 \in \partial_2 L(1,t) \cap \partial_2 L(-1,t)$. Then we find

$$0 \in \partial_2\big(\alpha L(1,t) + (1-\alpha)L(-1,t)\big) = \partial_2 C(\alpha,t)$$

for all $\alpha \in [0,1]$. This leads to $t \in F_L^*(\alpha)$ for all $\alpha \in [0,1]$ which contradicts the admissibility of L.
iii): Let us assume that (12) is false. Then for all $n \geq 1$ there exists $t_n \in A$ and $\delta_n, \delta'_n \in [-1/n, 1/n]$ with

$$0 \in \partial_2 L(1, t_n + \delta_n) \cap \partial_2 L(-1, t_n + \delta'_n) \; .$$

Since A is bounded we may assume without loss of generality that (t_n) converges to an element $t \in \mathbb{R}$ (otherwise we have to consider a convergent subsequence in the following). Then, given an arbitrary $\varepsilon > 0$ we find by Proposition 15

$$0 \;\in\; \partial_2 L(1, t_n + \delta_n) \cap \partial_2 L(-1, t_n + \delta'_n) \;\subset\; \big(\partial_2 L(1,t) + \varepsilon B_{\mathbb{R}}\big) \cap \big(\partial_2 L(-1,t) + \varepsilon B_{\mathbb{R}}\big)$$

for all sufficiently large n. This leads to

$$0 \in \bigcap_{\varepsilon > 0} \big(\partial_2 L(1,t) + \varepsilon B_{\mathbb{R}}\big) \;\cap\; \bigcap_{\varepsilon > 0} \big(\partial_2 L(-1,t) + \varepsilon B_{\mathbb{R}}\big).$$

Since the subdifferentials $\partial_2 L$ are compact the latter implies $0 \in \partial_2 L(1,t) \cap L(-1,t)$ which contradicts (11). ∎

Note that *i*) of the above Lemma was also observed by Bartlett et al. (2003). The next lemma collects some important facts about the solution operator F_L^* for convex admissible loss functions L:

Lemma 20 *For a convex admissible loss function L the following properties hold*

- *i)* $F_L^*(\alpha)$ *is a bounded, closed interval in* $\mathbb{R}$ *for all* $\alpha \in (0,1)$.
- *ii) for all* $\alpha \in [0,1]$ *and all* $t \in F_L^*(\alpha) \cap \mathbb{R}$ *there exist* $s_1 \in \partial_2 L(1,t)$ *and* $s_{-1} \in \partial_2 L(-1,t)$ *with* $s_1 \leq 0 \leq s_{-1}$.
- *iii) for all* $\alpha \in [0,1]$, *all* $t \in F_L^*(\alpha) \cap \mathbb{R}$ *and all* $\alpha' \in [0,1]$ *with* $\alpha' > \alpha$ *there exists an* $s \in \partial_2 C(\alpha', t)$ *with* $s < 0$.
- *iv)* $\alpha \mapsto F_L^*(\alpha)$ *is a monotone operator*
- *v)* $\operatorname{card} F_L^*(\alpha) > 1$ *for at most countably many* $\alpha \in [0,1]$.
- *vi) for all* $t \in F_L^*(1/2)$ *we have*

$$\begin{aligned} 0 \in \partial_2 L(1,t) \quad &\Rightarrow \quad t = \max F_L^*(1/2) \\ 0 \in \partial_2 L(-1,t) \quad &\Rightarrow \quad t = \min F_L^*(1/2) \end{aligned}$$

- *vii) let* $\alpha \in [0,1]$ *with* $0 \in \partial_2 L\big(1, F_L^*(\alpha)\big) \cap \partial_2 L\big(-1, F_L^*(\alpha)\big)$. *Then we have* $\alpha = 1/2$ *and* $\operatorname{card} F_L^*(1/2) > 1$.

Proof *i*): By Lemma 19 we know $s < 0$ for all $s \in \partial_2 L(1,0)$ and thus the definition of the subdifferential leads to $L(1,-\infty) = \infty$. Therefore, we find $-\infty \notin F_L^*(\alpha)$ for all $0 < \alpha < 1$. Analogously we can show $\infty \notin F_L^*(\alpha)$ for all $0 < \alpha < 1$. Moreover, $F_L^*(\alpha)$ is a compact subset of $\overline{\mathbb{R}}$ and therefore the previous considerations show that $F_L^*(\alpha)$ is closed and bounded for all $0 < \alpha < 1$. Since $C(\alpha,.)$ is convex it is also clear that $F_L^*(\alpha)$ is an interval.
ii): For given $t \in F_L^*(\alpha) \cap \mathbb{R}$ there exists an $s_1 \in \partial_2 L(1,t)$ and an $s_{-1} \in \partial_2 L(-1,t)$ with $0 = \alpha s_1 + (1-\alpha)s_{-1}$. If $\alpha = 1$ we find $s_1 = 0$ and $t > 0$ by the admissibility of L. The latter yields $s_{-1} > 0$ by the monotony of the subdifferential and Lemma 19. The case $\alpha = 0$ can be treated analogously. Hence, it suffices to consider the case $0 < \alpha < 1$. Then we have $s_{-1} = -\frac{\alpha}{1-\alpha} s_1$ which leads to either $s_{-1} \le 0 \le s_1$ or $s_1 \le 0 \le s_{-1}$. Since the monotony of the subdifferential and Lemma 19 yield that $s_1 \ge 0$ implies $t > 0$ and that $s_{-1} \le 0$ implies $t < 0$ we finally find the assertion.
iii): Let $\alpha \in [0,1]$ and $t \in F_L^*(\alpha) \cap \mathbb{R}$. Without loss of generality we may assume $\alpha < 1$. Let us fix $s_1 \in \partial_2 L(1,t)$ and $s_{-1} \in \partial_2 L(-1,t)$ according to *ii*). Then we find $s_1 - s_{-1} < 0$ by Lemma 19 and hence

$$s \;:=\; \alpha' s_1 + (1-\alpha')s_{-1} \;<\; \alpha s_1 + (1-\alpha)s_{-1} \;=\; 0\,.$$

Since the subdifferential is linear we also have $s \in \partial_2 C(\alpha', t)$ which shows the assertion.
iv): Let $0 \le \alpha < \alpha' \le 1$ as well as $t \in F_L^*(\alpha)$ and $t' \in F_L^*(\alpha')$. Since for $t' = \infty$ or $t = -\infty$ the assertion is trivial by *i*) we also assume $t, t' \in \mathbb{R}$. By *iii*) we find an $s \in \partial_2 C(\alpha', t)$ with $s < 0$. Then we obtain $t' \ge t$ since otherwise we observe $s' \le s < 0$ for all $s' \in \partial_2 C(\alpha', t')$ which contradicts $t' \in F_L^*(\alpha')$.
v): This is a direct consequence of *iv*).
vi): Let us suppose that there exists a $t \in F_L^*(1/2)$ with $0 \in \partial_2 L(1,t)$ and $t < \max F_L^*(1/2)$. We fix a $t' \in F_L^*(3/4)$. Since F_L^* is monotone we have $t' \ge \max F_L^*(1/2) > t$ and hence the monotony of $\partial_2 L(1,.)$ yields $\partial_2 L(1,t') \subset [0,\infty)$. Since $0 \in \partial_2 C(3/4, t')$ the latter implies that there exists an $s \in \partial_2 L(-1,t')$ with $s \le 0$. Therefore, by Lemma 19 *i*) and the monotony of $\partial_2 L(-1,.)$ we find $t' < 0$ which contradicts the admissibility of L. The second assertion can be proved analogously.
vii): By the assumption there exist $t, t' \in F_L^*(\alpha)$ with $0 \in \partial_2 L(1,t)$ and $0 \in \partial_2 L(-1,t')$. The monotony of $\partial_2 L(1,.)$ implies $t > 0$ and hence $\alpha \ge 1/2$ by the admissibility of L. Analogously, $0 \in \partial_2 L(-1,t')$ yields $\alpha \le 1/2$. The last assertion is a direct consequence of Lemma 19 *ii*). ∎

The next Lemma shows how we can approximate the set

$$S \;:=\; \Big\{(x,y) \in X_{cont} \times Y : 0 \notin \partial_2 L\big(y, F_L^*(P(1|x)) \cap \mathbb{R}\big)\Big\}$$

which was defined in (8):

Lemma 21 *Let L be an admissible and convex loss function. Then for*

$$S_\varepsilon \;:=\; \Big\{(x,y) \in X_{cont} \times Y : 0 \notin \partial_2 L\big(y, F_L^*(P(1|x)) \cap \mathbb{R} + \varepsilon B_{\mathbb{R}}\big)\Big\}$$

we have $S_\varepsilon \subset S$ and $S_\varepsilon \subset S_{\varepsilon'}$ for all $\varepsilon > \varepsilon' > 0$. Moreover, we have

$$\bigcup_{\varepsilon > 0} S_\varepsilon \;=\; S\,.$$

Proof Since the first two assertions are obvious it suffices to prove $S \subset \bigcup_{\varepsilon>0} S_\varepsilon$. Obviously, this follows once we have established

$$\bigcap_{\varepsilon>0} \bigcup_{\delta\in[-\varepsilon,\varepsilon]} \bigcup_{t\in F_L^*(\alpha)\cap\mathbb{R}} \partial_2 L(y,t+\varepsilon) \subset \bigcup_{t\in F_L^*(\alpha)\cap\mathbb{R}} \partial_2 L(y,t) \tag{13}$$

for all $\alpha \in [0,1]$, $y = \pm 1$. If $F_L^*(\alpha)\cap\mathbb{R} = \emptyset$ inclusion (13) is trivial. Therefore, we assume $F_L^*(\alpha)\cap\mathbb{R} \neq \emptyset$. Let us fix an element h of the left set in (13). Then for all $n \in \mathbb{N}$ there exist $\delta_n \in [-1/n, 1/n]$ and $t_n \in F_L^*(\alpha)\cap\mathbb{R}$ with $h \in \partial_2 L(y, t_n+\delta_n)$. If (t_n) is unbounded we observe $\alpha \in \{0,1\}$. Furthermore, we find $t_n+\delta_n \in F_L^*(\alpha)\cap\mathbb{R}$ for a sufficiently large n since $F_L^*(\alpha)$ is an interval by the convexity of L. Hence we have shown (13) in this case.
If (t_n) is bounded there exists a subsequence (t_{n_k}) of (t_n) converging to an element $t_0 \in F_L^*(\alpha)\cap\mathbb{R}$ by the compactness of $F_L^*(\alpha)$ in $\overline{\mathbb{R}}$. Now let us fix an $\varepsilon > 0$. Since $\partial_2 L(y,.) : \mathbb{R} \to 2^{\mathbb{R}}$ is upper semi-continuous by Proposition 15 we find

$$h \in \partial_2 L(y, t_{n_k}+\delta_{n_k}) \subset \partial_2 L(y,t_0) + \varepsilon B_{\mathbb{R}}$$

for a sufficiently large k. This yields

$$h \in \bigcap_{\varepsilon>0} \big(\partial_2 L(y,t_0) + \varepsilon B_{\mathbb{R}}\big)$$

and thus we finally find $h \in \partial_2 L(y,t_0)$ by the compactness of $\partial_2 L(y,t_0)$. ∎

3.3 Asymptotic Behaviour of the Solutions I

In order to describe the asymptotic behaviour of $f_{T,\lambda}$ and $\tilde{f}_{T,\lambda}+\tilde{b}_{T,\lambda}$ we have to introduce a "distance function" for $t \in \mathbb{R}$ and $B \subset \overline{\mathbb{R}}$:

$$\rho(t,B) := \begin{cases} \inf_{s\in B}|t-s| & \text{if } B\cap\mathbb{R} \neq \emptyset \\ \min\{1, \frac{1}{t_+}\} & \text{if } B = \{\infty\} \\ \min\{1, \frac{1}{(-t)_+}\} & \text{if } B = \{-\infty\} \\ \frac{1}{|t|} & \text{otherwise,} \end{cases}$$

where $s_+ := \max\{0,s\}$ for all $s \in \mathbb{R}$ and $1/0 := \infty$. Note that ρ reduces to the usual definition of the distance between a point t and a set B if the latter contains a real number. For brevity's sake we also write

$$E(f,\varepsilon) := \Big\{x \in X : \rho\big(f(x), F_L^*(P(1|x))\big) \geq \varepsilon\Big\}$$

for $\varepsilon > 0$ and measurable functions $f : X \to \mathbb{R}$. Note that if $F_L^*(\alpha)\cap\mathbb{R} \neq \emptyset$ holds for all $\alpha \in [0,1]$ then $E(f,\varepsilon)$ is the set of points where f differs more than ε from all functions minimizing $\mathcal{R}_{L,P}$. Now, we can state the following key result:

Theorem 22 *Let P be a Borel probability measure on $X \times Y$ and L be a loss function with* $\operatorname{card} F_L^*(\alpha) > 1$ *for at most countably many $\alpha \in [0,1]$. Furthermore, assume that for such α the set $F_L^*(\alpha)$ is an interval in $\overline{\mathbb{R}}$. Then for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all measurable functions $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,P}(f) \leq \mathcal{R}_{L,P} + \delta$ we have $P_X(E(f,\varepsilon)) \leq \varepsilon$.*

Proof Let $f : X \to \mathbb{R}$ be a measurable function and $f_{L,P} := f^*(P(1|.))$, where f^* is a measurable selection from F_L^*. Then for $E := E(f,\varepsilon)$ we find

$$\begin{aligned}\mathcal{R}_{L,P}(f) &\geq \int_{X\setminus E}\int_Y L\big(y, f_{L,P}(x)\big)P(dy|x)P_X(dx) + \int_E\int_Y L\big(y, f(x)\big)P(dy|x)P_X(dx)\\ &= \mathcal{R}_{L,P} + \int_E\int_Y \Big(L\big((y,f(x)\big) - L\big(y, f_{L,P}(x)\big)\Big)P(dy|x)P_X(dx)\ .\end{aligned}$$

Let $G_\varepsilon(\alpha) := \{s \in \mathbb{R} : \rho(s, F_L^*(\alpha)) \geq \varepsilon\}$ if there exists an $s \in \mathbb{R}$ with $\rho(s, F_L^*(\alpha)) \geq \varepsilon$, and $G_\varepsilon(\alpha) := \mathbb{R}$ otherwise. Denoting the closure of $G_\varepsilon(\alpha)$ in $\overline{\mathbb{R}}$ by $\overline{G_\varepsilon(\alpha)}^{\overline{\mathbb{R}}}$ there exists $f_*(\alpha) \in \overline{G_\varepsilon(\alpha)}^{\overline{\mathbb{R}}}$ with

$$C(\alpha, f_*(\alpha)) = \inf_{t\in G_\varepsilon(\alpha)} C(\alpha,t)$$

for all $\alpha \in [0,1]$. Moreover, by the assumptions on L we can assume that the function $f_* : [0,1] \to \overline{\mathbb{R}}$ is measurable. Our next step is to show

$$f_*(\alpha) \notin F_L^*(\alpha) \tag{14}$$

for all $\alpha \in [0,1]$ for which there exists an $s \in \mathbb{R}$ with $\rho(s, F_L^*(\alpha)) \geq \varepsilon$. Let us assume the converse, i.e. there is an $\alpha \in [0,1]$ with $f_*(\alpha) \in F_L^*(\alpha)$ and $\rho(s, F_L^*(\alpha)) \geq \varepsilon$ for a suitable $s \in \mathbb{R}$. If $f_*(\alpha) \in \mathbb{R}$ we have $f_*(\alpha) \in G_\varepsilon(\alpha)$ and $F_L^*(\alpha) \cap \mathbb{R} \neq \emptyset$. Hence we find

$$\inf_{s\in F_L^*(\alpha)} |f_*(\alpha) - s| = \rho(f_*(\alpha), F_L^*(\alpha)) \geq \varepsilon$$

which contradicts our assumption $f_*(\alpha) \in F_L^*(\alpha)$. Hence we have $f_*(\alpha) \in \{-\infty, \infty\}$. Without loss of generality we assume $f_*(\alpha) = \infty$. Since $f_*(\alpha) \in \overline{G_\varepsilon(\alpha)}^{\overline{\mathbb{R}}}$ there is a sequence $(t_n) \subset G_\varepsilon(\alpha)$ with $t_n \to \infty$. If $F_L^*(\alpha) \cap \mathbb{R} = \emptyset$ this shows $\rho(t_n, F_L^*(\alpha)) = 1/t_n \to 0$ for $n \to \infty$ which contradicts $(t_n) \subset G_\varepsilon(\alpha) = \{s \in \mathbb{R} : \rho(s, F_L^*(\alpha)) \geq \varepsilon\}$. Hence we have $F_L^*(\alpha) \cap \mathbb{R} \neq \emptyset$. Since $F_L^*(\alpha)$ is an interval we find $F_L^*(\alpha) = [a, \infty]$ for some $a \in [-\infty, \infty]$. For large n this implies $t_n \in F_L^*(\alpha)$, i.e. $\rho(t_n, F_L^*(\alpha)) = 0$, which also contradicts $(t_n) \subset G_\varepsilon(\alpha)$. Therefore we have established (14).

Now, the definition of f_* and our first estimate yields

$$\mathcal{R}_{L,P}(f) \geq \mathcal{R}_{L,P} + \int_E \Delta\, dP_X\ ,$$

where

$$\Delta(x) := \int_Y L\big(y, f_*(P(1|x))\big) - L(y, f_{L,P}(x))\, P(dy|x)\ .$$

Furthermore, since (14) guarantees $\Delta(x) > 0$ for all

$$x \in \tilde{X}_\varepsilon := \Big\{x \in X : \exists s \in \mathbb{R} \text{ with } \rho\big(s, F_L^*(P(1|x))\big) \geq \varepsilon\Big\}$$

the restrictions of the measures P_X and ΔdP_X onto $\tilde{X}_\varepsilon$ are absolutely continuous to each other. Now, the assertion easily follows from $E \subset \tilde{X}_\varepsilon$. ■

Remark 23 *The assumption* $\operatorname{card} F_L^*(\alpha) > 1$ *for at most countably many* $\alpha \in [0,1]$ *in the above theorem was only used to ensure the measurability of* f_*. *We suppose that this assumption is superfluous. As we have seen in Lemma 20 it is always satisfied for admissible convex loss functions. Using a slightly different definition of* ρ *the assumption "*$F_L^*(\alpha)$ *is an interval" can also be omitted. Since for convex loss functions* $F_L^*(\alpha)$ *is always an interval in* $\overline{\mathbb{R}}$ *we do not go into details.*

Remark 24 *As already pointed out in Section 2 it was shown by Steinwart (2003a) that there exist kernels and sequences of regularization parameters such that for the corresponding classifiers based on (1) and (2) we have* $\mathcal{R}_{L,P}(f_{T,\lambda_n}) \to \mathcal{R}_{L,P}$ *and* $\mathcal{R}_{L,P}(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}) \to \mathcal{R}_{L,P}$, *respectively. In this case, Theorem 22 e.g. yields*

$$P_X\big(E(f_{T,\lambda_n},\varepsilon)\big) \to 0$$

for all $\varepsilon > 0$. *In particular, if* $F_L^*(\alpha) \subset \mathbb{R}$ *and* $\operatorname{card} F_L^*(\alpha) = 1$ *hold for all* $\alpha \in [0,1]$ *then*

$$\|f_{T,\lambda_n} - f_{L,P}\|_0 \to 0 \tag{15}$$

holds in probability for $|T| = n \to \infty$. *Here*

$$\|f\|_0 := \int_X \min\{1, |f|\} dP_X$$

is a translation invariant metric which describes the convergence in probability with respect to P_X *in the space of all measurable functions* $L_0(P_X)$. *The aim of the following sections is to show that for convex and (strongly) admissible loss functions Theorem 22 and in particular (15) can be improved. Namely, we show that the set* $E(f_{T,\lambda},\varepsilon)$ *describing the* ε*-discrepancy of* f_{T,λ_n} *from* $f_{L,P}$ *is "essentially" independent of* T. *This will allow us to control the behaviour of* f_{T,λ_n} *on the samples of* T.

Remark 25 *Theorem 22 does not only apply to classifiers of SVM type. Indeed, it describes the limiting decision function and the corresponding convergence for every classifier minimizing a (modified) L-risk provided that the L-risks* $\mathcal{R}_{L,P}(f_T)$ *of its decision functions* f_T *converge to* $\mathcal{R}_{L,P}$. *Recall that the latter condition also ensures universal consistency for admissible loss functions.*

3.4 Stability

In this section we show that the decision functions of the classifiers based on (1) or (2) are concentrated around the minimizer of $\mathcal{R}^{reg}_{L,P,\lambda}$ if the loss function is convex. In order to unify the following considerations we define

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

for a RKHS H, a projection $A : H \to H$, a loss function L, $f \in H$ and $\lambda > 0$. Our first aim is to derive a formula for the subdifferential of $\mathcal{R}^{reg}_{L,P,\lambda,A}(.)$. Besides the calculus presented in the preliminaries we also need an integration rule in order to treat the integral $\mathcal{R}_{L,P}(.)$. Due to technical reasons it is convenient to split the latter: for a Borel probability measure P on $X \times Y$ and a measurable $B \subset X$ we define

$$\begin{aligned} P_X^+(B) &:= \int_X 1_B(x) P(1|x)\, P_X(dx) \\ P_X^-(B) &:= \int_X 1_B(x) P(-1|x)\, P_X(dx)\ , \end{aligned}$$

where 1_B denotes the indicator function of B. With the help of these measures we set

$$\begin{aligned}\mathcal{R}^+_{L,P}(f) &:= \int_X L(1, f(x))\, P^+_X(dx)\\ \mathcal{R}^-_{L,P}(f) &:= \int_X L(-1, f(x))\, P^-_X(dx)\end{aligned}$$

for admissible loss functions L and measurable functions $f : X \to \overline{\mathbb{R}}$. Obviously, we always have $\mathcal{R}_{L,P}(f) = \mathcal{R}^+_{L,P}(f) + \mathcal{R}^-_{L,P}(f)$. In the following proposition we collect some useful properties of $\mathcal{R}^\pm_{L,P}(.)$:

Proposition 26 *Let L be a convex and Lipschitz continuous loss function and P a Borel probability measure on $X \times Y$. Then the functionals $\mathcal{R}^\pm_{L,P} : L_2(P^\pm_X) \to [0,\infty]$ are convex, finite at 0 and continuous at 0. Furthermore, for all $h \in L_2(P)$ we have*

$$\partial \mathcal{R}^\pm_{L,P}(h) \;=\; \left\{ h^* \in L_2(P^\pm_X) \;:\; h^*(x) \in \partial L(\pm 1, h(x))\; P^\pm_X\text{-a.s} \right\}. \tag{16}$$

Proof We only have to consider $\mathcal{R}^+_{L,P}$. Using the notions of Rockafellar (1976) we first observe that $L(1,.)$ is a normal and convex integrand (see Rockafellar, 1976, p. 173). In particular, $\mathcal{R}^+_{L,P}$ is convex. Since $\mathcal{R}^+_{L,P}(0) = L(1,0)P^+_X(X) \in \mathbb{R}$ the equation (16) then follows (see Rockafellar, 1976, Cor. 3E.).

In order to prove the continuity at 0 let $(f_n) \subset L_2(P^+_X)$ be a sequence with $f_n \to 0$. Then for $\varepsilon > 0$ and $A^\varepsilon_n := \{x \in X : |f_n(x)| > \varepsilon\}$ one easily checks that there exists an integer n_0 such that for all $n \geq n_0$ we have both $P^+_X(A^\varepsilon_n) \leq \varepsilon$ and

$$\int_{A^\varepsilon_n} |f_n|\, dP^+_X \;\leq\; \varepsilon\,.$$

Moreover, the Lipschitz-continuity of L yields $L(1,t) \leq |L|_1|t| + L(1,0)$ for all $t \in \mathbb{R}$. Therefore we obtain

$$\begin{aligned}\mathcal{R}^+_{L,P}(f_n) &= \int_{A^\varepsilon_n} L(1, f_n)\, dP^+_X + \int_{X\setminus A^\varepsilon_n} L(1, f_n)\, dP^+_X\\ &\leq \int_{A^\varepsilon_n} |L|_1 |f_n| + L(1,0)\, dP^+_X + \int_{X\setminus A^\varepsilon_n} |L|_1|\varepsilon| + L(1,0)\, dP^+_X\\ &\leq 2\varepsilon|L|_1 + \mathcal{R}^+_{L,P}(0)\end{aligned}$$

and hence we find $\limsup_{n\to\infty} \mathcal{R}^+_{L,P}(f_n) \leq \mathcal{R}^+_{L,P}(0)$. In order to show $\mathcal{R}^+_{L,P}(0) \leq \liminf_{n\to\infty} \mathcal{R}^+_{L,P}(f_n)$ we observe that for $h = 0$ and $\varepsilon = 1$ we have $L(1, h(.) + a) \in L_2(P^+_X)$ for all $|a| \leq \varepsilon$. Hence, $\mathcal{R}^+_{L,P}$ is lower semi-continuous at 0 with respect to the weak topology of $L_2(P^+_X)$ (see Rockafellar, 1976, Cor. 3D. and Prop. 3G.). In particular, $\mathcal{R}^+_{L,P}$ is lower semi-continuous at 0 with respect to the norm, i.e. $\mathcal{R}^+_{L,P}(0) \leq \liminf_{n\to\infty} \mathcal{R}^+_{L,P}(f_n)$. ∎

Proposition 27 *Let k be a continuous kernel with RKHS H and feature map $\Phi : X \to H$. Moreover, let $A : H \to H$ be a projection, L be a convex and Lipschitz-continuous loss function and P be a Borel probability measure on $X \times Y$. Then for all $f \in H$ we have*

$$\partial \mathcal{R}^{reg}_{L,P,\lambda,A}(f) \;=\; 2\lambda A f + \left\{ \mathbb{E}_P h\Phi \;:\; h \in L_0(P),\; h(x,y) \in \partial_2 L(y, f(x))\; P\text{-a.s.} \right\}.$$

Note, that in the above proposition $\mathbb{E}_P h\Phi$ is a Hilbert space valued expectation. Such expectations are defined by Bochner integrals. For more information on Bochner integrals we refer to the book of Diestel and Uhl (1977).

Proof Let $I^{\pm} : H \to L_2(P_X^{\pm})$ be the natural inclusions, i.e. $I^{\pm} f := \langle f, \Phi(.) \rangle$ for all $f \in H$. Then we observe that $\mathcal{R}^{reg}_{L,P,\lambda,A}(f) = \lambda \langle Af, Af \rangle + \mathcal{R}^{+}_{L,P}(I^{+} f) + \mathcal{R}^{-}_{L,P}(I^{-} f)$ holds. The continuity of L and k ensures $\mathcal{R}^{\pm}_{L,P}(I^{\pm} f) \in \mathbb{R}$ for all $f \in H$. Furthermore, using Lebesgue's dominated convergence theorem we easily see that $\mathcal{R}^{\pm}_{L,P} \circ I^{\pm} : H \to \mathbb{R}$ are even continuous. Therefore, the linearity of the subdifferential and $\partial \|.\|^2_H(f) = 2f$ imply

$$\partial \mathcal{R}^{reg}_{L,P,\lambda,A}(f) = 2\lambda Af + \partial(\mathcal{R}^{+}_{L,P} \circ I^{+})(f) + \partial(\mathcal{R}^{-}_{L,P} \circ I^{-})(f) \,.$$

Now, $\mathcal{R}^{+}_{L,P} : L_2(P_X^{+}) \to [0,\infty]$ is continuous at 0. Hence, the chain rule of Proposition 18 together with Proposition 26 yields

$$\begin{aligned} \partial(\mathcal{R}^{+}_{L,P} \circ I^{+})(f) &= (I^{+})^{*} \partial \mathcal{R}^{+}_{L,P}(I^{+} f) \\ &= (I^{+})^{*} \big(\{ h^{+} \in L_2(P_X^{+}) \,:\, h^{+}(x) \in \partial L(1, f(x))\ P_X^{+}\text{-a.s} \} \big). \end{aligned}$$

Since the adjoint operator of I^{+} maps every $h \in L_2(P_X^{+})$ to $(I^{+})^{*} h = \mathbb{E}_{P_X^{+}} h\Phi$ we obtain

$$\partial(\mathcal{R}^{+}_{L,P} \circ I^{+})(f) \;=\; \big\{ \mathbb{E}_{P_X^{+}} h^{+}\Phi \,:\, h^{+} \in L_2(P_X^{+}),\ h^{+}(x) \in \partial L(1, f(x))\ P_X^{+}\text{-a.s.} \big\} \,.$$

Analogously, we get

$$\partial(\mathcal{R}^{-}_{L,P} \circ I^{-})(f) \;=\; \big\{ \mathbb{E}_{P_X^{-}} h^{-}\Phi \,:\, h^{-} \in L_2(P_X^{-}),\ h^{-}(x) \in \partial L(-1, f(x))\ P_X^{-}\text{-a.s.} \big\} \,.$$

Using the notation $h(x,1) := h^{+}(x)$ and $h(x,-1) := h^{-}(x)$ we thus find

$$\partial(\mathcal{R}^{+}_{L,P} \circ I^{+})(f) + \partial(\mathcal{R}^{-}_{L,P} \circ I^{-})(f) = \big\{ \mathbb{E}_P h\Phi \,:\, h \in L_2(P),\ h(x,y) \in \partial_2 L(y, f(x))\ P\text{-a.s.} \big\} \,.$$

Finally, $L_2(P)$ can be replaced by $L_0(P)$ since L is Lipschitz continuous. ■

The result of Proposition 27 has already been presented by Zhang (2001). However, the claim therein that Proposition 27 can be proved using subdifferential calculus on *finite* dimensional spaces is obviously not correct. For differentiable loss functions Proposition 27 is more or less trivial.

Now we are able to prove the main result of this subsection:

Theorem 28 *Let L be a convex loss function, H be a RKHS of a continuous kernel with feature map $\Phi : X \to H$, $A : H \to H$ be an orthogonal projection and P be a Borel probability measure on $X \times Y$. Assume that $\mathcal{R}^{reg}_{L,P,\lambda,A}$ can be minimized and that there exists a constant $c > 0$ such that $\|\hat{f}_{P,\lambda}\|_\infty \le c$ for all $\hat{f}_{P,\lambda} \in H$ minimizing $\mathcal{R}^{reg}_{L,P,\lambda,A}$. Then there exists a measurable function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \le \big|L_{|Y \times [-c,c]}\big|_1$ such that for all Borel probability measures Q and every element $\hat{f}_{Q,\lambda} \in H$ which minimizes $\mathcal{R}^{reg}_{L,Q,\lambda,A}$ and satisfies $\|\hat{f}_{Q,\lambda}\|_\infty \le c$ we have*

$$\|A\hat{f}_{P,\lambda} - A\hat{f}_{Q,\lambda}\|^2 \le \frac{\|\hat{f}_{P,\lambda} - \hat{f}_{Q,\lambda}\| \, \|\mathbb{E}_P h\Phi - \mathbb{E}_Q h\Phi\|}{\lambda} \,.$$

For the proof of Theorem 28 we need the following simple lemmas which will not be proved:

Lemma 29 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex and continuous function. Then f restricted to $[-a,a]$, $a > 0$, is Lipschitz continuous and we have*

$$\big|f_{|[-a,a]}\big|_1 \;\le\; \frac{2}{a}\big\|f_{|[-2a,2a]}\big\|_\infty \,.$$

Lemma 30 *Let $f : [a,b] \to \mathbb{R}^+$ be a convex and Lipschitz continuous function. Then there exists a convex extension $\tilde{f} : \mathbb{R} \to \mathbb{R}^+$ of f that is Lipschitz continuous with $|f|_1 = |\tilde{f}|_1$.*

Proof of Theorem 28 Let us first assume that L is Lipschitz continuous. Since $\hat{f}_{P,\lambda}$ minimizes $\mathcal{R}^{reg}_{L,P,\lambda,A}$ we observe $0 \in \partial\mathcal{R}^{reg}_{L,P,\lambda,A}(\hat{f}_{P,\lambda})$. Thus, by Proposition 27 there exists a function $h \in L_0(P)$ with $h(x,y) \in \partial_2 L(y, \hat{f}_{P,\lambda}(x))$ for P-almost all $(x,y) \in X \times Y$ and

$$0 \;=\; 2A\lambda\hat{f}_{P,\lambda} + \mathbb{E}_P h\Phi \,. \tag{17}$$

By the Lipschitz-continuity and Proposition 14 we actually have $\|h\|_\infty \le |L|_1$. Moreover, we can assume without loss of generality that $h(x,y) \in \partial_2 L(y, \hat{f}_{P,\lambda}(x))$ for *all* $(x,y) \in X \times Y$. Then we obtain

$$h(x,y)\big(\hat{f}_{Q,\lambda}(x) - \hat{f}_{P,\lambda}(x)\big) \;\le\; L(y, \hat{f}_{Q,\lambda}(x)) - L(y, \hat{f}_{P,\lambda}(x))$$

for all $(x,y) \in X \times Y$. Integration with respect to Q then yields

$$\mathbb{E}_{(x,y)\sim Q} L(y, \hat{f}_{P,\lambda}(x)) + \langle \hat{f}_{Q,\lambda} - \hat{f}_{P,\lambda}, \mathbb{E}_Q h\Phi\rangle \;\le\; \mathbb{E}_{(x,y)\sim Q} L(y, \hat{f}_{Q,\lambda}(x)).$$

Since $\lambda\|A\hat{f}_{P,\lambda}\|^2 + 2\lambda\langle A\hat{f}_{Q,\lambda} - A\hat{f}_{P,\lambda}, \hat{f}_{P,\lambda}\rangle + \lambda\|A\hat{f}_{P,\lambda} - A\hat{f}_{Q,\lambda}\|^2 = \lambda\|A\hat{f}_{Q,\lambda}\|^2$ the latter inequality implies

$$\mathcal{R}^{reg}_{L,Q,\lambda,A}(\hat{f}_{P,\lambda}) + \langle \hat{f}_{Q,\lambda} - \hat{f}_{P,\lambda}, \mathbb{E}_Q h\Phi + 2\lambda A^*\hat{f}_{P,\lambda}\rangle + \lambda\|A\hat{f}_{P,\lambda} - A\hat{f}_{Q,\lambda}\|^2 \;\le\; \mathcal{R}^{reg}_{L,Q,\lambda,A}(\hat{f}_{Q,\lambda}) \,.$$

Moreover, $\hat{f}_{Q,\lambda}$ minimizes $\mathcal{R}^{reg}_{L,Q,\lambda,A}$ and hence we have $\mathcal{R}^{reg}_{L,Q,\lambda,A}(\hat{f}_{Q,\lambda}) \le \mathcal{R}^{reg}_{L,Q,\lambda,A}(\hat{f}_{P,\lambda})$. This and $A^* = A$ yield

$$\begin{aligned}\lambda\|A\hat{f}_{P,\lambda} - A\hat{f}_{Q,\lambda}\|^2 &\le \langle \hat{f}_{P,\lambda} - \hat{f}_{Q,\lambda}, \mathbb{E}_Q h\Phi + 2\lambda A\hat{f}_{P,\lambda}\rangle \\ &\le \|\hat{f}_{P,\lambda} - \hat{f}_{Q,\lambda}\|\,\|\mathbb{E}_Q h\Phi + 2\lambda A\hat{f}_{P,\lambda}\| \,.\end{aligned}$$

With the help of (17) we can replace $2\lambda A\hat{f}_{P,\lambda}$ by $-\mathbb{E}_P h\Phi$ and thus the assertion follows.

In the general case we know by Lemma 29 that L restricted to $Y \times [-c,c]$ is Lipschitz continuous and thus there exists a Lipschitz continuous extension $\tilde{L}$ according to Lemma 30. Since $\mathcal{R}^{reg}_{\tilde{L},P,\lambda,A}$ and $\mathcal{R}^{reg}_{\tilde{L},Q,\lambda,A}$ coincide with $\mathcal{R}^{reg}_{L,P,\lambda,A}$ and $\mathcal{R}^{reg}_{L,Q,\lambda,A}$ on cB_H, respectively, we then obtain the assertion. ■

Remark 31 *Taking $P = Q$ in the previous theorem we immediately obtain that $A\hat{f}_{P,\lambda}$ is* unique. *In particular, the problem (1) has* always *a unique solution for convex loss functions. Furthermore, it is obvious that this also holds for L1- and L2-SVMs with offset since in these cases we have $\|\hat{f}_{P,\lambda}\|_\infty \le 2 + 2K\delta_\lambda$.*

Remark 32 *Equation (17) is a general form of the well-known representer theorem. Indeed, (17) reduces to*

$$A\hat{f}_{T,\lambda} \;=\; \sum_{i=1}^{n} \alpha_i k(x_i, .)$$

for training sets $T = \big((x_1,y_1),\ldots,(x_n,y_n)\big) \in (X \times Y)^n$ *and suitable coefficients* $\alpha_1,\ldots,\alpha_n \in \mathbb{R}$. *Furthermore, the above proof showed (7), i.e.*

$$\alpha_i \in -\frac{1}{2n\lambda}\partial_2 L(y_i, \hat{f}_{T,\lambda}(x_i))$$

for all $i = 1,\ldots,n$. *Therefore, a sample* x_i *must be a support vector of the above representation if* $0 \notin \partial_2 L(y_i, \hat{f}_{T,\lambda}(x_i))$. *In order to prove lower bounds on the number of support vectors it hence suffices to know the behaviour of* $\hat{f}_{T,\lambda}$ *on* T. *This will be our key idea in the following considerations.*

3.5 Asymptotic Behaviour of the Solutions II

In this part we refine the results of Subsection 3.3 concerning the asymptotic behaviour of the solutions of (1) and (2). We begin with:

Proposition 33 *Let* L *be a convex loss function,* H *be a RKHS of a continuous kernel and* P *be a Borel probability measure on* $X \times Y$. *Then for all* $\varepsilon > 0$, $\lambda > 0$ *and all* $n \geq 1$ *we have*

$$P^n\Big(T \in (X \times Y)^n : \big\| f_{T,\lambda} - f_{P,\lambda}\big\| \geq \varepsilon\Big) \;\leq\; 2\exp\Big(-\frac{\varepsilon^2\lambda^2 n}{8K^2|L_\lambda|_1^2 + 2\varepsilon\lambda K|L_\lambda|_1}\Big)\,.$$

For the proof we will need the following result which is a reformulation of Theorem 3.3.4 of Yurinsky (1995):

Lemma 34 *Let* $\eta_1,\ldots,\eta_n$ *be bounded i.i.d. random variables with values in a Hilbert space* H. *Assume* $\|\eta_i\|_\infty \leq M$ *for all* $i = 1,\ldots,n$. *Then for all* $\varepsilon > 0$ *and all* $n \geq 1$ *we have*

$$P\left(\Big\|\frac{1}{n}\sum_{i=1}^{n}(\eta_i - \mathbb{E}\eta_i)\Big\| \geq \varepsilon\right) \;\leq\; 2\exp\Big(-\frac{\varepsilon^2 n}{8M^2 + 4\varepsilon M}\Big)\,.$$

Proof Apply Theorem 3.3.4 of Yurinsky (1995) to $\xi_i := \eta_i - \mathbb{E}\eta_i$, $H := 2M$, $B := 2M\sqrt{n}$ and $x := \frac{\varepsilon\sqrt{n}}{2M}$. ∎

Proof of Proposition 33 By Theorem 28 we know $\lambda\|f_{P,\lambda} - f_{T,\lambda}\| \leq \|\mathbb{E}_P h\Phi - \mathbb{E}_T h\Phi\|$ for a suitable function $h : X \times Y \to \mathbb{R}$ independent on T. Moreover, our specific situation guarantees $\|h\|_\infty \leq |L_\lambda|_1$. Applying Lemma 34 to $\eta_i := h(x_i,y_i)\Phi(x_i)$, $i = 1,\ldots,n$, and $M = K|L_\lambda|_1$ we thus obtain

$$\begin{aligned} P^n\Big(T \in (X \times Y)^n : \big\| f_{T,\lambda} - f_{P,\lambda}\big\| \geq \varepsilon\Big) \;&\leq\; P^n\Big(T \in (X \times Y)^n : \|\mathbb{E}_T h\Phi - \mathbb{E}_P h\Phi\| \geq \varepsilon\lambda\Big) \\ &\leq\; 2\exp\Big(-\frac{\varepsilon^2\lambda^2 n}{8K^2|L_\lambda|_1^2 + 4\varepsilon\lambda K|L_\lambda|_1}\Big)\,, \end{aligned}$$

which is assertion. ∎

With the help of Proposition 33 we are now able to show that $E(f_{T,\lambda},\varepsilon)$ is essentially independent of T, i.e. it is contained in a small set which only depends on the training set size n and the accuracy ε. The precise result is stated in the following proposition:

Proposition 35 *Let L be an admissible and convex loss function, H be a RKHS of a universal kernel and P be a Borel probability measure on $X \times Y$. Let us further assume that (λ_n) is a sequence of strictly positive real numbers with $\lambda_n \to 0$ and $n\lambda_n^2/|L_{\lambda_n}|_1^2 \to \infty$. Then for all $\varepsilon \in (0,1)$ there exists a sequence of sets $E_n(\varepsilon) \subset X$ with $P_X(E_n(\varepsilon)) \to 0$ and*

$$P^n\Big(T \in (X \times Y)^n : E(f_{T,\lambda_n},\varepsilon) \subset E_n(\varepsilon)\Big) \to 1\,.$$

Proof Let T be a training set with $\|f_{T,\lambda_n} - f_{P,\lambda_n}\| \leq \frac{\varepsilon}{2K}$. Since $n\lambda_n^2/|L_{\lambda_n}|_1 \to \infty$ Proposition 33 ensures that the probability of such training sets tends to 1. We first show

$$E(f_{T,\lambda_n},\varepsilon) \subset E(f_{P,\lambda_n},\varepsilon/2) \tag{18}$$

Let us assume the converse, i.e. there is an $x \in E(f_{T,\lambda_n},\varepsilon)$ with $x \notin E(f_{P,\lambda_n},\varepsilon/2)$. If $F_L^*(P(1|x)) \cap \mathbb{R} \neq \emptyset$ the latter implies

$$\varepsilon/2 \;>\; \rho\big(f_{P,\lambda_n}(x), F_L^*(P(1|x))\big) \;=\; \inf_{s \in F_L^*(P(1|x))} |f_{P,\lambda_n}(x) - s|\,.$$

In particular there exists an $s^* \in F_L^*(P(1|x))$ with $|f_{P,\lambda_n}(x) - s^*| < \varepsilon/2$. Hence we find

$$\rho\big(f_{T,\lambda_n}(x), F_L^*(P(1|x))\big) \;\leq\; |f_{T,\lambda_n}(x) - s^*| \;\leq\; |f_{T,\lambda_n}(x) - f_{P,\lambda_n}(x)| + |f_{P,\lambda_n}(x) - s^*| < \varepsilon\,.$$

i.e. $x \notin E(f_{T,\lambda_n},\varepsilon)$ which contradicts our assumption. Therefore we have to consider the case $F_L^*(P(1|x)) \cap \mathbb{R} = \emptyset$. Without loss of generality we may assume $F_L^*(P(1|x)) = \{\infty\}$. Then $x \notin E(f_{P,\lambda_n},\varepsilon/2)$ implies

$$\frac{\varepsilon}{2} \;>\; \rho\big(f_{P,\lambda_n}(x), F_L^*(P(1|x))\big) \;=\; \min\Big\{1, \frac{1}{(f_{P,\lambda_n}(x))_+}\Big\}\,,$$

i.e. $f_{P,\lambda_n}(x) > 2/\varepsilon$. Hence we find

$$f_{T,\lambda_n}(x) \;=\; f_{T,\lambda_n}(x) - f_{P,\lambda_n}(x) + f_{P,\lambda_n}(x) \;>\; -\frac{\varepsilon}{2} + \frac{2}{\varepsilon} \;\geq\; \frac{1}{\varepsilon}\,.$$

This yields $\rho\big(f_{T,\lambda_n}(x), F_L^*(P(1|x))\big) = \min\big\{1, \frac{1}{(f_{T,\lambda_n}(x))_+}\big\} < \varepsilon$ which again contradicts our assumption $x \in E(f_{T,\lambda_n},\varepsilon)$. Therefore, we have shown (18).

Now, $\lambda_n \to 0$ yields $\mathcal{R}_{L,P}(f_{P,\lambda_n}) \to \mathcal{R}_{L,P}$ (see Steinwart, 2003a) and therefore, Theorem 22 shows that $E(f_{P,\lambda_n},\varepsilon/2)$ are the desired sets for large n. ∎

Remark 36 *Proposition 35 also holds for convex loss functions that satisfy the assumptions of Theorem 22.*

In the rest of this section we show that Proposition 35 also holds for classifiers based on (2). Unfortunately, it turns out that their treatment is a bit more technical. We begin with a result which is analogous to Proposition 33:

Proposition 37 *Let L be a regular and convex loss function, H be a RKHS of a continuous kernel, and P be a non-degenerate Borel probability measure on $X \times Y$. Then for all $\varepsilon > 0$ there exists a constant $c > 0$ such that for all $\lambda \in (0,1)$ and all $n \geq 1$ we have*

$$P^n\Big(T \in (X \times Y)^n : \|\tilde{f}_{T,\lambda} - \tilde{f}_{P,\lambda}\| \geq \varepsilon\Big) \;\leq\; 4\exp\Big(-c\,\frac{\varepsilon^4\lambda^3 n}{|L_\lambda|_1^2}\Big)\,.$$

Proof It was shown by Steinwart (2003a) that there exists a constant $\tilde{c} > 0$ with $|\tilde{b}_{P,\lambda}| \leq \tilde{c} + \delta_\lambda K$ for all $\lambda > 0$ such that

$$\Pr^*\big(T \in (X \times Y)^n : |\tilde{b}_{T,\lambda}| \leq \tilde{c} + \delta_\lambda K \text{ for all } \lambda > 0\big) \;\geq\; 1 - 2e^{-\tilde{c}n} \tag{19}$$

holds for all $n \geq 1$. We define $\tilde{L}_\lambda := L_{|Y\times[-a,a]}$, where $a := \tilde{c} + (1+K)\delta_\lambda$. Then we can apply Theorem 28 to the training sets considered in (19). This gives us a function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \leq |\tilde{L}_\lambda|_1$ and

$$\begin{aligned} &\Pr^*\left(T : \|\tilde{f}_{P,\lambda} - \tilde{f}_{T,\lambda}\|_H^2 \leq \frac{\|(\tilde{f}_{P,\lambda},\tilde{b}_{P,\lambda}) - (\tilde{f}_{T,\lambda},\tilde{b}_{T,\lambda})\|_{H\oplus_2\mathbb{R}}\,\|\mathbb{E}_P h\Phi - \mathbb{E}_T h\Phi\|_H}{\lambda}\right) \\ \geq\;& 1 - 2e^{-\tilde{c}n}, \end{aligned}$$

where $\|(f,b)\|_{H\oplus_2\mathbb{R}} := \sqrt{\|f\|_H^2 + |b|^2}$, $f \in H$, $b \in \mathbb{R}$, denotes the Hilbert space norm of the direct sum of H and $\mathbb{R}$. Moreover, for the training sets considered in (19) we always have

$$\|(\tilde{f}_{P,\lambda},\tilde{b}_{P,\lambda}) - (\tilde{f}_{T,\lambda},\tilde{b}_{T,\lambda})\|_{H\oplus_2\mathbb{R}} \;\leq\; \|\tilde{f}_{P,\lambda} - \tilde{f}_{T,\lambda}\|_H + |\tilde{b}_{P,\lambda} - \tilde{b}_{T,\lambda}| \;\leq\; 2\tilde{c} + 2(1+K)\delta_\lambda\,.$$

With $\tilde{\varepsilon} := \frac{\varepsilon^2\lambda}{2\tilde{c}+2(1+K)\delta_\lambda}$ we thus find

$$\begin{aligned} &P^n\Big(T \in (X \times Y)^n : \|\tilde{f}_{T,\lambda} - \tilde{f}_{P,\lambda}\| \geq \varepsilon\Big) \\ \leq\;& P^n\Big(T \in (X \times Y)^n : \|\mathbb{E}_T h\Phi - \mathbb{E}_P h\Phi\| \geq \tilde{\varepsilon}\Big) + 2e^{-\tilde{c}n} \\ \leq\;& 2\exp\Big(-\frac{\tilde{\varepsilon}^2 n}{8K^2|L_\lambda|_1^2 + 4\tilde{\varepsilon}K|L_\lambda|_1}\Big) + 2e^{-\tilde{c}n}. \end{aligned}$$

Using $\tilde{\varepsilon} \sim \varepsilon^2\lambda^{3/2}$ and $8K^2|L_\lambda|_1^2 + 4\tilde{\varepsilon}K|L_\lambda|_1 \preceq |L_\lambda|_1^2$ for fixed ε and $\lambda \to 0$ we then obtain the assertion. ∎

The following proposition essentially states the result of Proposition 35 for classifiers based on (2). Due to technical reasons we must restrict the class of probability measures for which the result holds. This lack will cause further technical difficulties in the proof of Theorem 10.

Proposition 38 *Let L be a strongly admissible, regular and convex loss function, H be a RKHS of a universal kernel and P be a non-degenerate Borel probability measure on $X \times Y$ with*

$$P_X\Big(x \in X : P(1|x) \notin \{0, 1/2, 1\}\Big) > 0 \,.$$

Let us further assume that (λ_n) is a sequence of strictly positive real numbers with $\lambda_n \to 0$, $n\lambda_n^3/|L_{\lambda_n}|_1^2 \to \infty$ and $n\lambda_n/\big(\|L_{\lambda_n}\|_\infty^2 |L_{\lambda_n}|_1^2 \log n\big) \to \infty$. Then for all sufficiently small $\varepsilon > 0$ we have

$$\Pr^*\big(T \in (X \times Y)^n : \|\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n} - \tilde{b}_{P,\lambda_n}\|_\infty \leq \varepsilon\big) \to 1 \,. \tag{20}$$

Moreover, for all sufficiently small $\varepsilon > 0$ there exists a sequence of sets $E_n(\varepsilon) \subset X$, $n \geq 1$, with $P_X(E_n(\varepsilon)) \to 0$ and

$$P^n\Big(T \in (X \times Y)^n : E(f_{T,\lambda_n}, \varepsilon) \subset E_n(\varepsilon)\Big) \to 1 \,. \tag{21}$$

Proof We define $\tilde{X} := \big\{x \in X : P(1|x) \notin \{0, 1/2, 1\}\big\}$ and fix an ε with $0 < \varepsilon < P_X(\tilde{X})$. Furthermore, for $\varepsilon/4$ we chose a $\delta > 0$ according to Theorem 22. Let us suppose that we have a training set T with $\mathcal{R}_{L,P}(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}) \leq \mathcal{R}_{L,P} + \delta$ and $\|\tilde{f}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n}\|_\infty \leq \varepsilon/4$. Recall that the probability of such training set converges to 1 (see Proposition 37 and Steinwart, 2003a). Now, the assumptions on T yield

$$P_X\big(x \in \tilde{X} : |\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} - f_{L,P}(x)| < \varepsilon/4\big) \;\geq\; \frac{2}{3} P_X(\tilde{X}) \,.$$

Note that unlike $\tilde{b}_{T,\lambda_n}$ the value $f_{L,P}(x) \in F_L^*(P(1|x))$ is uniquely determined for all $x \in \tilde{X}$ by the assumptions on L. Moreover, for sufficiently small λ_n we also have

$$P_X\big(x \in \tilde{X} : |\tilde{f}_{P,\lambda_n}(x) + \tilde{b}_{P,\lambda_n} - f_{L,P}(x)| < \varepsilon/4\big) \;\geq\; \frac{2}{3} P_X(\tilde{X}) \,.$$

Hence there exists an element $x_0 \in X$ with $|\tilde{f}_{T,\lambda_n}(x_0) + \tilde{b}_{T,\lambda_n} - f_{L,P}(x_0)| < \varepsilon/4$ and $|\tilde{f}_{P,\lambda_n}(x_0) + \tilde{b}_{P,\lambda_n} - f_{L,P}(x_0)| < \varepsilon/4$. Since this yields

$$\begin{aligned} |\tilde{b}_{T,\lambda_n} - \tilde{b}_{P,\lambda_n}| &\leq |\tilde{f}_{T,\lambda_n}(x_0) + \tilde{b}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n}(x_0) - \tilde{b}_{P,\lambda_n}| + \|\tilde{f}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n}\|_\infty \\ &\leq |\tilde{f}_{T,\lambda_n}(x_0) + \tilde{b}_{T,\lambda_n} - f_{L,P}(x_0)| + |\tilde{f}_{P,\lambda_n}(x_0) + \tilde{b}_{P,\lambda_n} - f_{L,P}(x_0)| + \varepsilon/4 \\ &\leq \frac{3}{4}\varepsilon \end{aligned}$$

we find (20). The second assertion can be shown as in the proof of Proposition 35. ∎

3.6 Proofs of the Main Theorems

In this final subsection we prove our main results including the lower bounds on $\mathcal{S}_{L,P}$. We begin with:

Proof of Lemma 1 Let H be the RKHS of k and $\Phi : X \to H$ be the associated feature map, i.e. $\Phi(x) = k(x, .)$, $x \in X$. Obviously, we only have to show that $\Phi(x_1), \ldots, \Phi(x_n)$ are linearly independent in H if and only if $x_1, \ldots, x_n$ are mutually different. Let us suppose that $x_1, \ldots, x_n$ are

mutually different but $\Phi(x_1),\ldots,\Phi(x_n)$ are linearly dependent. Then we may assume without loss of generality that there exists coefficients $\lambda_1,\ldots,\lambda_{n-1}\in\mathbb{R}$ with

$$\Phi(x_n) \;=\; \sum_{i=1}^{n-1}\lambda_i\Phi(x_i)\;.$$

Since k is universal there exists an element $w\in H$ with $\langle w,\Phi(x_n)\rangle<0$ and $\lambda_i\langle w,\Phi(x_i)\rangle\geq 0$ for all $i=1,\ldots,n-1$ (see Steinwart, 2001, Cor. 6). From this we easily get a contradiction. The other implication is trivial. ∎

Proof of Theorem 9 For brevity's sake we only prove the assertion in the case of $0\in\partial_2 L(1,F_L^*(1/2))\cap\partial_2 L(-1,F_L^*(1/2))$. The proof of the other case follows the same line but is slightly less technical. Obviously, it suffices to show the assertion for small $\varepsilon>0$. By Lemma 19 we find an $\varepsilon\in(0,1)$ with

$$0\notin\partial_2 L(1,t+\varepsilon B_{\mathbb{R}})\cap\partial_2 L(-1,t+\varepsilon B_{\mathbb{R}}) \tag{22}$$

for all $t\in F_L^*(1/2)+\varepsilon B_{\mathbb{R}}$. Moreover, we fix a $\delta\in(0,\varepsilon)$ with $P_X(S_\delta)\geq P_X(S)-\varepsilon/2$, where S_δ is the approximation of S defined in Lemma 21. Let us define

$$\begin{aligned} X_{n,\delta}^+ &:= \Big\{x\in X_0\cap X_{cont} : 0\notin\partial_2 L\big(1,\, f_{P,\lambda_n}(x)+\delta B_{\mathbb{R}}\big)\Big\}\\ X_{n,\delta}^- &:= \Big\{x\in X_0\cap X_{cont} : 0\notin\partial_2 L\big(-1,\, f_{P,\lambda_n}(x)+\delta B_{\mathbb{R}}\big)\Big\}\end{aligned}$$

for all $n\geq 1$. With the help of (22) we immediately obtain $(X_0\cap X_{cont})\setminus E(f_{P,\lambda_n},\delta)\subset X_{n,\delta}^+\cup X_{n,\delta}^-\subset X_0\cap X_{cont}$. Therefore, by Theorem 22 and some results of Steinwart (2003a) we find

$$P_X(X_{n,\delta}^+\cup X_{n,\delta}^-)\geq P_X(X_0\cap X_{cont})-\varepsilon/2$$

for all sufficiently large integers n. Hence, by the definition of δ we have

$$P_X(S_\delta)+\frac{1}{2}P_X(X_{n,\delta}^+\cup X_{n,\delta}^-)\geq\mathcal{S}_{L,P}-\frac{3}{4}\varepsilon \tag{23}$$

for all sufficiently large n. In order to consider "representative" training sets we define

$$\mathcal{C}_{T,\delta}:=\operatorname{card}\Big\{i:(x_i,y_i)\in S_\delta\setminus(E_n(\delta)\times Y)\text{ or }(x_i,y_i)\in X_{n,\delta}^+\times\{1\}\text{ or }(x_i,y_i)\in X_{n,\delta}^-\times\{-1\}\Big\}$$

for all training sets $T=((x_1,y_1),\ldots,(x_n,y_n))$, $n\geq 1$, where $E_n(\delta)$ are sets according to Proposition 35. Our above considerations together with Proposition 33, (23) and Hoeffding's inequality yield

$$\Pr^*\Big(T\in(X\times Y)^n:\mathcal{C}_{T,\delta}\geq(\mathcal{S}_{L,P}-\varepsilon)\,n, E(f_{T,\lambda_n},\delta)\subset E_n(\delta)\text{ and }\big\|f_{T,\lambda_n}-f_{P,\lambda_n}\big\|_\infty\leq\delta\Big)\to 1$$

for $n\to\infty$. Therefore, let us consider a training set T with $E(f_{T,\lambda_n},\delta)\subset E_n(\delta)$ and a sample (x_i,y_i) of T with $(x_i,y_i)\in S_\delta$ and $x_i\notin E_n(\delta)$. Then we have $x_i\notin E(f_{T,\lambda_n},\delta)$ and the definition of S_δ leads to

$$0\notin\partial_2 L\Big(y_i,F_L^*\big(P(1|x_i)\big)\cap\mathbb{R}+\delta B_{\mathbb{R}}\Big)\;.$$

Furthermore, the definition of $E(f_{T,\lambda_n},\delta)$ ensures

$$f_{T,\lambda_n}(x_i) \in F_L^*(P(1|x_i)) \cap \mathbb{R} + \delta B_{\mathbb{R}}$$

if $F_L^*(P(1|x_i)) \cap \mathbb{R} \neq \emptyset$. Hence we find $0 \notin \partial_2 L(y_i, f_{T,\lambda_n}(x_i))$ in this case, i.e. x_i is a support vector of the representation of f_{T,λ_n} as discussed in Remark 32. Moreover, since $x_i \in X_{cont}$ we also observe that the sample value of x_i occurs P^n-almost surely only once in T. Therefore, x_i is even P^n-almost surely a support vector in all minimal representations of f_{T,λ_n}. If $F_L^*(P(1|x_i)) \cap \mathbb{R} = \emptyset$ we have either $P(1|x_i) = 0$ or $P(1|x_i) = 1$ by Lemma 20 and the admissibility of L. Without loss of generality we may assume $P(1|x_i) = 1$. Then we have $F_L^*(1) = \{\infty\}$ and hence $0 \notin \partial_2 L(1,\mathbb{R})$. Therefore, the sample x_i is P^n-almost surely a support vector in all minimal representations whenever $y_i = 1$. The latter is P^n-almost surely fulfilled since $P(1|x_i) = 1$.
Now, let $(x_i,y_i) \in X_{n,\delta}^+ \times \{1\}$ be a sample of a training set T with $\|f_{T,\lambda_n} - f_{P,\lambda_n}\|_\infty \leq \delta$. Then we observe $f_{T,\lambda_n}(x_i) \in f_{P,\lambda_n}(x_i) + \delta B_{\mathbb{R}}$ and hence $0 \notin \partial_2 L(y_i, f_{T,\lambda_n}(x_i))$ by the definition of $X_{n,\delta}^+$. Again this shows that x_i is P^n-almost surely a support vector in all minimal representations of f_{T,λ_n}. Since the same argument can be applied for samples $(x_i,y_i) \in X_{n,\delta}^- \times \{-1\}$ we have shown the assertion. ■

Proof of Theorem 10 If P is a probability measure with

$$P_X\Big(x \in X : P(1|x) \notin \{0,1/2,1\}\Big) > 0 \tag{24}$$

the proof is analogous to the proof of Theorem 9 using Proposition 38 instead of Propositions 33 and 35. Therefore, let us suppose that (24) does not hold. In order to avoid technical notations we may also assume $X = X_{cont}$ without loss of generality. Furthermore, if $0 \notin \partial_2 L(Y,\mathbb{R})$ every sample $x_i \in X_{cont}$ of a training set $T \in (X \times Y)^n$ is P^n-a.s. a support vector in all minimal representations. Since $\mathcal{S}_{L,P} = P_X(X_{cont}) = 1$ the assertion is then a simple exercise. If $0 \in \partial_2 L(Y,\mathbb{R})$ we first assume that $0 \in \partial_2 L(1,\mathbb{R}) \cap \partial_2 L(-1,\mathbb{R})$. Then we have $S \subset X_0 \times Y$ P-almost surely and therefore samples $x_i \notin X_0$ can be neglected. Hence we may assume without loss of generality that $P_X(X_0) = 1$. In order to motivate the following construction let us first recall that we cannot control the behaviour of $\tilde{b}_{T,\lambda}$ in our situation. This makes it more difficult to define a subset $\tilde{X}_\varepsilon$ of X_0 such that *a)* $\tilde{X}_\varepsilon$ is "essentially" independent of T and *b)* $\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}$ maps into $F_L^*(1/2) + \varepsilon B_{\mathbb{R}}$ on $\tilde{X}_\varepsilon$.
Therefore, our first step is to construct such a set $\tilde{X}_\varepsilon$: for measurable $f : X \to \mathbb{R}$ and $\varepsilon,\delta > 0$ we define

$$\begin{aligned}
\overline{b}_{\varepsilon,\delta}(f) &:= \sup\Big\{b \in \mathbb{R} : P_X\big(x \in X : f(x) + b > \max F_L^*(1/2) + \varepsilon\big) \leq \delta\Big\} \\
\underline{b}_{\varepsilon,\delta}(f) &:= \inf\Big\{b \in \mathbb{R} : P_X\big(x \in X : f(x) + b < \min F_L^*(1/2) - \varepsilon\big) \leq \delta\Big\} .
\end{aligned}$$

It is easily checked that the supremum in the above definition is actually a maximum, i.e.

$$P_X\big(x \in X : f(x) + \overline{b}_{\varepsilon,\delta}(f) > \max F_L^*(1/2) + \varepsilon\big) \leq \delta . \tag{25}$$

The same holds for the infimum, i.e.

$$P_X\big(x \in X : f(x) + \underline{b}_{\varepsilon,\delta}(f) < \min F_L^*(1/2) - \varepsilon\big) \leq \delta . \tag{26}$$

Furthermore, we define

$$X_{\varepsilon,\delta}(f) := \Big\{x \in X : f(x) + \overline{b}_{\varepsilon,\delta}(f) \leq \max F_L^*(1/2) + \varepsilon \text{ and } f(x) + \underline{b}_{\varepsilon,\delta}(f) \geq \min F_L^*(1/2) - \varepsilon\Big\} .$$

Inequalities (25) and (26) yield

$$P_X(X_{\varepsilon,\delta}(f)) \;\geq\; 1-2\delta\,. \tag{27}$$

Moreover, if we have two bounded measurable functions $f,g : X \to \mathbb{R}$ with $\|f-g\|_\infty \leq \varepsilon$ we easily check

$$\begin{aligned} \overline{b}_{\varepsilon,\delta}(g)-\varepsilon &\leq \overline{b}_{\varepsilon,\delta}(f) \leq \overline{b}_{\varepsilon,\delta}(g)+\varepsilon \qquad (28)\\ \underline{b}_{\varepsilon,\delta}(g)-\varepsilon &\leq \underline{b}_{\varepsilon,\delta}(f) \leq \underline{b}_{\varepsilon,\delta}(g)+\varepsilon\,. \qquad (29) \end{aligned}$$

We find (see Steinwart, 2003a) that $\mathcal{R}_{L,P}(\tilde{f}_{T,\lambda_n}+\tilde{b}_{T,\lambda_n}) \to \mathcal{R}_{L,P}$ in probability for $n\to\infty$. Then Theorem 22 states that for all $\varepsilon>0$ and all $\delta>0$ we have

$$P^n\Big(T\in(X\times Y)^n : P_X\big(E(\tilde{f}_{T,\lambda_n}+\tilde{b}_{T,\lambda_n},\varepsilon)\big)\leq\delta\Big) \;\to 1 \tag{30}$$

for $n\to\infty$. Now, let us assume that we have a training set T of length n with

$$P_X\big(E(\tilde{f}_{T,\lambda_n}+\tilde{b}_{T,\lambda_n},\varepsilon)\big)\leq\delta \tag{31}$$

and

$$\|\tilde{f}_{T,\lambda_n}-\tilde{f}_{P,\lambda_n}\|_\infty \leq \varepsilon\,. \tag{32}$$

Recall, that the probability of such T also converges to 1 by Proposition 37. Then (31) yields $\underline{b}_{\varepsilon,\delta}(\tilde{f}_{T,\lambda_n}) \leq \tilde{b}_{T,\lambda_n} \leq \overline{b}_{\varepsilon,\delta}(\tilde{f}_{T,\lambda_n})$. By (28), (29) and (32) we hence find

$$\tilde{f}_{T,\lambda_n}(x)+\tilde{b}_{T,\lambda_n} \;\in\; F_L^*(1/2)+3\varepsilon B_{\mathbb{R}} \tag{33}$$

for all $x\in X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n})$, i.e. $X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n})$ is our desired set mentioned at the beginning. If $0\notin \partial_2 L(1,F_L^*(1/2))\cap\partial_2 L(-1,F_L^*(1/2))$ the rest of the proof is more or less canonical: fix a small $\delta>0$ and choose an $\varepsilon>0$ with $P(S_\varepsilon)\geq \mathcal{S}_{L,P}-\delta$. Then, consider only training sets T which are "representative on $X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n})\cap S_\varepsilon$ up to δ" and which fulfill both (31) and (32). For these T we find P^n-almost surely $\#SV(\tilde{f}_{T,\lambda_n}) \geq (\mathcal{S}_{L,P}-4\delta)n$.

As in the proof of Theorem 9 technical problems arise in the case of

$$0\in\partial_2 L(1,F_L^*(1/2))\cap\partial_2 L(-1,F_L^*(1/2))\,. \tag{34}$$

Even worse, the techniques used there cannot be applied in our situation since we cannot control the behaviour of $\tilde{b}_{T,\lambda_n}$. The key idea for solving these difficulties is the observation that for $t\in F_L^*(1/2)$ the subdifferentials $\partial_2 L(y,t)$ can only contain 0 at the boundary of $F_L^*(1/2)$ (cf. Lemma 20). Since we only have to prove the assertion for small $\varepsilon>0$ we fix an $\varepsilon>0$ with $\varepsilon < \big(\max F_L^*(1/2) - \min F_L^*(1/2)\big)/4$. Recall, that such ε actually exist by our assumption (34) and Lemma 19. For $\delta>0$ and $n\geq n_0$ we define

$$\begin{aligned} X^+_{\varepsilon,\delta,n} &:= \Big\{x\in X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) : f(x)+\overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) \in \max F_L^*(1/2)+\varepsilon B_{\mathbb{R}}\Big\}\\ X^-_{\varepsilon,\delta,n} &:= \Big\{x\in X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) : f(x)+\underline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) \in \min F_L^*(1/2)+\varepsilon B_{\mathbb{R}}\Big\}\\ X^0_{\varepsilon,\delta,n} &:= X_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n})\setminus\big(X^+_{\varepsilon,\delta,n}\cup X^-_{\varepsilon,\delta,n}\big)\,. \end{aligned}$$

Furthermore let us assume that we have a training set T of length n with $\|\tilde{f}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n}\|_\infty < \varepsilon/3$ and $P_X\big(E(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}, \varepsilon)\big) \leq \delta$. Let us suppose that we have a sample (x_i, y_i) of T with $x_i \in X^+_{\varepsilon,\delta,n}$. If $\tilde{b}_{T,\lambda_n} \geq \overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) - 2\varepsilon$ we get

$$\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} \;\geq\; \tilde{f}_{P,\lambda_n}(x) + \overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) - 3\varepsilon \;\geq\; \max F^*_L(1/2) - 4\varepsilon$$

and hence we find $\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} \in \max F^*_L(1/2) + 4\varepsilon B_{\mathbb{R}}$ by (33). Since $\min F^*_L(1/2) \notin \max F^*_L(1/2) + 4\varepsilon B_{\mathbb{R}}$ by the choice of ε the sample x_i is P^n-a.s. a support vector in all minimal representations of $\tilde{f}_{T,\lambda_n}$ if $y_i = -1$ (cf. Lemma 20). If $\tilde{b}_{T,\lambda_n} < \overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) - 2\varepsilon$ we find

$$\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} \;<\; \tilde{f}_{P,\lambda_n}(x) + \overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) - \varepsilon \;\leq\; \max F^*_L(1/2)\,.$$

Therefore, x_i is P^n-a.s. a support vector in all minimal representations of $\tilde{f}_{T,\lambda_n}$ if $y_i = 1$. Obviously, analogous considerations can be made for samples in $X^-_{\varepsilon,\delta,n}$. Finally, for a sample $x_i \in X^0_{\varepsilon,\delta,n}$ we obtain

$$\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} \;<\; \tilde{f}_{P,\lambda_n}(x) + \overline{b}_{\varepsilon,\delta}(\tilde{f}_{P,\lambda_n}) + \frac{2}{3}\varepsilon \;<\; \max F^*_L(1/2) - \varepsilon/3$$

and therefore x_i is P^n-a.s. a support vector of a minimal representation of $\tilde{f}_{T,\lambda_n}$ if $y_i = 1$. With the above considerations the proof can be finished as in the case $0 \notin \partial_2 L(1, F^*_L(1/2)) \cap \partial_2 L(-1, F^*_L(1/2))$.

Finally, the remaining case $0 \in \partial_2 L(Y, \mathbb{R})$ with $0 \notin \partial_2 L(1, \mathbb{R}) \cap \partial_2 L(-1, \mathbb{R})$ can be treated similarly to our considerations in the case $0 \in \partial_2 L(1, \mathbb{R}) \cap \partial_2 L(-1, \mathbb{R})$ with $0 \notin \partial_2 L(1, F^*_L(1/2)) \cap \partial_2 L(-1, F^*_L(1/2))$. ■

Proof of Proposition 11 It is easy to see that the assertion is a simple consequence of $0 \notin \partial_2 L\big(1, F^*_L(\alpha)\big) \cap \partial_2 L\big(-1, F^*_L(\alpha)\big)$ for all $\alpha \neq 1/2$ (cf. Lemma 20). ■

Proof of Proposition 12 In order to prove the assertion it suffices to show

$$0 \;\notin\; \partial_2 L(-1, F^*_L(\alpha) \cap \mathbb{R}) \;\cup\; \partial_2 L(1, F^*_L(\alpha) \cap \mathbb{R}).$$

for all $\alpha \in (0,1)$. Let us assume the converse, i.e. that there exists an $\alpha \in (0,1)$, a $y \in Y$ and a $t \in F^*_L(\alpha) \cap \mathbb{R}$ with $0 \in \partial_2 L(y,t)$. Without loss of generality we may assume $y = 1$. Since L is differentiable we have $\partial_2 L(1,t) = \{0\}$. Hence $0 \in \partial_2 C(\alpha, t)$ implies $0 \in \partial_2 L(-1,t)$ which contradicts Lemma 19. ■

Proof of Example 2 Due to space limitations we only sketch the proof: let $(\tilde{f}_{T,\lambda_n}, \tilde{b}_{T,\lambda_n})$ be a solution of (2) with a representation

$$\tilde{f}_{T,\lambda_n} \;=\; \sum_{i=1}^n y_i \alpha_i k(x_i, .)$$

found by solving the dual problem of (2) (see Cristianini and Shawe-Taylor, 2000, Ch. 6). Since $X_{cont} = X$ this representation is almost surely minimal. Furthermore, we have

$$0 \;=\; \sum_{i=1}^n y_i \alpha_i \;=\; \sum_{(x_i,y_i) \in X^1_1} \alpha_i \;-\; \sum_{(x_i,y_i) \in X^{-1}_1} \alpha_i \;+\; \sum_{(x_i,y_i) \in X^1_{-1}} \alpha_i \;-\; \sum_{(x_i,y_i) \in X^{-1}_{-1}} \alpha_i\,.$$

Without loss of generality we may assume $P(X_1^{-1}) \geq P(X_{-1}^1)$ and $\mathcal{R}_{L,P} > 0$. We fix a $\rho \in (0,1/3)$. Let us assume that we have a training set T that is representative on X_i^j, $i,j \in \{-1,1\}$ up to ρ and additionally satisfies both $P_X\big(E(\tilde{f}_{P,\lambda_n} + \tilde{b}_{P,\lambda_n}, \rho)\big) \leq \rho$ and $\|\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n} - \tilde{b}_{P,\lambda_n}\|_\infty \leq \rho$. Recall, that the probability of such training sets converge to 1 by Proposition 38. Then Remark 32 for the L1-SVM yield

$$\sum_{(x_i,y_i)\in X_1^{-1}} \alpha_i \;\geq\; n\big(P(X_1^{-1}) - \rho\big)\frac{1}{2\lambda_n n} \;\geq\; \frac{1}{2\lambda_n}\big(P(X_1^{-1}) - \rho\big)\,.$$

Analogously we find

$$\sum_{(x_i,y_i)\in X_{-1}^1} \alpha_i \;\leq\; n\big(P(X_{-1}^1) + \rho\big)\frac{1}{2\lambda_n n} \;\leq\; \frac{1}{2\lambda_n}\big(P(X_{-1}^1) + \rho\big)\,.$$

Together, both estimates almost surely lead to

$$\begin{aligned}
\frac{1}{2\lambda_n}\big(P(X_1^{-1}) - P(X_{-1}^1) - 2\rho\big) \;&\leq\; \sum_{(x_i,y_i)\in X_1^1} \alpha_i \;-\; \sum_{(x_i,y_i)\in X_{-1}^{-1}} \alpha_i \\
&\leq\; \sum_{\substack{(x_i,y_i)\in X_1^1 \\ \alpha_i > 0}} \alpha_i \\
&\leq\; \frac{1}{2\lambda_n}\mathrm{card}\,\big\{i : (x_i,y_i) \in X_1^1 \text{ is a support vector }\big\}\,.
\end{aligned}$$

Since up to ρn exceptions all samples in $X_1^{-1} \cup X_{-1}^1$ are support vectors the assertion then easily follows. ∎

Acknowledgments

I thank M.C. Wu for drawing the illustrations. Furthermore, I thank J.C. Scovel and D.R. Hush for carefully reading the revised introduction. This work was financially supported by the DFG grant *Ca 179/4-1*.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, University of California, Berkeley. Available online at `http://stat-www.berkeley.edu/tech-reports/638.pdf`, 2003.

C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York, 1984.

D.D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, 18:1676–1695, 1990.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

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

T. Downs, K.E. Gates, and A. Masters. Exact simplification of support vector solutions. *Journal of Machine Learning Research*, 2:293–297, 2001.

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

G.S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

R.R. Phelps. *Convex Functions, Monotone Operators and Differentiability*, volume 1364 of *Lecture Notes in Math.* Springer, 1986.

K. Ritter. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Math.* Springer, 2000.

R.T. Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture Notes in Math.*, pages 157–207, 1976.

G. Romano. New results in subdifferential calculus with applications to convex optimization. *Appl. Math. Optim.*, 32:213–234, 1995.

S. Saitoh. *Integral transforms, reproducing kernels and their applications*. Addison Wesley Longman, 1997.

B. Schölkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, volume 2111 of *Lecture Notes in Artificial Intelligence*, pages 416–426, 2001.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

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

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

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

I. Steinwart. Consistency of support vector machines and other regularized kernel machine. *IEEE Transactions on Information Theory*, accepted with minor revisions, 2003a.

I. Steinwart. On the optimal parameter choice for ν-support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1274–1284, 2003b.

I. Steinwart. Sparseness of support vector machines—some asymptotically sharp bounds. To appear in S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004. Available online at `http://www.c3.lanl.gov/~ingo/publications/nips-03.ps`.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

V. Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Math.* Springer, 1995.

T. Zhang. Convergence of large margin separable linear classification. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32, 2004.