Sparseness of Support Vector Machines

Ingo Steinwart*
Mathematisches Institut
Friedrich-Schiller-Universität
Ernst-Abbe-Platz 1-4
07743 Jena, Germany
steinwart@minet.uni-jena.de

January 27, 2003

Abstract

Support vector machines (SVM's) 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 SVM's. 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 SVM's using subdifferential calculus in the associated reproducing kernel Hilbert space.

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

1 Introduction and results

Given a sequence of pairs $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$, where X is a set and $Y = \{-1, 1\}$, the aim in binary classification is to predict the *label* y of a new unseen pair $(x, y) \in X \times Y$. The basic assumption in one of the most common models is that the *training set* T consists of i.i.d. pairs which are generated by an *unknown* distribution P on $X \times Y$ (cf. e.g. [4] for a thorough treatment). In order to predict new labels a (measurable) decision function $f_T : X \to \mathbb{R}$ is constructed by certain algorithms—the so-called *classifiers*. The prediction of f_T for the label of (x, y) is then $\operatorname{sign} f_T(x)$.

We will assume throughout this work that X is a compact topological Hausdorff space and P is a Borel probability measure on $X \times Y$, where Y is equipped with the discrete topology. The type of classifiers that we shall 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

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

respectively. Here, $T = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (X \times Y)^n$ 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 loss function (cf. the following section for precise definitions). The additional term b in

*Research was supported by the DFG grant *Ca 179/4*.

(2) is called *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) (cf. the following section for their existence). Various recently proposed algorithms including *regularization networks* and several variants of *support vector machines (SVM's)* belong to this type of classifiers.

As shown in [13], [17] and [15] these classifiers can "learn" under specific conditions on L, H and the behaviour of $\lambda = \lambda_n$. Here "learning" means that the probability for misclassifying a new sample (x, y) generated by P tends to the smallest possible value. To make this precise the *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 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. The works [13], [17] and [15] establish conditions under which the classifiers based on (1) and (2) are universally consistent.

In order to formulate our results, recall that by the well-known representer theorem (cf. [10] for the most general form) 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 and $T = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$. 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*. Moreover, for a function $f \in H$ the *minimal number of support vectors* is defined by

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

A representation of f is called *minimal* if it has $\#SV(f)$ support vectors. The next lemma characterizes minimal representations (cf. the following section for definitions and Section 3.6 for a proof):

Lemma 1.1 *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, considering the dual optimization problems for the hinge loss or the squared hinge loss leads to algorithms which do not construct minimal representation 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 summarize all coefficients α_i with $x_i = x$ and call the sum α'_l. Then chose 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.

We also have to define a quantity depending on L and P which gives a suitable lower bound for $\#SV$: for this we write $C(\alpha, t) := \alpha L(1, t) + (1 - \alpha) L(-1, t)$ 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 $f : X \to \overline{\mathbb{R}}$, namely

$$\mathcal{R}_{L,P}(f) \;:=\; \mathbb{E}_{(x,y)\sim P} L(y, f(x)) \;=\; \int_X C\big(P(1|x), f(x)\big)\, P_X(dx) \ .$$

Here, P is a Borel probability measure on $X \times Y$, P_X is the marginal distribution of P on X and $P(y|x)$ denotes a regular conditional probability (cf. also the next section). Therefore, in order to minimize the L-risk we have to minimize the function $C(\alpha, .)$ for every $\alpha \in [0,1]$. This leads to

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

for all $\alpha \in [0,1]$. Obviously, given a measurable selection f^* of F_L^* the function $f^*(P(1|.))$ actually minimizes the L-risk, i.e.

$$\mathcal{R}_{L,P}\big(f^*(P(1|.))\big) \;=\; \mathcal{R}_{L,P} \;:=\; \inf\{\mathcal{R}_{L,P}(f) \mid f : X \to \overline{\mathbb{R}} \text{ measurable}\} \;.$$

Moreover, it turns out that for all sequences of measurable functions $f_n : X \to \mathbb{R}$ with $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}_{L,P}$ we obtain that (f_n) tends to $F_L^*(P(1|.))$ in a certain sense (cf. Theorem 3.9 for details). In particular, this convergence holds for the solutions of (1) and (2) if k is a universal kernel and the regularization sequence (λ_n) converges "slowly enough" to 0. The latter was already claimed in [6] and [17] in order to explain the learning ability of SVM's.

Furthermore, we will see that for convex L the subdifferential calculus yields a representation (3) of $f_{T,\lambda}$ with

$$\alpha_i \in -\frac{1}{2n\lambda}\partial_2 L(y_i, f_{T,\lambda}(x_i)) \tag{4}$$

for all $i = 1, \ldots, n$. Here, $\partial_2 L$ denotes the subdifferential operator of L with respect to the second variable. Therefore, if $\partial_2 L(y_i, f_{T,\lambda}(x_i))$—i.e. approximately $\partial_2 L\big(y, F_L^*(P(1|x_i)) \cap \mathbb{R}\big)$—does not contain 0, the sample x_i is a support vector whenever its sample value occurs only once in T. Continuing our motivation the previous considerations lead to

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

where $X_{cont} := \{x \in X : P_X(\{x\}) = 0\}$. Now, for a convex loss function L and a Borel probability measure P on $X \times Y$ we can define

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

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 3.7). Recall, that 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.

In order to state our results we need some further technical notations: for a loss function L and a kernel k 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 $K := \sup\{\sqrt{k(x,x)} : x \in X\}$. Furthermore, every convex loss function L is locally 1-Hölder-continuous. In this case we denote the 1-Hölder constant of L_λ by $|L_\lambda|_1$ (cf. the following section for a precise definition).

Now, we are in a position to formulate our results. The first theorem treats classifiers based on (1). Its proof as well as the proofs of the following results can be found in Section 3.6.

Theorem 1.2 *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.

The next theorem establishes an analogous result for classifiers based on (2). Because of the offset we have to exclude *degenerated* probability measures P, i.e. measures with

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

for $y = 1$ or $y = -1$. 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 1.3 *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 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 1.3. This, together with the fact that every loss function of practical interest (cf. the examples below) is strongly admissible let us state the above theorem in the present form, only.

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

Proposition 1.4 *Let L be a convex admissible loss function and P be a Borel probability measure on $X \times Y$. Then we have*

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

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

Roughly speaking, the above result together with Theorem 1.2 and Theorem 1.3 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 by many experiments that the sparseness of SVM's heavily depends on the noise of the underlying distribution. The next proposition improves the lower bound on $\mathcal{S}_{L,P}$ for differentiable loss function:

Proposition 1.5 *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 1.5 also throws new light on the role of the margin in SVM's: 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 1.6 *L1-SVM's 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 1.2 reduce to $\lambda_n \to 0$ and $n\lambda_n^2 \to \infty$. Then, applying Proposition 1.5 yields lower bounds on the number of support vectors. In particular, the number of support vectors is asymptotically bounded from below by $n\mathcal{R}_P$ 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 $L1$-SVM's produce the same set of decision functions as the so-called ν-SVM's (cf. [11]). Furthermore, as shown in [14] 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 (cf. [11]) leads to our conjecture.

Example 1.7 *L1-SVM's with offset* are based on (2) and the hinge loss function. The corresponding conditions on (λ_n) of Theorem 1.3 can be unified to $\lambda_n \to 0$ and $n\lambda_n^3 \to \infty$. Of course, applying Proposition 1.5 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 1.3 we then obtain (cf. the end of Section 3.6 for a sketch of the proof)

$$\Pr^*\Big(T \in (X\times Y)^n : \#SV(\tilde{f}_{T,\lambda_n}) \geq \big(\mathcal{R}_{L,P} + |P(X_{-1}^1) - P(X_1^{-1})| - \varepsilon\big)n\Big) \to 1 \tag{5}$$

for $L1$-SVM's 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 (5) improves the result of Thereom 1.3. In the extremal cases $P(X_{-1}^1) = 0$ and $P(X_1^{-1}) = 0$ the lower bound in (5) becomes $2n\mathcal{R}_P$ which also corroborates our belief described in the previous example.

Example 1.8 *L2-SVM's 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 1.2 are $\lambda_n \to 0$ and $n\lambda_n^3 \to \infty$. The value of $\mathcal{S}_{L,P}$ can be estimated by Proposition 1.5.

Example 1.9 *L2-SVM's with offset* are based on the minimization problem (2) with the squared hinge loss function. The conditions on (λ_n) of Theorem 1.3 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 $n\lambda_n^4 \to \infty$ (cf. [15] for details). Again, the value of $\mathcal{S}_{L,P}$ can be estimated by Proposition 1.5.

Example 1.10 *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 $L2$-SVM's with offset. Furthermore, $\mathcal{S}_{L,P}$ is equal to the corresponding value for the squared hinge loss.

Example 1.11 *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 $L2$-SVM's without offset. Again, the value of $\mathcal{S}_{L,P}$ can be estimated by Proposition 1.5.

Example 1.12 *R1-SVM's 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 $0 \leq \varepsilon < 1$. Our results coincide with the results for the $L1$-SVM with or without offset, respectively.

Example 1.13 *R2-SVM's 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 $0 \leq \varepsilon < 1$. Our results coincide with the results for the $L2$-SVM with or without offset, respectively.

Example 1.14 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 3.19 we easily see that the lower bounds of Theorem 1.2 and Theorem 1.3 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 $\lambda > 0$.

The rest of this work is organized as follows: in Section 2 we introduce further notations and definitions. Section 3 which contains the (unfortunately very technical) proofs is divided into several subsections. In Subsection 3.1 we recall some known facts from the subdifferential calculus in Banach spaces. The following subsection establishes some useful results on convex admissible loss functions. In Subsection 3.3 we prove a result which in particular describes the convergence of f_{T,λ_n} and $\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}$ to $F_L^*(P(1|.))$. In the next subsection we first compute the subdifferential of $\mathcal{R}_{L,P}$. We then establish a result which describes the stability of the solutions f_{T,λ_n} and $\tilde{f}_{T,\lambda_n}$. In particular it turns out that *both* are *unique* for standard SVM's. We conclude this subsection by showing (4). In Subsection 3.5 we refine the results of Subsection 3.3 by showing that the described convergence is essentially independent of T for the solutions of (1) and (2). Finally, in the last subsection we prove our main results which were presented in the introduction.

2 Preliminaries

In the following let $\overline{\mathbb{R}} := [-\infty, \infty]$, $\mathbb{R}^+ := [0, \infty)$ and $\overline{\mathbb{R}}^+ := [0, \infty]$. Given two functions $g, h : (0, \infty) \to (0, \infty)$ we write $g \preceq h$ if there exists a constant $c > 0$ with $g(\varepsilon) \leq c\, h(\varepsilon)$ 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 (cf. [1] and [2, Ch. 3]) by H_k or simply H. For its closed unit ball we write B_H. Recall, that the *feature map* $\Phi : X \to H$, $x \mapsto k(x, .)$ fulfills $k(., .) = \langle \Phi(.), \Phi(.) \rangle_H$ 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 (cf. [12, Sect. 3]).

Recall, that for a given Borel probability measure P on $X \times Y$ there exists a map $x \mapsto P(\,.\,|x)$ from X into the set of all probability measures on Y such that P is the joint distribution of $(P(\,.\,|x))_x$ and of the marginal distribution P_X of P on X (cf. [5, Lem. 1.2.1.]).

Many important loss functions are not differentiable but convex. In order to treat these loss functions we recall the concept of subdifferentials:

Definition 2.1 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 in w is defined by

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

Given a subset A of H we often use the notation

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

For a geometric interpretation of subdifferentials we refer to [7, p. 6]. The following definition plays an important role in the investigation of subdifferentials:

Definition 2.2 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.

As shown in [15] it is important for ensuring universal consistency that $F_L^*(\alpha)$ only contains elements with a "correct" sign. This is formalized in the following definition:

Definition 2.3 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.* It is said to be an *admissible* loss function 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 card $F_L^*(\alpha) = 1$ for all $\alpha \in (0,1)$ with $\alpha \neq 1/2$.

As shown in [15] the admissibility of L is necessary and sufficient in order to get universally consistent classifiers based on (1). For classifiers based on (2) the admissibility is also necessary and sufficient apart from some technical conditions. Moreover, it was proved in [15] that for admissible loss functions there exists a measurable function $f^* : [0,1] \to \overline{\mathbb{R}}$ with $f^*(\alpha) \in F_L^*(\alpha)$ for all $\alpha \in [0,1]$.

An admissible loss function L is called *convex* if $L(y, .)$ is convex for $y = \pm 1$. A loss function is said to be *1-Hölder-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* 1-Hölder-continuous if $L_{|Y \times [-a,a]}$ is 1-Hölder-continuous for all $a > 0$. Recall, that convex loss functions are always locally 1-Hölder-continuous (cf. also Lemma 3.16). In order to treat classifiers that are based on (2) we need the following definition from [15]:

Definition 2.4 An admissible loss function L is called *regular* if L is locally 1-Hölder-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

$$\begin{aligned} \left| L_{|Y \times [-\gamma a, \gamma a]} \right|_1 \; &\leq \; c_\gamma \left| L_{|Y \times [-a,a]} \right|_1 && (6) \\ \left\| L_{|Y \times [-\gamma a, \gamma a]} \right\|_\infty \; &\leq \; c_\gamma \left\| L_{|Y \times [-a,a]} \right\|_\infty \; . && (7) \end{aligned}$$

Note, that convex admissible loss function are regular if (6) and (7) hold (cf. Subsection 3.2).

Given a RKHS H the *regularized L-risks* are defined by

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

for all $f \in H$, $b \in \mathbb{R}$ and all $\lambda > 0$. If P is an empirical measure with respect to $T \in (X \times Y)^n$ we write $\mathcal{R}_{L,T}(.)$, $\mathcal{R}^{reg}_{L,T,\lambda}(.)$ 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). The following lemmas show in particular that both optimization problems can be solved. The proofs can be found in [15].

Lemma 2.5 *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 $\|f_{P,\lambda}\| \leq \delta_\lambda$.

For classifiers based on (2) we have to exclude degenerated Borel probability measures in order to ensure that the offset is real:

Lemma 2.6 *Let L be a regular loss function and H be a RKHS of continuous functions. Then for all non-degenerated 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 $\|\tilde{f}_{P,\lambda}\| \leq \delta_\lambda$.

3 Proofs

3.1 Subdifferentials

In this part of the work 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 (cf. [7, Prop. 1.11.]):

Proposition 3.1 *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 $\|w^*\| \leq |F|_1$ 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. Its proof can be found in [7, Prop. 2.5]:

Proposition 3.2 *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. Its proof can be found in [7, Prop. 1.26]:

Proposition 3.3 *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 which can be found in e.g. [7, Thm. 3.16]:

Proposition 3.4 *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. A discussion of it can be found in [9]:

Proposition 3.5 *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 3.6 *Let L be a convex admissible loss function. Then L is locally 1-Hölder-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
$$0 \notin \partial_2 L(1,t) \cap \partial_2 L(-1,t) \ . \tag{8}$$

iii) for all bounded subsets $A \subset \mathbb{R}$ there exists an $\varepsilon > 0$ such that for all $t \in A$ we have
$$0 \notin \partial_2 L(1, t + \varepsilon B_{\mathbb{R}}) \cap \partial_2 L(-1, t + \varepsilon B_{\mathbb{R}}) \ . \tag{9}$$

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 (9) 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 3.2
$$0 \ \in \ \partial_2 L(1, t_n + \delta_n) \cap \partial_2 L(-1, t_n + \delta_n') \ \subset \ \big(\partial_2 L(1,t) + \varepsilon B_{\mathbb{R}}\big) \cap \big(\partial_2 L(-1,t) + \varepsilon B_{\mathbb{R}}\big)$$
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 (8). ■

The next lemma collects some important facts about the solution operator F_L^* for convex admissible loss functions L:

Lemma 3.7 *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*

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

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 3.6 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 3.6. 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 $s_{-1} \leq 0 \leq s_1$ or $s_1 \leq 0 \leq s_{-1}$. Since the monotony of the subdifferential and Lemma 3.6 yield that $s_1 \geq 0$ implies $t > 0$ and that $s_{-1} \leq 0$ 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 3.6 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 $0 \leq \alpha < \alpha' \leq 1$ 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 $t' \geq t$ since otherwise we observe $s' \leq s < 0$ 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 $t' \geq \max F_L^*(1/2) > t$ 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 $s \leq 0$. Therefore, by Lemma 3.6 *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 $\alpha \geq 1/2$ by the admissibility of L. Analogously, $0 \in \partial_2 L(-1,t')$ yields $\alpha \leq 1/2$. The last assertion is a direct consequence of Lemma 3.6 *ii)*. ■

Lemma 3.8 *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

$$\bigcap_{\varepsilon>0} \bigcup_{\delta\in[-\varepsilon,\varepsilon]} \bigcup_{t\in F_L^*(\alpha)\cap\mathbb{R}} \partial_2 L(y, t+\varepsilon) \subset \bigcup_{t\in F_L^*(\alpha)\cap\mathbb{R}} \partial_2 L(y,t) \tag{10}$$

for all $\alpha \in [0,1]$, $y = \pm 1$. If $F_L^*(\alpha) \cap \mathbb{R} = \emptyset$ inclusion (10) is trivial. Therefore, we assume $F_L^*(\alpha) \cap \mathbb{R} \neq \emptyset$. Let us fix an element h of the left set in (10). 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 (10) 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 3.2 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 the following "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 3.9 *Let P be a Borel probability measure on $X \times Y$ and L be a loss function with* card $F_L^*(\alpha) > 1$ *for at most countably many $\alpha \in [0,1]$. 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. Since in both cases $G_\varepsilon(\alpha)$ is closed in $\mathbb{R}$ there exists $f_*(\alpha) \in G_\varepsilon(\alpha) \cup \{\pm\infty\}$ 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. 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

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

Since our construction 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 to $\tilde{X}_\varepsilon$ are absolutely continuous to each other. Now, the assertion easily follows from $E \subset \tilde{X}_\varepsilon$. ■

Remark 3.10 The assumption 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.

Remark 3.11 As shown in [15] 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 3.9 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 card $F_L^*(\alpha) = 1$ hold for all $\alpha \in [0,1]$ then

$$\|f_{T,\lambda_n} - f_{L,P}\|_0 \to 0 \tag{11}$$

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 3.9 and in particular (11) 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 3.12 Theorem 3.9 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

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

where $\mathbf{1}_B$ 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 3.13 *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 in 0 and continuous in 0. Furthermore, for all $h \in L_2(P)$ we have*

$$\partial \mathcal{R}^\pm_{L,P}(h) \ = \big\{ h^* \in L_2(P^\pm_X)\ :\ h^*(x) \in \partial L(\pm 1, h(x))\ \ P^\pm_X\text{-}a.s \big\}\ . \tag{12}$$

Proof: We only have to consider $\mathcal{R}^+_{L,P}$. Using the notions of [8] we first observe that $L(1, .)$ is a normal and convex integrand (cf. [8, 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 (12) then follows by [8, Cor. 3E.].
In order to prove the continuity in 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

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

and hence $\limsup_{n\to\infty} \mathcal{R}^+_{L,P}(f_n) \leq \mathcal{R}^+_{L,P}(0)$. 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 $|a| \leq \varepsilon$. Therefore [8, Cor. 3D.] and [8, Prop. 3G.] yield the lower semi-continuity of $\mathcal{R}^+_{L,P}$ at 0 with respect to the weak topology of $L_2(P^+_X)$. 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 3.14 *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*

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

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

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

Now, $\mathcal{R}^{+}_{L,P} : L_2(P_X^+) \to [0,\infty]$ is continuous in 0. Hence, the chain rule of Proposition 3.5 together with Proposition 3.13 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 3.14 has already been presented in [16]. However, the claim therein that Proposition 3.14 can be proved using subdifferential calculus on *finite* dimensional spaces is obviously not correct. For differentiable loss functions Proposition 3.14 is more or less trivial.

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

Theorem 3.15 *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 3.15 we need the following simple lemmas which will not be proved:

Lemma 3.16 *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 3.17 *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 3.15: 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 3.14 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

$$0 \;=\; 2A\lambda\hat{f}_{P,\lambda} + \mathbb{E}_P h\Phi \;. \tag{13}$$

By the Lipschitz-continuity and Proposition 3.1 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 (13) 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 3.16 that L restricted to $Y \times [-c,c]$ is Lipschitz continuous and thus there exists a Lipschitz continuous extension $\tilde{L}$ according to Lemma 3.17. 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 3.18 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 $L2$-SVM's with offset since in these cases we have $\|\hat{f}_{P,\lambda}\|_\infty \le 2 + 2K\delta_\lambda$.

Remark 3.19 Equation (13) is a general form of the well-known representer theorem. Indeed, (13) 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 (4), 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 3.20 *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*

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

For the proof we will need the following result which is a reformulation of [18, Thm. 3.3.4]:

Lemma 3.21 *Let $\eta_1, \dots, \eta_n$ be bounded i.i.d. random variables with values in a Hilbert space H. Assume $\|\eta_i\|_\infty \leq M$ for all $i = 1, \dots, n$. 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 [18, Thm. 3.3.4] 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 3.20: By Theorem 3.15 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 3.21 to $\eta_i := h(x_i, y_i)\Phi(x_i)$, $i = 1, \dots, n$, and $M = K|L_\lambda|_1$ we thus obtain

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

■

With the help of Proposition 3.20 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 3.22 *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: For training sets T with $\|f_{T,\lambda_n} - f_{P,\lambda_n}\| \leq \frac{\varepsilon}{2K}$ we immediately obtain $E(f_{T,\lambda_n}, \varepsilon) \subset E(f_{P,\lambda_n}, \varepsilon/2)$. Since $n\lambda_n^2/|L_{\lambda_n}|_1 \to \infty$ Proposition 3.20 ensures that the probability of such training sets tends to 1. Finally, $\lambda_n \to 0$ yields $\mathcal{R}_{L,P}(f_{P,\lambda_n}) \to \mathcal{R}_{L,P}$ by [15, Prop. 3.2] and therefore, Theorem 3.9 shows that $E(f_{P,\lambda_n}, \varepsilon/2)$ are the desired sets for large n. ■

Remark 3.23 Proposition 3.22 also holds for convex loss functions satisfying the assumptions of Theorem 3.9.

In the rest of this section we show that Proposition 3.22 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 3.20:

Proposition 3.24 *Let L be a regular and convex loss function, H be a RKHS of a continuous kernel, and P be a non-degenerated 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 in [15, Lem. 2.6 and Lem. 5.2] 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

$$\Pr{}^*\big(T \in (X \times Y)^n : |\tilde{b}_{T,\lambda}| \leq \tilde{c} + \delta_\lambda K \text{ for all } \lambda > 0\big) \;\geq\; 1 - 2e^{-\tilde{c}n} \tag{14}$$

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 3.15 to the training sets considered in (14). This gives us a function $h : X \times Y \to \mathbb{R}$ with $\|h\|_\infty \leq |\tilde{L}_\lambda|_1$ and

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

Moreover, for the training sets considered in (14) 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 3.22 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 1.3.

Proposition 3.25 *Let L be a strongly admissible, regular and convex loss function, H be a RKHS of a universal kernel and P be a non-degenerated 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

$$\Pr{}^*\big(T \in (X \times Y)^n : \|\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n} - \tilde{b}_{P,\lambda_n}\|_\infty \leq \varepsilon\big) \to 1\;. \tag{15}$$

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

$$P^n\Big(T \in (X \times Y)^n : E(f_{T,\lambda_n}, \varepsilon) \subset E_n(\varepsilon)\Big) \to 1\;. \tag{16}$$

Proof: We define $\tilde{X} := \{x \in X : P(1|x) \notin \{0, 1/2, 1\}\}$ and fix an ε with $0 < \varepsilon < P_X(\tilde{X})$. Furthermore, for $\varepsilon/4$ we chose a $\delta > 0$ according to Theorem 3.9. 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 by [15, Cor. 3.19] and Proposition 3.24. 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 (15). The second assertion can be shown as in the proof of Proposition 3.22. ■

3.6 Proofs of the main theorems

Proof of Lemma 1.1: 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 $i = 1, \ldots, n-1$ (cf. [12, Cor. 6]). From this we easily get a contradiction. The other implication is trivial. ■

Proof of Theorem 1.2: 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 3.6 we find an $\varepsilon \in (0, 1)$ with

$$0 \notin \partial_2 L(1, t + \varepsilon B_\mathbb{R}) \cap \partial_2 L(-1, t + \varepsilon B_\mathbb{R}) \tag{17}$$

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 3.8. 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 (17) 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 3.9 and [15, Prop. 3.2] 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

$$P_X(S_\delta) + \frac{1}{2} P_X(X^+_{n,\delta} \cup X^-_{n,\delta}) \geq \mathcal{S}_{L,P} - \frac{3}{4}\varepsilon \tag{18}$$

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 3.22. Our above considerations together with Proposition 3.20, (18) and Hoeffding's inequality yield

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

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 3.19. 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 3.7 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 us consider a sample $(x_i, y_i) \in X^+_{n,\delta} \times \{1\}$ 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 1.3: If P is a probability measure with

$$P_X\Big(x \in X : P(1|x) \notin \{0, 1/2, 1\} \Big) > 0 \tag{19}$$

the proof is analogous to the proof of Theorem 1.2 using Proposition 3.25 instead of Propositions 3.20 and 3.22. Therefore, let us suppose that (19) 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

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

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

$$P_X\big(x \in X : f(x) + \overline{b}_{\varepsilon,\delta}(f) > \max F_L^*(1/2) + \varepsilon\big) \le \delta\,. \tag{20}$$

The same holds for the infimum, i.e.

$$P_X\big(x \in X : f(x) + \underline{b}_{\varepsilon,\delta}(f) < \min F_L^*(1/2) - \varepsilon\big) \le \delta\,. \tag{21}$$

Furthermore, we define

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

Inequalities (20) and (21) yield

$$P_X(X_{\varepsilon,\delta}(f)) \ \ge\ 1 - 2\delta\,. \tag{22}$$

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

$$\overline{b}_{\varepsilon,\delta}(g) - \varepsilon \ \le\ \overline{b}_{\varepsilon,\delta}(f) \ \le\ \overline{b}_{\varepsilon,\delta}(g) + \varepsilon \tag{23}$$

$$\underline{b}_{\varepsilon,\delta}(g) - \varepsilon \ \le\ \underline{b}_{\varepsilon,\delta}(f) \ \le\ \underline{b}_{\varepsilon,\delta}(g) + \varepsilon\,. \tag{24}$$

By [15, Lem. 3.18] we find $\mathcal{R}_{L,P}(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}) \to \mathcal{R}_{L,P}$ in probability for $n \to \infty$. Then Theorem 3.9 states that for all $\varepsilon > 0$ and all $\delta > 0$ we have

$$P^n\Big(T \in (X \times Y)^n : P_X\big(E(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}, \varepsilon)\big) \le \delta\Big) \ \to 1 \tag{25}$$

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

$$P_X\big(E(\tilde{f}_{T,\lambda_n} + \tilde{b}_{T,\lambda_n}, \varepsilon)\big) \le \delta \tag{26}$$

and

$$\|\tilde{f}_{T,\lambda_n} - \tilde{f}_{P,\lambda_n}\|_\infty \le \varepsilon\,. \tag{27}$$

Recall, that the probability of such T also converges to 1 by Proposition 3.24. Then (26) yields $\underline{b}_{\varepsilon,\delta}(\tilde{f}_{T,\lambda_n}) \le \tilde{b}_{T,\lambda_n} \le \overline{b}_{\varepsilon,\delta}(\tilde{f}_{T,\lambda_n})$. By (23), (24) and (27) we hence find

$$\tilde{f}_{T,\lambda_n}(x) + \tilde{b}_{T,\lambda_n} \ \in\ F_L^*(1/2) + 3\varepsilon B_{\mathbb{R}} \tag{28}$$

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 $P(S_\varepsilon) \ge \mathcal{S}_{L,P} - \delta$. 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 (26) and (27). For these T we find P^n-almost surely $\#SV(\tilde{f}_{T,\lambda_n}) \ge (\mathcal{S}_{L,P} - 4\delta)n$.
As in the proof of Theorem 1.2 technical problems arise in the case of

$$0 \in \partial_2 L(1, F_L^*(1/2)) \cap \partial_2 L(-1, F_L^*(1/2))\,. \tag{29}$$

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 3.7). 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 (29) and Lemma 3.6. 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 (28). 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 3.7). 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 hence 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 $0 \notin \partial_2 L(1, F_L^*(1/2)) \cap \partial_2 L(-1, F_L^*(1/2))$.
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 1.4: 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 3.7). ■

Proof of Proposition 1.5: In order to prove the assertion it suffices to show

$$0 \notin \partial_2 L(Y, F_L^*(\alpha) \cap \mathbb{R})$$

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 3.6. ■

Proof of Example 1.7: 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) (cf. [3, 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 3.25. Then Remark 3.19 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. ■

References

[1] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337-404

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

[3] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, 2000

[4] L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition", Springer, New York, 1997

[5] R.M. Dudley, Central limit theorems for empirical measures, *Ann. Probab.* **6** (1978), 899-929

[6] Y. Lin, Support vector machines and the Bayes rule in classification, *Data Mining and Knowledge Discovery* **6** (2002), 259-275

[7] R.R. Phelps, Convex Functions, Monotone Operators and Differentiability, *Lecture Notes in Math.* **1364** (1993)

[8] R.T. Rockafellar, Integral functionals, normal integrands and measurable selections, *Lecture Notes in Math.* **543** (1976), 157-207

[9] G. Romano, New results in subdifferential calculus with applications to convex optimization, *Appl. Math. Optim.* **32** (1995), 213-234

[10] B. Schölkopf, R. Herbrich, A.J. Smola, and R.C. Williamson, A generalized representer theorem, In "Proceedings of the 14th Annual Conference on Computational Learning Theory", *Lecture Notes in Artificial Intelligence* **2111** (2001), 416-426

[11] B. Schölkopf, A.J. Smola, R.C. Williamson and P.L. Bartlett, New support vector algorithms, *Neural Computation* **12** (2000), 1207-1245

[12] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, *Journal of Machine Learning Research* **2** (2001), 67-93

[13] I. Steinwart, Support vector machines are universally consistent, *J. Complexity* **18** (2002), 768-791

[14] I. Steinwart, On the optimal parameter choice for ν-support vector machines, submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*, available at `http://www.minet.uni-jena.de/~ingo/publications/pami-02.ps`

[15] I. Steinwart, Consistency of support vector machines and other regularized kernel machines, submitted to *IEEE Transactions on Information Theory*, available at `http://www.minet.uni-jena.de/~ingo/publications/info-02.ps`

[16] Tong 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", 357-363, MIT Press, 2001.

[17] Tong Zhang, Statistical behaviour and consistency of classification methods based on convex risk minimization, to appear in *Ann. Statist.*

[18] V. Yurinsky, Sums and Gaussian Vectors, *Lecture Notes in Math.* **1617** (1995)